Statistics in Data Analytics

Making Sense of Data Through Numbers and Trends

ULOH C. ARMSTRONG

What is Statistics in Data Analytics?

- Statistics is the branch of mathematics dealing with data collection, analysis, interpretation, and presentation.

Why Statistics Matters in Data Analytics

- Helps summarize complex datasets into meaningful insights.
- Enables data-driven decision-making in businesses.
- Provides tools to model uncertainty and make predictions.
- Measures the reliability of conclusions.

Types of Data

Data can be broadly categorized as:

1. Quantitative (e.g., sales figures).

2. Qualitative (e.g., customer feedback).

- Additionally, data can be continuous (e.g., height) or discrete (e.g., number of customers).

Descriptive Statistics

Descriptive statistics summarize and describe data features:

- Mean: Average value. Summarizes central tendency.
- Median: Middle value. The median is less affected by outliers.
- Mode: Most frequent value.
- Standard Deviation: Spread of data values.

Visualizing Data

Visualization is key to understanding data.

Common tools include:

- 1. Bar Charts: Compare categories.
- 2. Line Graphs: Show trends over time.
- 3. Scatter Plots: Explore relationships between variables.

Probability in Data Analytics

Probability is the likelihood of an event happening. It quantifies uncertainty, and predicts outcomes in analytics.

Example:

- Probability of a coin landing heads: 50%.
- Probability of rolling a 6 on a dice: 1/6 or 16.67%.
- Probability of drawing a red card from a deck: 26/52 or 50%.

Correlation vs. Causation

- Correlation shows a relationship between variables. Correlation shows a relationship between variables (e.g., ice cream sales and temperature).
- Causation proves one variable directly affects another. Causation indicates one variable directly affects another (e.g., flipping a light switch causes the light to turn on).

Example:

Ice cream sales and temperature are correlated but not causal. Ice cream sales and temperature are correlated but one doesn’t cause the other.

Correlation B

Measures the relationship between two variables.

Positive correlation, no correlation , and negative correlation are determined by the correlation coefficient. The correlation coefficient ranges from -1 to 1.

Examples:

Positive correlation: More study hours → Higher grades.

Negative correlation: More screen time → Less sleep.

Always remember that : Correlation ≠ causation

Correlation Coefficient

Correlation coefficient quantifies correlation (-1 to 1).

-1: Perfect negative

0: No correlation

1: Perfect positive.

Example: Correlation of 0.8 indicates a strong positive relationship

Regression Analysis

- Regression predicts relationships between variables.
- Example: Predicting house prices based on size, location, and age. Predicting sales based on advertising spend.

Identifying Outliers

Outliers are extreme values that differ from the rest of the data.

Example:

- A student scoring 100% when the average is 70%.
- A customer aged 160 years.

Statistical Inference

Statistical inference allows generalizing insights from a sample to a population. That is: drawing conclusions about a population based on a sample.

Key Tools: Estimation (e.g., confidence intervals) and hypothesis testing

Example:

Estimating the taste of a pot of soup by taking a spoon of it.

Estimating national average income using a survey.

Estimating the height of males in a country using a survey.

Hypothesis Testing A

A method to test assumptions about data. Testing assumptions about data through:

- Null Hypothesis (H0): No effect or change.
- Alternative Hypothesis (H1): Effect or change exists.

Example:

- Testing if a new drug improves recovery rates.
- Testing if a new marketing strategy increases sales.
- Are customers spending more after a promotional campaign?

Hypothesis Testing B

A method to test an assumption about a population parameter.

Key Steps:

- State null (H_0) and alternative (H_1) hypotheses.
- Choose a significance level (e.g., $\alpha = 0.05$).
- Conduct the test and interpret results.

P-Value

The p-value is the probability of obtaining results at least as extreme as observed, assuming H_0 is true.

Interpretation:

If $P < \alpha$: Reject H_0 (significant result).

If $P \geq \alpha$: Fail to reject H_0 (no significant evidence).

Example: "If $P = 0.03$, there's a 3% chance the result is due to random variation."

Confidence Interval

A range of values that is likely to contain the true population parameter.

Example: "We are 95% confident that the average height of adults is between 160 cm and 170 cm.“

Confidence interval is useful because:

It simplifies uncertainty.

It helps make predictions about a population based on a sample.

Statistical Tests

Common Tests:

T-test: Compares means of two groups.
Chi-square test: Examines relationships between categorical variables.
ANOVA: Compares means across multiple groups.
Regression: Explores relationships between variables.

Scenarios:
T-test: Comparing test scores of two classes.
Chi-square: Testing if gender influences purchasing habits.
ANOVA: Analyzing sales performance across regions.

Probability Distribution

Probability distribution is a function showing all possible values and their probabilities.

Types:

- Normal distribution: Bell-shaped curve.
- Binomial distribution: Success/failure outcomes.
- Poisson distribution: Rare events in a fixed interval.

etc

Conclusion

- Statistics is the backbone of data analytics, providing tools to make sense of complex data.
- Understanding these concepts ensures better decisions and insights.