**Product Metrics (15 questions)**

**1. What would be good metrics of success for an advertising-driven consumer product? (Buzzfeed, YouTube, Google Search, etc.) A service-driven consumer product? (Uber, Flickr, Venmo, etc.)**

* advertising-driven: Pageviews and daily actives, CTR, CPC (cost per click)
  + click-ads
  + display-ads
* service-driven: number of purchases, conversion rate

**2. What would be good metrics of success for a productiv- ity tool? (Evernote, Asana, Google Docs, etc.) A MOOC? (edX, Coursera, Udacity, etc.)**

* productivity tool: same as premium subscriptions
* MOOC: same as premium subscriptions, completion rate

**3. What would be good metrics of success for an e-commerce product? (Etsy, Groupon, Birchbox, etc.) A subscrip- tion product? (Net ix, Birchbox, Hulu, etc.) Premium subscriptions? (OKCupid, LinkedIn, Spotify, etc.)**

* e-commerce: number of purchases, conversion rate, Hourly, daily, weekly, monthly, quarterly, and annual sales, Cost of goods sold, Inventory levels, Site traffic, Unique visitors versus returning visitors, Customer service phone call count, Average resolution time
* subscription
  + churn, CoCA, ARPU, MRR, LTV
* premium subscriptions:

**4. What would be good metrics of success for a consumer product that relies heavily on engagement and interac- tion? (Snapchat, Pinterest, Facebook, etc.) A messaging product? (GroupMe, Hangouts, Snapchat, etc.)**

* heavily on engagement and interaction: uses AU ratios, email summary by type, and push notification summary by type, resurrection ratio
* messaging product:

**5. What would be good metrics of success for a product that o ered in-app purchases? (Zynga, Angry Birds, other gaming apps)**

* Average Revenue Per Paid User
* Average Revenue Per User

**6. A certain metric is violating your expectations by going down or up more than you expect. How would you try to identify the cause of the change?**

* breakdown the KPI’s into what consists them and find where the change is
* then further breakdown that basic KPI by channel, user cluster, etc. and relate them with any campaigns, changes in user behaviors in that segment

**7. Growth for total number of tweets sent has been slow this month. What data would you look at to determine the cause of the problem?**

**8. You’re a restaurant and are approached by Groupon to run a deal. What data would you ask from them in order to determine whether or not to do the deal?**

* for similar restaurants (they should define similarity), average increase in revenue gain per coupon, average increase in customers per coupon

**9. You are tasked with improving the e ciency of a subway system. Where would you start?**

* define efficiency

**10. Say you are working on Facebook News Feed. What would be some metrics that you think are important? How would you make the news each person gets more relevant?**

* rate for each action, duration users stay, CTR for sponsor feed posts
* ref. News Feed Optimization
  + Affinity score: how close the content creator and the users are
  + Weight: weight for the edge type (comment, like, tag, etc.). Emphasis on features the company wants to promote
  + Time decay: the older the less important

**11. How would you measure the impact that sponsored stories on Facebook News Feed have on user engagement? How would you determine the optimum balance between sponsored stories and organic content on a user’s News Feed?**

* AB test on different balance ratio and see

**12. You are on the data science team at Uber and you are asked to start thinking about surge pricing. What would be the objectives of such a product and how would you start looking into this?**

* there is a gradual step-function type scaling mechanism until that imbalance of requests-to-drivers is alleviated and then vice versa as too many drivers come online enticed by the surge pricing structure.
* I would bet the algorithm is custom tailored and calibrated to each location as price elasticities almost certainly vary across different cities depending on a huge multitude of variables: income, distance/sprawl, traffic patterns, car ownership, etc. With the massive troves of user data that Uber probably has collected, they most likely have tweaked the algos for each city to adjust for these varying sensitivities to surge pricing. Throw in some machine learning and incredibly rich data and you've got yourself an incredible, constantly-evolving algorithm.

**13. Say that you are Net ix. How would you determine what original series you should invest in and create?**

* Netflix uses data to estimate the potential market size for an original series before giving it the go-ahead.

**14. What kind of services would  nd churn (metric that tracks how many customers leave the service) helpful? How would you calculate churn?**

* subscription based services

**15. Let’s say that you’re are scheduling content for a content provider on television. How would you determine the best times to schedule content?Â**

**Communication (5 questions)**

**1. Explain to me a technical concept related to the role that you’re interviewing for.**

**2. Introduce me to something you’re passionate about.**

**3. How would you explain an A/B test to an engineer with no statistics background? A linear regression?**

* A/B testing, or more broadly, multivariate testing, is the testing of different elements of a user's experience to determine which variation helps the business achieve its goal more effectively (i.e. increasing conversions, etc..)  This can be copy on a web site, button colors, different user interfaces, different email subject lines, calls to action, offers, etc.

