QUESTIONS ASKED IN PLACEMENT INTERVIEWS

M.Stat Final Year Students, 2019-21 Batch

September 25, 2022

1 Optiver

1. Describe any project you want from your CV.

2. Estimate the number of schoolteachers in the US. Give 75%, 95%, 99% CI of your estimate.

3. There are 37 horses, you can have a race of 6 horses at one time, how many races do you need to determine the fastest 3 horses?

4. You toss a coin, I toss a coin, both are fair.

   1. HH: you pay me 6
   2. TT: you pay me 4
   3. HT or TH: i pay you 5

(a) What is your expected payoff from this game if you play n turns?

(b) Now you bias your coin by putting head probability = p. You tell me the value of p. I select a head probability for my coin q. Now we play the same game as in (a) but with unfair coins as above. Do you have a strategy that is optimal for you?

1. They asked about projects and interests.

2. You are given 12 coins and a weighing balance. One of the coins is lighter than the others. How will you find the lighter coin in minimum number of weighings.

3. Same question as second but this time we don't know whether the special coin is lighter or heavier. Find the coin in minimum weighings.

4. Estimate the number of daily bus passengers in USA. You just know that the population

of USA is 350 million.

2 JP Morgan Chase & Co.

1. You move 1km south, 1km east, 1km north and arrive at where you started from. Where on earth could you have started from?

2. You are in a queue. The manager comes out and says that the first person to share a birthday with someone in front of him gets a free ticket. Where would the optimal position in the queue?

1. Explain one of your projects.

2. A father wants to assign all his properties to one of his two sons. Both the son have their personal horses. He told them that a race should be conducted and the person whose horse will lose the race will get all the money but he thinks that he doesn't have that much time left in his life to conduct such a race. How should he decide immediately that whom should be given his properties?

3. You have 100 balls (50 black and 50 white) and 2 jars. You should distribute all the balls in those 2 jars and then you randomly pick one jar and pick one ball out of it. How should you distribute the balls such that the probability of a black ball being picked is maximized?

4. You have a dice. You can roll it at max 3 times. After every roll, you will be given choices of roll again or have the amount of money same as the outcome. The money gain is not additive. What is your expected gain from this game?

5. Do you know what R-Squared and Adjusted R-Squared are? Can R-Squared be negative?

6. Tell me some properties of normal distribution. (They wanted to know two things: (1) symmetric, (2) Mean=Median=Mode .)

7. Why do you want to join JPMC, more generally an investment bank, instead of any insurance company or any other financial companies?

First Round.

1. Out of all the projects and internships, explain the one you found most interesting. Follow up questions like bootstrap confidence interval, testing significance of the statistics used in the project.

2. Gauss-Markov theorem - assumptions, consistent statistics and efficiency definition.

3. Distribution of ratio of independent normals, squares of independent normals, sum of squares of independent normals,...

4. Birthday problem in probability and it's variations.

5. A chess puzzle based on knight's tour.

6. Time Series Analysis questions - Definitions of Stationarity, Ljung box test, Adf test.

7. 10 dice are rolled simultaneously. What is the probability of getting sum of 25?

Second Round.

1. Describe the work done in your internship.

2. Tell us about the assumptions of the Linear Regression model.

3. Explain the concepts of clustering, p-value in Layman's terms.

4. Questions on Martingale, Brownian motions.

5. What are the problems that occur when we forcibly fit linear model on non-linear data?

6. When is AIC preferred over BIC and vice-versa?

7. What is the interpretation of negative intercept in Logistic Regression model?

8. A simple probability problem on *Bayes Theorem*.

Only Round 1.

1. What is a random forest? What is the out of bag error?

2. CV description - mainly the internship project.

3 UBS

1. Suppose I don't have deep knowledge in Statistics and Mathematics. Explain me one of your projects.

2. Explain linear regression is layman's terms.

3. You have a deck of cards. You picked 4 cards one by one. What is the probability of them being 4 different suits?

4. Draw the PDF and CDF of standard normal distribution.

5. Give me an estimated value of $e^{0.01}$. Also explain the technique of estimation. (He wanted the value using expansion ignoring the higher order terms.)

First Round.

1. Assumptions and complexity of k-means clustering.

2. Testing the significance of regression coefficients and distribution of coefficients when

errors are not normally distributed.

3. $Corr(X, Y) = 0.99$. Suppose 99-th percentile of X is known. What can we say about 99-th percentile of $X + Y$?

