**Take-Home Assignment**

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

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

**3.⁠ ⁠What is the statistical power?**

**4.⁠ ⁠How do you control for biases?**

**5.⁠ ⁠What are confounding variables?**

**6.⁠ ⁠What is A/B testing?**

**7.⁠ ⁠What are confidence intervals?**

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

To assess statistical significance, you typically:

* Formulate a null hypothesis (e.g., no effect or no difference) and an alternative hypothesis.
* Choose a significance level (commonly α = 0.05).
* Conduct a statistical test (e.g., t-test, chi-squared test, ANOVA).
* Calculate the p-value, which represents the probability of observing your results (or more extreme) if the null hypothesis were true.
* Interpret the p-value:
  + If p < α, reject the null hypothesis → result is statistically significant.
  + If p ≥ α, fail to reject the null → not statistically significant.

Statistical significance tells you if an insight is unlikely to be due to random chance.

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

Central Limit Theorem (CLT) states that the sampling distribution of the sample mean (or sum) of a large number of independent, identically distributed (i.i.d.) random variables approaches a normal distribution, regardless of the original distribution, as the sample size increases (usually n ≥ 30 is sufficient).

Importance:

* Enables use of normal distribution assumptions for confidence intervals and hypothesis tests.
* Allows for statistical inference about population parameters using sample data.
* Underpins many machine learning algorithms and statistical models.

**3. What is statistical power?**

Statistical power is the probability that a test correctly rejects a false null hypothesis (i.e., detects a true effect).

Mathematically:  
Power = 1 − β, where β is the probability of a Type II error (false negative).

High power means:

* You're more likely to detect an effect when it exists.
* Reduces risk of missing important insights.

Factors affecting power:

* Sample size (larger = higher power)
* Effect size (larger effect = higher power)
* Significance level (α)
* Variability in the data (less variability = higher power)

**4. How do you control for biases?**

Biases can distort your results. To control for them:

* Randomization: Randomly assign subjects to groups to reduce selection bias.
* Blinding: Use single/double-blind setups to avoid placebo or observer bias.
* Control groups: Compare with a baseline to isolate treatment effects.
* Stratification: Group data by confounding variables to reduce imbalance.
* Data cleaning: Remove outliers or handle missing data appropriately.
* Cross-validation (in ML): Prevents overfitting by testing model on unseen data.
* Awareness and documentation: Recognize potential sources of bias and document them.

**5. What are confounding variables?**

Confounding variables are external variables that influence both the independent and dependent variables, potentially distorting the true relationship.

Example:  
If you're testing whether coffee causes heart disease, a confounder might be smoking—smokers may drink more coffee and also have higher heart disease risk.

Control methods:

* Randomization
* Matching groups
* Including confounders in statistical models (e.g., regression)
* Stratified analysis

**6. What is A/B testing?**

A/B testing is a controlled experiment comparing two variants (A and B) to evaluate which performs better on a specific metric (e.g., conversion rate).

Steps:

1. Split population randomly into two groups.
2. Group A gets the control; Group B gets the variant.
3. Measure outcomes and compare statistically (often using t-tests or proportion tests).
4. Determine if difference is statistically significant.

It’s widely used in web design, marketing, and product development for data-driven decision making.

**7. What are confidence intervals?**

A confidence interval (CI) provides a range of values that likely contains a population parameter (e.g., mean) with a given level of confidence, typically 95%.

Example:  
If a 95% CI for a mean is [4.5, 5.5], it means we are 95% confident that the true population mean lies between 4.5 and 5.5.

Interpretation tips:

* A wider CI = more uncertainty.
* A narrower CI = more precision (usually from larger sample size or less variance).
* Useful for understanding both effect size and uncertainty.