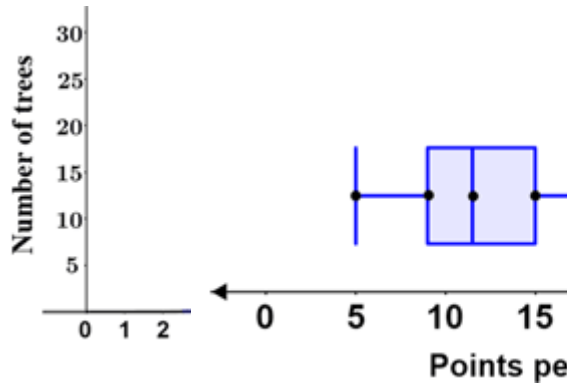
Statistics

Statistics is a discipline that involves collecting, organizing, displaying, analyzing, interpreting, and presenting data. It is widely used in scientific research, when considering social problems, and for industrial purposes, among many other applications.
On a base level, it involves proper data collection through sampling when population data is not known or cannot be determined, designing and conducting experimental and observational studies, and formulating conclusions or re-designing the studies based on the data. Two distinct branches of statistics are descriptive statistics and inferential statistics.

Descriptive statistics

Descriptive statistics is the branch of statistics concerned with summarizing data, be it in graphical, tabular, or some other form. A descriptive statistic is a summary statistic used to describe data. Examples of well-known descriptive statistics include the mean, median, and mode; these are classified as measures of central tendency and are one of the key types of descriptive statistics that provide information about a central or typical value in a probability distribution. Measures of variability are another classification of descriptive statistic; they describe the spread of the data (how stretched or squeezed the distribution is) and include statistics such as the standard deviation, variance, and more.
The figure below shows two types of figures used to depict descriptive statistics.

Histogram Box-and-whisker plot



In particular, the histogram and the curve fitted to it indicate a normal distribution, which is a commonly encountered probability distribution throughout statistics. Many natural phenomena exhibit a normal distribution, giving way to inferential statistics, which allows us to make inferences about data based on their probability distributions as well as other factors.

Inferential statistics

In the real world, it is often not possible, or highly impractical to collect large amounts of data from populations of interest. Ideally, we would be able to acquire all the data we need for a population and make informed decisions based on the descriptive statistics they provide.

Realistically, since this is rarely feasible, we instead make inferences about populations as a whole based on samples of said populations and the use of statistical methods; this is the goal of inferential statistics.
For example, we may want to know the mean score on the AP Physics exam for all high school students in the United States. Because of the large scale, it would be both difficult and expensive to obtain the results of every single student in the US. In such a case, inferential statistics can be used to estimate the mean score by collecting samples from the population of high school students, then using the sample data to make inferences or predictions about the mean score of the population as a whole.
When studying random phenomena, we may want to assess whether any observed differences can be attributed to some given input, or if the observed differences can be attributed fully to random chance. This is another area in which inferential statistics can be used through the process of statistical hypothesis testing. There are many different types of statistical hypothesis tests that can be used depending on the conditions of the experiment. In general, the process involves a statement of no difference, referred to as the null hypothesis, and a comparison of what is observed to what we would expect based on this null hypothesis. Through use of statistical methods, we can then draw conclusions about the significance of observed data.

Hypothesis Testing

Hypothesis testing is the backbone behind statistical inference and can be broken down into a couple of topics.
The first is the Central Limit Theorem, which plays an important role in studying large samples of data.
Other core elements of hypothesis testing are:

1. sampling distributions,
2. p-values,
3. confidence intervals,
4. and type I and II errors.

Lastly, it is worth looking at various tests involving proportions, and other hypothesis tests.

Most of these concepts play a crucial role in A/B testing, which is a commonly asked topic during interviews at consumer-tech companies like Facebook, Amazon, and Uber.

It's useful to not only understand the technical details but also conceptually how A/B testing operates, what the assumptions are, possible pitfalls, and applications to real-life products.

Hypothesis testing is the use of statistics to determine the probability that a given hypothesis is true.

- A hypothesis is a statement about something you are observing.
- Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data.
- A way of determining whether a pattern has occurred by chance is by performing a hypothesis test.
- The test provides evidence concerning the plausibility of the hypothesis, given the data.

- Statistical analysts test a hypothesis by measuring and examining a random sample of the population being analysed.

The hypothesis has to be substantiated with evidence provided from the data.
Statistical hypothesis tests provide a quantitative way of substantiating the belief or rejecting or modifying the original hypothesis.
In other words, a hypothesis test is a quantitative approach to determine whether your speculation can be substantiated.

The strength of the evidence can be measured and you can decide on whether or not to reject the hypothesis based on some risk measure, the risk that your decision may be incorrect.

The non-existence of any difference is your null hypothesis.

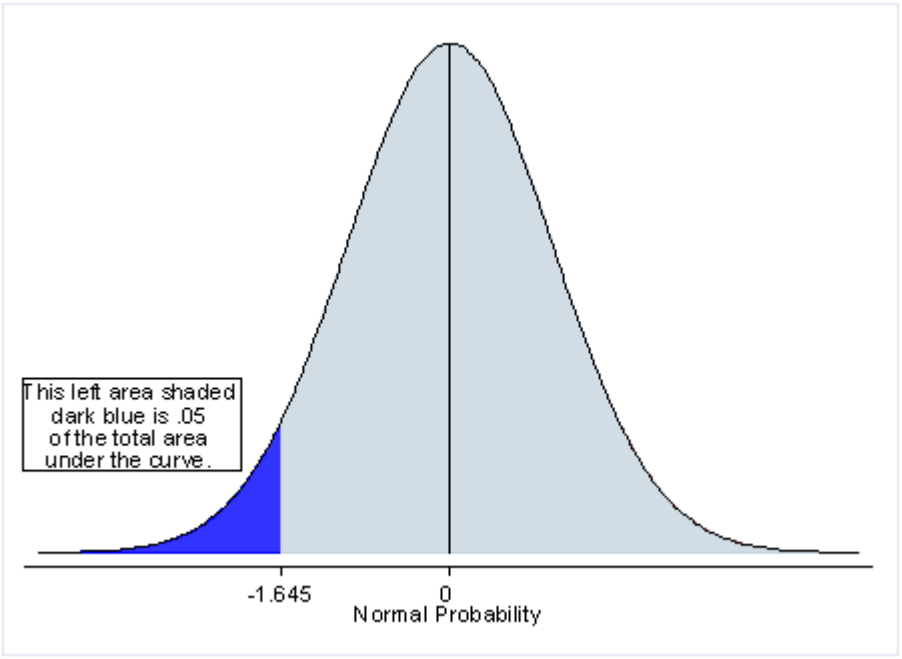
If there is evidence of a significant change, you can reject the prior belief, i.e., you can reject the null hypothesis for the alternative.

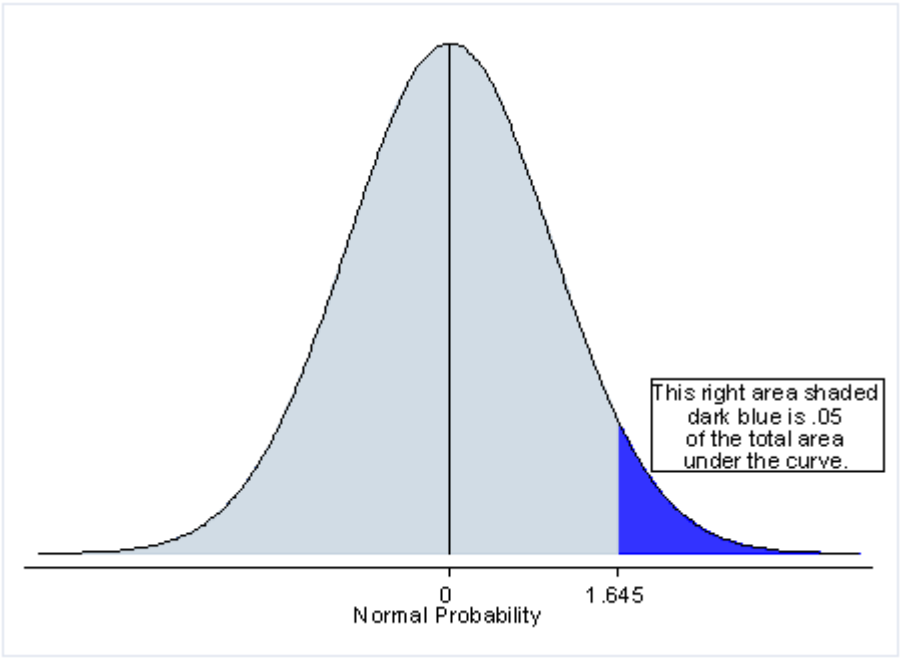
If there is no significant change, then, we cannot reject the null hypothesis. We continue to believe that the odd differences that were observed may be due to chance. This way, the change may be linked to its cause in a reverse fashion. Anyone who uses this data and the testing method should arrive at the same result.

“Beyond a reasonable doubt” is our credo.
Read More
Video

We are basing our decisions on the probability calculated from the sample data given a null hypothesis, we should say we are proving by low-probability.

Simple classification of the various types of hypothesis tests: one-sample tests and two or more sample tests.
What is a one-tailed test?
Next, let’s discuss the meaning of a one-tailed test. If you are using a significance level of .05, a one-tailed test allots all of your alpha to testing the statistical significance in the one direction of interest. This means that .05 is in one tail of the distribution of your test statistic. When using a one-tailed test, you are testing for the possibility of the relationship in one direction and completely disregarding the possibility of a relationship in the other direction. Let’s return to our example comparing the mean of a sample to a given value x using a t-test. Our null hypothesis is that the mean is equal to x. A one-tailed test will test either if the mean is significantly greater than x or if the mean is significantly less than x, but not both. Then, depending on the chosen tail, the mean is significantly greater than or less than x if the test statistic is in the top 5% of its probability distribution or bottom 5% of its probability distribution, resulting in a p-value less than 0.05. The one-tailed test provides more power to detect an effect in one direction by not testing the effect in the other direction. A discussion of when this is an appropriate option follows.





When is a one-tailed test appropriate?

Because the one-tailed test provides more power to detect an effect, you may be tempted to use a one-tailed test whenever you have a hypothesis about the direction of an effect. Before doing so, consider the consequences of missing an effect in the other direction. Imagine you have developed a new drug that you believe is an improvement over an existing drug. You wish to maximise your ability to detect the improvement, so you opt for a one-tailed test. In doing so, you fail to test for the possibility that the new drug is less effective than the existing drug. The consequences in this example are extreme, but they illustrate a danger of inappropriate use of a one-tailed test.

So when is a one-tailed test appropriate? If you consider the consequences of missing an effect in the untested direction and conclude that they are negligible and in no way

irresponsible or unethical, then you can proceed with a one-tailed test. For example, imagine again that you have developed a new drug. It is cheaper than the existing drug and, you believe, no less effective. In testing this drug, you are only interested in testing if it less effective than the existing drug. You do not care if it is significantly more effective. You only wish to show that it is not less effective. In this scenario, a one-tailed test would be appropriate.

When is a one-tailed test NOT appropriate?

Choosing a one-tailed test for the sole purpose of attaining significance is not appropriate. Choosing a one-tailed test after running a two-tailed test that failed to reject the null hypothesis is not appropriate, no matter how "close" to significant the two-tailed test was. Using statistical tests inappropriately can lead to invalid results that are not replicable and highly questionable–a steep price to pay for a significant star in your results table!

A one-sample hypothesis is a statement about the parameter of the population; or, it is a statement about the probability distribution of a random variable.

Our discussion today is on whether or not a certain proportion of subjects taking the memory-boosting mocha improve their memory. The test is to see if this proportion is significantly different from 50%. We are verifying whether the parameter (proportion, p) is equal to or different from 50%. So it is a one-sample hypothesis test.

The value that we compare the parameter on can be based on experience or knowledge of the process, based on some theory, or based on some design considerations or obligations. If it is based on experience or prior knowledge of the process, then we are verifying whether or not the parameter has changed.

If it is based on some theory, then we are testing the theory. Our coffee example will fall under this criterion.

We know that random chance means a 50% probability of improving (or not) the memory. So we test the proportion against this model; p = 0.5. If the parameter is compared against a value based on some design consideration or obligation, then we are testing for compliance.

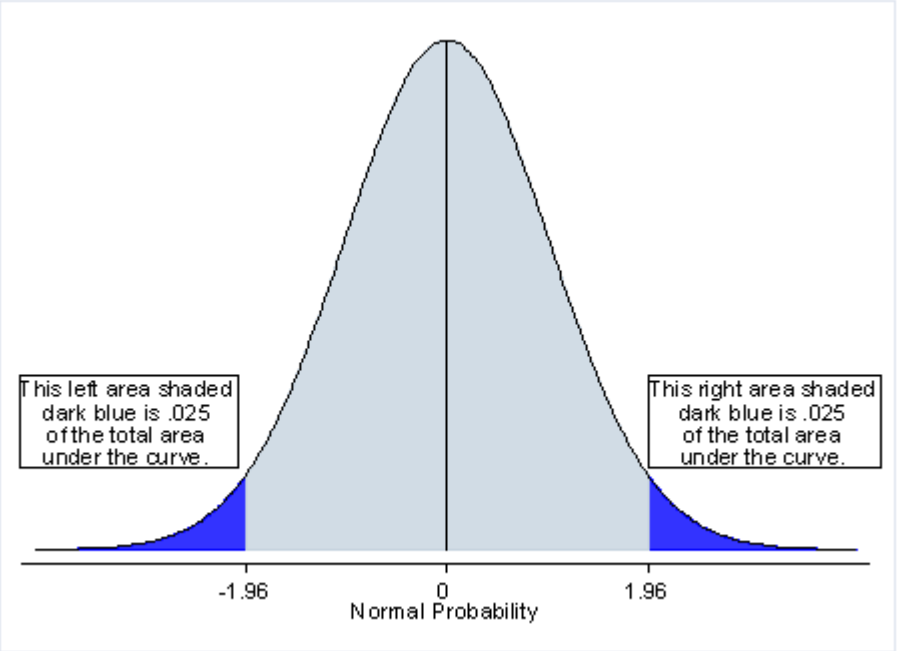
What is a two-tailed test?

Sometimes, we have to test one sample against another sample. For example, people who take the memory-boosting test from New York City may be compared with people taking the test from San Francisco.

This type of test is a two or multiple sample hypothesis test where we determine whether a random variable differs in its parameter among the two or more groups.

First let's start with the meaning of a two-tailed test. If you are using a significance level of 0.05, a two-tailed test allots half of your alpha to testing the statistical significance in one direction and half of your alpha to testing statistical significance in the other direction. This means that .025 is in each tail of the distribution of your test statistic. When using a two-tailed test, regardless of the direction of the relationship you hypothesise, you are testing for the possibility of the relationship in both directions. For example, we may wish to compare the mean of a sample to a given value x using a t-test. Our null hypothesis is that

the mean is equal to x. A two-tailed test will test both if the mean is significantly greater than x and if the mean significantly less than x. The mean is considered significantly different from x if the test statistic is in the top 2.5% or bottom 2.5% of its probability distribution, resulting in a p-value less than 0.05.



Now, for any of these two types, we can further classify them into parametric tests or nonparametric tests.

