**Interview questions**

What are some of the downsides to using a Normal distribution to predict the maximum number of calls in a day?

- The number of calls you can expect in a day is a discrete variable, but the Normal distribution is continuous.

- The number of calls in a day can't be negative, but the Normal distribution will include negative values!

- The Normal distribution is symmetric, but if we plotted the number of calls each day, there's a good chance it won't be symmetric.

In general: \*\*"With confidence level 95%, the true population mean lies in the confidence interval."\*\*

For this example: \*\*"With confidence level 95%, the true population percentage of people who think our country is on the right track is between 24.55% to 51.45%."\*\*

- Generally, we would say:

- "I am {confidence level} % confident

- that the true population {parameter}

- is between {lower confidence bound} and {upper confidence bound}."

Two common misconceptions:

1. There is \*not\* a 95% probability that the true parameter lies within a particular confidence interval. Make sure you do not use the word probability! Instead, we are confident that over a large number of samples, 95% of them will contain the population statistic.

2. As the number of samples increases, the standard deviation of the sampling distribution decreases. However, a small standard deviation by itself does not imply that the mean is accurate. (For example, units matter!)

So how do we make the decision?

**Remember that ALPHA is our level of significance.**

- If p-value < ALPHA, then there is **evidence to reject the null hypothesis**, so you accept that $H\_0$ is incorrect and therefore $H\_A$ is correct.

- i.e., a statistically significant difference between the two groups!

- This is like saying there is enough evidence to say our dog isn't innocent... so we say our dog is guilty.

- If p-value > ALPHA, then there **is insufficient evidence to reject the null hypothesis** and you cannot accept that either $H\_0$ or $H\_A$ is correct.

- i.e., there is no statistical difference between your two groups.

- This is like saying there is not enough evidence to say our dog isn't innocent. We can't totally determine that our dog is innocent, but we haven't determined that our dog is guilty, either.

You are analyzing attrition (employees leaving a company). You have a dataset of all employees, with specific features on their role (department, position, salary, etc.) and whether or not they left the company. Your analysis should indicate how many individuals left the organization and identify any trends associated with attrition (ex. concentration, etc.) What visualizations would you build to communicate your findings to the client? (Answers can - and should - vary!)

- Since **"left company"** is a **categorical variable**, we could generate a **heatmap** comparing these values with other categorical variables like department, position, etc.

- We could **generate stacked histograms comparing "left company" for quantitative variables**. For example, visualize salary for those who stayed versus those who left.

What are the three types of error in a Machine Learning model? Briefly describe them.

1. **Bias** - error caused by choosing an algorithm that does a poor job at modeling the signal in the data, i.e. the model is bad. For example, using linear regression to model highly non-linear data would result in error due to bias.

2. **Variance** - error caused by a model not generalizing well to new data or being overfit to the training data.

3. **Irreducible error** - error caused by noise in the data that cannot be removed through modeling.

What is the bias-variance trade-off?

Bias occurs when your model is to simple and is not picking up on the complexities in the dataset (underfit). Variance occurs when your model is too complex and is modeling too much noise in the data, therefore not generalizing well to new data (overfit). The bias-variance trade off is the trade off between underfitting and overfitting. The goal of building a good machine learning model should be a balance between bias and variance: good enough to get accurate predictions but general enough to perform well on unseen data.

What is the difference between a classification and a regression problem?

- A classification problem has a categorical $Y$ variable. A regression problem has a numeric Y variable.

What are some of the benefits of logistic regression as a classifier?

OR

The original interview question was "If you're comparing decision trees and logistic regression, what are the pros and cons of each?"

(Answers may vary; this is not an exhaustive list!)

- Logistic regression is a classification algorithm that shares similar properties to linear regression.

- The coefficients in a logistic regression model are interpretable. (They represent the change in log-odds caused by the input variables.)

- Logistic regression is a very fast model to fit and generate predictions from.

- It is by far the most common classification algorithm.

What is the ROC curve?

- The ROC curve is a plot of the True Positive Rate (sensitivity) vs. the False Positive Rate (1 - specificity) for all possible decision thresholds.

Let's say you were building a search engine and wanted to build a classification model that would recommend articles based on the search input. What metric would you want to optimize for and why?

- You could make a case for wanting to minimize false positives (stories that weren't relevant), in which case you'd want to optimize for precision.

- You could make a case for wanting to minimize false negatives (not passing along possibly useful content), in which case you'd want to optimize for recall.

- Alumni Comment: "The interviewer seemed more interested in seeing if I knew what the metrics were and explaining what priorities would lead me to optimize for one over the other."

More interview practice questions on these topics [here](<https://kiwidamien.github.io/interview-practice-with-precision-and-recall.html>)!

What is the difference between hyperparameters and statistical parameters?

- **Statistical** parameters are quantities that a **model can learn or estimate.** Examples include $\beta\_0$ and $\beta\_1$ in a linear model.

- **Hyperparameters** are quantities our model cannot **learn but affect the fit of our model**. Examples include $k$ in $k$-nearest neighbors and ALPHA in regularization.