Week 7 : Take - Home Assignment

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

To assess statistical significance:

a. Formulate hypotheses: define a null hypothesis (H_0) and an alternative hypothesis (H_1).

b. Select a significance level (α), commonly 0.05.

c. Conduct a statistical test (e.g., t-test, chi-square test, ANOVA).

d. Calculate a p-value.

e. Compare p-value to α:

- If $p \leq \alpha$: reject $H_0 \rightarrow$ statistically significant.
- If $p > \alpha$: fail to reject $H_0 \rightarrow$ not statistically significant.

Always complement p-values with effect size and confidence intervals for practical significance.

Example: If you're testing whether a new website design improves conversion rate, and the p-value from your A/B test is 0.03 with $\alpha = 0.05$, you can conclude the new design has a statistically significant effect.

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

Definition: The Central Limit Theorem states that the distribution of the sample mean of a large number of independent and identically distributed (i.i.d.) variables tends to follow a normal distribution, regardless of the original distribution, as long as the sample size is sufficiently large.

Mathematically: If $X_1, X_2, \ldots, X_n$ are i.i.d. with mean μ and variance σ^2, then:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \rightarrow N(0,1) \quad \text{as } n \rightarrow \infty$$

Importance:

- Enables the use of normal-based statistical methods on non-normal data.
- Supports confidence intervals, hypothesis testing, and A/B testing.
- Forms the foundation for inferential statistics.

Example: Even if customer purchase amounts follow a skewed distribution, the average purchase amount from 50 randomly selected customers will approximate a normal distribution due to the CLT.

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):

$$\text{Power} = 1 - \beta$$

Where β is the probability of a Type II error (false negative).

Factors affecting power:

- Sample size (larger $\rightarrow$ higher power)
- Effect size (larger effects $\rightarrow$ easier to detect)
- Significance level (α)
- Variability in data

Example: A drug trial with 1000 patients has more power to detect the drug's effect than one with 100 patients because the sample size is larger.

4. How do you control for biases?

- **Randomization**: reduces selection bias.
- **Blinding**: prevents observer and participant bias.
- **Matched sampling or stratification**: ensures comparability.
- **Statistical controls**: regression or ANCOVA.
- **Data cleaning**: remove outliers, duplicates, errors.
- **Cross-validation**: reduces overfitting and selection bias.

Also watch for confirmation bias, survivorship bias, and sampling bias.

Example: In a clinical trial, double-blinding prevents doctors and patients from knowing who receives the treatment, reducing placebo and observer bias.

5. What are confounding variables?

Confounding variables are third variables that affect both the independent and dependent variables, leading to a spurious association.

Example: Ice cream sales and drowning deaths are positively correlated, but the confounding variable is temperature—hotter weather increases both.

Controlling confounding:

- Randomization
- Stratification or matching
- Including confounders in regression models

Another Example: In studying whether exercise reduces heart disease risk, age could be a confounder if older people exercise less and also have higher heart disease risk.

6. What is A/B testing?

A/B testing is a method to compare two versions (A and B) to determine which performs better on a specific metric.

Steps:

1. Define a hypothesis and metric (e.g., conversion rate).
2. Randomly assign users to group A (control) or B (variant).
3. Collect data with sufficient sample size.
4. Use statistical tests (e.g., t-test) to compare.
5. Evaluate significance and make decisions.

Applications: Web design, marketing, feature testing, pricing, etc.

Example: If version B of a product page increases user sign-ups from 10% to 13% with $p = 0.02$, it's statistically significant, and B is preferred.

7. What are confidence intervals?

A confidence interval (CI) gives a range of plausible values for an unknown population parameter.

Example: A 95% CI for a mean might be [45, 55], meaning we are 95% confident the true mean lies within this interval.

Formula (for known σ):

$$\bar{x} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Notes:

- Wider CIs = less precision
- Narrower CIs = more precision (larger sample or lower variability)

Another Example: After surveying 1000 voters, a political candidate's approval rating is estimated at 60% with a 95% CI of [58%, 62%].