Summary

Probability and Probability Distributions

Exploratory data analysis (EDA) helps you discover patterns in data using various techniques and approaches. As you learnt, EDA is one of the most important steps in the data analysis process. It is also the part on which data analysts spend most of their time.

However, sometimes, you may require a huge amount of data for your analysis, which would need a lot of time and resources. In such situations, you are forced to work with a sample of the data instead of the entire data.

Situations like these arise all the time in big companies, such as Amazon. For example, the Amazon QC (Quality Control) department wants to know what proportion of the products in its warehouses is defective. So, instead of going through all of its products (which would be a lot), the Amazon QC team can simply check a small sample of 1,000 products and find the defect rate for it (i.e., the proportion of defective products). Then, based on the defect rate of this sample, the team can 'infer' the defect rate for all the products in its warehouses. This process of 'inferring' insights from sample data is called 'inferential statistics'.

Introduction to Probability

Inferential statistics is used to make an inference about a population based on a sample. Note that a population is a complete set of data, whereas a sample is a subset of a population.

Probability refers to the likelihood of an event occurring in an experiment. It can be calculated using the following formula:

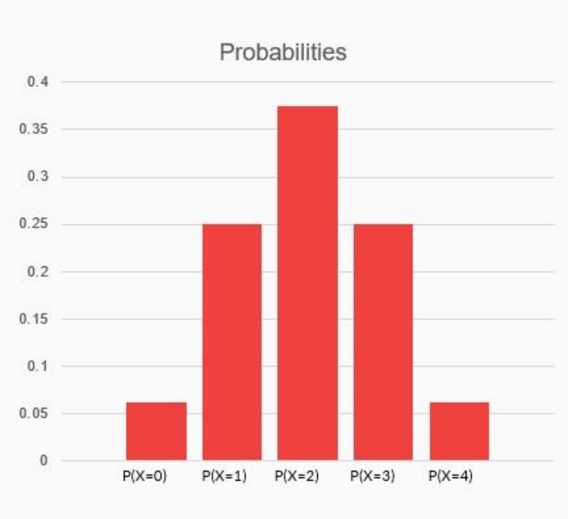
$$\text{Probability of A} = \frac{\text{Number of Favourable Outcomes of experiment A}}{\text{Total Number of Outcomes of experiment A}}$$

A random variable is a variable whose values depend on the outcomes of a random experiment. The sample space represents all the possible values of a random variable of a particular experiment.

Probability Distributions

A probability distribution for X is a form of representation that tells you the probability of all the possible values of X. It could be a table, a chart or an equation.

A probability distribution looks similar to a frequency distribution, but with a different scale. For example, the probability distribution and the frequency distribution (histogram) for the upGrad ball game are depicted in the graph provided below.



Expected Value

The expected value of a variable, X, is the value of X that you would 'expect' to get after performing an experiment once. It is also called the expectation, average or mean value. Mathematically, for a random variable X that can take the values x1, x2, x3...xn, the expected value (EV) can be given as follows:

EV = x1 × P(X = x1) + x2 × P(X = x2) + x3 × P(X = x3) +.....+ xn × P(X = xn)

Suppose you are trying to find the expected value of the number of red balls in the upGrad ball game. The random variable X, which is the number of red balls that the player gets after playing the game once, can take the values 0, 1, 2, 3 and 4. So, the expected value of the number of red balls will be calculated as follows:

EV(X) = 0 * P(X = 0) + 1 * P(X = 1) + 2 * P(X = 2) + 3 * P(X = 3) + 4 * P(X = 4)

EV(X) = 0 * (0.0625) + 1 * (0.25) + 2 * (0.375) + 3 * (0.25) + 4 * (0.0625) = 2

However, you may or may not get two red balls in one game. Also, the expected value, i.e., the value that you would 'expect' to get after playing one game, does not have to be the value that will appear in that game. So, the expected value is actually the average value of X that you will get after playing the game an infinite number of times.

Hence, the expected value should be interpreted as the average value that you get after the experiment has been conducted an infinite number of times. For example, the expected value of the number of red balls is two; this means that if you conduct the experiment (play the game) an infinite number of times, then the average number of red balls that you will get after playing one game will be two.

Discrete Probability Distribution

Recall the upGrad ball game that you played and analysed in the previous session. In this game, to quantify the problem, you initially defined the random variable X as the number of red balls obtained, and the values that X could take were 0, 1, 2, 3 and 4. Not that these values are discrete, i.e., the random variable X can take only discrete values and can never take continuous values in between, such as 0.99 or 1.01. Such events where the random variable can take only discrete values are represented using discrete probability distributions.

You learnt about the following two types of discrete probability distributions:

1. Uniform distribution
2. Binomial distribution

A discrete uniform distribution is a probability distribution that has 'n' discrete outcomes, and the probability of each of these outcomes is the same, that is, 1/n. Hence, the probability of each outcome is exactly the same. The most basic example in this case is rolling a die. When you roll an unbiased die, the chances of getting any of the numbers are equally probable.

You also learnt about another way of representing probabilities: cumulative probability distribution.

A cumulative probability is represented as follows:

$F(X) = P(X \leq x)$

For example, $F(2) = P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$

Binomial Probability Distribution

Now, let's understand the binomial experiment with the help of an example. Suppose you want to find the probability of getting heads twice on tossing a coin five times. Considering this example, the properties of a binomial experiment are as follows:

1. The experiment consists of a sequence of n identical and independent trials. Therefore, none of the trials affects the probability of the subsequent trial. In the coin-tossing experiment, each trial is a single coin toss, and each coin toss is independent of the other coin tosses.

2. Essentially, each trial has two possible outcomes: success and failure. In the coin-tossing experiment, there are two possible outcomes, heads and tails, where heads denotes 'success' and tails denotes 'failure'.

3. The probability of success, which is equal to p, is the same for all the trials. Consequently, the probability of failure, which is equal to (1 - p), is constant for each trial. In the coin-tossing experiment, the probability of getting heads is p = 0.5, and the probability of getting tails is (1 - p) = 0.5.

Calculating Binomial Probability

The objective of this experiment is to find the probability of getting exactly x successes in n trials. Let x be the random variable for this distribution, which is defined as the number of successes in n trials. So, the values are as follows:

P(x) = Probability of x successes in n trials

As x is a random variable, it can take discrete values between 0 and n. Let's perform the following steps to calculate the probability of getting two successes in five trials.

1. The first step is to find the different combinations of getting two successes and three failures. Now, let S denote success and F denote failure. You can find the following 10 combinations:
   {SSFFF, SFSFF, SFFSF, SFFFS, FSSFF, FSFSF, FSFFS, FFSSF, FFSFS, FFFSS}

2. You can directly calculate the number of combinations using the combination formula given below:

   $$^{n}C_{k} = \frac{n!}{k!(n-k)!}$$

   where n = number of trails and k = number of successes.

   On substituting n = 5 and k = 2 from the example, we get the value as 10.

3. Since getting each combination is equally likely, you can directly calculate the probability by multiplying the probability of a single combination by 10. So, the probability of getting two successes in five trials will be 10 x P(SSFFF).

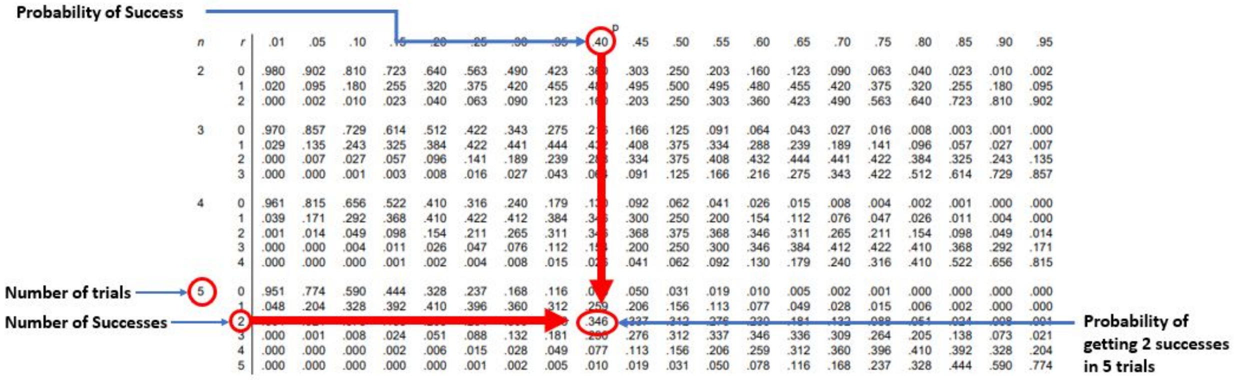
4. As you have already learnt the probability for independent trials can be obtained by multiplying the probability of each trial. So, the equation will be as follows:

10 x P(SSFFF) = 10 x P(S) x P(S) x P(F) x P(F) x P(F) = 10 x $P(S)^2$ x $(1 - P(S))^3$

Knowing that the value of P(S) = 0.4, you can substitute this value and calculate the probability of getting two successes in five trials, which is equal to 10 x 0.4^2 x 0.6^3 = 0.3456.

