Data Science Bootcamp
Week 7

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

Answer:

- Statistical significance is assessed by formulating a null hypothesis (no effect or difference) and an alternative hypothesis (effect or difference exists).
- We then compute a p-value using appropriate statistical tests (e.g., t-test, chi-square test, ANOVA) based on the data.
- If the p-value is below a predetermined threshold, commonly 0.05, we reject the null hypothesis and consider the insight statistically significant.
- Confidence intervals and effect sizes are used to provide context beyond just p-values.

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

Answer:

- The Central Limit Theorem states that the sampling distribution of the sample mean approaches a normal distribution as the sample size becomes large, regardless of the population's original distribution (assuming finite mean and variance).
- In simpler terms, if we take enough samples and compute their means, those means will follow a normal curve.
- Importance:
    - It justifies the use of normal-based methods (e.g., z-tests, t-tests) even when data itself is not normally distributed.
    - It enables confidence intervals and hypothesis tests to be reliably performed even on non-normal data with large samples.

3. What is the statistical power?

Answer:

- Statistical power is the probability of correctly rejecting a false null hypothesis.
- It is calculated as 1 - β, where β is the probability of a Type II error (failing to detect a true effect).
- High power (typically 80% or higher) means there's a good chance the test will catch true differences or effects when they exist.
- Power depends on the sample size, effect size, significance level (alpha), and variability in the data.

4. How do you control for biases?

Answer:

- Randomization: Randomly assign subjects to groups to avoid selection bias.

- Blinding: Use single-blind or double-blind procedures to prevent measurement and observer bias.
- Control groups: Compare outcomes against a baseline to account for placebo effects.
- Standardized procedures: Apply consistent methods of measurement and treatment.
- Statistical controls: Adjust for biases during analysis (e.g., using regression to control for covariates).
- Sampling techniques: Use stratified sampling or oversampling for underrepresented groups to reduce sampling bias.

5. What are confounding variables?
Answer:

- Confounding variables are external factors that are related to both the independent and dependent variables, distorting the observed relationship.
- They make it difficult to determine causality because they offer alternative explanations for the results.
- Example: In studying whether exercise reduces heart disease, age might be a confounder if older individuals both exercise less and have higher heart disease rates.

6. What is A/B testing?
Answer:

- A/B testing is a method to compare two versions (A and B) of a product, webpage, or feature to determine which performs better.
- Users are randomly assigned to either group A (control) or group B (treatment).
- Key metrics (e.g., conversion rate, click-through rate) are compared statistically to determine if observed differences are significant.
- It is widely used in marketing, product design, and UX optimization.

7. What are confidence intervals?
Answer:

- A confidence interval (CI) is a range of values that is likely to contain the true population parameter (like the mean) with a specified level of confidence (usually 95%).
- It reflects both the estimate and the uncertainty around it.
- Example: If a 95% CI for a mean salary is ($50,000, $60,000), we are 95% confident that the true mean lies within that range.
- Narrower intervals suggest more precise estimates, while wider intervals indicate greater uncertainty.