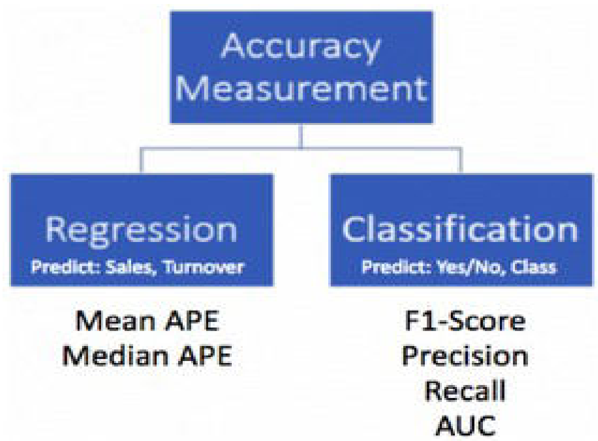
Q: How did you measure your predictive model accuracy?

Accuracy is measured differently for Regression and Classification predictive models. Depending on what type of project you did, mention how accuracy was measured and how the model was deployed in production.

Regression: 100-MAPE or 100- (Median APE)

Classification: F1-Score, Precision, Recall, AUC, ROC

All these terms are explained one by one in the following section.



Measuring predictive model accuracy

Q: What is MAPE?

Mean Absolute Percent Error.

This value gives an idea that, on an average, how much error the predictive model is doing for each of the predictions.

Original	Prediction	Absolute Difference	APE
20	25	5	25.0
31	30	1	3.2
33	29	4	12.1
38	40	2	5.3
		MAPE	**11.4**
		Median APE	**8.7**

How to calculate MAPE

APE is calculated by finding the absolute percentage difference between the predicted and original values. Take a look at the above example, in the first row the APE is 25% because the absolute difference between Original and Prediction is 5 and when we divide it by Original value 20 it gives 25%.

In a way you can say the prediction was 25% away from the original, hence the accuracy for this prediction was 100-25 = 75%. Now if we need to understand the overall accuracy, then we take the average of error for each of the predictions. This is known as Mean Absolute Percent Error.

Accuracy of the model is 100- MAPE

Q. What is the Median APE?

Median Absolute Percent Error.

Original	Prediction	Absolute Difference	APE
20	25	5	25.0
31	30	1	3.2
33	29	4	12.1
38	40	2	5.3
		MAPE	**11.4**
		Median APE	**8.7**

How to calculate Median APE

You can see MAPE is higher because of the outlier ‘25.0’, which means there is one prediction which has 25% error.

The Median APE is used because Mean APE is affected by outliers and can go above 100% also, which will make the accuracy value(100-MAPE) negative.

This is helpful to analyze the central tendency of the error committed by the predictive model. For example, if the Median APE is 5% then it tells that if there are 60 total predictions done, 30 of those will have an error value of less than 5%.

Q. What is RMSE?

Root Mean Squared Error

1. Find out the difference between original and predicted values for each row.
2. Square the differences
3. Sum all squared differences
4. Take the average of the above sum
5. Take the square root of above average

Original	Predicted	Residual	Difference	Abs Difference	Squared Difference		Absolute Percent Error(APE)
10	12	10-12	-2	2	4		20.00%
14	13	14-13	1	1	1		7.14%
18	15	18-15	3	3	9		16.67%
20	23	20-23	-3	3	9		15.00%
11	15	11-15	-4	4	16		36.36%
				SUM	39	**MAPE**	**19.03%**
				Mean(SUM)	7.8		
				RMSE	**2.79**		

How to calculate RMSE

Q. Which one is better RMSE or MAPE?

In terms of interpretation, MAPE is better because it is easy to visualize it. It represents the on an average error committed by the predictive model. RMSE does not have such a clear visualization.

In terms of penalizing the large error (Outliers), RMSE is better. MAPE gets affected by the outliers.

Q. How to measure accuracy for Regression Models?

Subtract the Mean Absolute Percent Error from 100 and the value is accuracy. All the below calculations will fetch accuracy values for the predictive model.

- 100-MAPE
- 100-(Median APE)
- 100-RMSE

Q. What is R-Squared value in Regression?

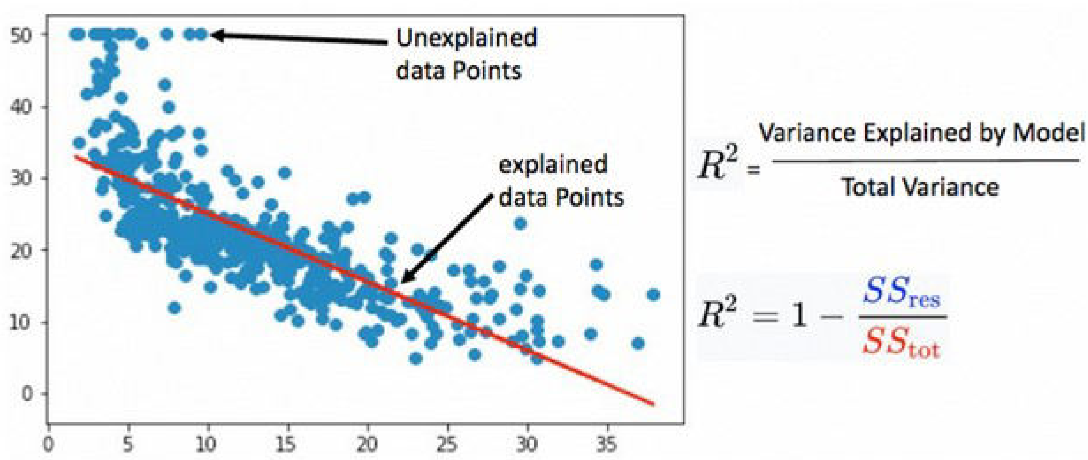
R2 value measures the goodness of fit. It is NOT the Accuracy of the model. Accuracy is measured using 100-MAPE value.

It tells how many data points are being explained by the model out of all the data points. That means variance explained by the model Vs Total Variance.

