1. How do you assess the statistical significance of an insight?
To assess the statistical significance, we need to first design the hypothesis: H_0: the null hypothesis and H_1: the alternative hypothesis.
Then we can choose an appropriate test for the hypothesis such as z-test, student t-test, χ^2-test and F-test. With the test, we can calculate the corresponding statistic and use this statistic value to get the p-value, i.e., the probability of having such data under H_0.
Now we can choose the significance level $1-\alpha$ as we want, e.g., 95
If $p \leq \alpha$, we reject H_0 (statistically significant).

2. What is the Central Limit Theorem? Explain it. Why is it important?
CLT:
Let $X_1, \cdots, X_n$ be iid random variables with $\mathbb{E}[X] = \mu < \infty$ and $Var(X) = \sigma^2 < \infty$.
Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
Then

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

Reason for why it is important:
CLT is universal. No matter what the population distribution is, the above statistic will converge in distribution to standard normal as long as n is large enough.

3. What is the statistical power?

$$Power = \mathbb{P}(reject\ H_0 | H_1\ true) = 1 - \beta,$$

where β is the type II error.
High power means low false negative rate.

4. How do you control for biases?

- Collecting data stage: Random sampling: reduce the selection bias; Make sure the source of data is unbiased.

- Preprocessing data stage: Appropriately dual with the missing values, e.g., don't directly delete the records with missing values, but use mean/median/mode or time series interpolation to replace them.
- Modeling stage: Using models with high interpretablity, e.g. decision tree; Constraint model: adding penalty into the model.
- Assessment stage: Avoid the trap of accuracy: extremely high accuracy may mean overfit, insteadly, we can use also precision, recall, score function, etc.

5. What are confounding variables?
Confounding variable is the variable that influences both exposure X and outcome Y.
For example, we want to explore what makes more shark attack occur and we find that the better icecream sales, the more shark attack occur. It is unreasonable to have such a causual relationship, so there must be a confounding variable behind this relationship, e.g., weather, # of tourists, etc.

6. What is A/B testing?
A/B test is a method to determine which version (e.g., version A and version B) is better through comparing the performance of different versions.

7. What are confidence intervals?
Confidence interval is a closed interval that, with repeating sampling, this interval will contain the target parameter θ in the proportion of $1 - \alpha$.