- Design of Experiments is related to statistics, you should make sure to review concepts like a p-value and statistical tests before attempting these questions!

- Google wants to add a new button to their main search page. How can they determine whether or not people enjoy this new button feature?

- This type of question is common, essentially just testing to see if you understand the concept of an A/B test.
- A|B testing is a very large topic, so be sure to study it well if the position posting asked for it specifically.

- First we decide on metrics to compare across the dual versions of the site we will serve.
- Common metrics are DAU,MAU,CTR,Impressions, Engagement, etc...

- Second step is to prepare two (or more) versions of your site/page.
- One version without the feature change, another version with the new feature change.

- You will then serve the different versions of the page to the populations.
- You need to randomly split the population, being careful not to introduce biases when randomly splitting.

- Once this set-up is done you can begin to define your statistical hypothesis test.
- Create a null hypothesis for your hypothesis testing, for example, the metric chosen will be the same for both groups.

- For the hypothesis test you will pre-define your acceptable Type I Error (FPR or alpha)
- Using this you can use the p-value of your test to decide whether the results are significant.

- We will also want to assess the power of the statistical test.
- This is 1 - Beta , where Beta is the Type II Error.

- In practice you will also want to consider factors such as how long to run the experiment (confidence intervals can be used for this).
- You will also need to decide on a protocol for outliers, such as truncation.

- How can you know if a sample is biased? What different kind of biases should you be aware of when using a sample?

- If you know the true mean of your population from which you sampled, you can take samples of your sample multiple times and check if the mean of these samples are normally distributed around the true mean of the population.
- Check out the resources on Bootstrapping

- Selection Bias
  - By some error, you have excluded a specific part of the population
  - Sampling for average weight of USA by only sampling one state.

- Measurement Bias
  - The method of measurement creates observations that are different than the true value coming in.

- Measurement Bias
    - Sampling from a stream of information, but your sampling rate is lower than the rate of change in the stream.

- Facebook is testing different designs of the user homepage. They have come up with 50 variations. They would like to test all of them and choose the best. How would you setup this experiment? What metrics would you calculate and how would you report the results?

- This is slightly different than a straight A|B test because we are dealing with more than just 2 variations.
- There are multiple statistical techniques that exist for this sort of problem, let's discuss some of them.

- T-test Among Pairs of Treatment
  - This is similar to an AB test except we randomly assign all 50 versions to users.
  - There are a few things to consider with this approach however...

- T-test Among Pairs of Treatment
    - With so many versions you will need to make sure you have enough users to get a statistically significant result (this usually is not an issue for Facebook)

- T-test Among Pairs of Treatment
    - With multiple hypotheses test you increase the likelihood of a rare event occurring, meaning you need to adjust your alpha value you use to compare your p-value against for significance accordingly.

- T-test Among Pairs of Treatment
    - Since this is a multiple inference problem you can correct your alpha value using the Bonferroni correction method.

- T-test Among Pairs of Treatment
  - Typically we set alpha to 0.05, but with the Bonferroni correction we divide this by the number of tests.
  - In our case it's 0.05/50 = 0.001
  - Clearly we'll need a large population for testing!

- What is the definition of power for a statistical test? What factors affect the power and how does the power relate to the p-value?

- The power of any test of statistical significance is defined as the probability that it will reject a false null hypothesis.
- Statistical power is inversely related to beta or the probability of making a Type II error. In short, power = 1 – β.

- Statistical power is the likelihood that a study will detect an effect when there is an effect there to be detected. If statistical power is high, the probability of making a Type II error, or concluding there is no effect when, in fact, there is one, goes down.

- Statistical power is affected chiefly by the size of the effect and the size of the sample used to detect it.
- Bigger effects are easier to detect than smaller effects, while large samples offer greater test sensitivity than small samples.

- If you perform a test, the p-value is the smallest $\alpha$ for which the test would reject the null hypothesis.
- It is the value of the significance level for which the test statistic would have been on the boundary of the rejection region.

- By comparing the p-value with Type I error we can decide if we are going to reject the null hypothesis.
- If we set a larger $\alpha$ we then in turn allow for larger p-values to be considered significant, meaning we increase the power of the test.