

# Machine Learning Fundamentals

Q: What are some of the steps for data wrangling and data cleaning before applying machine learning algorithms?

There are many steps that can be taken when data wrangling and data cleaning. Some of the most common steps are listed below:

- **Data profiling:** Almost everyone starts off by getting an understanding of their dataset. More specifically, you can look at the shape of the dataset with `.shape` and a description of your numerical variables with `.describe()`.
- **Data visualizations:** Sometimes, it's useful to visualize your data with histograms, boxplots, and scatterplots to better understand the relationships between variables and also to identify potential outliers.
- **Syntax error:** This includes making sure there's no white space, making sure letter casing is consistent, and checking for typos. You can check for typos by using `.unique()` or by using bar graphs.
- **Standardization or normalization:** Depending on the dataset you're working with and the machine learning method you decide to use,

it may be useful to standardize or normalize your data so that different scales of different variables don't negatively impact the performance of your model.

- **Handling null values:** There are a number of ways to handle null values including deleting rows with null values altogether, replacing null values with the mean/median/mode, replacing null values with a new category (eg. unknown), predicting the values, or using machine learning models that can deal with null values. *Read more [here](#).*
- **Other things include:** removing irrelevant data, removing duplicates, and type conversion.

*Read more on the **Amazon machine learning interview and questions** [here](#).*

Q: How to deal with unbalanced binary classification?

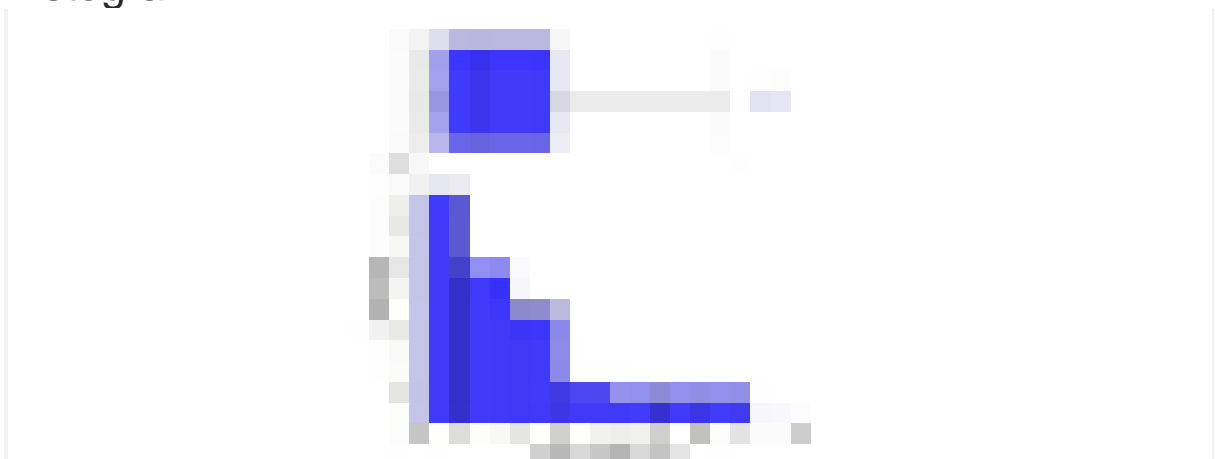
There are a number of ways to handle unbalanced binary classification (assuming that you want to identify the minority class):

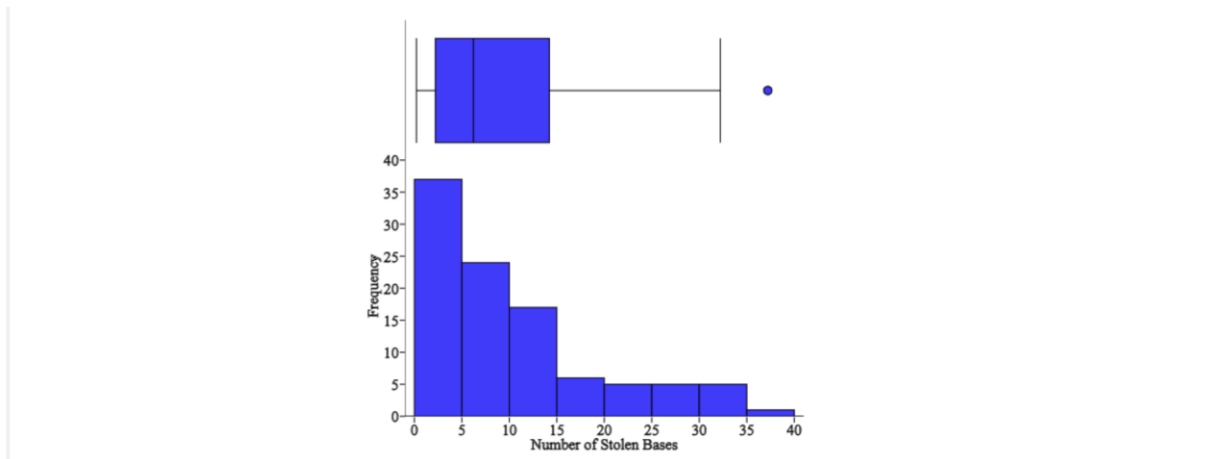
- First, you want to reconsider the **metrics** that you'd use to evaluate your model. The accuracy of your model might not be the best metric to look at because and I'll use an example to explain

why. Let's say 99 bank withdrawals were not fraudulent and 1 withdrawal was. If your model simply classified every instance as "not fraudulent", it would have an accuracy of 99%! Therefore, you may want to consider using metrics like precision and recall.

- Another method to improve unbalanced binary classification is by **increasing the cost of misclassifying** the minority class. By increasing the penalty of such, the model should classify the minority class more accurately.
- Lastly, you can improve the balance of classes by **oversampling** the minority class or by **undersampling** the majority class. You can read more about it [here](#).

Q: What is the difference between a box plot and a histogram?



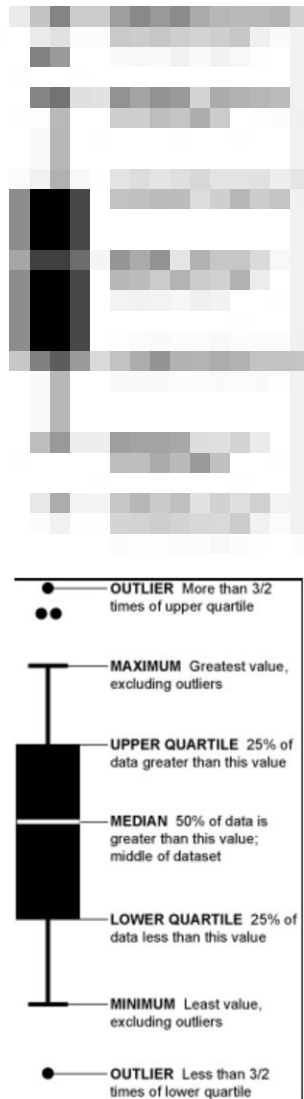


Boxplot vs Histogram

While boxplots and histograms are visualizations used to show the distribution of the data, they communicate information differently.

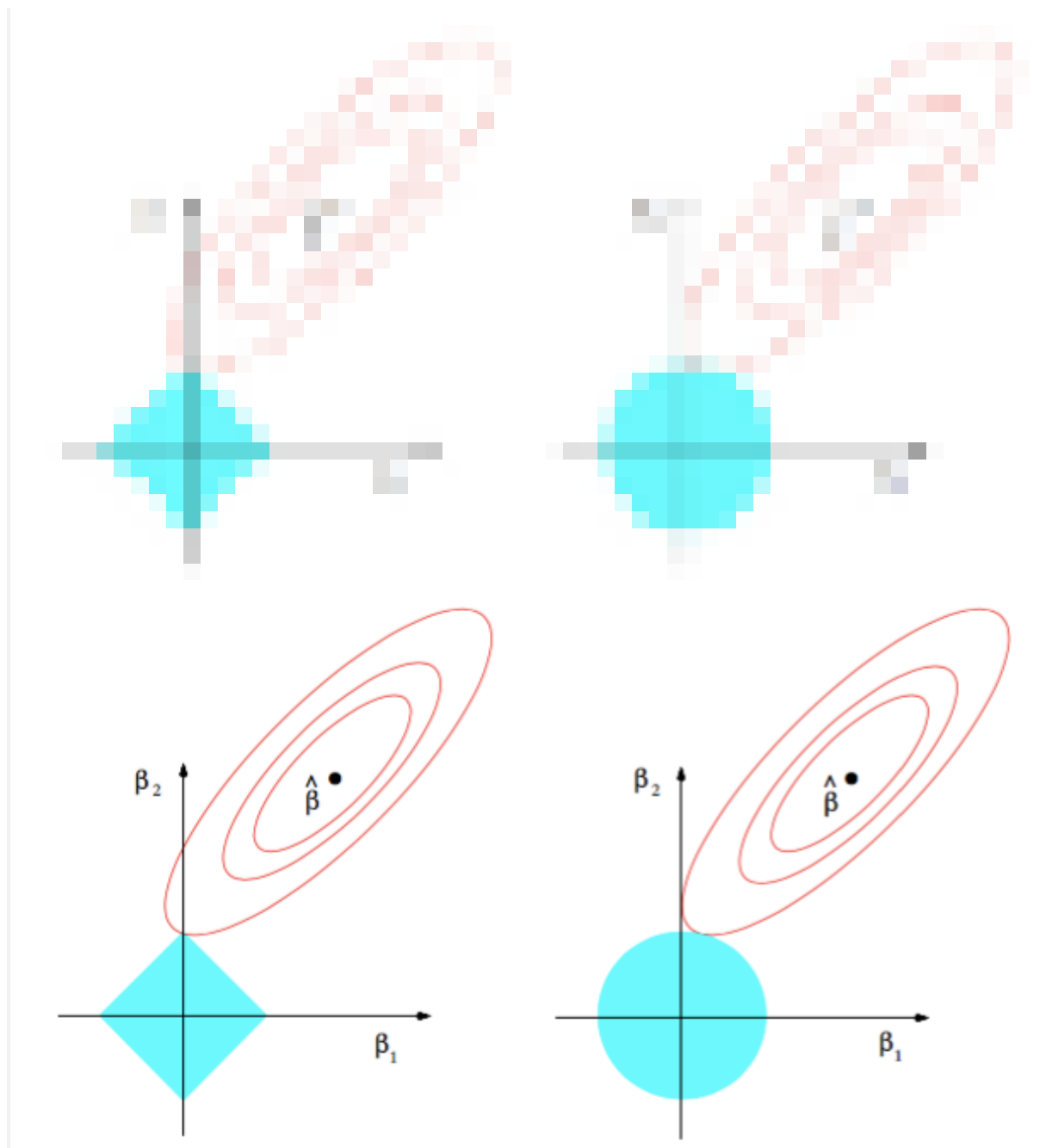
Histograms are bar charts that show the frequency of a numerical variable's values and are used to approximate the probability distribution of the given variable. It allows you to quickly understand the shape of the distribution, the variation, and potential outliers.

Boxplots communicate different aspects of the distribution of data. While you can't see the shape of the distribution through a box plot, you can gather other information like the quartiles, the range, and outliers. Boxplots are especially useful when you want to compare multiple charts at the same time because they take up less space than histograms.



How to read a boxplot

Q: Describe different regularization methods, such as L1 and L2 regularization?



Both L1 and L2 regularization are methods used to reduce the overfitting of training data. Least Squares minimizes the sum of the squared residuals, which can result in low bias but high variance.

L2 Regularization, also called ridge regression, minimizes the sum of the squared residuals **plus lambda times the slope**

**squared**. This additional term is called the **Ridge Regression Penalty**. This increases the bias of the model, making the fit worse on the training data, but also decreases the variance.

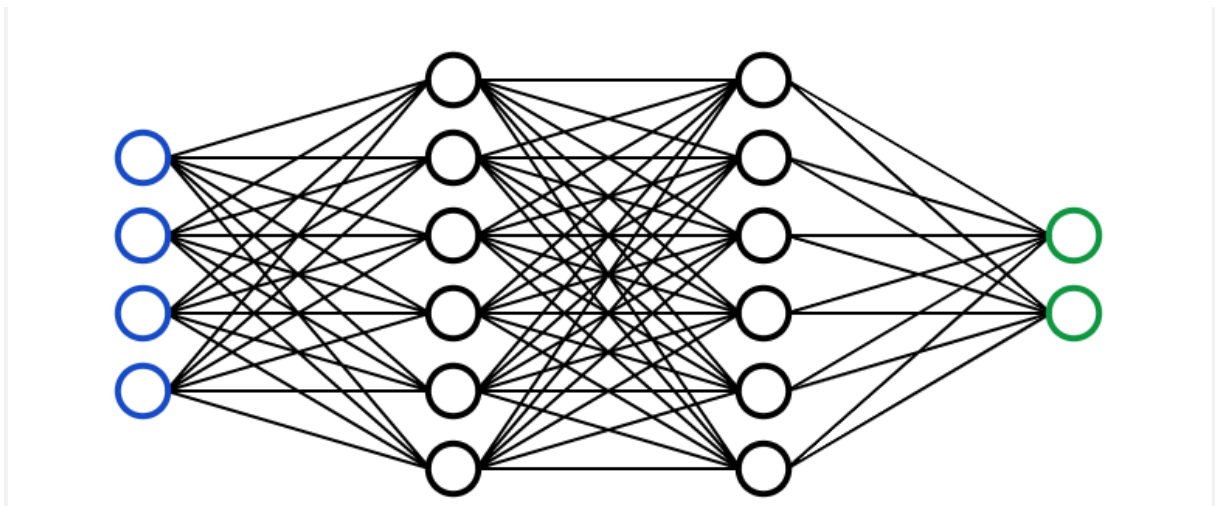
If you take the ridge regression penalty and replace it with the **absolute** value of the slope, then you get Lasso regression or L1 regularization.

L2 is less robust but has a stable solution and always one solution. L1 is more robust but has an unstable solution and can possibly have multiple solutions.

StatQuest has an amazing video on Lasso and Ridge regression [here](#).

Q: Neural Network Fundamentals





A **neural network** is a multi-layered model inspired by the human brain. Like the neurons in our brain, the circles above represent a node. The blue circles represent the **input layer**, the black circles represent the **hidden layers**, and the green circles represent the **output layer**. Each node in the hidden layers represents a function that the inputs go through, ultimately leading to an output in the green circles. The formal term for these functions is called the **sigmoid activation function**.

If you want a step by step example of creating a neural network, check out Victor Zhou's article [here](#).

If you're a visual/audio learner, 3Blue1Brown has an amazing series on neural networks and deep learning on YouTube [here](#).

Q: What is cross-validation?

Cross-validation is essentially a technique used to assess how well a model performs on a new independent dataset. The simplest example of cross-validation is when you split your data



into two groups: training data and testing data, where you use the training data to build the model and the testing data to test the model.

Q: How to define/select metrics?

There isn't a one-size-fits-all metric. The metric(s) chosen to evaluate a machine learning model depends on various factors:

- Is it a regression or classification task?
- What is the business objective? Eg. precision vs recall
- What is the distribution of the target variable?

There are a number of metrics that can be used, including adjusted r-squared, MAE, MSE, accuracy, recall, precision, f1 score, and the list goes on.