**4. How would you explain a confidence interval to an engi- neer with no statistics background? What does 95% confidence mean?**

* [link](https://www.quora.com/What-is-a-confidence-interval-in-laymans-terms) <https://www.quora.com/What-is-a-confidence-interval-in-laymans-terms>

**5. How would you explain to a group of senior executives why data is important?**

**Data Analysis (27 questions)**

**1. (Given a Dataset) Analyze this dataset and tell me what you can learn from it.**

**2. What is R2? What are some other metrics that could be better than R2 and why?**

* goodness of fit measure. variance explained by the regression / total variance
* the more predictors you add the higher R^2 becomes.
  + hence use adjusted R^2 which adjusts for the degrees of freedom
  + or train error metrics

**3. What is the curse of dimensionality?**

* High dimensionality makes clustering hard, because having lots of dimensions means that everything is "far away" from each other.
* For example, to cover a fraction of the volume of the data we need to capture a very wide range for each variable as the number of variables increases
* All samples are close to the edge of the sample. And this is a bad news because prediction is much more difficult near the edges of the training sample.
* The sampling density decreases exponentially as p increases and hence the data becomes much more sparse without significantly more data.
* We should conduct PCA to reduce dimensionality

**4. Is more data always better?**

* Statistically,
  + It depends on the quality of your data, for example, if your data is biased, just getting more data won’t help.
  + It depends on your model. If your model suffers from high bias, getting more data won’t improve your test results beyond a point. You’d need to add more features, etc.
* Practically,
  + Also there’s a tradeoff between having more data and the additional storage, computational power, memory it requires. Hence, always think about the cost of having more data.

**5. What are advantages of plotting your data before per- forming analysis?**

* 1. Data sets have errors.  You won't find them all but you might find some. That 212 year old man. That 9 foot tall woman.

1. Variables can have skewness, outliers etc.  Then the arithmetic mean might not be useful. Which means the standard deviation isn't useful.
2. Variables can be multimodal!  If a variable is multimodal then anything based on its mean or median is going to be suspect.

**6. How can you make sure that you don’t analyze something that ends up meaningless?**

* Proper exploratory data analysis.

In every data analysis task, there's the exploratory phase where you're just graphing things, testing things on small sets of the data, summarizing simple statistics, and getting rough ideas of what hypotheses you might want to pursue further.

Then there's the exploitatory phase, where you look deeply into a set of hypotheses.

The exploratory phase will generate lots of possible hypotheses, and the exploitatory phase will let you really understand a few of them. Balance the two and you'll prevent yourself from wasting time on many things that end up meaningless, although not all.

**7. What is the role of trial and error in data analysis? What is the the role of making a hypothesis before diving in?**

* data analysis is a repetition of setting up a new hypothesis and trying to refute the null hypothesis.
* The scientific method is eminently inductive: we elaborate a hypothesis, test it and refute it or not. As a result, we come up with new hypotheses which are in turn tested and so on. This is an iterative process, as science always is.

**8. How can you determine which features are the most im- portant in your model?**

* run the features though a Gradient Boosting Machine or Random Forest to generate plots of relative importance and information gain for each feature in the ensembles.
* Look at the variables added in forward variable selection

**9. How do you deal with some of your predictors being missing?**

* Remove rows with missing values - This works well if 1) the values are missing randomly (see [Vinay Prabhu's answer](https://www.quora.com/How-can-I-deal-with-missing-values-in-a-predictive-model/answer/Vinay-Prabhu-7) for more details on this) 2) if you don't lose too much of the dataset after doing so.
* Build another predictive model to predict the missing values - This could be a whole project in itself, so simple techniques are usually used here.
* Use a model that can incorporate missing data - Like a random forest, or any tree-based method.

**10. You have several variables that are positively correlated with your response, and you think combining all of the variables could give you a good prediction of your response. However, you see that in the multiple linear regression, one of the weights on the predictors is negative. What could be the issue?**

* Multicollinearity refers to a situation in which two or more explanatory variables in a [multiple regression](https://en.wikipedia.org/wiki/Multiple_regression) model are highly linearly related.
* Leave the model as is, despite multicollinearity. The presence of multicollinearity doesn't affect the efficiency of extrapolating the fitted model to new data provided that the predictor variables follow the same pattern of multicollinearity in the new data as in the data on which the regression model is based.
* principal component regression

**11. Let’s say you’re given an unfeasible amount of predictors in a predictive modeling task. What are some ways to make the prediction more feasible?**

* PCA

**12. Now you have a feasible amount of predictors, but you’re fairly sure that you don’t need all of them. How would you perform feature selection on the dataset?**

* ridge / lasso / elastic net regression
* Univariate Feature Selection where a statistical test is applied to each feature individually. You retain only the best features according to the test outcome scores
* "Recursive Feature Elimination":
  + First, train a model with all the feature and evaluate its performance on held out data.
  + Then drop let say the 10% weakest features (e.g. the feature with least absolute coefficients in a linear model) and retrain on the remaining features.
  + Iterate until you observe a sharp drop in the predictive accuracy of the model.

