Data Science Bootcamp - Week 7 Follow Up Questions - Pooja Mahesh

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

To assess statistical significance:

- **Formulate Hypotheses**: Define the null hypothesis (H_0) and the alternative hypothesis (H_1).
- **Choose a Significance Level (α)**: Typically 0.05 (5%).
- **Select an Appropriate Test**: Example: t-test, chi-square test, ANOVA, etc.
- **Calculate the Test Statistic**: Compute the statistic (like t-value) based on the data.
- **Find the p-value**: Determine the probability of observing the data (or something more extreme) assuming the null hypothesis is true.
- **Compare p-value and α**:
  - If **$p < \alpha$**, reject H_0 → statistically significant.
  - If **$p \geq \alpha$**, fail to reject H_0 → not statistically significant.

🔍 **Key point**: Statistical significance tells us how unlikely the result is under the null hypothesis but does not directly confirm the truth of the alternative hypothesis.

---

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

The **Central Limit Theorem (CLT)** states:

> *If you take sufficiently large random samples from any population (regardless of the population's original distribution), the distribution of the sample means will approximate a normal distribution.*

✅ **Why is it important?**

- Enables the use of **normal-based statistical tests** (like t-tests and z-tests) even when the population is not normally distributed.
- Supports **confidence intervals** and **hypothesis testing**.
- Allows generalization from a sample to the population.

---

3. What is statistical power?

Statistical power is:

> *The probability of correctly rejecting the null hypothesis when it is false (i.e., detecting a true effect).*

Mathematically:
Power = 1 - β, where **β** is the probability of a Type II error (failing to reject a false null hypothesis).

✅ **Why is it important?**

- Higher power reduces the risk of false negatives.
- Common target for power is **80% or higher**, meaning there's an 80% chance of detecting an effect if it exists.
- Influenced by:
    - Sample size (larger samples = higher power).
    - Effect size (larger effect = higher power).
    - Significance level (α).
    - Variance in the data.

---

4. How do you control for biases?

To control for biases:

- **Randomization**: Randomly assign participants to groups to reduce selection bias.
- **Blinding**:
  - Single-blind: Participant doesn't know which group they're in.
  - Double-blind: Neither participant nor experimenter knows group assignment.
- **Control Groups**: Include placebo or baseline groups for comparison.
- **Matching**: Match participants on key variables (e.g., age, gender).
- **Statistical Controls**: Use regression techniques to control for covariates.
- **Pre-registration**: Document analysis plans before data collection to avoid data dredging.

---

5. What are confounding variables?

Confounding variables are:

> *External factors that influence both the independent variable (IV) and the dependent variable (DV), creating a false association between them.*

✅ **Why are they problematic?**

- They can lead to incorrect conclusions about causality.
- Example: Studying the effect of exercise on heart health without controlling for diet.

✅ **How to handle confounders?**

- Randomization.
- Stratification.
- Statistical adjustment (like ANCOVA or multiple regression).
- Matching groups on confounding variables.

---

6. What is A/B testing?

A/B Testing is:

> *A randomized controlled experiment where two versions (A and B) are compared to see which performs better.*

✅ **How it works**:

- **A**: Control group (current version).
- **B**: Treatment group (new version).
- Users are **randomly assigned** to either group.
- Outcome (like click rate, conversion rate) is measured.
- Use statistical tests (e.g., t-test) to determine if the difference is significant.

---

7. What are confidence intervals?

A **confidence interval (CI)** is:

> *A range of values, derived from the sample data, that is likely to contain the true population parameter with a specified probability (e.g., 95%).*

✅ **Interpretation**:

- A **95% CI** means that if we repeated the experiment many times, **95% of the calculated intervals would contain the true parameter**.

✅ **Why use CIs?**

- Provides a **range estimate** (not just a single value like the mean).
- Gives information about the **precision and uncertainty** of the estimate.
- Complements p-values by showing the size of the effect and the reliability of the estimate.