*Check out questions related to modeling metrics on [Interview Query](#)*

Q: Explain what precision and recall are

**Recall** attempts to answer “What proportion of actual positives was identified correctly?”

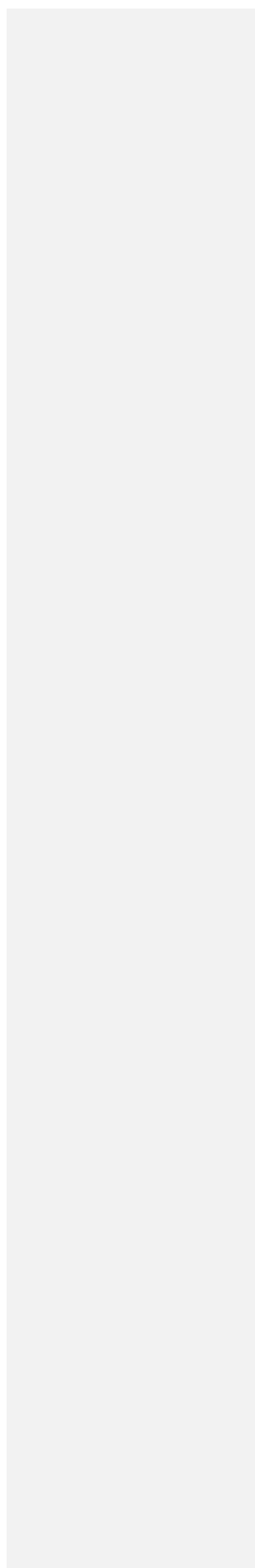
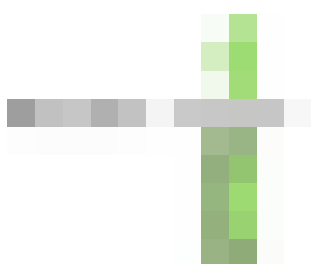
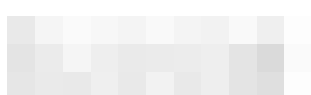
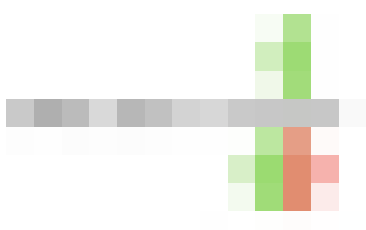
$$\text{Recall} = \frac{TP}{TP + FN}$$

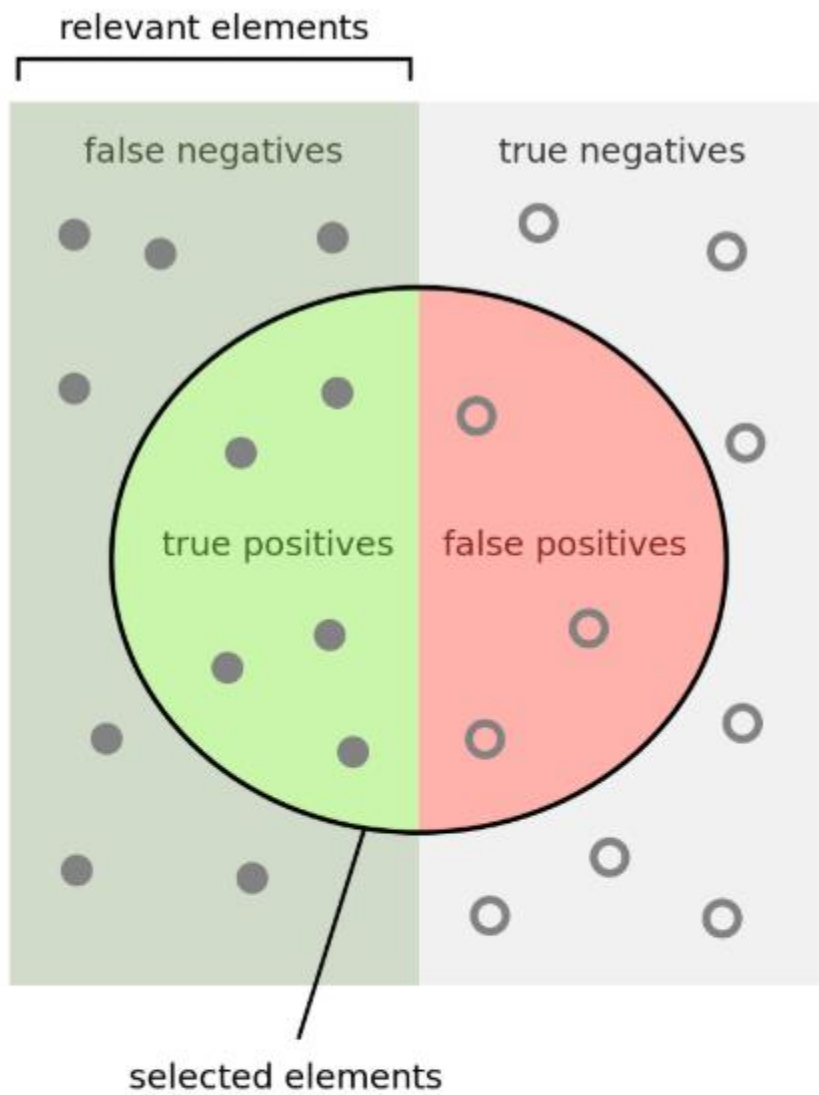
$$\text{Recall} = \frac{TP}{TP + FN}$$

**Precision** attempts to answer “What proportion of positive identifications was actually correct?”

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$





How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

Taken from Wikipedia

Q: Explain what a false positive and a false negative are. Why is it important these from each other? Provide examples when false positives are more important than false negatives, false negatives are more important than false positives and when these two types of errors are equally important

A **false positive** is an incorrect identification of the presence of a condition when it's absent.

A **false negative** is an incorrect identification of the absence of a condition when it's actually present.

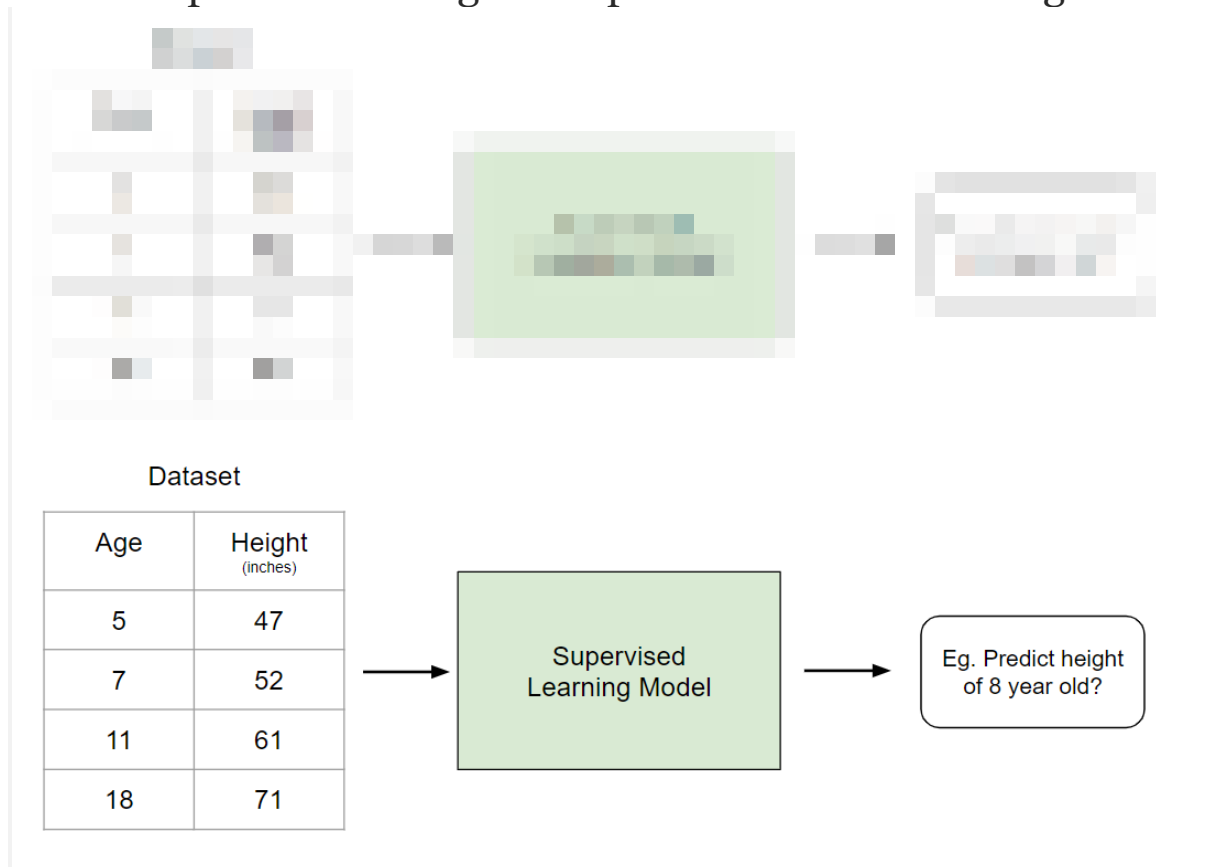
An example of when false negatives are more important than false positives is when screening for cancer. It's much worse to say that someone doesn't have cancer when they do, instead of saying that someone does and later realizing that they don't.

This is a subjective argument, but false positives can be worse than false negatives from a psychological point of view. For example, a false positive for winning the lottery could be a worse outcome than a false negative because people normally don't expect to win the lottery anyways.

Q: What is the difference between supervised learning and unsupervised learning? Give concrete examples

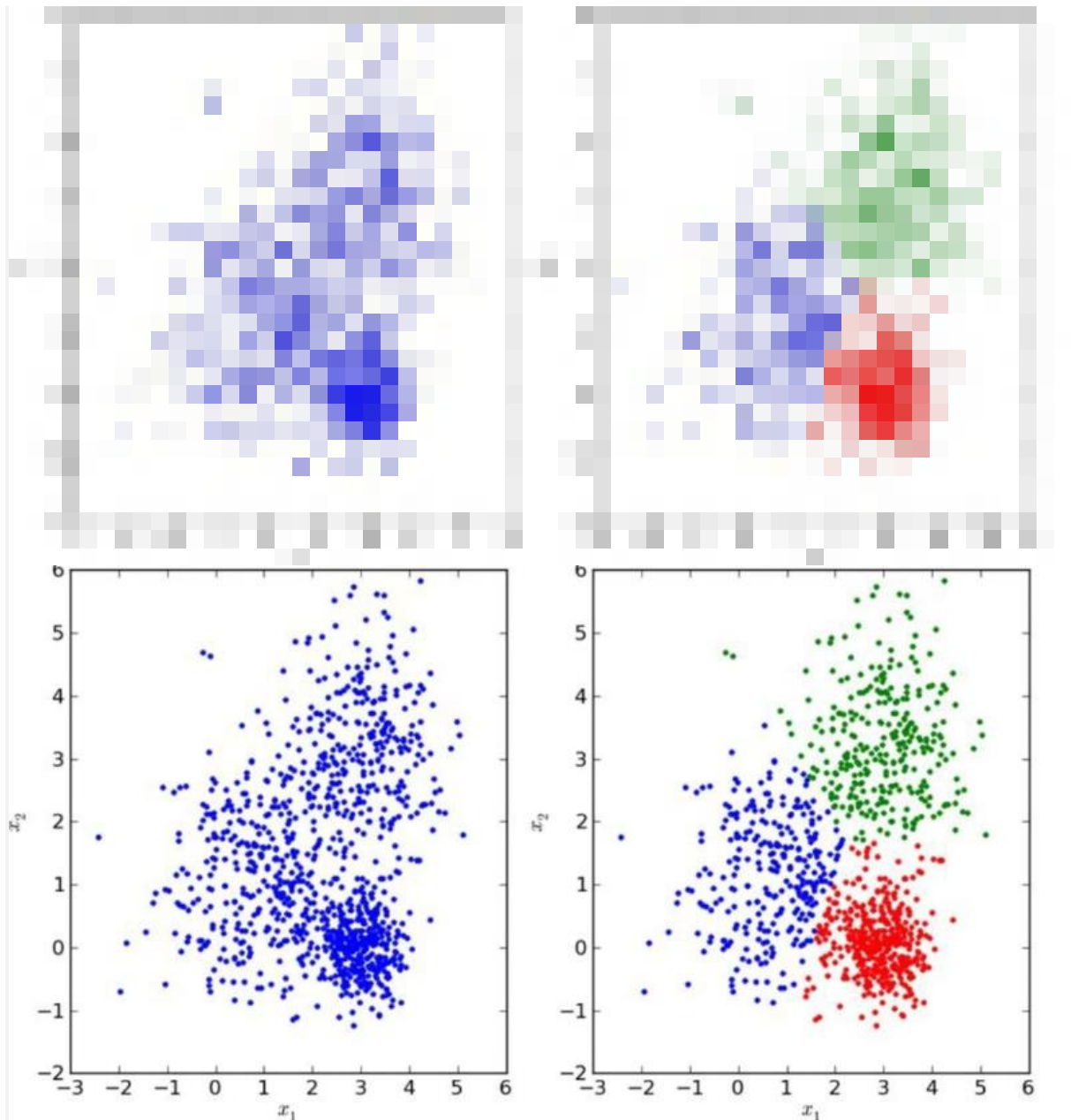
**Supervised learning** involves learning a function that maps an input to an output based on example input-output pairs [1].

For example, if I had a dataset with two variables, age (input) and height (output), I could implement a supervised learning model to predict the height of a person based on their age.



Created by author

Unlike supervised learning, **unsupervised learning** is used to draw inferences and find patterns from input data without references to labeled outcomes. A common use of unsupervised learning is grouping customers by purchasing behavior to find target markets.



Check out my article [‘All Machine Learning Models Explained in Six Minutes’](#) if you’d like to learn more about this!

Q: Assume you need to generate a predictive model using multiple regression. Explain how you intend to validate this model

There are two main ways that you can do this:

## A) Adjusted R-squared.

R Squared is a measurement that tells you to what extent the proportion of variance in the dependent variable is explained by the variance in the independent variables. In simpler terms, while the coefficients estimate trends, R-squared represents the scatter around the line of best fit.

However, every additional independent variable added to a model **always** increases the R-squared value — therefore, a model with several independent variables may seem to be a better fit even if it isn't. This is where adjusted  $R^2$  comes in. The adjusted  $R^2$  compensates for each additional independent variable and only increases if each given variable improves the model above what is possible by probability. This is important since we are creating a multiple regression model.

## B) Cross-Validation

A method common to most people is cross-validation, splitting the data into two sets: training and testing data. *See the answer to the first question for more on this.*

Q: What does NLP stand for?

NLP stands for **Natural Language Processing**. It is a branch of artificial intelligence that gives machines the ability to read and understand human languages.

Q: When would you use random forests Vs SVM and why?



There are a couple of reasons why a random forest is a better choice of model than a support vector machine:

- Random forests allow you to determine the feature importance. SVM's can't do this.
- Random forests are much quicker and simpler to build than an SVM.
- For multi-class classification problems, SVMs require a one-vs-rest method, which is less scalable and more memory intensive.

Q: Why is dimension reduction important?

Dimensionality reduction is the process of reducing the number of features in a dataset. This is important mainly in the case when you want to reduce variance in your model (overfitting).

Wikipedia states four advantages of dimensionality reduction ([see here](#)):

1. *It reduces the time and storage space required*
2. *Removal of multi-collinearity improves the interpretation of the parameters of the machine learning model*
3. *It becomes easier to visualize the data when reduced to very low dimensions such as 2D or 3D*

#### 4. *It avoids the curse of dimensionality*

Q: What is principal component analysis? Explain the sort of problems you would use PCA for.

In its simplest sense, PCA involves project higher dimensional data (eg. 3 dimensions) to a smaller space (eg. 2 dimensions). This results in a lower dimension of data, (2 dimensions instead of 3 dimensions) while keeping all original variables in the model.

PCA is commonly used for compression purposes, to reduce required memory and to speed up the algorithm, as well as for visualization purposes, making it easier to summarize data.

Q: Why is Naive Bayes so bad? How would you improve a spam detection algorithm that uses naive Bayes?

One major drawback of Naive Bayes is that it holds a strong assumption in that the features are assumed to be uncorrelated with one another, which typically is never the case.

One way to improve such an algorithm that uses Naive Bayes is by decorrelating the features so that the assumption holds true.

Q: What are the drawbacks of a linear model?

There are a couple of drawbacks of a linear model:

- A linear model holds some strong assumptions that may not be true in application. It assumes a

linear relationship, multivariate normality, no or little multicollinearity, no auto-correlation, and homoscedasticity

- A linear model can't be used for discrete or binary outcomes.
- You can't vary the model flexibility of a linear model.

Q: Do you think 50 small decision trees are better than a large one? Why?

Another way of asking this question is “Is a random forest a better model than a decision tree?” And the answer is yes because a random forest is an ensemble method that takes many weak decision trees to make a strong learner. Random forests are more accurate, more robust, and less prone to overfitting.

Q: Why is mean square error a bad measure of model performance? What would you suggest instead?

Mean Squared Error (MSE) gives a relatively high weight to large errors — therefore, MSE tends to put too much emphasis on large deviations. A more robust alternative is MAE (mean absolute deviation).

Q: What are the assumptions required for linear regression? What if some of these assumptions are violated?

The assumptions are as follows:

1. The sample data used to fit the model is **representative of the population**
2. The relationship between X and the mean of Y is **linear**
3. The variance of the residual is the same for any value of X (**homoscedasticity**)
4. Observations are **independent** of each other
5. For any value of X, Y is **normally distributed**.

Extreme violations of these assumptions will make the results redundant. Small violations of these assumptions will result in a greater bias or variance of the estimate.

Q: What is collinearity and what to do with it? How to remove multicollinearity?

Multicollinearity exists when an independent variable is highly correlated with another independent variable in a multiple regression equation. This can be problematic because it undermines the statistical significance of an independent variable.

You could use the Variance Inflation Factors (VIF) to determine if there is any multicollinearity between independent variables — a standard benchmark is that if the VIF is greater than 5 then multicollinearity exists.

Q: How to check if the regression model fits the data well?

