1. We use hypothesis testing against a sample of a population to check if a certain assumption regarding that sample is true or not which results in us accepting or rejecting that null hypothesis. This is when we take the alternate hypothesis to be true.
2. Tests like **t-tests, z tests** and **annova tests**, which assume data is from normal distribution.

Z test A z-test is a statistical test used to determine whether two population means are different when the variances are known and the sample size is large.

T test is same as Z test but it is used when the population parameters (mean and standard deviation) are not known.

1. P value is like a threshold value below which the null hypothesis will be rejected. (significance value). So it helps us determine how likely it is to get a particular result when the null hypothesis is assumed to be true
2. **Type 1** errors – often assimilated with false positives happen in hypothesis testing when the null hypothesis is true but rejected.

Eg. Suppose a person invests money in stocks and you incorrectly predict that it’ll crash, you’ve done a type 1 error

**Type 2** errors happen when the null hypothesis is false and you subsequently fail to reject it.

Eg. Suppose a person has cancer but your model predicts that he/she does not have and this gives a False negative.

So it depends from case to case depending on the severity of the situation.

1. Chi-square test is used to compare two categorical variables. Calculating the Chi-Square statistic value and comparing it against a critical value from the Chi-Square distribution allows to assess whether the observed frequency are significantly different from the expected frequency.
2. A picture containing text

   Description automatically generated

This is the formula of Chi square test.

It is primarily used for categorical data to check if each observed value and expected value have some sort of a relation. We cannot use it for numerical data because there are other tests like Z-test, F-test, T-test and others.

1. Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples.

F= **((SSE1 — SSE2)/m)/ SSE2/n-k,** where

SSE = residual sum of squares

m = number of restrictions

k = number of independent variables

1. We use t test instead of Z test in the following scenarios

* We do not know the population variance
* Our sample size is small, n < 30

let’s say we want to determine if on average, boys score 15 marks more than girls in the exam. We do not have the information related to variance (or standard deviation) for girls’ scores or boys’ scores. To perform a t-test. we randomly collect the data of 10 girls and boys with their marks. We choose our ⍺ value (significance level) to be 0.05 as the criteria for Hypothesis Testing.

Here if the Standard deviation and variance is given (i.e we assume its normally distributed) we can use Z test, else we use T test.

1. When we make generalizations from a population by taking a sample mean, mode or median. we use inferential statistics to make judgments of the probability that an observed difference between groups is a dependable one or one that might have happened by chance in this study.
2. When you have an entire population and calculate any parameter (like the population variance or population standard deviation), your results will be accurate. That’s because you have all the data about your population. However, when you work with a sample, you’ve only got a small fraction of the population to work with. Therefore, your answers aren’t going to be as accurate as those you would have got, if you had the entire set of data to work with.
3. When the mean is greater than the median it is called being right skewed (positive skewness). This is when there are more outliers on the right hand side of the graph.

Eg. Wealth distribution follows right skewness

1. Normal distribution/Gaussian distribution – This is when the data distribution is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve. In either one of the std. deviation (68.7%) is covered. In 2 std. deviations it becomes (approx. 87%)

Standard normal distribution – Same as normal distribution except that mean is zero and standard deviation of 1

uniform distribution refers to a type of probability distribution in which all outcomes are equally likely. A deck of cards has within it uniform distributions because the likelihood of drawing a heart, a club, a diamond, or a spade is equally likely.

1. Different types of distributions are

* Bernoulli Distribution
* Uniform Distribution
* Binomial Distribution
* Normal Distribution
* Poisson Distribution
* Exponential Distribution

1. if it is a mirror image about its middle then we can call is symmetric dataset
2. So there are different mathematical transformations that can be done on the dataset to check if it can be converted to symmetry, two popular ones are normalization and standardization.

Percentile Linearization, Log Max Root(log normal transformation), Percentile Linearization are also ways in which we can overcome the skewness

1. Skew = 3 \* (Mean – Median) / Standard Deviation.
2. No I haven’t but I can try to explain what it means,

It is similar to the normal distribution except that tail heaviness is determined by a parameter of the T distribution called degrees of freedom, with smaller values giving heavier tails, and with higher values making the T distribution resemble a standard normal distribution with a mean of 0, and a standard deviation of 1. The T distribution is also known as "Student's T Distribution."

1. It's the science of collecting, exploring and presenting large amounts of data to discover underlying patterns and trends. Statistics are applied every day – in research, industry and government – to become more scientific about decisions that need to be made.

Its like making a pie chart, graph, line chart and other visual representations of data to come to a conclusion about an assertion. I haven’t done in any projects although will implement it very soon

1. a fixed number (n) of trials. each trial must be independent of the others. each trial has just two possible outcomes
2. p(x=10) = ((50/3) power 10 \* e -(50/3)) / 10! = 0.026332
3. There is no existing condition provided to answer this question. Like saying how many times the person cracked the interview in his history.
4. Its just another name for normal distribution where the data is surrounded around the mean of the data and there are standard deviations present where each step in either direction increases the probability of data being inside like 68%, 87% 97% and so on.
5. Each step from across the mean accords to one standard deviation
6. Variance is a measure of dispersion, meaning it is a measure of how far a set of numbers is spread out from their average value. Its calculated by finding

Standard deviation = square root of the variance

1. High variance tends to overfitting the model.

* Try getting additional features, you are generalizing the datasets
* Try adding polynomial features, make the model more complicated.

