Interview Questions

From 120 Data Science Interview Questions – Statistical Inference

1. In an A/B test, how can you check if assignment to the various buckets was truly at random?
    - Demographics such as age, gender, user-type, device type, etc. should be distributed similarly in each bucket. For example, if the proportion of iPhone users in bucket 1 is 35%, you should expect the proportion of iPhone users in bucket 2 to also be close to 35%. So, to decide whether assignment to buckets was actually at random, you could compare the distributions of relevant demographics between the two buckets.
2. What be the benefits of running an A/A test, where you have two buckets who are exposed to the exact same product?
    - The benefit here is mainly to evaluate whether you believe your randomization procedure is working.
    - To confirm this, you could do what was suggested above, but you could also conduct a hypothesis test to formally compare your metric of interest across the two groups. You should not reject the null hypothesis of equality. If you do, there's a problem with your randomization mechanism.
3. What would be the hazards of letting users sneak a peek at the other bucket in an A/B test?
    - When users become aware of both conditions your experiment becomes *contaminated*. This is a problem because a user may behave differently in condition A if they also saw (or became aware of) condition B, and this different behavior prevents the experimenter from clearly understanding the effect of the experimental condition. In other words, a contaminated user's behavior is no longer representative of what their experience would have been if they had only ever been aware of one condition, and this biases the results.
4. What would be some issues if blogs decide to cover one of your experimental groups?
    - The main issue here, like above, is contamination. Users may behave differently if they know they're being experimented on, and this confounds your causal inference. Did they behave the way they did because of the condition they were in? Or did they behave the way they did because the blog made them aware of the experiment? Or is the cause of their behavior a mixture of these two? We can't know.
5. How would you conduct an A/B test on an opt-in feature?
    - Naively you could run this experiment where one group has the opportunity to opt-in (the treatment group) and the other group does not have this opportunity (the control group). The problem, however, is that the data you collect on the treatment group will only be on individuals that chose to opt-in, and these people, in general, are not representative of your whole user population (which means your results will not generalize).

- For this reason, I would not run this as a formal experiment. I would instead run an observational study (sometimes called a *natural experiment*) where you use propensity score matching (or some similar method) to match opt-in user to users who did not have the opportunity to opt-in and compare the behavior of these two matched groups. This doesn't result in perfect causal inference, but it does a better job of approximating the counterfactual than the previously described experiment does.

6. How would you run an A/B test for many variants, say 20 or more?
    - From a practical standpoint the running of the experiment is no different than if there were only two variants. That said, we have to be very careful with multiple testing here and perform Bonferroni (or similar) corrections in the analysis, and account for such corrections during sample size calculations.
7. How would you run an A/B test if your observations are extremely right-skewed?
    - In this case, t-tests (which assume your observations are normally distributed) are not appropriate. As an alternative you can use nonparametric approaches such as randomization / permutation tests which do not make any assumptions about the distribution of the underlying data.
    - If you wanted to perform a parametric test that is more appropriate for right-skewed data, you could model the data in each condition with a gamma distribution and perform a likelihood ratio test to decide whether the gamma distributions were the same or not.
8. I have two different experiments that both the change the sign-up button to my website. I want to test them at the same time. What kinds of things should I keep in mind?
    - What you should be worried about here is your ability to isolate the effect of each experiment. Can you say for certain that the effect you observe in your experiment is not due to the changes your colleague made in his/her experiment?
    - In general, it is best not to run different experiments simultaneously because their effects may become confounded. An alternative, if you're planning two experiments that change different things about the same button, is to combine them into one multi-factor experiment where you can explicitly control and account for the effects of the different factors.
9. What is a p-value? What is the difference between type-1 and type-2 error?
    - A p-value is a probability of observing a result at least as extreme as the one you did if the null hypothesis is true. In other words, it's a probability that quantifies how likely it is that effect you observe occurred just by chance.
    - Type I errors occur when a true null hypothesis is rejected (i.e., a false positive).
    - Type II errors occur when a false null hypothesis is accepted (i.e., a false negative)
10. You are AirBnb and you want to test the hypothesis that a greater number of photographs increases the chances that a buyer selects the listing. How would you test this hypothesis?

- Choose the metric that will answer this question. Something like number of bookings per some time period would likely be sufficient.
- Your design factor here is number of photographs – choose a number of different levels for this factor (i.e., 3, 5, 10) and use these to define different experiment conditions.
- Listers would then be assigned to one of the conditions and would be appropriately constrained in the number of photos they could post. It would be important here to incorporate blocking so as to compare only listings that are similar to one another. Possible nuisance factors include house size, neighborhood, rental price, etc.
- You would then want to, within a block, compare the number of bookings per listing per time frame across the different conditions.