Hence, the probability of getting two heads on tossing a coin five times is 0.3456.

5. You can match your values with those provided in the table given below.


Probability of Success
p
Number of trials
Number of Successes
Probability of getting 2 successes in 5 trials

n	r	.01	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95
2	0	.980	.902	.810	.723	.640	.563	.490	.423	.360	.303	.250	.203	.160	.123	.090	.063	.040	.023	.010	.002
	1	.020	.095	.180	.255	.320	.375	.420	.455	.480	.495	.500	.495	.480	.455	.420	.375	.320	.255	.180	.095
	2	.000	.002	.010	.023	.040	.063	.090	.123	.160	.203	.250	.303	.360	.423	.490	.563	.640	.723	.810	.902
3	0	.970	.857	.729	.614	.512	.422	.343	.275	.216	.166	.125	.091	.064	.043	.027	.016	.008	.003	.001	.000
	1	.029	.135	.243	.325	.384	.422	.441	.444	.432	.408	.375	.334	.288	.239	.189	.141	.096	.057	.027	.007
	2	.000	.007	.027	.057	.096	.141	.189	.239	.288	.334	.375	.408	.432	.444	.441	.422	.384	.325	.243	.135
	3	.000	.000	.001	.003	.008	.016	.027	.043	.064	.091	.125	.166	.216	.275	.343	.422	.512	.614	.729	.857
4	0	.961	.815	.656	.522	.410	.316	.240	.179	.130	.092	.062	.041	.026	.015	.008	.004	.002	.001	.000	.000
	1	.039	.171	.292	.368	.410	.422	.412	.384	.346	.300	.250	.200	.154	.112	.076	.047	.026	.011	.004	.000
	2	.001	.014	.049	.098	.154	.211	.265	.311	.346	.368	.375	.368	.346	.311	.265	.211	.154	.098	.049	.014
	3	.000	.000	.004	.011	.026	.047	.076	.112	.154	.200	.250	.300	.346	.384	.412	.422	.410	.368	.292	.171
	4	.000	.000	.000	.001	.002	.004	.008	.015	.026	.041	.062	.092	.130	.179	.240	.316	.410	.522	.656	.815
5	0	.951	.774	.590	.444	.328	.237	.168	.116	.078	.050	.031	.019	.010	.005	.002	.001	.000	.000	.000	.000
	1	.048	.204	.328	.392	.410	.396	.360	.312	.259	.206	.156	.113	.077	.049	.028	.015	.006	.002	.000	.000
	2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	.346	.337	.312	.276	.230	.181	.132	.088	.051	.024	.008	.001
	3	.000	.001	.008	.024	.051	.088	.132	.181	.230	.276	.312	.337	.346	.336	.309	.264	.205	.138	.073	.021
	4	.000	.000	.000	.002	.006	.015	.028	.049	.077	.113	.156	.206	.259	.312	.360	.396	.410	.392	.328	.204
	5	.000	.000	.000	.000	.000	.001	.002	.005	.010	.019	.031	.050	.078	.116	.168	.237	.328	.444	.590	.774

Essentially, the formula for calculating the binomial probability is as follows:

$$P(X) =^n C_r p^r (1-p)^{n-r},$$

where,
n = number of trials,
r = number of success trials, and
p = probability of success.

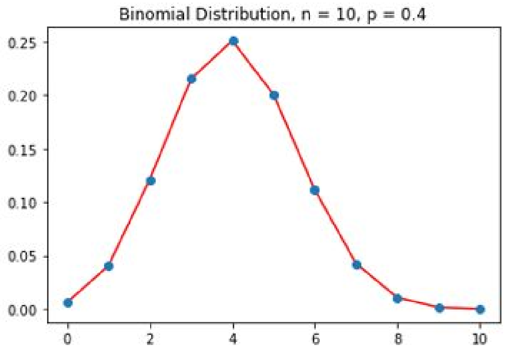
The mean (expected value), variance and the standard deviation for this distribution are as follows:

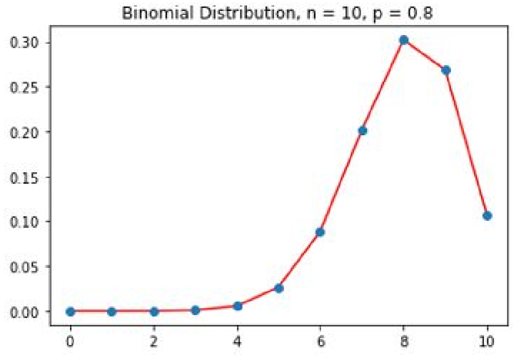
Mean (μ) = E(X) = np

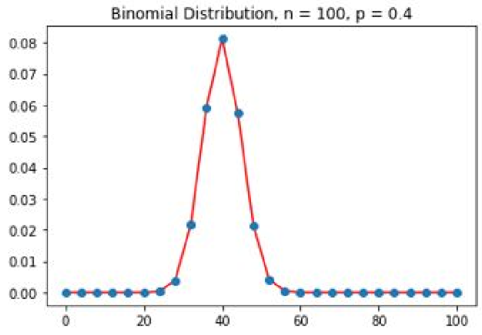
Var = npq

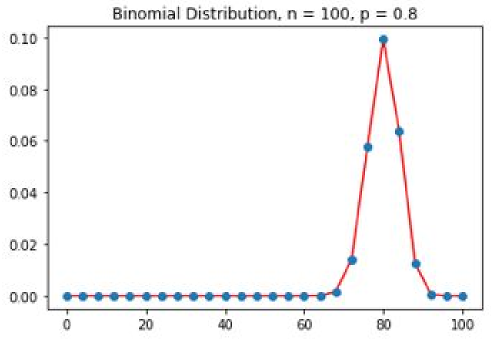
Standard deviation (σ) = $\sqrt{npq}$ = $\sqrt{np(1-p)}$

The plots of the binomial distribution for different values of n and p are depicted in the graphs given below. Since it is a discrete distribution, the probability is defined only for whole number values of x between 0 and n.







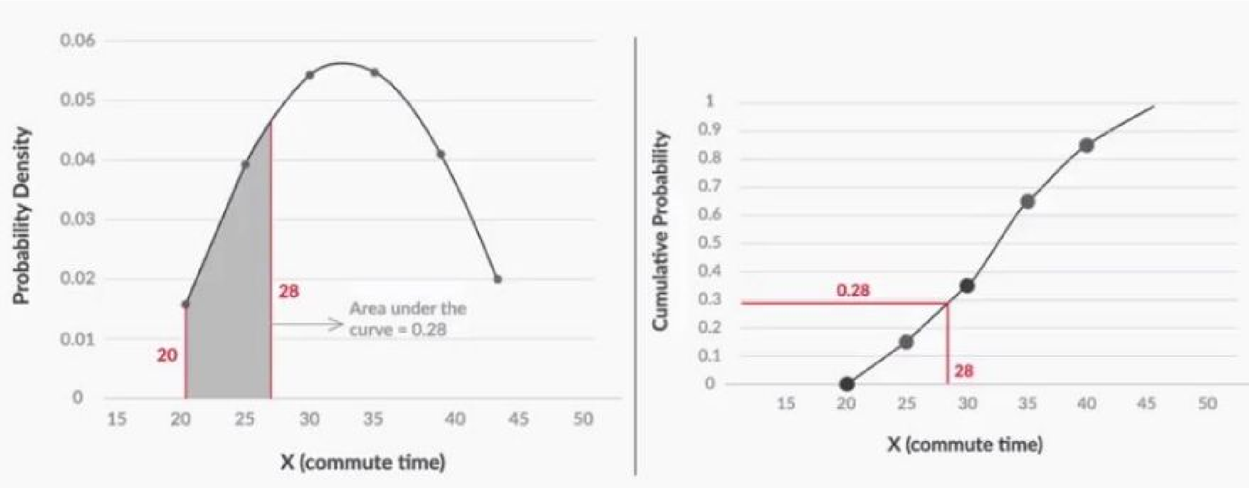


Continuous Distributions

For a continuous random variable, the probability of getting an exact value is zero. Hence, to find the probability of continuous random variables, you need to consider only intervals.

