**A/B Testing and Statistics Questions:**

What is the **Central Limit Theorem**?

* The central limit theorem tells us that if we take large enough samples from a population, the mean of the samples (aka sampling distribution), will fall within a normal distribution. Regardless of the distribution of the original population.
* In practice, let’s say you have a population that has some sort or distribution that does not have to be normal. You take 1000 samples of size N and then you take the mean of each of these 1000 samples. The resulting distribution of means is called your sampling distribution and as you increase N, that distribution will become more and more normally distributed. This is important because it tells us that if we take a sample from the population with a sufficiently large N, the mean of that sample would be a good representation of our population mean.
* Also, the variance of our sampling distribution of 1000 samples = population variance / N <- not as practical because usually we aren’t taking multiple samples, we are usually just taking 1.

Why is the Central Limit Theorem Important?

* Imagine a population that is ugly and messy and not normal at all. Now you don’t know the mean and variance of the population, but you need to use the mean to do some comparisons. The central limit theorem claims that if you take a sample of 30 from the population and the resulting mean would be a good representation of your population mean.

In classification, what do you usually use for **performance metric?**

* It depends on what you care about more.
* Precision is what percentage of positive predictions are correct. TP/TP+FP. You use precision when the cost of false positives is high. For example, if you’re convicting a criminal from your model, you care a lot about avoiding false positives. In my work, precision matters more because throwing a perfectly good chip away can be extremely costly. We have safeguards later on in line to catch false negatives anyways.
* Recall is what percentage of TP is described as positive by the model. TP/TP+FN = TP/(actual positives). In this case, you use recall when the cost of false negatives is high. For example, your model predicts whether or not the patient has cancer. You want to avoid false negatives.
  + Don’t want to maximize either precision or recall because there’s a trade off. Perfect precision might give you 0 recall. Use it as guidelines not to maximize.
* F1 score is the harmonic mean between precision and recall. Best F1 score is at 1 and worst is at 0.
* Accuracy – Ratio of all correct predictions over all observations. TP+TN / TP+TN+FP+FN. This is a really good and intuitive performance metric if you have balanced labels. If your labels are not balanced, you can just predict the majority case and end up with a high accuracy. Biased on test size and distribution.
* ROC curve – True Positive Rate TPR: TP/TP+FN (Recall) (y axis) vs. False Positive Rate FPR: FP/FP+TN (1 - specificity) (x axis).
  + Let’s say you want to find planes in the sky. Your model can be super picky and not label anything in the sky as a plane. Then your TPR will be very low. Your FPR will also be very low because you won’t be making any false positive guesses.
  + Your model can be super loose and label everything flying by as a plane. In that case, you will have a high True Positive Rate because you’re catching all the planes. But your False Positive Rate will also be very high because you’re calling all the birds a plane.
  + The ROC curve plots different classification thresholds from loose to strict and you’ll have a rather smooth curve telling you your model’s TRP and FPR at each threshold.
  + If your model randomly calls an object a plane or not, your ROC curve will be a straight line and the Area under the curve would be 0.5
  + You want your model to have higher TPR than FPR so you want to see the ROC curve bow outwards and give you an AUC that is > 0.5.
* Area under the curve – Literally the area below the ROC curve.
  + Tells you how well the model is able to separate the classes. Between 0 and 1, the probability that your model would predict a positive as positive over a negative. In other words, it measures the performance of your model based on how many true positives it finds vs. how may false positives it finds.
  + Benefits of AUC is that the results does not depend on the classification threshold. You can use ROC to help determine your classification threshold.
  + Also, it measures how well your predictions are ranked, rather than their absolute values. Good to use when comparing the results of different modeling algorithms.
* Shapley Value –

Consider a classification model that separates email into two categories: “spam” or “not spam”. If you raise the classification threshold, what will happen to precision?

* Precision is TP/TP + FP. Increasing the classification threshold means you predict fewer positives and more negatives. That means your # of TP will decrease or stay the same and the number of FP will also decrease or stay the same. Therefore, in most cases, your precision will probably increase.

Consider a classification model that separates email into two categories: “spam” or “not spam”. If you raise the classification threshold, what will happen to recall?

* Recall is TP/TP + FN. Increasing the classification threshold means you predict fewer positives and more negatives. That means your # of TP will decrease or stay the same and your # of FN will increase or stay the same. Therefore, your recall will either decrease or stay the same.

Explain how Two Sample T-Tests work

* The null hypothesis is that the difference in means of the two samples is 0
* You calculate the difference in means and a 95% confidence interval around the difference in means. This should be diff in mean +- 1.96\*SE.
* This means that if you were to repeat this experiment many times, 95% of the resulting difference in means will fall within this interval.
* So if the 95% confidence interval does not contain 0, you can reject the null hypothesis that the difference in means is 0.

Explain how Paired Sample T-Tests work

* While a 2 sample T-Test deals with 2 independent data sets, a Paired Sample T-Test deals with 2 samples where both samples share the same subjects or entities. For example, you want to see if there’s a difference in means in the performance of your employees before and after training.