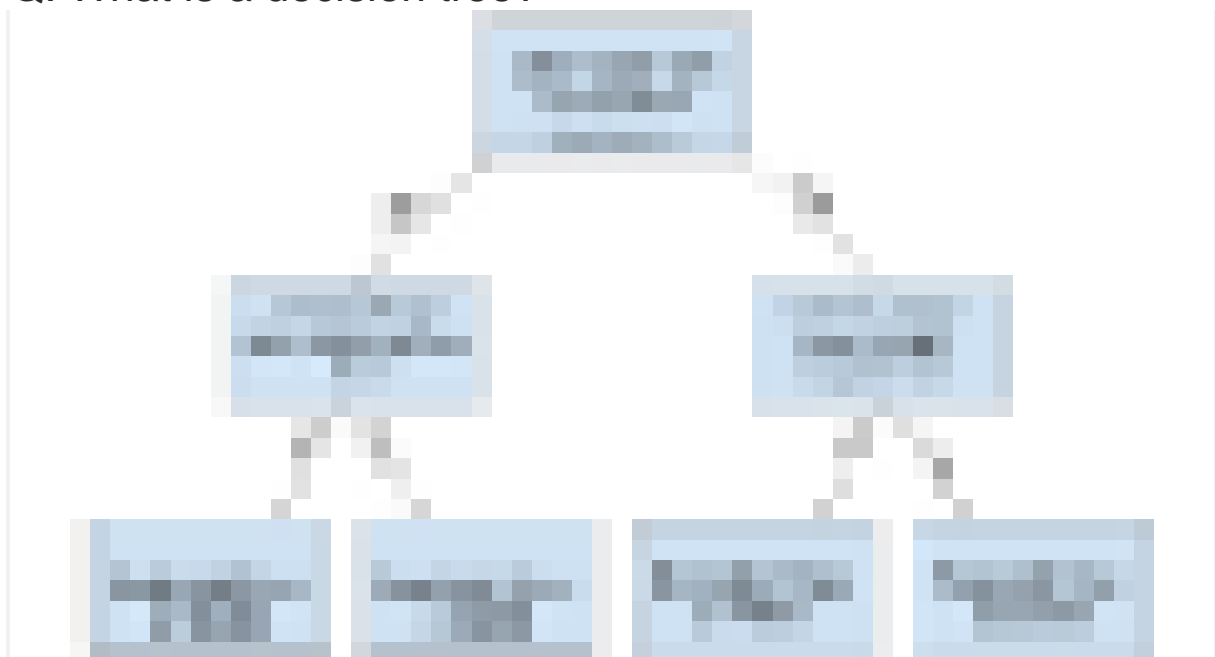
There are a couple of metrics that you can use:

**R-squared/Adjusted R-squared:** Relative measure of fit. *This was explained in a previous answer*

**F1 Score:** Evaluates the null hypothesis that all regression coefficients are equal to zero vs the alternative hypothesis that at least one doesn't equal zero

**RMSE:** Absolute measure of fit.

Q: What is a decision tree?



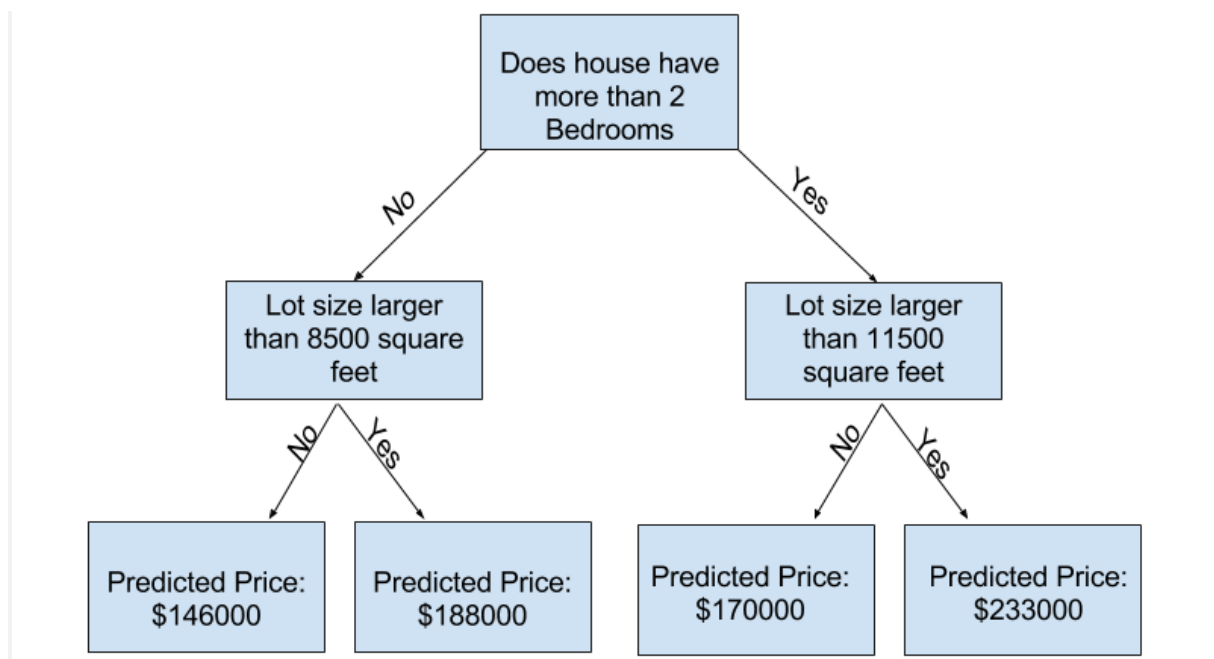
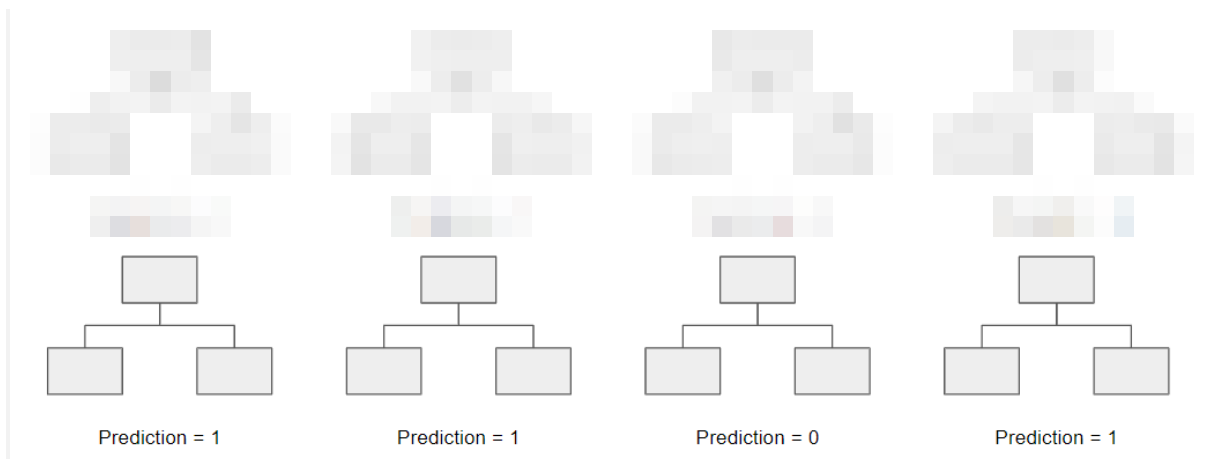


Image taken from Kaggle

**Decision trees** are a popular model, used in operations research, strategic planning, and machine learning. Each square above is called a **node**, and the more nodes you have, the more accurate your decision tree will be (generally). The last nodes of the decision tree, where a decision is made, are called the **leaves** of the tree. Decision trees are intuitive and easy to build but fall short when it comes to accuracy.

Q: What is a random forest? Why is it good?

Random forests are an [ensemble learning](#) technique that builds off of decision trees. Random forests involve creating multiple decision trees using [bootstrapped datasets](#) of the original data and randomly selecting a subset of variables at each step of the decision tree. The model then selects the mode of all of the predictions of each decision tree. By relying on a “majority wins” model, it reduces the risk of error from an individual tree.



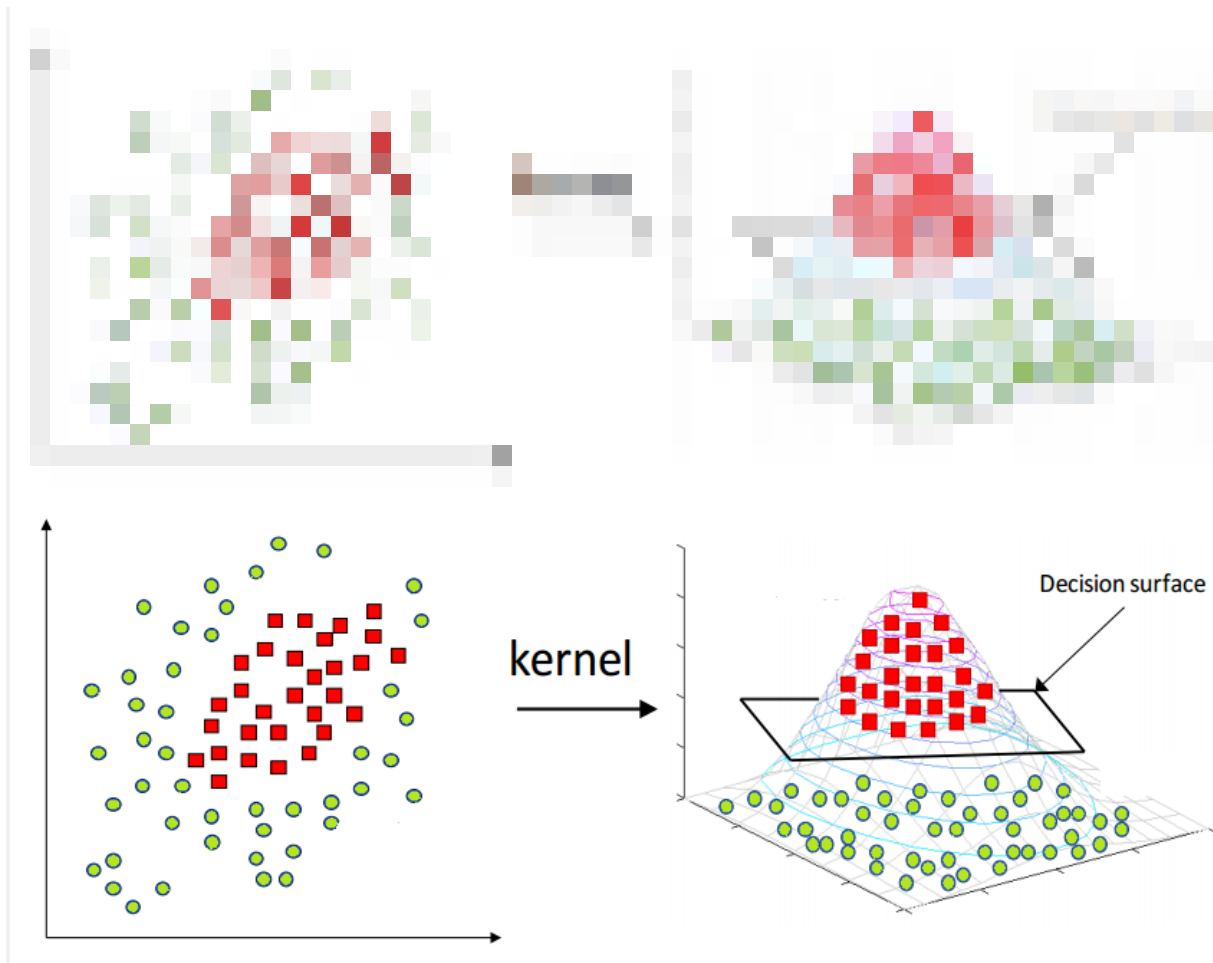
For example, if we created one decision tree, the third one, it would predict 0. But if we relied on the mode of all 4 decision trees, the predicted value would be 1. This is the power of random forests.

Random forests offer several other benefits including strong performance, can model non-linear boundaries, no cross-validation needed, and gives feature importance.

Q: What is a kernel? Explain the kernel trick

A kernel is a way of computing the dot product of two vectors  $\mathbf{x}$  and  $\mathbf{y}$  in some (possibly very high dimensional) feature space, which is why kernel functions are sometimes called “generalized dot product” [2]

The kernel trick is a method of using a linear classifier to solve a non-linear problem by transforming linearly inseparable data to linearly separable ones in a higher dimension.



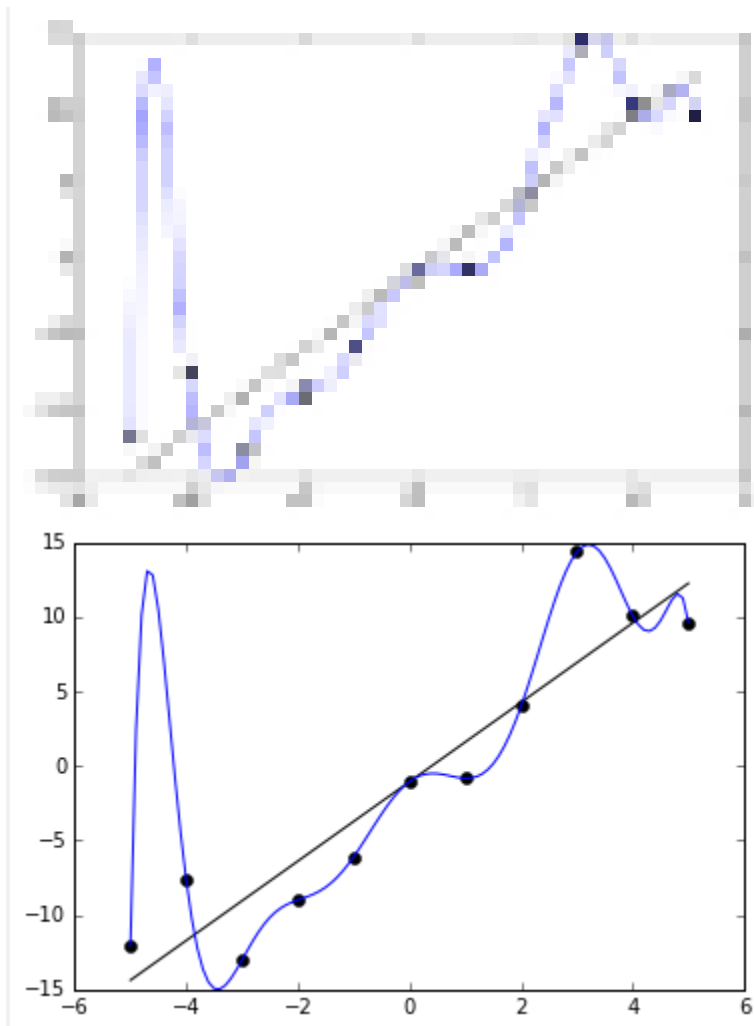
Taken from Analytics Vidhya

Q: Is it beneficial to perform dimensionality reduction before fitting an SVM? Why or why not?

When the number of features is greater than the number of observations, then performing dimensionality reduction will generally improve the SVM.

Q: What is overfitting?





Taken from Wikipedia

Overfitting is an error where the model ‘fits’ the data too well, resulting in a model with high variance and low bias. As a consequence, an overfit model will inaccurately predict new data points even though it has a high accuracy on the training data.

Q: What is boosting?

Boosting is an ensemble method to improve a model by reducing its bias and variance, ultimately converting weak learners to strong learners. The general idea is to train a weak learner and sequentially iterate and improve the model by

learning from the previous learner. *You can learn more about it [here](#).*

## Statistics, Probability, and Mathematics

Q: The probability that item an item at location A is 0.6, and 0.8 at location B. What is the probability that item would be found on Amazon website?

We need to make some assumptions about this question before we can answer it. **Let's assume that there are two possible places to purchase a particular item on Amazon and the probability of finding it at location A is 0.6 and B is 0.8. The probability of finding the item on Amazon can be explained as so:**

We can reword the above as  $P(A) = 0.6$  and  $P(B) = 0.8$ .

Furthermore, let's assume that these are independent events, meaning that the probability of one event is not impacted by the other. We can then use the formula...

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

$$P(A \text{ or } B) = 0.6 + 0.8 - (0.6 * 0.8)$$

$$P(A \text{ or } B) = 0.92$$

*Check out the Amazon data scientist interview guide [here](#).*

Q: You randomly draw a coin from 100 coins — 1 unfair coin (head-head), 99 fair coins (head-tail) and roll it 10 times. If the result is 10 heads, what is the probability that the coin is unfair?