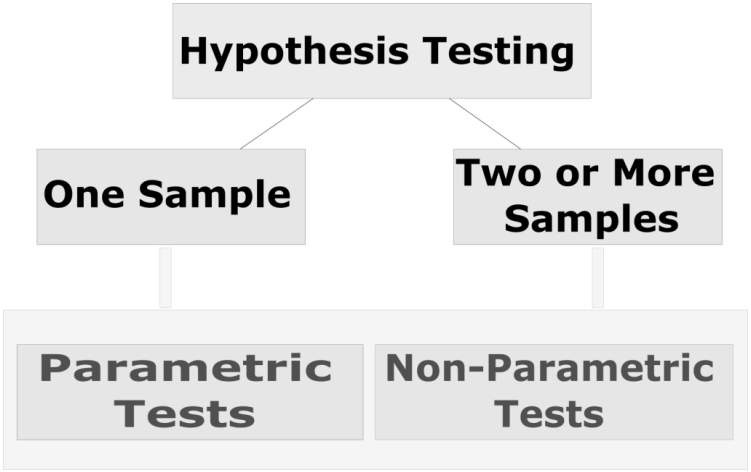
If we assume that the data has a particular probability distribution, the test can be developed based on this probability distribution. These are called parametric tests.

If a probability distribution is appropriate for the data, then, the information contained in the data can be summarised using the parameters of this distribution; like the mean, standard deviation, proportion, etc.
The hypothesis test can be designed using these parameters. The entire process becomes very efficient since we already know the mathematical formulations. In our case, since we are testing for proportion, we can assume a binomial distribution to derive the probabilities.

If we make incorrect assumptions regarding the probability distributions, the parameters that we use to summarise the data are at best, a poor representation of the data, which will result in incorrect conclusions.

There are hypothesis tests that do not require the assumption that the data follow a particular probability distribution.

These types of tests are called nonparametric hypothesis tests. Information is efficiently extracted from the data without summarising them into their statistics or parameters.



We can follow these five steps for any hypothesis test:

1. Choose the appropriate test; one-sample or two-sample and parametric or nonparametric.

2. Establish the null and alternative hypotheses.

3. Decide on an acceptable rate of error or rejection rate (α).

4. Compute the test statistic and its corresponding p-value from the observed data.

5. Make the decision; Reject the null hypothesis if the p-value is less than the acceptable rate of error, α.

If we are comfortable with the assumption of a probability distribution for the data, a parametric test may be used.

If there is little information about the prior process, then it is beneficial to use the nonparametric tests.
Nonparametric tests are also especially appropriate for small data sets.

Example of Distribution
Suppose ten people take the test, the probabilities can be derived from a binomial distribution with n = 10 and p = 0.5.
The null distribution, i.e., what may happen by chance is a binomial distribution with n = 10 and p = 0.5, and we can check how far out on this distribution is our observed proportion.

The null hypothesis (H_{0}) is that p = 0.5
The alternate hypothesis (H_{A}) is that p > 0.5.

The null hypothesis is usually denoted as H_{0}, and the alternate hypothesis is denoted as H_{A}.

The null hypothesis (H_0) is what is assumed to be true before any evidence from data. It is usually the null situation that has to be disproved otherwise. Null has the meaning of “no effect,” or “of no consequence.”
H_0 is identified with the hypothesis of no change from the current belief.

The alternate hypothesis (H_A) is the situation that is anticipated to be true if the data (empirical evidence) shows that the null hypothesis (H_0) is unlikely.

The alternate hypothesis can be of two types, the one-sided alternative or the two-sided alternative.

The two-sided alternative can be considered when evidence in either direction (values larger than or smaller than the accepted level) would cause the rejection of the null hypothesis. The one-sided alternative is considered when the departures in one direction (either less than or greater than) are sufficient to reject H_0.

Our test is a one-sided alternative hypothesis test. The proportion of people who would benefit from the memory-booster coffee is greater than the proportion who would claim benefit randomly.

It is usually the case that the null hypothesis is the favoured claim. The onus of proof is on the alternative, i.e., we will continue to believe in H_0, the status quo unless the experimental evidence strongly contradicts it; proof by low-probability.

We will either reject the null hypothesis or accept the null hypothesis.

We know H_0 is true, but, based on the sample, we had to reject it. We committed an error.

This kind of error is called a Type I error. Let’s call this error, the rejection rate α. There is a certain probability that this will happen, and we select this rejection rate.

Assume $\alpha = 5\%$. A 5% rejection rate implies that we are rejecting the null hypothesis 5% of the times when in fact H_0 is true.

Now, in reality, we will not know whether or not H_0 is true. The choice of α is the risk taken by us for rejecting the truth. If we choose $\alpha = 5\%$, a 5% rejection rate, we choose to reject the null hypothesis 5% of the time.

In hypothesis tests, it is a common practice to set α at 5%. However, α can also be chosen to have a higher or lower rejection rate.

Suppose $\alpha = 1\%$, we will only reject the null hypothesis 1% of the time. There needs to be greater proof to reject the null. If you want to save yourself that extra dollar, you would like to see a greater proof, a lower rejection rate. The coffee shop would perhaps like to choose

$\alpha = 10\%$. They want to reject the null hypothesis more often, so they can show value in their new product.

There is another kind of error, the second type, Type II. It is the probability of not rejecting the null hypothesis when it is false.

Type II error is also called the lack of power in the test.

Some attention to these two Types shows that Type I and Type II errors are inversely related.
If Type I error is high, i.e., if we choose high α, then Type II error will be low. Alternately, if we want a low α value, then Type II error will be high.

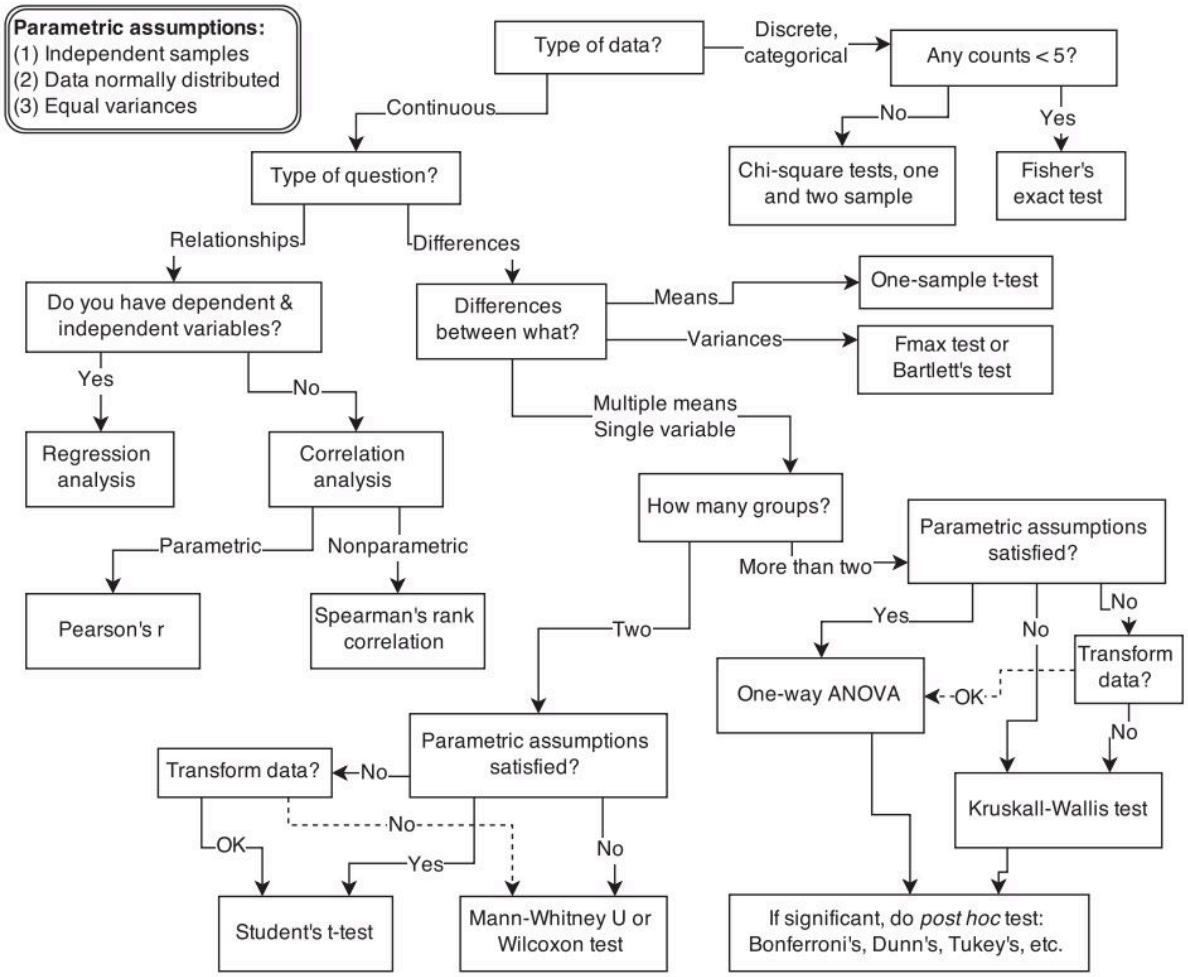
		Truth (unknown)	
		H_0 is True	**H_0 is False**
Test Outcome (our decision)	**Reject H_0**	Error (Type I)	Correct Decision
	Fail to Reject H_0	Correct Decision	Error (Type II)

The test statistic summarises the information in the data. For example, suppose out of ten people who took the test, 9 reported a positive effect, we would take nine as the test statistic, and compute $P(X \geq 9)$ as the p-value.
In a Binomial null distribution with n = 10 and p = 0.5, what is the probability of getting a value that is as large or greater than 9?
If the value has a sufficiently low probability, we cannot say that it may occur by chance.

If this statistic, 9, is not significantly different from what is expected in the null hypothesis, then H_0 cannot be rejected.

The p-value is the probability of obtaining the computed test statistics under the null hypothesis. It is the evidence or lack thereof against the null hypothesis.
The smaller the p-value, the less likely the observed statistic under the null hypothesis – and stronger evidence of rejecting the null.

And remember,
The null hypothesis is never “accepted,” or proven to be true. It is assumed to be true until proven otherwise and is “not rejected” when there is insufficient evidence to do so.



Hypothesis testing for data scientists

The frequentist approach
Hypothesis testing is a common statistical tool used in research and data science to support the certainty of findings. The aim of testing is to answer how probable an apparent effect is detected by chance given a random data sample. This article provides a detailed explanation of the key concepts in Frequentist hypothesis testing using problems from the business domain as examples.

What is a hypothesis?

A hypothesis is often described as an “educated guess” about a specific parameter or population. Once it is defined, one can collect data to determine whether it provides enough evidence that the hypothesis is true.

Hypothesis testing

In hypothesis testing, two mutually exclusive statements about a parameter or population (hypotheses) are evaluated to decide which statement is best supported by sample data.

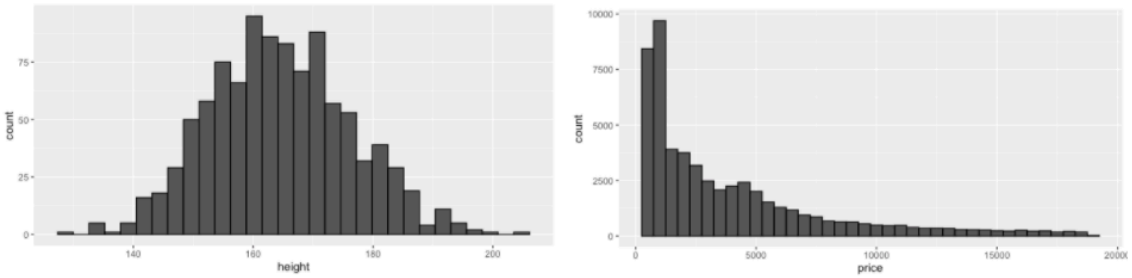
Parameters and statistics

In statistics, a parameter is a description of a population, while a statistic describes a small portion of a population (sample). For example, if you ask everyone in your class (population) about their average height, you receive a parameter, a true description about the population

since everyone was asked. If you now want to guess the average height of people in your grade (population) using the information you have from your class (sample), this information turns into a statistic.

Hypothesis tests including a specific parameter are called parametric tests. In parametric tests, the population is assumed to have a normal distribution (e.g., the height of people in a class).

Non-parametric tests
In contrast, non-parametric tests (also distribution-free tests) are used when parameters of a population cannot be assumed to be normally distributed. For example, the price of diamonds seems exponentially distributed (below right). Non-parametric doesn't mean that you do not know anything about a population but rather that it is not normally distributed.



Left: example of normally distributed data. Right: example of non-normal data distribution.

For simplicity, I will focus on parametric tests in the following, with a few mentions on where to look further if a normal distribution cannot be assumed.

Real-world examples
An often-used example to explain hypothesis tests is the fair coin example. It is an excellent way to explain the basic concepts of a test but also very abstract.

More tangible examples of possible hypotheses in business that one can ask itself could be:
Hypothesis 1: Average order value has increased since last financial year
Parameter: Mean order value
Test type: one-sample, parametric test (assuming the order value follows a normal distribution)
— —
Hypothesis 2: Investing in A brings a higher return than investing in B
Parameter: Difference in mean return
Test type: two-sample, parametric test, also AB test (assuming the return follows a normal distribution)
— —
Hypothesis 3: The new user interface converts more users into customers than the expected 30%
Parameter: none
Test type: one-sample, non-parametric test (assuming number of customers is not normally distributed)

One-sample, two-sample, or more-sample test
When testing hypotheses, it is distinguished between one-sample, two-sample or more-sample tests.
This is not to be confused with one- and two-sided tests. In a one-sample test, a sample (average order value this year) is compared to a known value (average order value of last year).
In a two-sample test, two samples (investment A and B) are compared to each other.

Basic Steps of a hypothesis test

Several steps are used to test a hypothesis and verify its significance. In the following, I will explain each step in detail and will use the examples from above to explain all concepts.

1. Null & Alternative hypothesis
The null and alternative hypotheses are the two mutually exclusive statements about a parameter or population mentioned.
The null hypothesis (often abbreviated as H0) claims that there is no effect or no difference.
The alternative hypothesis (often abbreviated as H1 or HA) is what you want to prove.

Using one of the examples from above:
H0: There is no difference in the mean return from A and B, or the difference between A and B is zero.
H1: There is a difference in the mean return from A and B or the difference between A and B > zero.

One-sided and two-sided (one-tailed and two-tailed) tests
The example hypotheses above describe a so-called two-tailed test. In a two-tailed test, you are testing in both directions, meaning it is tested whether the mean return from A is significantly greater and significantly less than the mean return from B.

In a one-tailed test, you are testing in one direction, meaning it is tested either if the mean return from A is significantly greater or significantly less than the mean return from B.
In this case, the alternative hypothesis would change to:
H1: The mean return of A is greater than the mean return of B. OR
H1: The mean return of A is lower than the mean return of B.

2. Selection of an appropriate test statistic
To test your claims, you need to decide on the right test or test statistic. Often discussed tests are the t-test, z-test, or F-test, which all assume a normal distribution. However, in business, a normal distribution often cannot be assumed. Therefore, I will briefly explain the main concepts you need to know to find the proper test for your hypothesis.