**13. Your linear regression didn’t run and communicates that there are an infinite number of best estimates for the regression coefficients. What could be wrong?**

* p > n.
* If some of the explanatory variables are perfectly correlated (positively or negatively) then the coefficients would not be unique.

**14. You run your regression on different subsets of your data, and find that in each subset, the beta value for a certain variable varies wildly. What could be the issue here?**

* The dataset might be heterogeneous. In which case, it is recommended to cluster datasets into different subsets wisely, and then draw different models for different subsets. Or, use models like non parametric models (trees) which can deal with heterogeneity quite nicely.

**15. What is the main idea behind ensemble learning? If I had many different models that predicted the same response variable, what might I want to do to incorporate all of the models? Would you expect this to perform better than an individual model or worse?**

* The assumption is that a group of weak learners can be combined to form a strong learner.
* Hence the combined model is expected to perform better than an individual model.
* Assumptions:
  + average out biases
  + reduce variance
* Bagging works because some underlying learning algorithms are unstable: slightly different inputs leads to very different outputs. If you can take advantage of this instability by running multiple instances, it can be shown that the reduced instability leads to lower error. If you want to understand why, the original bagging paper( [http://www.springerlink.com/cont...](http://www.springerlink.com/content/l4780124w2874025/)) has a section called "why bagging works"
* Boosting works because of the focus on better defining the "decision edge". By reweighting examples near the margin (the positive and negative examples) you get a reduced error (see <http://citeseerx.ist.psu.edu/vie>...)
* Use the outputs of your models as inputs to a meta-model.

For example, if you're doing binary classification, you can use all the probability outputs of your individual models as inputs to a final logistic regression (or any model, really) that can combine the probability estimates.

One very important point is to make sure that the output of your models are out-of-sample predictions. This means that the predicted value for any row in your dataframe should NOT depend on the actual value for that row.

**16. Given that you have wi  data in your o ce, how would you determine which rooms and areas are underutilized and overutilized?**

* If the data is more used in one room, then that one is over utilized! Maybe account for the room capacity and normalize the data.

**17. How could you use GPS data from a car to determine the quality of a driver?**

**18. Given accelerometer, altitude, and fuel usage data from a car, how would you determine the optimum acceleration pattern to drive over hills?**

**19. Given position data of NBA players in a season’s games, how would you evaluate a basketball player’s defensive ability?**

**20. How would you quantify the influence of a Twitter user?**

* like page rank with each user corresponding to the webpages and linking to the page equivalent to following.

**21. Given location data of golf balls in games, how would construct a model that can advise golfers where to aim?**

**22. You have 100 mathletes and 100 math problems. Each mathlete gets to choose 10 problems to solve. Given data on who got what problem correct, how would you rank the problems in terms of di culty?**

* One way you could do this is by storing a "skill level" for each user and a "difficulty level" for each problem.  We assume that the probability that a user solves a problem only depends on the skill of the user and the difficulty of the problem.\*  Then we maximize the likelihood of the data to find the hidden skill and difficulty levels.
* The Rasch model for dichotomous data takes the form:  
  {\displaystyle \Pr\{X\_{ni}=1\}={\frac {\exp({\beta \_{n}}-{\delta \_{i}})}{1+\exp({\beta \_{n}}-{\delta \_{i}})}},}  
  where  is the ability of person  and  is the difficulty of item}.

**23. You have 5000 people that rank 10 sushis in terms of saltiness. How would you aggregate this data to estimate the true saltiness rank in each sushi?**

* Some people would take the mean rank of each sushi.  If I wanted something simple, I would use the median, since ranks are (strictly speaking) ordinal and not interval, so adding them is a bit risque (but people do it all the time and you probably won't be far wrong).