For example, for a particular company, the probability of an employee’s commute time being exactly equal to 35 minutes was zero, but the probability of the employee’s commute time being between 35 and 40 minutes was 0.2.

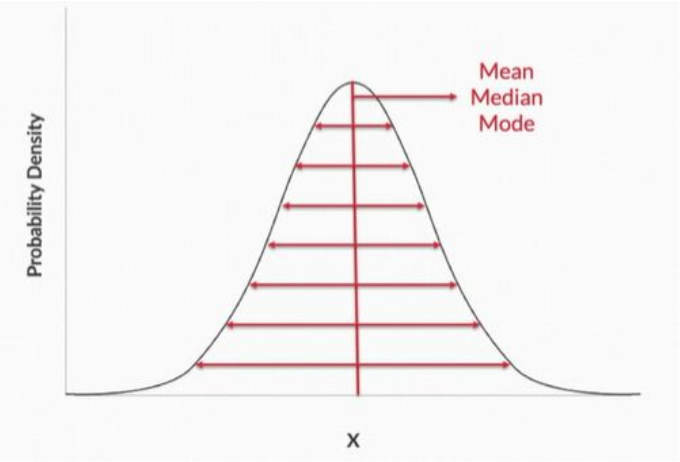
Hence, for continuous random variables, probability density functions (PDFs) and cumulative distribution functions (CDFs) are used instead of a bar chart, which is a type of distribution used for calculating the probability of discrete random variables. These functions are preferred because they talk about probability in terms of intervals.



To find the cumulative probability using a CDF, you only need to check the value of the graph. For example, F(28), that is, the probability of an employee's commute time being less than or equal to 28 minutes, is given by the value of the CDF at X = 28. In the PDF, it is given by the area under the graph, which is between X = 20 (the lowest value) and X = 28.

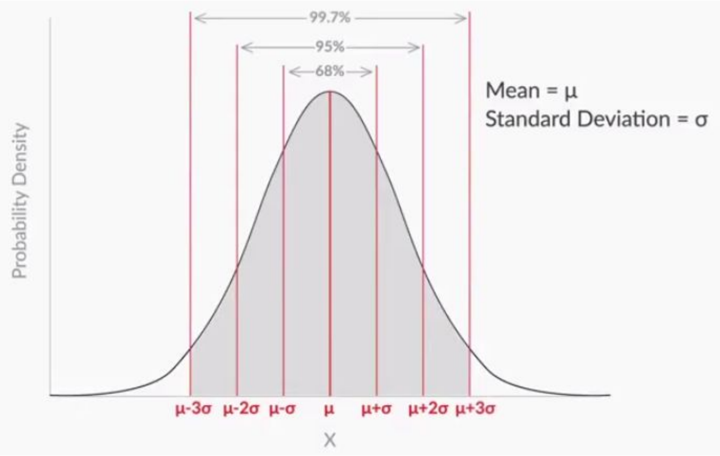
Normal Distribution

A commonly used probability density function is the normal distribution. It is a symmetric distribution, and its mean, median and mode lie at the centre. A normal distribution is the most commonly occurring probability distribution in day-to-day life.

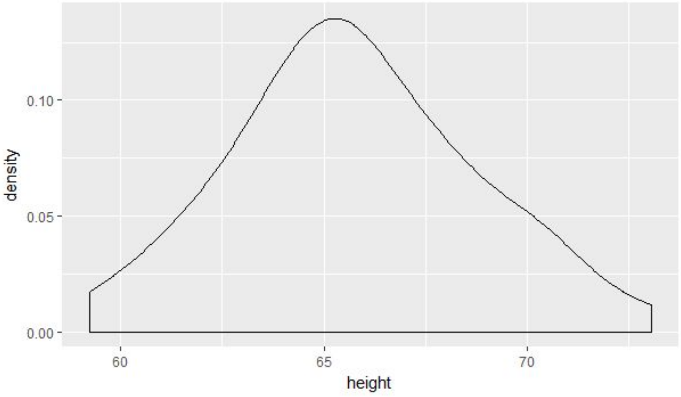


Also, a variable that is normally distributed follows the 1-2-3 rule, which states that there is a:

1. 68% probability of the variable lying within one standard deviation of the mean,
2. 95% probability of the variable lying within two standard deviations of the mean, and
3. 99.7% probability of the variable lying within three standard deviations of the mean.



A large number of naturally occurring variables are normally distributed. For example, the height of a group of adult men would be normally distributed. To try this out, we asked 50 male employees at the upGrad office for their height and then plotted the probability density function using the data. The graph for the same is given below.



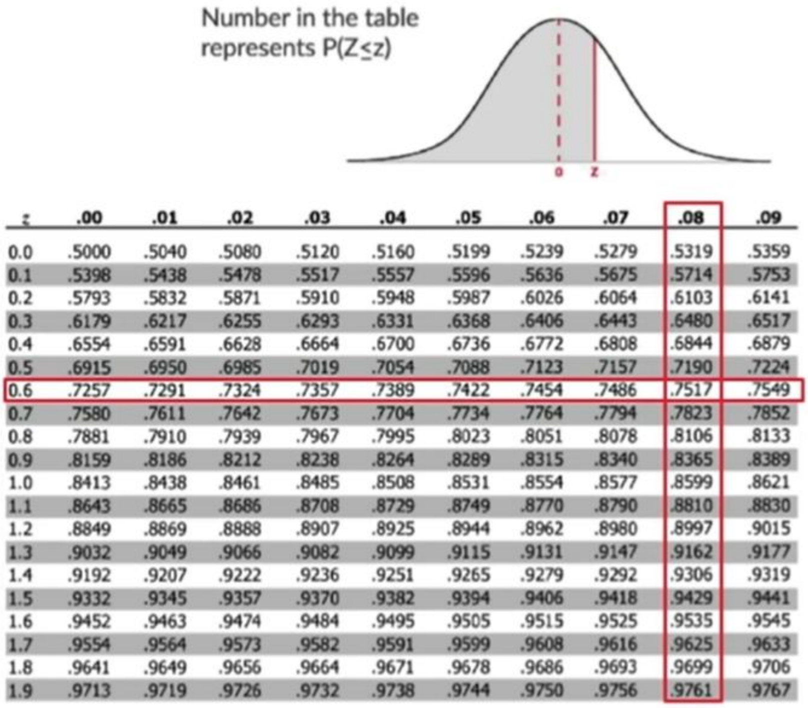
As you can see, the data is roughly normal.

Standard Normal Distribution

In order to find the probability of a normal variable, you do not need to know the value of the mean or the standard deviation; you only need to know how many standard deviations away from the mean your random variable is. This is given by the following formula:

$$z = \frac{\bar{x} - \mu}{\sigma}$$

This is called the z-score, or the standard normal variable. In fact, you can use the z-table given below to find the cumulative probability for various values of Z. Suppose you want to find the cumulative probability for Z = 0.68 using the z-table.



Number in the table represents P(Z≤z)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767

The intersection of row ‘0.6’ and column ‘0.08’, which is 0.7517, is the answer.

Summary

Central Limit Theorem and Confidence Interval

In this session, you learnt about samples and sampling distributions. A sampling distribution, which is the distribution of the sample means of a population, has some interesting properties; these are collectively called the central limit theorem (CLT). You also learnt how to estimate the population mean using the sample mean and the properties of the central limit theorem.

Samples and Sampling

Instead of finding the mean and standard deviation for the entire population, it is often beneficial to find the mean and standard deviation for only a small representative sample. You may have to do this because of time and/or money constraints.