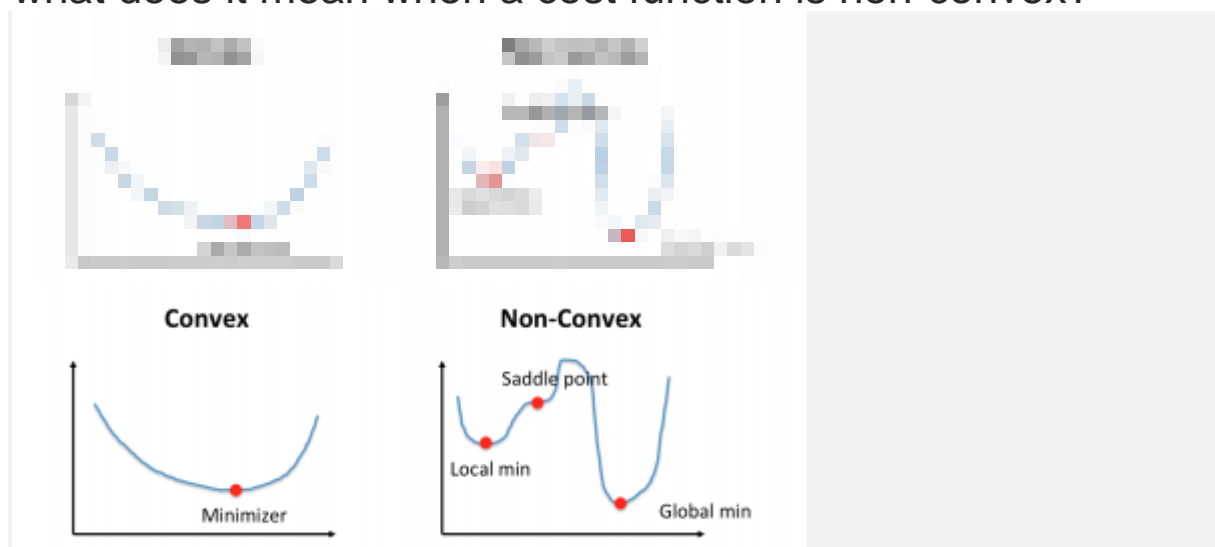
This can be answered using the Bayes Theorem. The extended equation for the Bayes Theorem is the following:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\neg A)P(\neg A)}$$

Assume that the probability of picking the unfair coin is denoted as  $P(A)$  and the probability of flipping 10 heads in a row is denoted as  $P(B)$ . Then  $P(B|A)$  is equal to 1,  $P(B|\neg A)$  is equal to  $0.5^{10}$ , and  $P(\neg A)$  is equal to 0.99.

If you fill in the equation, then  $P(A|B) = 0.9118$  or 91.18%.

Q: Difference between convex and non-convex cost function; what does it mean when a cost function is non-convex?



Taken from Cho-Jui Hsieh, UCLA

A **convex function** is one where a line drawn between any two points on the graph lies on or above the graph. It has one minimum.

A **non-convex function** is one where a line drawn between any two points on the graph may intersect other points on the graph. It is characterized as “wavy”.

When a cost function is non-convex, it means that there's a likelihood that the function may find local minima instead of the global minimum, which is typically undesired in machine learning models from an optimization perspective.

Q: Walk through the probability fundamentals

*For this, I'm going to look at the eight rules of probability laid out [here](#) and the four different counting methods (see more [here](#)).*

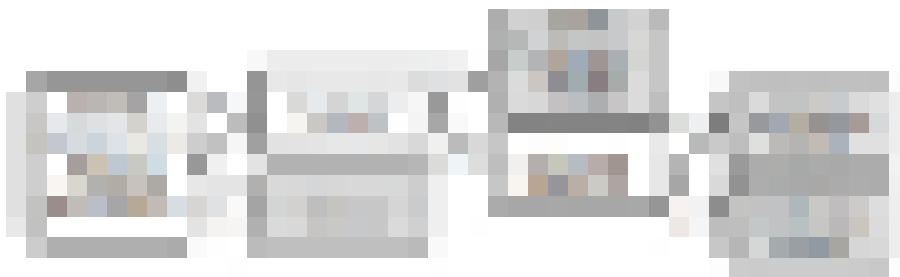
## **Eight rules of probability**

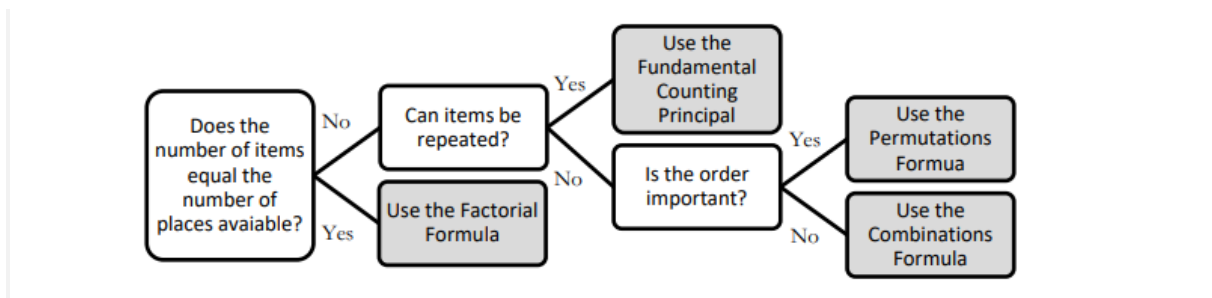
- Rule #1: For any event A,  $0 \leq P(A) \leq 1$ ; in other words, the probability of an event can range from 0 to 1.
- Rule #2: The sum of the probabilities of all possible outcomes always equals 1.
- Rule #3:  $P(\text{not } A) = 1 - P(A)$ ; This rule explains the relationship between the

*probability of an event and its complement event. A complement event is one that includes all possible outcomes that aren't in A.*

- Rule #4: If A and B are disjoint events (mutually exclusive), then  **$P(A \text{ or } B) = P(A) + P(B)$** ; *this is called the addition rule for disjoint events*
- Rule #5:  **$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$** ; *this is called the general addition rule.*
- Rule #6: If A and B are two independent events, then  **$P(A \text{ and } B) = P(A) * P(B)$** ; *this is called the multiplication rule for independent events.*
- Rule #7: The conditional probability of event B given event A is  **$P(B|A) = P(A \text{ and } B) / P(A)$**
- Rule #8: For any two events A and B,  **$P(A \text{ and } B) = P(A) * P(B|A)$** ; *this is called the general multiplication rule*

## Counting Methods





### **Factorial Formula: $n! = n \times (n - 1) \times (n - 2) \times \dots \times 2 \times 1$**

Use when the number of items is equal to the number of places available.

*Eg. Find the total number of ways 5 people can sit in 5 empty seats.*

$$= 5 \times 4 \times 3 \times 2 \times 1 = 120$$

### **Fundamental Counting Principle (multiplication)**

This method should be used when repetitions are allowed and the number of ways to fill an open place is not affected by previous fills.

*Eg. There are 3 types of breakfasts, 4 types of lunches, and 5 types of desserts. The total number of combinations is  $= 5 \times 4 \times 3 = 60$*

### **Permutations: $P(n,r) = n! / (n-r)!$**

This method is used when replacements are not allowed and order of item ranking matters.

*Eg. A code has 4 digits in a particular order and the digits range from 0 to 9. How many permutations are there if one digit can only be used once?*

$$P(n,r) = 10! / (10-4)! = (10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1) / (6 \times 5 \times 4 \times 3 \times 2 \times 1) = 5040$$

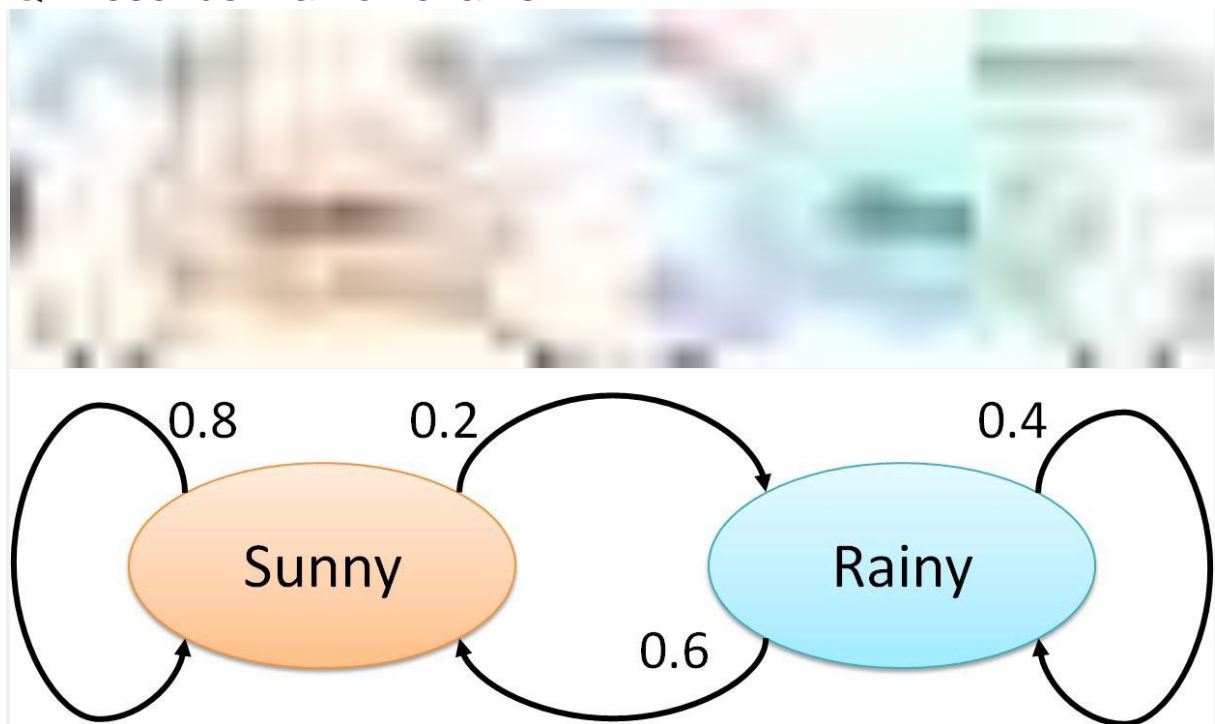
### Combinations Formula: $C(n,r) = \frac{n!}{(n-r)!r!}$

This is used when replacements are not allowed and the order in which items are ranked does not matter.

*Eg. To win the lottery, you must select the 5 correct numbers in any order from 1 to 52. What is the number of possible combinations?*

$$C(n,r) = 52! / (52-5)!5! = 2,598,960$$

Q: Describe Markov chains?



Brilliant provides a great definition of Markov chains ([here](#)):

“A Markov chain is a mathematical system that experiences transitions from one state to another according to certain [probabilistic](#) rules. The defining characteristic of a Markov chain is that no matter how the [process](#) arrived at its present state, the possible future states are fixed. In other

words, the probability of transitioning to any particular state is dependent solely on the current state and time elapsed.”

The actual math behind Markov chains requires knowledge on linear algebra and matrices, so I’ll leave some links below in case you want to explore this topic further on your own.

See more [here](#) or [here](#).

Q: A box has 12 red cards and 12 black cards. Another box has 24 red cards and 24 black cards. You want to draw two cards at random from one of the two boxes, one card at a time. Which box has a higher probability of getting cards of the same color and why?

The box with 24 red cards and 24 black cards has a higher probability of getting two cards of the same color. Let’s walk through each step.

Let’s say the first card you draw from each deck is a red Ace.

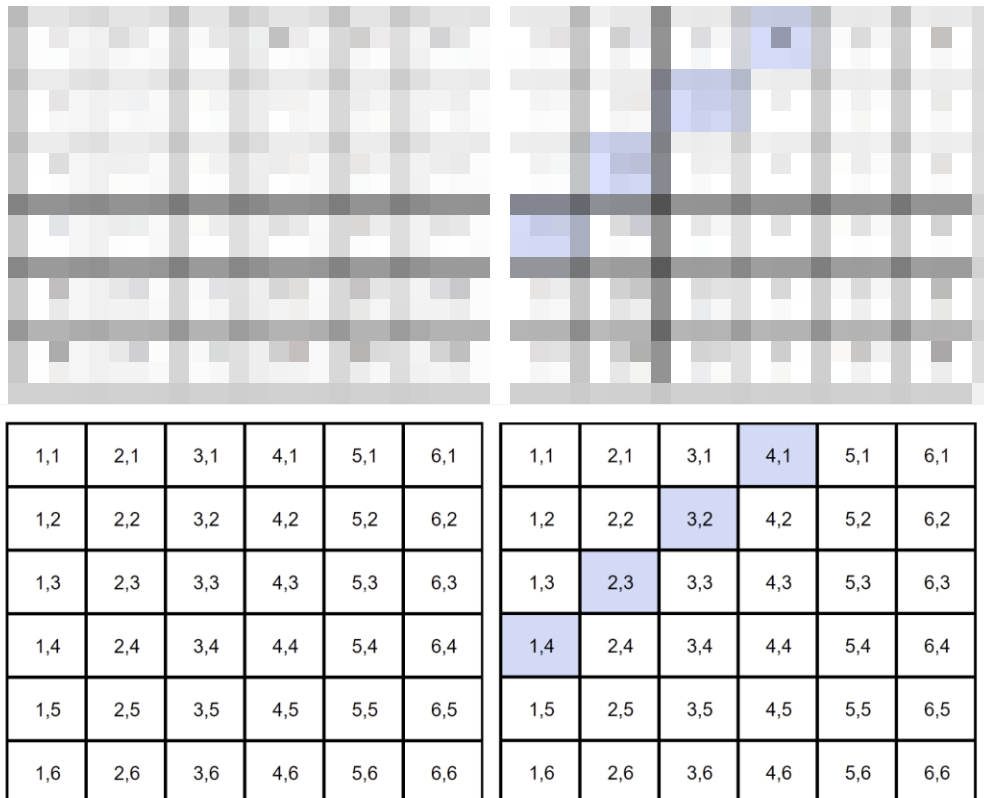
This means that in the deck with 12 reds and 12 blacks, there’s now 11 reds and 12 blacks. Therefore your odds of drawing another red are equal to  $11/(11+12)$  or  $11/23$ .

In the deck with 24 reds and 24 blacks, there would then be 23 reds and 24 blacks. Therefore your odds of drawing another red are equal to  $23/(23+24)$  or  $23/47$ .



Since  $23/47 > 11/23$ , the second deck with more cards has a higher probability of getting the same two cards.

Q: You are at a Casino and have two dice to play with. You win \$10 every time you roll a 5. If you play till you win and then stop, what is the expected payout?



- Let's assume that it costs \$5 every time you want to play.
- There are 36 possible combinations with two dice.
- Of the 36 combinations, there are 4 combinations that result in rolling a five (*see blue*). This means that there is a  $4/36$  or  $1/9$  chance of rolling a 5.

- A  $1/9$  chance of winning means you'll lose eight times and win once (theoretically).
- Therefore, your expected payout is equal to  $\$10.00 * 1 - \$5.00 * 9 = -\$35.00$ .

*Edit: Thank you guys for commenting and pointing out that it should be -\$35!*

Q: How can you tell if a given coin is biased?

This isn't a trick question. The answer is simply to perform a hypothesis test:

1. The null hypothesis is that the coin is not biased and the probability of flipping heads should equal 50% ( $p=0.5$ ). The alternative hypothesis is that the coin is biased and  $p \neq 0.5$ .
2. Flip the coin 500 times.
3. Calculate Z-score (if the sample is less than 30, you would calculate the t-statistics).
4. Compare against alpha (two-tailed test so  $0.05/2 = 0.025$ ).
5. If  $p\text{-value} > \alpha$ , the null is not rejected and the coin is not biased.  
If  $p\text{-value} < \alpha$ , the null is rejected and the coin is biased.

Learn more about hypothesis testing [here](#).

Q: Make an unfair coin fair

Since a coin flip is a binary outcome, you can make an unfair coin fair by flipping it twice. If you flip it twice, there are two outcomes that you can bet on: heads followed by tails or tails followed by heads.

$$P(\text{heads}) * P(\text{tails}) = P(\text{tails}) * P(\text{heads})$$

This makes sense since each coin toss is an **independent event**. This means that if you get heads → heads or tails → tails, you would need to reflip the coin.

Q: You are about to get on a plane to London, you want to know whether you have to bring an umbrella or not. You call three of your random friends and ask each one of them if it's raining. The probability that your friend is telling the truth is 2/3 and the probability that they are playing a prank on you by lying is 1/3. If all 3 of them tell that it is raining, then what is the probability that it is actually raining in London.

You can tell that this question is related to Bayesian theory because of the last statement which essentially follows the structure, "What is the probability A is true **given** B is true?" Therefore we need to know the probability of it raining in London on a given day. Let's assume it's 25%.

