**STATISTICS**

## **Descriptive Statistics**

### **1. What is the difference between population and sample?**

* **Population** refers to the entire group being studied, while a **sample** is a subset of the population used for analysis.

### **2. What are mean, median, and mode?**

* **Mean**: The average of all values.
* **Median**: The middle value when arranged in order.
* **Mode**: The most frequently occurring value.

### **3. What is the difference between variance and standard deviation?**

* **Variance** measures how far data points are from the mean, squared.
* **Standard deviation** is the square root of variance, making it more interpretable in the same units as the data.

### **4. What are skewness and kurtosis?**

* **Skewness**: Measures the asymmetry of the data distribution.
  + Positive skew → right tail is longer.
  + Negative skew → left tail is longer.
* **Kurtosis**: Measures the "tailedness" of the distribution.
  + High kurtosis → more outliers (leptokurtic).
  + Low kurtosis → fewer outliers (platykurtic).

### **5. What is the difference between a parameter and a statistic?**

* **Parameter**: A measure that describes the entire population (e.g., population mean).
* **Statistic**: A measure calculated from a sample (e.g., sample mean).

## **Probability & Distributions**

### **6. What is the difference between discrete and continuous variables?**

* **Discrete**: Countable values (e.g., number of students).
* **Continuous**: Infinite possible values within a range (e.g., height, weight).

### **7. What are different probability distributions?**

* **Normal Distribution**: Symmetric, bell-shaped curve.
* **Binomial Distribution**: Discrete distribution for independent trials (e.g., coin flips).
* **Poisson Distribution**: Counts of rare events in a fixed time/space.
* **Exponential Distribution**: Time between events in a Poisson process.

### **8. What is the Central Limit Theorem (CLT)?**

* CLT states that the **sampling distribution** of the sample mean approaches a **normal distribution** as the sample size increases, regardless of the original population's distribution.

### **9. What is the Law of Large Numbers?**

* As the sample size increases, the sample mean converges to the population mean.

### **10. What is Bayes’ Theorem and its formula?**

* It describes the probability of an event based on prior knowledge: P(A∣B)=P(B∣A)P(A)P(B)P(A | B) = \frac{P(B | A) P(A)}{P(B)}

## **Inferential Statistics & Hypothesis Testing**

### **11. What is the difference between Type I and Type II errors?**

* **Type I Error (False Positive)**: Rejecting a true null hypothesis.
* **Type II Error (False Negative)**: Failing to reject a false null hypothesis.

### **12. What is p-value?**

* The p-value is the probability of obtaining results at least as extreme as the observed data, assuming the null hypothesis is true.

### **13. What is the confidence interval?**

* A range within which the true population parameter is expected to fall, with a certain probability (e.g., 95% confidence interval).

### **14. What are different types of hypothesis tests?**

* **Z-test**: For large sample sizes (n > 30) with known population variance.
* **T-test**: For small sample sizes (n < 30) or unknown variance.
* **Chi-square test**: For categorical data relationships.
* **ANOVA**: Compares means across multiple groups.

### **15. What are one-tailed and two-tailed tests?**

* **One-tailed test**: Tests for effect in one direction (greater or lesser).
* **Two-tailed test**: Tests for effect in both directions (different but unspecified).

### **16. What is the difference between correlation and causation?**

* **Correlation**: Two variables move together but do not imply a cause-effect relationship.
* **Causation**: One variable directly influences another.

## **Regression Analysis**

### **17. What is linear regression?**

* A statistical technique that models the relationship between a dependent variable **Y** and an independent variable **X** using the equation: Y=β0+β1X+ϵY = \beta\_0 + \beta\_1 X + \epsilon

### **18. What is R-squared?**

* A measure of how well the regression model fits the data. Values range from **0 to 1**, where **1 indicates a perfect fit**.

### **19. What are multicollinearity and how to detect it?**

* **Multicollinearity**: When independent variables are highly correlated.
* Detection methods: **Variance Inflation Factor (VIF)**, correlation matrix.

### **20. What is heteroskedasticity?**

* Unequal variance of residuals across values of an independent variable in regression, violating the assumption of homoscedasticity.

