**Data Science Interview Questions**

**1. Statistics Interview Questions**

1. What is the Central Limit Theorem and why is it important?
   * “Suppose that we are interested in estimating the average height among all people. Collecting data for every person in the world is impossible. While we can’t obtain a height measurement from everyone in the population, we can still sample some people. The question now becomes, what can we say about the average height of the entire population given a single sample. The Central Limit Theorem addresses this question exactly.” *Read more*[*here*](https://spin.atomicobject.com/2015/02/12/central-limit-theorem-intro/)*.*
2. What is sampling? How many sampling methods do you know?
   * “Data sampling is a statistical analysis technique used to select, manipulate and analyze a representative subset of data points to identify patterns and trends in the larger data set being examined.” *Read the full answer* [*here*](https://searchbusinessanalytics.techtarget.com/definition/data-sampling)*.*
3. What is the difference between type I vs type II error?
   * “A type I error occurs when the null hypothesis is true, but is rejected. A type II error occurs when the null hypothesis is false, but erroneously fails to be rejected.” *Read the full answer* [*here*](https://www.datasciencecentral.com/profiles/blogs/understanding-type-i-and-type-ii-errors)*.*
4. What is linear regression? What do the terms p-value, coefficient, and r-squared value mean? What is the significance of each of these components?
   * A linear regression is a good tool for quick predictive analysis: for example, the price of a house depends on a myriad of factors, such as its size or its location. In order to see the relationship between these variables, we need to build a linear regression, which predicts the line of best fit between them and can help conclude whether or not these two factors have a positive or negative relationship. *Read more* [*here*](https://www.springboard.com/blog/linear-regression-in-python-a-tutorial/) *and*[*here.*](http://blog.minitab.com/blog/adventures-in-statistics/how-to-interpret-regression-analysis-results-p-values-and-coefficients)
5. What are the assumptions required for linear regression?
   * There are four major assumptions: 1. There is a linear relationship between the dependent variables and the regressors, meaning the model you are creating actually fits the data, 2. The errors or residuals of the data are normally distributed and independent from each other, 3. There is minimal multicollinearity between explanatory variables, and 4. Homoscedasticity. This means the variance around the regression line is the same for all values of the predictor variable.
6. What is a statistical interaction?
   * ”Basically, an interaction is when the effect of one factor (input variable) on the dependent variable (output variable) differs among levels of another factor.” *Read more*[*here*](http://icbseverywhere.com/blog/mini-lessons-tutorials-and-support-pages/statistical-interactions/)*.*
7. What is selection bias?
   * “Selection (or ‘sampling’) bias occurs in an ‘active,’ sense when the sample data that is gathered and prepared for modeling has characteristics that are not representative of the true, future population of cases the model will see. That is, active selection bias occurs when a subset of the data are systematically (i.e., non-randomly) excluded from analysis.” *Read more* [*here*](https://www.elderresearch.com/blog/selection-bias-in-analytics)*.*
8. What is an example of a data set with a non-Gaussian distribution?
   * “The Gaussian distribution is part of the Exponential family of distributions, but there are a lot more of them, with the same sort of ease of use, in many cases, and if the person doing the machine learning has a solid grounding in statistics, they can be utilized where appropriate.” *Read more* [*here*](https://www.quora.com/Most-machine-learning-datasets-are-in-Gaussian-distribution-Where-can-we-find-the-dataset-which-follows-Bernoulli-Poisson-gamma-beta-etc-distribution)*.*
9. What is the Binomial Probability Formula?
   * “The binomial distribution consists of the probabilities of each of the possible numbers of successes on N trials for independent events that each have a probability of π (the Greek letter pi) of occurring.” *Read more* [*here*](http://onlinestatbook.com/2/probability/binomial.html)*.*

<https://www.youtube.com/watch?v=5JZsSNLXXuE>

1. **How do you build a random forest model?**
   1. Randomly select k features from total of m features k< m
   2. Among the k features calculate the best node d using the best split
   3. Split the node into daughter nodes using the best split
   4. Repeat steps 2 &3 until leaf nodes finalized
   5. Build forest by steps 1 to 4 for n number of times to create n number of trees
   6. Has a vote at the end who ever gets most votes wins
2. **How can you avoid overfitting of your model?**
   1. Keep the model simple – take into account fewer variables by removing some of the noise
   2. Use cross validations techniques such as k-folds cross-validation
   3. Use regularization techniques such as lasso that penalize certain parameters if they’re likely to cause overfitting