**Introduction: Inferential Statistics**

Welcome to the module on ‘**Inferential Statistics**’. In the last module on **EDA**, you learnt how to explore data and derive insights from its exploration.

**In this module**

Exploratory data analysis helped you understand how to **discover patterns in data**using various techniques and approaches. As you've learnt, EDA is one of the most important parts of the data analysis process. It is also the part on which data analysts spend most of their time.

However, sometimes, you may require a very large amount of data for your analysis which may need too much time and resources to acquire. In such situations, you are forced to work with a **smaller sample of the data**, instead of having the entire data to work with.

Situations like these arise all the time at big companies like Amazon. For example, say the Amazon QC department wants to know what proportion of the products in its warehouses are defective. Instead of going through all of its products (which would be a lot!), the Amazon QC team can just check a small sample of 1,000 products and then find, for this sample, the defect rate (i.e. the proportion of defective products). Then, based on this sample's defect rate, the team can **"infer"**what the defect rate is for all the products in the warehouses.

**This process of “inferring” insights from sample data is called “Inferential Statistics”.**

Note that even after using inferential statistics, you would only be able to estimate the population data from the sample data, but not find the exact values. This is because when you don't have the exact data, you can only make reasonable estimates about it with a limited level of certainty. Therefore, when certainty is limited, we talk in terms of probability.

## Prerequisites

You’ll need to brush up on your concepts of **probability** before you begin this module, specifically the following concepts:

* **Basic definition** of probability
* **Multiplication rule** of probability
* **Addition rule**of probability
* ^{n}C^{\; }_{r} (**Combinatorics**)

You can brush up on these topics using the links given below -

# Introduction: Basics of Probability

Welcome to the session on ‘**Basics of Probability**’.

## In this session

In this session, you will learn a few basic concepts of **probability** through a specific example. The broad agenda for this session is as follows:

* Random variables
* Probability distributions
* Expected value

# Random Variables

Welcome to the first session on inferential statistics! This will be a very interactive session, with a lot of questions that will compel you to think about a concept, helping you explore it more actively.

So, let’s get started.

Recall the original question: In the long run (i.e. if it is played a lot of times), is this game profitable for the players or for the house? Or will everybody break even in the long run?

Recall that we established a three-step process for answering this question:

1. Find all the possible combinations
2. Find the probability of each combination
3. Use the probabilities to estimate the profit/loss per player

We have completed step 1, i.e. finding all the possible combinations. Now, let’s proceed to step 2, i.e. finding the probability of each combination. What are the steps involved in finding the probability? Let’s hear more from Professor Tricha on this:

So, the **random variable X** basically converts outcomes of experiments to something measurable.

For example, let’s say as a Data analyst at a bank, you are trying to find out which of the customers will default on their loan, i.e. stop paying their loans. Based on some data, you have been able to make the following predictions:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Customer No. | Yearly Income (in rupees) | Amount of Loan Due (in rupees) | Number of Dependents | Default Prediction (Yes/No) |
| 1 | ₹10 lakh | ₹75 lakh | 3 | Yes |
| 2 | ₹15 lakh | ₹50 lakh | 2 | No |
| 3 | ₹20 lakh | ₹40 lakh | 1 | No |

Now, instead of processing the yes/no response, it will be much easier if you define a random variable X, indicating whether the customer is predicted to default or not. The values will be assigned according to this rule:

X = 1, if the customer defaults

X = 0, if the customer does not default

Now, the data changes to the following:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Customer No. | Yearly Income (in rupees) | Amount of Loan Due (in rupees) | Number of Dependents | X (random variable) |
| 1 | ₹10 lakh | ₹75 lakh | 3 | 1 |
| 2 | ₹15 lakh | ₹50 lakh | 2 | 0 |
| 3 | ₹20 lakh | ₹40 lakh | 1 | 0 |

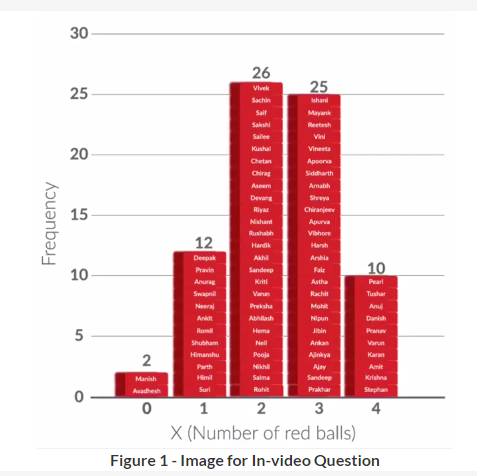
Now, in this form, the table is entirely **quantified**, i.e. converted to numbers. Now that the data is entirely in quantitative terms, it becomes possible to perform a number of different kinds of statistical analyses on it.

# Probability Distributions - I

Recall that, in our UpGrad game example, we need to find out if the game would be profitable for the players or for us (i.e. the house) in the long run. The three-step process for this is:

1. Find all the possible combinations
2. Find the probability of each combination
3. Use the probabilities to estimate the profit/loss per player

So far, we have completed step 1, and are on step 2, i.e. finding the probability of each combination. For this purpose, we defined a random variable X which helped us convert the outcomes of our experiment to something measurable. Now, let’s actually find the probability of each of these combinations.



