**UNIT – IV: Sampling theory**

**Estimation- Percentiles- The Bootstrap - Confidence Intervals- Using Confidence Intervals - The SD and the Normal Curve - The Central Limit Theorem - point and interval estimation, Prediction- Correlation -The Regression Line -The Method of Least Squares - Least Squares**

**Estimation**

Estimation refers to the process of inferring the value of a population parameter based on sample data. There are two main types of estimation:

1. Point Estimation: Provides a single value (point) as an estimate of a population parameter.

2. Interval Estimation: Provides a range (interval) within which the parameter is expected to lie, often accompanied by a confidence level.

**Percentiles**

Percentiles are values that divide a dataset into 100 equal parts. For example, the 50th percentile (median) is the value below which 50% of the observations fall.

**Example:** In a test score dataset, if the score at the 90th percentile is 85, it means that 90% of the students scored below 85.

**The Bootstrap**

**Bootstrapping** is a resampling technique used to estimate the distribution of a sample statistic by repeatedly sampling with replacement from the data.

**Steps:**

1. Take a sample from the population.

2. Resample the data with replacement to create multiple bootstrap samples.

3. Calculate the statistic of interest for each bootstrap sample.

4. Use the distribution of the bootstrap statistics to make inferences about the population parameter.

**Example:** If we want to estimate the mean of a small dataset, we can create several bootstrap samples, calculate their means, and then use those means to create a confidence interval for the population mean.

**Confidence Intervals**

A confidence interval is a range of values used to estimate a population parameter. It is calculated from sample data and provides an interval estimate with a specified level of confidence (e.g., 95%).

\*Formula for Confidence Interval\* (for the mean):

\[\text{CI} = \bar{x} \pm z \left( \frac{s}{\sqrt{n}} \right)\]

Where:

- \( \bar{x} \) = sample mean

- \( z \) = z-value from the standard normal distribution (for 95% confidence, \( z \approx 1.96 \))

- \( s \) = sample standard deviation

- \( n \) = sample size

**Example:** If a sample of 30 students has a mean test score of 78 and a standard deviation of 10, the 95% confidence interval for the mean test score would be:

\[CI = 78 \pm 1.96 \left( \frac{10}{\sqrt{30}} \right) \approx 78 \pm 3.58 \rightarrow (74.42, 81.58)

\]

**Using Confidence Intervals**

Confidence intervals are used to:

- Estimate population parameters.

- Make decisions based on the range of plausible values.

- Assess the precision of the estimate: a narrower interval indicates more precision.

**The Standard Deviation and the Normal Curve**

The standard deviation (SD) measures the dispersion or spread of a dataset. It is calculated as:

\[s = \sqrt{\frac{\sum (x\_i - \bar{x})^2}{n - 1}}\]

The normal curve is a bell-shaped distribution characterized by its mean (µ) and standard deviation (σ). Approximately 68% of the data falls within one standard deviation from the mean, about 95% falls within two standard deviations, and about 99.7% falls within three standard deviations (Empirical Rule).

**The Central Limit Theorem (CLT)**

The Central Limit Theorem states that the distribution of sample means approaches a normal distribution as the sample size increases, regardless of the population's distribution. This is crucial for inference because it allows us to use normal distribution properties for hypothesis testing and confidence intervals.

**Implication:** If you take sufficiently large random samples from any population, the sampling distribution of the sample mean will be normally distributed.

**Point and Interval Estimation**

**- Point Estimation:** Involves providing a single best estimate of a population parameter. For example, using the sample mean as a point estimate for the population mean.

**- Interval Estimation:** Provides a range of values (confidence interval) that likely contains the population parameter. This accounts for sampling variability.

**Prediction**

Prediction involves forecasting future observations based on existing data. Linear regression models are often used for making predictions.

**Example:** If we have a regression model that predicts test scores based on study hours, we can input the number of study hours to predict a student's score.

**Correlation**

Correlation measures the strength and direction of the linear relationship between two variables. It is quantified using the \*correlation coefficient (r)\*, which ranges from -1 to 1.

- \( r = 1 \): Perfect positive correlation

- \( r = -1 \): Perfect negative correlation

- \( r = 0 \): No correlation

**Example:** In a study measuring hours studied and test scores, a correlation of 0.8 indicates a strong positive relationship: as study hours increase, test scores tend to increase.

**The Regression Line**

The regression line represents the best-fitting line through the data points in a scatterplot. It can be expressed in the form:

\[y = mx + b\]

Where:

- \( y \) = dependent variable (response)

- \( m \) = slope of the line (change in \( y \) for a one-unit change in \( x \))

- \( x \) = independent variable (predictor)

- \( b \) = y-intercept (the value of \( y \) when \( x = 0 \))

**The Method of Least Squares**

The method of least squares is used to determine the best-fitting line by minimizing the sum of the squared differences between observed values and the values predicted by the line.

**Example:** Given a set of data points, the least squares method will find the line that minimizes the squared residuals (the vertical distances between the points and the line).

**Least Squares**

In regression analysis, the least squares method provides:

1. The estimated coefficients (slope and intercept).

2. The ability to make predictions based on the regression model.

3. Insights into the relationship between variables.

\*Example\*: If we find that the regression equation for predicting test scores from hours studied is:

\[\text{Test Score} = 50 + 10 \times \text{Hours}\]

This indicates that for each additional hour studied, the test score is expected to increase by 10 points, starting from a base score of 50.

**Estimation**

Estimation involves inferring the value of a population parameter based on sample data. It is crucial in statistics as it allows researchers to make educated guesses about characteristics of a larger group without having to survey every individual.

Example: Estimating the Average Salary of Employees

Imagine a company wants to know the average salary of all its employees (population). Instead of surveying all 1,000 employees, they randomly select 100 employees and calculate their average salary.