Test statistic
Parametric or non-parametric test, each test has a test statistic. A test statistic is a numerical summary of a sample. It is a random variable as it is derived from a random sample. In hypothesis tests, it compares the sample statistic to the expected result of the null hypothesis. The selection of the test statistic is dependent on:

- Parametric vs. non-parametric
- Number of samples (one, two, multiple)

- Discrete (e.g. number of customers) or continuous variable (e.g. order value)

Let's assume that the mean average order value AOV in your web shop used to be $20. After hiring a new web designer with promising skills, the AOV increased to $22. You want to test whether the mean AOV has significantly increased:

Parameter: mean AOV (continuous variable, assumed to be normally distributed)
Sample statistic: $22 (one sample)
Expected value: $20
Test statistic: t-score
Test: one-sample t-test

3. Selection of the appropriate significance level
When testing hypotheses, we cannot always test it on the whole population but only on randomly selected data samples.

Can we, therefore, say that our conclusions are always 100% true for the population? Not really.

There are two types of errors that we can make:
Type I error: Rejecting the null hypothesis when it is true.
Type II error: Accepting the null hypothesis when it is false.

Alpha is the probability of the type I error and the chance of making a mistake by rejecting the null hypothesis when it is true.
The lower the alpha, the better. It is, therefore, used as a threshold to make decisions.

Before starting a hypothesis test, you generally pick an error level you are willing to accept.

For example, you are willing to accept a 5% chance that you're mistaken when you reject the null hypothesis.

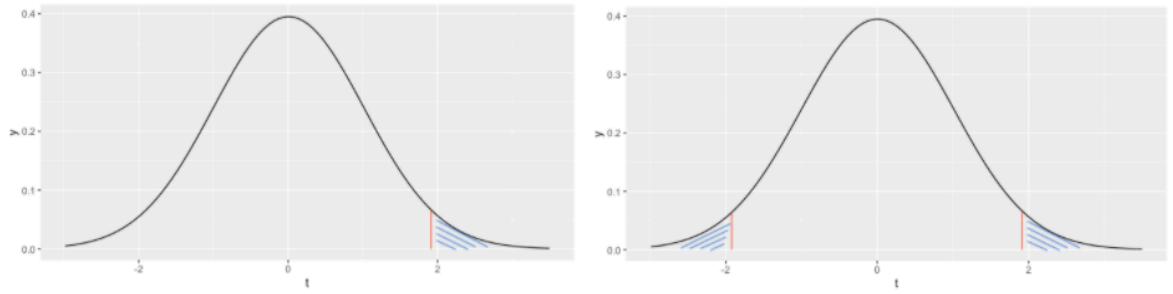
But, wouldn't I always want to be 100% confident that I didn't make a mistake, so alpha = 0%?

Power of a test
Yes, this is where it gets problematic. Because next to alpha, we also have beta, the probability of the type II error. 1-beta is the probability of not making a Type II error and defined as the power of a test. The lower the beta, the higher the power. Naturally, you would like to keep both errors as low as possible.

However, it is essential to note that both errors somewhat work against each other: Assume you want to minimise error I or the mistake of rejecting the null hypothesis when it is true. Then, the easiest way would be to just always accept it. But this would then work directly against the type II error, namely accepting it when it is not true.

Therefore, commonly used significance (alpha) levels 0.01, 0.05, or 0.10 serve as a good balance and should be determined before data collection.

Note here that in a two-tailed test, the alpha level is split in half and applied to both sides of the sampling distribution of a statistic.



Left: example of a sampling distribution with a rejection region on only one side. Right: example of a sampling distribution with a rejection region on both sides. Image by author.

4. Data collection

To run a hypothesis test, we need a portion of the true population of interest, a random sample. The sample should be randomly selected to avoid any bias or undesirable effects. The question about the optimal sample size is not an easy one to answer.

Generally, it is safe to say: the more data, the better. However, there are cases where this is hard to achieve due to budget or time constraints or just the nature of data.

There are several formulas available that help finds the right sample size:

- Cochran sample size formula
- Slovins's formula

Moreover, some tests can be used with small sample sizes, like e.g., the t-test or non-parametric tests which generally need fewer sample sizes as they do not require a normal distribution.

Sample size in two-sample or multi-sample tests

When conducting a two-sample or multi-sample test, be aware that your chosen test may require a similar sample size unless it is robust to different sample sizes. For example, a test like the t-test may not be appropriate anymore as an unequal sample size can affect the Type 1 error. In this case, it is best to search for a robust alternative (e.g., Welch's t-test).

5. Calculation of the test statistics and the p-value

Once the data is collected, the chosen test statistic and the corresponding p-value can be calculated.
Both values can be used to make your final decision on inference and are retrieved from the probability distribution from the test statistic (also sampling distribution).

How to calculate the test statistic?
You can calculate the test statistic traditionally using its formula (can be found online), or through statistical software like SPSS or using R/python.
For one of our examples from before, assuming a sample size (n) of 20 and a sample standard deviation (s) of 1.5, our test statistic is:

$$t = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$$

$$t = \frac{22 - 20}{\frac{1.5}{\sqrt{20}}} \approx 5.96$$

One-sample t-test

The p-value
The p-value (short for probability value) is the most critically viewed number in statistics. It is defined as the probability of receiving a result at least as extreme as the one observed, given that the null hypothesis is true.

My favourite resource of describing the p-value in simple words is by Cassie Kozyrkov:

The p-value tells you, given the evidence that you have (data), if the null hypothesis looks ridiculous or not […]
The lower the p-value, the more ridiculous the null hypothesis looks.

The p-value is a value between 0% and 100% and can be retrieved from the null hypothesis, sampling distribution, and the data.
Generally, it is calculated with the help of statistical software or reading off a distribution table with set parameters (degrees of freedom, alpha level etc.). Distribution tables with the most common parameters can be found online for most test statistics, like t-score, chi-squared score, or Wilcoxon-rank-sum.

6. Decision
To decide on inference, either the test statistic is compared to a critical value (critical value approach), or the p-value is compared to the alpha-level (p-value approach).

Critical value
The critical value splits the sampling distribution into a “rejection region” and “acceptance region”. If the test statistic is greater than the critical value, then the null hypothesis is rejected in favour of the alternative hypothesis with a confidence level of 1-alpha.

If the test statistic is smaller than the critical value, the null hypothesis is not rejected. Critical values are found with the sampling distribution and the alpha-level. However, a more common approach for making a test decision is the p-value approach.
P-value vs. alpha level

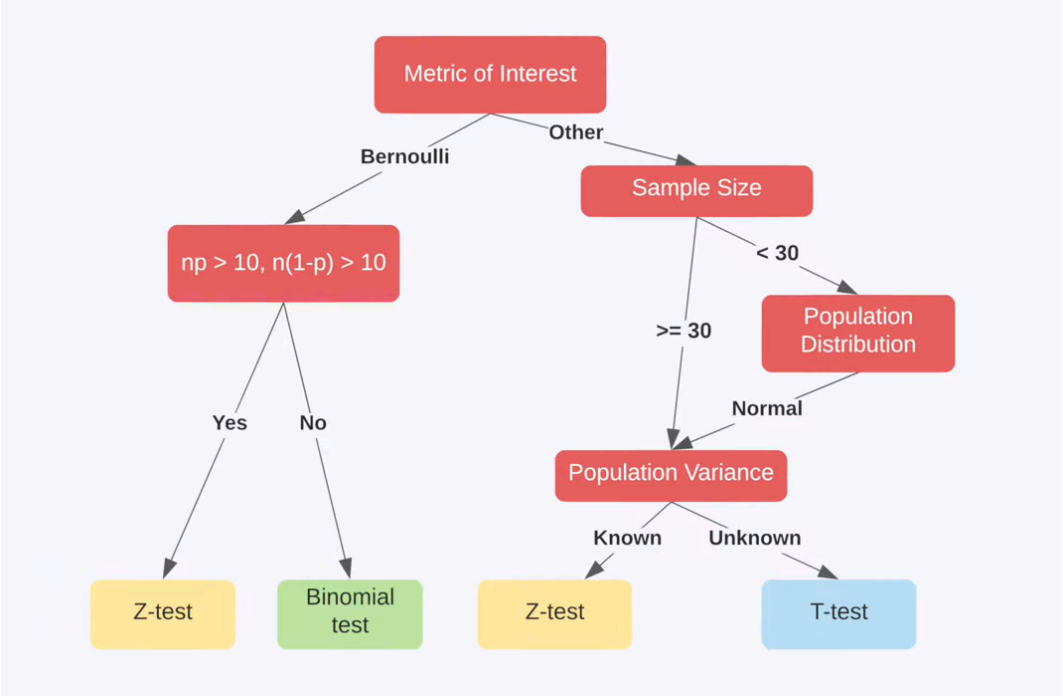
Given your alpha level, if the p <= alpha, the null hypothesis is rejected in favour of the alternative hypothesis with confidence level 1-alpha. If the p-value is greater than the alpha-level, the null hypothesis is accepted.

Summary

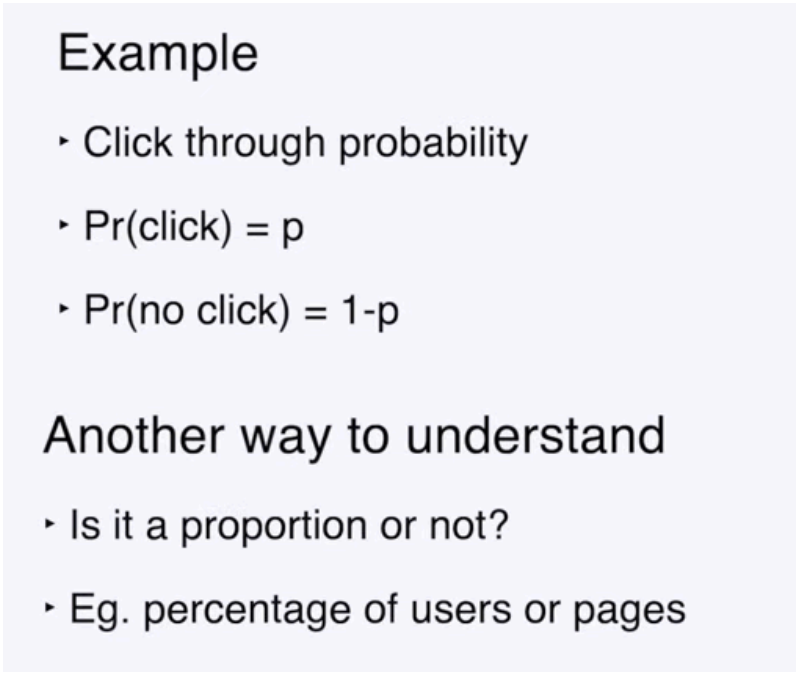
In hypothesis testing, two mutually exclusive statements about a population are tested using a random data sample. It comprises many concepts and steps that greatly impact the results, like formulating the hypotheses or selecting the test statistic, alpha-level, and sample size.

Resources

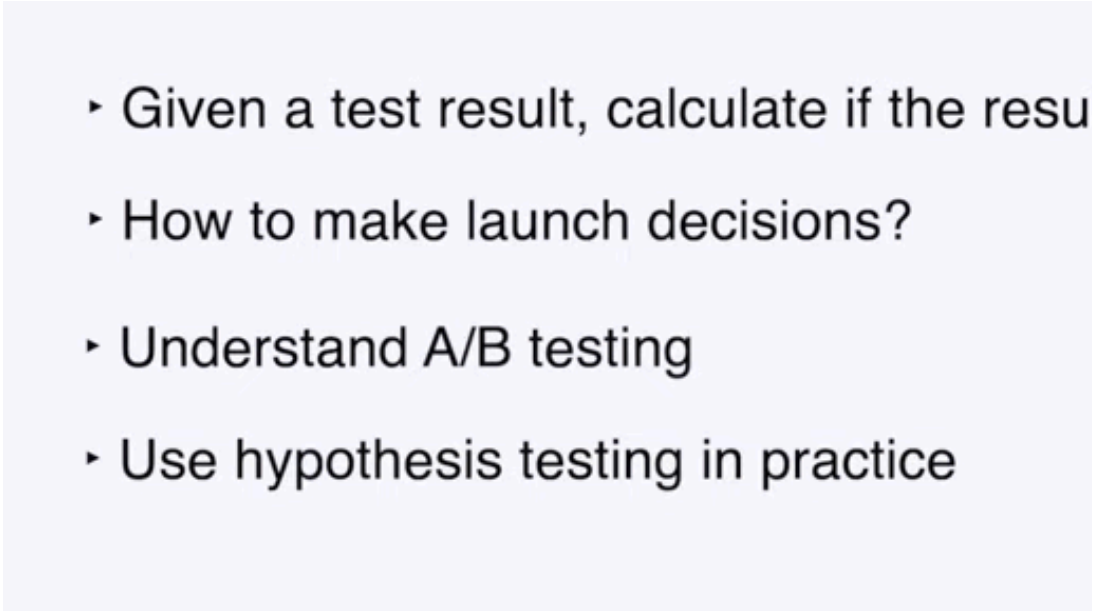
- https://greenteapress.com/thinkstats2/thinkstats2.pdf
- https://www.statisticshowto.com/
- https://openstax.org/books/introductory-business-statistics/pages/10-1-comparing-two-independent-population-means
- https://stats.libretexts.org/Courses/Highline_College/Book%3A_Statistics_Using_Technology_(Kozak)/07%3A_One-Sample_Inference/7.01%3A_Basics_of_Hypothesis_Testing

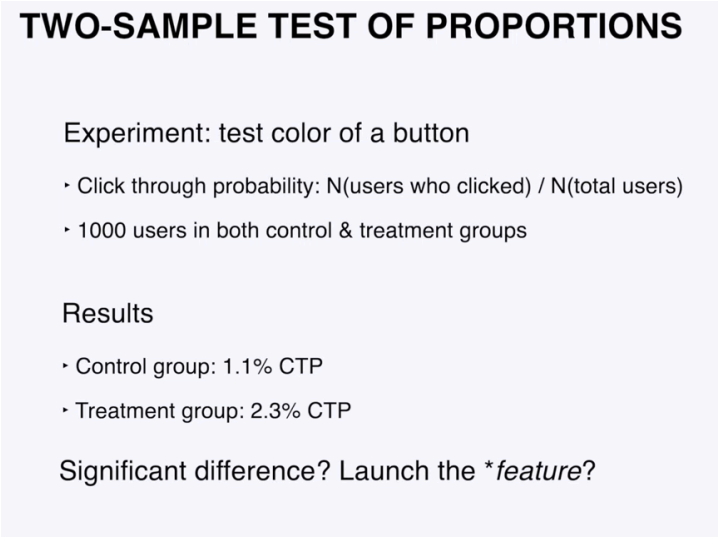


Bernouilli Distribution



[Youtube Video]
[Video]
[Binomial Tests]





Upper-tailed, Lower-tailed, Two-tailed Tests

The research or alternative hypothesis can take one of three forms. An investigator might believe that the parameter has increased, decreased or changed. For example, an investigator might hypothesise:

1. H_1: $\mu > \mu_0$, where μ_0 is the comparator or null value (e.g., μ_0 =191 in our example about weight in men in 2006) and an increase is hypothesised - this type of test is called an upper-tailed test;
2. H_1: $\mu < \mu_0$, where a decrease is hypothesised and this is called a lower-tailed test; or
3. H_1: $\mu \neq \mu_0$, where a difference is hypothesised and this is called a two-tailed test.