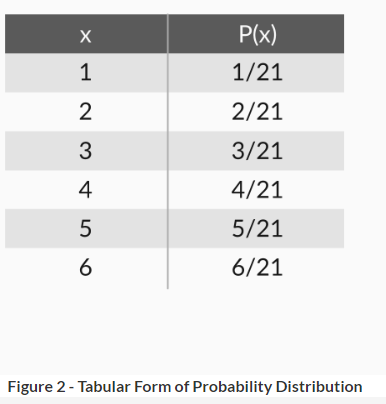
So, we performed the experiment (i.e. played the game) 75 times, and then made the **frequency distribution (histogram)**. Now, you may be thinking, “Well, I want to try that out too.” Unfortunately, even if you do have a bag with 3 red balls and 2 blue balls, playing the game 300 times would be tedious and difficult. Well, that’s not a problem. You can simulate the whole experiment in R. Let’s hear from Kshitij on how that can be done.

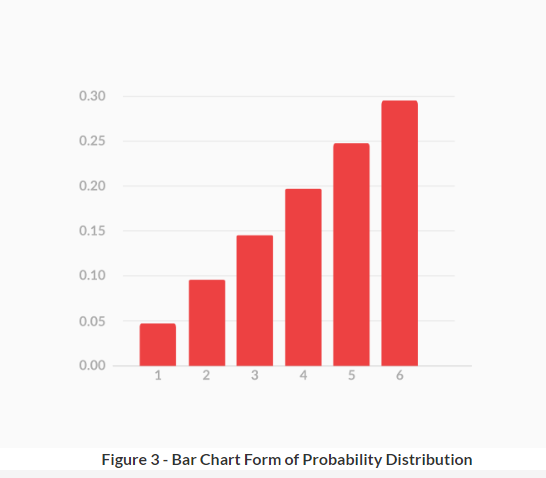
*(This whole process is just being shown for your understanding, you will not be tested in this module on how to simulate the experiment and make histograms in R.)*

**robability Distributions - II**

So basically, a **probability distribution** is ANY form of representation that tells us the probability for all possible values of X. It could be any of the following:

* A table





Now, let’s say that a company’s management is pondering over investing in a certain project. Before doing this, it wants to use probability to find whether it can safely expect to make a profit. Whether the company makes a profit or not will actually depend on which economic cycle is going on, i.e. recession, boom, and so on.

Based on the opinions of some experts, the following table is created:

|  |  |
| --- | --- |
| Economic Cycle | Probability |
| Recession | 0.1 |
| Normal | 0.7 |
| Boom | 0.2 |

Suppose, as an analyst in the investment division, you have been asked to find the answer to the question: “Can the company expect to make a profit or not? Should it invest in this project?”

However, in this form, the table is of no help at all. Hence, let’s quantify it using a random variable. Since you are interested in whether the company will profit or not, let’s define X as X = Net revenue of the project.

Now, through some calculations, a fellow analyst of the company has found what the net revenue would be for each of these scenarios. She creates a probability distribution with this data:

|  |  |
| --- | --- |
| X (Net Revenue of Project, in ₹ crores) | P(x) |
| -305 | 0.1 |
| +15 | 0.7 |
| +95 | 0.2 |

Now, you finally have a probability distribution for X, the net revenue of the project. Using this probability distribution, you can find the answer to our original question - “Can the company expect a profit from this project? Or should it expect a loss?” However, to answer it, you will have to learn the concept of expected value, which is what we will cover next.

# Expected Value - I

Again, let’s go back to the three-step process we followed to find whether the UpGrad red ball game was profitable for the players or for the house:

1. Find all the possible combinations
2. Find the probability of each combination
3. Use the probabilities to estimate the profit/loss per player

Now that we have completed steps 1 and 2, let’s move on to step 3 where we will use the probabilities we calculated to estimate the profit/loss per player.

So, the **expected value** for a variable X is the value of X we would “expect” to get after performing the experiment once. It is also called the **expectation**, **average**, and **mean value**. Mathematically speaking, for a random variable X that can take values x1,x2,x3,...........,xnx1,x2,x3,...........,xn, the expected value (EV) is given by:

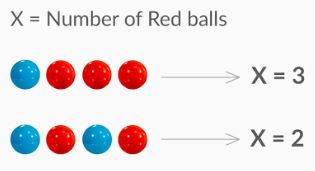
EV(X)=x1∗P(X=x1)+x2∗P(X=x2)+x3∗P(X=x3)+...........+xn∗P(X=xn)EV(X)=x1∗P(X=x1)+x2∗P(X=x2)+x3∗P(X=x3)+...........+xn∗P(X=xn)

As you may recall, for our red ball game, the expected value came out to be **2.385**. What does that mean? How does that even help us with our original question, which was how much money, on average, are the players expected to make?

# Summary: Basics of Probability

In the first section, you learnt how to **quantify the outcomes** of events by using **random variables**.

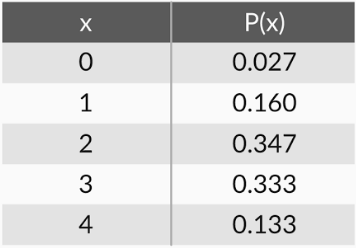
For example, recall that we quantified the colours of the balls we would get after playing our game by assigning a value of X to each outcome. We did so by defining **X as the number of red balls** we would get after playing the game once.



**Figure 4 - Quantifying Using Random Variables**

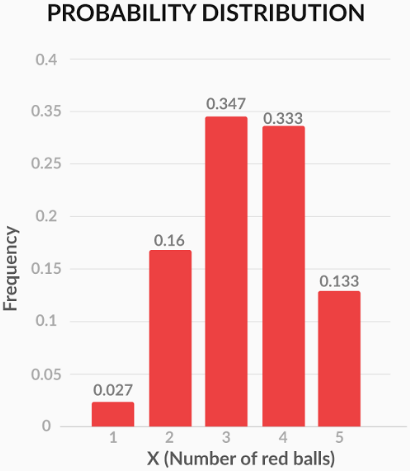
Next, we found the **probability distribution**, which was a **distribution giving us the probability for all possible values of X**.

