Advanced Certification Programme in Data Science Business Analytics

E & I C T
IIT Guwahati

Week 18
Inferential Statics (part 1)

Topics Covered

- Sampling Distribution
- The Central Limit Theorem
- Hypothesis Testing Basics
- Hands-on Practice
- Q & A

Async Recap

1. Understand Python Basics
Learn how to install Python, explore code editors like VS Code or Jupyter, and begin writing and executing simple programmes.

2. Master Core Syntax and Logic
Grasp foundational elements such as data types, variables, and conditional statements to construct logical Python scripts.

3. Automate Using Loops and Functions
Use for and while loops to repeat tasks and define reusable logic using functions.

4. Work with Python Data Structures
Organise data efficiently using lists, tuples, sets, and dictionaries to store and retrieve information effectively.

5. Perform Data Analysis with NumPy and Pandas
Apply NumPy for numerical operations and use Pandas to clean, structure, and analyse datasets for meaningful insights.

Sampling Distribution

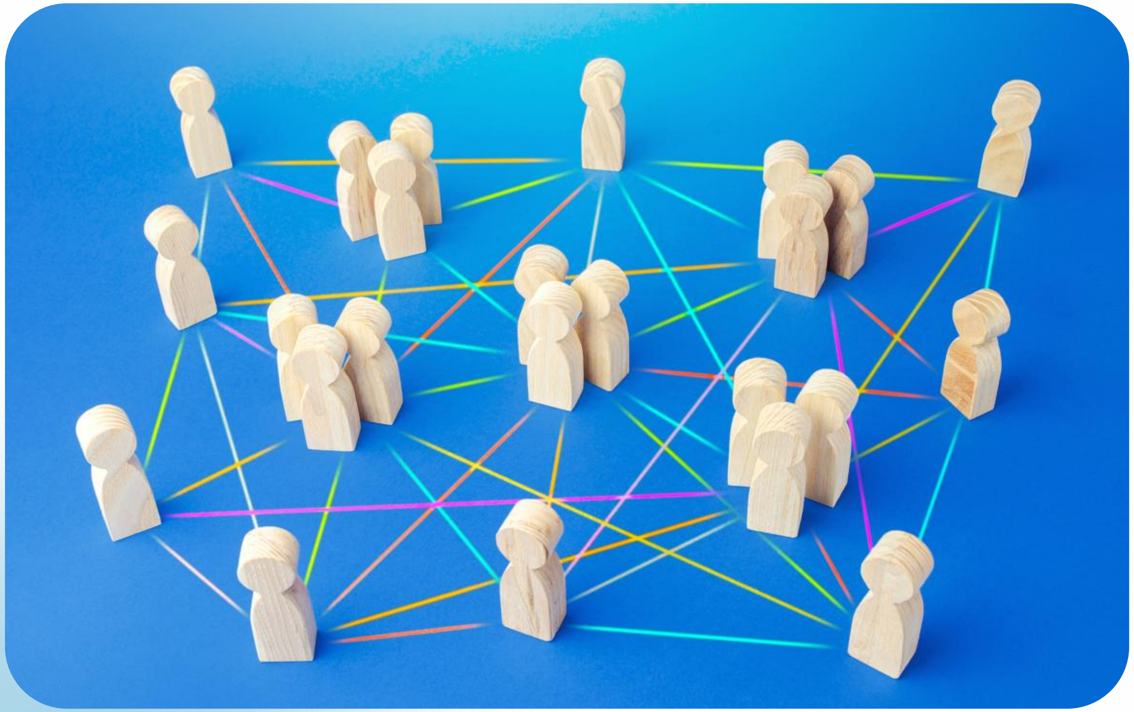
Understanding Sampling Distribution

Key Concept in Statistical Interference

- Shows the sampling distribution of a sample statistic (e.g., mean, proportion, variance)
- Based on repeated samples from the same population
- Measures variability in sample estimates
- Essential for confidence intervals and hypothesis testing

Key Concepts: Population vs Sample

Foundation for Statistical Inference



- **Population:** Entire group of interest
- **Sample:** Subset taken from the population
- **Parameter:** Value that describes the population (e.g. μ, σ)
- **Statistic:** Value calculated from a sample (e.g. mean $\bar{x}$)

Role of Sampling Distribution in Inference

Link Between Sample Data and Population Insight

Core idea

- Shows how a statistic varies from sample to sample

Applications

- Estimating population parameters
- Testing hypotheses about population characteristics