For example, we wanted to find the average commute time for 30,000 employees working in an office. So, instead of asking all the employees for their commute time, we asked only 100 of them and found out that the mean was equal to 36.6 minutes and the standard deviation was equal to 10 minutes.

However, we believed that it would not be fair to infer that the population mean is exactly equal to the sample mean. This is because the flaws of the sampling process must have led to some error. Hence, the value of the sample mean has to be reported with some error margin.

For example, the mean commute time for those 30,000 employees will be equal to 36.6 + 3 minutes, 36.6 + 1 minute or 36.6 + 10 minutes or, for that matter, 36.6 minutes + some error margin.

However, in order to find this margin, it is important to understand what sampling distributions are, as their properties help in finding this margin.

The notation and formulae related to samples and populations are summarised in the table given below.

Population/Sample	Term	Notation	Formula
Population $(X_1, X_2, X_3, \ldots\ldots, X_N)$	Population Size	N	Number of items/elements in the population
	Population Mean	μ	$\frac{\sum_{i=1}^{i=N} X_i}{N}$
	Population Variance	σ^2	$\frac{\sum_{i=1}^{i=N}(X_i - \mu)^2}{N}$
Sample $(X_1, X_2, X_3, \ldots\ldots, X_n)$ (Sample of Population)	Sample Size	n	Number of items/elements in the sample
	Sample Mean	$\bar{X}$	$\frac{\sum_{i=1}^{i=n} X_i}{n}$
	Sample Variance	S^2	$\frac{\sum_{i=1}^{i=n}(X_i - \bar{X})^2}{n-1}$

Sampling Distributions and Central Limit Theorem

A sampling distribution, specifically the sampling distribution of the sample means, is a probability distribution function for the sample means of a population.

A sampling distribution, which is the distribution of the sample means of a population, has some interesting properties, which are collectively called the central limit theorem. This theorem states that irrespective of the distribution of the original population, the sampling distribution will follow the following three properties:

1. Sampling Distribution's Mean ($\mu_{\bar{x}}$) = Population Mean (μ).
2. Sampling Distribution's Standard Deviation (Standard Error) = $\sigma / \sqrt{n}$, where σ is the standard deviation and n is the sample size of the population.
3. When n > 30, the sampling distribution becomes a normal distribution.

Interval Estimation

Using the CLT, you can estimate the population mean from the sample mean and the standard deviation.

For example, to estimate the mean commute time of 30,000 employees working in an office, you took a sample of 100 employees and found their mean commute time. For this sample, the sample mean $\bar{X}$ = 36.6 minutes and the sample standard deviation S = 10 minutes.

Using the CLT, you can infer that the sampling distribution for the mean commute time will have:

1. Mean = µ {unknown},
2. Standard error = $\sigma \sqrt{n}$ = $S \sqrt{n}$ = 10/ $\sqrt{100}$ = 1, and
3. Since n(100) > 30, the sampling distribution is a normal distribution.

Using these properties, you can claim that the probability that the population mean µ lies between 34.6 (36.6 - 2) minutes and 38.6 (36.6 + 2) minutes is 95.4%.

Also, the terminology related to this claim is as follows:

1. The probability associated with the claim is called confidence level (Here, it is 95.4%).
2. The maximum error made in a sample mean is called margin of error (Here, it is two minutes).
3. The final interval of values is called confidence interval {Here, it is the range – (34.6, 38.6)}.

In fact, you can generalise the entire process. Suppose you have a sample with sample size n, mean $\bar{X}$ and standard deviation S. Now, the y% confidence interval (i.e., confidence interval corresponding to y% confidence level) for µ will be given by the following range:

Confidence interval = $(\bar{X} - Z*S/\sqrt{n}, \bar{X} + Z*S/\sqrt{n})$

where Z* is the z-score associated with y% confidence level.

For example, the 90% confidence interval for the mean commute time will be as follows:

μ = $(\bar{X} - Z*S\sqrt{n}, \bar{X} + Z*S\sqrt{n})$

Here,

$\bar{X}$ = 36.6 minutes

S = 10 minutes

n = 100

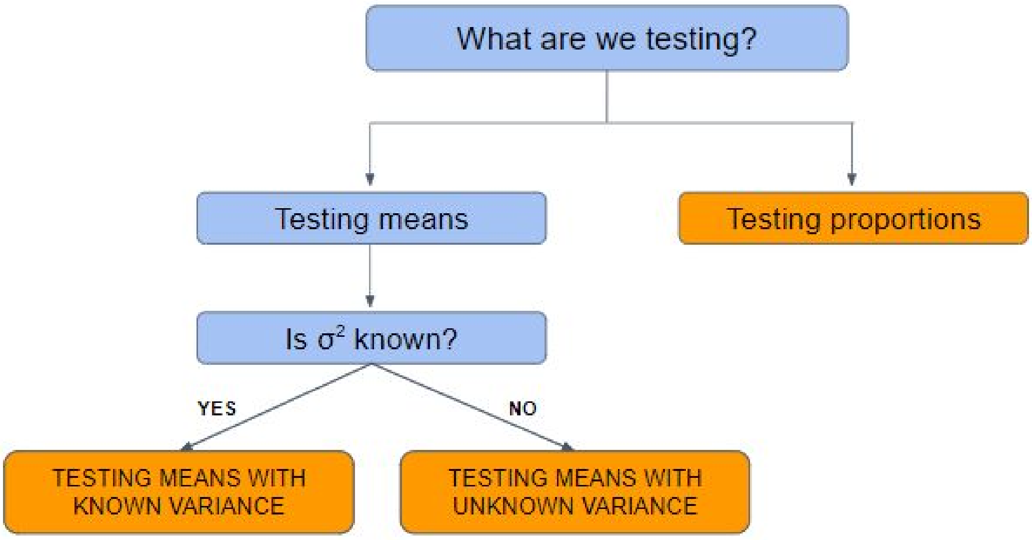
Z* = 1.65 (Z* corresponding to a 90% confidence level)

So, the confidence interval is μ = (34.95 minutes, 38.25 minutes).

Summary

Hypothesis Testing With a Single Population

The statistical analyses that you learnt in the module on 'Inferential Statistics' enabled you to make inferences about population mean and other population data from the sample data. However, you could not confirm the conclusions that you made about the population. This is where hypothesis testing comes into the picture. In this session, you learnt about the one sample test and the test statistic for each type of test. The tests are shown in the flow chart given below.



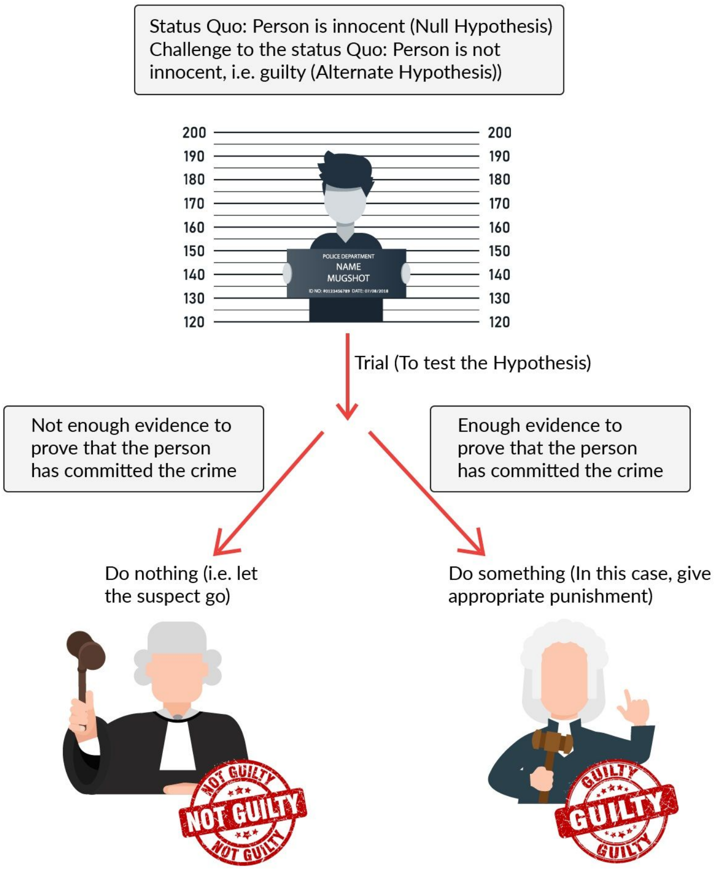
Framing a Hypothesis

The statistical analysis that you learnt in the module on ‘Inferential Statistics’ enabled you to make inferences about the population mean from the sample data when you did not know anything about the population mean. However, sometimes, you may have some starting assumptions about the population mean, and you may want to confirm those assumptions using the sample data. This is where hypothesis testing comes into the picture.