We created this distribution in a **tabular form:**



**Figure 5 - Tabular Form of Probability Distribution**

We also created it in a **bar chart form:**



**Figure 6 - Bar Chart Form of Probability Distribution**

You saw that, in the bar chart form, we were able to visualise the probability in a much better way., Thus, this form is used more widely as it helps you see trends easily.

Then, we went on to find the **expected value** for X, the money won by a player after playing the game once. The expected value (EV) for X was calculated using the formula:

EV (X) = x_{1}*P(X=x_{1})+x_{2}*P(X=x_{2})+....................+x_{n}*P(X=x_{n})

Another way of writing this is

EV(X) = \sum_{i=1}^{i=n}x_{i}*P(X=x_{i})

Calculating the answer this way, we found the expected value to be +11.28.

In other words, if we conduct the experiment (play the game) **infinite times**, the **average money**won by a player would be ₹11.28. Hence, we decided that we should either decrease the prize money or increase the penalty to make the expected value of X negative. A negative expected value would imply that, on average, a player would be expected to lose money and the house would profit.

# Introduction: Discrete Probability Distributions

Welcome to the session on ‘**Discrete Probability Distributions**’. In the last session, you learnt some basic concepts of probability, such as **random variables**, **probability distributions**, and **expected value**. Let’s now learn some slightly more advanced concepts.

## In this session

In this session, you will learn about some probability distributions that are commonly used for discrete random variables, such as the **binomial probability distribution** and the **uniform probability distribution**. Also, you will learn the concept of **cumulative probability** which will be very useful in our next session on continuous probability distributions.

## Prerequisites

It has been assumed in this session that you have knowledge of some basic rules of probability, i.e. **multiplication rule** and **addition rule**. Also, it is assumed that you have read upon **combinatorics**and know what the expression nCrnCr represents.

# Probability Without Experiment - I

In the last session, we found the probability for certain events by conducting experiments.

Specifically, we asked 75 people to play the UpGrad red ball game. Based on the data of these people, we created a histogram or, in other words, a frequency distribution. Then, using this histogram, we created the probability distribution.

However, this is a lengthy process. Is there a shorter process for finding the probabilities, perhaps one that doesn’t need repeated experiments? Let’s see.

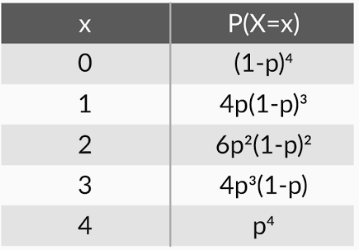
Hence, we have gone through the exercise of finding the **probability without conducting any experiment**. You saw that these theoretical (calculated) values of probability are actually quite close to the experimental values that we got. The small differences that you can notice exist because of the low number of experiments done.

# Binomial Distribution

Previously, we found the theoretical probability for our game and compared it with the experimental one. Finding the probability without conducting an experiment means that we can find the probability using just pen and paper and with minimal effort.

Now, let’s try to generalise it — let’s say that the probability of getting one red ball in one trial is equal to p. In that case, what would be the probability of all 4 balls being red?

So, now we have this probability distribution for X (the number of red balls drawn after 4 trials), if the **probability of getting a red ball in 1 trial is p**:



**Figure 1 - Probability Distribution for General Probability p**

So, the formula for finding **binomial probability** is given by -

P(X=r)=P(X=r)= nCr(p)r(1−p)n−rnCr(p)r(1−p)n−r

Where **n** is **no. of trials**, **p** is **probability of success** and **r** is **no. of successes after n trials**.

However, as Prof. Tricha said, there are some **conditions** that need to be followed in order for us to be able to apply the formula.

1. **Total number** of trials is **fixed** at n
2. Each trial is **binary**, i.e., has **only two possible outcomes** - success or failure
3. **Probability of success** is **same** in all trials, denoted by p

# Binomial Distribution (Examples)

In the previous section, we listed down some conditions that are to be fulfilled for the binomial distribution to be applicable. Let’s take a few examples to understand these conditions in detail.

|  |  |
| --- | --- |
| **Binomial Distribution Applicable** | **Binomial Distribution Not Applicable** |
| Tossing a coin 20 times to see how many tails occur | Tossing a coin until a heads occurs |
| Asking 200 randomly selected people if they are older than 21 or not | Asking 200 randomly selected people how old they are |
| Drawing 4 red balls from a bag, putting each ball back after drawing it | Drawing 4 red balls from a bag, not putting each ball back after drawing it |

If you toss a coin 20 times to see how many tails occur, you are following all the conditions required for a binomial distribution. The total number of trials is fixed (20), and you can only have two outcomes, i.e. a tails or a heads. The probability of getting a tails is equal to 0.5 each time you toss the coin.

In a way, this is similar to drawing 20 balls out of a bag, replacing each ball after drawing it, and seeing how many of the balls are red. Here, the probability of getting a red ball in one trial is 0.5.

When you toss a coin until a heads occurs, the total number of trials is not fixed. This is similar to taking out balls from the bag repeatedly until you draw a red ball. You can still find the probability of getting a heads in 1 trial, in 2 trials, in 3 trials etc. and so on, but you cannot use binomial distribution to find that probability.

In the second example, where binomial distribution is not applicable, the experiment does not have only two outcomes, but many. It is similar to taking out balls from a bag that contains red, blue, black, orange, and other colours of balls. The probability distribution for this experiment cannot be made using binomial distribution.

In the final example, the probability of the trials is not equal to each other. For example, the probability of drawing a red ball in the first trial is 3535. Now, in the second trial, the probability of drawing a red ball would be equal to 2424 not 3535, as the red ball taken out in the first trial was not put back. Hence, the probability of getting the combination Red-Red-Red-Blue, for example, would be 3535\*2424\*1313\*2222, which is not the value we got while deriving binomial distribution (3535\*3535\*3535\*2525). Again, you cannot use binomial distribution to find the probability in this case.