Visualisation

- Imagine plotting means from repeated samples to form a distribution

The Central Limit Theorem

Central Limit Theorem (CLT)

Normality of Sample Means with Large Samples

Distribution of sample means

- Sample means follow a normal distribution as sample size grows
- Applies even if the population is not normally distributed

Key formula:

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

- $\bar{X}$ = Sample mean
- μ = Population mean
- σ = Population standard deviation
- n = Sample size

Properties of the Central Limit Theorem

Predictable Behaviour of Sample Means

Mean convergence:

- The sample mean ($\bar{X}$) approaches the population mean (μ) as the sample size increases

Standard error:

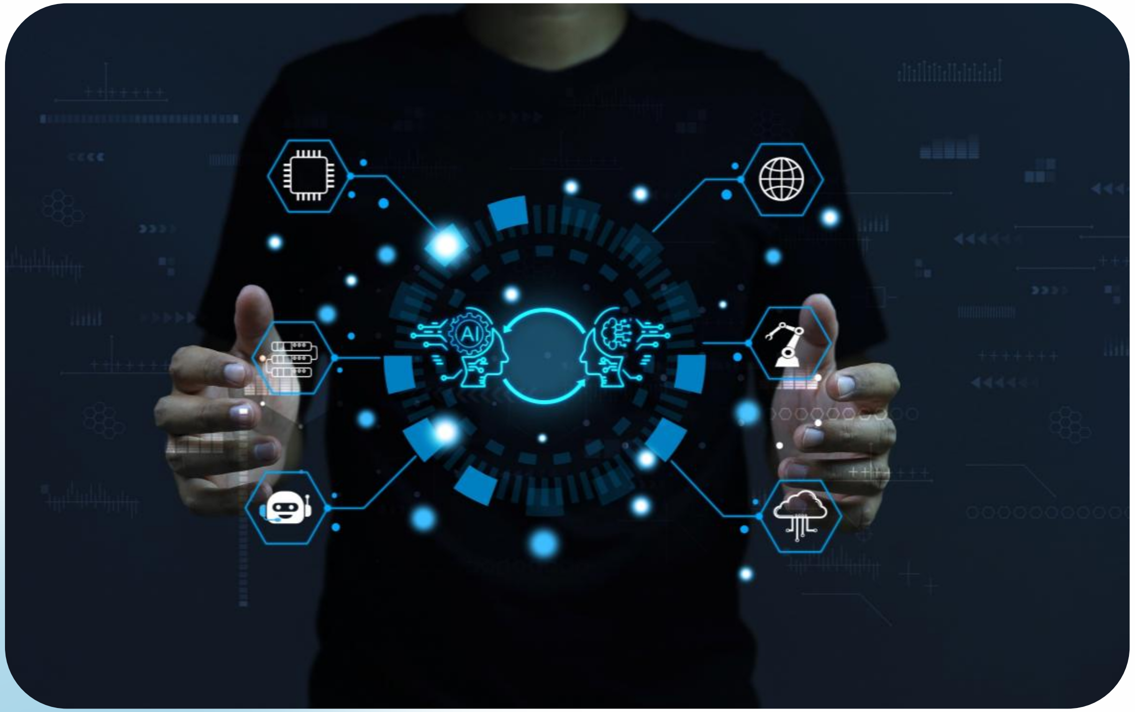
- The standard error of the mean (SEM) decreases as the sample size increases
- SEM = $\sigma / \sqrt{n}$

Confidence levels:

- ~68% within 1 SEM of μ
- ~95% within 2 SEM of μ
- ~99.7% within 3 SEM of μ

Conditions for Applying the Central Limit Theorem

When CLT Holds True



- Select randomly and independent of each other
- Select sample size with $n \geq 30$ for the Central Limit Theorem to apply
- Select population with finite variance
- Applies to both discrete and continuous distributions

Applications in Statistical Inference

Practical Uses of Sampling Distributions

- **Parameter estimation:** Estimate population parameters like mean or SD
- **Confidence interval construction:** Construct confidence intervals for likely value ranges
- **Hypothesis testing:** Conduct hypothesis tests (e.g. t-test, z-test) to assess the validity of statistical claims
- **Parametric test applicability:** Apply parametric tests even with non-normal population

Real-World Examples of CLT in Action

How CLT Powers Everyday Decisions

Election polling: Predict outcomes from voter samples, for randomly selected voters

Manufacturing quality control: Monitor sample means to ensure consistent production quality

Medical trials: Analyse sample means to assess drug efficacy across patient groups

