Homework 7

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

You assess **statistical significance** by using **hypothesis testing**. The process typically includes:

- **Formulating a null hypothesis (H_0)** (e.g., "There is no effect or difference").
- **Calculating a p-value** from your test statistic.
- **Comparing the p-value to a chosen significance level (α)**, often 0.05.

If p-value < α, you reject the null hypothesis — meaning the result is statistically significant and unlikely due to chance.

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

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

When independent random variables are added, their sum (or average) tends toward a **normal distribution**, regardless of the original distribution, **as the sample size increases**.

For large enough sample sizes (usually $n \geq 30$), the **sampling distribution of the mean** is approximately normal, even if the population itself is not.

Importance:

- Allows us to use **normal distribution-based techniques** (like confidence intervals and z-tests).
- Makes statistical inference possible even when data isn't normally distributed.

3. What is statistical power?

Statistical power is the **probability of correctly rejecting the null hypothesis** when it is false.

Mathematically:

Power = $1 - \beta$
Where β = probability of Type II error (failing to reject a false null hypothesis)

Higher power means you're more likely to detect a true effect when one exists.

Good experiments aim for power ≥ 0.8 (80%).

4. How do you control for biases?

To **control for biases**, you can:

- **Randomize** data collection (random sampling, random assignment).
- Use **control groups** in experiments.
- **Blind or double-blind** study designs (participants and/or researchers don't know conditions).
- **Standardize procedures** to reduce variability.
- Use **statistical techniques** (e.g., matching, stratification, regression) to adjust for known biases.
- Always **inspect and clean data** to identify bias in data collection or entry.

5. What are confounding variables?

A **confounding variable** is an outside factor that affects both the **independent variable** and the **dependent variable**, potentially giving a false impression of the relationship between them.

To control for confounders, use techniques like stratification, matching, or multiple regression.

6. What is A/B testing?

A/B Testing is a type of **randomized controlled experiment** where you compare two versions (A and B) to see which performs better.

- **Group A (control):** current version
- **Group B (variant):** new version

You measure a **key metric** (e.g., click-through rate, conversion rate), and then use **statistical tests** (e.g., t-test) to determine if the difference is significant.

7. What are confidence intervals?

A **confidence interval (CI)** gives a **range of values** within which the true population parameter is likely to fall, **with a certain level of confidence** (usually 95%).

CI = point estimate ± margin of error
The **wider** the interval, the **less precise** the estimate.