In other words, binomial distribution is applicable in situations where **there are a fixed number of yes or no questions, with the probability of a yes or a no remaining the same in all the questions.**

# Cumulative Probability

In the previous example, we only discussed the probability of getting an exact value. For example, we know the probability of X = 4 (4 red balls). But what if the house wants to know the probability of getting < 3 red balls, as the house knows that for < 3 red balls, the players will lose and they will make money?

Sometimes, talking in **less than** is more useful, for example — how many employees can get to work in less than 40 minutes? Let’s explore how you can find the probability for such cases.

**Clarification -** The cumulative probability distribution table, shown in the above video (02:22 to 03:11) should be -

|  |  |
| --- | --- |
| **x** | **F(x) = P(X<x)** |
| 0 | 0.0256 |
| 1 | 0.1792 |
| 2 | 0.5248 |
| 3 | 0.8704 |
| 4 | 1.0000 |

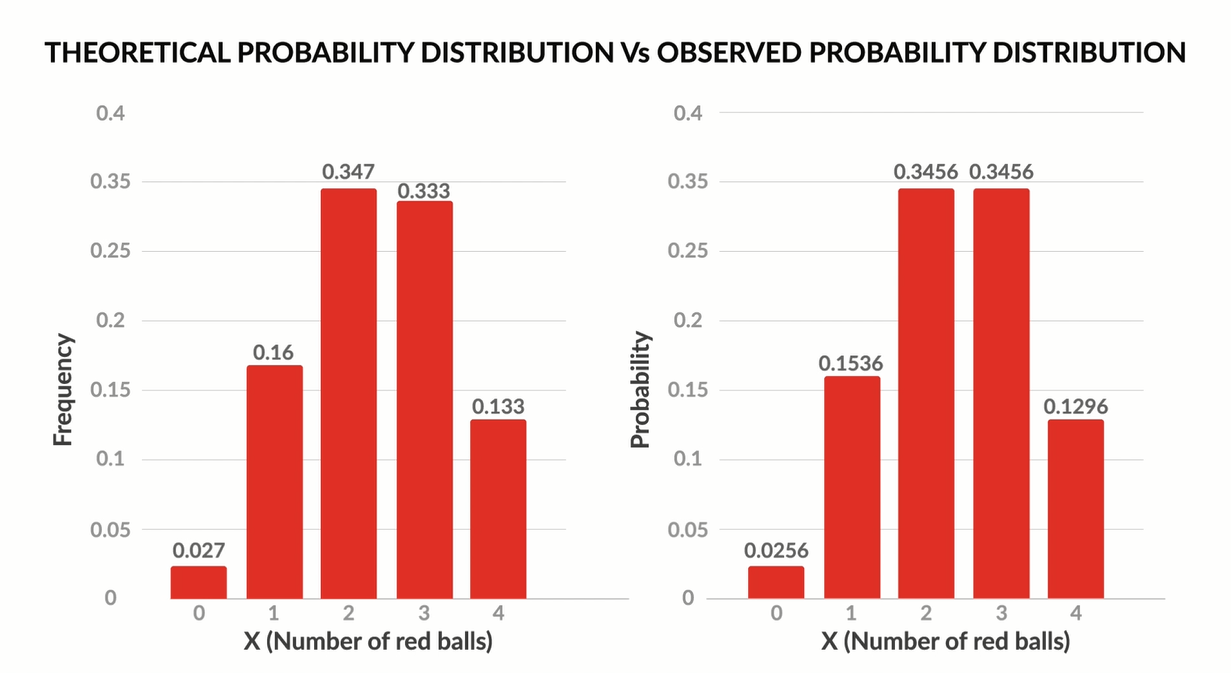
So, the **cumulative probability of X**, denoted by **F(x)**, is defined as **the probability of the variable being less than or equal to x**.

In mathematical terms, you would write cumulative probability **F(x) = P(X<x)**. For example, F(4) = P(X<4), F(3) = P(X<3).

# Summary: Discrete Probability Distributions

We started with learning how to find **probability without experiment**, using basic concepts such as the addition and the multiplication rule of probability.

As a demonstration, we calculated the probability of getting 0, 1, 2, 3 and 4 red balls for our UpGrad red ball game. Then, we compared the values we got by theoretical calculations with the ones we got after the experiments in the first session. Recall that the values were quite similar.

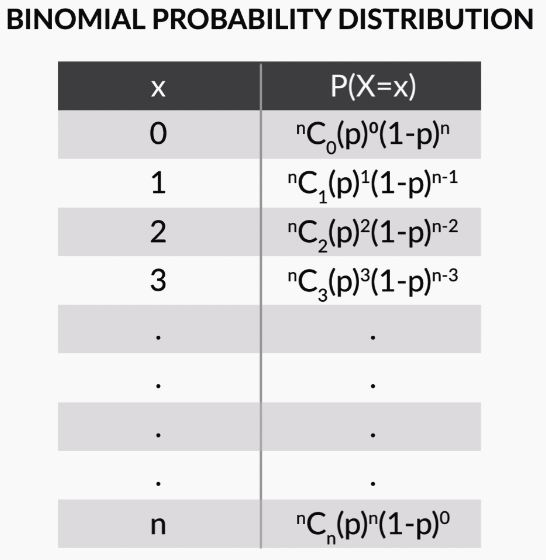


**Figure 2 - Observed vs Theoretical Probability Distribution**

However, they were not exactly the same, due to the low number of experiments we conducted (75). If we had done more experiments, the values would have been exactly the same (the values are exactly the same if the number of experiments approaches infinity).