**24. Given data on congressional bills and which congressional representatives co-sponsored the bills, how would you determine which other representatives are most similar to yours in voting behavior? How would you evaluate who is the most liberal? Most republican? Most bipartisan?**

* collaborative filtering. you have your votes and we can calculate the similarity for each representatives and select the most similar representative
* for liberal and republican parties, find the mean vector and find the representative closest to the center point

**25. How would you come up with an algorithm to detect plagiarism in online content?**

* reduce the text to a more compact form (e.g. fingerprinting, bag of words) then compare those with other texts by calculating the similarity

**26. You have data on all purchases of customers at a grocery store. Describe to me how you would program an algorithm that would cluster the customers into groups. How would you determine the appropriate number of clusters to include?**

* KNN
* choose a small value of k that still has a low SSE (elbow method)
* <https://bl.ocks.org/rpgove/0060ff3b656618e9136b>

**27. Let's say you're building the recommended music engine at Spotify to recommend people music based on past listening history. How would you approach this problem?**

* [collaborative filtering](https://en.wikipedia.org/wiki/Collaborative_filtering)

**Predictive Modeling (19 questions)**

**1. (Given a Dataset) Analyze this dataset and give me a model that can predict this response variable.**

* Start by fitting a simple model (multivariate regression, logistic regression), do some feature engineering accordingly, and then try some complicated models. Always split the dataset into train, validation, test dataset and use cross validation to check their performance.
* Determine if the problem is classification or regression
* Favor simple models that run quickly and you can easily explain.
* Mention cross validation as a means to evaluate the model.
* Plot and visualize the data.

**2. What could be some issues if the distribution of the test data is significantly different than the distribution of the training data?**

* The model that has high training accuracy might have low test accuracy. Without further knowledge, it is hard to know which dataset represents the population data and thus the generalizability of the algorithm is hard to measure. This should be mitigated by repeated splitting of train vs test dataset (as in cross validation).
* When there is a change in data distribution, this is called the dataset shift. If the train and test data has a different distribution, then the classifier would likely overfit to the train data.
* This issue can be overcome by using a more general learning method.
* This can occur when:
  + P(y|x) are the same but P(x) are different. (covariate shift)
  + P(y|x) are different. (concept shift)
* The causes can be:
  + Training samples are obtained in a biased way. (sample selection bias)
  + Train is different from test because of temporal, spatial changes. (non-stationary environments)
* Solution to covariate shift
  + importance weighted cv

**3. What are some ways I can make my model more robust to outliers?**

* We can have regularization such as L1 or L2 to reduce variance (increase bias).
* Changes to the algorithm:
  + Use tree-based methods instead of regression methods as they are more resistant to outliers. For statistical tests, use non parametric tests instead of parametric ones.
  + Use robust error metrics such as MAE or Huber Loss instead of MSE.
* Changes to the data:
  + Winsorizing the data
  + Transforming the data (e.g. log)
  + Remove them only if you’re certain they’re anomalies not worth predicting

**4. What are some differences you would expect in a model that minimizes squared error, versus a model that minimizes absolute error? In which cases would each error metric be appropriate?**

* MSE is more strict to having outliers. MAE is more robust in that sense, but is harder to fit the model for because it cannot be numerically optimized. So when there are less variability in the model and the model is computationally easy to fit, we should use MAE, and if that’s not the case, we should use MSE.
* MSE: easier to compute the gradient, MAE: linear programming needed to compute the gradient
* MAE more robust to outliers. If the consequences of large errors are great, use MSE
* MSE corresponds to maximizing likelihood of Gaussian random variables

**5. What error metric would you use to evaluate how good a binary classifier is? What if the classes are imbalanced? What if there are more than 2 groups?**

* Accuracy: proportion of instances you predict correctly. Pros: intuitive, easy to explain, Cons: works poorly when the class labels are imbalanced and the signal from the data is weak
* AUROC: plot fpr on the x axis and tpr on the y axis for different threshold. Given a random positive instance and a random negative instance, the AUC is the probability that you can identify who's who. Pros: Works well when testing the ability of distinguishing the two classes, Cons: can’t interpret predictions as probabilities (because AUC is determined by rankings), so can’t explain the uncertainty of the model
* logloss/deviance: Pros: error metric based on probabilities, Cons: very sensitive to false positives, negatives
* When there are more than 2 groups, we can have k binary classifications and add them up for logloss. Some metrics like AUC is only applicable in the binary case.

**6. What are various ways to predict a binary response variable? Can you compare two of them and tell me when one would be more appropriate? What’s the difference between these? (SVM, Logistic Regression, Naive Bayes, Decision Tree, etc.)**

* Things to look at: N, P, linearly seperable?, features independent?, likely to overfit?, speed, performance, memory usage
* Logistic Regression
  + features roughly linear, problem roughly linearly separable
  + robust to noise, use l1,l2 regularization for model selection, avoid overfitting
  + the output come as probabilities
  + efficient and the computation can be distributed
  + can be used as a baseline for other algorithms
  + (-) can hardly handle categorical features
* SVM
  + with a nonlinear kernel, can deal with problems that are not linearly separable
  + (-) slow to train, for most industry scale applications, not really efficient
* Naive Bayes
  + computationally efficient when P is large by alleviating the curse of dimensionality
  + works surprisingly well for some cases even if the condition doesn’t hold
  + with word frequencies as features, the independence assumption can be seen reasonable. So the algorithm can be used in text categorization
  + (-) conditional independence of every other feature should be met
* Tree Ensembles
  + good for large N and large P, can deal with categorical features very well
  + non parametric, so no need to worry about outliers
  + GBT’s work better but the parameters are harder to tune
  + RF works out of the box, but usually performs worse than GBT
* Deep Learning
  + works well for some classification tasks (e.g. image)
  + used to squeeze something out of the problem

**7. What is regularization and where might it be helpful? What is an example of using regularization in a model?**

* Regularization is useful for reducing variance in the model, meaning avoiding overfitting . For example, we can use L1 regularization in Lasso regression to penalize large coefficients.