- Max value of R2 is 1
- Min Value of R2 is 0

An Ideal range of R2 value is between 0.6 to 0.9. This means the predictive model is able to explain a good amount of variance in the data and can be taken into consideration for testing and accuracy calculation on Test Data.

- R2 < 0.5 means tending towards **Underfitting** of the model
- R2 > 0.9 means tending towards **Overfitting** of the model



What is R-Squared value

From a visual perspective How many points are closer to the line of best fit Vs how many points which are far away from the line.

Q. How R-Squared value is calculated?

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

How R2 is calculated?

Consider below an example of five predictions and original values

	Original	Predicted	Distance from Mean	Residual	SS res	SS total
	10	12	10-14.6	10-12	39	75.2
	14	13	14-14.6	14-13		
	18	15	18-14.6	18-15		
	20	23	20-14.6	20-23		
	11	15	11-14.6	11-15		
Mean	**14.6**					

How to calculate R2

SSres means the Sum of Squared Residuals.

In the above example, SSres is 39.

SStotal means the Sum of Squared distance of each point from the mean value.

In the above example, SStotal is 75.2

The calculation for SSres and SStot can be seen below.

- SSres = $(12-10)^2 + (14-13)^2 + (18-15)^2 + (20-33)^2 + (11-15)^2$

- SStot = $(10-14.6)^2 + (14.6-14)^2 + (18-14.6)^2 + (20-14.6)^2 + (11-14.6)^2$

This equates to R2 = 1- (39/75.2) = 0.48

This means, for the given data, model was able to explain 48% of variance out of total variance.

Q. What is Adjusted R-Squared Value?

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

How to calculate Adjusted R2

- p =Total number of explanatory variables in the model
- n= number of rows in training data.

The adjusted R2 value is always less than R2 and It can be negative also.

Adjusted R2 takes into account the addition of new predictors to the model. It adjusts the value and does not allow the variance explained to increase just for adding new predictor.

The Adjusted R2 value increases only if the new predictor is significant and helpful to predict the target variable. Whereas, R2 increases with every new predictor's addition to the model.

Hence Adjusted R2 value is more accurate while judging the goodness of fit for regression models.

Q. How to create a Confusion Matrix?

A Confusion matrix is created by comparing original values with predicted values in a classification model.

- True Positive(TP): How many times Yes was predicted as Yes
- True Negative(TN): How many times No was predicted as No
- False Positive(FP): How many times No was predicted as Yes
- False Negative(FN): How many times Yes was predicted as No

In below example, all of the above have been counted and the resultant matrix is known as a confusion matrix.

		Predicted	
		Yes	No
Original	Yes	3	1
	No	2	2

		Predicted	
		Yes	No
Original	Yes	TP	FN
	No	FP	TN

Original	Predicted	Comparison
Yes	Yes	Correct
Yes	No	Incorrect
No	Yes	Incorrect
No	No	Correct
No	No	Correct
Yes	Yes	Correct
Yes	Yes	Correct
No	Yes	Incorrect

Q. What is Precision?

How many correct predictions were done for a class out of all predictions for that class?

Precision for 'Yes' class will tell out of all the 'Yes' predicted by the algorithm, how many were correct? Similarly, Precision for 'No' class will tell out of all the 'No' predicted by the algorithm, how many were correct?

i.e how precise the prediction is for that class.

A Good range for precision is 0.7-0.9

$$\text{Precision (Yes)} = \frac{TP}{TP + FP}$$

$$\text{Precision (No)} = \frac{TN}{TN + FN}$$

How to calculate Precision for any class?

Q. What is Recall?

How many actual values were correctly *recalled* by the model? In other terms, how many predictions were correct out of all the original values for that class.

Recall for 'Yes' will tell out of all the *Actual* 'Yes' values how many were correctly predicted by the model.

Recall for 'No' will tell out of all the *Actual* 'No' values how many were correctly predicted by the model.

A good range for the recall is 0.7-0.9.

$$\text{Recall(Yes)} = \frac{TP}{TP + FN}$$

$$\text{Recall(No)} = \frac{TN}{TN + FP}$$

How to calculate recall for any class?

Q. What is F1-Score?

F1-Score is the harmonic mean of Precision and recall.

It is the accuracy of classification predictive model. It tells how efficient the model is while predicting *Yes* as Yes and *No* as No.

A good range for F1-Score is 0.7-0.9

$$\text{F1-Score} = \frac{2 * (\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}}$$

How to calculate F1-Score

Q. What is Sensitivity?

Recall(Yes) is also known as Sensitivity. The True Positive Rate (TPR)

$$\text{Sensitivity} = \text{Recall(Yes)} = \frac{TP}{TP + FN}$$

How to calculate Sensitivity

Q. What is Specificity?

Recall(No) class is also known as Specificity. The True Negative Rate (TNR)

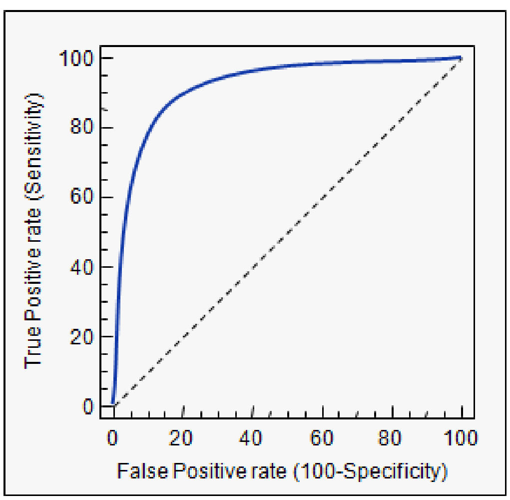
$$\text{Specificity} = \text{Recall(No)} = \frac{TN}{TN + FP}$$

How to calculate Specificity

Q. What is ROC?

The curve between True Positive Rates(TPR) in Y-Axis and False Positive Rates(FPR) in X-Axis is known as the ROC curve. ROC stands for **Receiver Operating Characteristic.**

The plot is generated by capturing (TPR, FPR) values for multiple iterations of sampling and predictions.



Q. What is AUC?

Area Under the Curve (AUC)

The amount of area covered under the ROC curve. Perfect classification will have its value as 1. A good range for AUC is 0.6-0.9. Which helps to understand the performance of the model. Higher the AUC the better it is.

