Ruchi Jha

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

Answer 1-To assess the statistical significance of an insight, we typically perform hypothesis testing. This involves defining:

- Null Hypothesis (H_0): A default assumption (e.g., no effect or difference).
- Alternative Hypothesis (H_1): Represents the insight we want to test.

We calculate a test statistic (like a t-score or z-score), then compute a p-value, which is the probability of observing our results (or more extreme) if H_0 were true.

If the p-value $< \alpha$ (commonly $\alpha = 0.05$), we reject the null hypothesis, concluding the result is statistically significant.

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

Answer 2-The Central Limit Theorem (CLT) states that the sampling distribution of the sample mean approaches a normal distribution as the sample size increases, regardless of the population's original distribution.

If $X_1, X_2, ..., X_n$ are independent, identically distributed random variables with mean μ and standard deviation σ, then:

$\bar{X} \approx N(\mu, \sigma^2 / n)$ as $n \to \infty$

This means the distribution of sample means becomes approximately normal, with:

- Mean: μ
- Standard Error: $\sigma / \sqrt{n}$

It enables us to use normal-based inference (z-tests, confidence intervals) even when the population is not normal.It's foundational for statistical estimation and hypothesis testing.

Question 3. What is statistical power?

Answer 3-Statistical power is the probability of correctly rejecting the null hypothesis when it is actually false. It is the ability of a test to detect an effect if there is one.

Power $= 1 - \beta$

Where:

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

Higher power (usually ≥ 0.8) means a lower risk of missing real effects. Power increases with:

- Larger sample size
- Larger effect size
- Higher significance level (α)
- Lower variability

Question 4. How do you control for biases?

Answer 4-To control for biases, you must design your study or experiment carefully to minimize systematic errors. Techniques include:

- Randomization: Randomly assigning participants to groups to reduce selection bias.
- Blinding: Hiding group assignments from participants or experimenters (single-blind or double-blind).
- Controlling confounders: Using statistical adjustments or design strategies to isolate the effect of the independent variable.
- Standardization: Keeping procedures and measurements consistent across conditions.

These steps help ensure the results are due to the variables being studied, not hidden biases.

Question 5. What are confounding variables?

Answer 5-Confounding variables are external factors that affect both the independent and dependent variables, potentially distorting the apparent relationship between them.

For example if you observe that people who carry lighters tend to have lung cancer, a confounder could be smoking, which influences both.

Confounders must be identified and accounted for through:

- Study design (e.g., randomization)
- Statistical techniques (e.g., regression, stratification)

Question 6. What is A/B testing?

Answer 6-A/B testing is a type of controlled experiment where users are randomly split into two groups:

- Group A (control): sees the current version.
- Group B (variant): sees the new version.

The performance metric (e.g., click-through rate, purchase rate) is measured, and a statistical test (often a t-test or z-test) is used to evaluate whether the difference is statistically significant.

It's widely used in marketing, UX design, and product development to make data-driven decisions.

Question 7. What are confidence intervals?

Answer 7-A confidence interval (CI) gives a range of values within which we expect the true population parameter (like the mean) to lie, with a specified level of confidence.

For a 95% CI of the population mean:

$CI = \bar{X} \pm z \times (\sigma / \sqrt{n})*$

Where:

- $\bar{X}$ = sample mean
- z^* = critical value from z-distribution (1.96 for 95% CI)
- σ = standard deviation
- n = sample size

This means we are 95% confident that the true population mean lies within this interval. Confidence intervals quantify uncertainty in estimation and are used alongside p-values in reporting results.