1. **Where you have used Hypothesis Testing in your Machine learning Solution.**

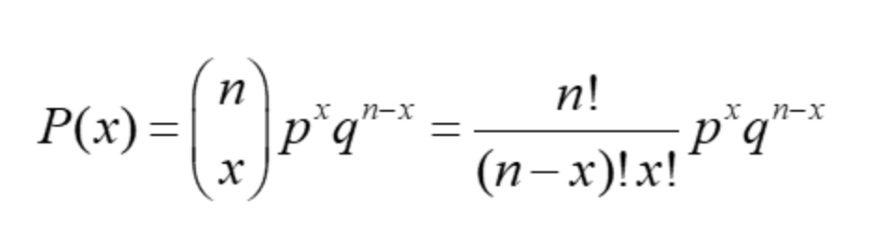
Suppose you are working on a machine learning project, for which you want to predict if a set of patients have or not a mortal disease, based on several features on your dataset as blood pressure, heart rate, pulse and others.

Sounds like a serious project, for which you’ll need to really trust your model and predictions, right? That’s why you got hundreds of samples, that your local hospital very gently allowed you to collect, given the importance and the seriousness of the topic. But how do you know if your sample is representative of the whole population? And how can we know how much difference might be reasonable? For example, assume that thanks to some previous studies we know that the real probability for any given patient of not having this particular disease is 99%. Now suppose that our sample says that 95% of the patients don’t have the disease. Well, 4% difference doesn’t sound like a significant difference that may lead us to SUCH bad modelling, right? It might not be the same, but it kind of sounds like it may be representative. To confirm this, we need to build a better understanding of the theoretical background.

Let’s start by what we know…the real probability of not having the disease:

P (not having the disease) = 99%

Now let’s assume that we find a new group of 100 people and we test all of them to check if any has this disease we’re studying. Can we be sure that 99 of these folks won’t have the disease? Maybe, but there’s also a possibility that none of them has the disease, or even that several may have it. What we have here, is a binomial probability problem. The objective of this story is not to talk about probabilities, however, in simple words, the binomial probability is no more than a given chance of something happening a fixed number of times, given a prior probability for each independent event. We can find it by just applying the following equation:



Where:

* n = the number of trials (or the number being sampled)
* x = the number of successes desired
* p = probability of getting a success in one trial
* q = 1 — p = the probability of getting a failure in one trial

So if we want to know what is the probability that in our sample of 100 guys, we don’t have any of them infected with the disease, we may just fill in the blanks to find that the probability is of 36.6%. And if we want to know the probability of having 99% folks NOT infected, we fill in the blanks again, to find out it is approximately 37.0%. And this sounds reasonable: getting 100 out of 100 not infected doesn’t sound very unlikely if every single case has 99% of not being infected. And in this line, it also sounds reasonable that having 99 not infected out 100 folks might be a little more likely.

We could keep going and find the probability of having even less people not infected in our sample of 100 people:

* P (not infected 98 out of 100) = 18.5%
* P (not infected 97 out of 100) = 6.0%
* P (not infected 96 out of 100) = 1.5%
* P (not infected 95 out of 100) = 0.3%

Now, let’s go back to the sample we had from our friendly local hospital, which says that 95% of the guys in our sample are NOT infected by this horrible mortal disease. Well, even though it might sound like the difference in between 95% and 99% is not relevant, given that we would not be working with a random sample of folks, but instead these guys belong to the same population that we know has a 99% probability of not being infected, we’d be setting a hypothesis that our sample is representative when in reality we would have only 0.3% chance of obtaining a sample with 95 out of 100 people not infected. Therefore we should reject our hypothesis and not proceed.

1. **What kind of statistical tests you have performed in your ML Application**

**Ans: In statics we have a lot of tests like t-tets, Z-test, anova test, Welch’s test etc. we can’t say in every problem statement and every project we can use same test. No.**

**It’s depends on our dataset, problem statements and goal. we can use and perform.**

**3. What do you understand by P Value? And what is use of it in ML?**

**P-value** helps us determine how likely it is to get a particular result when the [null hypothesis](https://en.wikipedia.org/wiki/Null_hypothesis) is assumed to be true. It is the probability of getting a sample like ours or more extreme than ours if the null hypothesis is correct. Therefore, if the null hypothesis is assumed to be true, the p-value gives us an estimate of how “strange” our sample is.