$P(A)$  = probability of it raining = 25%

$P(B)$  = probability of all 3 friends say that it's raining

$P(A|B)$  probability that it's raining given they're telling that it is raining

$P(B|A)$  probability that all 3 friends say that it's raining given it's raining  $= (2/3)^3 = 8/27$

*Step 1: Solve for  $P(B)$*

$P(A|B) = P(B|A) * P(A) / P(B)$ , can be rewritten as

$P(B) = P(B|A) * P(A) + P(B|\text{not } A) * P(\text{not } A)$

$P(B) = (2/3)^3 * 0.25 + (1/3)^3 * 0.75 = 0.25 * 8/27 + 0.75 * 1/27$

*Step 2: Solve for  $P(A|B)$*

$P(A|B) = 0.25 * (8/27) / (0.25 * 8/27 + 0.75 * 1/27)$

$P(A|B) = 8 / (8 + 3) = 8/11$

Therefore, if all three friends say that it's raining, then there's an 8/11 chance that it's actually raining.

Q: You are given 40 cards with four different colors- 10 Green cards, 10 Red Cards, 10 Blue cards, and 10 Yellow cards. The cards of each color are numbered from one to ten. Two cards are picked at random. Find out the probability that the cards picked are not of the same number and same color.

Since these events are not independent, we can use the rule:

$P(A \text{ and } B) = P(A) * P(B|A)$ , which is also equal to

$P(\text{not } A \text{ and not } B) = P(\text{not } A) * P(\text{not } B | \text{not } A)$

For example:

$$P(\text{not 4 and not yellow}) = P(\text{not 4}) * P(\text{not yellow} | \text{not 4})$$

$$P(\text{not 4 and not yellow}) = (36/39) * (27/36)$$

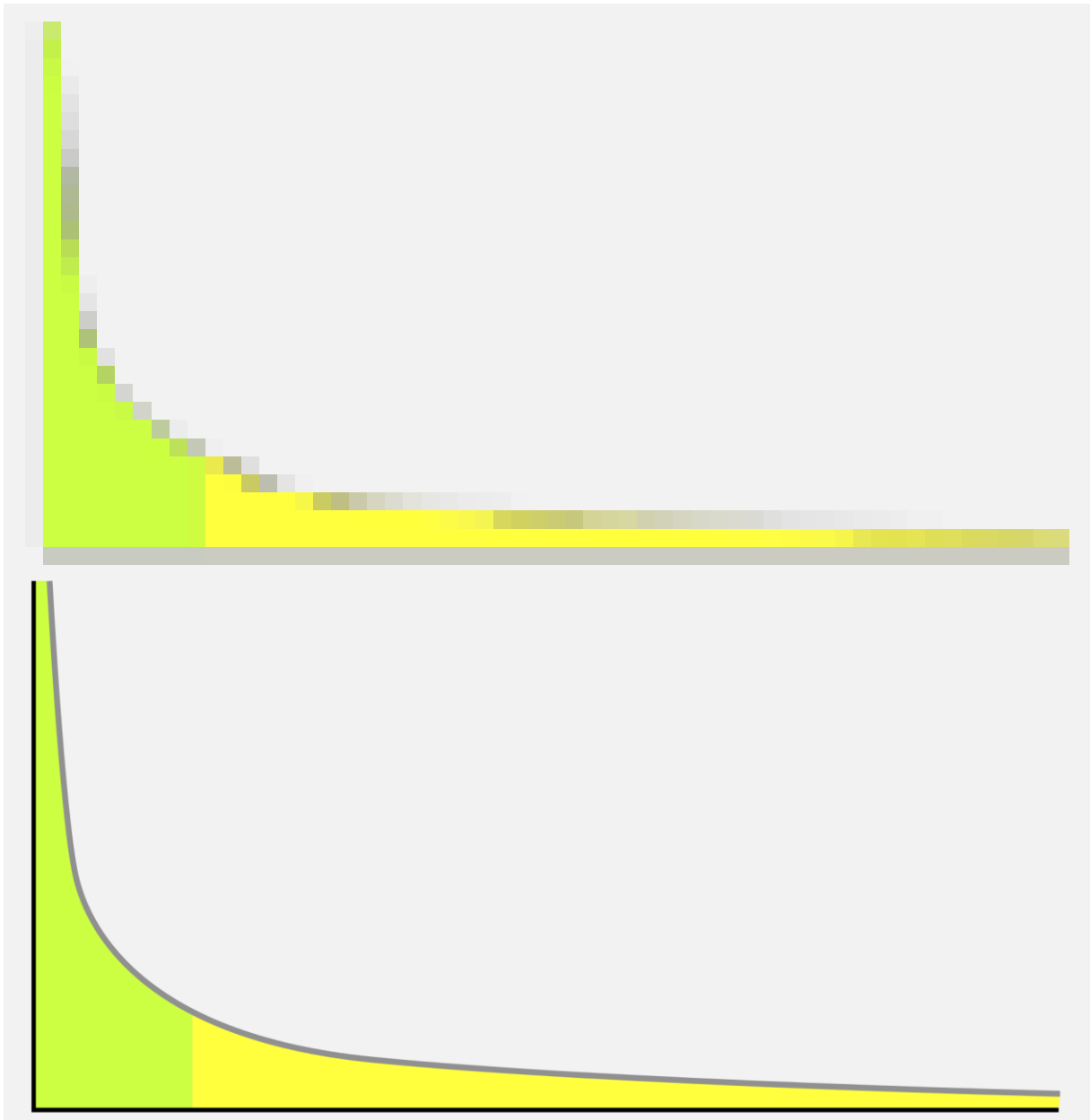
$$P(\text{not 4 and not yellow}) = 0.692$$

Therefore, the probability that the cards picked are not the same number and the same color is 69.2%.

Q: How do you assess the statistical significance of an insight?

You would perform hypothesis testing to determine statistical significance. First, you would state the null hypothesis and alternative hypothesis. Second, you would calculate the p-value, the probability of obtaining the observed results of a test assuming that the null hypothesis is true. Last, you would set the level of the significance (alpha) and if the p-value is less than the alpha, you would reject the null — in other words, the result is statistically significant.

Q: Explain what a long-tailed distribution is and provide three examples of relevant phenomena that have long tails. Why are they important in classification and regression problems?



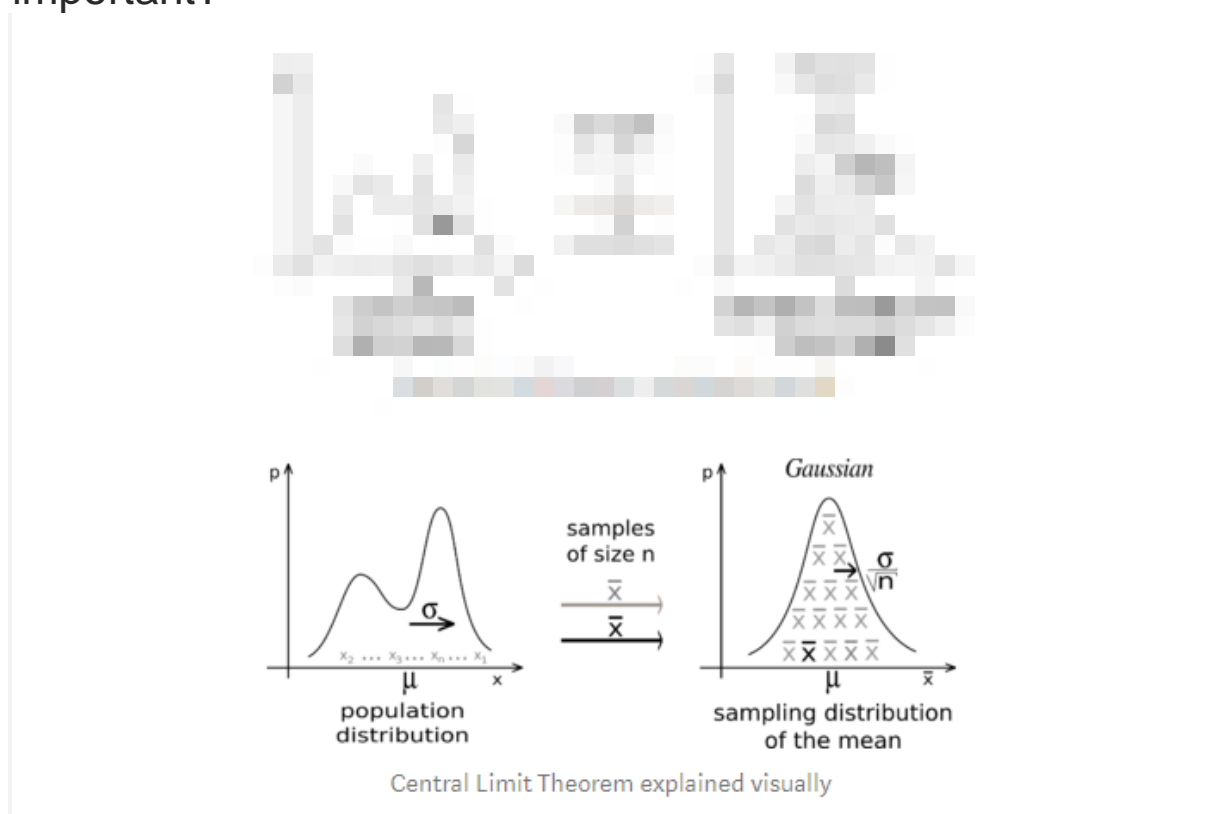
Example of a long tail distribution

A **long-tailed distribution** is a type of heavy-tailed distribution that has a tail (or tails) that drop off gradually and asymptotically.

3 practical examples include the power law, the Pareto principle (more commonly known as the 80–20 rule), and product sales (i.e. best selling products vs others).

It's important to be mindful of long-tailed distributions in classification and regression problems because the least frequently occurring values make up the majority of the population. This can ultimately change the way that you deal with outliers, and it also conflicts with some machine learning techniques with the assumption that the data is normally distributed.

Q: What is the Central Limit Theorem? Explain it. Why is it important?



From Wikipedia

Statistics How To provides the best definition of CLT, which is:

“The central limit theorem states that the sampling distribution of the sample mean approaches a normal distribution as the

sample size gets larger no matter what the shape of the population distribution.” [1]

The central limit theorem is important because it is used in hypothesis testing and also to calculate confidence intervals.

Q: What is the statistical power?

‘Statistical power’ refers to the power of a binary hypothesis, which is the probability that the test rejects the null hypothesis given that the alternative hypothesis is true. [2]



$$\text{Power} = P(\text{reject Null} \mid \text{alternative is true})$$

Q: Explain selection bias (with regard to a dataset, not variable selection). Why is it important? How can data management procedures such as missing data handling make it worse?

**Selection bias** is the phenomenon of selecting individuals, groups or data for analysis in such a way that proper randomization is not achieved, ultimately resulting in a sample that is not representative of the population.

Understanding and identifying selection bias is important because it can significantly skew results and provide false insights about a particular population group.



Types of selection bias include:

- **sampling bias:** a biased sample caused by non-random sampling
- **time interval:** selecting a specific time frame that supports the desired conclusion. e.g. conducting a sales analysis near Christmas.
- **exposure:** includes clinical susceptibility bias, protopathic bias, indication bias. *Read more [here](#).*
- **data:** includes cherry-picking, suppressing evidence, and the fallacy of incomplete evidence.
- **attrition:** attrition bias is similar to survivorship bias, where only those that ‘survived’ a long process are included in an analysis, or failure bias, where those that ‘failed’ are only included
- **observer selection:** related to the Anthropic principle, which is a philosophical consideration that any data we collect about the universe is filtered by the fact that, in order for it to be observable, it must be compatible with the conscious and sapient life that observes it. [3]

Handling missing data can make selection bias worse because different methods impact the data in different ways. For

example, if you replace null values with the mean of the data, you adding bias in the sense that you're assuming that the data is not as spread out as it might actually be.

Q: Provide a simple example of how an experimental design can help answer a question about behavior. How does experimental data contrast with observational data?

**Observational data** comes from observational studies which are when you observe certain variables and try to determine if there is any correlation.

**Experimental data** comes from experimental studies which are when you control certain variables and hold them constant to determine if there is any causality.

An example of experimental design is the following: split a group up into two. The control group lives their lives normally. The test group is told to drink a glass of wine every night for 30 days. Then research can be conducted to see how wine affects sleep.

Q: Is mean imputation of missing data acceptable practice? Why or why not?

**Mean imputation** is the practice of replacing null values in a data set with the mean of the data.

Mean imputation is generally bad practice because it doesn't take into account feature correlation. For example, imagine we have a table showing age and fitness score and imagine that an

eighty-year-old has a missing fitness score. If we took the average fitness score from an age range of 15 to 80, then the eighty-year-old will appear to have a much higher fitness score than he actually should.

Second, mean imputation reduces the variance of the data and increases bias in our data. This leads to a less accurate model and a narrower confidence interval due to a smaller variance.

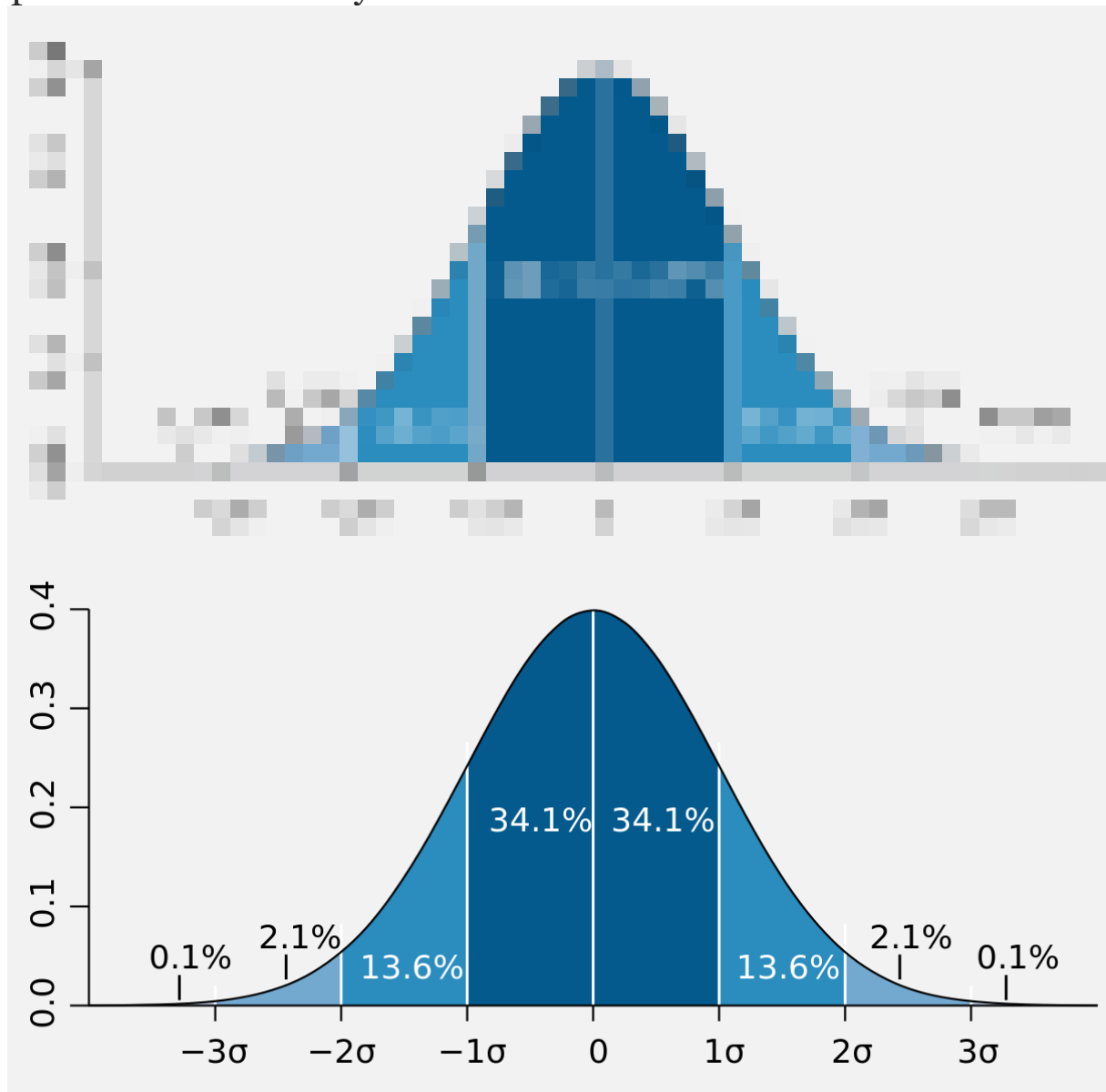
Q: What is an outlier? Explain how you might screen for outliers and what you would do if you found them in your dataset. Also, explain what an inlier is and how you might screen for them and what you would do if you found them in your dataset.