Next, we **generalised this probability**. Specifically, we talked out the probability of getting **r red balls** after drawing **n balls** from a bag. Here, the probability of drawing **a red ball in 1 trial** was equal to **p**.

The probability distribution for this case is given by the following table (X = number of red balls drawn after playing the game once).



**Figure 3 - Binomial Distribution**

In general,

P(X=r)=P(X=r)=  nCr(p)r(1−p)n−rnCr(p)r(1−p)n−r

This distribution is called the **binomial distribution**. It can be used to find the probability of any kind of event, if that event is **a series of yes or no questions, with the probability of yes being the same for all questions**.

Phrasing the conditions more formally, the binomial distribution can be used if, for an experiment:

* The **total number** of trials is **fixed**
* Each trial is **binary**, i.e. has **only two possible outcomes**, success and failure
* The **probability of success** is the **same** for all the trials

Next, we discussed the **cumulative probability of x**, denoted by **F(x)**, which is the **probability that the random variable X takes a value less than x**.

For example, we found F(2), the probability of getting 2 or fewer red balls in our UpGrad game. It was calculated as:

F(2) = P(X < 2) = P(X = 0) + P(X = 1) + P(X = 2) = 0.0256 + 0.1536 + 0.3456 = 0.5248

Cumulative probability is a concept that we will use extensively in our next session on continuous random variables.

# Introduction: Continuous Probability Distributions

Welcome to the session on ‘**Continuous Probability Distributions**’. In the last session, you learnt about the **binomial distribution** and the **uniform distribution**. Also, you learnt the concept of **cumulative probability**.

## In this session

In this session, you will learn about **cumulative probability** in a little more depth. You will see how the **probability of a continuous variable** is expressed and how it is different from the way the probability of a discrete variable is expressed. You will then learn about the **normal distribution**, which is a commonly used probability distribution among continuous random variables.

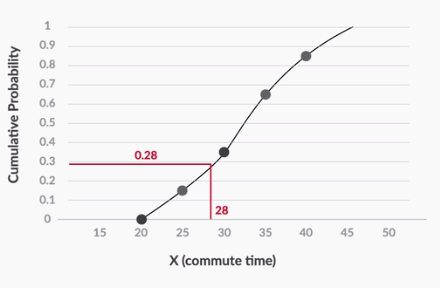
# Probability Density Functions - I

In the last section, you saw how to find the probability of certain events using multiplication and addition rules of probability. Also, for some specific cases, you saw that probability distributions like the binomial distribution and the uniform distribution can be used to find the probability.

However, so far we have only been talking about discrete random variables, e.g. number of balls, number of patients, cars, wickets, pasta packets, etc. What happens when we talk about the **probability of continuous random variables**, such as time, weight etc.? Is there any difference? Let’s see.

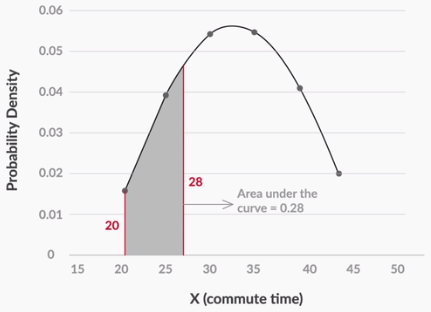
So, now you know what a **CDF** is and what a **PDF** is. Since these two functions talk about probabilities **in terms of intervals** rather than exact values, it is advisable to use them when talking about continuous random variables, and not the bar chart distribution that we used for discrete variables.

Just to recall, a **CDF**, or a **cumulative distribution function**, is a distribution which plots the cumulative probability of X against X.



**Figure 1 - CDF (Cumulative Distribution Function)**

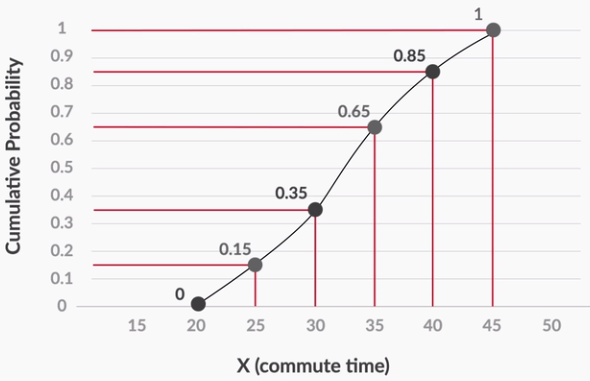
A **PDF**, or **Probability Density Function**, however, is a function in which the area under the curve, gives you the cumulative probability.



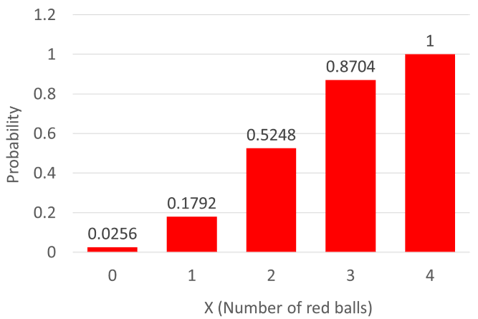
**Figure 2 - PDF (Probability Density Functions)**

For example, the area under the curve, between 20, the smallest possible value of X and 28, gives the cumulative probability for X = 28.