4. What will you do if you have a deadline tomorrow morning and tonight you find the data file corrupted?

5. Main difference between the two shrinkage methods Lasso and Ridge.

6. Prove that cdf of a continuous random variable follows $U(0, 1)$ distribution.

Second Round.

1. How many zeros are at the end of 1000! At the end of $1^1 * 2^2 * ..... * 10^{10}$?

2. Numerical approximation of $e^{0.01}$ up to three decimal places.

3. $X, Y \overset{i.i.d}{\sim} N(0, 1)$. Find $P(X > 0, Y < 0 | X + Y > 0)$.

4. What can you conclude about bias of a coin if you get 6 heads out of 10 tosses? 6000 heads out of 10000 tosses?

5. Probability of obtaining four cards of different suites while drawing randomly from a deck of cards.

6. $X \sim N(0, 1)$. Which of $F(x)$ and $f(x)$ converges to 0 faster as $x \to -\infty$?

4 Bank of America

First Round.

1. Asked about the internship and then some detailed follow up questions on ARIMA and GARCH modeling.

2. They mentioned a few time series models and then asked which were stationary and which were not.

3. 10 dice are rolled simultaneously. What is the probability that the sum is divisible by 5?

4. Buses arrive at random in a bus stand at rate 5 buses/hour. What is the probability next bus arrives after at least 5 mins given the last bus went 5 mins ago?

5. A stick of length 1 is broken at random at two places. What is the probability that the three broken parts form acute angled triangle?

6. Egg-dropping puzzle. 2 eggs, 100 floors.

Second Round.

1. Plot $x^2 + y^2 = 1$. How would the plot change if Euclidean distance was replaced by L_∞ norm?

2. Which is larger e^{π} or π^{e}?

3. Given a square frame of reference, how would you estimate the area of the map of India?

4. 20 cards are removed at random from a deck of cards. Two cards are drawn randomly from the remaining pile. What is the probability that both are Aces?

5. n dice are rolled simultaneously. What is the probability that maximum of them is 4?

6. A fair coin is tossed repeatedly. What is the expected number of tosses till we get HH? Till we get n heads in a row?

5 Eli Lilly and Company

1. How shall you explain heteroscedasticity to a layman?

2. Suppose you have two models: a simpler model and a more complex model. Which one you would prefer to choose if inference is your purpose? (They probably wanted to hear the word "interpretability".)

3. Suppose you have two vector observations. How can you measure the distance between them? what were the advantages and disadvantages of using euclidean distance?

4. Suppose you are in a team and your colleague is working from US. How would you manage a work relationship with him?

5. What are the things you look for in a job?

6. What are the differences between randomized study and observational studies? How to infer causal relationships in a randomized study?

6 BlackRock

First Round.

1. Explain the logistic regression model. How to estimate the coefficients? Distribution of errors in logistic regression.

2. How to proceed when errors violate the homoscedasticity condition in linear regression?

3. Explain power transforms (specifically wanted the Box-cox transform).

4. Fair coin is repeatedly tossed. What is the probability of obtaining the sequence HTT before THH?

5. How to test the equality of variances of two samples? How to simultaneously test equality for n samples?

6. In a game of chess each side moves twice. Can black have winning strategy?

Second Round.

1. Why are American Options priced more than European ones? Why shouldn't we exercise American options before expiration date?

2. What is VaR? What are the flaws in VaR?

3. What is the effect on coefficient estimates of Linear Regression when predictors suffer from Multicolinearity?

4. What is the change in Bias-Variance of a k-nn classifier as $k \to \infty$?

5. Explain Simpson's paradox and the effect of confounding variables.

6. Asked few R codes which could be easily done using while loops but they wanted to see how the candidate handles lapply and sapply methods.

Third Round.

Generic HR questions.

7 Dr. Reddy's Laboratories

First Round.

1. Explain projects from CV.

2. What is PCA? How do we compute PCA? Why do we use PCA? What is the motivation behind it?

3. If you have a dataset of 9000 variables can you always use PCA for dimension reduction? If not what are the other methods we can use?

4. What is LDA?

5. Tell us about any clustering method you know. (the candidate told them about k-means) Can you guarantee global optimization by using k-means? If no what are the methods you can use to ensure that?

6. What is SVM? Why do we use kernel in SVM? Do you know about RBF kernel?

Second Round.

1. Explain anyone of your projects. What were the most difficult parts of the project?

2. Where do you want to see yourself in 4-5 years?

3. Do you have any questions from us?

1. Imbalanced data. You have 6 classes of which first class has 70% and rest have 30%. How do you classify?

2. What are the missing value imputation techniques (from CV)?

3. Why 10 fold cross validation is better than 1-fold cross validation?

8 Alphonso Inc.

Code

You are given n points on the x axis and m line segments (given left and right endpoints) which are also on the x axis. You have to print the number of line segments each of the n points lie in.