The exact form of the research hypothesis depends on the investigator's belief about the parameter of interest and whether it has possibly increased, decreased or is different from the null value. The research hypothesis is set up by the investigator before any data are collected.

Z-Test for Proportions

One Proportion Z Test

The test statistic is a z-score (z) defined by the following equation. $z = \frac{(p-P)}{\sigma}$

where P is the hypothesized value of population proportion in the null hypothesis, p is the sample proportion, and σ is the standard deviation of the sampling distribution.

Test Statistics is defined and given by the following function:

Formula

$$z = \frac{\hat{p} - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}}$$

Where –

- z = Test statistics
- n = Sample size
- p_o = Null hypothesized value
- $\hat{p}$ = Observed proportion

Problem Statement:

A survey claims that 9 out of 10 doctors recommend aspirin for their patients with headaches. To test this claim, a random sample of 100 doctors is obtained. Of these 100 doctors, 82 indicate that they recommend aspirin. Is this claim accurate? Use alpha = 0.05.

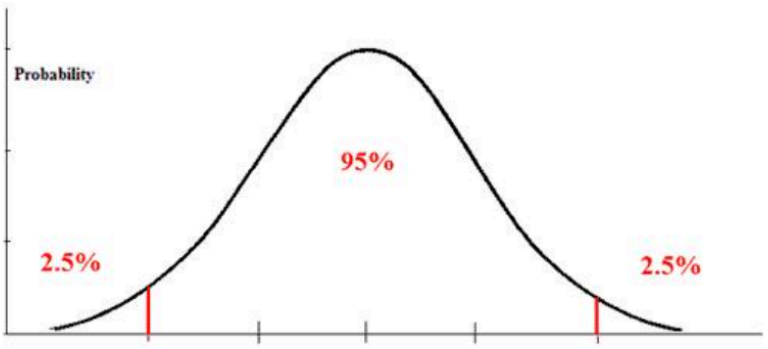
Solution:

Define Null and Alternative Hypotheses

$$H_0; p = .90$$
$$H_0; p \neq .90$$

Here Alpha = 0.05. Using an alpha of 0.05 with a two-tailed test, we would expect our distribution to look something like this:

Here Alpha = 0.05. Using an alpha of 0.05 with a two-tailed test, we would expect our distribution to look something like this:



Here we have 0.025 in each tail. Looking up 1 - 0.025 in our z-table, we find a critical value of 1.96. Thus, our decision rule for this two-tailed test is: If Z is less than -1.96, or greater than 1.96, reject the null hypothesis.Calculate Test Statistic:

$$z = \frac{\hat{p} - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}}$$

$$\hat{p} = .82$$

$$p_o = .90$$

$$n = 100$$

$$z_o = \frac{.82 - .90}{\sqrt{\frac{.90(1-.90)}{100}}}$$

$$= \frac{-.08}{0.03}$$

$$= -2.667$$

As z = -2.667 Thus as result we should reject the null hypothesis and as conclusion, The claim that 9 out of 10 doctors recommend aspirin for their patients is not accurate, z = -2.667, p < 0.05.

Two Sample Z Test of Proportions

Two-sample Z test of proportions is the test to determine whether the two populations differ significantly on specific characteristics. In other words, compare the proportion of two different populations that have some single characteristic. It calculates the range of values that is likely to include the difference between the population proportions.

TWO-SAMPLE TEST OF PROPORTIONS

1. Which hypothesis test to use?
2. What is the null hypothesis?
3. Is the result statistically significant?
4. Is the result practically significant?
5. Make decisions

Which hypothesis test to use?

- Bernoulli population: either clicks or doesn't click
- Control group: n * p̂ = 1000 * 1.1% = 11
- Treatment group: n * p̂ = 1000 * 2.3% = 23
- Test statistic follows Z-distribution

Measurements

- Users clicked x_{ct}, x_{tr}
- Total number of users n_{ct}, n_{tr}

$$\hat{p}_{ct} = \frac{X_{ct}}{n_{ct}} = \frac{11}{1000}$$

$$\hat{p}_{tr} = \frac{X_{tr}}{n_{tr}} = \frac{23}{1000}$$

What is the null hypothesis?

$$d = \hat{p}_{tr} - \hat{p}_{ct}$$

Null hypothesis

$$H_0 : p_{ct} = p_{tr}, d = 0$$

$$\hat{d} \sim N(0, SE^2)$$

Is result **statistically** significant?

- critical z-score (α: 0.05) = 1.96
- TS > 1.96 or TS < -1.96, reject null hypothesis

In this example

- TS = 2.076 > 1.96
- Test is statistically significant

Is result **practically** significant?

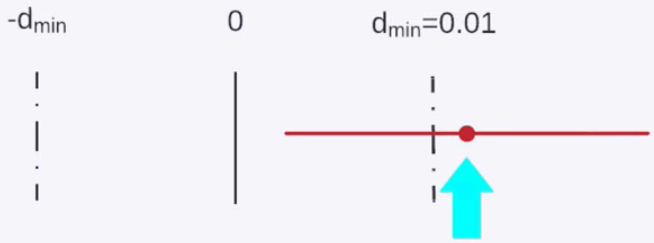
Confidence interval of d

- Center of C.I. = 0.012
- Width of C.I. (margin of error)

$$m = Z * S_{pool} = 1.96 * 0.00578 = 0.0113$$

$$\text{CI of d: } 0.012 \pm 0.0113 = 0.0007 \sim 0.0233$$

Is result **practically** significant?



Best guess: there is a practical significant change

Checking statistical significance

- Check if C.I. overlaps with 0
  - If it does, result is not statistically significant
- Equivalent to comparing TS with critical value

When to use Two Sample Z Proportion test

The purpose of two sample Z test is to compare the random samples of two populations. Use two sample z test of proportion for large sample size and Fisher exact probability test is an excellent non-parametric test for small sample sizes.

Assumptions of the Two Sample Z Proportion Hypothesis Tests

- The data are simple random values from both the populations
- Both populations follow a binomial distribution
- Samples are independent of each other
- Test results are accurate when np and n(1-p) are greater than 5

Hypothesis of two sample Z proportion test

- Null hypothesis: The difference between population proportions is equal to hypothesised difference
- Alternative hypothesis: The difference between population proportions is not equal to hypothesised difference (two -tailed)
- The difference between population proportions is greater than hypothesised difference (right-tailed)
- The difference between population proportions is less than hypothesised difference (left -tailed)

Pooled (vs Unpooled)

The pooling refers to the way in which the standard error is estimated when calculating terms for a hypothesis test.
In the pooled version, the two proportions are averaged, and only one proportion is used to estimate the standard error. ASQ, Villanova, and most other organisations favour pooled calculations.
In the unpooled version, the two proportions are used separately. IASSC generally favours unpooled.

Two Sample Z Test of Proportions Variations

There are 2 ways to compute the two sample Z test of proportions i.e pooled or unpooled.
Pooled Z test of proportions formula

$$Z = \frac{\widehat{p_1} - \widehat{p_2} - 0}{\sqrt{p_0(1-p_0)(\frac{1}{n_1} + \frac{1}{n_2})}}$$

Pooled Calculation

Un-pooled Z test of proportions formula

$$Z_{1-\alpha/2} = \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{(\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2})}}$$

Unpooled Calculation

Procedure to execute Two Sample Proportion Hypothesis Test

- State the null hypothesis and alternative hypothesis
- State alpha, in other words determine the significance level
- Compute the test statistic
- Determine the critical value (from critical value table)
- Define the rejection criteria
- Finally, interpret the result. If the test statistic falls in critical region, reject the null hypothesis

Two Sample Z Test of Proportions Pooled (ASQ, Villanova)

$$Z = \frac{\widehat{p_1} - \widehat{p_2} - 0}{\sqrt{p_0(1-p_0)(\frac{1}{n_1} + \frac{1}{n_2})}}$$

Where

$$p_0 = \frac{X_1 + X_2}{n_1 + n_2}$$

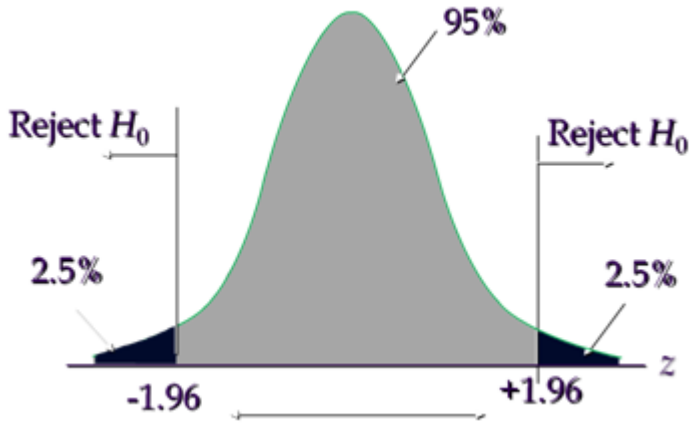
- z is test statistic
- p̂1 and p̂2 are observed proportion of events in the two samples
- n1 and n2 are sample sizes
- X1 and X2 are number of trails

Example of Two Sample Z Proportion Test (pooled)
Example: A car manufacturer aims to improve the quality of the products by reducing the defects and also increase the customer satisfaction. Therefore, he monitors the efficiency of two assembly lines in the shop floor. In line A there are 18 defects reported out of 200 samples. While line B shows 25 defects out of 600 cars. At α 5%, are the differences between two assembly procedures significant?
Define Null and Alternative hypothesis

- Null Hypothesis: Two proportions are the same
- Alternative Hypothesis: Two proportions are not the same

α=0.05
State decision rule

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756



Critical value is ±1.96, hence reject the null hypothesis if the calculated value is less than -1.96 or greater than +1.96

Calculate Test Statistic

- Line A= p̂1=18/200= 0.09 = 9%
- Line B= p̂2= 25/600 = 0.0416 = 4.16%

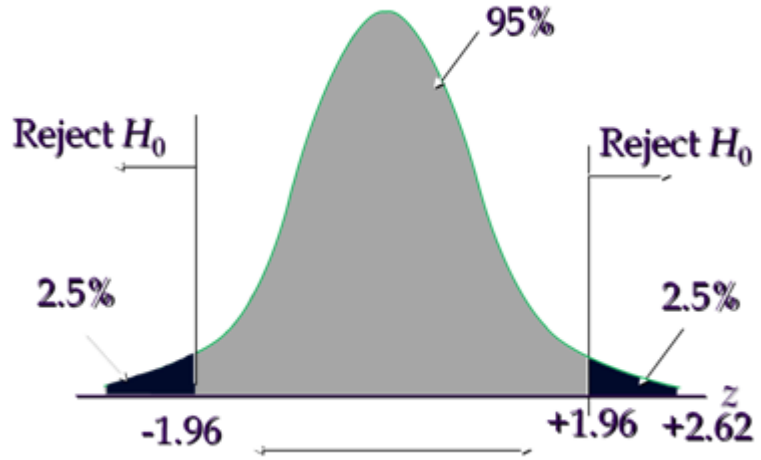
$$p_0 = \frac{X_1+X_2}{n_1+n_2}$$

- First compute p0 = 18+25/200+600=43/800=0.0537 =5.37%
- Now p̂1 – p̂2 = 0.09 – 0.0416 = 0.0484
- p0 * (1 – p0) = 0.0537 * (1-0.0537) = 0.0537 * 0.9463 = 0.0508
- And (1/n1) + (1/n2) = 0.006667

$$z = \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{p_0(1-p_0)(\frac{1}{n_1}+\frac{1}{n_2})}}$$

$$z = \frac{0.09 - 0.0416}{\sqrt{0.0537(1-0.0537)(\frac{1}{200}+\frac{1}{600})}}$$

- So, Z = (0.0484)/ SQRT ((0.0508) * (0.006667))
- Z = (0.0484)/ SQRT (0.000339)) = (0.0484)/ (0.018406) = 2.62



Interpret the results:

Compare Zcalc to Zcritical . In hypothesis testing, a critical value is a point on the test distribution compared to the test statistic to determine whether to reject the null hypothesis. Calculated test statistic value 2.62 and it is in critical region, hence reject the null hypothesis, so, there is a significant difference in two line assembly procedures.

Note that statistical significance is directly impacted by sample size. Recall that there is an inverse relationship between sample size and the standard error (i.e., standard deviation of the sampling distribution). Very small differences will be statistically significant with a very large sample size.

Thus, when results are statistically significant it is important to also examine practical significance. Practical significance is not directly influenced by sample size.
Practical significance refers to the magnitude of the difference, which is known as the effect size. Results are practically significant when the difference is large enough to be meaningful in real life. What is meaningful may be subjective and may depend on the context.

Effect Size

For some tests there are commonly used measures of effect size. For example, when comparing the difference in two means we often compute Cohen's d which is the difference between the two observed sample means in standard deviation units:

$$d = \frac{\overline{x}_1 - \overline{x}_2}{s_p}$$

Where s_p is the pooled standard deviation

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

Below are commonly used standards when interpreting Cohen's d:

Cohen's d	Interpretation
0 - 0.2	Little or no effect
0.2 - 0.5	Small effect size
0.5 - 0.8	Medium effect size
0.8 or more	Large effect size

For a single mean, you can compute the difference between the observed mean and hypothesized mean in standard deviation units: $d = \frac{\overline{x} - \mu_0}{s}$

For correlation and regression we can compute r^2 which is known as the coefficient of determination. This is the proportion of shared variation. We will learn more about r^2 when we study simple linear regression and correlation at the end of this course.