If the p-value is very small (<0.05 is considered generally), then our sample is “strange,” and this means that our assumption that the null hypothesis is correct is most likely to be false. Thus, we reject it.

**When and how is p-value is used?**

p-values are often reported whenever you perform a [statistical significance test](https://www.machinelearningplus.com/statistics/statistical-significance-tests-r/) (like t-test, chi-square test etc). These tests typically return a computed test statistic and the associated p-value. This reported value is used to establish the statistical significance of the relationships being tested.

So, whenever you see a p-value, there is an associated statistical test.

That means, there is a Hypothesis testing being conducted with a defined Null Hypothesis (H0) and a corresponding Alternate hypothesis (HA).

The p-value reported is used to make a decision on whether the null hypothesis being tested can be rejected or not.

**Let’s understand a little bit more about the null and alternate hypothesis.**

Now, how to frame a Null hypothesis in general?

While the null hypothesis itself changes with every statistical test, there is a general principle to frame it:

**The null hypothesis assumes there is ‘no effect’ or ‘relationship’ by default**.

**Significance of P-values:**

* If p > 0.10 : the observed difference is “not significant”
* If p ≤ 0.10 : the observed difference is “marginally significant”
* If p ≤ 0.05 : the observed difference is “significant”
* If p ≤ 0.01 : the observed difference is “highly significant.”

**For example**: if you are testing if a drug treatment is effective or not, then the null hypothesis will assume there is not difference in outcome between the treated and untreated groups. Likewise, if you are testing if one variable influence another (say, car weight influences the mileage), then null hypothesis will postulate there is no relationship between the two.

**4. Which type of error is severe Error, Type 1 or Type 2? And why with example.**

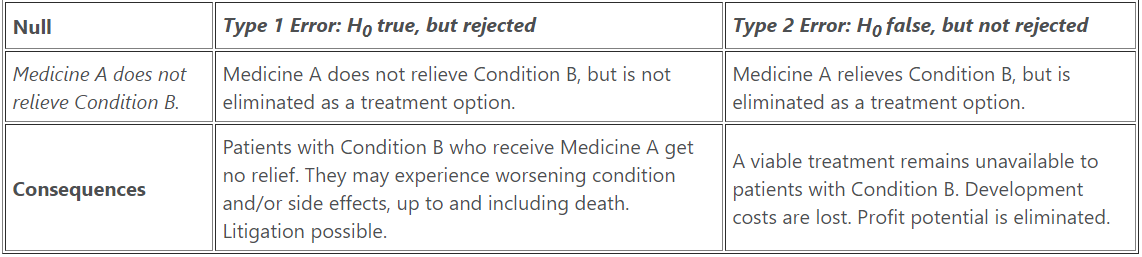
**Ans: It’s depends, based on the researcher and instructor** Type 1 (false positive) is worse than a Type 2 (false negative) error. The rationale boils down to the idea that if you stick to the status quo or default assumption, at least you're not making things worse. And in many cases, that's true.  But like so much in statistics, in application it's not really so black or white. The analogy of the defendant is great for teaching the concept, but when we try to make it a rule of thumb for which type of error is worse in practice, it falls.

In one instance, the Type I error may have consequences that are less acceptable than those from a Type II error. In another, the Type II error could be less costly than a Type I error. And sometimes, as Dan Smith pointed out in [*Significance*](http://magazine.amstat.org/blog/2013/11/01/mathmyopia/) a few years back with respect to Six Sigma and quality improvement, "neither" is the only answer to which error is worse:

Most Six Sigma students are going to use the skills they learn in the context of business. In business, whether we cost a company $3 million by suggesting an alternative process when there is nothing wrong with the current process or we fail to realize $3 million in gains when we should switch to a new process but fail to do so, the end result is the same. The company failed to capture $3 million in additional revenue.

**POTENTIAL CONSEQUENCES**

Since there's not a clear rule of thumb about whether Type 1 or Type 2 errors are worse, our best option when using data to test a hypothesis is to look very carefully at the fallout that might follow both kinds of errors. Several experts suggest using a table like the one below to detail the consequences for a Type 1 and a Type 2 error in your particular analysis.

