# Questions

**1. How would you test if survey responses were filled at random by certain individuals, as opposed to truthful selections?**

If at random I’d expect there to be an equal spread between answers. Could use a t-test to determine if the % of a certain answer is significantly different from an equal split.

If it’s possible to design the survey so that the same question (or similar questions) is repeated twice or some reverse questions, that could be a way to test this (if the same question is answered differently by the same person, they’re doing it at random).

**2. Suppose you send out a survey to gather more information about the population, but the response rate is very low (less than 1% responded). How might you analyze the data you collected?**

Check that you have enough data to complete the analysis. Check to make sure the group of respondents is representative of the larger population (to ensure that analysis can be generalized).

Could use bootstrapping or another resampling technique to get a larger sample size.

**3. Explain selection bias.**

Selection bias occurs when the people volunteering for a study share a trait that would make them different than the population as a whole.

**4. Suppose a survey about voting preference for a candidate is done over multiple counties/districts in a particular state. Should you worry about clustering? How would it affect the analysis?**

Yes it’s possible that clustering could affect the analysis here if the sample isn’t from all counties.(the result from samples may have bias with population). People living in the same county may be more likely to vote the same or have the same preference for a candidate.

**5. How would you build and test a metric to compare two users’ rankings of products on Amazon’s platform?**

Collect all user rankings and calculate average standard deviation, split up by product type. This gives us good values to compare the ranking styles of the two users. Take purchase behavior into account as well.

**6. How would you build and test a metric to compare between two products in the same category on Amazon’s platform?**

Calculate average ranking and standard deviation for each. We could also look at how often each product is bought. Construct a metric out of a combination of these observations, such as sales, rates, etc.

**7. How would you design a test to filter out sponsored reviews from Amazon?**

Since reviewers who receive a product for free to review are required to state that in their review, I would filter out reviews that include words saying that. It could also be helpful to examine all other reviews left by a particular account. Feed a lot of sponsored reviews into an ML model and run each of the reviews past that to see how similar they are?

Sponsored reviews may mean that sellers pay some people for their reviews so that they can get what they want.

Since the data we need to collect is text data, we use Naive Bayesian Model. 1. Collect data and label 2. Randomly select train data 2. Calculate the probability of sponsored reviews and non-sponsored reviews, and Conditional Probability of every word 3. Use test data to test the model. 4. Predict the classification of new reviews

**8. Tell me how you would train a neural net to predict user ratings of a product on Amazon.**

I would start with random weights, and then train with relevant data from each user such as sales, past ratings, other products purchased, while minimizing loss(gradient descent).

**9. When running an A/B test for a new product that runs over the course of the day and the data is collected over time, what aspect of the statistical analysis should you worry about, especially with respect to p-hacking?**

If you run enough t-tests over time, you’re going to get a result that is statistically significant (p-hacking). P-hacking can result in dishonest results. Data may be the time series data, so it may be serially correlated, which is contrary to the independence assumption.

One option would be to analyze this as time series data (if the time has an effect on the data).

**10. How would you know if one algorithm is better than the other?**

I would use a set of testing data (where I know the results of what the algorithm is trying to predict) and have each algorithm predict a result for everything in that dataset. Then I would compare the error rate across algorithms.

Could also be repeated again with training data to see if results differ and get an idea for how robust the algorithm is.

**11. You're about to get on a plane to Seattle. You want to know if you should bring an umbrella. You call 3 random friends of yours who live there and ask each independently if it's raining. Each of your friends has a 2/3 chance of telling you the truth and a ⅓ chance of messing with you by lying. All 3 friends tell you that "Yes" it is raining. What is the probability that it's actually raining in Seattle? (Apparently, Facebook likes these types of probability and Bayesian questions...)**

Prob that all three friends are lying: ⅓\*⅓\*⅓ = 1/27

Prob that all three friends are telling the truth: ⅔\*⅔\*⅔ = 8/27

So prob that it’s actually raining is (8/27)/(9/27) = 8/9

Another answer is about Bayesian : X has Binomial Distribution, Y is the probability of raining

P(X=1) = (3 1)(⅔)(⅓)^2 = 2/9, P(Y|X) = 1/3

P(X=2) = (3 2)(⅔)^2(⅓) = 4/9, P(Y|X) = 2/3

P(X=3) = (3 3)(⅔)^3 = 8/27, P(Y|X) = 1

P(Y) = ⅓\*2/9+⅔\*4/9 + 8/27 = 2/3

**12. Tell me about various MCMC algorithms, especially their strengths and weaknesses.**

Metropolis-Hastings: generalizable but needs to be tuned manually, only in one-dimension, Efficiency may be low, can get sample from any distribution.