11. How would you design an experiment to determine the impact of latency on user engagement?
    - Choose some metric that measures engagement such as average time on page, bounce rate, etc.
    - Construct different experimental conditions with artificially imposed page load times (250ms, 500ms, 1000ms, 1500ms, 2000ms, 2500ms) and randomly assign users to each condition.
    - Compare your engagement metric across these conditions.
12. What is maximum likelihood estimation? Could there be any case where it doesn't exist?
    - Maximum likelihood estimation is an approach to parameter estimate which provides as the estimate the value most likely to have produced the data that you actually observed.
    - ML estimation requires that you believe some underlying model or distribution generated the data you observed. If, on the other hand, you do not want to perform parametric inference then MLE, in a sense, does not exist.
13. What's the difference between a MAP, MOM, MLE estimator? In which cases would you want to use each?
    - MAP (maximum a posteriori), MOM (method of moments) and MLE (maximum likelihood estimation) are all different methods of parameter estimation. MAP is Bayesian, MOM/MLE are frequentist, MOM is nonparametric and MLE/MAP are parametric. Depending on whether you work from a Bayesian vs. frequentist and/or parametric vs. nonparametric perspective will dictate your choice as to which estimation procedure to use.
14. What is a confidence interval and how do you interpret it?
    - A confidence interval is a range of plausible values for some parameter of interest. A point estimate of a parameter provides someone's best guess at the unknown parameter value, but a confidence interval provides a range within which we can reasonably sure the unknown true value lies.
    - The interpretation of a 95% confidence interval is that if infinitely many random samples were drawn from the population, and a confidence interval is

constructed from each of them, 95% of those infinitely many confidence intervals would contain the unknown true parameter value.

15. What is unbiasedness as a property of an estimator? Is this always a desirable property when performing inference? What about in data analysis of predictive modeling?
    - Unbiased of an estimator means that if infinitely random samples were drawn from the population, and the parameter estimate was calculated from each of them, the average of those infinitely many estimates would be the unknown true parameter value.
    - This is always a desirable property. Period.

Miscellaneous interview questions asked of current and former students:

1. How would you design an experiment to pick the best set of hyper-parameters for a machine learning model, say random forest or gradient boosted trees?

2. Explain multi-armed bandit experiments and their use.

3. Show why Bonferroni corrections work and explain when/why we need them.

4. Given a metric to measure, properly identify the statistical test used to formally test hypotheses concerning it.

5. How do you determine sample size?

6. Suppose the marketing team runs a test and calculates a p-value of 0.06 and the significance level of the test is 0.05. Would you recommend rejecting or not rejecting the null hypothesis?

7. Given a business use-case, design / define relevant metrics.

8. Explain what a p-value is in layman's terms.

9. You find users in your control group are contaminated (e.g., users see and ad but they're not supposed to). How should we handle this? What are the possible methodologies when randomization is impossible to achieve?

10. Suppose that in your experiment you find the click-through-rate in your experimental group is 0.07 which is much better than in control, and so you roll out the experimental condition to all users. After doing so you find the click-through-rate is significantly lower than the 0.07 you observed in the experiment. How would you explain / investigate this?

11. The ideal experiment should have 50% male and 50% female in both the control and experimental group. However, after collecting data, you find that your experimental group is 80% male and 20% female. In this case, how do you analyze the experiment?

12. If we want to experiment with a re-designed email and we have 2 factors to change, how would you design the experiment? What metrics will you use for this experiment? Why?

13. What is the relationship between sample size and other variables?

14. If your business partner sees a significant result and wants to stop the test early, how do you convince him this is a bad idea (in plain English)?

15. If a test has a small p-value and also low power, how will you describe the result of the test? Will increasing power change anything?

16. The Driver Experience team has just finished redesigning the Uber Partner app. The new version expands the purpose of the app beyond just driving. It includes additional information on earnings, ratings, and provides a unified platform for Uber to communicate with its partners.
    a. Propose and define the primary success metric of the redesigned app. What are 2-3 additional tracking metrics that will be important to monitor in addition to the success metric defined above?
    b. Outline a testing plan to evaluate if the redesigned app performs better (according to the metrics you outlined). How would you balance the need to deliver quick results, with statistical rigor, and while still monitoring for risks?
    c. Explain how you would translate the results from the testing plan into a decision on whether to launch the new design or roll it back.

17. Given different layouts of results from a Google search for some product, how would you design an experiment to maximize revenue for Google?

Part 2. Suppose an advertiser wants to measure different creative types on our platform. Suppose they have three different creative types (e.g., one has logo, one has brand name instead of a logo, control has no mention of the brand). Further suppose that we want test the effectiveness of different ad formats -- static pins (control), cinematic (ad moves when scrolling), or 30-second auto-play video pins.

1. **Assume we have the ability to randomly assign Pinners to different experiences. What kind of experimental design would you propose? What would the different experimental cells be? Be sure to explain your design and how it will allow us to assess the effectiveness of creative types and ad formats.**

2. **What metric will you choose to measure the effectiveness?**

3. **What statistical significance test would you run to compare the designs?**