****

**5. Where we can use chi square and have used this test anywhere in your application**

A chi-square test is a statistical test used to compare observed results with expected results. The purpose of this test is to determine if a difference between observed data and expected data is due to chance, or if it is due to a relationship between the variables you are studying. Therefore, a chi-square test is an excellent choice to help us better understand and interpret the relationship between our two categorical variables.

Use cases example:

A research scholar is interested in the relationship between the placement of students in the statistics department of a reputed University and their C.G.P.A (their final assessment score).

He obtains the placement records of the past five years from the placement cell database (at random). He records how many students who got placed fell into each of the following C.G.P.A. categories – 9-10, 8-9, 7-8, 6-7, and below 6.

If there is no relationship between the placement rate and the C.G.P.A., then the placed students should be equally spread across the different C.G.P.A. categories (i.e. there should be similar numbers of placed students in each category).

However, if students having C.G.P.A more than 8 are more likely to get placed, then there would be a large number of placed students in the higher C.G.P.A. categories as compared to the lower C.G.P.A. categories. In this case, the data collected would make up the observed frequencies.

**So the question is, are these frequencies being observed by chance or do they follow some pattern?**

Here enters the chi-square test! **The chi-square test helps us answer the above question by comparing the observed frequencies to the frequencies that we might expect to obtain purely by chance.**

**Reference:** <https://www.analyticsvidhya.com/blog/2019/11/what-is-chi-square-test-how-it-works/>

**6. Can we use Chi square with Numerical dataset? If yes, give example. If no, give Reason?**

**7. What do you understand by ANOVA Testing?**

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.

We can use ANOVA to prove/disprove if all the medication treatments were equally effective or not.

In simpler and general terms, it can be stated that the ANOVA test is used to identify which process, among all the other processes, is better. The fundamental concept behind the Analysis of Variance is the “[Linear Model](https://www.analyticssteps.com/blogs/multiple-linear-regression)”.

## ****Example of ANOVA****

An example to understand this can be prescribing medicines.

* Suppose, there is a group of patients who are suffering from fever.
* They are being given three different medicines that have the same functionality i.e., to cure fever.
* To understand the effectiveness of each medicine and choose the best among them, the ANOVA test is used.

You may wonder that a t-test can also be used instead of using the ANOVA test. You are probably right, but, since t-tests are used to compare only two things, you will have to run multiple t-tests to come up with an outcome. While that is not the case with the ANOVA test.

That is why the ANOVA test is also reckoned as an extension of t-test and z-tests.

## ****Types of ANOVA Test****

The ANOVA test is generally done in three ways depending on the number of Independent Variables (IVs) included in the test. Sometimes the test includes one IV, sometimes it has two IVs, and sometimes the test may include multiple IVs.

We have three known types of ANOVA test:

1. **One-Way ANOVA**
2. **Two-Way ANOVA**
3. **N-Way ANOVA (MANOVA)**

**Example:** Suppose medical researchers want to find the best diabetes medicine and they have to choose from four medicines. They can choose 20 patients and give them each of the four medicines for four months.

The researchers can take note of the sugar levels before and after medication for each medicine and then to understand whether there is a statistically significant difference in the mean results from the medications, they can use one-way ANOVA.

The type of medicine can be a factor and reduction in sugar level can be considered the response. Researchers can then calculate the p-value and compare if they are lower than the significance level.

If the results reveal that there is a statistically significant difference in mean sugar level reductions caused by the four medicines, the [post hoc tests](https://www.statology.org/anova-post-hoc-tests/) can be run further to determine which medicine led to this result.

**8. Give me a scenario where you can use Z test and T test.**

z-tests are used when we have large sample sizes **(n > 30)**, whereas t-tests are most helpful with a smaller sample size (n < 30). Both methods assume a normal distribution of the data, but the z-tests are most useful when the standard deviation is known.

Z-test is the statistical test, used to analyze whether two population means are different or not when the variances are known and the sample size is large.

This test statistic is assumed to have a normal distribution, and standard deviation must be known to perform an accurate z-test.