Economics: Analyse consumer behaviour and market trends using consumer data samples

Biology: Analyse sample means to study genetic traits and protein levels in populations

Limits of the Central Limit Theorem

Key Conditions Where the CLT May Not Hold True

- **Small sample size:** Small samples ($n < 30$) may reduce reliability
- **Independence assumption:** Observations must be independent
- **Infinite variance:** Not valid for distributions with infinite variance such as Cauchy distribution
- **Skewness impact:** Skewed data may affect accuracy, especially with highly skewed distributions
- **Assumption verification:** Always verify assumptions before applying CLT

Inferential Statistics in Action

Core uses of the Central Limit Theorem

- Enables estimation of population parameters
- Acts as a foundation for hypothesis testing (z-tests, t-tests, ANOVA)
- Construct confidence intervals
- Conduct power analysis for sample size

Real-World Applications of the CLT

How the Central Limit Theorem Shapes Decision-Making

- Guides sample size in research design
- Quantifies uncertainty in estimates
- Supports quality control processes
- Improves polling and market insights
- Aids risk assessment in finance and insurance

Hypothesis Testing Basics

What is Hypothesis Testing?

Shaping Business Intelligence and Decision-Making

- **Definition:** Make inferences about a population parameter based on sample data
- **Purpose:** Assess evidence provided by data against a stated hypothesis
- **Importance:** Supports decision-making in research, business, healthcare, and many other fields

Essential Hypothesis Testing Terms

Key concepts for interpreting test results

- **Null hypothesis (H_0):** Assumes no effect or difference
- **Alternative hypothesis (H_1):** Contradicts the null assumption
- **Significance level (α):** Cut-off for rejecting H_0 (often 0.05)

- **p-value:** Assume H_0 holds, compute likelihood of observed or more extreme result
- **Test statistic:** Standardised value used to evaluate H_0

Steps in Hypothesis Testing

A Structured Approach to Testing Assumptions

1. Formulate hypothesis

State H_0 and H_1 clearly

2. Choose a significance level:

Determine the risk threshold (commonly 5% or 1%)

3. Collect and summarise data

Use descriptive statistics to capture sample characteristics

4. Calculate the test statistic

Use a formula based on the sample data

5. Make a decision

Compare the test statistic to a critical value or use the p-value

6. Interpret the results

State the conclusion in the context of the research question

Null vs Alternative Hypotheses

Statistical Testing Framework

Null hypothesis (H_0):

- Assumes no effect or difference exists
- Example: "no difference in mean sales before/after marketing campaign"

Alternative hypothesis (H_1):

- Predicts presence of an effect or difference
- Example: "difference exists in mean sales before/after marketing campaign"

Purpose:

- Establishes clear baseline (H_0) for objective evidence testing

Test Statistics and p-Values

Core Components of Statistical Testing

Test statistic:

- Converts observed data into standardised form
- Enables comparison to theoretical distribution under h_0

p-value:

- Measures strength of evidence against H_0

Decision rule:

- If p-value $\leq \alpha$, reject H_0
- If p-value $> \alpha$, fail to reject H_0

Note: These concepts do not require detailed knowledge of the underlying method to be understood in principle

Type I and Type II Errors

Understanding Statistical Decision-Making Risks

Type I error (False positive):

- Rejects H_0 when it is actually true
- Probability equals significance level (α)

Type II error (False negative):

- Fails to reject H_0 when H_1 is true
- Probability denoted by β

Error trade-off:

- Reduces one type of error often increases the other
- Set appropriate significance levels for balanced decisions

Confidence Intervals and Hypothesis Testing

Linking Parameter Estimation with Statistical Decisions

Definition:

- Range of values where true population parameter is expected to lie with certain confidence level

Hypothesis testing link:

- Indicate H_0 should be rejected when parameter value lies outside

Visual decision tool:

- Overlap between CI and hypothesised value aids decision-making

Interpreting and Communicating Results

Effective Reporting and Communication

Result interpretation:

- State if H_0 was rejected and explain the practical implications

Reporting p-values and confidence intervals:

- Provide context and limitations of the results

Communicating uncertainty:

- Discuss potential errors and reliability of findings

Common Misconceptions and Conceptual Example

Avoiding Pitfalls in Statistical Interpretation

Common misconceptions:

- Low p-value ≠ large effect size
- "Failing to reject" ≠ "accepting" H_0 (insufficient evidence, not proof)
- Statistical significance ≠ practical importance

