**Sutherland interview Question and Answer**

1. **Central Limit Theorem (CLT)**

**Let’s start with one example,** Consider that there are 15 sections in the science department of a university and each section hosts around 100 students. Our task is to calculate the average weight of students in the science department. Sounds simple, right?

The approach I get from aspiring data scientists is to simply calculate the average:

* First, measure the weights of all the students in the science department
* Add all the weights
* Finally, divide the total sum of weights with a total number of students to get the average

But what if the size of the data is humongous? Does this approach make sense? Not really – measuring the weight of all the students will be a very tiresome and long process. So, what can we do instead? Let’s look at an alternate approach.

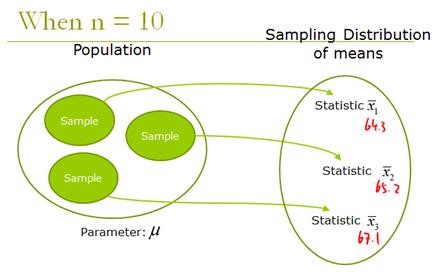
* First, draw groups of students at random from the class. We will call this a sample. We’ll draw multiple samples, each consisting of 30 students.
* Calculate the individual mean of these samples
* Calculate the mean of these sample means
* This value will give us the approximate mean weight of the students in the science department
* Additionally, the histogram of the sample mean weights of students will resemble a bell curve (or normal distribution)

**Significance of the Central Limit Theorem**

The central limit theorem has both statistical significance as well as practical applications. Isn’t that the sweet spot we aim for when we’re learning a new concept?

We’ll look at both aspects to gauge where we can use them.

### **Statistical Significance of CLT**



* Analyzing data involves statistical methods like hypothesis testing and constructing confidence intervals. These methods assume that the population is normally distributed. In the case of unknown or non-normal distributions, we treat the sampling distribution as normal according to the central limit theorem
* If we increase the samples drawn from the population, the standard deviation of sample means will decrease. This helps us estimate the population mean much more accurately
* Also, the sample mean can be used to create the range of values known as a confidence interval (that is likely to consist of the population mean)

**Assumptions Behind the Central Limit Theorem**

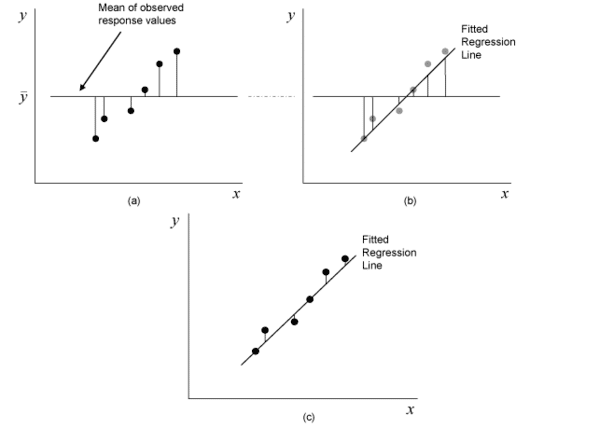
1. The data must follow the randomization condition. It must be sampled randomly
2. Samples should be independent of each other. One sample should not influence the other samples
3. Sample size should be not more than 10% of the population when sampling is done without replacement
4. The sample size should be sufficiently large. Now, how we will figure out how large this size should be? Well, it depends on the population. When the population is skewed or asymmetric, the sample size should be large. If the population is symmetric, then we can draw small samples as well.

**Question 2. Explain f-test.**

Sometimes we want to compare a model that we have calculated to a mean. For example, let’s say that you have calculated a [linear regression](https://magoosh.com/statistics/2018/01/04/what-is-the-regression-equation/) model. Remember that the [mean](https://magoosh.com/statistics/2017/12/21/statistics-basics-heres-what-you-need-to-know/)is also a model that can be used to explain the data.

The *F*-Test is a way that we compare the model that we have calculated to the overall mean of the data. Similar to the *t*-test, if it is higher than a critical value then the model is better at explaining the data than the mean is.

Before we get into the nitty-gritty of the *F*-test, we need to talk about the sum of squares. Let’s take a look at an example of some data that already has a line of best fit on it.