Gibbs: doesn’t require tuning but doesn’t work well with highly correlated parameters, higher dimension, only work in the direction parallel to axis

**13. Tell me about Hamiltonian Monte Carlo.**

**Hamiltonian Monte Carlo corresponds to an instance of the** [**Metropolis–Hastings algorithm**](https://en.wikipedia.org/wiki/Metropolis%E2%80%93Hastings_algorithm)**, with a** [**Hamiltonian dynamics**](https://en.wikipedia.org/wiki/Hamiltonian_mechanics) **evolution simulated using a** [**time-reversible**](https://en.wikipedia.org/wiki/Time_reversibility) **and volume-preserving numerical integrator (typically the** [**leapfrog integrator**](https://en.wikipedia.org/wiki/Leapfrog_integration)**) to propose a move to a new point in the state space**

**(From Wikipedia)**

**The algorithm requires a positive integer for number of leap frog steps L and a positive number for the step size . Suppose the chain is at . Get the initial position and momentum from a normal distribution , where M is a matrix related to the Hamilton's equations. Then using the Leap frog algorithm, the particle will run under Hamiltonian dynamics for time and get and . Then transit to using the Metropolis Hastings step.**

Limited applications of HMC because we need to be able to compute gradients, but still generally more efficient than Metropolis-Hastings (even though each step takes more computation) because distance between points is larger and more likely to go to a new state.

**14. Tell me about Metropolis Hastings and Gibbs sampling. How are they similar? How do they differ from each other?**

Metropolis-Hastings: generalizable but needs to be tuned manually. Uses a full joint density distribution. Inefficient in higher dimensions

Gibbs: doesn’t require tuning but doesn’t work well with highly correlated parameters. Relies on conditional distributions. Gibbs is a special case of M-H for when we don’t know the join distribution (or it is difficult to sample from) but we know the conditional distribution of each variable. Has a burn in period (first few samples taken shouldn’t be counted)

Metropolis Hastings: Choose a proposal distribution. Give the initial value and then get the samples from the following iteration. Suppose we have got the t samples. Take advantage of the sample generated from the proposal distribution to calculate an acceptance probability and get the t+1 sample based on the acceptance probability.

For the distribution :

(1) Choose a proposal/instrumental distribution

(2) Give an initial value

(3) Given :

(3.1) generate

(3.1) take

where:

Gibbs: Give the initial value the vector (joint distribution) and then sample each component of the vector iteratively. Get each component through the marginal distribution and use the new value to update the the marginal distribution.

For the distribution :

(1) Give an initial value of

(2) Given , generate

...

p.

Gibbs sampling was introduced to break the curse of dimensionality (which impacts both rejection and importance sampling) by producing a sequence of low dimension simulations that still converge to the right target. Gibbs sampling is a special case of Metropolis-Hasting algorithm with a probability of acceptance of one.

Usually, Gibbs sampling [understood as running a sequence of low-dimensional conditional simulations] is favoured in settings where the decomposition into such conditionals is easy to implement and fast to run. In settings where such decompositions induce multimodality and hence a difficulty to move between modes (latent variable models like mixture models come to mind), using a more global proposal in a Metropolis-Hasting algorithm may produce a higher efficiency. But the drawback stands with choosing the proposal distribution in the Metropolis-Hasting algorithm.

**15. How would you obtain user-related statistics (e.g. time spent on Facebook, number of likes total in a month) over the Facebook graph?**

Ask what kind of data is collected in the facebook graph? Though it was just friends/mutual friends?

Can consider about using network to build a user-related graph.

1. Collect users information, such as their ID, time spent on Facebook, their nationality, sex, age, etc
2. Build the network, the nodes represent different users and the lines represent their information or features.
3. Use lines to connect the nodes. If two users have every one same feature, they will be connected by one line. Thus, we can see that if two nodes are connected by more lines, they may be stronger related.

**16. Facebook is planning to launch a new set of like buttons (now known as love, smileys, etc.). How would you go about testing this new set of like buttons?**

A/B testing to check use (one group of users has the buttons and one group doesn’t have them). One way ANOVA could be used to test this as well

**17. Facebook wants to run ad campaigns on behalf of its advertising partners. Suppose there are two ads that they are interested in running. How would you design an experiment to test which ad performs better?**

Conduct A/B test with both ads with a small group of users and measure/test click through rate

**18. How would you build an anomaly detection algorithm for potentially false Uber requests from passengers?**

There are several anomaly situations:

1. The boarding point location in Uber's map is different from the actual location which the passengers would like to choose.