For example: points 3 7 10 2

line segments (2,5) (3,4), (1,10)

Answer will be 3 1 1 2

Data Science related questions

What is linear regression and what are its assumptions? How do we check homoscedasticity? What to do if we have heteroscedastic residuals? What if the normality of errors assumption is not satisfied?

What is a decision tree? Explain how it is trained for classification. Difference between bagging decision trees and random forest.

How to use a categorical variable if it has a lot of categories, lets say 500.(One hot encoding or dummy variables) Will there be any negative effect due to this categorical variable during training of random forest?

If we do an experiment for cancer and we collect data in a retrospective manner i.e., our data contains equal proportion of both classes (has cancer or doesn't have cancer, will the coefficients for logistic regression be similar to the ones if we had collected data in a prospective manner?

First Round.

Asked about expertise in CS related area as the candidate had done two courses (Design and Analysis of Algorithms and Special Topics in Algorithms).

Given an array with integer values having duplicate elements, how to filter them out? Explained hashing. Asked to code on C++ hashing algorithm for this problem (he wanted the exact code)

Difference between binary tree and binary search tree, what is an AVL tree ,why does a tree require balancing, bit about heaps. Also a few more questions on complexity of a few

algorithms in trees.

A small question on how strings work in python.

Questions on p-value, linear regression etc

How does decision tree work? Wanted the explanation of how decision boundaries are chosen in a two way classification using the concept of entropy. Also asked about the visual representation of boundaries (basically parallel to the surface)

Basics of hierarchical clustering

Given a sample, how will you get 95% confidence interval of the median? (Basically wanted median of bootstrapped samples and choose 2.5 and 97.5 percentile)

Will logistic regression work on a completely separable data?

A question on posterior distribution which was also asked in the test. Also related questions regarding posterior distributions etc

Question on Retrospective sampling.

Second Round.

For 1 hour they asked every possible detail about the internship. It included why that project was required, what were the steps and the intuition behind them.

Third Round.

Cultural Fit Interview. It wasn't exactly a HR question answer round. One can consider it as a 50 min long conversation regarding general topics. There was no technical questions. It would be helpful to know what the company does and why do you want to join it.

9 Tata Digital.

Round 1.

1. Explain us one of your projects in your CV. Also asked some follow-up questions about internship.

2. How did you solve the problem regarding the expected number of rolls to get each of the faces of a dice?

3. When analysing any data for your projects what difficulties did you face and how did you solve them?

Round 2.

1. Introduce yourself. What are your hobbies and all?

2. So you are interested in music. Let us consider an app like google music. How do you model the recommendations for a particular user? What all variables you would like to consider as the data?

3. How does logistic regression work? What if there are multiple classes?

4. What are the clustering techniques you know?

5. How do you classify a new observation in KNN method? What are the distance metrics used? Which metric is favoured when?

6. How do you handle imbalanced dataset?

7. Consider the Amazon Prime platform. What data you would want to analyse from a data scientist point of view? And what betterment would you suggest?

8. What makes you to choose TATA digital?

9. Tell us 3 qualities which you think a Data Scientist should have apart from knowing statistics and all.

10. What skill do you think you should improve during the behavioural training camp in TATA if you get a chance?

10 Standard Chartered

Round 1.

1. CV description.

2. rain/not rain prediction probability, of today/ yesterday weather. What is the probability of rain in day after tomorrow given yesterday is rain?

3. What is your best performance work wise according to you?

4. Which software is used to handle time based data better, give a code to extract a January set from all the time data of every day of 10 years?

5. Have you handled under pressure work? Give instances.

Round 2.

1. Ant on three vertices of triangle, probability of colliding.

2. 4x4 square(made of 1x1 unit square) , color on the outside of the bigger, so how many are with one side surface colored in these 16 i unit squares.

Goldman Sachs

Position: Analyst Location: Bengaluru

1 Selection Test Questions

The test was taken virtually in Hackerrank platform. There were 17 multiple choice questions (8 on quant, 7 on computer science and 2 on coding), one coding problem and the following subjective question:

> Describe the time when your excitement and passion about a project or goal motivated you to achieve outstanding results on the project.