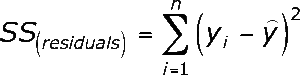


The *F*-test compares what is called the mean sum of squares for the *residuals* of the model and and the overall mean of the data. Party fact, the residuals are the difference between the actual, or observed, data point and the predicted data point.

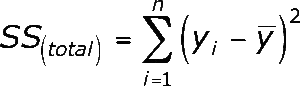
In the case of graph (a), you are looking at the residuals of the data points and the overall sample mean. In the case of graph (c), you are looking at the residuals of the data points and the model that you calculated from the data. But in graph (b), you are looking at the residuals of the *model* and the overall sample mean.

The sum of squares is a measure of how the residuals compare to the model or the mean, depending on which one we are working with. There are three that we are concerned with.

The *sum of squares of the residuals* (SSR) is the sum of the squares of the residuals between the data points and the actual regression lines, like graph (c). They are squared to compensate for the negative values. SSR is calculated by



The sum of squares of the total (SST) is the sum of the squares of the residuals between the data points and the mean of the sample, like graph (a). They are squared to compensate for the negative values. SST is calculated by



It is important to note that while the equations may look the same at first glance, there is an important distinction. The SSR equation involves the predicted value, so the second Y has a little carrot over it (pronounced Y-hat). The SST equation involves the sample mean, so the second Y has a little bar over it (pronounced Y-bar). Don’t forget this very important distinction.

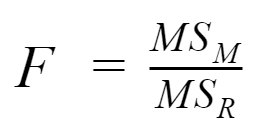
The difference between the two (SSR – SST) will tell you the overall sum of squares for the model itself, like graph (b). This is what we are after in order to finally start to calculate the actual *F* value.

These sum of squares values give us a sense of how much the model varies from the observed values, which comes in handy in determining if the model is really any good for prediction. The next step in the *F*-test process is to calculate the mean of squares for the residuals and for the model.

To calculate the *mean of squares of the model*, or MSM, you need to know the degrees of freedom for the model. Thankfully, it is pretty straightforward. The degrees of freedom for the model is the number of variables in the model! Then follow the formula MSM = SSM ÷ dfmodel

To calculate the *mean of squares of the residuals*, or MSR, you need to know the degrees of freedom in the sample size. The degrees of freedom in the sample size is always N – 1. Then simply follow the formula MSR = SSR ÷ dfresiduals

Ok, you have done a whole lot of math so far. I’m proud of you because I know that it is not super fun. But it is super important to know where these values come from because it helps understand how they work. Because now we are actually going to see how the F-statistic is actually calculated!



This calculation gives you a ratio of the model’s prediction to the regular mean of the data. Then you compare this ratio to an [F-distribution table](http://www.socr.ucla.edu/applets.dir/f_table.html) as you would the t-statistic. If the calculated value exceeds the critical value in the table, then the model is significantly different from the mean of the data, and therefore better at explaining the patterns in the data.

**Question 3 Explain t test.**

The *t*-test is a test statistic that compares the means of two different groups. There are a bunch of cases in which you may want to compare group performance such as test scores, clinical trials, or even how happy different types of people are in different places. Of course, different types of groups and setups call for different types of tests. The type of *t*-test that you may need depends on the type of sample that you have.

If your two groups are the same size *and* you are taking a sort of before-and-after experiment, then you will conduct what is called a Dependent or Paired Sample *t*-test.

If the two groups are different sizes or you are comparing two separate event means, then you conduct a Independent Sample *t*-test.

**Question 4. Explain ANOVA.**

ANOVA (Analysis of Variance) is used to check if at least one of two or more groups have statistically different means. Now, the question arises – Why do we need another test for checking the difference of means between independent groups? Why can we not use multiple t-tests to check for the difference in means?

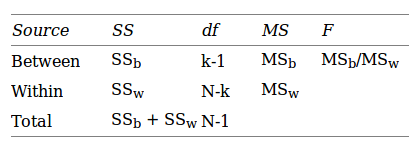
The answer is simple. Multiple t-tests will have a compound effect on the error rate of the result. Performing t-test thrice will give an error rate of ~15% which is too high, whereas ANOVA keeps it at 5% for a 95% confidence interval.