The main difference between the cumulative probability distribution of a continuous random variable and a discrete one, is the way you plot them. While the continuous variables’ cumulative distribution is a curve, the distribution for discrete variables looks more like a bar chart:



**Figure 3 - Cumulative Probability Distribution for Continuous Variables (Commute Time)**



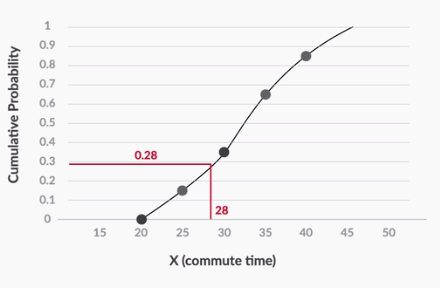
**Figure 4 - Cumulative Probability Distribution for Discrete Variables (Number of Red Balls)**

The reason for showing both of these so differently is that, for discrete variables, the cumulative probability does not change very frequently. In the discrete example, we only care about what the probability is for 0, 1, 2, 3 and 4. This is because the cumulative probability will not change between, say, 3 and 3.999999. For all values between these two, the cumulative probability is equal to 0.8704.

However, for the continuous variable, i.e. the daily commute time, you have a different cumulative probability value for every value of X. For example, the value of cumulative probability at 21 will be different from its value at 21.1, which will again be different from the one at 21.2 and so on. Hence, you would show its cumulative probability as a continuous curve, not a bar chart.

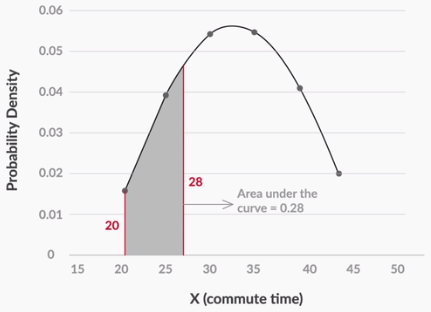
So, now you know what a **CDF** is and what a **PDF** is. Since these two functions talk about probabilities **in terms of intervals** rather than exact values, it is advisable to use them when talking about continuous random variables, and not the bar chart distribution that we used for discrete variables.

Just to recall, a **CDF**, or a **cumulative distribution function**, is a distribution which plots the cumulative probability of X against X.



**Figure 1 - CDF (Cumulative Distribution Function)**

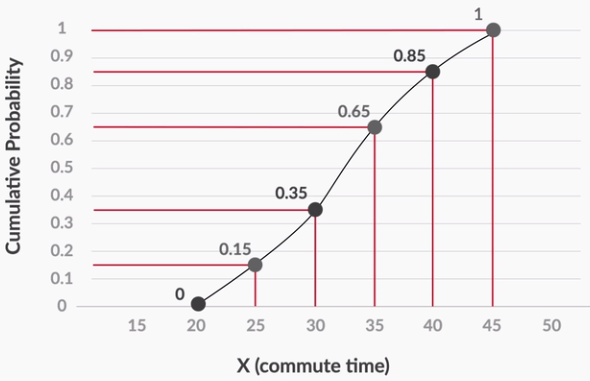
A **PDF**, or **Probability Density Function**, however, is a function in which the area under the curve, gives you the cumulative probability.



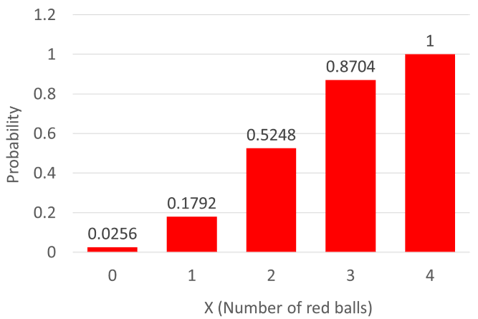
**Figure 2 - PDF (Probability Density Functions)**

For example, the area under the curve, between 20, the smallest possible value of X and 28, gives the cumulative probability for X = 28.

The main difference between the cumulative probability distribution of a continuous random variable and a discrete one, is the way you plot them. While the continuous variables’ cumulative distribution is a curve, the distribution for discrete variables looks more like a bar chart:



**Figure 3 - Cumulative Probability Distribution for Continuous Variables (Commute Time)**



**Figure 4 - Cumulative Probability Distribution for Discrete Variables (Number of Red Balls)**

The reason for showing both of these so differently is that, for discrete variables, the cumulative probability does not change very frequently. In the discrete example, we only care about what the probability is for 0, 1, 2, 3 and 4. This is because the cumulative probability will not change between, say, 3 and 3.999999. For all values between these two, the cumulative probability is equal to 0.8704.

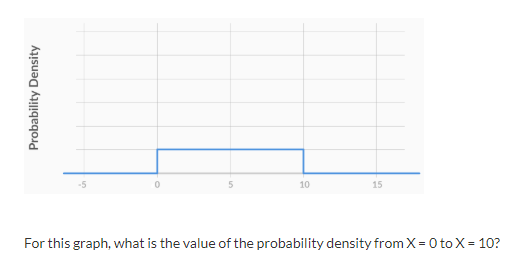
However, for the continuous variable, i.e. the daily commute time, you have a different cumulative probability value for every value of X. For example, the value of cumulative probability at 21 will be different from its value at 21.1, which will again be different from the one at 21.2 and so on. Hence, you would show its cumulative probability as a continuous curve, not a bar chart.

# Probability Density Functions - II

A commonly observed type of distribution among continuous variables is the **uniform distribution**. For a continuous random variable following a uniform distribution, the value of probability density is equal for all possible values

**Uniform Distribution**

In a uniform PDF, all the possible values have the same probability density. The figure below shows such a uniform PDF, where the possible values are 0 to 10.



For this graph, what is the value of the probability density from X = 0 to X = 10?