Checking statistical significance

- Check if C.I. overlaps with 0
  - If it does, result is not statistically significant
- Equivalent to comparing TS with critical value

TWO-SAMPLE TEST OF MEANS

Experiment: if a new feature changes avg. number of posts

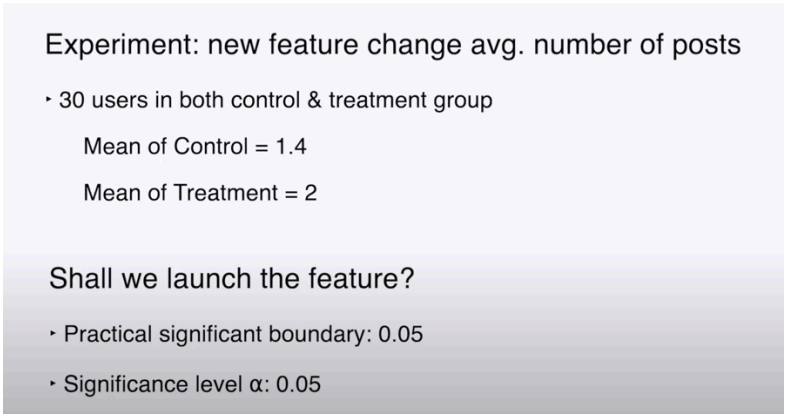
- 30 users in both control & treatment groups

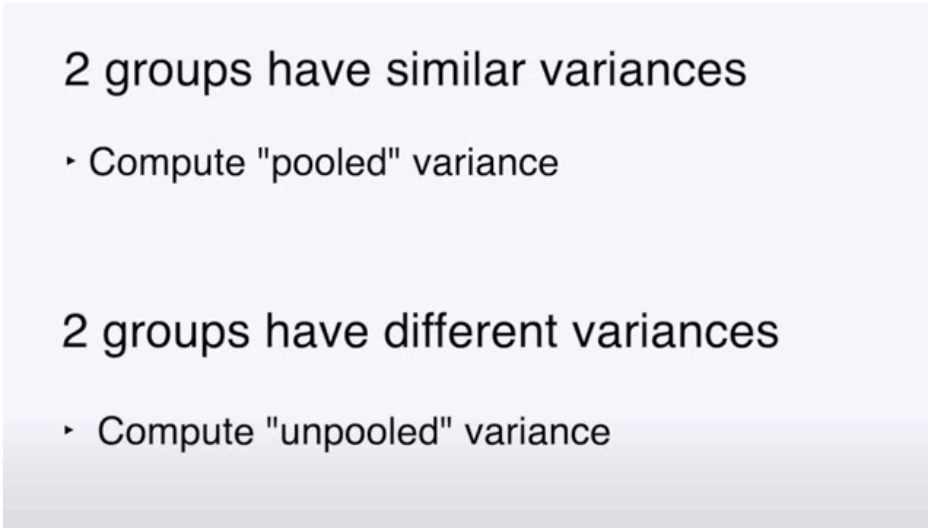
Control:

[1, 0, 1, 3, 2, 1, 0, 1, 3, 2, 1, 0, 1, 3, 2,
1, 0, 1, 3, 2, 1, 0, 1, 3, 2, 1, 0, 1, 3, 2]

Treatment:

[0, 1, 3, 2, 1, 2, 1, 3, 2, 1, 0, 2, 3, 2, 1,
0, 2, 3, 2, 1, 0, 2, 3, 2, 1, 0, 2, 3, 2, 4]





Two-Sample z-test for Comparing Two Means

[Read More]

Requirements: Two normally distributed but independent populations, σ is known

Hypothesis test

Formula: $$z = \frac{\bar{x}_1 - \bar{x}_2 - \Delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the means of the two samples, Δ is the hypothesised difference between the population means (0 if testing for equal means), σ_1 and σ_2 are the standard deviations of the two populations, n n_1 and n_2 are the sizes of the two samples.

The amount of a certain trace element in blood is known to vary with a standard deviation of 14.1 ppm (parts per million) for male blood donors and 9.5 ppm for female donors. Random samples of 75 male and 50 female donors yield concentration means of 28 and 33 ppm, respectively. What is the likelihood that the population means concentrations of the element are the same for men and women?

Null hypothesis: H_0: $\mu_1 = \mu_2$

or H_0: $\mu_1 - \mu_2 = 0$

alternative hypothesis: H_a : $\mu_1 \neq \mu_2$

$$z = \frac{28-33-0}{\sqrt{\frac{14.1^2}{75}+\frac{9.5^2}{50}}} = \frac{-5}{\sqrt{2.65+1.81}} = -2.37$$

or: $H_a : \mu_1 - \mu_2 \neq 0$

The computed z-value is negative because the (larger) mean for females was subtracted from the (smaller) mean for males. But because the hypothesised difference between the populations is 0, the order of the samples in this computation is arbitrary— $\bar{x}_1$ could just as well have been the female sample mean and $\bar{x}_2$ the male sample means, in which case z would be 2.37 instead of –2.37. An extreme z-score in either tail of the distribution (plus or minus) will lead to the rejection of the null hypothesis of no difference.

The area of the standard normal curve corresponding to a z-score of –2.37 is 0.0089. Because this test is two-tailed, that figure is doubled to yield a probability of 0.0178 that the population means are the same. If the test had been conducted at a pre-specified significance level of $\alpha < 0.05$, the null hypothesis of equal means could be rejected. If the specified significance level had been the more conservative (more stringent) $\alpha < 0.01$, however, the null hypothesis could not be rejected.

Chi-Square (X2)

The statistical procedures that we have reviewed thus far are appropriate only for numerical variables.

The chi-square (χ^2) test can be used to evaluate a relationship between two categorical variables. It is one example of a nonparametric test. Nonparametric tests are used when assumptions about normal distribution in the population cannot be met. These tests are less powerful than parametric tests.

Suppose that 125 children are shown three television commercials for breakfast cereal and are asked to pick which they liked best. The results are shown in Table 1.

You would like to know if the choice of favourite commercial was related to whether the child was a boy or a girl or if these two variables are independent. The totals in the margins will allow you to determine the overall probability of (1) liking commercial A, B, or C, regardless of gender, and (2) being either a boy or a girl, regardless of favourite commercial. If the two variables are independent, then you should be able to use these probabilities to predict approximately how many children should be in each cell. If the actual count is very different from the count that you would expect if the probabilities are independent, the two variables must be related.

Table 1. Commercial Preference for Boys and Girls

	A	B	C	Totals
Boys	30	29	16	75
Girls	12	33	5	50
Totals	42	62	21	125

Consider the upper-right cell of the table. The overall probability of a child in the sample being a boy is 75 ÷ 125 = 0.6. The overall probability of liking Commercial A is 42 ÷ 125 = 0.336. The multiplication rule states that the probability of both of two independent events occurring is the product of their two probabilities. Therefore, the probability of a child both being a boy and liking Commercial A is 0.6 × 0.336 = 0.202. The expected number of children in this cell, then, is 0.202 × 125 = 25.2.

There is a faster way of computing the expected count for each cell: Multiply the row total by the column total and divide by n. The expected count for the first cell is, therefore, (75 × 42) ÷ 125 = 25.2. If you perform this operation for each cell, you get the expected counts (in parentheses) shown in Table 2.

Table 2. Chi-Square Results for Table

	A	B	C	Totals
Boys	30 (25.2)	29 (37.2)	16 (12.6)	75
Girls	12 (16.8)	33 (24.8)	5 (8.4)	50
Totals	42	62	21	125

Note that the expected counts properly add up to the row and column totals. You are now ready for the formula for χ^2, which compares each cell's actual count to its expected count:

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

The formula describes an operation that is performed on each cell and which yields a number. When all the numbers are summed, the result is χ^2. Now, compute it for the six

$$\chi^2 = \frac{(30-25.2)^2}{25.2} + \frac{(29-37.2)^2}{37.2} + \frac{(16-12.6)^2}{12.6} + \frac{(12-16.8)^2}{16.8} + \frac{(33-24.8)^2}{24.8} + \frac{(5-8.4)^2}{8.4}$$
$$= 0.914 + 1.808 + 0.917 + 1.371 + 2.711 + 1.376$$

cells in the example: $= 9.098$

The larger χ^2, the more likely that the variables are related; note that the cells that contribute the most to the resulting statistic are those in which the expected count is very different from the actual count.

Chi-square has a probability distribution, the critical values for which are listed in Table 4 in "Statistics Tables." As with the t-distribution, χ^2 has a degrees-of-freedom parameter, the formula for which is

(number of rows – 1) × (number of columns – 1)

or in your example:

(2 – l) × (3 – 1) = 1 × 2 = 2

In Table 4 in "Statistics Tables," a chi-square of 9.097 with two degrees of freedom falls between the commonly used significance levels of 0.05 and 0.01. If you had specified an alpha of 0.05 for the test, you could, therefore, reject the null hypothesis that gender and favourite commercials are independent. At a = 0.01, however, you could not reject the null hypothesis.

The χ^2 test does not allow you to conclude anything more specific than that there is some relationship in your sample between gender and commercial liked (at $\alpha = 0.05$). Examining the observed versus expected counts in each cell might give you a clue as to the nature of the relationship and which levels of the variables are involved. For example, Commercial B appears to have been liked more by girls than boys. But χ^2 tests only the very general null hypothesis that the two variables are independent.

Sometimes a chi-square test of homogeneity of populations is used. It is very similar to the test for independence. In fact the mechanics of these tests are identical. The real difference is in the design of the study and the sampling method.

Modelling

Modelling relies on a strong understanding of probability distributions and hypothesis testing. Since it is a broad term, we will refer to modelling as the areas which have a strong statistical intersection with Machine Learning.

This includes topics such as linear regression, maximum likelihood estimation, & bayesian statistics.

For interviews focused on modelling and machine learning, knowing these topics is essential.

The distinction between probability and likelihood is fundamentally important:

- Probability attaches to possible results;
- likelihood attaches to hypotheses.

Possible results are mutually exclusive and exhaustive. Suppose we ask a subject to predict the outcome of each of 10 tosses of a coin. There are only 11 possible results (0 to 10 correct predictions).
The actual result will always be one and only one of the possible results. Thus, the probabilities that attach to the possible results must sum to 1.

Hypotheses, unlike results, are neither mutually exclusive nor exhaustive.

Suppose that the first subject we test predicts 7 of the 10 outcomes correctly. I might hypothesise that the subject just guessed, and you might hypothesise that the subject may be somewhat clairvoyant, by which you mean that the subject may be expected to correctly predict the results at slightly greater than chance rates over the long run.

These are different hypotheses, but they are not mutually exclusive, because you hedged when you said: “maybe.” You thereby allowed your hypothesis to include mine. In technical terminology, my hypothesis is nested within yours.

Someone else might hypothesise that the subject is strongly clairvoyant and that the observed result underestimates the probability that her next prediction will be correct.

Another person could hypothesise something else altogether. There is no limit to the hypotheses one might entertain.

The first step towards problem-solving in data science projects isn’t about building machine learning models.
That distinction belongs to hypothesis generation – the step where we combine our problem solving skills with our business intuition.
It’s a truly crucial step in ensuring a successful data science project.

Let’s be honest – all of us think of a hypothesis almost everyday. Let us consider the example of a famous sport in India – cricket. It is that time of the year when IPL fever is high and we are all absorbed in predicting the winner.

If you have been guessing which team would win based on various factors like the size of the stadium and batsmen present in the team with six hitting capabilities or batsmen with high T20 averages, then kudos to you all.
You have all been making an educated guess and generating hypotheses based on your domain knowledge of the sport.

Similarly, the first step towards solving any business problem using machine learning is hypothesis generation. Understanding the problem statement with good domain knowledge is important and formulating a hypothesis will further expose you to newer ideas of problem-solving.

So in this article, let’s dive into what hypothesis generation is and figure out why it is important for every data scientist.

What is Hypothesis Generation?

Hypothesis generation is an educated “guess” of various factors that are impacting the business problem that needs to be solved using machine learning.
In framing a hypothesis, the data scientist must not know the outcome of the hypothesis that has been generated based on any evidence.
“A hypothesis may be simply defined as a guess. A scientific hypothesis is an intelligent guess.” – Isaac Asimov

Hypothesis generation is a crucial step in any data science project. If you skip this or skim through this, the likelihood of the project failing increases exponentially.

Hypothesis Generation vs. Hypothesis Testing

This is a very common mistake data science beginners make.

Hypothesis generation is a process beginning with an educated guess whereas hypothesis testing is a process to conclude that the educated guess is true/false or the relationship between the variables is statistically significant or not.

This latter part could be used for further research using statistical proof. A hypothesis is accepted or rejected based on the significance level and test score of the test used for testing the hypothesis.
To understand more about hypothesis testing in detail, you can read about it here

How Does Hypothesis Generation Help?

Here are 5 key reasons why hypothesis generation is so important in data science:

- Hypothesis generation helps in comprehending the business problem as we dive deep in inferring the various factors affecting our target variable
- You will get a much better idea of what are the major factors that are responsible to solve the problem
- Data that needs to be collected from various sources that are key in converting your business problem into a data science-based problem
- Improves your domain knowledge if you are new to the domain as you spend time understanding the problem
- Helps to approach the problem in a structured manner

When Should you Perform Hypothesis Generation?

The million-dollar question – when in the world should you perform hypothesis generation?

- The hypothesis generation should be made before looking at the dataset or collection of the data
- You will notice that if you have done your hypothesis generation adequately, you would have included all the variables present in the dataset in your hypothesis generation
- You might also have included variables that are not present in the dataset

Hypothesis Generation Based On Various Factors

Case Study: Hypothesis Generation on “New York City Taxi Trip Duration Prediction”

To predict the duration of a trip so that the company can assign the cabs that are free for the next trip. This will help in reducing the wait time for customers and will also help in earning customer trust.

1. Distance/Speed based Features

Let us try to come up with a formula that would have a relation with trip duration and would help us in generating various hypotheses for the problem:

TIME=DISTANCE/SPEED

Distance and speed play an important role in predicting the trip duration.

We can notice that the trip duration is directly proportional to the distance travelled and inversely proportional to the speed of the taxi. Using this we can come up with a hypothesis based on distance and speed.

- Distance: More the distance travelled by the taxi, the more will be the trip duration.
- Interior drop point: Drop points to congested or interior lanes could result in an increase in trip duration
- Speed: Higher the speed, the lower the trip duration

2. Features based on Car

Cars are of various types, sizes, brands, and these features of the car could be vital for commuting not only on the basis of the safety of the passengers but also for the trip duration. Let us now generate a few hypotheses based on the features of the car.

- Condition of the car: Good conditioned cars are unlikely to have breakdown issues and could have a lower trip duration
- Car Size: Small-sized cars (Hatchback) may have a lower trip duration and larger-sized cars (XUV) may have higher trip duration based on the size of the car and congestion in the city

3. Type of the Trip

Trip types can be different based on trip vendors – it could be an outstation trip, single or pool rides. Let us now define a hypothesis based on the type of trip used.

- Pool Car: Trips with pooling can lead to higher trip duration as the car reaches multiple places before reaching your assigned destination

4. Features based on Driver Details

A driver is an important person when it comes to commute time. Various factors about the driver can help in understanding the reason behind trip duration and here are a few hypotheses for this.

- Age of driver: Older drivers could be more careful and could contribute to higher trip duration
- Gender: Female drivers are likely to drive slowly and could contribute to higher trip duration
- Driver experience: Drivers with very less driving experience can cause higher trip duration
- Medical condition: Drivers with a medical condition can contribute to higher trip duration

5. Passenger details

Passengers can influence the trip duration knowingly or unknowingly. We usually come across passengers requesting drivers to increase the speed as they are getting late and there could be other factors to hypothesise which we can look at.

- Age of passengers: Senior citizens as passengers may contribute to higher trip duration as drivers tend to go slow in trips involving senior citizens

- Medical conditions or pregnancy: Passengers with medical conditions contribute to a longer trip duration
- Emergency: Passengers with an emergency could contribute to a shorter trip duration
- Passenger count: Higher passenger count leads to shorter duration trips due to congestion in seating

6. Date-Time Features

The day and time of the week are important as New York is a busy city and could be highly congested during office hours or weekdays. Let us now generate a few hypotheses on the date and time-based features.

Pickup Day:

- Weekends could contribute to more outstation trips and could have a higher trip duration
- Weekdays tend to have higher trip duration due to high traffic
- If the pickup day falls on a holiday then the trip duration may be shorter
- If the pickup day falls on a festive week then the trip duration could be lower due to lesser traffic

Time:

- Early morning trips have a lesser trip duration due to lesser traffic
- Evening trips have a higher trip duration due to peak hours

7. Road-based Features

Roads are of different types and the condition of the road or obstructions in the road are factors that can't be ignored. Let's form some hypotheses based on these factors.