If the value of AUC is less than 0.5 then it means the predictive model is not able to discriminate between the classes.

Q. What is Hypothesis testing?

Hypothesis means assumption.

To test whether our assumption is correct based on given data is Hypothesis testing.

Consider a scenario from a tire factory. The radius of the ideal tire must be 16 inches. However, even if there is a deviation of 8% then it is accepted. Hence in this scenario, we can apply hypothesis testing like below using some dummy values for the explanation.

1. **Define the Null Hypothesis (H0)**: The radius of the tire= 16 Inch
2. **Define the alternate Hypothesis(Ha)**: The radius of the tire != 16 Inch
3. **Define the error tolerance limit**: 8%
4. **Conduct the test**: Chosen T-Test
5. **Look at the P-value generated by the test**: P-value= 0.79
6. **If P-Value > 0.05 then accept the Null Hypothesis otherwise reject it**. : Accept the Hypothesis, Hence, The tire produced is of good quality

Q. What is P-Value?

P-Value is the probability of H0 being True.

The higher the P-value, the better the chances of our assumption(H0) to be true. The Textbook threshold to reject a Null Hypothesis is 5%. So, if P-Value is less than 0.05, this means there is less than 5% chance of Null Hypothesis being true, hence it is rejected. Otherwise, if P-Value is more than 0.05, then the Null Hypothesis is accepted.

Q. What is Alpha-Value?

The acceptable error threshold. Also known as Level of significance.

In the above tire example, the acceptable error amount was 8%

Q. What is Confidence Interval?

The range of values which are acceptable based on the error threshold (Alpha Value).

In the above example if the Alpha Value is 8%. So the acceptable error is 16*0.08= 1.28 Inches. This means if a tire is produced with a radius between [16-1.28, 16+1.28] Then it is a good tire with 92% confidence. Further, the 92% confidence interval is [16-1.28, 16+1.28].

Q. What is Type-1 error?

A type-1 error, also known as an error of the first kind, occurs when the **null hypothesis (H0) is true but is rejected.**

Q. What is Type-2 error?

A type II error, also known as an error of the second kind, occurs when the **null hypothesis is false, but it is erroneously accepted as true**.

Q. What is Standard Deviation?

Standard Deviation tells us the overall spread of the values by giving us the average distance of each point from the mean value. In other terms, on an average how far each point is from the mean.

All we need to do is take the square root of the variance. We call this the standard deviation. If this value is large, it means the data is very scattered, if this is small then the data is consolidated and close to each other in value. More details about standard deviation can be found in this blog post.

Q. What is an Outlier?

Certain values which are extremely low or extremely high compared to all other values in a dataset are called outliers.

Eg (1, 2, 3, 4, 5, 6, 50), here 50 is an outlier because it is abnormally large than most of the values in the dataset.

Q. What is Correlation?

Correlations are mathematical relationships between variables. You can identify correlations on a scatter diagram by the distinct patterns they form. The correlation is said to be linear if the scatter diagram shows the points lying in an approximately straight line. Let's take a look at a few common types of correlation between two variables:

Positive linear correlation (r=0 to 1)

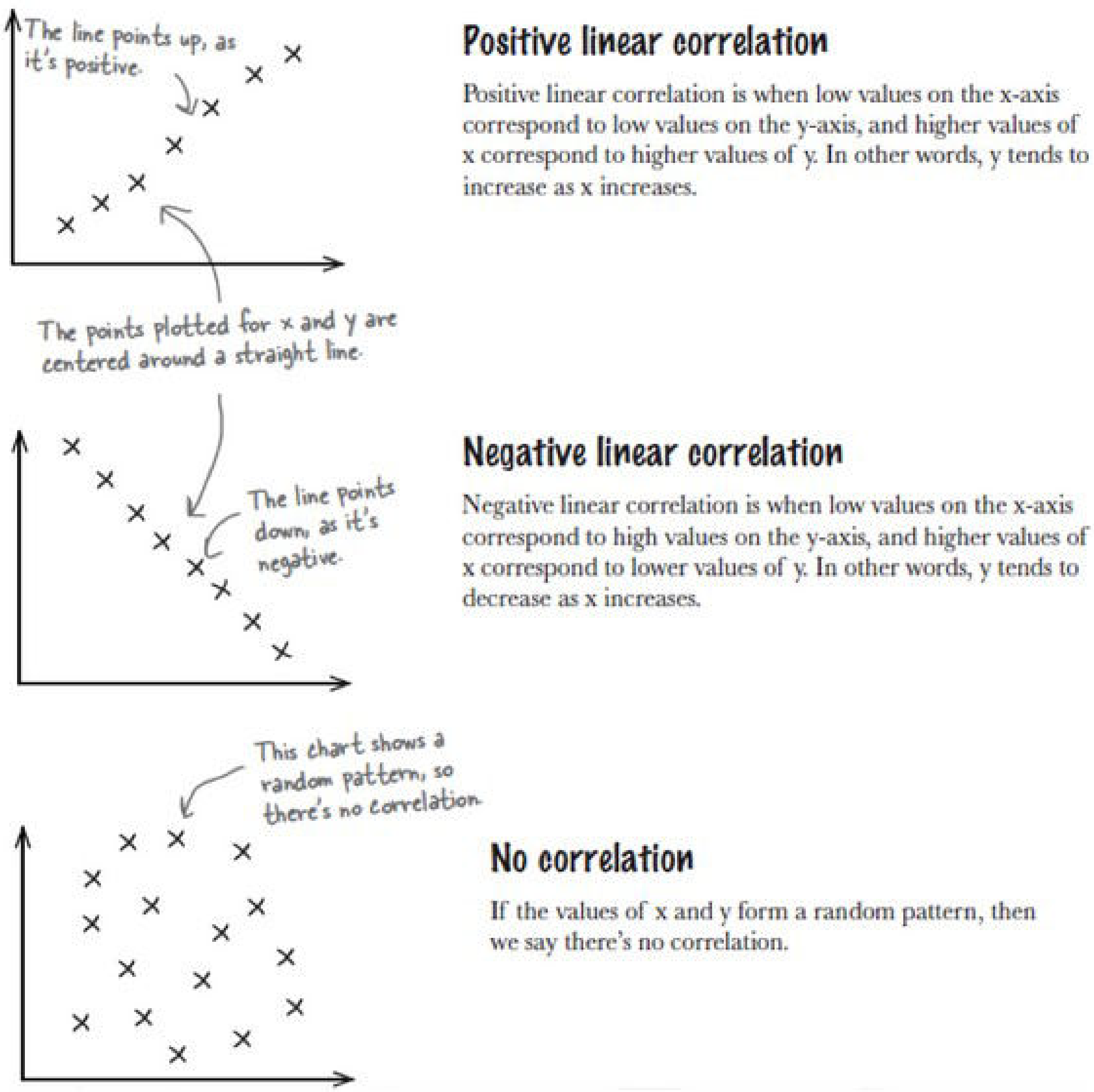
Positive linear correlation is when low values on the x-axis correspond to low values on the y-axis, and higher values of x correspond to higher values of y. In other words, y tends to increase as x increases.