An **outlier** is a data point that differs significantly from other observations.

Depending on the cause of the outlier, they can be bad from a machine learning perspective because they can worsen the accuracy of a model. If the outlier is caused by a measurement error, it's important to remove them from the dataset. There are a couple of ways to identify outliers:

**Z-score/standard deviations:** if we know that 99.7% of data in a data set lie within three standard deviations, then we can calculate the size of one standard deviation, multiply it by 3, and identify the data points that are outside of this range. Likewise, we can calculate the z-score of a given point, and if it's equal to  $\pm 3$ , then it's an outlier.

Note: that there are a few contingencies that need to be considered when using this method; the data must be normally distributed, this is [not applicable for small data sets](#), and the presence of too many outliers can throw off z-score.



**Interquartile Range (IQR):** IQR, the concept used to build boxplots, can also be used to identify outliers. The IQR is equal to the difference between the 3rd quartile and the 1st quartile. You can then identify if a point is an outlier if it is less than  $Q1 -$

$1.5 \times \text{IQR}$  or greater than  $Q3 + 1.5 \times \text{IQR}$ . This comes to approximately 2.698 standard deviations.

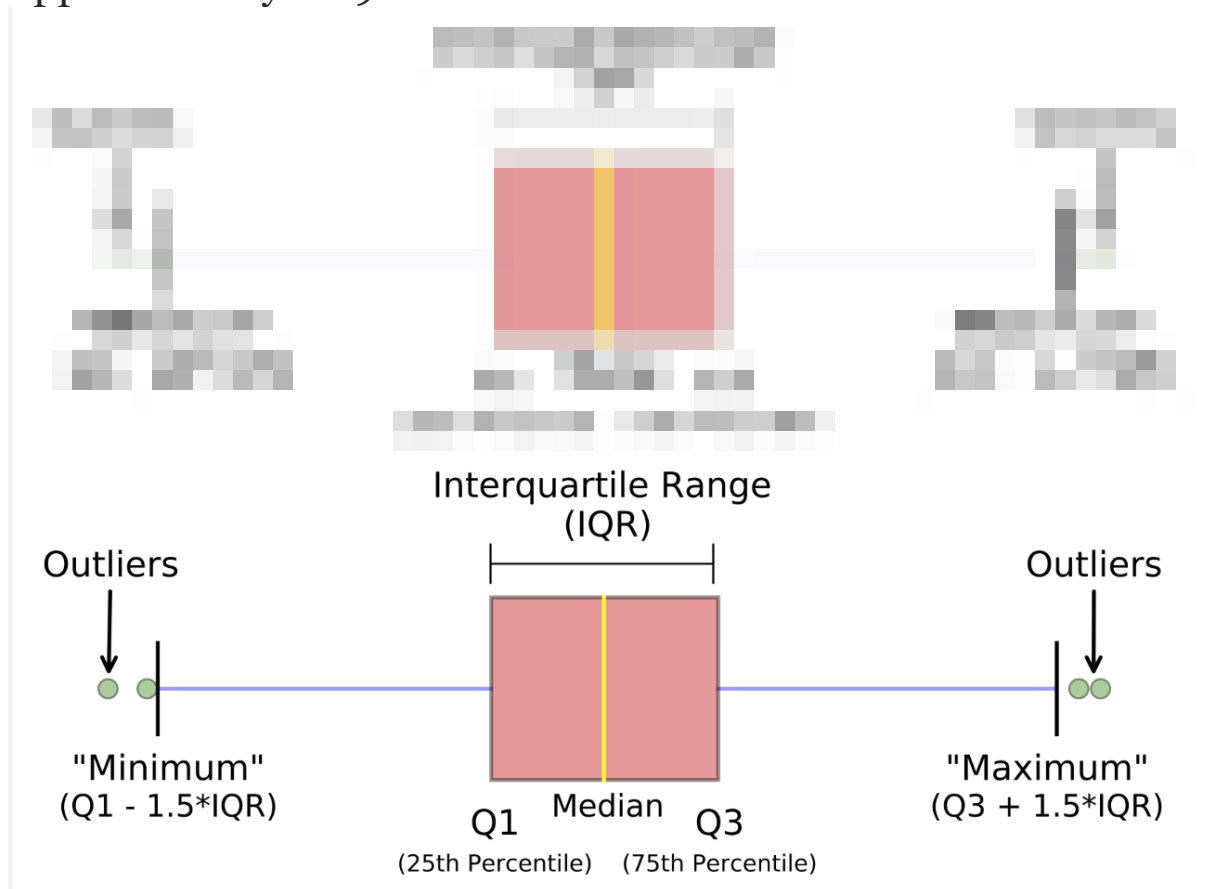


Photo from Michael Galarnyk

Other methods include DBScan clustering, Isolation Forests, and Robust Random Cut Forests.

An **inlier** is a data observation that lies within the rest of the dataset and is unusual or an error. Since it lies in the dataset, it is typically harder to identify than an outlier and requires external data to identify them. Should you identify any inliers, you can simply remove them from the dataset to address them.

Q: How do you handle missing data? What imputation techniques do you recommend?

There are several ways to handle missing data:

- Delete rows with missing data
- Mean/Median/Mode imputation
- Assigning a unique value
- Predicting the missing values
- Using an algorithm which supports missing values, like random forests

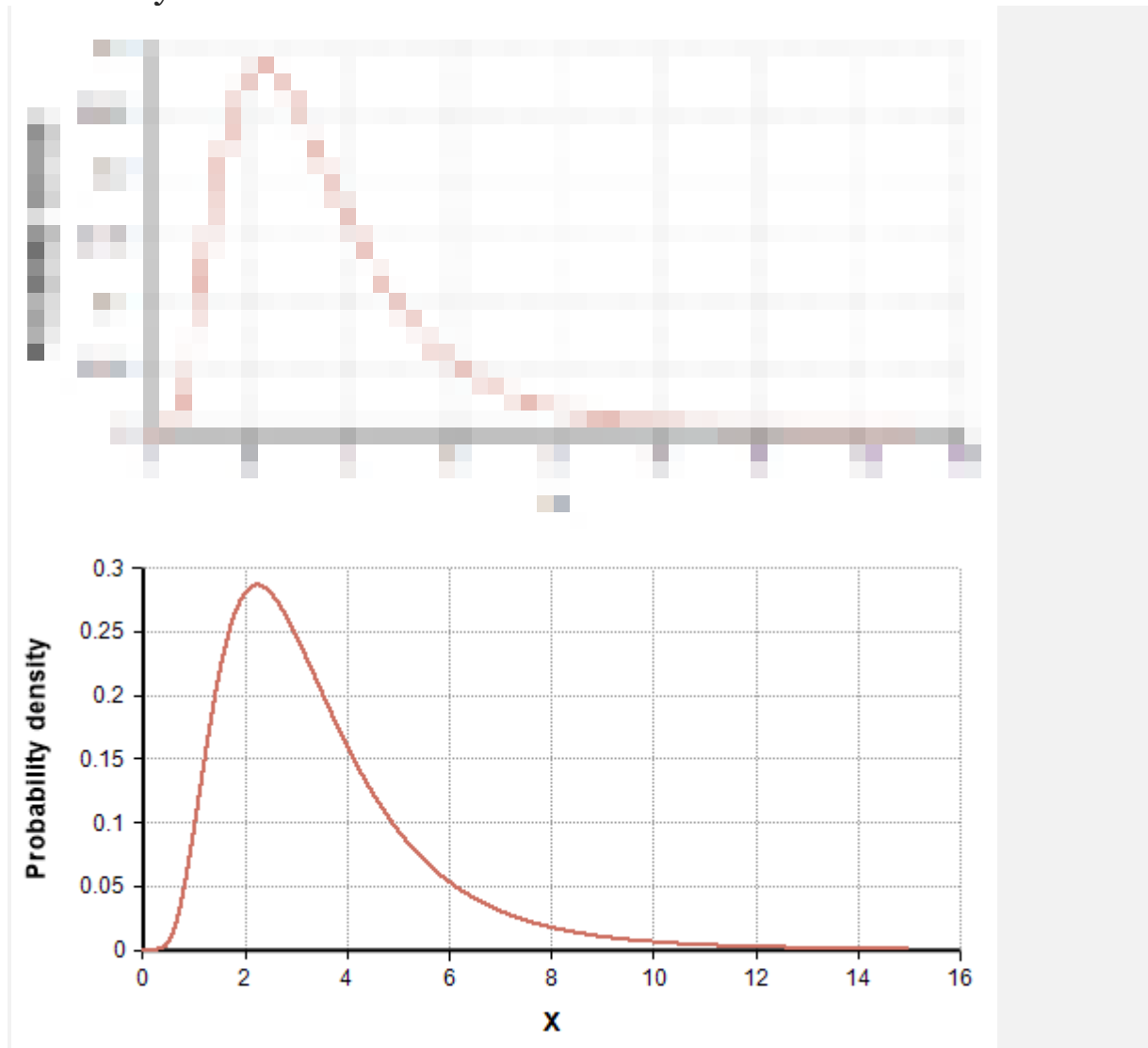
The best method is to delete rows with missing data as it ensures that no bias or variance is added or removed, and ultimately results in a robust and accurate model. However, this is only recommended if there's a lot of data to start with and the percentage of missing values is low.

Q: You have data on the duration of calls to a call center. Generate a plan for how you would code and analyze these data. Explain a plausible scenario for what the distribution of these durations might look like. How could you test, even graphically, whether your expectations are borne out?

First I would conduct EDA — Exploratory Data Analysis to clean, explore, and understand my data. *See my article on EDA [here](#).* As part of my EDA, I could compose a histogram of the duration of calls to see the underlying distribution.

My guess is that the duration of calls would follow a lognormal distribution (see below). The reason that I believe it's positively

skewed is because the lower end is limited to 0 since a call can't be negative seconds. However, on the upper end, it's likely for there to be a small proportion of calls that are extremely long relatively.



Lognormal Distribution Example

You could use a QQ plot to confirm whether the duration of calls follows a lognormal distribution or not. See [here](#) to learn more about QQ plots.

Q: Explain likely differences between administrative datasets and datasets gathered from experimental studies. What are

likely problems encountered with administrative data? How do experimental methods help alleviate these problems? What problem do they bring?

Administrative datasets are typically datasets used by governments or other organizations for non-statistical reasons.

Administrative datasets are usually larger and more cost-efficient than experimental studies. They are also regularly updated assuming that the organization associated with the administrative dataset is active and functioning. At the same time, administrative datasets may not capture all of the data that one may want and may not be in the desired format either. It is also prone to quality issues and missing entries.

Q: You are compiling a report for user content uploaded every month and notice a spike in uploads in October. In particular, a spike in picture uploads. What might you think is the cause of this, and how would you test it?

There are a number of potential reasons for a spike in photo uploads:

1. A new feature may have been implemented in October which involves uploading photos and gained a lot of traction by users. For example, a feature that gives the ability to create photo albums.



2. Similarly, it's possible that the process of uploading photos before was not intuitive and was improved in the month of October.
3. There may have been a viral social media movement that involved uploading photos that lasted for all of October. Eg. Movember but something more scalable.
4. It's possible that the spike is due to people posting pictures of themselves in costumes for Halloween.

The method of testing depends on the cause of the spike, but you would conduct hypothesis testing to determine if the inferred cause is the actual cause.

Q: Give examples of data that does not have a Gaussian distribution, nor log-normal.

- Any type of categorical data won't have a gaussian distribution or lognormal distribution.
- Exponential distributions — eg. the amount of time that a car battery lasts or the amount of time until an earthquake occurs.

Q: What is root cause analysis? How to identify a cause vs. a correlation? Give examples

**Root cause analysis:** a method of problem-solving used for identifying the root cause(s) of a problem [5]

**Correlation** measures the relationship between two variables, range from -1 to 1. **Causation** is when a first event appears to have caused a second event. Causation essentially looks at direct relationships while correlation can look at both direct and indirect relationships.

Example: a higher crime rate is associated with higher sales in ice cream in Canada, aka they are positively correlated. However, this doesn't mean that one causes another. Instead, it's because both occur more when it's warmer outside.

You can test for causation using hypothesis testing or A/B testing.

Q: Give an example where the median is a better measure than the mean

When there are a number of outliers that positively or negatively skew the data.

Q: Given two fair dices, what is the probability of getting scores that sum to 4? to 8?

There are 4 combinations of rolling a 4 (1+3, 3+1, 2+2):

$$P(\text{rolling a 4}) = 3/36 = 1/12$$

There are combinations of rolling an 8 (2+6, 6+2, 3+5, 5+3, 4+4):


$$P(\text{rolling an 8}) = 5/36$$

Q: What is the Law of Large Numbers?

The Law of Large Numbers is a theory that states that as the number of trials increases, the average of the result will become closer to the expected value.

Eg. flipping heads from fair coin 100,000 times should be closer to 0.5 than 100 times.

Q. How do you calculate the needed sample size?


$$ME = t * \frac{S}{\sqrt{n}} \quad - \text{ or } - \quad ME = z * \frac{\sigma}{\sqrt{n}}$$

Formula for margin of error

You can use the margin of error (ME) formula to determine the desired sample size.

- $t/z$  =  $t/z$  score used to calculate the confidence interval
- ME = the desired margin of error
- S = sample standard deviation

Q: When you sample, what bias are you inflicting?

Potential biases include the following:

- **Sampling bias:** a biased sample caused by non-random sampling

- **Under coverage bias:** sampling too few observations
- **Survivorship bias:** error of overlooking observations that did not make it past a form of selection process.

Q: How do you control for biases?

There are many things that you can do to control and minimize bias. Two common things include **randomization**, where participants are assigned by chance, and **random sampling**, sampling in which each member has an equal probability of being chosen.

Q: What are confounding variables?

A confounding variable, or a confounder, is a variable that influences both the dependent variable and the independent variable, causing a spurious association, a mathematical relationship in which two or more variables are associated but not causally related.

Q: What is A/B testing?

A/B testing is a form of hypothesis testing and two-sample hypothesis testing to compare two versions, the control and variant, of a single variable. It is commonly used to improve and optimize user experience and marketing.

[Check out my article, A Simple Guide to A/B Testing for Data Science.](#)

Q: How do you prove that males are on average taller than females by knowing just gender height?

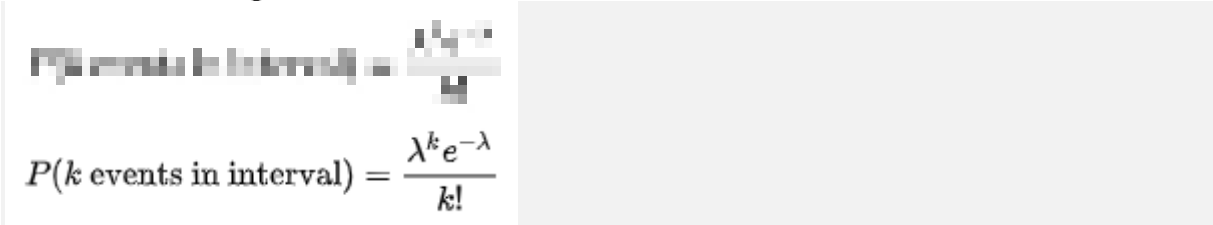
You can use hypothesis testing to prove that males are taller on average than females.

The null hypothesis would state that males and females are the same height on average, while the alternative hypothesis would state that the average height of males is greater than the average height of females.

Then you would collect a random sample of heights of males and females and use a t-test to determine if you reject the null or not.

Q: Infection rates at a hospital above a 1 infection per 100 person-days at risk are considered high. A hospital had 10 infections over the last 1787 person-days at risk. Give the p-value of the correct one-sided test of whether the hospital is below the standard.

Since we looking at the number of events (# of infections) occurring within a given timeframe, this is a Poisson distribution question.


$$P(k \text{ events in interval}) = \frac{\lambda^k e^{-\lambda}}{k!}$$

The probability of observing k events in an interval

Null (H<sub>0</sub>): 1 infection per person-days

Alternative (H<sub>1</sub>): >1 infection per person-days

k (actual) = 10 infections

lambda (theoretical) = (1/100)\*1787

p = 0.032372 or 3.2372% *calculated using .poisson() in excel or ppois in R*

Since p-value < alpha (assuming 5% level of significance), we reject the null and conclude that the hospital is below the standard.

Q: You roll a biased coin (p(head)=0.8) five times. What's the probability of getting three or more heads?