*A z-statistic, or z-score, is a number representing the value’s relationship to the mean of a group of values, it is measured with population parameters such as population standard deviation and used to validate a hypothesis*.

For example, the null hypothesis is “sample mean is the same as the population mean”, and alternative hypothesis is “the sample mean  is not the same as the population mean”.

**T-test:**

In order to know how significant the difference between two groups are, T-test is used, basically it tells that difference (measured in means) between two separate groups could have occurred by chance.

This test assumes to have a normal distribution while based on t-distribution, and population parameters such as mean, or standard deviation are unknown.

The ratio between the difference between two groups and the difference within the group is known as T-score. Greater is the t-score, more is the difference between groups, and smaller is the t-score, more similarities are there among groups.

For example, a t-score value of 2 indicates that the groups are two times as different from each other as they are with each other.

Also, after running t-test, if the larger t-value is obtained, it is highly likely that the outcomes are more repeatable, such that

* A larger t-score states that groups are different
* A smaller t-score states that groups are similar.

Mainly, there are three types of t-test:

1. **An Independent Sample t-test,**compare the means for two groups.
2. **A Paired Sample t-test,**compare means from the same group but at different times, such as six months apart.
3. **A One Sample t-test,** test a mean of a group against the known mean.

**9. What do you understand by inferential Statistics?**

· Inferential statistics is work with a random sample of data taken from a population to illustrate and make inferences about the population.

· Inferential statistics are valuable when working with of each member of an entire population is not convenient or possible.

· It’s help us get to the conclusions and make predictions based on our data.

· Inferential statistics understands the whole population from sample taken from it.

· In Inferential statistics we use a random sample, so we can generalize outcome from the sample to the large population.

· In Inferential statistics, we can calculate the mean, standard deviation, and proportion for our random sample data from population.

The following types of inferential statistics are mostly used and quite easy to interpret:

· Conditional Probability

· Probability Distribution and Distribution function

· Probability

· Regression Analysis

· Central Limit Theorem

· Hypothesis Testing

· T- Test

· Z- Test

· Sampling Distribution

· Chi-square test

· Confidence Interval

· ANOVA (Analysis of variance)

**10. When you are trying to calculate Std Deviation or Variance, why you used N-1 in Denominator?**

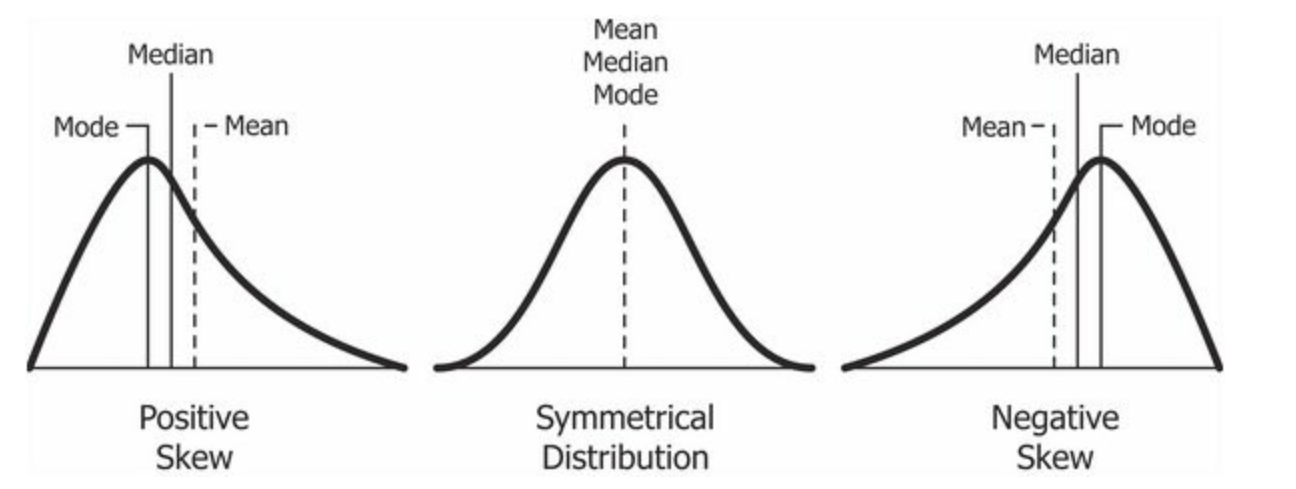
The n-1 equation is used in the common situation where you are analyzing a sample of data and wish to make more general conclusions. The SD computed this way (with n-1 in the denominator) is your best guess for the value of the SD in the overall population.