Negative linear correlation(r= -1 to 0)

Negative linear correlation is when low values on the x-axis correspond to high values on the y-axis, and higher values of x correspond to lower values of y. In other words, y tends to decrease as x increases.

No correlation(r=0)

If the values of x and y form a random pattern, then we say there's no correlation.



Scatter Plot Analysis for Correlation

Q. What is central limit theorem?

If you repeatedly take large(more than 30 values) samples of size n from a population, then the mean values of all those samples will follow a normal distribution. i.e if you plot its histogram then it will form a bell curve.

Q. What is T-Test?

The T-Test is one of the many Tests employed in Hypothesis testing.

It is Used to see if the mean of the population is statistically different from an assumed value(Null Hypothesis).

Consider below example where we are selecting some random number of gumballs from a jar.

Randomly selected gums sample from a jar

Assumption: The average size of all gumballs inside the jar is **25mm (μ_0)**

If you randomly select some 20 gumballs from the jar then the average size of those gumballs should be 25mm. However, it can vary a little bit due to manufacturing defects so let us say the average came out to be **24.3mm($\bar{X}$)**

Assuming Standard Deviation of sizes of all gumballs: **0.1mm(sd)d**

$$\text{T-Value} = (\bar{X} - \mu_0) / (s / \sqrt{n})$$

Here in our case : T-Value= (24.3 – 25)/ (0.1 /√20) = -31.30

Higher the absolute T-Value, the difference between the mean of population and sample will be statistically significant.

Lower the absolute T-Value, the difference between the mean of population and sample will NOT be statistically significant. I.e the means are equal to each other from both sources.

The t-test is also used in **Linear Regression** to test which variable is helping to predict the target variable and which is not.

H0: The variable is not helping

The t-test is conducted for each of the variables and it produces a T-Value and a probability. If the probability (p-value) is less than 0.05 then we reject the hypothesis(H0). That means the variable ***is*** helping and our assumption was wrong. So we select the variable in the model.

Q. What is Z-Test?

Z-Test is same as T-Test. We use Z-Test when the sample size is MORE than 30 and otherwise T-test is used.

The t-test is used when the sample size is LESS than 30.

In Z-test, if the variance σ is not known, then it is approximated using the sample values as sd/√(n).

Z-**score** can be calculated from the following formula.

$$\mathbf{z} = (X - \mu) / \sigma$$

Where **z** is the **z**-score, X is the value of the element, μ is the population mean, and σ is the standard deviation.

Q. What is F-Test?

F-Test is used to check if the VARIANCES are equal for two populations.

It is also used in Linear Regression, where the Null Hypothesis(H0) is: Model cannot be created. if the P-Value < 0.05 that means the H0 was incorrect and hence rejected and the model is accepted.

Q. What is ANOVA test?

Analysis of Variance is used when sample means from more than 3 populations are to be compared. The F-Test is employed to do the comparison.

ANOVA stands for Analysis of Variances. It is used when means of more than 3 groups are to be compared. The t-test cannot be used here since T-Test can compare means from a maximum of 2 Groups. Hence When we have more than 2 groups we use ANOVA is performed using the F-Test.

Q. What is Chi-Square test?

Chi-Square test is used to check if there is any relationship between two categorical variables. We cannot compute the correlation value between two categorical variables hence Chi-square test is used.

$$\chi 2 = (\text{Observed} - \text{Expected})^2 / \text{Expected}$$

Q. What is AIC?

AIC The Akaike Information Criterion (AIC) provides a method for assessing the quality of your predictive model through comparison of related models. The number itself is not meaningful. If you have more than one model then you should **select the model that has the smallest AIC**.

AIC is used in Logistic Regression to perform goodness of fit test since there is no R2 for Logistic Regression.

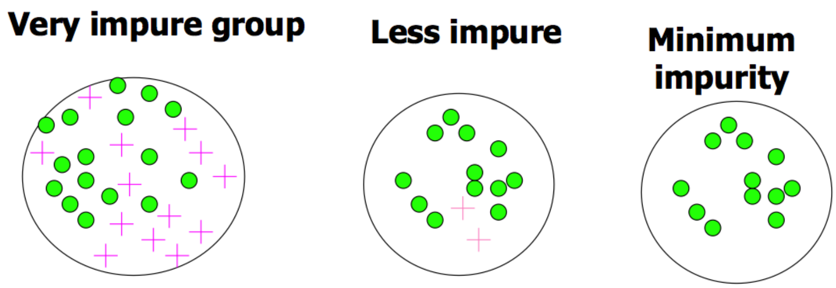
Q. What is BIC?

The formula for the Bayesian information criterion (BIC) is similar to the formula for AIC, but with a different penalty for the number of parameters. Unlike the AIC, the BIC penalizes free parameters more strongly.

Q. What is Entropy?

Entropy means randomness. It is used to measure the randomness/impurity in a group.

In simple terms, if all the entities in a group are of the same type then its pure and its randomness is also less, hence the entropy is Zero.