**8. Why might it be preferable to include fewer predictors over many?**

* When we add irrelevant features, it increases model's tendency to overfit because those features introduce more noise. When two variables are correlated, they might be harder to interpret in case of regression, etc.
* curse of dimensionality
* adding random noise makes the model more complicated but useless
* computational cost
* Ask someone for more details.

**9. Given training data on tweets and their retweets, how would you predict the number of retweets of a given tweet after 7 days after only observing 2 days worth of data?**

* Build a time series model with the training data with a seven day cycle and then use that for a new data with only 2 days data.
* Ask someone for more details.
* Build a regression function to estimate the number of retweets as a function of time t
* to determine if one regression function can be built, see if there are clusters in terms of the trends in the number of retweets
* if not, we have to add features to the regression function
* features + # of retweets on the first and the second day -> predict the seventh day
* <https://en.wikipedia.org/wiki/Dynamic_time_warping>

**10. How could you collect and analyze data to use social media to predict the weather?**

* We can collect social media data using twitter, Facebook, instagram API’s. Then, for example, for twitter, we can construct features from each tweet, e.g. the tweeted date, number of favorites, retweets, and of course, the features created from the tweeted content itself. Then use a multi variate time series model to predict the weather.
* Ask someone for more details.

**11. How would you construct a feed to show relevant content for a site that involves user interactions with items?**

* We can do so using building a recommendation engine. The easiest we can do is to show contents that are popular other users, which is still a valid strategy if for example the contents are news articles. To be more accurate, we can build a content based filtering or collaborative filtering. If there’s enough user usage data, we can try collaborative filtering and recommend contents other similar users have consumed. If there isn’t, we can recommend similar items based on vectorization of items (content based filtering).

**12. How would you design the people you may know feature on LinkedIn or Facebook?**

* Find strong unconnected people in weighted connection graph
  + Define similarity as how strong the two people are connected
  + Given a certain feature, we can calculate the similarity based on
    - friend connections (neighbors)
    - Check-in’s people being at the same location all the time.
    - same college, workplace
    - Have randomly dropped graphs test the performance of the algorithm
* ref. News Feed Optimization
  + Affinity score: how close the content creator and the users are
  + Weight: weight for the edge type (comment, like, tag, etc.). Emphasis on features the company wants to promote
  + Time decay: the older the less important

**13. How would you predict who someone may want to send a Snapchat or Gmail to?**

* for each user, assign a score of how likely someone would send an email to
* the rest is feature engineering:
  + number of past emails, how many responses, the last time they exchanged an email, whether the last email ends with a question mark, features about the other users, etc.
* Ask someone for more details.
* People who someone sent emails the most in the past, conditioning on time decay.

**14. How would you suggest to a franchise where to open a new store?**

* build a master dataset with local demographic information available for each location.
  + local income levels, proximity to traffic, weather, population density, proximity to other businesses
  + a reference dataset on local, regional, and national macroeconomic conditions (e.g. unemployment, inflation, prime interest rate, etc.)
  + any data on the local franchise owner-operators, to the degree the manager
* identify a set of KPIs acceptable to the management that had requested the analysis concerning the most desirable factors surrounding a franchise
  + quarterly operating profit, ROI, EVA, pay-down rate, etc.
* run econometric models to understand the relative significance of each variable
* run machine learning algorithms to predict the performance of each location candidate

**15. In a search engine, given partial data on what the user has typed, how would you predict the user’s eventual search query?**

* Based on the past frequencies of words shown up given a sequence of words, we can construct conditional probabilities of the set of next sequences of words that can show up (n-gram). The sequences with highest conditional probabilities can show up as top candidates.
* To further improve this algorithm,
  + we can put more weight on past sequences which showed up more recently and near your location to account for trends
  + show your recent searches given partial data

**16. Given a database of all previous alumni donations to your university, how would you predict which recent alumni are most likely to donate?**

* Based on frequency and amount of donations, graduation year, major, etc, construct a supervised regression (or binary classification) algorithm.