The duration of the test was probably 2 hours or 2.25 hours. Some multiple choice questions are mentioned below:

1. Let $A = \frac{x}{1-x^2} + \frac{x^2}{1-x^4} + \ldots + \frac{x^{32}}{1-x^{64}}$ where $x \neq 1$. Simplify A.

2. $p + q + r + s = 0$, $p^2 + q^2 + r^2 + s^2 = 46$, and $p^3 + q^3 + r^3 + s^3 = 898$. What is a possible value of $pqrs$? [options were 72, 60, 40 and can not be determined.]

3. Let A be a 4×4 matrix such that $\forall(i, j)$, A_{ij} is either 0 or 1. Sum of each row of A is odd and also sum of each column of A is odd. How many such A are possible? [options were 100, 250, 500 and 1000.]

Apart from these, there were mcqs on Binary Search Trees, Selection Sort, Encapisulation, Process Matrix, Allocation Matrix, Picking aces from a deck of cards, Polynomials, Equation solving, Colored tiles, GCD etc.

2 Interview Questions

There were two rounds of technical interviews.

Algorithms.

Input is n, and a list of n numbers from 1 to n with one number missing, and one repeated twice. Find the missing number.

(a) What is the time complexity of your algorithm?

(b) What is the space complexity of your compiling time, (extra space used).

(c) Can you reduce space complexity?

Statistics

1. What is mle? If I toss a coin 10 times and get 4 heads, what is the mle of the head probability?

2. Describe the algorithm for k-means. when should we not use k-means? (This was from CV)

3. What is the expected number of tosses till you get two consecutive heads. (Link)

4. What is the expected number of throws of a die till you see all 6 numbers? (Link) See Coupon Collector Problem.

Round 1

1. You are given a circle. If you pass a line through it you divide it into 2 regions. If you take a second line, then you can divide it into a maximum of 4 regions. You can similarly calculate for 3 lines. Give me a solution for maximum regions that a circle can be divided using n lines.

2. You are given a bag with 1000 coins. One of the coin is biased and gives heads with probability 1. You take out 1 coin from the bag and toss it 10 times. Find the probability of it being a fair coin, given you observed all 10 heads.

3. Code: You are given an array of integers A. Print another array B such that i'th element of B is the product of all elements of A except A's i'th element. For example, $A = [2, 3, 5]$, then $B = [15, 10, 6]$.

4. What is a stationary time series? Difference between weakly stationary and strictly stationary?(The candidate told that he mentioned a project related to time series in his CV)

5. You keep rolling a dice till you observe 3 sixes consecutively. What is the expected number of throws? (Ans. 258, Link)

Round 2

1. Write code to find the n'th term of Fibonacci sequence (both recursive and dynamic programming solution). Which solution has less time complexity? Which one has less space complexity?

2. You are given 10 bags with infinite coins and a weighing machine. Each coin weighs 10 grams. One of the bags is fake and the coins in that weigh 9 grams. How will you find the bag with fake coins while minimizing the number of uses of weighing machine. (It can be done by using the weighing machine just once!)

3. Coupon Collector Problem.

4. Suppose there are 3 doors and one of them has a prize behind them. You choose one door. The host opens one other door and shows it is empty. Should you switch to one of the other 2 doors or not? (Yes, because the probability of each of the other two doors containing the

prize is 3/8 now) See Monty Hall Problem.

1. A series of Questions on projects mentioned in CV.

2. State assumptions of Linear Regression. How would you check normality assumption? How would you assess model fit?

3. What is Dynamic programming? Questions on Fibonacci sequence. Can dynamic programming reduce both time complexity and space complexity together?

4. What is the expected number of tosses till you get two consecutive sixes? What is the expected number of tosses till you get *three* consecutive sixes?

5. Monty Hall Problem.

1. Monty Hall Problem.

2. What is the expected number of throws till you get *three* consecutive sixes?

3. What is the expected number of throws till you get all the six faces?

4. You throw a die three times. What is the probability that the three faces shown are in decreasing order? What is this probability that they are in increasing order?

This pdf has a nice collection of dice problems (and their solutions too). Give a look at the problems 1, 2, 5, 8, 11, 12, 13. For tricky problems on probability, Frederick Mosteller's *Fifty Challenging Problems in Probability with Solutions* is a classic.

Position: Data Scientist

Location: Mumbai

1 Selection Test Questions

The test was taken virtually in Hackerrank platform. There were 12 short-answer type questions. The duration of the test was 75 minutes. Some questions are mentioned below:

1. Why Ridge regression can be thought as a smooth version of Principal Component Regression (PCR)?

Both PCR and Ridge regression operate via the principal components of the input matrix. Ridge regression shrinks the coefficients of the principal components, shrinking more depending on the size of the corresponding eigenvalue. PCR discards the $p - M$ smallest eigenvalue components (here M is the number of principal components and p is the original number of features).