Solution: Capture the passengers’ current location minute by minute, compare whether the locations’ coordinate in an interval time is equal to coordinate of the boarding point location.

1. Passengers select the wrong car model.

Solution: Select users information about their trip information, such as destination, starting place, and car type corresponding to travel. Then predict the car model they may choose on this trip.

1. Wrong Click. (users do not want to use Uber, just click unconsciously)

Solution: Collect information about the current location, if they just move in a small range of the boarding point, they may not click unconsciously. Otherwise, they may click wrong.

Try to detect outliers

**19. How would you build a prediction algorithm to let Uber drivers know that certain areas of the city will have high demand?**

Ask about which kind of data Uber collects (how frequently). Divide data up by neighborhood (if possible?). Build a time series model based on past data by neighborhood. Identify day of the week and time of day trends. Factor in holidays and events (there might be a spike in rides requested after a concert or sporting event).

Regression tree. Divide the city into diverse areas. Analyze the features of each area. Build up a regression tree model based on the features for those areas. Given the certain area, use the model to predict whether it will have high demand. Use the data of the past to train the model. Use the latest data to update the model. (I think the time series answer is better than this).

**20. Devise a way to generate any n random samples of a particular continuous distribution F from a given n random samples from the uniform[0,1].**

**generate samples from the U[0,1]**

**F^{-1}(U)**

One way: Acceptance-Rejection Method: First find a known pdf g(x), s.t. f(x)/g(x)<=c, where c is the sup{f(x)/g(x)}(we want c to be close to 1).

1. generate random value Y distributed as g(x)
2. generate U from uniform [0,1]
3. If U<=f(y)/cg(y), we accept X=Y, otherwise back to 1

Two way: Monte Carlo method

[*inverse transform sampling*](https://en.wikipedia.org/wiki/Inverse_transform_sampling). Let *U* be a uniformly distributed random number between zero and one. Then to sample a random number with a (possibly nonuniform) probability distribution function *f(x)*, do the following:

1. Normalize the function *f(x)* if it isn’t already normalized.
2. Integrate the normalized PDF *f(x)* to compute the CDF, *F(x)*.
3. Invert the function *F(x)*. The resulting function is the *inverse cumulative distribution function* or *quantile function* *F-1(x)*. Because we’ve already normalized *f(x)*, we could also clarify by calling this the *inverse normal cumulative distribution function*, or simply the inverse normal CDF.
4. Substitute the value of the uniformly distributed random number *U* into the inverse normal CDF.

From https://www.comsol.com/blogs/sampling-random-numbers-from-probability-distribution-functions/

**21. Given n samples from a uniform distribution of whole numbers {1,...,M}, find an estimator for M.**

Estimate for M is max of the samples

**22. Given n samples from a uniform distribution between [a,b], a<b, find an estimate for a and b.**

MLE: Estimate for a = min of the samples

Estimate for b = max of the samples

**23. How do you diagnose your model for a continuous outcome using multiple linear regression? What are the assumptions for a linear regression model?**

Check residual plots for assumptions like normality, independence, heteroskedasticity, linearity

Assumptions: Linearity,normality, independence,link function

**24. Suppose you are fitting a multiple linear regression model and you have some missing values in your predictors. How would you deal with them?**

1. If there’s only a couple missing in a large data set, drop them.
2. If they can be imputed, Median imputation methods (see 26).
3. If missing a lot, random forest? Look for patterns in the missingness, or data that’s missing because the engineers messed something up
4. Interpolation method :Lagrange polynomial

**25. Suppose you are fitting a multiple linear regression model and you have some missing values in your outcomes. How would you deal with them?**

First I would look to see if I can discern a pattern with which values are missing.

1. If they’re just missing at random, I could attempt of predict them based on the predictors.
2. Or simply remove them from the dataset

**26. Tell me about imputation methods.**

Median imputation: replace each missing value with the median in that area

Multiple imputation: use multivariate methods to “guess” missing value (when columns are related)

**27. How do you interpret the coefficient in a multiple linear regression?**

With other explanatory variables fixed, for every one unit increase in the explanatory variable, this much (coefficient variable) is how much the response variable/outcome increases.

**28. Explain how one fits a logistic regression.**

1. Use the likelihood function to estimate the probability of observing the data, given the unknown parameters, and then get log-likelihood function.
2. Solve the unknown parameters from log-likelihood function



1. See the Null deviance and residual deviance. If residual deviance/its freedom is equal to 1, the model is fitted with assumption. Use (Null deviance -residual deviance)/the number of explanatory variables(the difference of their freedom), if it is larger than 1, the explanatory variable are useful.
2. Use accuracy rate to see the fit of model.

