Section 2: Inferential Statistics

12. Why Inferential Statistics?

- **Explain the difference between Correlation and Causation with an example.**

- Inferential statistics is a branch of statistics that focuses on using sample data to make generalizations (inferences) about a larger population.

- It involves drawing conclusions, making predictions, and testing hypotheses based on data collected from a subset of the population.

CORRELATION:

- Correlation measures the strength and direction of the relationship between two variables.

Range: Correlation coefficients (e.g., Pearson's r) range from −1 to +1:

r=+1: Perfect positive correlation.

r=−1: Perfect negative correlation. r=0: No correlation.

CAUSATION:

- Causation implies that a change in one variable directly causes a change in another variable.

Requirements:

Temporal precedence: The cause must occur before the effect.

Association: There must be a correlation between the variables.

Elimination of confounding variables: Other factors that could explain the relationship must be ruled out.

Example:

Smoking causes lung cancer (established through controlled studies and eliminating confounding factors).

Key Difference:

Correlation does not imply causation. Just because two variables are correlated does not mean that one causes the other. Establishing causation requires additional evidence, such as controlled experiments or longitudinal studies.

13. Population vs. Sample:

- **Why do we need sampling? Provide a real-world example.**

- ***POPULATION:***

It is often impractical or impossible to collect data from an entire population (e.g., all people in a country, all products in a factory).

- ***SAMPLE:***

Instead, we collect data from a sample (a smaller, manageable subset of the population) and use inferential statistics to generalize the findings to the population.

Why do we need sampling?

Sampling is the process of selecting a subset of individuals, items, or data points from a larger population to make inferences about the entire population. Here's why sampling is essential:

1) Practicality:

It is often impractical or impossible to study an entire population due to time, cost, or logistical constraints. Sampling allows researchers to collect data efficiently.

2) Cost-Effectiveness:

Studying a sample is less expensive than studying an entire population.

3) Time Efficiency:

Sampling reduces the time required to collect and analyze data.

4) Feasibility:

In some cases, the population is too large or inaccessible (e.g., all humans, all stars in the galaxy), making sampling the only viable option.

5) Accuracy: With proper sampling techniques, researchers can obtain accurate and reliable results that generalize to the population.

Real_ World Example:

Example: Political Opinion Polls
During an election, polling organizations aim to predict the outcome of the vote. Instead of surveying every eligible voter (which would be impractical due to the sheer size of the population), they use sampling. For instance:

- A polling agency might select a random sample of 1,000 voters from different demographics, regions, and age groups.
- By analyzing the responses of this sample, they can estimate the overall voting behavior of the entire population.
- This approach saves time and money while providing a reliable snapshot of public opinion.

Sampling ensures that the results are representative of the larger population, provided the sample is chosen carefully and without bias.

14. Hypothesis Testing Concepts:

- **Define Null Hypothesis, Alternate Hypothesis, Significance Level (α), and P-value**

Hypothesis testing is a statistical method used to make decisions or inferences about a population based on sample data. It involves testing a claim or assumption (hypothesis) about a population parameter (e.g., mean, proportion, variance).

Null Hypothesis (H0)

The null hypothesis is a statement of no effect or no difference. It represents the default or status quo assumption.

The goal of hypothesis testing is often to determine whether there is enough evidence to **reject the null hypothesis**.

It is denoted as H0, and it is pronounced H-naught.

Alternate Hypothesis(H1)

The alternate hypothesis is a statement that contradicts the null hypothesis. It represents the research question or the effect we want to test.

Significance Level (α)

- The significance level (α) is the probability of rejecting the null hypothesis when it is true (Type I error).
- Common values for α are 0.05 (5%), 0.01 (1%), and 0.10 (10%).
- It determines the threshold for deciding whether the observed result is statistically significant.

Example:

If α = 0.05, we reject H0 if the p-value is less than 0.05.

- Typically used values are 0.05 (e.g. e-commerce) and 0.01 (e.g. in fields like medicine)

P-Value

The P-Test is not a distinct test but relates to the calculation and interpretation of the p

value, a key concept in hypothesis testing.

Key Concepts

• P-Value Definition: – A measure of the probability that the observed data (or something more extreme) occurred under the null hypothesis. – Expressed as:

P=P(Observed outcome or more extreme Null Hypothesis true)

• Threshold for Significance: – The p-value is compared against a significance level (α), often 0.05 or 0.01.

• P-value<α: Reject the null hypothesis.

• P-value>α: Fail to reject the null hypothesis.

Usage

• P-values are calculated in conjunction with specific tests (e.g., T-tests, Z-tests).

• A smaller p-value indicates stronger evidence against the null hypothesis

20. Summary and Insights:

● Summarize the key takeaways from the analysis performed above and describe how descriptive and inferential statistics can be used in real-world data analysis

- ***Key takeaways:***

- **Descriptive statistics used to summarize data whereas inferential statistics helps us to make predictions or inferences about a population.**
- **Measures mean, median, and standard deviation describe central tendency and dispersion.**
- **Hypothesis testing helps to determine if observed effects are statistically significant or not.**

- ***Real-world Applications:***

- **Descriptive statistics are used in reporting and data visualization.**
- **Inferential statistics are used in research, quality control, and decision-making.**