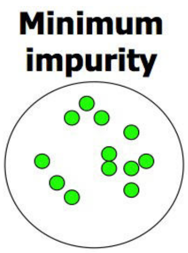
Entropy means randomness in a group

$$\text{Entropy} = \sum_i -p_i \log_2 p_i$$

Formula to calculate Entropy

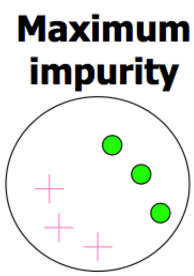
pi is the probability of class i.

The entropy of a group in which all examples belong to the same class is Zero. It means minimum randomness.



entropy= -1 log2 (1)=0

The entropy of a group in which 50% of examples belong to the same class is 1. It means maximum randomness.



entropy = -0.5 log20.5 – 0.5 log20.5 =1

Machine Learning Algorithms ID3 (**Iterative Dichotomiser 3),C4.5, C5.0** all of these uses Entropy in order to find the best root node and split further nodes.

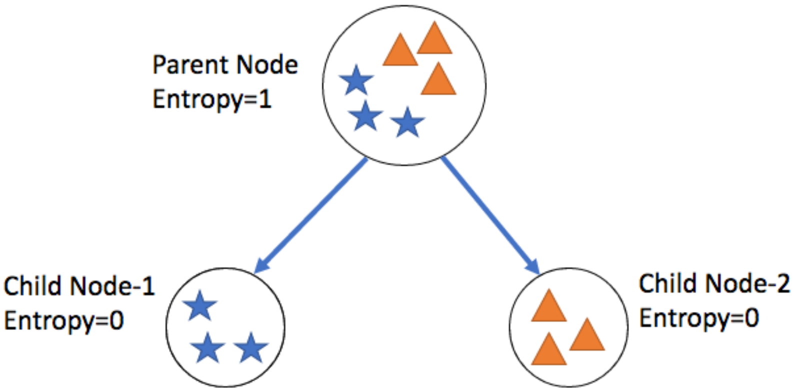
Q. What is Information Gain?

How much information is gained if a node is split in a decision tree? In other terms how much discriminative power is gained if the node is split.

Formally it is defined as below. The total entropy of Parent node minus the weighted average of Entropy of all child nodes.

Information Gain= Entropy(Parent Node) – Average(Entropy(All Child Nodes))

In the best case scenario, the parent node will have highest Entropy=1 and all the child nodes will have Entropy=0. The information gain in this scenario will be the highest. The value will be equal to 1.



Information Gain = 1 – [(3/6) * 0 + (3/6) *0] = 1

Machine Learning Algorithm **CART** (Classification and Regression Trees) uses Information Gain or Gini Index to find the best root node and split further nodes.

Q. What is Gini-Index?

Gini Index is similar to Entropy. If all values are same, then the value of Gini Index will be zero. Otherwise, it will be some positive value calculated using the below formula.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

How to calculate Gini Index

For example, if there are two classes (Binary Classification) and both of them are present 50/50 then the value calculated will be as below:

Gini= 1 – (1/2)^2—(1/2)^2 = 0.5

Q. What is Multicollinearity?

Collinearity means a linear relationship between *two* variables.

Two variables are perfectly collinear if there is an exact linear relationship between them. For example, V2 = a* V1 +b. If there is such a relation, then V1 and V2 are collinear.

Multicollinearity refers to a situation in which two or more explanatory variables in a Multiple Regression model are highly linearly related.

More commonly, the issue of multicollinearity arises when there is an approximately linear relationship between two or more independent variables(Predictors).

In simple terms Two Predictor variables have a high correlation value will generate **Multicollinearity**.

It is bad for R2 value since it inflates it. This happens because the model thinks it is explaining a lot of variances, but it is actually explaining the same variance twice (High Correlation between Predictors).

Q. How to remove Multicollinearity in Data?

1. Check the VIF of all the Predictor variables using **vif()** function from the **library(car) in R**. OR the **variance_inflation_factor()** function present in **statsmodel lib in python**
2. If any variable Has VIF>5 then remove it from the regression equation.
3. Re-check the VIF
4. Repeat Steps 1-3 till all variables have VIF<5

Q. What is VIF

Variance Inflation Factor. It is used to detect multicollinearity in data.

$$\text{tolerance} = 1 - R_j^2, \quad \text{VIF} = \frac{1}{\text{tolerance}}$$

Formula to calculate VIF

R^2j is the R-Squared of regression of Predictor *j* on all the other Predictors.

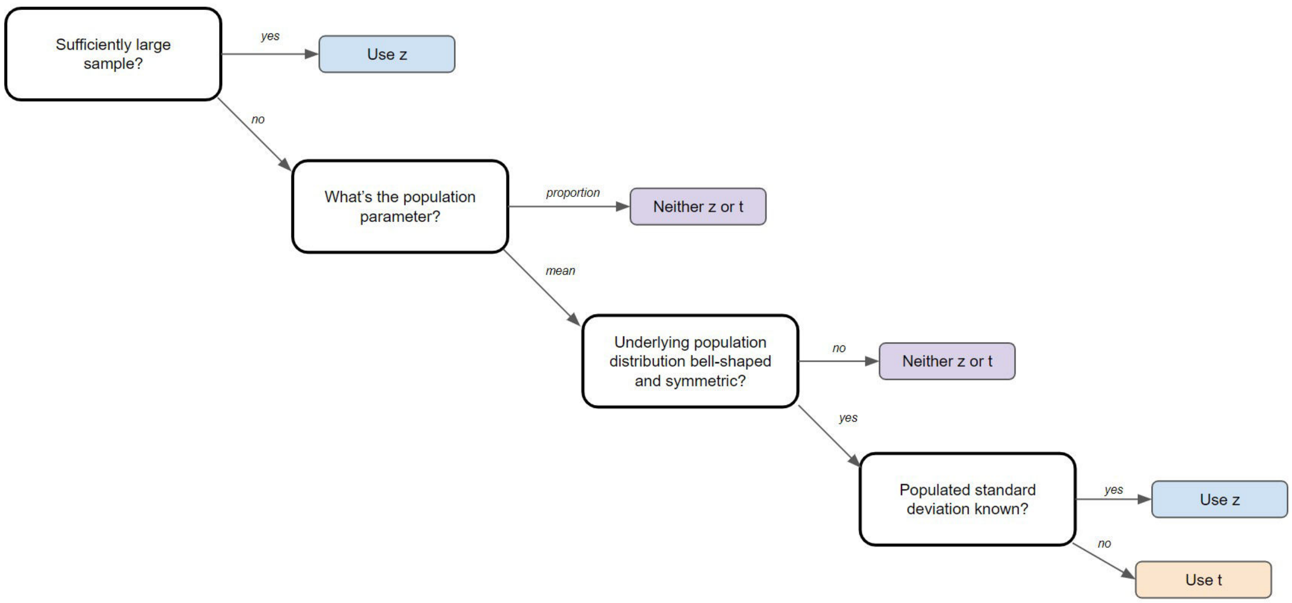
- A tolerance of less than 0.20 or 0.10 and/or a **VIF of 5 or 10** and above indicates a multicollinearity problem.
- If VIF is found with multiple Predictors, The Predictor with Highest VIF is removed and the test is conducted again.
- **library(car)** in R has a function called vif() to calculate VIF for each of the predictors.
- **variance_inflation_factor()** function present in **statsmodel lib in python.**