A hypothesis is only a starting point that is open to a test, and it can be accepted or rejected in light of strong evidence. The two terms that you came across are as follows:

1. Null hypothesis (H0): This refers to the status quo.
2. Alternative hypothesis (HA): This refers to the challenge to the status quo.

You also looked at the example of the justice system to understand the concepts of null and alternate hypotheses better.


Status Quo: Person is innocent (Null Hypothesis)
Challenge to the status Quo: Person is not innocent, i.e. guilty (Alternate Hypothesis))
200
190
180
170
160
150
140
130
120
POLICE DEPARTMENT
NAME
MUGSHOT
Trial (To test the Hypothesis)
Not enough evidence to prove that the person has committed the crime
Enough evidence to prove that the person has committed the crime
Do nothing (i.e. let the suspect go)
Do something (In this case, give appropriate punishment)
NOT GUILTY
GUILTY

Apart from this, you also learnt that the general convention for framing the hypothesis is as follows:

1. Null hypothesis (H0): It always contains an equality (=, ≤, ≥).

2. Alternate hypothesis (HA): It always contains an inequality (≠, <, >).

Hypothesis Testing Process and P-Value

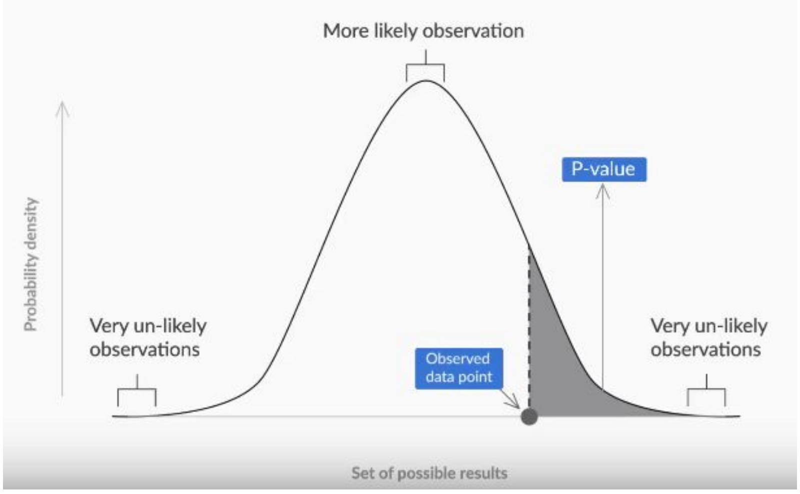
Next, you learnt about the five steps involved in any hypothesis test. These steps are as follows:

1. Begin by assuming that H0, i.e., the status quo, is true. In our case, the status quo is that the average spending is less than ₹120, as this does not require any action.
2. Put the onus of contradicting H0 beyond a **reasonable doubt** on the data. In our case, the sample mean that we got was ₹130. So, we need to prove that this mean is significantly different from ₹120. This means that using this sample mean of ₹130, we need to prove that the chances of the population mean being less than ₹120 are **very low**.
3. Then, quantify the 'reasonable doubt' or 'very low' highlighted above. For that, we use the p-value.
4. Calculate the actual probability of observing the sample, i.e., the p-value.
5. Conclude and take appropriate action: either reject or fail to reject the null hypothesis.

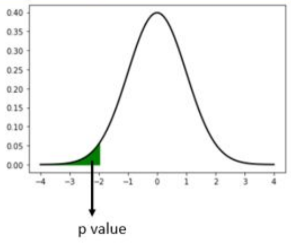
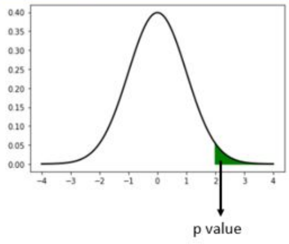
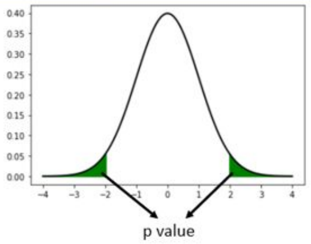
To find the correct p-value from the z-score, you first need to find the cumulative probability by considering the z-table, which gives you the area under the curve until that point.

P-Value Approach

P-value is the area in the tail of a distribution based on the test statistic.



Reject H0 if p-value $\leq \alpha$.

	Lower Tail Test	Upper Tail Test	Two-Tailed Test
Hypothesis	H0 : $\mu \geq \mu0$ Ha : $\mu < \mu0$	H0 : $\mu \leq \mu0$ Ha : $\mu > \mu0$	H0 : $\mu = \mu0$ Ha : $\mu \neq \mu0$
p value rule	p value $\leq$ α	p value $\leq$ α	p value $\leq$ α
z score is determined by the test statistic			

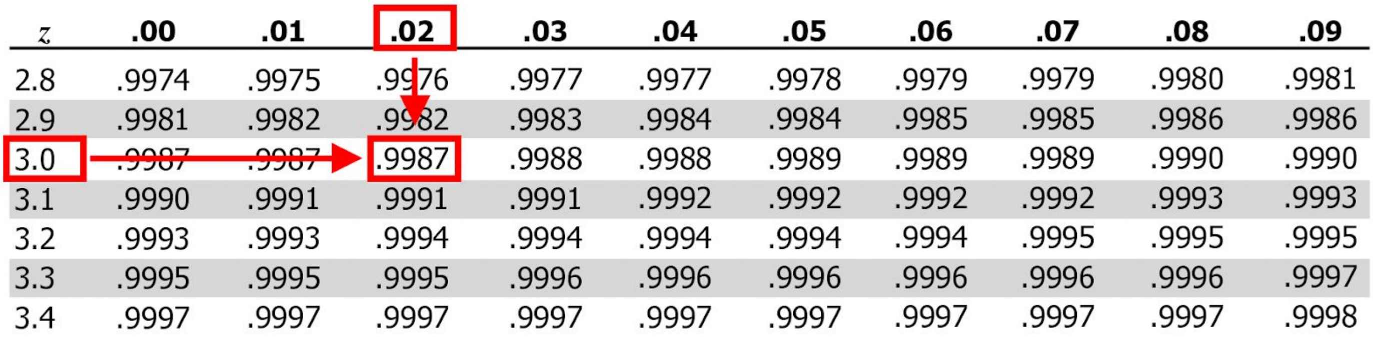
Note: For a two-tailed test, we can write either:

- P-value < α , or
- P-value at any tail < α /2.

Note: If α is not given, then you can assume it to be 0.05.

Situation 1: The sample mean is on the right side of the distribution mean, i.e., the z-score is positive (right-tailed test).

Example: z-score for the sample point = +3.02



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

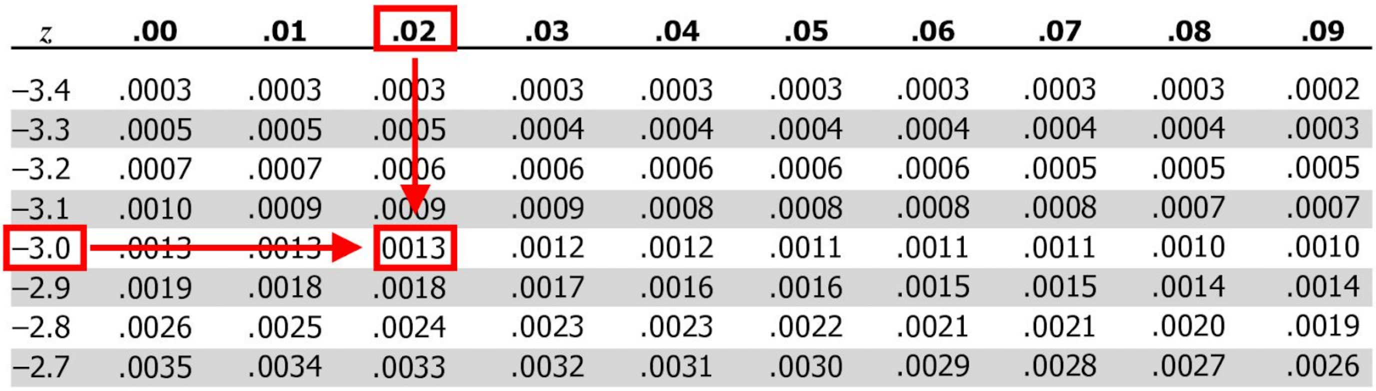
Hence, the cumulative probability of the sample point is 0.9987.