Use the General Binomial Probability formula to answer this question:



$$P(k \text{ out of } n) = \frac{n!}{k! (n - k)!} * p^k (1 - p)^{(n-k)}$$

General Binomial Probability Formula

p = 0.8

n = 5

k = 3,4,5

P(3 or more heads) = P(3 heads) + P(4 heads) + P(5 heads)  
= **0.94 or 94%**

Q: A random variable X is normal with mean 1020 and a standard deviation 50. Calculate  $P(X > 1200)$

Using Excel...

`p = 1 - norm.dist(1200, 1020, 50, true)`

**p = 0.000159**

Q: Consider the number of people that show up at a bus station is Poisson with mean 2.5/h. What is the probability that at most three people show up in a four hour period?

$x = 3$

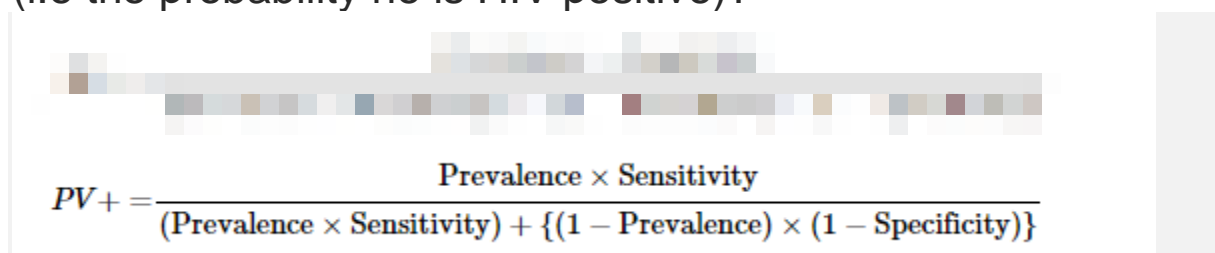
$\text{mean} = 2.5 \times 4 = 10$

using Excel...

`p = poisson.dist(3, 10, true)`

**p = 0.010336**

Q: An HIV test has a sensitivity of 99.7% and a specificity of 98.5%. A subject from a population of prevalence 0.1% receives a positive test result. What is the precision of the test (i.e the probability he is HIV positive)?


$$PV+ = \frac{\text{Prevalence} \times \text{Sensitivity}}{(\text{Prevalence} \times \text{Sensitivity}) + \{(1 - \text{Prevalence}) \times (1 - \text{Specificity})\}}$$

Equation for Precision (PV)

Precision = Positive Predictive Value = PV

$$PV = (0.001 * 0.997) / [(0.001 * 0.997) + ((1 - 0.001) * (1 - 0.985))]$$

$$PV = 0.0624 \text{ or } 6.24\%$$

See more about this equation [here](#).

Q: You are running for office and your pollster polled hundred people. Sixty of them claimed they will vote for you. Can you relax?

- Assume that there's only you and one other opponent.
- Also, assume that we want a 95% confidence interval. This gives us a z-score of 1.96.

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

Confidence interval formula

$$\hat{p} = 60/100 = 0.6$$

$$z^* = 1.96$$

$$n = 100$$

This gives us a confidence interval of [50.4, 69.6]. Therefore, given a confidence interval of 95%, if you are okay with the worst scenario of tying then you can relax. Otherwise, you cannot relax until you got 61 out of 100 to claim yes.



Q: Geiger counter records 100 radioactive decays in 5 minutes. Find an approximate 95% interval for the number of decays per hour.

- Since this is a Poisson distribution question, mean =  $\lambda$  = variance, which also means that standard deviation = square root of the mean
- a 95% confidence interval implies a z score of 1.96
- one standard deviation = 10

Therefore the confidence interval =  $100 \pm 19.6 = [964.8, 1435.2]$

Q: The homicide rate in Scotland fell last year to 99 from 115 the year before. Is this reported change really noteworthy?

- Since this is a Poisson distribution question, mean =  $\lambda$  = variance, which also means that standard deviation = square root of the mean
- a 95% confidence interval implies a z score of 1.96
- one standard deviation =  $\sqrt{115} = 10.724$

Therefore the confidence interval =  $115 \pm 21.45 = [93.55, 136.45]$ . Since 99 is within this confidence interval, we can assume that this change is not very noteworthy.

Q: Consider influenza epidemics for two-parent heterosexual families. Suppose that the probability is 17% that at least one of the parents has contracted the disease. The probability that the father has contracted influenza is 12% while the probability that both the mother and father have contracted the disease is 6%. What is the probability that the mother has contracted influenza?

Using the General Addition Rule in probability:

$$P(\text{mother or father}) = P(\text{mother}) + P(\text{father}) - P(\text{mother and father})$$
$$P(\text{mother}) = P(\text{mother or father}) + P(\text{mother and father}) - P(\text{father})$$
$$P(\text{mother}) = 0.17 + 0.06 - 0.12$$
$$P(\text{mother}) = 0.11$$

Q: Suppose that diastolic blood pressures (DBPs) for men aged 35–44 are normally distributed with a mean of 80 (mm Hg) and a standard deviation of 10. About what is the probability that a random 35–44 year old has a DBP less than 70?

Since 70 is one standard deviation below the mean, take the area of the Gaussian distribution to the left of one standard deviation.

$$= 2.3 + 13.6 = 15.9\%$$

Q: In a population of interest, a sample of 9 men yielded a sample average brain volume of 1,100cc and a standard deviation of 30cc. What is a 95% Student's T confidence interval for the mean brain volume in this new population?

$$\bar{X} \pm t \frac{s}{\sqrt{n}}$$

$$\bar{X} \pm t \frac{s}{\sqrt{n}}$$

Confidence interval for sample

Given a confidence level of 95% and degrees of freedom equal to 8, the t-score = 2.306

Confidence interval = 1100 +/- 2.306\*(30/3)

Confidence interval = [1076.94, 1123.06]

Q: A diet pill is given to 9 subjects over six weeks. The average difference in weight (follow up — baseline) is -2 pounds. What would the standard deviation of the difference in weight have to be for the upper endpoint of the 95% T confidence interval to touch 0?

Upper bound = mean + t-score\*(standard deviation/sqrt(sample size))

$$0 = -2 + 2.306*(s/3)$$

$$2 = 2.306 * s / 3$$

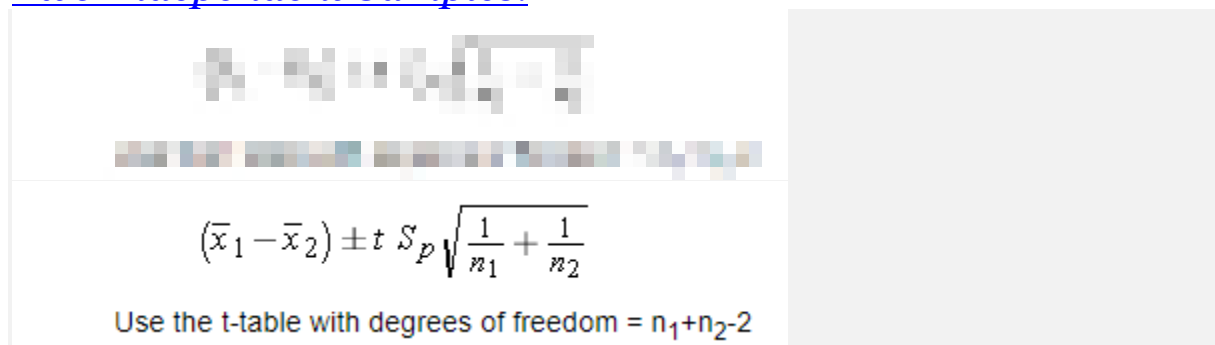
$$s = 2.601903$$

Therefore the standard deviation would have to be at least

approximately 2.60 for the upper bound of the 95% T confidence interval to touch 0.

Q: In a study of emergency room waiting times, investigators consider a new and the standard triage systems. To test the systems, administrators selected 20 nights and randomly assigned the new triage system to be used on 10 nights and the standard system on the remaining 10 nights. They calculated the nightly median waiting time (MWT) to see a physician. The average MWT for the new system was 3 hours with a variance of 0.60 while the average MWT for the old system was 5 hours with a variance of 0.68. Consider the 95% confidence interval estimate for the differences of the mean MWT associated with the new system. Assume a constant variance. What is the interval? Subtract in this order (New System — Old System).

[See here for full tutorial on finding the Confidence Interval for Two Independent Samples.](#)

A screenshot of a webpage showing a confidence interval formula and a note. The formula is  $(\bar{x}_1 - \bar{x}_2) \pm t S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ . Below the formula, it says "Use the t-table with degrees of freedom =  $n_1 + n_2 - 2$ ".

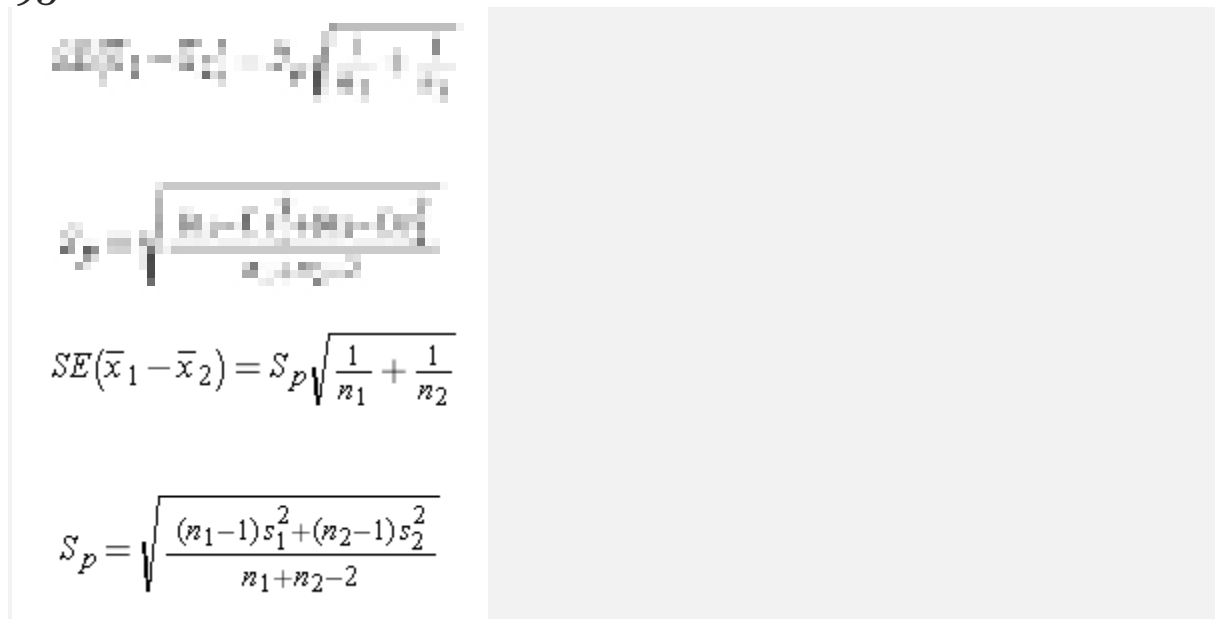
$$(\bar{x}_1 - \bar{x}_2) \pm t S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Use the t-table with degrees of freedom =  $n_1 + n_2 - 2$

Confidence Interval = mean +/- t-score \* standard error (see above)

mean = new mean — old mean = 3–5 = -2

t-score = 2.101 given df=18 (20-2) and confidence interval of 95%



The image shows handwritten mathematical formulas for calculating a confidence interval and standard error for two independent groups with unequal variances. The formulas are as follows:

$$SE(\bar{x}_1 - \bar{x}_2) = t_{\alpha/2} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$S_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$SE(\bar{x}_1 - \bar{x}_2) = S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$S_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

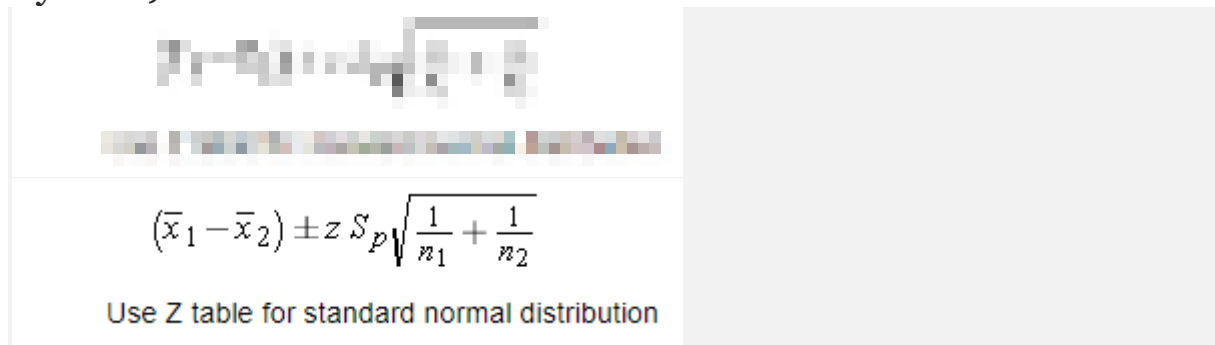
standard error =  $\sqrt{((0.62^2 * 9) + (0.68^2 * 9)) / (10 + 10 - 2)}$  \*  
 $\sqrt{1/10 + 1/10}$   
 standard error = 0.352

confidence interval = [-2.75, -1.25]

Q: To further test the hospital triage system, administrators selected 200 nights and randomly assigned a new triage system to be used on 100 nights and a standard system on the remaining 100 nights. They calculated the nightly median waiting time (MWT) to see a physician. The average MWT for the new system was 4 hours with a standard deviation of 0.5 hours while the average MWT for the old system was 6 hours with a standard deviation of 2 hours. Consider the hypothesis of a decrease in the mean MWT associated with the new treatment. What does the 95% independent group confidence interval with unequal variances suggest vis a vis this

hypothesis? (Because there's so many observations per group, just use the Z quantile instead of the T.)

Assuming we subtract in this order (New System — Old System):



The screenshot shows a presentation slide with a title bar at the top. Below the title bar, there is a formula for the confidence interval of the difference between two means:

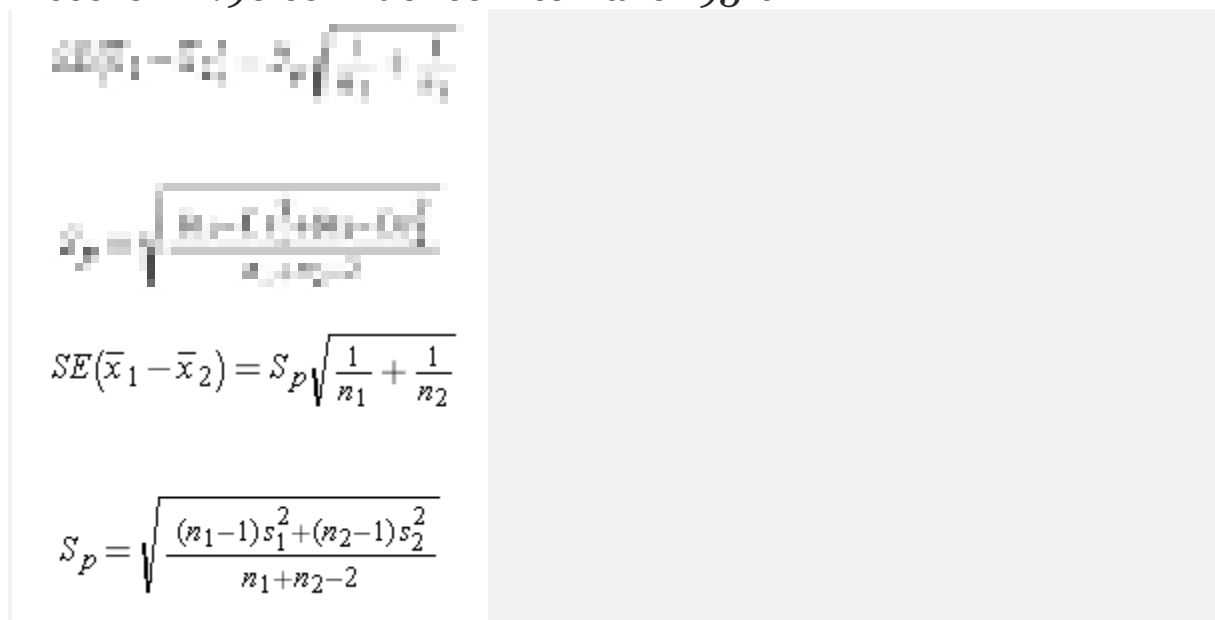
$$(\bar{x}_1 - \bar{x}_2) \pm z S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Below the formula, it says "Use Z table for standard normal distribution".