**17. You’re Uber and you want to design a heatmap to recommend to drivers where to wait for a passenger. How would you approach this?**

* Based on the past pickup location of passengers around the same time of the day, day of the week (month, year), construct
* Ask someone for more details.
* Based on the number of past pickups
  + account for periodicity (seasonal, monthly, weekly, daily, hourly)
  + special events (concerts, festivals, etc.) from tweets

**18. How would you build a model to predict a March Madness bracket?**

* One vector each for team A and B. Take the difference of the two vectors and use that as an input to predict the probability that team A would win by training the model. Train the models using past tournament data and make a prediction for the new tournament by running the trained model for each round of the tournament
* Some extensions:
  + Experiment with different ways of consolidating the 2 team vectors into one (e.g concantenating, averaging, etc)
  + Consider using a RNN type model that looks at time series data.

**19. You want to run a regression to predict the probability of a flight delay, but there are flights with delays of up to 12 hours that are really messing up your model. How can you address this?**

* This is equivalent to making the model more robust to outliers.
* See Q3

**Probability (19 questions)**

**1. Bobo the amoeba has a 25%, 25%, and 50% chance of producing 0, 1, or 2 o spring, respectively. Each of Bobo’s descendants also have the same probabilities. What is the probability that Bobo’s lineage dies out?**

* p=1/4+1/4*p+1/2*p^2 => p=1/2

**2. In any 15-minute interval, there is a 20% probability that you will see at least one shooting star. What is the proba- bility that you see at least one shooting star in the period of an hour?**

* 1-(0.8)^4. Or, we can use Poisson processes

**3. How can you generate a random number between 1 - 7 with only a die?**

* Launch it 3 times: each throw sets the nth bit of the result.
* For each launch, if the value is 1-3, record a 0, else 1. The result is between 0 (000) and 7 (111), evenly spread (3 independent throw). Repeat the throws if 0 was obtained: the process stops on evenly spread values.

**4. How can you get a fair coin toss if someone hands you a coin that is weighted to come up heads more often than tails?**

* Flip twice and if HT then H, TH then T.

**5. You have an 50-50 mixture of two normal distributions with the same standard deviation. How far apart do the means need to be in order for this distribution to be bimodal?**

* more than two standard deviations

**6. Given draws from a normal distribution with known parameters, how can you simulate draws from a uniform distribution?**

* plug in the value to the CDF of the same random variable

**7. A certain couple tells you that they have two children, at least one of which is a girl. What is the probability that they have two girls?**

* 1/3

**8. You have a group of couples that decide to have children until they have their first girl, after which they stop having children. What is the expected gender ratio of the children that are born? What is the expected number of children each couple will have?**

* gender ratio is 1:1. Expected number of children is 2. let X be the number of children until getting a female (happens with prob 1/2). this follows a geometric distribution with probability 1/2

**9. How many ways can you split 12 people into 3 teams of 4?**

* the outcome follows a multinomial distribution with n=12 and k=3. but the classes are indistinguishable

**10. Your hash function assigns each object to a number between 1:10, each with equal probability. With 10 objects, what is the probability of a hash collision? What is the expected number of hash collisions? What is the expected number of hashes that are unused.**

* the probability of a hash collision: 1-(10!/10^10)
* the expected number of hash collisions: 1-10\*(9/10)^10
* the expected number of hashes that are unused: 10\*(9/10)^10

**11. You call 2 UberX’s and 3 Lyfts. If the time that each takes to reach you is IID, what is the probability that all the Lyfts arrive first? What is the probability that all the UberX’s arrive first?**

* Lyfts arrive first: 2!\*3!/5!
* Ubers arrive first: same

**12. I write a program should print out all the numbers from 1 to 300, but prints out Fizz instead if the number is divisible by 3, Buzz instead if the number is divisible by 5, and FizzBuzz if the number is divisible by 3 and 5. What is the total number of numbers that is either Fizzed, Buzzed, or FizzBuzzed?**

* 100+60-20=140

**13. On a dating site, users can select 5 out of 24 adjectives to describe themselves. A match is declared between two users if they match on at least 4 adjectives. If Alice and Bob randomly pick adjectives, what is the probability that they form a match?**

* 24C5\*(1+5(24-5))/24C5\*24C5 = 4/1771

**14. A lazy high school senior types up application and envelopes to n different colleges, but puts the applications randomly into the envelopes. What is the expected number of applications that went to the right college?**

* 1

**15. Let’s say you have a very tall father. On average, what would you expect the height of his son to be? Taller, equal, or shorter? What if you had a very short father?**

* Shorter. Regression to the mean

**16. What’s the expected number of coin flips until you get two heads in a row? What’s the expected number of coin flips until you get two tails in a row?**