For one-tailed test → p-value = 1 - 0.9987 = 0.0013

For two-tailed test → p-value = 2 (1 - 0.9987) = 2 * 0.0013 = 0.0026

Situation 2: The sample mean is on the left side of the distribution mean, i.e., the z-score is negative (Left-tailed test).

Example: z-score for the sample point = -3.02



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026

Hence, the cumulative probability of the sample point is 0.0013.

For a one-tailed test → p = 0.0013

For a two-tailed test → p = 2 * 0.0013 = 0.0026

Testing Means With Known Variance

This type of problem involves testing the mean of a population against a certain value based on a sample. Here, since you are testing the population mean, this problem is a 'testing means problem'. Now, suppose you were able to obtain a good estimate of the population variance or the standard deviation prior to sampling. In such a case, the population standard deviation (σ) can be considered known. Hence, the type of problem now becomes 'testing means with known variance'.

The steps for solving this problem are as follows:

Step 1:

Identify the test statistic. The test statistic is the z-score that you learnt about in the previous segment. Calculate the z-score using the following formula:

$$z = \frac{\bar{x} - \mu}{\sigma}$$

where,

$\bar{x}$ is the sample mean,

μ is the population mean,

σ is the population standard deviation, and

n is the sample size.

Step 2:

From the z-score, calculate the z-value using the z-table.

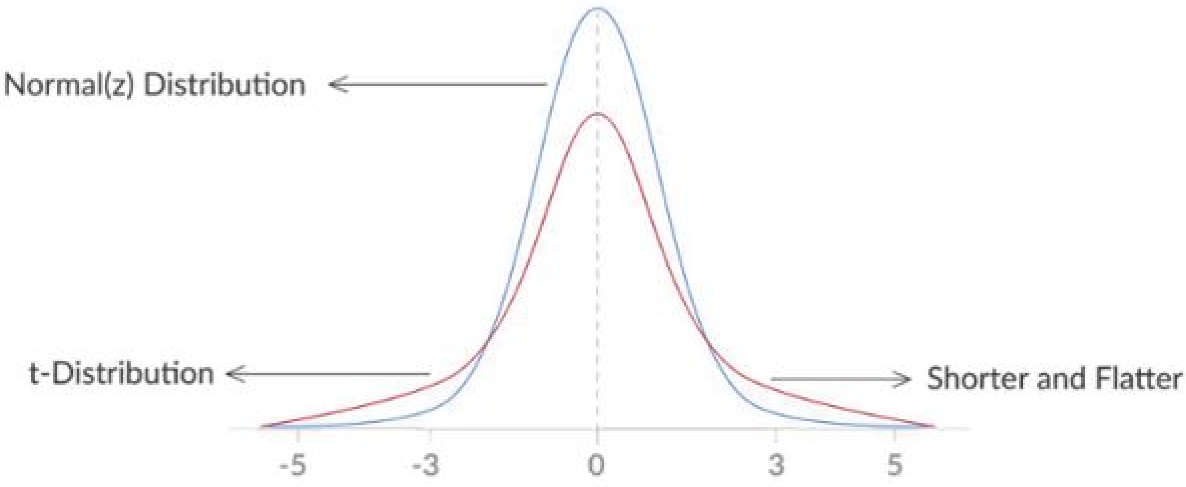
Step 3:

Calculate the p-value from the problem.

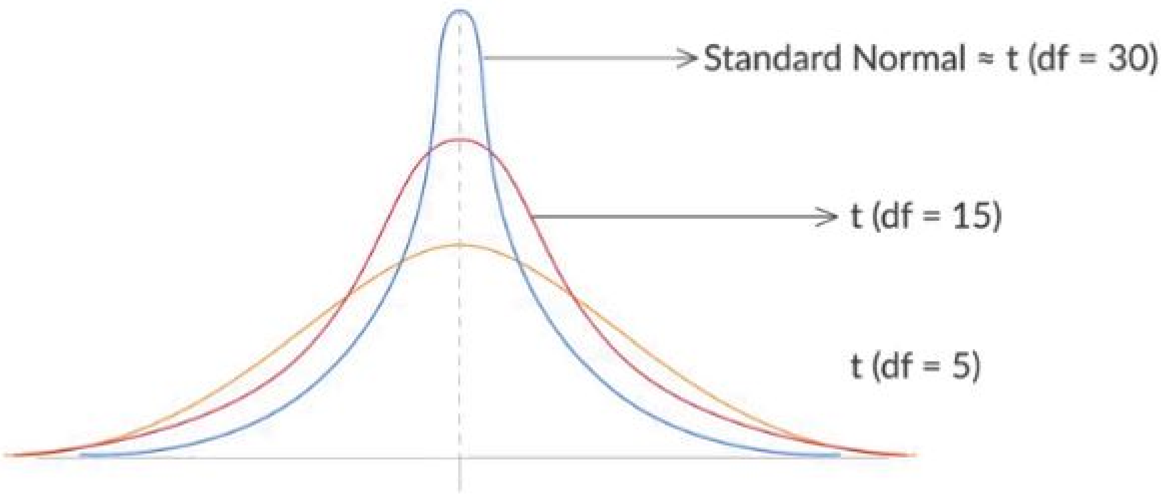
Testing Means With Unknown Variance

So far, you have performed hypothesis testing using a z-test, i.e., the normal distribution. But, many times, either your sample size is small (<30) or the population parameters are not known. In such cases, you need to evaluate a hypothesis test using a t-distribution instead of a z-distribution.

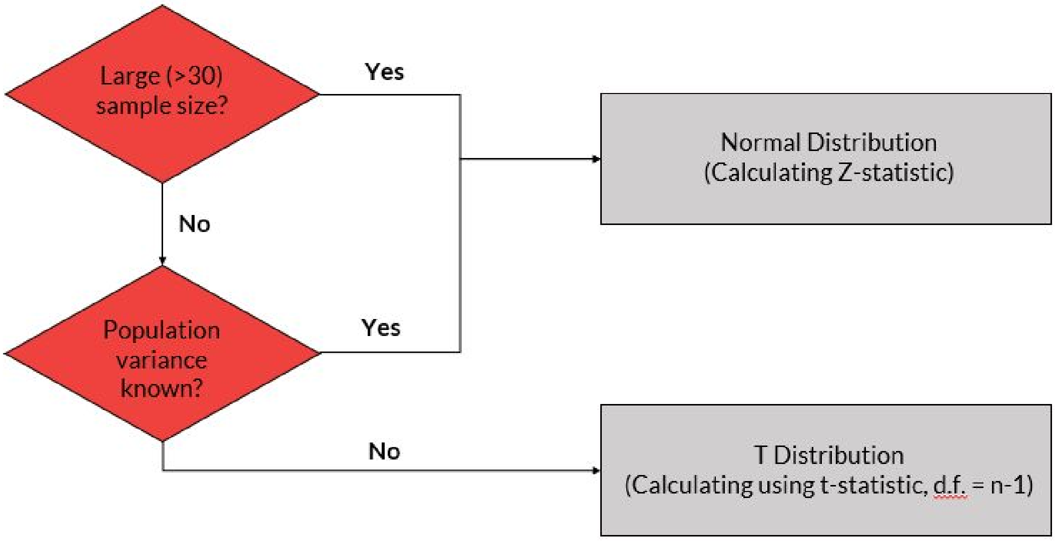
A t-distribution is similar to a z-distribution, except that the t-curve is shorter and flatter than the z-curve.



