Week 7 Homework – Aishwarya Kandam Vitta

---

1. How do you assess the statistical significance of an insight?

Statistical significance is assessed to determine whether an observed effect is likely due to chance or represents a real effect in the data. The common steps include:

- **Formulating a hypothesis**: Define a null hypothesis (H_0) and an alternative hypothesis (H_1).
- **Selecting a significance level (α\alpha)**: Typically set at 0.05 (5%), it represents the probability of rejecting the null hypothesis when it is true.
- **Performing a statistical test**: Use tests like t-tests, chi-square tests, or ANOVA, depending on the data and the type of analysis.
- **Calculating the p-value**: The p-value indicates the probability of observing the data if H_0 is true. If p≤αp \leq \alpha, the result is statistically significant.
- **Interpreting the results**: If statistically significant, the result supports rejecting H_0 in favor of H_1.

---

2. What is the Central Limit Theorem? Explain it. Why is it important?

The Central Limit Theorem (CLT) states that:

For a sufficiently large sample size, the sampling distribution of the sample mean will approximate a normal distribution, regardless of the original population's distribution.

Key Points:

- **Large sample size**: Typically n≥30n \geq 30.
- **Mean**: The mean of the sampling distribution equals the population mean (μ\mu).
- **Variance**: The variance of the sampling distribution equals the population variance divided by the sample size (σ2/n\sigma^2 / n).

Importance:

- Facilitates hypothesis testing and confidence interval construction for population parameters.
- Allows the use of parametric tests based on the normality assumption, even if the population is not normally distributed.

Week 7 Homework – Aishwarya Kandam Vitta

---

3. What is the statistical power?

Statistical power is the probability of correctly rejecting the null hypothesis (H_0) when it is false. It is defined as:

Power=1−β\text{Power} = 1 - \beta

Where β\beta is the probability of a Type II error (failing to reject a false H_0).

Factors affecting statistical power:

- **Sample size**: Larger samples increase power.
- **Effect size**: Larger effects are easier to detect.
- **Significance level (α\alpha)**: A higher α\alpha increases power but raises the risk of a Type I error.
- **Variance in data**: Lower variance increases power.

Why is it important?

High statistical power reduces the likelihood of failing to detect a true effect, improving the reliability of study results.

---

4. How do you control for biases?

Biases are systematic errors that can distort results. They can be controlled through:

- **Randomization**: Randomly assign participants or data points to groups to minimize selection bias.
- **Blinding**: Implement single- or double-blind designs to reduce observer and participant bias.
- **Standardization**: Use consistent procedures and protocols for data collection.
- **Matching**: Pair subjects with similar characteristics to control for confounding variables.
- **Statistical adjustments**: Use techniques like regression analysis to control for biases in observational data.

Week 7 Homework – Aishwarya Kandam Vitta

5. What are confounding variables?

A confounding variable is an extraneous factor that influences both the independent variable and the dependent variable, potentially distorting the true relationship.

Example:

In a study on exercise (independent variable) and heart health (dependent variable), age could be a confounding variable since it affects both exercise habits and heart health.

Addressing confounders:

- **Randomization**: Balances confounders across groups.
- **Stratification**: Analyze subgroups based on confounding variables.
- **Statistical methods**: Use regression models to adjust for confounders.

---

6. What is A/B testing?

A/B testing, also known as split testing, is an experimental method used to compare two versions of a variable to determine which performs better.

Process:

1. **Define goals**: Identify the metric to improve (e.g., click-through rate).
2. **Create variants**: Develop version A (control) and version B (treatment).
3. **Random assignment**: Assign users randomly to each version.
4. **Measure outcomes**: Collect data on the performance of each version.
5. **Statistical analysis**: Use hypothesis testing to evaluate differences between groups.

Applications:

- Website optimization
- Marketing campaigns
- Feature testing in software development

---

7. What are confidence intervals?

Week 7 Homework – Aishwarya Kandam Vitta

A confidence interval (CI) is a range of values derived from sample data that likely contains the population parameter with a specified level of confidence (e.g., 95%).

Formula:

$$CI = \hat{\mu} \pm Z \left(\frac{\sigma}{\sqrt{n}} \right)$$

Where:

- $\hat{\mu}$: Sample mean
- Z: Z-score corresponding to the confidence level (e.g., 1.96 for 95%)
- σ: Population standard deviation (or estimate)
- n: Sample size

Importance:

- Indicates the reliability of an estimate.
- Helps quantify uncertainty in measurements.
- Widely used in reporting research findings.

For example, a 95% CI of [2.5, 4.0] suggests there is a 95% chance the true parameter lies within that range.