**17. Let’s say we play a game where I keep flipping a coin until I get heads. If the first time I get heads is on the nth coin, then I pay you 2n-1 dollars. How much would you pay me to play this game?**

* less than $3

**18. You have two coins, one of which is fair and comes up heads with a probability 1/2, and the other which is biased and comes up heads with probability 3/4. You randomly pick coin and flip it twice, and get heads both times. What is the probability that you picked the fair coin?**

* 4/13

**19. You have a 0.1% chance of picking up a coin with both heads, and a 99.9% chance that you pick up a fair coin. You flip your coin and it comes up heads 10 times. What’s the chance that you picked up the fair coin, given the information that you observed?**

* Events: F = "picked a fair coin", T = "10 heads in a row"
* (1) P(F|T) = P(T|F)P(F)/P(T) (Bayes formula)
* (2) P(T) = P(T|F)P(F) + P(T|¬F)P(¬F) (total probabilities formula)
* Injecting (2) in (1): P(F|T) = P(T|F)P(F)/(P(T|F)P(F) + P(T|¬F)P(¬F)) = 1 / (1 + P(T|¬F)P(¬F)/(P(T|F)P(F)))
* Numerically: 1/(1 + 0.001 \* 2^10 /0.999).
* With 2^10 ≈ 1000 and 0.999 ≈ 1 this simplifies to 1/2

**20. What is a P-Value ?**

* The probability to obtain a similar or more extreme result than observed when the null hypothesis is assumed.
* ⇒ If the p-value is small, the null hypothesis is unlikely

**Programming (14 questions)**

**1. Write a function to calculate all possible assignment vectors of 2n users, where n users are assigned to group 0 (control), and n users are assigned to group 1 (treatment).**

* Recursive programming (sol in code)

**2. Given a list of tweets, determine the top 10 most used hashtags.**

* Store all the hashtags in a dictionary and get the top 10 values

**3. Program an algorithm to find the best approximate solution to the knapsack problem1 in a given time.**

* Greedy solution (add the best v/w as much as possible and move on to the next)

**4. Program an algorithm to find the best approximate solution to the travelling salesman problem2 in a given time.**

* Greedy

**5. You have a stream of data coming in of size n, but you don’t know what n is ahead of time. Write an algorithm that will take a random sample of k elements. Can you write one that takes O(k) space?**

* <https://en.wikipedia.org/wiki/Reservoir_sampling>

**6. Write an algorithm that can calculate the square root of a number.**

* <https://www.quora.com/What-is-the-method-to-calculate-a-square-root-by-hand?redirected_qid=664405>
* <https://en.wikipedia.org/wiki/Newton's_method#Square_root_of_a_number>

**7. Given a list of numbers, can you return the outliers?**

* sort then select the highest and the lowest 2.5%

**8. When can parallelism make your algorithms run faster?**

When could it make your algorithms run slower?

* Ask someone for more details.
* compute in parallel when communication cost < computation cost
  + ensemble trees
  + minibatch
  + cross validation
  + forward propagation
  + minibatch
  + not suitable for online learning

**9. What are the different types of joins? What are the differences between them?**

* (INNER) JOIN: Returns records that have matching values in both tables LEFT (OUTER) JOIN: Return all records from the left table, and the matched records from the right table RIGHT (OUTER) JOIN: Return all records from the right table, and the matched records from the left table FULL (OUTER) JOIN: Return all records when there is a match in either left or right table

**10. Why might a join on a subquery be slow? How might you speed it up?**

* Change the subquery to a join.

**11. Describe the difference between primary keys and foreign keys in a SQL database.**

* Primary keys are columns whose value combinations must be unique in a specific table so that each row can be referenced uniquely. Foreign keys are columns that references columns (often primary keys) in other tables.

**12. Given a COURSES table with columns course\_id and course\_name, a FACULTY table with columns faculty\_id and faculty\_name, and a COURSE\_FACULTY table with columns faculty\_id and course\_id, how would you return a list of faculty who teach a course given the name of a course?**

* select faculty\_name from faculty\_id c join (select faculty\_id from (select course\_id from COURSES where course\_name=xxx) as a join COURSE\_FACULTY b on a.course\_id = b.course\_id) d on c.faculty\_id = d.faculty\_id

**13. Given a IMPRESSIONS table with ad\_id, click (an indicator that the ad was clicked), and date, write a SQL query that will tell me the click-through-rate of each ad by month.**

* select ad\_id, average(click) from IMPRESSIONS group by ad\_id, month(date)

**14. Write a query that returns the name of each department and a count of the number of employees in each:**

EMPLOYEES containing: Emp\_ID (Primary key) and Emp\_Name  
EMPLOYEE\_DEPT containing: Emp\_ID (Foreign key) and Dept\_ID (Foreign key)  
DEPTS containing: Dept\_ID (Primary key) and Dept\_Name

* select Dept\_Name, count(1) from DEPTS a right join EMPLOYEE\_DEPT b on a.Dept\_id = b.Dept\_id group by Dept\_Name

**Statistical Inference (15 questions)**

**1. In an A/B test, how can you check if assignment to the various buckets was truly random?**

* Plot the distributions of multiple features for both A and B and make sure that they have the same shape. More rigorously, we can conduct a permutation test to see if the distributions are the same.
* MANOVA to compare different means

**2. What might be the benefits of running an A/A test, where you have two buckets who are exposed to the exact same product?**

* Verify the sampling algorithm is random.