1. A low bias and high variance problem is overfitting. A high bias and low variance problem is underfitting.
2. We use the normal distribution approach where mean = 0 and std. = 1
3. Z score tells how far away a particular point is away from the mean. This leads to the question as to what the difference between Z score and standard deviation could be. Standard deviation is a reflection of the amount of variability within a given data set, so it talks about the whole data set in general.
4. (1/ sigma \* root(2 \* pi) )\* e^ (-1/2)(x-mu/sigma)^2

Where sigma = standard deviation

Mu = mean

1. The critical region is the region of values that corresponds to the rejection of the null hypothesis at some chosen probability level.
2. A/B testing is basically same as null and alternate hypothesis where u make an assumption about a population by deriving a certain sample parameter (like mean, median or anything else). Then u test for significance of that metric u just calculated by performing anyone of the tests and seeing the results. If the significance value (p value) Is greater than the specified value then the null hypothesis is rejected and the opposite is true as well.
3. The null and alternative hypotheses are two mutually exclusive statements about a population.

So no we cannot mean the both to be the same.

1. Confusion matrix gives values such as TP,TN,FP & FN. But the methods to do it in multi classification statement in slightly different

TP (True positive) The actual value and predicted value should be the same

FN (False negative) - The False-negative value for a class will be the sum of values of corresponding rows except for the TP value. So if it is a row of values, all values (of that row) will be chose except TP.

FP The sum of values of corresponding column except the TP value.

TN The sum of values of all columns and row except the values of that class that we are calculating the values for.

1. False negative – if someone had cancer but the model predicted that he did not have. Type 2 error. This could prove to be fatal.
2. Precision = TP/(TP+FP) 🡪 Here we focus on minimizing FP

Recall = TP(TP+FN) 🡪 Here we focus on minimizing FN

F-Measure provides a way to combine both precision and recall into a single measure that captures both properties

Formula, 2 \* (p \* r) / (p \* r)

where,

P = precision

R = recall

1. I’m not sure of the answer but I’d say something like this.

* We need to have a written commitment from the client that the data format must be agreed upon so that u don’t end up facing issues with your model (consistency)
* Start and end date of the project and setting clear goals in the start.

1. a probability density function that is used especially in analysis of variance and is a function of the ratio of two independent random variables each of which has a chi-square distribution and is divided by its number of degrees of freedom

v1/u1 / v2/u2

where v1 – Independent random variable

u1 – degree of freedom

this is for 1 chi square test result

1. The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems. It is a probability curve that plots the TPR against FPR at various threshold values and essentially separates the ‘signal’ from the ‘noise’. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.

lemme get a few terms right here, TPR = True positive rate

and FPR = False positive rate

True positive rateSensitivity formula

False positive rate = 1 - Specificity formula

AUC value of closer to 1 is desirable

1. It actually depends on the type of problem being solved at hand.

If it classification I’ll go with confusion matrix, AUC-ROC Curve.

If it is regression ill go with MSE (Mean squared error), MAE (Mean absolute error) or even RMSE (Root MSE)

1. 1 tailed test A one-tailed test is a statistical test in which the critical area of a distribution is one-sided so that it is either greater than or less than a certain value

2 tailed test corresponds to checking both ends of the bell curve. Like you’ll ensure that if sample mean Is within the range of the p value specified (not too high, not too low)

1. The power of a test is the probability of rejecting the null hypothesis when it is false; in other words, it is the probability of avoiding a type II error. The power may also be thought of as the likelihood that a particular study will detect a deviation from the null hypothesis given that one exists
2. This information is provided from the domain expert.
3. t-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related in certain features. The t-test is one of many tests used for the purpose of hypothesis testing in statistics.
4. Statistical models are more of a concept which you check assumptions against a sample of a population to derive insights. So answer would be that if we parameterize our data it would be possible to productionize out statistical model
5. Every time there is a new data incoming which is of sufficient threshold we can productionize the model
6. T-tests are used when comparing the results for two classes in particular. ANOVA and MANOVA measures are used to equate the means of the two classes.

In a Z-test, the distribution of the sample is presumed to be natural. A z-score is determined using population parameters such as "population mean" and "population standard deviation" and is used to verify an inference that the sample being drawn belongs to the same population.

For evaluating categorical variables, the Chi-square method is used. It has 2 types.

The goodness of fit test determines whether a sample matches the population.

With two independent variables, a chi-square fit test is used to evaluate two variables in a contingency table to determine if the data matches.