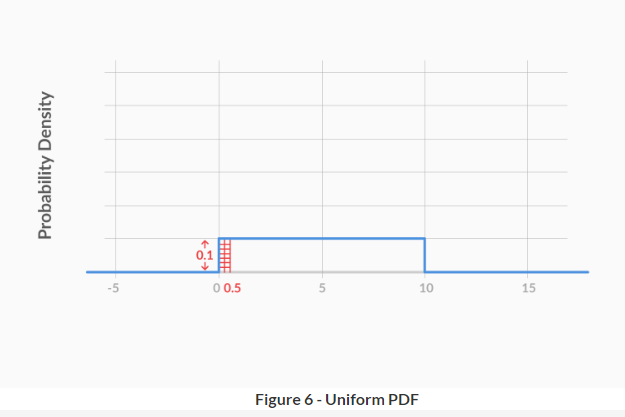
Top of Form



**Feedback :**The correct answer is 0.1. Let's see why, in the text below.

Bottom of Form

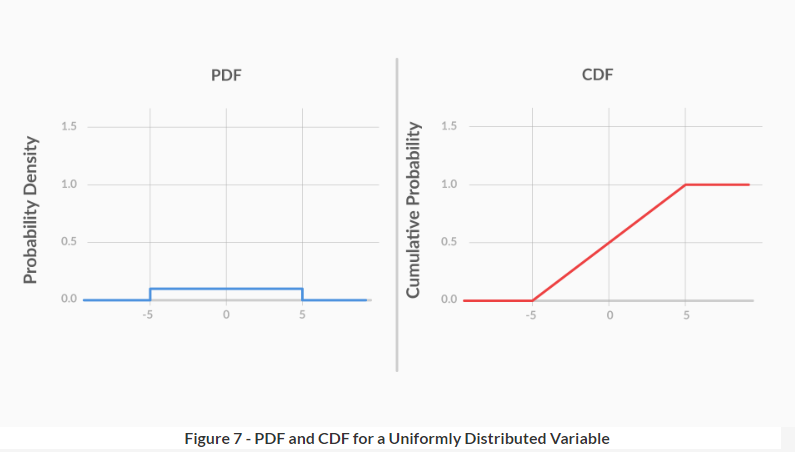
The cumulative probability for X = 0.5 is equal to the area under the curve between X = 0, the lowest possible value, and X = 0.5.



This area = 0.1\*0.5 = 0.05.

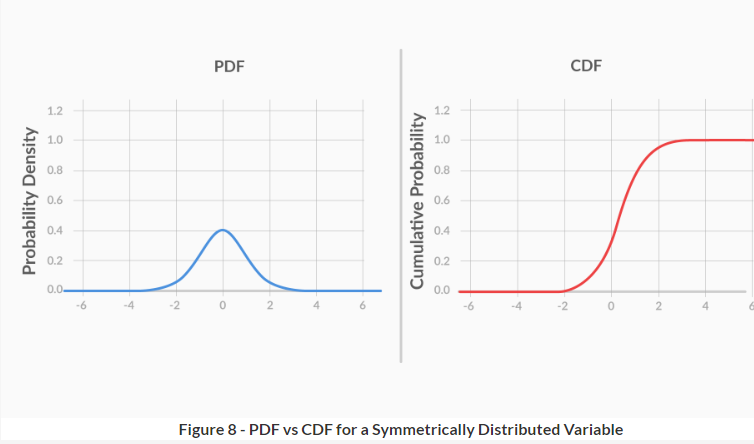
Now, I’m sure you are wondering, when to use PDFs and when to use CDFs? They are both good for continuous variables, but which one is used more in real life analysis?

Well, PDFs are more commonly used in real life. The reason is that it is much **easier to see patterns in PDFs** as compared to CDFs. For example, here are the PDF and the CDF of a uniformly distributed continuous random variable:



The **PDF clearly shows uniformity**, as the probability density’s value remains constant for all possible values. However, the **CDF does not show any trends** that help you identify quickly that the variable is uniformly distributed.

Now, let’s see the PDF and the CDF of a symmetrically distributed continuous random variable:



Again, it is clear that the symmetrical nature of the variable is much more apparent in the PDF than in the CDF.

Hence, generally, PDFs are used more commonly that CDFs.

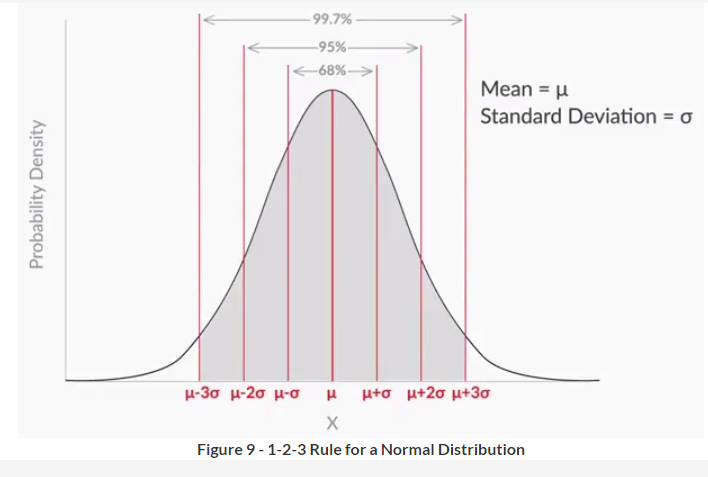
# Normal Distribution

You’ve seen how the probability distributions of continuous random variables differ from those of discrete random variables.

But can you think of some examples of continuous distributions? Which is the most commonly used continuous probability distribution? Which distribution occurs most commonly in nature? Let’s hear from Prof. Tricha on this.