## **Machine Learning Statistics**

### **21. What is logistic regression?**

* A regression model used for binary classification (e.g., spam vs. not spam), using the **sigmoid function** to predict probabilities.

### **22. What is the difference between parametric and non-parametric tests?**

* **Parametric tests**: Assume a normal distribution (e.g., t-test, ANOVA).
* **Non-parametric tests**: Do not assume normality (e.g., Mann-Whitney U test).

### **23. What is the difference between Precision and Recall?**

* **Precision**: True Positives / (True Positives + False Positives).
* **Recall**: True Positives / (True Positives + False Negatives).

### **24. What is the F1-score?**

* Harmonic mean of precision and recall: F1=2×Precision×RecallPrecision+RecallF1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}

### **25. What is A/B testing?**

* A statistical experiment comparing two versions (A & B) to determine which performs better using hypothesis testing.

### **26. What is overfitting and how to prevent it?**

* **Overfitting**: When a model learns noise instead of the actual pattern.
* **Prevention**: Regularization (L1/L2), cross-validation, pruning (decision trees).

### **27. What is the difference between Bagging and Boosting?**

* **Bagging**: Combines multiple weak models in parallel to reduce variance (e.g., Random Forest).
* **Boosting**: Sequentially trains models to focus on errors (e.g., AdaBoost, XGBoost).

### **28. What is the curse of dimensionality?**

* When the number of features increases, data points become sparse, reducing model performance.

### **29. What is the p-value threshold typically used?**

* A p-value < 0.05 indicates statistical significance.

### **30. How do you check if data is normally distributed?**

* **Visual methods**: Histogram, Q-Q plot.
* **Statistical tests**: Shapiro-Wilk, Kolmogorov-Smirnov test.

**Hypothesis Testing**

## **1. Analysis of Variance (ANOVA)**

**Purpose:**

* ANOVA tests **differences in means** across **three or more** independent groups.
* It determines if at least one group mean is significantly different but does not specify which one.

**Types of ANOVA:**

* **One-Way ANOVA** → Compares one categorical independent variable with multiple groups.
* **Two-Way ANOVA** → Compares two independent variables with multiple groups.
* **Repeated Measures ANOVA** → Used for dependent samples (e.g., same subjects tested multiple times).

**Assumptions:**

1. Data is **normally distributed**.
2. **Homogeneity of variance** (equal variance across groups) → Tested using Levene’s test.
3. Observations are **independent**.

**Example Scenario:**

* Comparing the test scores of students in three different teaching methods.

**Post-hoc test:**

* If ANOVA is significant, use **Tukey’s HSD** or **Bonferroni correction** to identify which groups differ.

## **2. Chi-Square Test**

**Purpose:**

* Used to **test relationships between categorical variables**.

**Types:**

* **Chi-Square Test for Independence**: Determines if two categorical variables are related.
* **Chi-Square Goodness of Fit Test**: Checks if a sample follows a specific expected distribution.

**Assumptions:**

1. Data is **categorical** (nominal or ordinal).
2. Observations are **independent**.
3. Expected frequency in each cell is **≥ 5** for validity.

**Example Scenario:**

* Testing if **gender (Male/Female)** is related to **voting preference (Party A/Party B)**.

**Formula:**

χ2=∑(O−E)2E\chi^2 = \sum \frac{(O - E)^2}{E}

where **O** = observed frequency, **E** = expected frequency.

## **3. Mann-Whitney U Test (Non-Parametric Equivalent of Independent t-test)**

**Purpose:**

* Used to compare **two independent groups** when **data is not normally distributed**.
* It checks if one group tends to have higher/lower values than the other.

**Assumptions:**

1. Data is **ordinal or continuous**.
2. Observations in the two groups are **independent**.

**Example Scenario:**

* Comparing customer satisfaction scores between two different products.

## **4. Kruskal-Wallis Test (Non-Parametric Equivalent of ANOVA)**

**Purpose:**

* Used to compare **three or more** independent groups when **data is not normally distributed**.