- Condition of the road: The duration of the trip is more if the condition of the road is bad
- Road type: Trips in concrete roads tend to have a lower trip duration
- Strike on the road: Strikes carried out on roads in the direction of the trip causes the trip duration to increase

8. Weather Based Features

Weather can change at any time and could possibly impact the commute if the weather turns bad. Hence, this is an important feature to consider in our hypothesis.

- Weather at the start of the trip: Rainy weather condition contributes to a higher trip duration

Q: Explain the central limit theorem and give examples of when you can use it in a real-world problem.

The central limit theorem states that if any random variable, regardless of the distribution, is sampled a large enough time, the sample mean will be approximately normally distributed. This allows for studying the properties of any statistical distribution as long as there is a large enough sample size.

We can rely on the CLT with means (because it applies to any unbiased statistic) only if expressing data in this way makes sense. And it makes sense ONLY in the case of unimodal and symmetric data, coming from additive processes.

So forget skewed, multi-modal data with mixtures of distributions, coming from multiplicative processes, and non-trivial mean-variance relationships. Those are the places where arithmetic means are meaningless. Thus, using the CLT of e.g. bootstrap will give some valid answers to an invalid question.
the distribution of means isn't enough. Every single kind of inference requires the entire test statistic to follow a certain distribution. And the test statistic consists also of the estimate of variance. Never assume the same sample size sufficient for means will suffice for the entire test statistic. Especially do not believe in magic numbers like N=30.

Think first about how to sensible describe your data, state the hypothesis of interest and then apply a valid method.

Examples of real-world usage of CLT:

1. The CLT can be used at any company with a large amount of data. Companies like Uber/Lyft want to test whether adding a new feature will increase the booked rides or not using hypothesis testing. So if we have a large number of individual rides X, which in this case is a Bernoulli random variable (since the rider will book a ride or not), we can estimate the statistical properties of the total number of bookings. Understanding and estimating these statistical properties play a significant role in applying hypothesis testing to your data and knowing whether adding a new feature will increase the number of booked riders or not.
2. Manufacturing plants often use the central limit theorem to estimate how many products produced by the plant are defective.

Q. What is the Central Limit Theorem?
Central Limit Theorem is the cornerstone of statistics. It states that the distribution of a sample from a population comprising a large sample size will have its mean normally distributed. In other words, it will not have any effect on the original population distribution. Central Limit Theorem is widely used in the calculation of confidence intervals and hypothesis testing. Here is an example – We want to calculate the average height of people in the world, and we take some samples from the general population, which serves as the data set. Since it is hard or impossible to obtain data regarding the height of every person in the world, we will simply calculate the mean of our sample.
By multiplying it several times, we will obtain the mean and their frequencies which we can plot on the graph and create a normal distribution. It will form a bell-shaped curve that will closely resemble the original data set.

Q. What is the assumption of normality?
The assumption of normality dictates that the mean distribution across samples is normal. This is true across independent samples as well.

Q: Briefly explain the A/B testing and its application? What are some common pitfalls encountered in A/B testing?
A/B testing helps us to determine whether a change in something will cause a change in performance significantly or not. So in other words you aim to statistically estimate the impact of a given change within your digital product (for example). You measure success and counter metrics on at least 1 treatment vs 1 control group (there can be more than 1 XP group for multivariate tests).

Applications:

1. Consider the example of a general store that sells bread packets but not butter, for a year. If we want to check whether its sale depends on the butter or not, then suppose the store also sells butter and sales for next year are observed. Now we can determine whether selling butter can significantly increase/decrease or doesn't affect the sale of bread.
2. While developing the landing page of a website you create 2 different versions of the page. You define a criteria for success eg. conversion rate. Then define your hypothesis Null hypothesis(H): No difference between the performance of the 2 versions. Alternative hypothesis(H'): version A will perform better than B.

NOTE: You will have to split your traffic randomly(to avoid sample bias) into 2 versions. The split doesn't have to be symmetric, you just need to set the minimum sample size for each version to avoid undersample bias.

Now if version A gives better results than version B, we will still have to statistically prove that results derived from our sample represent the entire population. Now one of the very common tests used to do so is 2 sample t-tests where we use values of significance level (alpha) and p-value to see which hypothesis is right. If p-value<alpha, H is rejected.

Common pitfalls:

1. Wrong success metrics inadequate to the business problem
2. Lack of counter metric, as you might add friction to the product regardless along with the positive impact
3. Sample mismatch: heterogeneous control and treatment, unequal variances
4. Underpowered test: too small sample or XP running too short 5. Not accounting for network effects (introduce bias within measurement)

Q: Describe briefly the hypothesis testing and p-value in layman's terms? And give a practical application for them ?

In Layman's terms:

- Hypothesis test is where you have a current state (null hypothesis) and an alternative state (alternative hypothesis). You assess the results of both of the states and see some differences. You want to decide whether the difference is due to the alternative approach or not.

You use the p-value to decide this, where the p-value is the likelihood of getting the same results the alternative approach achieved if you keep using the existing approach. It's the probability to find the result in the gaussian distribution of the results you may get from the existing approach.

The rule of thumb is to reject the null hypothesis if the p-value < 0.05, which means that the probability to get these results from the existing approach is <95%. But this % changes according to task and domain.

To explain the hypothesis testing in Layman's term with an example, suppose we have two drugs A and B, and we want to determine whether these two drugs are the same or different.

This idea of trying to determine whether the drugs are the same or different is called hypothesis testing. The null hypothesis is that the drugs are the same, and the p-value helps us decide whether we should reject the null hypothesis or not.
p-values are numbers between 0 and 1, and in this particular case, it helps us to quantify how confident we should be to conclude that drug A is different from drug B. The closer the p-value is to 0, the more confident we are that the drugs A and B are different.

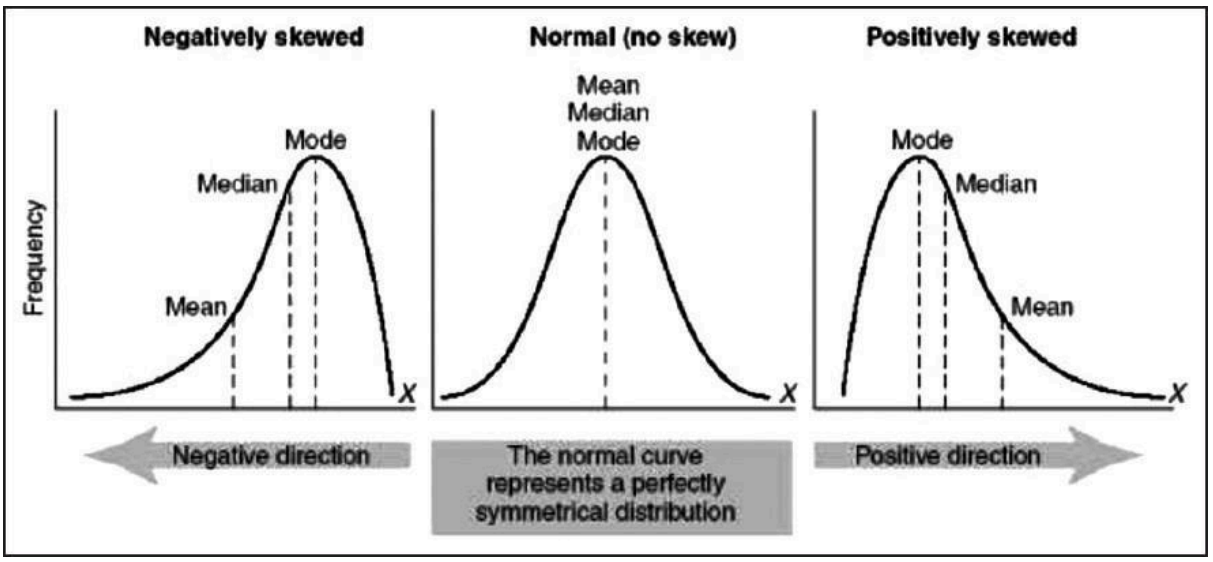
Q. Describe Hypothesis Testing. How is the statistical significance of an insight assessed?
Hypothesis Testing in statistics is used to see if a certain experiment yields meaningful results. It essentially helps to assess the statistical significance of insight by determining the odds of the results occurring by chance. The first thing is to know the null hypothesis and then state it. Then the p-value is calculated, and if the null hypothesis is true, other values are also determined. The alpha value denotes the significance and is adjusted accordingly. If the p-value is less than alpha, the null hypothesis is rejected, but if it is greater than alpha, the null hypothesis is accepted. The rejection of the null hypothesis indicates that the results obtained are statistically significant.

Q. What are observational and experimental data in statistics?
Observational data is derived from the observation of certain variables from observational studies. The variables are observed to determine any correlation between them.
Experimental data is derived from those experimental studies where certain variables are kept constant to determine any discrepancy or causality.

Q: Given a left-skewed distribution that has a median of 60, what conclusions can we draw about the mean and the mode of the data?
Left skewed distribution means the tail of the distribution is to the left and the tip is to the right. So the mean which tends to be near outliers (very large or small values) will be shifted towards the left or in other words, towards the tail.

While the mode (which represents the most repeated value) will be near the tip and the median is the middle element independent of the distribution skewness, therefore it will be smaller than the mode and more than the mean.

Mean < 60 Mode > 60



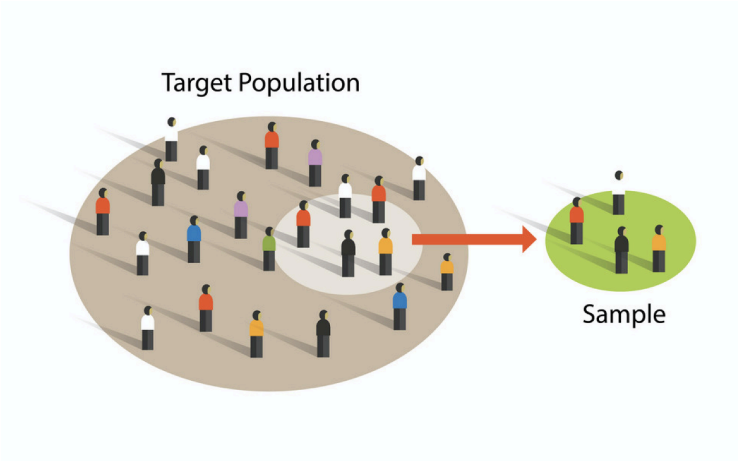
Q: What is the meaning of selection bias and how to avoid it?
Sampling bias is the phenomenon that occurs when a research study design fails to collect a representative sample of a target population. This typically occurs because the selection criteria for respondents failed to capture a wide enough sampling frame to represent all viewpoints.

The cause of sampling bias almost always owes to one of two conditions.

1. Poor methodology: In most cases, non-representative samples pop up when researchers set improper parameters for survey research. The most accurate and repeatable sampling method is simple random sampling where a large number of respondents are chosen at random. When researchers stray from random sampling (also called probability sampling), they risk injecting their own selection bias into recruiting respondents.
2. Poor execution: Sometimes data researchers craft scientifically sound sampling methods, but their work is undermined when field workers cut corners. By reverting to convenience sampling (where the only people studied are those who are easy to reach) or giving up on reaching non-responders, a field worker can jeopardise the careful methodology set up by data scientists.

The best way to avoid sampling bias is to stick to probability-based sampling methods.

These include simple random sampling, systematic sampling, cluster sampling, and stratified sampling. In these methodologies, respondents are only chosen through processes of random selection—even if they are sometimes sorted into demographic groups along the way.



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

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

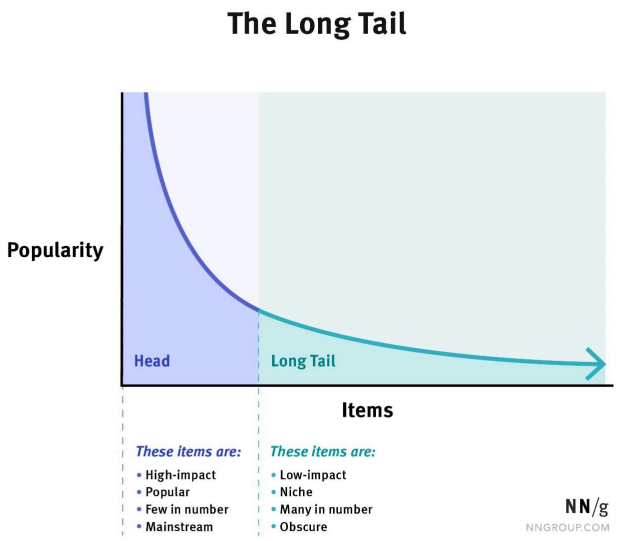
Three examples of relevant phenomena that have long tails:

1. Frequencies of languages spoken
2. Population of cities
3. Pageviews of articles

All of these follow something close to the 80-20 rule: 80% of outcomes (or outputs) result from 20% of all causes (or inputs) for any given event. This 20% forms the long tail in the distribution.

It's important to be mindful of long-tailed distributions in classification and regression problems because the least frequently occurring values make up the majority of the population.

This can ultimately change the way that you deal with outliers, and it also conflicts with some machine learning techniques with the assumption that the data is normally distributed.



Q: What is the meaning of KPI in statistics

KPI stands for key performance indicator, a quantifiable measure of performance over time for a specific objective. KPIs provide targets for teams to shoot for, milestones to gauge progress, and insights that help people across the organisation make better decisions. From finance and HR to marketing and sales, key performance indicators help every area of the business move forward at the strategic level.

KPIs are an important way to ensure your teams are supporting the overall goals of the organisation. Here are some of the biggest reasons why you need key performance indicators.

- Keep your teams aligned: Whether measuring project success or employee performance, KPIs keep teams moving in the same direction.
- Provide a health check: Key performance indicators give you a realistic look at the health of your organisation, from risk factors to financial indicators.
- Make adjustments: KPIs help you clearly see your successes and failures so you can do more of what's working, and less of what's not.
- Hold your teams accountable: Make sure everyone provides value with key performance indicators that help employees track their progress and help managers move things along.

Types of KPIs Key performance indicators come in many flavours. While some are used to measure monthly progress against a goal, others have a longer-term focus. The one thing all KPIs have in common is that they're tied to strategic goals. Here's an overview of some of the most common types of KPIs.

- Strategic: These big-picture key performance indicators monitor organisational goals. Executives typically look to one or two strategic KPIs to find out how the organisation is doing at any given time. Examples include return on investment, revenue and market share.

- Operational: These KPIs typically measure performance in a shorter time frame, and are focused on organisational processes and efficiencies. Some examples include sales by region, average monthly transportation costs and cost per acquisition (CPA).
- Functional Unit: Many key performance indicators are tied to specific functions, such finance or IT. While IT might track time to resolution or average uptime, finance KPIs track gross profit margin or return on assets. These functional KPIs can also be classified as strategic or operational.
- Leading vs Lagging: Regardless of the type of key performance indicator you define, you should know the difference between leading indicators and lagging indicators. While leading KPIs can help predict outcomes, lagging KPIs track what has already happened. Organisations use a mix of both to ensure they're tracking what's most important.