Teaching method example

- **H_0:** New method has no effect on performance
- **H_1:** New method has an effect

Steps:

- **Data Collection:** Gather test scores from a sample of students
- **Calculation:** Determine the test statistic and p-value
- **Decision:** Use the p-value to decide on H_0
- **Discussion:** How the evidence is weighed, without specifying the testing method

Hands-on Practice

Hands-on Practice

Confidence Intervals and Central Limit Theorem

CI calculations:

- Computing confidence intervals for normal distribution samples

Sample size comparison:

- Analysing how different sample sizes affect confidence intervals

CLT simulation:

- Visualising Central Limit Theorem using exponential distribution

Calculate Confidence Interval Using Python

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import t

def exercise_confidence_interval(sample, alpha=0.05):
    """
    Given a sample, compute the 95% (or given alpha) confidence interval.
    Returns the sample mean, sample standard deviation, margin of error, and the confidence
interval.
    """
    sample_mean = np.mean(sample)
    sample_std = np.std(sample, ddof=1)
    n = len(sample)
    t_crit = t.ppf(1 - alpha/2, df=n-1)
    margin_of_error = t_crit * (sample_std / np.sqrt(n))
    ci_lower = sample_mean - margin_of_error
    ci_upper = sample_mean + margin_of_error
    return sample_mean, sample_std, margin_of_error, (ci_lower, ci_upper)
```

Compare Confidence Intervals By Sample Size

```
def compare_confidence_intervals():
    print("=== Comparing Confidence Intervals for Different Sample Sizes ===")
    alpha = 0.05

    # For sample size n = 30
    np.random.seed(0)
    sample_n30 = np.random.normal(loc=100, scale=15, size=30)
    mean_n30, std_n30, me_n30, ci_n30 = exercise_confidence_interval(sample_n30, alpha=alpha)
    print("\nFor n = 30:")
    print("  Sample Mean:", mean_n30)
    print("  Sample Standard Deviation:", std_n30)
    print("  Margin of Error:", me_n30)
    print("  95% Confidence Interval: [{:.2f}, {:.2f}]".format(ci_n30[0], ci_n30[1]))

    # For sample size n = 100
    # Reset the seed to keep the same population parameters
    np.random.seed(0)
    sample_n100 = np.random.normal(loc=100, scale=15, size=100)
    mean_n100, std_n100, me_n100, ci_n100 = exercise_confidence_interval(sample_n100, alpha=alpha)
    print("\nFor n = 100:")
    print("  Sample Mean:", mean_n100)
    print("  Sample Standard Deviation:", std_n100)
    print("  Margin of Error:", me_n100)
    print("  95% Confidence Interval: [{:.2f}, {:.2f}]".format(ci_n100[0], ci_n100[1]))
```

Simulate Central Limit Theorem Using Samples

```
def simulate_clt(sample_size=30, num_samples=1000):
    """
    Simulate the Central Limit Theorem by drawing repeated samples from an exponential
distribution.
    Returns an array of sample means.
    """
    # Simulate a non-normal population: exponential distribution.
    population = np.random.exponential(scale=1.0, size=10000)
    means = []
    for _ in range(num_samples):
        sample = np.random.choice(population, size=sample_size, replace=True)
        means.append(np.mean(sample))
    return means
```

Plot Sample Means Histogram

```
def plot_sample_means(means, sample_size):
    """
    Plot the histogram of sample means.
    """
    plt.figure(figsize=(8, 6))
    plt.hist(means, bins=30, edgecolor='black', alpha=0.7)
    plt.title(f"Distribution of Sample Means (n = {sample_size})")
    plt.xlabel("Sample Mean")
    plt.ylabel("Frequency")
    plt.show()
```

Run Confidence Interval Demo

```
def main():
    # Part 1: Confidence Intervals Practice
    print("=== Part 1: Confidence Intervals Practice ===\n")

    # Exercise 1: Calculating a 95% Confidence Interval for a sample of size 50.
    print("Exercise 1: Calculating a 95% Confidence Interval")
    np.random.seed(0)
    sample = np.random.normal(loc=100, scale=15, size=50)
    mean, std, margin, ci = exercise_confidence_interval(sample)
    print("  Sample Mean:", mean)
    print("  Sample Standard Deviation:", std)
    print("  Margin of Error:", margin)
    print("  95% Confidence Interval: [{:.2f}, {:.2f}]".format(ci[0], ci[1]))

if __name__ == '__main__':
    main()
```

Q & A

Thank you