**29. How do you interpret the coefficient from a logistic regression?**

Log odds of a particular outcome based on the input factor, the answer like in the LM

**30. How does one-way ANOVA with two factors compare to two-sample t-tests?**

1. One way ANOVA assumes a common variance (pooled); with two sample t tests we can have different variances.
2. ANOVA can compare more than two levels of treatment while t test can only compare one

**31. What are the differences/similarities between fixed and random effects in ANOVA-based models?**

Similarities:

1. They can also test the difference among different levels.
2. Both types of effects can have an impact on the results(y).

Difference:

1. Fixed effects are chosen by the investigators and deliberately set ahead of time. Random effects are not chosen and are random variables with a common mean meant to represent the population.
2. For random effects, different levels are independent, but observations with the same i (in the same group) are dependent.

**32. Suppose your outcome is a rating from 1 to 10 and you want to predict this rating based on a set of predictors. Would you use logistic regression or multiple linear regression or neither? Explain.**

multinominal logistic regression ( no order) or Ordinal logistic regression(ordered). Something handy about multinomial is that assumptions of normality, homoscedasticity, linearity do not need to be met (in contrast to an alternative: LDA or discriminant function analysis)

**33. What is shrinkage and penalization in regression? Why is it useful?**

Shrinkage and penalization is shrinking the value of a coefficient so much that it’s close to nonexistent in the model. Penalization can help with collinearity and overfitting.

Since we get the restrictions to the coefficients, the value of coefficients will not keep increasing.

**34. How would you fit a model to predict X-quarter sales based on past sales data?**

Use time series tools to build a model based on past data (maybe ARIMA model or seasonal ARMA model?). Take sales and holidays into account if relevant.

**35. Suppose you are testing the efficacy of vaccine A versus vaccine B and you measure whether the person was infected after vaccination. Propose a way to test the difference between the two vaccines.**

1. Use a two-sample t test. Pooled variance assumption because it’s difficult to calculate standard deviation for a binary variable like infected/not infected.
2. Use Four grid table and Chi-square test

**36. Suppose are testing the relationship between a categorical variable, say Male and Female, with a binary outcome, say Yes and No on a survey. What are some statistical tests to the relationship between the two variables? What are their strengths and weaknesses?**

Log-linear analysis (extension chi-square)

Chi-square test (chi square:logistic regression :: ANOVA:regression)

Chi-square is just a descriptive test (like a correlation), not a modeling technique

Four grid table

Logistic regression - more general because not restricted to a dichotomous variable, also not limited to a single predictor? (tho don’t know if that’s relevant here)

**37. What is statistical power? Why is it important?**

The probability of a test finding an effect if there is an effect to be found. Power can be used to help determine the sample size needed when designing an experiment.

Probability of rejecting a null hypothesis when it is false (1 - beta, beta is error type II).

Important: power tell how likely you are to detect a real effect. On the other hand, it means how much probability we will miss the real effect.

**38. What is a p-value? What are some properties of the p-values?**

P-value is the probability of getting results as extreme (or more extreme) than the observations. This is under the assumption that the null hypothesis holds. If a p-value is very small, this indicates that the results obtained are very unlikely and the null hypothesis is potentially not true.

**39. What is multiple testing? How can it be used in A/B testing?**

Multiple testing is the practice of running multiple tests simultaneously. Can be used in A/B testing but be careful about interference between the tests.

**40. What are resampling methods and how can they be used? What are their strengths and limitations?**

Bootstrap: Samples drawn from dataset with replacement

* Doesn’t require independent samples；It is useful for small sample, but not for large sample. And when get the estimation of SE difficulty, use it. Increasing the number of resampling will not increase the information of data, just get more stable result. Depend on the computer power.

Monte Carlo: I think it is just a kind of method of random sampling, but not resampling.

K-fold cross validation: partition data into k sets, keep one as test data and use the rest as validation or training.

Strengths:

1. It matters less how the data gets divided
2. Every data point gets to be in a test set exactly once, and gets to be in a training set k-1 times.
3. The variance of the resulting estimate is reduced as k is increased.

Limitations:

1. The training algorithm has to be rerun from scratch k times, which means it takes k times as much computation to make an evaluation.