confidence interval formula for two independent samples

mean = new mean — old mean = 4–6 = -2

z-score = 1.96 confidence interval of 95%



The screenshot shows a presentation slide with a title bar at the top. Below the title bar, there are four formulas related to the confidence interval for two independent samples:

$$SE(\bar{x}_1 - \bar{x}_2) = z \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$S_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$SE(\bar{x}_1 - \bar{x}_2) = S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$S_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

st. error = sqrt((0.52\*99+22\*99)/(100+100–2)) \*  
sqrt(1/100+1/100)

standard error = 0.205061

lower bound =  $-2 - 1.96 * 0.205061 = -2.40192$

upper bound =  $-2 + 1.96 * 0.205061 = -1.59808$

confidence interval =  $[-2.40192, -1.59808]$

## SQL Practice Problems

### PROBLEM #1: Second Highest Salary

*Write a SQL query to get the second highest salary from the `Employee` table. For example, given the `Employee` table below, the query should return `200` as the second highest salary. If there is no second highest salary, then the query should return `null`.*

Id	Salary
1	100
2	200
3	300

### SOLUTION A: Using IFNULL, OFFSET

- **IFNULL(expression, alt)** : ifnull() returns the specified value if null, otherwise returns the expected value. We'll use this to return null if there's no second-highest salary.
- **OFFSET** : offset is used with the ORDER BY clause to disregard the top n rows that you specify. This will be useful as you'll want to get the second row (2nd highest salary)

```

SELECT
    IFNULL (
        (SELECT DISTINCT Salary
         FROM Employee
         ORDER BY Salary DESC
         LIMIT 1 OFFSET 1
        ), null) as SecondHighestSalary
FROM Employee
LIMIT 1

```

## SOLUTION B: Using MAX()

This query says to choose the MAX salary that isn't equal to the MAX salary, which is equivalent to saying to choose the second-highest salary!

```

SELECT MAX(salary) AS SecondHighestSalary
FROM Employee
WHERE salary != (SELECT MAX(salary) FROM Employee)

```

[Here are three SQL concepts to review before your next interview!](#)

## PROBLEM #2: Duplicate Emails

*Write a SQL query to find all duplicate emails in a table named `Person`.*

```

+-----+-----+
| Id | Email |
+-----+-----+
| 1 | a@b.com |
| 2 | c@d.com |
| 3 | a@b.com |
+-----+-----+

```

## SOLUTION A: COUNT() in a Subquery

First, a subquery is created to show the count of the frequency of each email. Then the subquery is filtered WHERE the count is greater than 1.



```
SELECT Email
FROM (
    SELECT Email, count(Email) AS count
    FROM Person
    GROUP BY Email
) as email_count
WHERE count > 1
```

## SOLUTION B: HAVING Clause

- **HAVING** is a clause that essentially allows you to use a WHERE statement in conjunction with aggregates (GROUP BY).

```
SELECT Email
FROM Person
GROUP BY Email
HAVING count(Email) > 1
```

## PROBLEM #3: Rising Temperature

Given a `Weather` table, write a SQL query to find all dates' Ids with higher temperature compared to its previous (yesterday's) dates.

Id (INT)	RecordDate (DATE)	Temperature (INT)
1	2015-01-01	10
2	2015-01-02	25
3	2015-01-03	20
4	2015-01-04	30

## SOLUTION: DATEDIFF()

- **DATEDIFF** calculates the difference between two dates and is used to make sure we're comparing today's temperature to yesterday's temperature.

In plain English, the query is saying, Select the Ids where the temperature on a given day is greater than the temperature yesterday.

```
SELECT DISTINCT a.Id
FROM Weather a, Weather b
WHERE a.Temperature > b.Temperature
AND DATEDIFF(a.Recorddate, b.Recorddate) = 1
```

## PROBLEM #4: Department Highest Salary

The `Employee` table holds all employees. Every employee has an `Id`, a salary, and there is also a column for the department `Id`.

Id	Name	Salary	DepartmentId
1	Joe	70000	1
2	Jim	90000	1
3	Henry	80000	2
4	Sam	60000	2
5	Max	90000	1

The `Department` table holds all departments of the company.

Id	Name
1	IT
2	Sales

Write a SQL query to find employees who have the highest salary in each of the departments. For the above tables, your SQL query should return the following rows (order of rows does not matter).

Department	Employee	Salary
IT	Max	90000
IT	Jim	90000
Sales	Henry	80000

## SOLUTION: IN Clause

- The **IN** clause allows you to use multiple OR clauses in a WHERE statement. For example WHERE country = 'Canada' or country = 'USA' is the same as WHERE country IN ('Canada', 'USA').
- In this case, we want to filter the Department table to only show the highest Salary per Department (i.e. DepartmentId). Then we can join the two tables WHERE the DepartmentId and Salary is in the filtered Department table.

```
SELECT
    Department.name AS 'Department',
    Employee.name AS 'Employee',
    Salary
FROM Employee
INNER JOIN Department ON Employee.DepartmentId = Department.Id
WHERE (DepartmentId , Salary)
    IN
    (
        SELECT
            DepartmentId, MAX(Salary)
        FROM
            Employee
        GROUP BY DepartmentId
    )
```

## PROBLEM #5: Exchange Seats

*Mary is a teacher in a middle school and she has a table `seat` storing students' names and their corresponding seat ids. The column **id** is a continuous increment. Mary wants to change seats for the adjacent students.*

*Can you write a SQL query to output the result for Mary?*

```
+-----+-----+
|      id      | student |
```

1	Abbot	
2	Doris	
3	Emerson	
4	Green	
5	Jeames	

*For the sample input, the output is:*

id	student	
1	Doris	
2	Abbot	
3	Green	
4	Emerson	
5	Jeames	

### **Note:**

*If the number of students is odd, there is no need to change the last one's seat.*

### **SOLUTION: CASE WHEN**

- Think of a CASE WHEN THEN statement like an IF statement in coding.
- The first WHEN statement checks to see if there's an odd number of rows, and if there is, ensure that the id number does not change.
- The second WHEN statement adds 1 to each id (eg. 1,3,5 becomes 2,4,6)
- Similarly, the third WHEN statement subtracts 1 to each id (2,4,6 becomes 1,3,5)

```
SELECT
CASE
```

```

        WHEN((SELECT MAX(id) FROM seat)%2 = 1) AND id = (SELECT
MAX(id) FROM seat) THEN id
        WHEN id%2 = 1 THEN id + 1
        ELSE id - 1
    END AS id, student
FROM seat
ORDER BY id

```

## Miscellaneous

Q: If there are 8 marbles of equal weight and 1 marble that weighs a little bit more (for a total of 9 marbles), how many weighings are required to determine which marble is the heaviest?

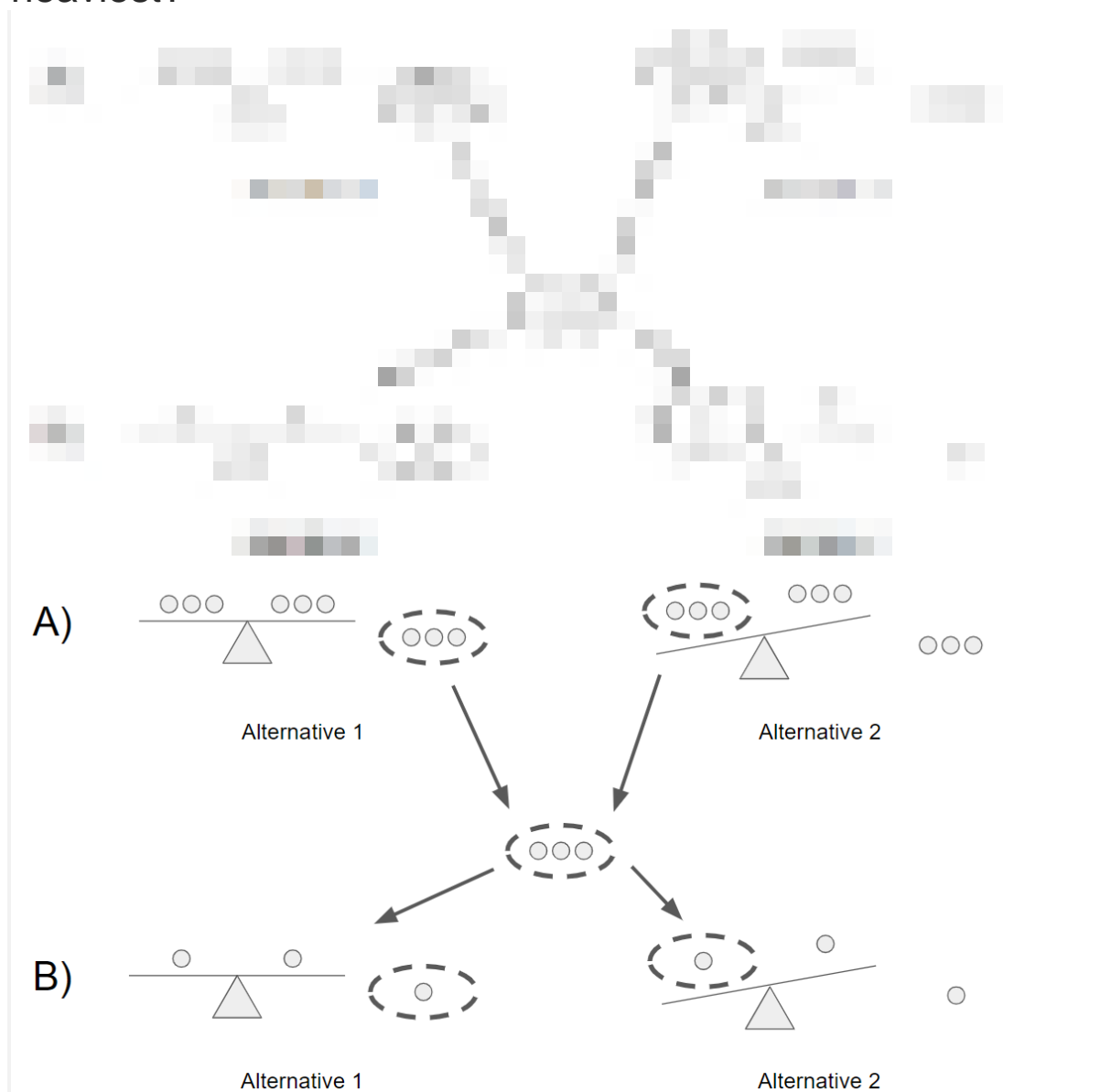


Image created by author

Two weighings would be required (see part A and B above):

1. You would split the nine marbles into three groups of three and weigh two of the groups. If the scale balances (alternative 1), you know that the heavy marble is in the third group of marbles. Otherwise, you'll take the group that is weighed more heavily (alternative 2).
2. Then you would exercise the same step, but you'd have three groups of one marble instead of three groups of three.

Q: How would the change of prime membership fee affect the market?

*I'm not 100% sure about the answer to this question but will give my best shot!*

Let's take the instance where there's an increase in the prime membership fee — there are two parties involved, the buyers and the sellers.

For the buyers, the impact of an increase in a prime membership fee ultimately depends on the price elasticity of demand for the buyers. If the price elasticity is high, then a given increase in price will result in a large drop in demand and vice versa. Buyers that continue to purchase a membership fee

are likely Amazon's most loyal and active customers — they are also likely to place a higher emphasis on products with prime.

Sellers will take a hit, as there is now a higher cost of purchasing Amazon's basket of products. That being said, some products will take a harder hit while others may not be impacted. It is likely that premium products that Amazon's most loyal customers purchase would not be affected as much, like electronics.

Q: If 70% of Facebook users on iOS use Instagram, but only 35% of Facebook users on Android use Instagram, how would you investigate the discrepancy?

There are a number of possible variables that can cause such a discrepancy that I would check to see:

- The demographics of iOS and Android users might differ significantly. For example, according to [Hootsuite](#), 43% of females use Instagram as opposed to 31% of men. If the proportion of female users for iOS is significantly larger than for Android then this can explain the discrepancy (or at least a part of it). This can also be said for age, race, ethnicity, location, etc...
- Behavioral factors can also have an impact on the discrepancy. If iOS users use their phones more heavily than Android users, it's more likely that they'll indulge in Instagram and other apps

than someone who spent significantly less time on their phones.

- Another possible factor to consider is how Google Play and the App Store differ. For example, if Android users have significantly more apps (and social media apps) to choose from, that may cause greater dilution of users.
- Lastly, any differences in the user experience can deter Android users from using Instagram compared to iOS users. If the app is more buggy for Android users than iOS users, they'll be less likely to be active on the app.

*Check out more Facebook data science interview questions [here](#)*

Q: Likes/user and minutes spent on a platform are increasing but total number of users are decreasing. What could be the root cause of it?

Generally, you would want to probe the interviewer for more information but let's assume that this is the only information that he/she is willing to give.

Focusing on likes per user, there are two reasons why this would have gone up. The first reason is that the engagement of users has generally increased on average over time — this makes sense because as time passes, active users are more likely to be loyal users as using the platform becomes a habitual practice. The other reason why likes per user would increase is that the



denominator, the total number of users, is decreasing.

Assuming that users that stop using the platform are inactive users, aka users with little engagement and fewer likes than average, this would increase the average number of likes per user.

The explanation above can also be applied to minutes spent on the platform. Active users are becoming more engaged over time, while users with little usage are becoming inactive. Overall the increase in engagement outweighs the users with little engagement.

To take it a step further, it's possible that the 'users with little engagement' are bots that Facebook has been able to detect. But over time, Facebook has been able to develop algorithms to spot and remove bots. If there were a significant number of bots before, this can potentially be the root cause of this phenomenon.

Q: Facebook sees that likes are up 10% year over year, why could this be?

The total number of likes in a given year is a function of the total number of users and the average number of likes per user (which I'll refer to as engagement).

Some potential reasons for an increase in the total number of users are the following: users acquired due to international expansion and younger age groups signing up for Facebook as they get older.

Some potential reasons for an increase in engagement are an increase in usage of the app from users that are becoming more and more loyal, new features and functionality, and an improved user experience.

Q: If we were testing product X, what metrics would you look at to determine if it is a success?

The metrics that determine a product's success are dependent on the business model and what the business is trying to achieve through the product. The book Lean analytics lays out a great framework that one can use to determine what metrics to use in a given scenario:



	E-commerce	2-sided market	SaaS	Mobile app	User-generated content	Media
<b>Empathy</b>	Interviews; qualitative results; quantitative scoring; surveys					
<b>Stickiness</b>	Loyalty, conversion	Inventory, listings	Engagement, churn	Downloads, churn, virality	Content, spam	Traffic, visits, returns
<b>Virality</b>	CAC, shares, reactivation	SEM, sharing	Inherent virality, CAC	WoM, app ratings, CAC	Invites, sharing	Content virality, SEM
	(Money from transactions)		(Money from active users)		(Money from ad clicks)	
<b>Revenue</b>	Transaction, CLV	Transactions, commission	Upselling, CAC, CLV	CLV, ARPDAU	Ads, donations	CPE, affiliate %, eyeballs
<b>Scale</b>	Affiliates, white-label	Other verticals	API, magic #, mktplace	Spinoffs, publishers	Analytics, user data	Syndication, licenses

Framework from Lean Analytics

Q: If a PM says that they want to double the number of ads in Newsfeed, how would you figure out if this is a good idea or not?

You can perform an A/B test by splitting the users into two groups: a control group with the normal number of ads and a test group with double the number of ads. Then you would choose the metric to define what a “good idea” is. For example, we can say that the null hypothesis is that doubling the number of ads will reduce the time spent on Facebook and the alternative hypothesis is that doubling the number of ads won’t have any impact on the time spent on Facebook. However, you can choose a different metric like the number of active users or the churn rate. Then you would conduct the test and determine the statistical significance of the test to reject or not reject the null.

Q: What is: lift, KPI, robustness, model fitting, design of experiments, 80/20 rule?

**Lift:** lift is a measure of the performance of a targeting model measured against a random choice targeting model; in other words, lift tells you how much better your model is at predicting things than if you had no model.

**KPI:** stands for Key Performance Indicator, which is a measurable metric used to determine how well a company is achieving its business objectives. Eg. error rate.

**Robustness:** generally robustness refers to a system's ability to handle variability and remain effective.

**Model fitting:** refers to how well a model fits a set of observations.

**Design of experiments:** also known as DOE, it is the design of any task that aims to describe and explain the variation of information under conditions that are hypothesized to reflect the variable. [4] In essence, an experiment aims to predict an outcome based on a change in one or more inputs (independent variables).

**80/20 rule:** also known as the Pareto principle; states that 80% of the effects come from 20% of the causes. Eg. 80% of sales come from 20% of customers.

Q: Define quality assurance, six sigma.

**Quality assurance:** an activity or set of activities focused on maintaining a desired level of quality by minimizing mistakes and defects.

**Six sigma:** a specific type of quality assurance methodology composed of a set of techniques and tools for process improvement. A six sigma process is one in which 99.99966% of all outcomes are free of defects.