**Assumptions:**

1. Data is **ordinal or continuous**.
2. Groups are **independent**.
3. It does not assume equal variances.

**Example Scenario:**

* Comparing customer ratings across three different restaurants.

**Post-hoc Test:**

* If significant, **Dunn’s test** is used to find which groups differ.

## **5. Independent T-Test (Unpaired T-Test)**

**Purpose:**

* Compares **two independent groups** for **significant differences in means**.

**Assumptions:**

1. **Normality** → Data follows a normal distribution.
2. **Homogeneity of variance** → Checked using **Levene’s test**.
3. Groups are **independent**.

**Example Scenario:**

* Comparing **average heights** of male and female students.

**Formula:**

t=Xˉ1−Xˉ2s12n1+s22n2t = \frac{\bar{X}\_1 - \bar{X}\_2}{\sqrt{\frac{s\_1^2}{n\_1} + \frac{s\_2^2}{n\_2}}}

## **6. Related (Paired) T-Test**

**Purpose:**

* Compares **two related (paired) groups**, like **before-and-after** measurements on the same subjects.

**Assumptions:**

1. Differences between pairs are **normally distributed**.
2. The sample pairs are **dependent** (e.g., same person tested twice).

**Example Scenario:**

* Measuring **blood pressure** before and after a medication trial.

**Formula:**

t=DˉsD/nt = \frac{\bar{D}}{s\_D / \sqrt{n}}

where **D** is the difference between paired observations.

### **Summary of When to Use Each Test**

|  |  |  |  |
| --- | --- | --- | --- |
| **Test** | **Parametric/Non-Parametric** | **Purpose** | **Example** |
| One-Way ANOVA | Parametric | Compare **3+ groups** | Test scores of students in different classes |
| Two-Way ANOVA | Parametric | Compare **2 factors** with multiple levels | Examining effect of gender & study method on scores |
| Chi-Square | Non-Parametric | Test **association between categorical** variables | Gender vs. Voting Preference |
| Mann-Whitney U | Non-Parametric | Compare **2 independent groups** | Customer satisfaction scores of two stores |
| Kruskal-Wallis | Non-Parametric | Compare **3+ independent groups** | Ratings of three restaurants |
| Independent T-Test | Parametric | Compare **2 independent groups** | Height of males vs. females |
| Paired T-Test | Parametric | Compare **before-after** in **same group** | Blood pressure before vs. after medicine |

## **1. What is a point estimate?**

### **Answer:**

A **point estimate** is a single value used to estimate an unknown population parameter.

* Example: The sample mean (xˉ\bar{x}) is a point estimate of the population mean (μ\mu).

## **2. What is an interval estimate, and how is it different from a point estimate?**

### **Answer:**

An **interval estimate** provides a range of values that is likely to contain the population parameter, whereas a **point estimate** gives a single value.

* Example: A **confidence interval** (CI) is an interval estimate for a population mean.

## **3. How do you calculate a confidence interval for the population mean when the population standard deviation is known?**

### **Solution:**

The confidence interval (CI) formula for a population mean is:

CI=xˉ±Zα/2×σnCI = \bar{x} \pm Z\_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}

Where:

* xˉ\bar{x} = sample mean
* Zα/2Z\_{\alpha/2} = Z-score from standard normal distribution
* σ\sigma = population standard deviation
* nn = sample size

**Example:** Suppose xˉ=50\bar{x} = 50, σ=10\sigma = 10, n=100n = 100, and we want a **95% confidence interval**.

* Z0.025=1.96Z\_{0.025} = 1.96 (from Z-table)

CI=50±1.96×10100CI = 50 \pm 1.96 \times \frac{10}{\sqrt{100}} CI=50±1.96×1=50±1.96CI = 50 \pm 1.96 \times 1 = 50 \pm 1.96 CI=(48.04,51.96)CI = (48.04, 51.96)

**Interpretation:** We are 95% confident that the true population mean is between **48.04 and 51.96**.

## **4. What happens to the confidence interval if we increase the sample size?**

### **Answer:**

* As the **sample size increases**, the confidence interval **becomes narrower** because the standard error σn\frac{\sigma}{\sqrt{n}} decreases.
* This means **higher precision** in estimating the population parameter.