**3. What would be the hazards of letting users sneak a peek at the other bucket in an A/B test?**

* The user might not act the same suppose had they not seen the other bucket. You are essentially adding additional variables of whether the user peeked the other bucket, which are not random across groups.

**4. What would be some issues if blogs decide to cover one of your experimental groups?**

* Same as the previous question. The above problem can happen in larger scale.

**5. How would you conduct an A/B test on an opt-in feature?**

* Ask someone for more details.

**6. How would you run an A/B test for many variants, say 20 or more?**

* one control, 20 treatment, if the sample size for each group is big enough.
* Ways to attempt to correct for this include changing your confidence level (e.g. Bonferroni Correction) or doing family-wide tests before you dive in to the individual metrics (e.g. Fisher's Protected LSD).

**7. How would you run an A/B test if the observations are extremely right-skewed?**

* lower the variability by modifying the KPI
* cap values
* percentile metrics
* log transform
* <https://www.quora.com/How-would-you-run-an-A-B-test-if-the-observations-are-extremely-right-skewed>

**8. I have two different experiments that both 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?**

* exclusive -> ok

**9. What is a p-value? What is the di erence between type-1 and type-2 error?**

* type-1 error: rejecting Ho when Ho is true
* type-2 error: not rejecting Ho when Ha is true

**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?**

* For randomly selected listings with more than 1 pictures, hide 1 random picture for group A, and show all for group B. Compare the booking rate for the two groups.
* Ask someone for more details.

**11. How would you design an experiment to determine the impact of latency on user engagement?**

* The best way I know to quantify the impact of performance is to isolate just that factor using a slowdown experiment, i.e., add a delay in an A/B test.

**12. What is maximum likelihood estimation? Could there be any case where it doesn’t exist?**

* A method for parameter optimization (fitting a model). We choose parameters so as to maximize the likelihood function (how likely the outcome would happen given the current data and our model).
* maximum likelihood estimation (MLE) is a method of [estimating](https://en.wikipedia.org/wiki/Estimator) the [parameters](https://en.wikipedia.org/wiki/Statistical_parameter) of a [statistical model](https://en.wikipedia.org/wiki/Statistical_model) given observations, by finding the parameter values that maximize the [likelihood](https://en.wikipedia.org/wiki/Likelihood) of making the observations given the parameters. MLE can be seen as a special case of the [maximum a posteriori estimation](https://en.wikipedia.org/wiki/Maximum_a_posteriori_estimation) (MAP) that assumes a [uniform](https://en.wikipedia.org/wiki/Uniform_distribution_(continuous)) [prior distribution](https://en.wikipedia.org/wiki/Prior_probability) of the parameters, or as a variant of the MAP that ignores the prior and which therefore is [unregularized](https://en.wikipedia.org/wiki/Regularization_(mathematics)).
* for gaussian mixtures, non parametric models, it doesn’t exist

**13. What’s the di erence between a MAP, MOM, MLE estima- tor? In which cases would you want to use each?**

* MAP estimates the posterior distribution given the prior distribution and data which maximizes the likelihood function. MLE is a special case of MAP where the prior is uninformative uniform distribution.
* MOM sets moment values and solves for the parameters. MOM is not used much anymore because maximum likelihood estimators have higher probability of being close to the quantities to be estimated and are more often unbiased.

**14. What is a confidence interval and how do you interpret it?**

* For example, 95% confidence interval is an interval that when constructed for a set of samples each sampled in the same way, the constructed intervals include the true mean 95% of the time.
* if confidence intervals are constructed using a given confidence level in an infinite number of independent experiments, the proportion of those intervals that contain the true value of the parameter will match the confidence level.

**15. What is unbiasedness as a property of an estimator? Is this always a desirable property when performing inference? What about in data analysis or predictive modeling?**

* Unbiasedness means that the expectation of the estimator is equal to the population value we are estimating. This is desirable in inference because the goal is to explain the dataset as accurately as possible. However, this is not always desirable for data analysis or predictive modeling as there is the bias variance tradeoff. We sometimes want to prioritize the generalizability and avoid overfitting by reducing variance and thus increasing bias.