You also learnt about the degrees of freedom associated with the t-distribution. Degrees of freedom are calculated as the sample size minus 1. For example, if your sample size is n, then the degrees of freedom will be n - 1. Then, you learnt that as you keep increasing the sample size, the t-distribution becomes narrower and steeper. In fact, for a large sample size (> 30), it can be closely approximated to a normal distribution.



Finally, you learnt when you should use a z-test and a t-test.If your sample size is small (< 30) and the population standard deviation is not known, then you can use a t-test. In all other cases, you can use a z-test. You can use the following flowchart to decide whether you should use a t-test or a z-test.


Large (>30) sample size?
Yes
No
Population variance known?
Yes
No
Normal Distribution (Calculating Z-statistic)
T Distribution (Calculating using t-statistic, d.f. = n-1)

Testing Proportions

In this segment, you learnt how to test population proportions against a certain hypothesised proportion. Suppose the hypothesised proportion is p_0 and the population proportion is p. Therefore, you could formulate the following possible hypotheses.

$$\begin{array}{lll} \text{H0}: p \geq p0 & \text{H0}: p \leq p0 & \text{H0}: p = p0 \\ \text{Ha}: p < p0 & \text{Ha}: p > p0 & \text{Ha}: p \neq p0 \end{array}$$.

Now, in order to test these hypotheses, you can collect a sample and determine the sample proportion, $\overline{p}$. Using this, you can calculate the test statistic using the z-distribution and compare it with the critical values.

1. The standard error or the standard deviation for the sample is given by the following formula.

$$\sigma = \sqrt{\frac{p0\ (1 - p0)}{n}}$$.

2. The test statistic is defined as follows:

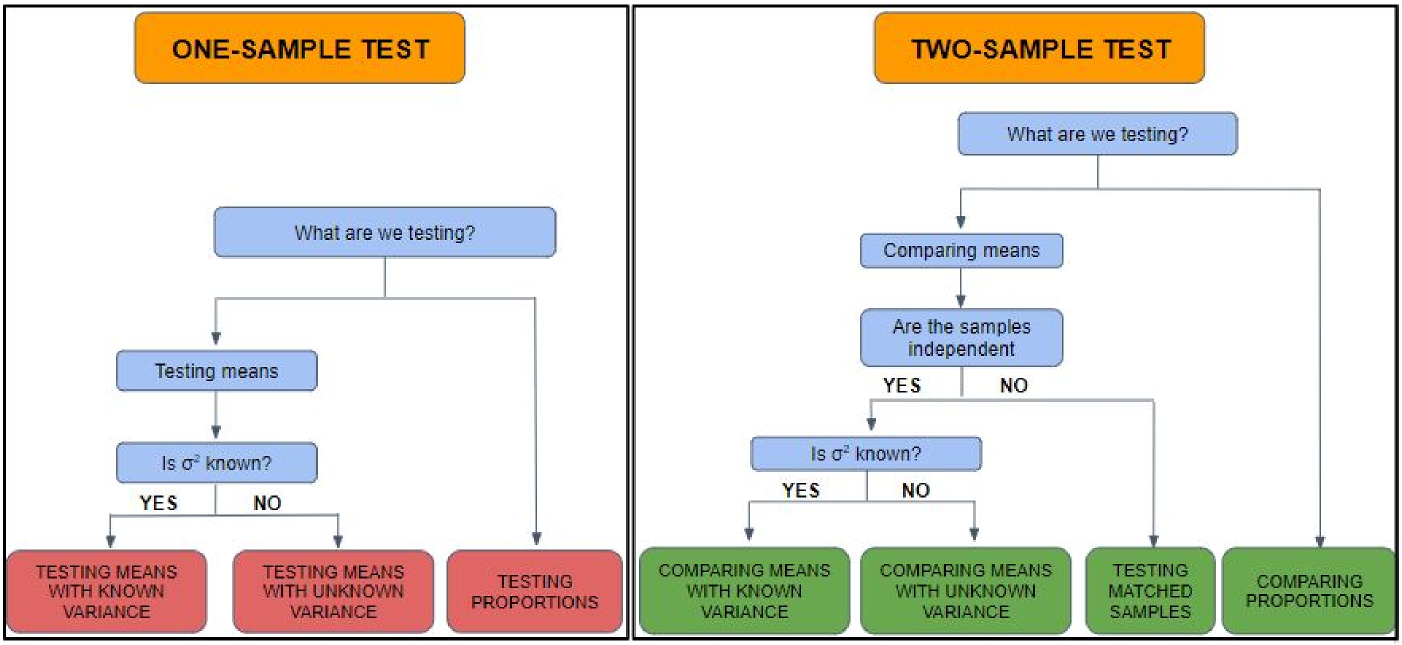
$$z = \frac{\bar{p} - p0}{\sigma_p} = \frac{\bar{p} - p0}{\sqrt{\frac{p0\ (1 - p0)}{n}}}$$.

Using the test statistic, calculate the p-value and compare it with the level of significance.

Summary

Hypothesis Testing With Two Populations

In this session, you learnt how to solve problems on hypothesis testing related to two samples. The tests covered in this session have been listed in the flow chart given below under ‘two-sample test’.



Comparing Means With Known Variance

In this segment, you learnt how to conduct hypothesis tests on the difference between the means of two independent populations, where standard deviations are assumed to be known.

Now, let’s select sample 1 from population 1 and sample 2 from population 2 separately and independently. The notations for the same are given below.

μ_1 = Mean of population 1

σ_1 = Standard deviation of population 1

μ_2 = Mean of population 2

σ_2 = Standard deviation of population 2

Let d_o denote the hypothesised difference between the two populations. Now, you can formulate the following possible hypotheses.

$$\begin{array}{lll} \text{H0}: \mu_1 - \mu_2 \geq d_0 & \text{H0}: \mu_1 - \mu_2 \leq d_0 & \text{H0}: \mu_1 - \mu_2 = d_0 \\ \text{Ha}: \mu_1 - \mu_2 < d_0 & \text{Ha}: \mu_1 - \mu_2 > d_0 & \text{Ha}: \mu_1 - \mu_2 \neq d_0 \end{array}$$

The test statistic is as follows:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

Once you have calculated the value of the test statistic, the rest of the process remains the same. Based on the level of significance, you can check whether or not the calculated p-value is less than α. Based on this information, you can decide whether or not to reject the null hypothesis.

Comparing Means With Unknown Variance

This test is also known as an independent sample test. Since the population variances are unknown, you will use the sample variances as point estimators for the population variances. Now, suppose you have selected sample 1 from population 1 and sample 2 from population 2 separately and independently.

The standard deviations for both populations are estimated by the sample standard deviations for the two populations. Let’s consider the standard deviations as s1 and s2. The test statistic for this test is given by the following formula:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2) - \mathrm{d0}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The degrees of freedom for this test are given by the following formula.
Note: Degrees of freedom are computed with the help of statistical softwares; hence, do not try to remember the following formula.

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

Comparing Matched Samples

A matched-samples test is usually conducted in the following scenarios:

1. For a single sample of players, each player is measured twice, i.e., before and after a certain intervention.
2. Given a single sample of players, each player is measured twice under two different experimental conditions.
3. Remember that the sample size for both the samples will always be equal.

So, once you have identified the test as a matched-samples test, you must compute the difference between the matched samples. Then, you can consider the 'Difference' column as one sample and test it against the hypothesised value, d_0. Assuming the mean of the difference between the population is given by d, you can formulate the following hypotheses.

H0 : $d \geq d_0$	H0 : $d \leq d_0$	H0 : $d = d_0$
Ha : $d < d_0$	Ha : $d > d_0$	Ha : $d \neq d_0$.

Since you are considering the 'Difference' column as your new sample for this test, you need to recalculate the sample mean and sample standard deviation for this test.

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$$

$$s_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n - 1}}$$

The test statistic for this test is given by the following formula:

$$t = \frac{\bar{d} - d_0}{s_d / \sqrt{n}}$$

.

where degrees of freedom (df) = n - 1.

Comparing a Single Proportion

Suppose you have selected sample 1 from population 1 and sample 2 from population 2 separately and independently. Assuming that each event in the population is classified as a success or a failure, your aim is to compare the proportion of successes in these two samples.

Now, let:

p_1 = Proportion of successes in population 1;

p_2 = Proportion of successes in population 2;

$\overline{p_1}$ = Proportion of successes in sample 1 (selected from population 1); and

$\overline{p_2}$ = Proportion of successes in sample 2 (selected from population 2).

