Data science homework 7 - Mayank Sharma - ms14662

1. **How do you assess the statistical significance of an insight?** Statistical significance is assessed using hypothesis testing. The key steps are:

    - Formulate the **null hypothesis** (H_0) and **alternative hypothesis** (H_1).
    - Select an appropriate test (e.g., t-test, chi-square test, KS test) based on the data.
    - Calculate the **p-value**. If $p < \alpha$ (e.g., $\alpha = 0.05$), the result is statistically significant.
    - This process determines whether observed results are due to random chance or a real effect.
2. **What is the Central Limit Theorem? Explain it. Why is it important? Central Limit Theorem (CLT)** states:

    - The sampling distribution of the mean will approximate a **normal distribution** as the sample size increases, regardless of the population's original distribution.

    Importance:

    - It allows us to use the normal distribution to make inferences about population parameters (e.g., confidence intervals, hypothesis testing).
    - Critical for statistical tests like t-tests, z-tests, etc., where assumptions of normality are required.
3. **What is the statistical power? Statistical power** is the probability of correctly rejecting the null hypothesis (H_0) when the alternative hypothesis (H_1) is true.

    - High power reduces the likelihood of a **Type II error** (false negative).
    - Power increases with:
        - Larger sample size
        - Higher effect size
        - Lower variability in the data
        - Increased significance level (α).
4. **How do you control for biases?** To control biases in experiments:

    - **Randomization**: Randomly assign participants to groups.
    - **Blinding**:
        - Single-blind: Participants don't know their group assignment.
        - Double-blind: Both participants and experimenters are unaware.
    - **Control Groups**: Use baseline or placebo groups for comparison.
    - **Statistical Methods**: Adjust for biases using regression or stratification techniques.
5. **What are confounding variables?** A **confounding variable** is an external factor that influences both the **independent variable (IV)** and **dependent variable (DV)**, creating a

false association.

Example: In a study on exercise (IV) and weight loss (DV), diet is a confounding variable because it affects weight loss independently.

6. **What is A/B testing? A/B testing** is a controlled experiment where two versions (A and B) of a variable are compared to determine which one performs better.

   Steps:

   - Split the population into two groups.
   - Present **Version A** to Group 1 and **Version B** to Group 2.
   - Measure and compare the outcome (e.g., click rates, sales).
   - Use statistical tests to determine which version is more effective.

7. **What are confidence intervals?** A **confidence interval (CI)** is a range of values likely to contain the **population parameter** (e.g., mean) with a certain confidence level, such as 95%.

   Example:

   - A 95% confidence interval for a sample mean might be [50, 60].
   - This means that, if we repeated the experiment many times, 95% of the intervals would contain the true population mean.

   Formula:
   $CI = \bar{x} \pm z \cdot \frac{\sigma}{\sqrt{n}}$
   Where:

   - $\bar{x}$: Sample mean
   - z: Z-score corresponding to the confidence level
   - σ: Standard deviation
   - n: Sample size