Q: When should you use a t-test vs a z-test?

A **Z-test** is a hypothesis test with a normal distribution that uses a **z-statistic**. A z-test is used when you know the population variance or if you don’t know the population variance but have a large sample size.

A **T-test** is a hypothesis test with a t-distribution that uses a **t-statistic**. You would use a t-test when you don’t know the population variance and have a small sample size.

You can see the image below as a reference to guide which test you should use:



Q: How would you describe what a ‘p-value’ is to a non-technical person?

The best way to describe the p-value in simple terms is with an example. In practice, if the p-value is less than the alpha, say of 0.05, then we’re saying that there’s a probability of less than 5% that the result could have happened by chance. Similarly, a p-value of 0.05 is the same as saying “5% of the time, we would see this by chance.”

Q: What is cherry-picking, P-hacking, and significance chasing?

Cherry picking refers to the practice of only selecting data or information that supports one’s desired conclusion.

P-hacking refers to when one manipulates his/her data collection or analysis until non-significant results become significant. This includes deciding mid-test to not collect anymore data.

Significance chasing refers to when a researcher reports insignificant results as if they’re “almost” significant.

Q: What is the assumption of normality?

The assumption of normality is the the sampling distribution is normal and centers around the population parameter, according to the central limit theorem.

Q: What is the central limit theorem and why is it so important?

The central limit theorem is very powerful — it states that the distribution of sample means approximates a normal distribution.

To give an example, you would take a sample from a data set and calculate the mean of that sample. Once repeated multiple times, you would plot all your means and their frequencies onto a graph and see that a bell curve, also known as a normal distribution, has been created. The mean of this distribution will closely resemble that of the original data.

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

Q: What is the empirical rule?

The empirical rule states that if a dataset is normally distributed, 68% of the data will fall within one standard deviation, 95% of the data will fall within two standard deviations, and 99.7% of the data will fall within 3 standard deviations.

Q: What general conditions must be satisfied for the central limit theorem to hold?

1. The data must be sampled randomly
2. The sample values must be independent of each other
3. The sample size must be sufficiently large, generally it should be greater or equal than 30

Q: What is the equation for confidence intervals for means vs for proportions?

$$CI\ for\ means : (\ \bar{x} \pm Z\frac{\sigma}{\sqrt{n}}\ )$$

$$CI\ for\ proportions : (p_{hat} \pm Z\sqrt{\frac{p_{hat}(1-p_{hat})}{n}}\ )$$

Q: What is the difference between a combination and a permutation?

A permutation of n elements is any arrangement of those n elements in a **definite order**. There are n factorial (n!) ways to arrange n elements. *Note the bold: order matters!* **The number of permutations of n things taken r-at-a-time** is defined as the number of r-tuples that can be taken from n different elements and is equal to the following equation:

$$P_{n,r} = \frac{n!}{(n-r)!}$$

On the other hand, combinations refer to the number of ways to choose r out of n objects where **order doesn't matter**. **The number of combinations of n things taken r-at-a-time** is defined as the number of subsets with r elements of a set with n elements and is equal to the following equation:

$$C_r^n = \frac{n!}{(n-r)!\,r!}$$

Q: How many permutations does a license plate have with 6 digits?

$$P_{9,6} = \frac{9!}{(9-6)!} = 60480$$

Q: *How many ways can you draw 6 cards from a deck of 52 cards?*

$$C_6^{52} = \frac{52!}{(52-6)!\,6!} = 20358520$$

If you want more technical interview questions like this, you can find more here!

Q: How are confidence tests and hypothesis tests similar? How are they different?

Confidence intervals and hypothesis testing are both tools used for to make statistical inferences.

The confidence interval suggests a range of values for an unknown parameter and is then associated with a confidence level that the true parameter is within the suggested range of. Confidence intervals are often very important in medical research to provide researchers with a stronger basis for their estimations. A confidence interval can be shown as "10 +/- 0.5" or [9.5, 10.5] to give an example.

Hypothesis testing is the basis of any research question and often comes down to trying to prove something did not happen by chance. For example, you could try to prove when rolling a dye, one number was more likely to come up than the rest.

Q: What is the difference between observational and experimental 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.

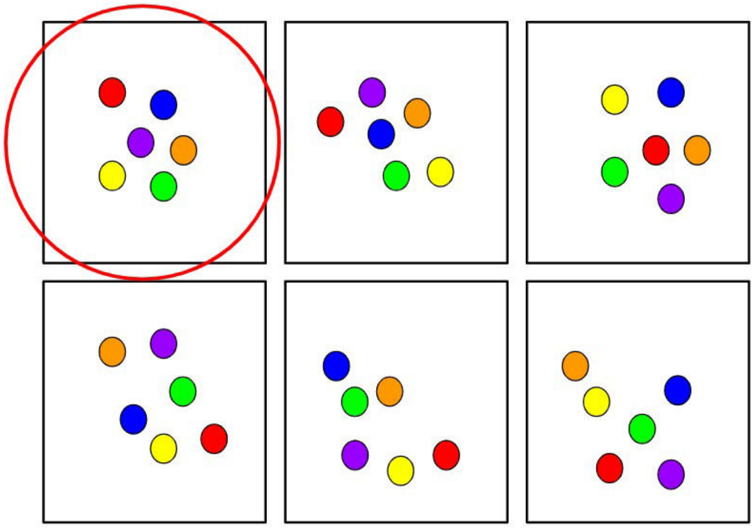
An example of experimental design is the following: split a group up into two. The control group lives their lives normally. The test group is told to drink a glass of wine every night for 30 days. Then research can be conducted to see how wine affects sleep.