Q: Say you flip a coin 10 times and observe only one head. What would be the null hypothesis and p-value for testing whether the coin is fair or not?
The null hypothesis is that the coin is fair, and the alternative hypothesis is that the coin is biased. The p-value is the probability of observing the results obtained given that the null hypothesis is true, in this case, the coin is fair.
In total for 10 flips of a coin, there are 2^10 = 1024 possible outcomes and in only 10 of them there are 9 tails and one head.
Hence, the exact probability of the given result is the p-value, which is 10/1024 = 0.0098.
Therefore, with a significance level set, for example, at 0.05, we can reject the null hypothesis.

Q: You are testing hundreds of hypotheses, each with a t-test. What considerations would you take into account when doing this?
The main consideration when we have a large number of tests is that the probability of getting a significant test due to chance alone increases. This will increase the type 1 error (rejecting the null hypothesis when it's actually true).

Therefore we need to consider the Bonferroni Effect which happens when we make many tests. Ex. If our significance level is 0.05 but we made a 100 test it means that the probability of getting a value inside the rejection rejoin is 0.0005, not 0.05 so here we need to use another significance level which's called alpha star = significance level /K Where K is the number of the tests.

Q: What general conditions must be satisfied for the central limit theorem to hold?
In order to apply the central limit theorem, there are four conditions that must be met:
1.** Randomization:** The data must be sampled randomly such that every member in a population has an equal probability of being selected to be in the sample.
2. Independence: The sample values must be independent of each other.
3. The 10% Condition: When the sample is drawn without replacement, the sample size should be no larger than 10% of the population.
4. Large Sample Condition: The sample size needs to be sufficiently large.

Q: What is skewness discussing two methods to measure it?

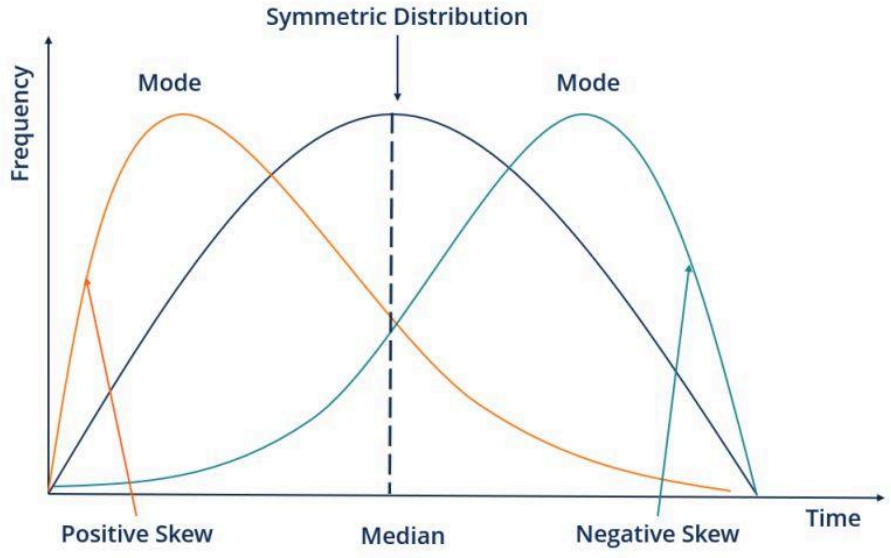
Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data. If the curve is shifted to the left or to the right, it is said to be skewed.Skewness can be quantified as a representation of the extent to which a given distribution varies from a normal distribution.
There are two main types of skewness:

- A negative skew which refers to a longer or fatter tail on the left side of the distribution
- A positive skew refers to a longer or fatter tail on the right.

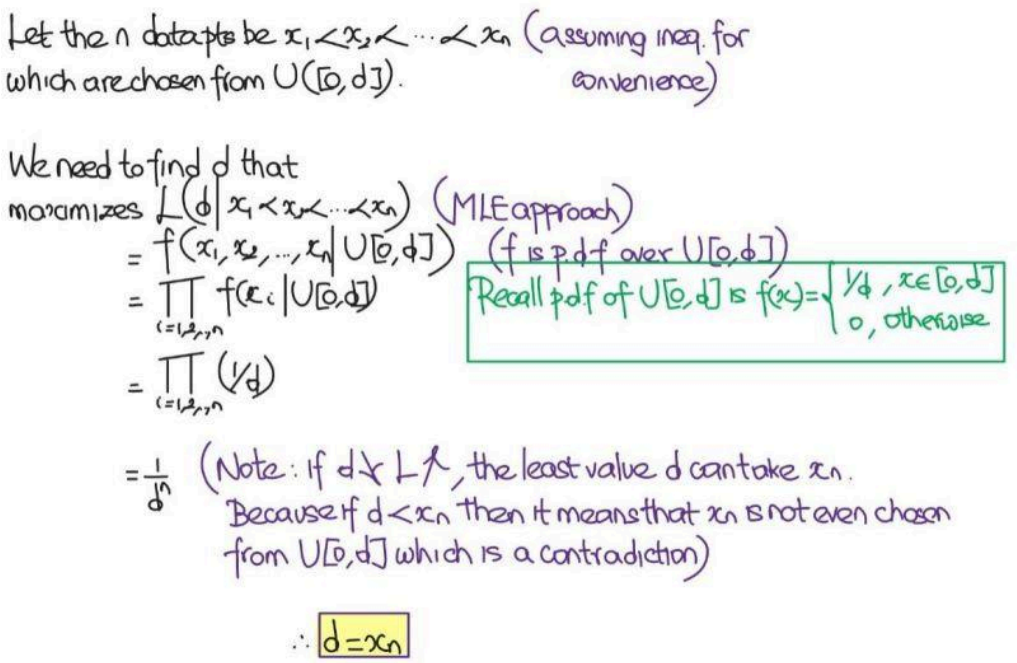
These two skews refer to the direction or weight of the distribution.

The mean of positively skewed data will be greater than the median. In a negatively skewed distribution, the exact opposite is the case: the mean of negatively skewed data will be less than the median. If the data graphs symmetrically, the distribution has zero skewness, regardless of how long or fat the tails are.

There are several ways to measure skewness. Pearson's first and second coefficients of skewness are two common methods. Pearson's first coefficient of skewness, or Pearson mode skewness, subtracts the mode from the mean and divides the difference by the standard deviation.
Pearson's second coefficient of skewness, or Pearson median skewness, subtracts the median from the mean, multiplies the difference by three, and divides the product by the standard deviation.



Q. You sample from a uniform distribution [0, d] n times. What is your best estimate of d?
Intuitively it is the maximum of the sample points. Here's the mathematical proof is in the figure below:



Q: Discuss the Chi-square, ANOVA, and t-test

Chi-square test A statistical method is used to find the difference or correlation between the observed and expected categorical variables in the dataset.

Example: A food delivery company wants to find the relationship between gender, location, and food choices of people.

It is used to determine whether the difference between 2 categorical variables is:

- Due to chance or
- Due to relationship

Analysis of Variance (ANOVA) is a statistical formula used to compare variances across the means (or average) of different groups. A range of scenarios uses it to determine if there is any difference between the means of different groups.

t_test is a statistical method for the comparison of the mean of the two groups of the normally distributed sample(s).

It comes in various types such as:

1. One sample t-test:

Used to compare the mean of a sample and the population.

2. Two sample t-tests:

Used to compare the mean of two independent samples and whether their population is statistically different.

3. Paired t-test:

Used to compare means of different samples from the same group.

Q: Say you have two subsets of a dataset for which you know their means and standard deviations. How do you calculate the blended mean and standard deviation of the total dataset? Can you extend it to K subsets?

Read here

Q: What is the relationship between the significance level and the confidence level in Statistics?

Confidence level = 1 - significance level.
It's closely related to hypothesis testing and confidence intervals.
Significance Level according to the hypothesis testing literature means the probability of Type-I error one is willing to tolerate.

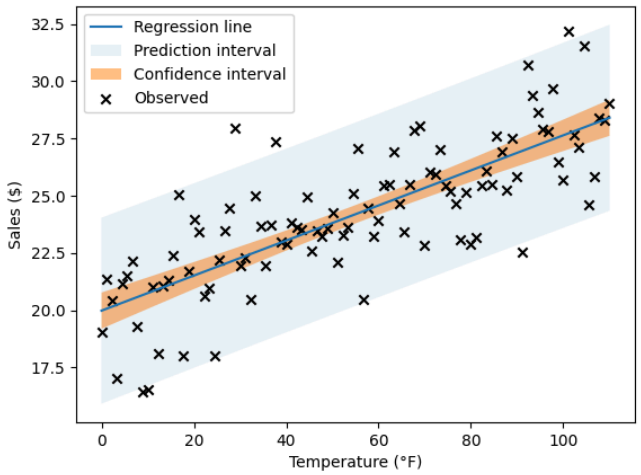
Confidence Level according to the confidence interval literature means the probability in terms of the true parameter value lying inside the confidence interval. They are usually written in percentages.

Q: What is the Law of Large Numbers in statistics and how can it be used in data science ?
The law of large numbers states that as the number of trials in a random experiment increases, the average of the results obtained from the experiment approaches the expected value. In statistics, it's used to describe the relationship between sample size and the accuracy of statistical estimates.

In data science, the law of large numbers is used to understand the behaviour of random variables over many trials. It's often applied in areas such as predictive modelling, risk assessment, and quality control to ensure that data-driven decisions are based on a robust and accurate representation of the underlying patterns in the data.
The law of large numbers helps to guarantee that the average of the results from a large number of independent and identically distributed trials will converge to the expected value, providing a foundation for statistical inference and hypothesis testing.

Q: What is the difference between a confidence interval and a prediction interval, and how do you calculate them?
A confidence interval is a range of values that is likely to contain the true value of a population parameter with a certain level of confidence. It is used to estimate the precision or accuracy of a sample statistic, such as a mean or a proportion, based on a sample from a larger population.
For example, if we want to estimate the average height of all adults in a certain region, we can take a random sample of individuals from that region and calculate the sample mean height. Then we can construct a confidence interval for the true population mean height, based on the sample mean and the sample size, with a certain level of confidence, such as 95%. This means that if we repeat the sampling process many times, 95% of the resulting intervals will contain the true population mean height.
The formula for a confidence interval is: confidence interval = sample statistic +/- margin of error
The margin of error depends on the sample size, the standard deviation of the population (or the sample, if the population standard deviation is unknown), and the desired level of confidence. For example, if the sample size is larger or the standard deviation is smaller, the margin of error will be smaller, resulting in a narrower confidence interval.
A prediction interval is a range of values that is likely to contain a future observation or outcome with a certain level of confidence. It is used to estimate the uncertainty or variability of a future value based on a statistical model and the observed data.
For example, if we have a regression model that predicts the sales of a product based on its price and advertising budget, we can use a prediction interval to estimate the range of possible sales for a new product with a certain price and advertising budget, with a certain

level of confidence, such as 95%. This means that if we repeat the prediction process many times, 95% of the resulting intervals will contain the true sales value.
The formula for a prediction interval is: prediction interval = point estimate +/- margin of error
The point estimate is the predicted value of the outcome variable based on the model and the input variables. The margin of error depends on the residual standard deviation of the model, which measures the variability of the observed data around the predicted values, and the desired level of confidence. For example, if the residual standard deviation is larger or the level of confidence is higher, the margin of error will be larger, resulting in a wider prediction interval.



Q. What is an outlier?
Outliers can be defined as the data points within a data set that varies largely in comparison to other observations. Depending on its cause, an outlier can decrease the accuracy as well as the efficiency of a model. Therefore, it is crucial to remove them from the data set.

Q. How to screen for outliers in a data set?
There are many ways to screen and identify potential outliers in a data set. Two key methods are described below –

- Standard deviation/z-score – Z-score or standard score can be obtained in a normal distribution by calculating the size of one standard deviation and multiplying it by 3. The data points outside the range are then identified. The Z-score is measured from the mean. If the z-score is positive, it means the data point is above average.

If the z-score is negative, the data point is below average.
If the z-score is close to zero, the data point is close to average.
If the z-score is above or below 3, it is an outlier and the data point is considered unusual.
The formula for calculating a z-score is –
z= data point−mean/standard deviation OR z=x−μ/ σ

- Interquartile range (IQR) – IQR, also called midspread, is a method to identify outliers and can be described as the range of values that occur throughout the length of the middle of 50% of a data set. It is simply the difference between two extreme data points within the observation.

IQR=Q3 – Q1
Other methods to screen outliers include Isolation Forests, Robust Random Cut Forests, and DBScan clustering.

Q. What is the meaning of an inlier?
An Inliner is a data point within a data set that lies at the same level as the others. It is usually an error and is removed to improve the model accuracy. Unlike outliers, inlier is hard to find and often requires external data for accurate identification.

Q. What is the meaning of six sigma in statistics?
Six sigma in statistics is a quality control method to produce an error or defect-free data set. Standard deviation is known as Sigma or σ. The more the standard deviation, the less likely that process performs with accuracy and causes a defect. If a process outcome is 99.99966% error-free, it is considered six sigma. A six sigma model works better than 1σ, 2σ, 3σ, 4σ, 5σ processes and is reliable enough to produce defect-free work.

Q. What is the meaning of KPI in statistics?
KPI is an acronym for a key performance indicator. It can be defined as a quantifiable measure to understand whether the goal is being achieved or not. KPI is a reliable metric to measure the performance level of an organisation or individual with respect to the objectives. An example of KPI in an organisation is the expense ratio.

Q. What is the Pareto principle?
Also known as the 80/20 rule, the Pareto principle states that 80% of the effects or results in an experiment are obtained from 20% of the causes. A simple example is – 20% of sales come from 80% of customers.

Q. What is the Law of Large Numbers in statistics?
According to the law of large numbers, an increase in the number of trials in an experiment will result in a positive and proportional increase in the results coming closer to the expected value. As an example, let us check the probability of rolling a six-sided dice three times. The expected value obtained is far from the average value. And if we roll a dice a large number of times, we will obtain the average result closer to the expected value (which is 3.5 in this case).

Q. What are some of the properties of a normal distribution?
Also known as Gaussian distribution, Normal distribution refers to the data which is symmetric to the mean, and data far from the mean is less frequent in occurrence. It appears as a bell-shaped curve in graphical form, which is symmetrical along the axes.
The properties of a normal distribution are –

- Symmetrical – The shape changes with that of parameter values
- Unimodal – Has only one mode.
- Mean – the measure of central tendency
- Central tendency – the mean, median, and mode lie at the centre, which means that they are all equal, and the curve is perfectly symmetrical at the midpoint.

Q. How would you describe a ‘p-value’?
P-value in statistics is calculated during hypothesis testing, and it is a number that indicates the likelihood of data occurring by a random chance. If a p-value is 0.5 and is less than alpha, we can conclude that there is a probability of 5% that the experiment results occurred by chance, or you can say, 5% of the time, we can observe these results by chance.

