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

* Formulate a Hypothesis: Null hypothesis and alternative hypothesis
* Choose a statistical test
* Compute p-value and compare against significance level
* Consider effect size and confidence intervals

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

CLT given a sufficiently large sample size, the sampling distribution of sample mean of iid random variables approaches a normal distribution, regardless of the original distribution of the population.

It allows us to make statistical inferences using the normal distribution even if the data is not normally distributed. So it actually simplifies the analysis, and also support many practical tools like z-scores, t-tests etc.

3.⁠ ⁠What is the statistical power?

Statistical power = 1- \beta

It is the probability that a statistical test correctly rejects a false null hypothesis. It can help avoid false negatives, like low power means you might miss real effects, leading to incorrect conclusions.

4.⁠ ⁠How do you control for biases?

* Randomly assign subjects to groups or randomly sample from the population
* Use regression models or matching techniques to control for known confounding variables
* Use the same tools, procedures and training across all subject or time perio to standardize data collection
* Use cross-validation for models on unseen data to avoid overfitting

5.⁠ ⁠What are confounding variables?

Confounding variables are variables that influence both the independent variable and the dependent variable, potentially giving a false impression of a causal relationship between them.

6.⁠ ⁠What is A/B testing?

It is a controlled experiment used to compare two versions of something.

First split audience randomly into two groups: A (control group, can see the original version), B (treatment group, see a new version). Then measure the performance on a key metric, and use statistical testing to assess whether the observed difference is statistically significant.

7.⁠ ⁠What are confidence intervals?

CI gives a range that is likely to contain the true value of the population parameter, based on sample data. For example, a 95% confidence interval means that if you were to repeat the sampling process many times, about 95% of those intervals would contain the true parameter.