The test statistic is calculated as follows:

$$z = \frac{\bar{p_1} - \bar{p_2}}{s_{\bar{p_1} - \bar{p_2}}}$$

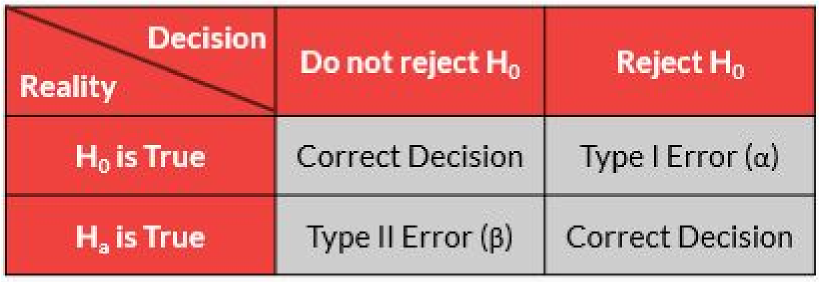
where,

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\bar{p}(1 - \bar{p})\left(\frac{1}{n}, + \frac{1}{n_2}\right)} \quad \text{and} \quad \bar{p} = \frac{n_1\bar{p_1} + n_2\bar{p_2}}{n_1 + n_2}$$

.

Types of Errors in Hypothesis Testing

You can commit two types of errors while performing a hypothesis test. They are mentioned in the table given below.



Decision Reality	Do not reject H_0	Reject H_0
H_0 is True	Correct Decision	Type I Error (α)
H_a is True	Type II Error (β)	Correct Decision

These two errors are explained below.

1. Type-I error: α is the acceptable probability of making a type-I error (also called the significance level). Alternatively, $(1 - \alpha)$ is called the confidence level (Recall that you learnt about confidence level in the module on 'Inferential Statistics'.). This occurs when your null hypothesis is actually true but you reject it.

2. Type-II error: β is the probability of making a type-II error. Alternatively, $(1 - \beta)$ is called the power of the test. This occurs when your alternate hypothesis is true but you fail to reject your null hypothesis.

Now, let's take a look at the business implications of a type-I error. In this case, you are rejecting your null hypothesis even if it is true, i.e., you are accepting the alternative hypothesis and taking an action. However, since this action requires time, money and other resources, you cannot afford to make a mistake often, and this is why managers are more concerned about a type-I error.

Summary

A/B Testing Case Study

In this session, you learnt about hypothesis testing on Python and A/B testing. You also went through a case study on an e-commerce company, which involved performing exploratory data analysis, hypothesis testing and A/B testing.

Hypothesis Testing in Python

ttest_1samp(a, popmean[, axis, nan_policy])	Calculate the t-statistic for the mean of one group of scores.
ttest_ind(a, b[, axis, equal_var, nan_policy])	Conduct a t-test on the means of two independent samples of scores.
ttest_rel(a, b[, axis, nan_policy])	Conduct a t-test on two related samples of scores a and b.
statsmodels.stats.proportion.proportions_ztest	Test proportions based on a z-test (one- or two-sample).

A/B Testing

While developing an e-commerce website, there could be different opinions on the choices of various elements, such as the shape of the buttons, the text on the call-to-action buttons, the colour of various UI elements and the copy on the website.

The choice of these elements is often subjective, and it is difficult to predict which option will perform better. To resolve such conflicts, you can use A/B testing. A/B testing provides you a way to test two different versions of the same element and determine which version performs better. An example of a situation where you can conduct an A/B test is as follows:

Can a change in the design of a car lead to better mileage?

When you run a multivariate test, you change multiple factors in order to test the change in a metric. Depending on the variation of individual variables, there will be more than two versions to compare. This is called multivariate testing. Suppose, in the car design example, you also change the fuel type as well as the fuel technology.

The steps to conduct an A/B test are as follows:

1. Defining the objective
2. Defining the KPIs

   Note: KPI stands for 'Key Performance Indicator'. KPIs are performance indicators that are used to evaluate how effectively an organisation is achieving its key business objectives.

   To know more about KPIs, you can refer to the link given below.
   Additional Link on KPIs

3. Identifying the drivers for the KPIs
4. Prioritising the driver that you want to test
5. Segmenting the population
6. Interpreting the results

We have considered the following industries and shown the steps for conducting an A/B test on each of them.

1. E-commerce
2. Telecom

E-commerce Company

The methodology for conducting an A/B test for an e-commerce website is summarised in the table given below.

Defining the objective	To increase the number of daily orders
Defining the KPIs	Session conversion

	- Within a session, customers visit a web page, search for a product and decide whether or not to buy it. - Session conversion reflects the fraction of customers who purchased a product in a particular session.
Identifying the drivers for the KPIs	Some drivers for session conversion are as follows: - Discounts - Search relevancy - Product reviews
Prioritising the driver that you want to test	Search relevancy - Solution: Develop a new algorithm
Segmenting the population	Select 5%–10% of the total population visiting the website
Interpreting the results	Conduct an A/B test to determine whether or not the new algorithm improves session conversion

Telecom Company

The methodology for conducting an A/B test for a telecom company is summarised in the table given below.

Defining the objective	To increase the number of recharges done per month
Defining the KPIs	Reduce customer attrition or onboard new customers Note: Customer attrition refers to the loss of customers in a business.
Identifying the drivers for the KPIs	Some drivers for preventing customer attrition are as follows:

	● Competitor price ● Service quality
Prioritising the driver that you want to test	Service quality is the driver that has been prioritised
Segmenting the population	Select 5%–10% of the total population (for the given circle) and put them in the treatment group
Interpreting the results	Conduct an A/B test to determine whether or not the new algorithm works

Case Study

Problem Statement:

1. This e-commerce company is evaluating its customer experience. It conducts a survey among its customers to get their feedback.

2. The feedback shows that late delivery is one of the most frequent problems faced by the customers.

3. However, the consumer experience manager says that the late delivery complaint percentage is similar to the industry average.

4. In order to solve this issue, the company has come up with a solution. It decides to build an intelligent system that predicts the delivery date by taking into account several factors.

5. Once the algorithm is built, it is rolled out for 10% of the customers to check whether or not it is successful.

Approach

1. Establish the fact that the organisation has a poor delivery record.

2. Conduct a hypothesis test to compare the organisation’s late delivery complaint percentage with the industry average of 6%.

3. Compare the late delivery complaint percentage before and after the algorithm was rolled out.

Solution

1. A matched-samples hypothesis test was performed to compare the actual and estimated delivery times, and it proved that a significant difference exists between them.

2. Using some exploratory data analysis, it was established that most of the reviews were inclined towards a late delivery experience.

3. As you saw in the problem statement, the consumer experience manager said that the late delivery percentage is similar to the industry average. Therefore, a one-sample proportion test was conducted to check whether or not the industry average was within 6%. This test established the fact that the organisation has a poor delivery record.

4. The company had built an intelligent system that predicts the delivery date by taking into account several factors. It intends to test this algorithm on a certain segment of the population. A two-sample proportion test was conducted to prove that a significant difference existed between the treatment and control groups. This test proved that with the help of the new algorithm, a significant decrease was achieved in the proportion of late delivery reviews.

Disclaimer: All content and material on the upGrad website is copyrighted material, either belonging to upGrad or its bonafide contributors, and is purely for the dissemination of education. You are permitted to access, print and download extracts from this site purely for your own education only and on the following basis:

- You can download this document from the website for self-use only.
- Any copy of this document, in part or full, saved to disc or to any other storage medium, may only be used for subsequent, self-viewing purposes or to print an individual extract or copy for non-commercial personal use only.
- Any further dissemination, distribution, reproduction, copying of the content of the document herein or uploading thereof on other websites or use of content for any other commercial/unauthorised purposes in any way which could infringe the intellectual property rights of upGrad or its contributors, is strictly prohibited.
- No graphics, images or photographs from any accompanying text in this document will be used separately for unauthorised purposes.
- No material in this document may be modified, adapted or altered in any way.
- No part of this document or upGrad content may be reproduced or stored in any other website or included in any public or private electronic retrieval system or service without upGrad's prior written permission.
- Any right not expressly granted in these terms are reserved.