1. Sensitivity is defined as the true-positive recognition rate: number of true positives / (number of true positives + number of false negatives)

Sensitivity analysis provides an approach to quantifying the relationship between model performance and dataset size for a given model and prediction problem. How to perform a sensitivity analysis of dataset size and interpret the results.

1. If you have plenty of computational resources, you can test multiple algorithms and parameter settings. In this approach, the main question is how to estimate and compare the performance of the algorithms in a reliable way.
2. Logistic regression is a calculation used to predict a binary outcome: either something happens, or does not. This can be exhibited as Yes/No, Pass/Fail, Alive/Dead, etc.

So we cannot use it for multiple classes. I would suggest KNN algorithm for such an application

1. K means clustering and data set would need latitude and longitude value
2. I don’t know the answer to this question
3. I don’t know the answer to this question
4. We can maybe develop a CV solution which will describe the objects in close proximity through hearing aid.
5. R square helps us to predict how well a particular point fits in a line. Adjusted R square also does the same but with one difference. It penalizes u if the a useless point is added and hence decreases the R square value

R2 increases with every predictor added to a model. As R2 always increases and never decreases, it can appear to be a better fit with the more terms you add to the model. This can be completely misleading.

Similarly, if your model has too many terms and too many high-order polynomials you can run into the problem of over-fitting the data. When you over-fit data, a misleadingly high R2 value can lead to misleading projections.

1. We apply regularization mainly if the model is overfitting (doing well in the train data but extremely poor score in the test data). It is a term added to mean squared error to increase the penalty of extreme values.

A picture containing text, clock, watch

Description automatically generated

The above is an example of L2 regularization

L1 (Lasso) regression would have the same equation except for squaring the final term we take an absolute value.

Lambda is a tuning parameter tweaking it would result in varying levels of penalty.

1. Homoscedasticity means that the residuals have constant variance no matter the level of the dependent variable.

This means that the error data that is present all of them have same level of variance

Multicollinearity exists whenever an independent variable is highly correlated with one or more of the other independent variables in a multiple regression equation. Multicollinearity is a problem because it undermines the statistical significance of an independent variable.

1. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y, denoted E(y | x), and has been used to describe nonlinear phenomena

Equation 🡪 y= b0+b1x + b2x2+ b3x3+....+ bnxn

The main steps involved in Polynomial Regression are given below:

* Data Pre-processing
* Build a Linear Regression model and fit it to the dataset
* Build a Polynomial Regression model and fit it to the dataset
* Visualize the result for Linear Regression and Polynomial Regression model.
* Predicting the output.

We can keep increase the degrees of freedom to check If the line will fit the model or not

1. I’m not aware of the answer to this.

Maybe we can use NLP methods to differentiate between each type of speech

1. Multiple linear regression is a regression model that estimates the relationship between a quantitative dependent variable and two or more independent variables using a straight line.

Multiple regressions are based on the assumption that there is a linear relationship between both the dependent and independent variables. It also assumes no major correlation between the independent variables.

In the case of 5 features it will draw 5 statistical lines over the input (dependent features) for the independent feature.

1. I know very few clustering algorithms but after searching in google I got a basic understanding of a few of them like,

* DBSCAN
* K-Means
* Gaussian Mixture Model

1. Three important factors by which clustering can be evaluated are
2. Clustering tendency (b) Number of clusters, k (c) Clustering quality

Something which comes to my mind about clustering methods is,

Elbow method:-

Within-cluster variance is a measure of compactness of the cluster. Lower the value of within cluster variance, higher the compactness of cluster formed.

1. Random state ensures that the splits that you generate are reproducible. Scikit-learn uses random permutations to generate the splits. The random state that you provide is used as a seed to the random number generator. This ensures that the random numbers are generated in the same order
2. To the best of my knowledge,

U must check for null values first, get that sorted by either dropping the column or replacing with mean/median/mode values.

Then u could check the correlation between the independent features and ensure that each variable is In the right scale. (standardization and normalization)

So, data cleaning 🡪 feature engineering 🡪 feature selection 🡪 model building 🡪 Evaluating the model.

Then I believe u can check for the best fit model for the data and test it maybe using the ensemble techniques

1. I have not sir, looking forward to ineurons machine learning internship
2. Machine Learning is a set of methods which enables the computer to take decisions or infer conclusions without us guiding it.

“So if we don’t guide the computer, how does it learn ?”

Just like a human, a computer can learn from three sources. One is Observing what others did in similar situations. The other is observing a situation and trying to come up with best possible logic on the spot to decide/conclude . The third is learning from previous mistakes/success . These three methods correspond to three branches of Machine learning, Supervised, Unsupervised and Reinforcement learning respectively

1. To the best of my limited knowledge, this is how I would answer..

Where are we installing this surveillance system ? What is the area being covered ? What is the method of detection of intrusion ? Is it only through cameras or motion detector or heat sensor or laser beam ? Is it surveilling only at a particular time or its it 24/7 ? How are we gathering the data for this project ? What is the format of data ?

These are the initial questions I would ask the client to ensure that we are on the same page.

Then