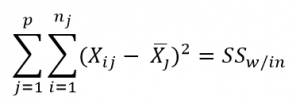
To perform an ANOVA, you must have a continuous response variable and at least one categorical factor with two or more levels. ANOVA requires data from approximately normally distributed populations with equal variances between factor levels. However, ANOVA procedures work quite well even if the normality assumption has been violated unless one or more of the distributions are highly skewed or if the variances are quite different.

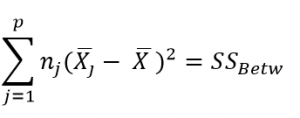
ANOVA is measured using a statistic known as F-Ratio. It is defined as the ratio of Mean Square (between groups) to the Mean Square (within group).

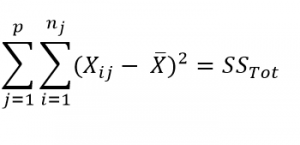
Mean Square (between groups) = Sum of Squares (between groups) / degree of freedom (between groups)

Mean Square (within group) = Sum of Squares (within group) / degree of freedom (within group)









Here, **p** = represents the number of groups

**n =**represents the number of observations in a group

=  represents the mean of a particular group

**X (bar)** = represents the mean of all the observations

Now, let us understand the degree of freedom for within group and between groups respectively.

Between groups : If there are k groups in ANOVA model, then k-1 will be independent. Hence, k-1 degree of freedom.

Within groups : If N represents the total observations in ANOVA (∑n over all groups) and k are the number of groups then, there will be k fixed points. Hence, N-k degree of freedom.

**Question 5. random forest overfitting avoidance technique.**

Random Forests are less likely to overfit but it is still something that you want to make an explicit effort to avoid. the main thing you need to do is optimize a tuning parameter that governs the number of features that are randomly chosen to grow each tree from the bootstrapped data. Typically, you do this via kk-fold cross-validation, where k∈{5,10}k∈{5,10}, and choose the tuning parameter that minimizes test sample prediction error. In addition, growing a larger forest will improve predictive accuracy, although there are usually diminishing returns once you get up to several hundreds of trees.

**Question 6. SQL code to print duplicates.**

Input

|  |  |
| --- | --- |
| ID | NAME |
| 1 | iNeuron |
| 2 | One neuron |
| 3 | iNeuron |

Expected Output:

|  |  |
| --- | --- |
| NAME | num |
| iNeuron | 1 |
| One neuron | 2 |

Duplicated NAME existed more than one time, so to count the times each NAME exists, we can use the following code:

select NAME, count(NAME) as num

from Person

group by NAME;

**Question 7. regularisation technique for feature selection and how are features reduced.**

Regularisation consists in adding a **penalty to** the different parameters of the machine learning model to reduce the freedom of the model and in other words to avoid overfitting. In linear model regularisation, the penalty is applied over the coefficients that multiply each of the predictors.

There is some techniques to do a feature selection.

* **Wrapper methods (forward, backward, and stepwise selection)**,
* Filter methods (ANOVA, Pearson correlation, variance thresholding),
* Embedded methods (Lasso, Ridge, Decision Tree)

**Question 8. Heteroskedasticity and how does it happen in regression.**

Heteroscedasticity is a systematic change in the spread of the residuals over the range of measured values. Heteroscedasticity is a problem because ordinary least squares (OLS) regression assumes that all residuals are drawn from a population that has a constant variance (homoscedasticity).

**There are three common ways to fix heteroscedasticity in regression:**

1. Transform the dependent variable. One way to fix heteroscedasticity is to transform the dependent variable in some way.
2. Redefine the dependent variable. Another way to fix heteroscedasticity is to redefine the dependent variable.
3. Use weighted regression.

**Question 9. odd function in logistic regression**

The important thing to remember about the odds ratio is that an odds ratio greater than 1 is a positive association (i.e., higher number for the predictor means group 1 in the outcome), and an odds ratio less than 1 is negative association (i.e., higher number for the predictor means group 0 in the outcome.