Permutation tests: To test the difference between two small sample

Strengths:

1. Permutation tests exist for any test statistic, regardless of whether or not its distribution is known.
2. Used for analyzing unbalanced designs

Limitations:

1. The observations are exchangeable

**41. What are random forests? How would one implement a random forest? When should you use a random forest compared to other methods?**

Classification method consisting of lots of decision trees.

Implementation:

1. Preset hyperparameters，such how many trees or how many stratification
2. Sampling randomly for every tree.
3. Enter the sample to be tested in trees, and integrate all results: mean for regression or mode for classification.

Strengths: 1. Deal with high-dimensional data faster; 2. More robust, not easily overfit, and not sensitive to outliers. 3. Every trees are easy to be explained.

Weaknesses: Model may be too general to deal with some samples, for example in a classification question, some samples are hard to be classified.

Bagging : Sampling with replacement. RF is improved model based on bagging.

Compared with bagging, RF is more efficient, since RF just considered about a part of features.

**42. What are some clustering algorithms? What are their strengths and weaknesses?**

K-means clustering:

Strengths: fast, easy to calculate

Weaknesses: We must know in advance how many classes / groups of data there are. It may lump things together that it doesn’t make sense to lump or split larger groups down when they don’t need to be (based on the number it is fed)

kNN（k-nearest-neighbor):

Strengths: 1. Easy to be understood; 2. Can be used for nonlinear classification

Weaknesses: 1. The amount of calculation is large, especially when the number of features is very large; 2. In some situations, hard to explain the meaning of different clusterings

Hierarchical Clustering:

Strengths: 1. Can be used for any types of data; 2. Not sensitive to the order of data points.

Weaknesses: 1. Irreversibility

**43. Tell me about the QR decomposition and the SVD. How would you implement them? What are they useful for? How are they similar/different?**

QR is a decomposition of matrix A into A=QR, where Q is orthogonal matrix and R is a upper triangular matrix.

SVD is a decomposition of matrix M into M=U\SigmaV’.

They both transform high-dimension, sparse or nonsingular matrix into some matrices which have better statistical nature.

Different : 1. QR divided into two matrix and SVD divided into three.

2. QR is used for LM (requires matrix to be square, but SVD doesn’t). SVD is used for image processing or classification, data sparse (matrix).

**44. Tell me about the differences between a PCA and ICA.**

PCA is dimensionality reduction where data is projected on to a lower dimension subspace and the “principal components” (orthogonal vectors) are the axes of this subspace. When coming up with these components we want to maximize variance so that we capture as much information about the data as possible

ICA is independent component analysis, a version of PCA.

PCA focuses on maximizing variance while ICA focuses on independence

PCA has normal distribution while ICA doesn’t require this assumption. ICA assumptions: components we are trying to uncover are independent and non-Gaussian

PCA compresses information while ICA separates it. This is why it’s good to do PCA before performing ICA

PCA looks at second order info, ICA can go into higher orders; so PCA removes correlations, but not higher order dependence, while ICA removes both

PCA vectors are orthogonal while ICA vectors are not

PCA some components are more important than others (ordered) and ICA all components are equally important.

Similarity: both require data to be autoscaled (subtract each column by its mean and divide by its standard deviation)

**45. What is reinforcement learning? Tell me about an algorithm that is typical in reinforcement learning. How is it different from supervised, semi-supervised, or unsupervised learning?**

Reinforcement learning allows a machine to train itself with rewards and punishments. This allows the focus to shift (from traditional ML) from pattern recognition to decision making.

Recurrent Neural Network (characterized by its “memory”; used for NLP) and Markov Decision Process are examples

Supervised learning: Need feature and label

* Uses labeled data, typically classification (linear classifiers, random forests, decision trees, SVMs) or regression (linear regression, logistic regression, polynomial regression)

Unsupervised learning: Only feature

* Analyzes and clusters unlabeled datasets (clustering, association, dimensionality reduction)

Semi-Supervised learning: Only part of data has label, most does not have.

* Some labeled data, some unlabeled. Example where this is useful: medical imaging

Reinforcement learning: No label, train itself with rewards and punishments, learn in an interactive environment, it is a dynamic learning, but others are static.

Real world examples of reinforcement learning: https://neptune.ai/blog/reinforcement-learning-applications

**46. Explain, concisely, the law of large numbers and the central limit theorem.**

LLN: As the sample size moves towards infinity, the sample mean will equal the population mean

CLT: As the sample size moves towards infinity, the distribution will be normal.

# Tips

* Goal for interviewer is to see your thought process, especially in solving new problems
* Organized response
* Use specific examples/details/numbers
* Be complete and play to your strengths
* Ask questions and engage with interviewer