Q. How can you calculate the p-value using MS Excel?
The formula used in MS Excel to calculate p-value is –
=tdist(x,deg_freedom,tails)
The p-value is expressed in decimals in Excel. Here are the steps to calculate it –

- Find the Data tab
- On the Analysis tab, click on the data analysis icon
- Select Descriptive Statistics and then click OK
- Select the relevant column
- Input the confidence level and other variables

Q. What are the types of biases that you can encounter while sampling?
Sampling bias occurs when you lack the fair representation of data samples during an investigation or a survey. The six main types of biases that one can encounter while sampling are –

- Undercoverage bias
- Observer Bias
- Survivorship bias
- Self-Selection/Voluntary Response Bias
- Recall Bias
- Exclusion Bias

Q. What is cherry-picking, P-hacking, and significance chasing?
Cherry-picking can be defined as the practice in statistics where only that information is selected which supports a certain claim and ignores any other claim that refutes the desired conclusion.
P-hacking refers to a technique in which data collection or analysis is manipulated until significant patterns can be found who have no underlying effect whatsoever.
Significance chasing is also known by the names of Data Dredging, Data Fishing, or Data Snooping. It refers to the reporting of insignificant results as if they are almost significant.

Q. What is the difference between type I vs type II errors?
A type 1 error occurs when the null hypothesis is rejected even if it is true. It is also known as false positive.
A type 2 error occurs when the null hypothesis fails to get rejected, even if it is false. It is also known as a false negative.

Q. What is a statistical interaction?
A statistical interaction refers to the phenomenon which occurs when the influence of an input variable impacts the output variable. A real-life example includes the interaction of adding sugar to the stirring of tea. Neither of the two variables has an impact on sweetness, but it is the combination of these two variables that do.

Q. Give an example of a data set with a non-Gaussian distribution?
A non-Gaussian distribution is a common occurrence in many processes in statistics. This happens when the data naturally follows a non-normal distribution with data clumped on one side or the other on a graph. For example, the growth of bacteria follows a non-Gaussian or exponential distribution naturally and a Weibull distribution.

Q. What is the Binomial Distribution Formula?
The binomial distribution formula is:
b(x; n, P) = nCx * Px * (1 – P)n – x
Where:
b = binomial probability
x = total number of “successes” (pass or fail, heads or tails, etc.)
P = probability of success on an individual trial
n = number of trials

Q. What are the criteria that Binomial distributions must meet?
Here are the three main criteria that Binomial distributions must meet –

- The number of observation trials must be fixed. It means that one can only find the probability of something when done only a certain number of times.
- Each trial needs to be independent. It means that none of the trials should impact the probability of other trials.
- The probability of success remains the same across all trials.

Q. What is linear regression?
In statistics, linear regression is an approach that models the relationship between one or more explanatory variables and one outcome variable. For example, linear regression can be used to quantify or model the relationship between various predictor variables such as age, gender, genetics, and diet on height, outcome variables.

Q. What are the assumptions required for linear regression?
Four major assumptions for linear regression are as under –

- There’s a linear relationship between the predictor (independent) variables and the outcome (dependent) variable. It means that the relationship between X and the mean of Y is linear.
- The errors are normally distributed with no correlation between them. This process is known as Autocorrelation.
- There is an absence of correlation between predictor variables. This phenomenon is called multicollinearity.
- The variation in the outcome or response variable is the same for all values of independent or predictor variables. This phenomenon of assumption of equal variance is known as homoscedasticity.

Q. What are some of the low and high-bias Machine Learning algorithms?
Some of the widely used low and high-bias Machine Learning algorithms are –
Low bias -Decision trees, Support Vector Machines, k-Nearest Neighbors, etc.
High bias -Linear Regression, Logistic Regression, Linear Discriminant Analysis, etc.
Check out the free course on Statistical Methods For Decision Making.

Q. When should you use a t-test vs a z-test?
The z-test is used for hypothesis testing in statistics with a normal distribution. It is used to determine population variance in the case where a sample is large.
The t-test is used with a t-distribution and used to determine population variance when you have a small sample size.

In case the sample size is large or n>30, a z-test is used. T-tests are helpful when the sample size is small or n<30.

Q. What is the equation for confidence intervals for means vs for proportions?
To calculate the confidence intervals for mean, we use the following equation –
For n > 30
Use the Z table for the standard normal distribution.
For n<30
Use the t table with df=n-1
Confidence Interval for the Population Proportion –

Q. What is the empirical rule?
In statistics, the empirical rule states that every piece of data in a normal distribution lies within three standard deviations of the mean. It is also known as the 68–95–99.7 rule. According to the empirical rule, the percentage of values that lie in a normal distribution follow the 68%, 95%, and 99.7% rule. In other words, 68% of values will fall within one standard deviation of the mean, 95% will fall within two standard deviations, and 99.75 will fall within three standard deviations of the mean.

Q. How are confidence tests and hypothesis tests similar? How are they different?
Confidence tests and hypothesis tests both form the foundation of statistics.
The confidence interval holds importance in research to offer a strong base for research estimations, especially in medical research. The confidence interval provides a range of values that helps in capturing the unknown parameter.

Hypothesis testing is used to test an experiment or observation and determine if the results did not occur purely by chance or luck using the below formula where ‘p’ is some parameter. Confidence and hypothesis testing are inferential techniques used to either estimate a parameter or test the validity of a hypothesis using a sample of data from that data set. While the confidence interval provides a range of values for an accurate estimation of the precision of that parameter, hypothesis testing tells us how confident we are inaccurately drawing conclusions about a parameter from a sample. Both can be used to infer population parameters in tandem.
In case we include 0 in the confidence interval, it indicates that the sample and population have no difference. If we get a p-value that is higher than alpha from hypothesis testing, it means that we will fail to reject the null hypothesis.

Q. What general conditions must be satisfied for the central limit theorem to hold?
Here are the conditions that must be satisfied for the central limit theorem to hold –

- The data must follow the randomization condition which means that it must be sampled randomly.
- The Independence Assumptions dictate that the sample values must be independent of each other.
- Sample sizes must be large. They must be equal to or greater than 30 to be able to hold CLT. Large sample size is required to hold the accuracy of CLT to be true.

Q. What is Random Sampling? Give some examples of some random sampling techniques.

Random sampling is a sampling method in which each sample has an equal probability of being chosen as a sample. It is also known as probability sampling.
Let us check four main types of random sampling techniques –

- Simple Random Sampling technique – In this technique, a sample is chosen randomly using randomly generated numbers. A sampling frame with the list of members of a population is required, which is denoted by 'n'. Using Excel, one can randomly generate a number for each element that is required.
- Systematic Random Sampling technique -This technique is very common and easy to use in statistics. In this technique, every k'th element is sampled. For instance, one element is taken from the sample and then the next while skipping the pre-defined amount or 'n'.

In a sampling frame, divide the size of the frame N by the sample size (n) to get 'k', the index number. Then pick every k'th element to create your sample.

- Cluster Random Sampling technique -In this technique, the population is divided into clusters or groups in such a way that each cluster represents the population. After that, you can randomly select clusters to sample.
- Stratified Random Sampling technique – In this technique, the population is divided into groups that have similar characteristics. Then a random sample can be taken from each group to ensure that different segments are represented equally within a population.

Q. What is the difference between population and sample in inferential statistics?
A population in inferential statistics refers to the entire group we take samples from and are used to draw conclusions. A sample, on the other hand, is a specific group we take data from and this data is used to calculate the statistics. Sample size is always less than that of the population.

Q. What are descriptive statistics?
Descriptive statistics are used to summarise the basic characteristics of a data set in a study or experiment. It has three main types –

- Distribution – refers to the frequencies of responses.
- Central Tendency – gives a measure or the average of each response.
- Variability – shows the dispersion of a data set.

Q. What are quantitative data and qualitative data?
Qualitative data is used to describe the characteristics of data and is also known as Categorical data. For example, how many types. Quantitative data is a measure of numerical values or counts. For example, how much or how often. It is also known as Numeric data.

Q. How to calculate range and interquartile range?
The range is the difference between the highest and the lowest values whereas the Interquartile range is the difference between upper and lower medians.
Range (X) = Max(X) – Min(X)

IQR = Q3 – Q1
Here, Q3 is the third quartile (75 percentile)
Here, Q1 is the first quartile (25 percentile)

Q. What is the meaning of standard deviation?
Standard deviation gives the measure of the variation of dispersion of values in a data set. It represents the differences of each observation or data point from the mean.
$(\sigma) = \sqrt{(\sum(x-\mu)2 / n)}$
Where the variance is the square of standard deviation.

Q. What is the relationship between mean and median in normal distribution?
In a normal distribution, the mean and the median are equal.

Q. What is the left-skewed distribution and the right-skewed distribution?
In the left-skewed distribution, the left tail is longer than the right side.
Mean < median < mode
In the right-skewed distribution, the right tail is longer. It is also known as positive-skew distribution.
Mode < median < mean

Q. How to convert normal distribution to standard normal distribution?
Any point (x) from the normal distribution can be converted into standard normal distribution (Z) using this formula –
$Z(\text{standardized}) = (x-\mu) / \sigma$
Here, Z for any particular x value indicates how many standard deviations x is away from the mean of all values of x.

Q. What can you do with an outlier?
Outliers affect A/B testing and they can be either removed or kept according to what situation demands or the data set requirements.
Here are some ways to deal with outliers in data –

- Filter out outliers especially when we have loads of data.
- If a data point is wrong, it is best to remove the outliers.
- Alternatively, two options can be provided – one with outliers and one without.
- During post-test analysis, outliers can be removed or modified. The best way to modify them is to trim the data set.
- If there are a lot of outliers and results are critical, then it is best to change the value of the outliers to other variables. They can be changed to a value that is representative of the data set.
- When outliers have meaning, they can be considered, especially in the case of mild outliers.

Q. How to detect outliers?
The best way to detect outliers is through graphical means. Apart from that, outliers can also be detected through the use of statistical methods using tools such as Excel, Python, SAS, among others. The most popular graphical ways to detect outliers include box plot and scatter plot.

Q. Why do we need sample statistics?
Sampling in statistics is done when population parameters are not known, especially when the population size is too large.

Q. What is the relationship between standard error and margin of error?
Margin of error = Critical value X Standard deviation for the population
and
Margin of error = Critical value X Standard error of the sample.
The margin of error will increase with the standard error.

Q. What is the proportion of confidence intervals that will not contain the population parameter?
Alpha is the probability in a confidence interval that will not contain the population parameter.
$\alpha = 1 - CL$
Alpha is usually expressed as a proportion. For instance, if the confidence level is 95%, then alpha would be equal to 1-0.95 or 0.05.

Q. What is skewness?
Skewness provides the measure of the symmetry of a distribution. If a distribution is not normal or asymmetrical, it is skewed. A distribution can exhibit positive skewness or negative skewness if the tail on the right is longer and the tail on the left side is longer, respectively.

Q. What is the meaning of covariance?
In statistics, covariance is a measure of association between two random variables from their respective means in a cycle.

Q. What is a confounding variable?
A confounding variable in statistics is an ‘extra’ or ‘third’ variable that is associated with both the dependent variable and the independent variable, and it can give a wrong estimate that provides useless results.
For example, if we are studying the effect of weight gain, then lack of workout will be the independent variable, and weight gain will be the dependent variable. In this case, the amount of food consumption can be the confounding variable as it will mask or distort the effect of other variables in the study. The effect of weather can be another confounding variable that may later the experiment design.

Q. What does it mean if a model is heteroscedastic?
A model is said to be heteroscedastic when the variation in errors comes out to be inconsistent. It often occurs in two forms – conditional and unconditional.

Q. What is selection bias and why is it important?
Selection bias is a term in statistics used to denote the situation when selected individuals or a group within a study differ in a manner from the population of interest that they give systematic error in the outcome.
Typically selection bias can be identified using bivariate tests apart from using other methods of multiple regression such as logistic regression.
It is crucial to understand and identify selection bias to avoid skewing results in a study. Selection bias can lead to false insights about a particular population group in a study.
Different types of selection bias include –

- Sampling bias – It is often caused by non-random sampling. The best way to overcome this is by drawing from a sample that is not self-selecting.

- Participant attrition – The dropout rate of participants from a study constitutes participant attrition. It can be avoided by following up with the participants who dropped off to determine if the attrition is due to the presence of a common factor between participants or something else.
- Exposure – It occurs due to the incorrect assessment or the lack of internal validity between exposure and effect in a population.
- Data – It includes dredging of data and cherry-picking and occurs when a large number of variables are present in the data causing even bogus results to appear significant.
- Time-interval – It is a sampling error that occurs when observations are selected from a certain time period only. For example, analyzing sales during the Christmas season.
- Observer selection- It is a kind of discrepancy or detection bias that occurs during the observation of a process and dictates that for the data to be observable, it must be compatible with the life that observes it.

Q. What does autocorrelation mean?
Autocorrelation is a representation of the degree of correlation between the two variables in a given time series. It means that the data is correlated in a way that future outcomes are linked to past outcomes. Autocorrelation makes a model less accurate because even errors follow a sequential pattern.

Q. What does Design of Experiments mean?
The Design of Experiments or DOE is a systematic method that explains the relationship between the factors affecting a process and its output. It is used to infer and predict an outcome by changing the input variables.

Q. What is Bessel's correction?
Bessel's correction advocates the use of n-1 instead of n in the formula of standard deviation. It helps to increase the accuracy of results while analysing a sample of data to derive more general conclusions.

Q. What types of variables are used for Pearson's correlation coefficient?
Variables (both the dependent and independent variables) used for Pearson's correlation coefficient must be quantitative. It will only test for the linear relationship between two variables.

Q. What is the use of Hash tables in statistics?
In statistics, hash tables are used to store key values or pairs in a structured way. It uses a hash function to compute an index into an array of slots in which the desired elements can be searched.

Q. Does symmetric distribution need to be unimodal?
Symmetrical distribution does not necessarily need to be unimodal, they can be skewed or asymmetric. They can be bimodal with two peaks or multimodal with multiple peaks.

Q. What is the benefit of using box plots?

Boxplot is a visually effective representation of two or more data sets and facilitates quick comparison between a group of histograms.

Q. What is the meaning of sensitivity in statistics?
Sensitivity refers to the accuracy of a classifier in a test. It can be calculated using the formula –
Sensitivity = Predicted True Events/Total number of Events

Q. What is the difference between the first quartile, the second quartile, and the third quartile?
The first quartile is denoted by Q1 and it is the median of the lower half of the data set.
The second quartile is denoted by Q2 and is the median of the data set.
The third quartile is denoted by Q3 and is the median of the upper half of the data set.
About 25% of the data set lies above Q3, 75% lies below Q3 and 50% lies below Q2. The Q1, Q2, and Q3 are the 25th, 50th, and 75th percentile respectively.

Q. What is kurtosis?
Kurtosis is a measure of the degree of the extreme values present in one tail of distribution or the peaks of frequency distribution as compared to the others. The standard normal distribution has a kurtosis of 3 whereas the values of symmetry and kurtosis between -2 and +2 are considered normal and acceptable. The data sets with a high level of kurtosis imply that there is a presence of outliers. One needs to add data or remove outliers to overcome this problem. Data sets with low kurtosis levels have light tails and lack outliers.

Q. What is a bell-curve distribution?
A bell-curve distribution is represented by the shape of a bell and indicates normal distribution. It occurs naturally in many situations especially while analysing financial data. The top of the curve shows the mode, mean and median of the data and is perfectly symmetrical. The key characteristics of a bell-shaped curve are –

- The empirical rule says that approximately 68% of data lies within one standard deviation of the mean in either of the directions.
- Around 95% of data falls within two standard deviations and
- Around 99.7% of data fall within three standard deviations in either direction.