**- Point Estimation:** If the average salary of the 100 sampled employees is $60,000, this is a point estimate of the average salary for all employees.

**- Interval Estimation:** The company might also calculate a confidence interval, for example, $58,000 to $62,000, indicating they are 95% confident the true average salary lies within this range.

**Percentiles**

Percentiles divide a dataset into 100 equal parts, helping to understand the distribution of data points.

**Example: Exam Scores**

Consider a dataset of exam scores for 1,000 students. The 90th percentile score (P90) indicates that 90% of students scored below this value.

- If the 90th percentile score is 92, it means that 900 students scored less than 92. This helps educators understand how students are performing relative to one another.

**The Bootstrap**

Bootstrapping is a resampling method used to estimate the distribution of a statistic by repeatedly sampling with replacement from the data.

**Example: Estimating Median Income**

Suppose a researcher collects income data from 30 households and wants to estimate the median income. They can use bootstrapping:

1. Randomly sample from the 30 households with replacement to create multiple "bootstrap" samples (e.g., 1,000 samples).

2. Calculate the median income for each sample.

3. Use the distribution of these medians to create a confidence interval for the population median.

**Confidence Intervals**

A confidence interval provides a range of values for an estimated population parameter, giving an idea about the uncertainty of the estimate.

**Example: Polling Data**

Consider a political poll conducted among 1,200 voters to estimate support for a candidate. If 52% of respondents support the candidate, the pollster might calculate a 95% confidence interval:

\[\text{CI} = 52\% \pm 1.96 \times \sqrt{\frac{52\% \times (1 - 52\%)}{1200}} \approx 52\% \pm 2.8\%\]

This results in a confidence interval of approximately (49.2%, 54.8%), indicating that the true support level is likely between these percentages.

**Using Confidence Intervals**

Confidence intervals can be used to:

- Estimate population parameters (mean, proportion).

- Make decisions based on the range of plausible values.

- Assess the precision of estimates, as narrower intervals indicate greater precision.

**The Standard Deviation and the Normal Curve**

The standard deviation (SD) measures the dispersion of data points around the mean. The \*normal curve\* represents the probability distribution of a continuous random variable.

**Example: Heights of Students**

If we measure the heights of 200 college students, we may find:

- Mean height = 170 cm

- Standard deviation = 10 cm

Using the normal distribution (bell curve), we can predict that:

- Approximately 68% of students will have heights between 160 cm and 180 cm (mean ± 1 SD).

- Approximately 95% will fall between 150 cm and 190 cm (mean ± 2 SD).

**The Central Limit Theorem (CLT)**

The Central Limit Theorem states that the sampling distribution of the sample mean will be normally distributed, regardless of the population's distribution, provided the sample size is sufficiently large (usually n ≥ 30).

**Example: Average Daily Temperature**

If we take multiple samples of daily temperatures from a city over a year, the average temperature of each sample will tend to follow a normal distribution, even if the daily temperatures themselves do not. This allows meteorologists to make predictions about future temperatures based on sample data.

**Point and Interval Estimation**

**- Point Estimation:** Involves providing a single best estimate of a population parameter.

**- Interval Estimation:** Provides a range of values (confidence interval) that likely contains the population parameter.

**Example: Drug Efficacy Study**

In a clinical trial evaluating a new drug, researchers find that the mean reduction in symptoms is 15 units (point estimate). They can also compute a 95% confidence interval for this estimate, say (12, 18), indicating they are confident the true mean reduction lies within this interval.

**Prediction**

Prediction involves forecasting future observations based on existing data, often utilizing regression models.

**Example: Housing Prices**

A real estate analyst builds a regression model to predict house prices based on features like size (square footage), number of bedrooms, and location.

- If the regression equation is:

\[\text{Price} = 50,000 + 200 \times \text{Size} + 10,000 \times \text{Bedrooms}\]

This equation allows the analyst to predict the price of a house based on its size and number of bedrooms.

**Correlation**

Correlation measures the strength and direction of the linear relationship between two variables, quantified using the correlation coefficient (r).

**Example: Study Hours vs. Exam Scores**

In a study examining the relationship between hours studied and exam scores, researchers find a correlation coefficient of 0.85. This strong positive correlation indicates that as study hours increase, exam scores tend to increase as well.

**The Regression Line**

The regression line is the line of best fit through a scatterplot of data points, representing the relationship between the independent variable (predictor) and the dependent variable (response).

**Example: Sales Prediction**

A company analyzes its advertising spend and sales data. The regression line might be expressed as:

\[\text{Sales} = 10,000 + 5 \times \text{Ad Spend}\]

This means for every additional dollar spent on advertising, sales increase by $5.

**The Method of Least Squares**

The method of least squares minimizes the sum of the squared differences between observed values and the values predicted by the regression line.

**Example: Fitting a Line to Data Points**

If a researcher has data points representing hours studied and corresponding test scores, the least squares method will find the line that minimizes the squared vertical distances (residuals) between the points and the line.

**Least Squares**

In regression analysis, the least squares method provides estimated coefficients (slope and intercept) and enables predictions.

**Example: Analyzing Sales Data**

If the least squares regression analysis of sales data yields a slope of 3 and an intercept of 7, the regression equation would be:

\[\text{Sales} = 7 + 3 \times \text{Advertising Spend}\]

This indicates that if no money is spent on advertising, sales would start at 7, and for each dollar spent, sales would increase by 3 units.

**Conclusion**

These concepts offer a comprehensive understanding of statistical estimation, confidence intervals, regression analysis, and prediction. They are essential in various fields, including business, healthcare, and social sciences, allowing researchers and practitioners to make data-driven decisions and forecasts.