Q: Give some examples of some random sampling techniques

Simple random sampling requires using randomly generated numbers to choose a sample. More specifically, it initially requires a **sampling frame**, a list or database of all members of a population. You can then randomly generate a number for each element, using Excel for example, and take the first n samples that you require.

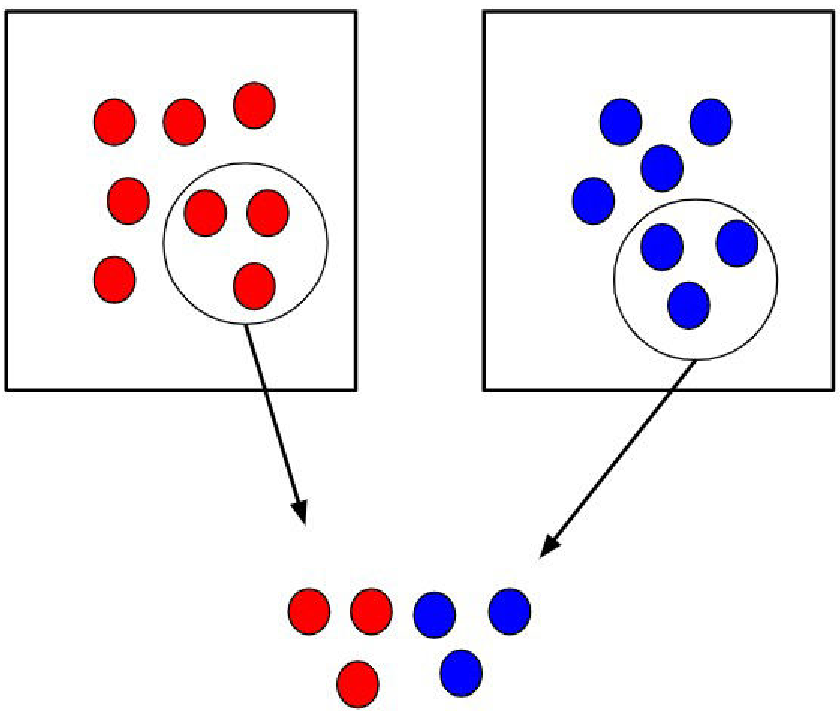
Systematic sampling can be even easier to do, you simply take one element from your sample, skip a predefined amount (n) and then take your next element. Going back to our example, you could take every fourth name on the list.

Cluster sampling starts by dividing a population into groups, or **clusters.** What makes this different that stratified sampling is that each cluster must be representative of the population. Then, you randomly selecting entire clusters to sample. For example, if an elementary school had five different grade eight classes, cluster random sampling might be used and only one class would be chosen as a sample, for example.



Example of cluster sampling

Stratified random sampling starts off by dividing a population into groups with similar attributes. Then a random sample is taken from each group. This method is used to ensure that different segments in a population are equally represented. To give an example, imagine a survey is conducted at a school to determine overall satisfaction. It might make sense here to use stratified random sampling to equally represent the opinions of students in each department.



Example of stratified random sampling

Q: What is the difference between type 1 error and type 2 error?

A **type 1 error** is when you incorrectly reject a true null hypothesis. It's also called a false positive.

A **type 2 error** is when you don't reject a false null hypothesis. It's also called a false negative.

Q: What is the power of a test? What are two ways to increase the power of a test?

The power of a test is the probability of rejecting the null hypothesis when it's false. It's also equal to 1 minus the beta.

To increase the power of the test, you can do two things:

1. You can increase alpha, but it also increases the chance of a type 1 error
2. Increase the sample size, n. This maintains the type 1 error but reduces type 2.

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: What is the Pareto principle?

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

Q: What is a confounding variable?

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 are the assumptions required for linear regression?

There are four major assumptions:

1. There is a linear relationship between the dependent variables and the regressors, meaning the model you are creating actually fits the data
2. The errors or residuals of the data are normally distributed and independent from each other
3. There is minimal multicollinearity between explanatory variables
4. Homoscedasticity. This means the variance around the regression line is the same for all values of the predictor variable.

Q: What does it mean if a model is heteroscedastic? what about homoscedastic?

A model is heteroscedastic when the variance in errors is **not** consistent. Conversely, a model is homoscedastic when the variances in errors is consistent.

Q: What does interpolation and extrapolation mean? Which is generally more accurate?

Interpolation is a prediction made using inputs that lie within the set of observed values. Extrapolation is when a prediction is made using an input that's outside the set of observed values.

Generally, interpolations are more accurate.

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.

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: 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 that 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 does autocorrelation mean?

Autocorrelation is when future outcomes depend on previous outcomes. When there is autocorrelation, the errors show a sequential pattern and the model is less accurate.

Q: When you sample, what potential biases can you be 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.