## **5. What is the relationship between confidence level and margin of error?**

### **Answer:**

* Higher **confidence level** (e.g., 99% vs. 95%) → **Larger margin of error** (wider interval).
* Lower **confidence level** (e.g., 90% vs. 95%) → **Smaller margin of error** (narrower interval).

**Example:** A **99% confidence interval** is wider than a **95% confidence interval** because higher confidence requires a larger range.

## **6. What is a p-value, and how is it interpreted?**

### **Answer:**

* The **p-value** is the probability of obtaining the observed data (or more extreme results) **if the null hypothesis is true**.
* A **small p-value (< 0.05)** suggests strong evidence **against** the null hypothesis (**reject H0H\_0**).
* A **large p-value (> 0.05)** suggests **insufficient evidence** to reject the null hypothesis.

**Example:** A p-value of **0.02** in a hypothesis test means there is only a **2% chance** of getting these results if H0H\_0 were true. Since 0.02<0.050.02 < 0.05, we reject H0H\_0.

## **7. What is the level of significance (α\alpha) and how does it relate to hypothesis testing?**

### **Answer:**

* The **level of significance** (α\alpha) is the **probability of making a Type I error** (rejecting a true null hypothesis).
* Common values:
  + **0.05 (5%)** → Most common
  + **0.01 (1%)** → More strict
  + **0.10 (10%)** → Less strict

**Example:** If α=0.05\alpha = 0.05, we **reject H0H\_0** if p<0.05p < 0.05.

## **8. How do you calculate a confidence interval for a population proportion?**

### **Solution:**

The confidence interval (CI) formula for a proportion is:

CI=p^±Zα/2×p^(1−p^)nCI = \hat{p} \pm Z\_{\alpha/2} \times \sqrt{\frac{\hat{p} (1 - \hat{p})}{n}}

Where:

* p^\hat{p} = sample proportion
* Zα/2Z\_{\alpha/2} = Z-score for confidence level
* nn = sample size

**Example:** If a survey of 400 people finds that **60%** (p^=0.60\hat{p} = 0.60) support a policy, and we want a **95% confidence interval**:

* Z0.025=1.96Z\_{0.025} = 1.96
* CI=0.60±1.96×(0.60)(0.40)400CI = 0.60 \pm 1.96 \times \sqrt{\frac{(0.60)(0.40)}{400}}
* CI=0.60±1.96×0.0245CI = 0.60 \pm 1.96 \times 0.0245
* CI=0.60±0.048CI = 0.60 \pm 0.048
* CI=(0.552,0.648)CI = (0.552, 0.648)

**Interpretation:** We are 95% confident that the true proportion is between **55.2% and 64.8%**.

## **9. What is the margin of error in a confidence interval?**

### **Answer:**

* **Margin of error (ME)** is the range around the point estimate in a confidence interval.
* It is given by: ME=Zα/2×σnME = Z\_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}
* A **smaller margin of error** means a **more precise** estimate.

## **10. What factors affect the width of a confidence interval?**

### **Answer:**

1. **Sample size (nn)** → Larger nn → **Narrower CI**.
2. **Confidence level (1−α1 - \alpha)** → Higher confidence (e.g., 99%) → **Wider CI**.
3. **Standard deviation (σ\sigma)** → More variability → **Wider CI**.

**Example:**

* If we increase nn from **100 to 500**, the CI becomes **narrower**.
* If we increase confidence from **95% to 99%**, the CI becomes **wider**.

### **Summary Table**

|  |  |
| --- | --- |
| **Concept** | **Explanation** |
| **Point Estimate** | A single value estimating a population parameter (e.g., sample mean). |
| **Interval Estimate** | A range of values likely to contain the true parameter (e.g., confidence interval). |
| **Confidence Level** | The probability that the interval contains the true parameter (e.g., 95%). |
| **P-Value** | Probability of obtaining the observed result if H0H\_0 is true. |
| **Level of Significance (α\alpha)** | The threshold for rejecting H0H\_0 (e.g., 0.05). |
| **Margin of Error** | The maximum expected difference between the true parameter and the estimate. |