If you simply want to quantify the variation in a particular set of data, and don't plan to extrapolate to make wider conclusions, then you can compute the SD using n in the denominator. The resulting SD is the SD of those particular values. It makes no sense to compute the SD this way if you want to estimate the SD of the population from which those points were drawn. It only makes sense to use n in the denominator when there is no sampling from a population, there is no desire to make general conclusions.

**11. What do you understand by right skewness, Give example?**

**Ans:**  skewness is the measure of how much the probability distribution of a random variable deviates from the [normal distribution](https://www.analyticsvidhya.com/blog/2020/04/statistics-data-science-normal-distribution/?utm_source=blog&utm_medium=what-is-skewness-statistics). Now, you might be thinking – why am I talking about normal distribution here?

Well, the normal distribution is the probability distribution without any skewness. You can look at the image below which shows symmetrical distribution that’s basically a normal distribution and you can see that it is symmetrical on both sides of the dashed line. Apart from this, there are two types of skewness:



The probability distribution with its tail on the right side is a positively skewed distribution and the one with its tail on the left side is a negatively skewed distribution. If you’re finding the above figures confusing, that’s alright. We’ll understand this in more detail later.

**12. What is difference between Normal distribution and Std Normal Distribution and Uniform Distribution?**

The only thing similar about the two is that they are both continuous distributions with two parameters. Differences include:

1. Normal has infinite support, uniform has finite support
2. Normal has a single most likely value, uniform has every allowable value equally likely
3. Uniform has a piecewise constant density, normal has a continuous bell-shaped density
4. Normal distributions arise from the central limit theorem, uniforms do not.

**13. What is different kind of Probabilistic distributions you heard of?**

**Probability:** Simply put, probability is an intuitive concept. We use it on a daily basis without necessarily realising that we are speaking and applying probability to work.

Life is full of uncertainties. We don’t know the outcomes of a particular situation until it happens. Will it rain today? Will I pass the next math test? Will my favourite team win the toss? Will I get a promotion in next 6 months? All these questions are examples of uncertain situations we live in. Let us map them to few common terminologies which we will use going forward.

**Experiment** – are the uncertain situations, which could have multiple outcomes. Whether it rains on a daily basis is an experiment.

**Outcome** is the result of a single trial. So, if it rains today, the outcome of today’s trial from the experiment is “It rained”

**Event** is one or more outcome from an experiment. “It rained” is one of the possible events for this experiment.

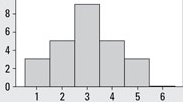
**Probability** is a measure of how likely an event is. So, if it is 60% chance that it will rain tomorrow, the probability of Outcome “it rained” for tomorrow is 0.6

**Types of Distributions**

* 1. Bernoulli Distribution
  2. Uniform Distribution
  3. Binomial Distribution
  4. Normal Distribution
  5. Poisson Distribution
  6. Exponential Distribution

**14. What do you understand by symmetric dataset?**

If the data are symmetric, **they have about the same shape on either side of the middle**. In other words, if you fold the histogram in half, it looks about the same on both sides. Below figure shows an example of symmetric data. With symmetric data, the mean and median are close together.

****

**15. In your last project, were you using symmetric data or Asymmetric Data, if its asymmetric, what kind of EDA**

**you have performed?**

**Ans: Recently I have done one Chatbot project, here I have dataset in the from of CSV. During the EDA we have some lots of steps because our dataset is a imbalanced and in our dataset we have lots of outliers.**

* 1. **Here, we are plotting a Box plot and scatter plot for checking a outlier.**
  2. **We are plotting a count plot for checking a how many total values available in our every features.**
  3. **As we know our dataset is imbalanced so here we are used a under sampling for handling a imbalanced datasets.**
  4. **Before handling a imbalanced data and outliers we are getting 43% accuracy but after handling this we are getting 83% using DistilBERT.**