2. Describe a situation when false positives are more important than false negatives.

This is the case when "positive" means a really positive situation. Suppose we have an AI which predicts when should I invest in a particular stock. Suppose it gives false positive alert, i.e, it tells me to invest in a bad situation. This might prove costly! Suppose it gives false negative alert, i.e, I missed an investment opportunity. This should not affect much!

3. Between L1 and L2 regularization, which one leads to a unique solution? Which one is more stable?

4. Let $X \sim Poisson(\lambda)$. Suppose $P(X = 2) = 3 \cdot P(X = 4)$. $\lambda =$?

5. What does it mean for R^2 to be negative?

6. Given an unfair coin with $P(head) \neq .5$, how can you make a list of random 0's and 1's?

7. Suppose you have observations $0, 1, \ldots, n$ with respective frequencies $\binom{n}{0}, \binom{n}{1}, \ldots, \binom{n}{n}$. Compute the variance of this distribution.

8. How would you use machine learning to improve a placement process?

9. Suppose you have observations $y_1, \ldots, y_n$. Let R be the sample range and s^2 be the sample variance (with $n-1$ in denominator). Show that $s^2 \leq \frac{n}{n-1} R^2$.

Apart from these, there were problems on hypothesis testing and gambling.

IN CRF, the following was mentioned:

Must Have:

• Strong fundamentals in applied statistics and probability.

• Understands the math and theory behind basic statistical models, statistical tests as well as deep learning models.

• Should be comfortable with end to end machine learning lifecycle from Research, EDA, transformations, modelling, error analytics, versioning and scalability.

• Hypothesis testing & experimental design.

• In Depth understanding of ML algorithms (Linear models, k-means, k-NN, Decision Trees, Random Forest, Support Vector Machines, Gradient Boosted Machines).

• Proficient in Python, R, Julia.

Good to Have:

• Implemented a few good Statistical Learning or Machine Learning papers from scratch.

• Understand concepts such as cost sensitive learning, tweaking of cost functions, deep learning concepts and implementations.

Position: Analyst – Quantitative Research & Trading

1 Selection Test Questions

The test was taken virtually in Hackerrank platform. There were 19 (12 multiple choice questions with exactly one correct option, 3 mcqs with possibly several correct options, 3 "complete the sentence" type questions and 1 coding problem) questions. There was no negative marking. The duration of the test was 90 minutes.

Multiple choice questions with exactly one correct option:

1. $X_1, \ldots, X_{75} \overset{iid}{\sim} (\mu = 2, \sigma^2 = 3)$. Approximate $P\left(\sum_{i=1}^{75} X_i > 120\right)$. You may assume $P(|Z| < 2) = 0.95$.

2. A rider is waiting for a bus which is always X hours late, $X \sim U[0, A]$. Prior on A: $\frac{5}{7}, \frac{4}{9}, \frac{3}{4}, \frac{1}{2}$ with equal probabilities. Suppose, on one particular day, $X = 12$ is observed. What is the most likely value of A, given this observation?

3. What is the pdf of $Y = e^{aZ+b}$ where $Z \sim N(0, 1)$?

4. There are N billiard balls numbered 1 to N. Four are randomly selected, having labels $3, 4, 6$ and 11. What is the MLE of N?

5. What is the probability of getting two aces when two cards are drawn at random from a deck of cards?

6. Which of the following is true?

    1. Outliers decrease correlation.
    2. Outliers increase correlation.
    3. Outliers have no effect on correlation.
    4. Outliers could increase or could decrease correlation.

7. Two friends decide to meet between 9 pm and 9.30 pm on a given day. Whoever arrives first will wait for the other for 15 minutes or upto 9.30 (whichever is earlier). What is the probability that they will meet that day?

8. Which distribution would you use to model bimodal data?

There were questions on hypothesis testing and neural networks also.

Multiple choice questions with several correct options:

1. Which of the following statistics does cross validation reduce? Pick one or more options.

    1. Bias

2. Variance
3. Type I Error
4. Type II Error

2. Which of the following are valid forms of regularization in Neural Networks?

1. Convolution
2. L1 regularization
3. Pooling
4. Dropout
5. Early stopping

---

1. You are throwing a fair die multiple times. What is the expected number of throws to get 3 consecutive sixes?

2. What is the maximum value of $(X+1)^2 + (Y-1)^2$ subject to $(X-2)^2 + (Y+3)^2 \leq 64$?