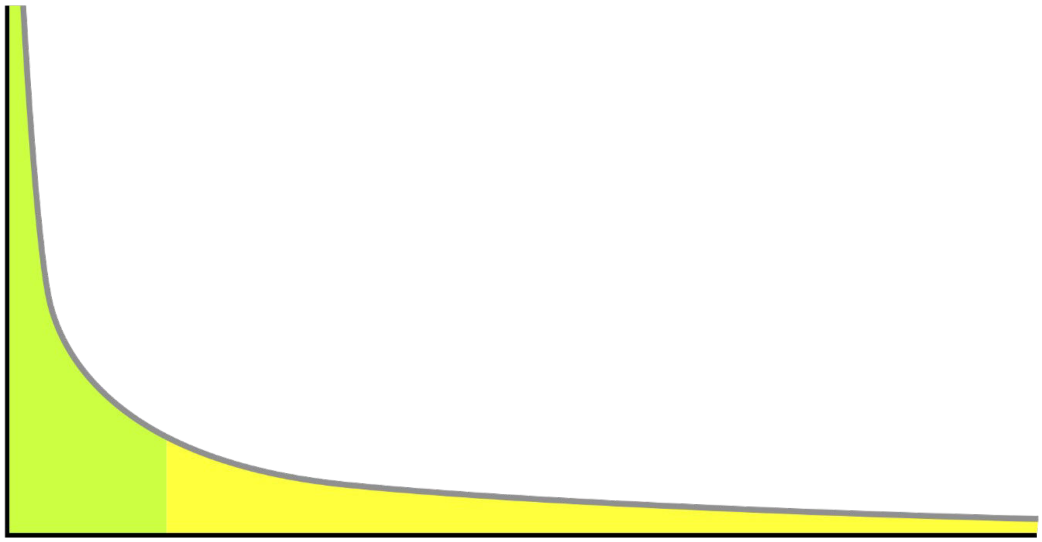
Q: How do you assess the statistical significance of an insight?

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 an outlier? Explain how you might screen for outliers and what would you do if you found them in your dataset. Also, explain what an inlier is and how you might screen for them and what would you 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 +/- 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*IRQ or greater than Q3 + 1.5*IQR. This comes to approximately 2.698 standard deviations.

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

Q: What is an inlier?

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: What does the Poisson distribution represent?

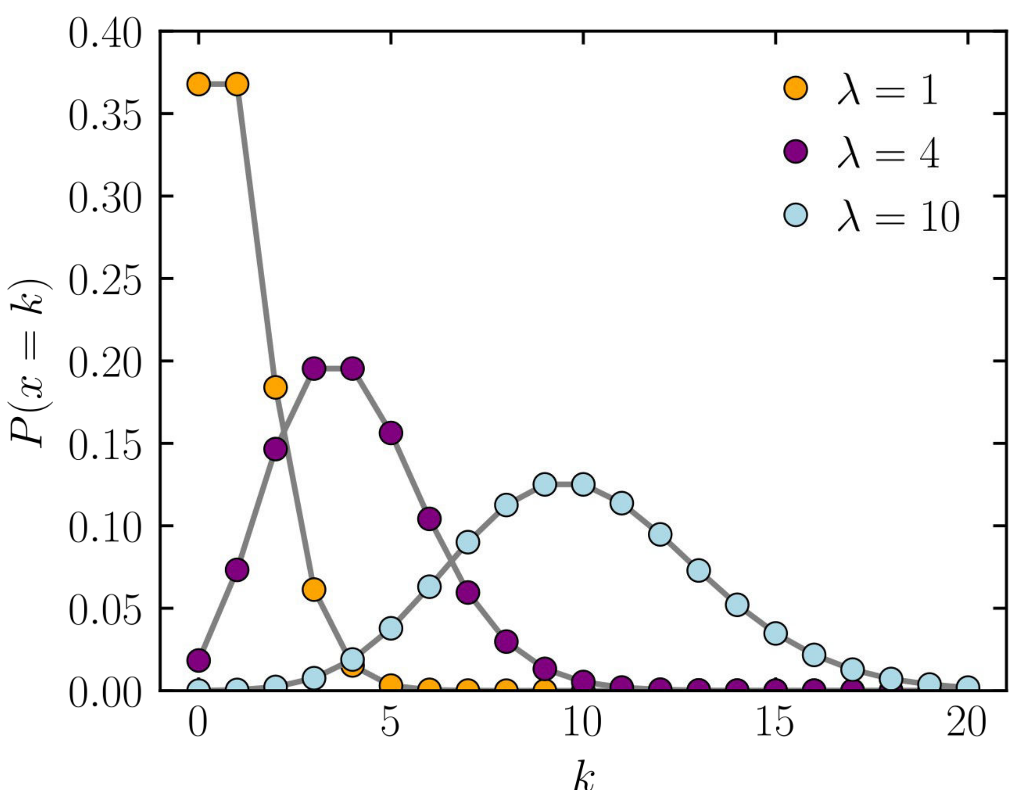


Image taken from Wikimedia

The Poisson distribution is a discrete distribution that gives the probability of the number of independent events occurring in a fixed time. An example of when you would use this is if you want to determine the likelihood of X patients coming into a hospital in a given hour.

The mean and variance are both equal to λ.

Q: What does Design of Experiments mean?

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. In essence, an experiment aims to predict an outcome based on a change in one or more inputs (independent variables).

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: 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 (H0): 1 infection per person-days
Alternative (H1): >1 infection per person-days

k (actual) = 10 infections
lambda (theoretical) = (1/100)*1787
p = 0.032372 or 3.2372%

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:

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

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
mean = 2.5*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 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.

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

Confidence interval formula

p-hat = 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: 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+/- 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(mother or father) = P(mother) + P(father) — P(mother and father)
P(mother) = P(mother or father) + P(mother and father) — P(father)
P(mother) = 0.17 + 0.06–0.12
P(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?

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

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.

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

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%

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

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

standard error = sqrt((0.^{62}*9+0.^{682}*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):

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

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%

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

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

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]

Q: There’s one box — has 12 black and 12 red cards, 2nd box has 24 black and 24 red; if you want to draw 2 cards at random from one of the 2 boxes, which box has the higher probability of getting the same color? Can you tell intuitively why the 2nd box has a higher probability

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: 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(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(rolling an 8) = 5/36

Q: If a distribution is skewed to the right and has a median of 30, will the mean be greater than or less than 30?

If the given distribution is a right-skewed distribution, then the mean should be greater than 30, while the mode remains to be less than 30.

Q: You're about to get on a plane to Seattle. You want to know if you should bring an umbrella. You call 3 random friends of yours who live there and ask each independently if it's raining. Each of your friends has a 2/3 chance of telling you the truth and a 1/3 chance of messing with you by lying. All 3 friends tell you that "Yes" it is raining. What is the probability that it's actually raining in Seattle?

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|not A) * P(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.