All data that is normally distributed follows the **1-2-3 rule**. This rule states that there is a -

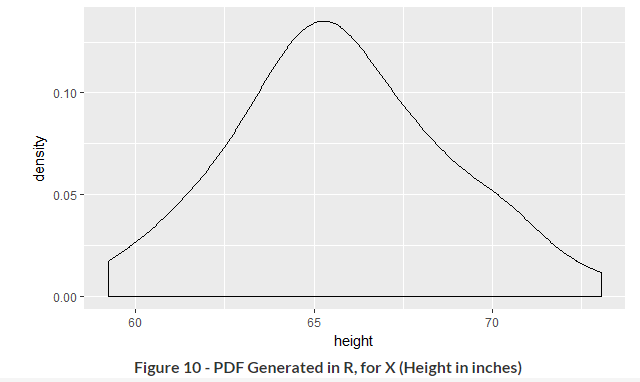
1. **68%** probability of the variable lying **within 1 standard deviation** of the mean
2. **95%** probability of the variable lying **within 2 standard deviations** of the mean
3. **99.7%** probability of the variable lying **within 3 standard deviations** of the mean



This is actually like saying that, if you buy a loaf of bread everyday and measure it, then - (mean weight = 100 g, standard deviation = 1 g)

1. For 5 days every week, the weight of the loaf you bought that day will be within 99 g (100-1) and 101 g (100+1).
2. For 20 days every 3 weeks, the weight of the loaf you bought that day will be within 98 g (100-2) and 102 g (100+2).
3. For 364 days every year, the weight of the loaf you bought that day will be within 97 g (100-3) and 103 g (100+3).

A lot of naturally occurring variables are normally distributed. For example, the heights 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 that data.



As you can see, the data is roughly normal.

You can visualise the PDF and CDF for different normal distributions by using the interactive app given below. From the various options in the drop-down menu, select **“Normal”**. You will then get to see the probability distribution for a normal distribution with µ = 0 and σ = 1. In fact, you can **play around with the value of µ and σ** to see how that changes the distribution. Using the **green slider**below the distribution, you can visualise the distribution’s **cumulative probability**.

In fact, you can select **“Uniform”** in the drop-down menu and visualise the CDF and PDF for the uniform distribution too. Be sure to **play around with a and b**, which give the lowest and highest possible values, respectively, for the random variable.

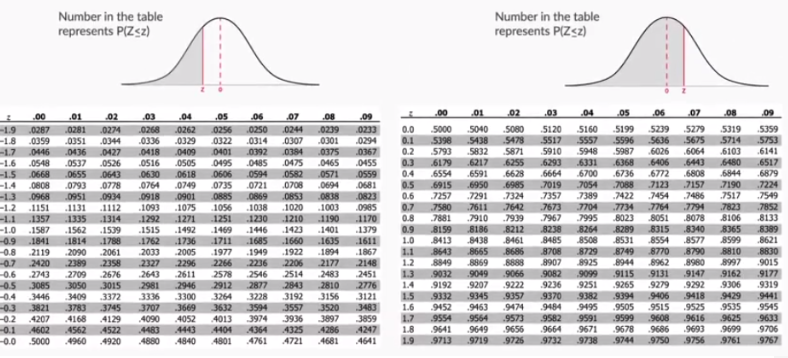
# Standard Normal Distribution

As you learnt in the previous question, it doesn’t matter what the value of µ and σ is. All you need to know, if you want to find the probability, is how far the value of X is from µ — specifically, what **multiple of σ** is the **difference between X and µ**.

As you just learnt, the **standardised random variable** is an important parameter. It is given by:

Z=X−μσZ=X−μσ

Basically, it tells you **how many standard deviations away from the mean** your random variable is. As you just saw, you can find the cumulative probability corresponding to a given value of Z, using the **Z table**:



Alternatively, you can use the following equation to find the cumulative probability:

F(Z)=1√2π∫Z−∞e−t22dtF(Z)=12π∫−∞Ze−t22dt

However, I’ve a feeling that you will prefer the table! :-)

Not only that, you can also use **Excel** or **R** to find the cumulative probability for Z. For example, let’s say you want to find the cumulative probability for Z = 1.5. In Excel, you would type:

= NORM.S.DIST(1.5, TRUE)

Basically, the syntax is:

= NORM.S.DIST(z, TRUE)

Here, z is the value of the Z score for which you want to find the cumulative probability. TRUE = find cumulative probability, FALSE = find probability density.

Also, you can find the probability without standardising. Let’s say that X is normally distributed, with mean (μ) = 35 and standard deviation (σ) = 5. Now, if you want to find the cumulative probability for X = 30, you would type:

= NORM.DIST(30, 35, 5, TRUE)

Basically, the syntax is:

= NORM.DIST(x, mean, standard\_dev, TRUE)

In R, you would type pnorm(z) to find the probability for standard normal variable. For example, the cumulative probability for Z = 1.5 can be found by typing:

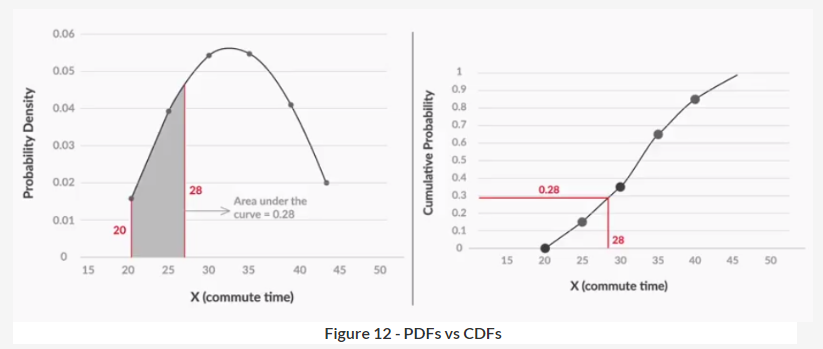
pnorm(1.5)

# Summary: Continuous Probability Distributions

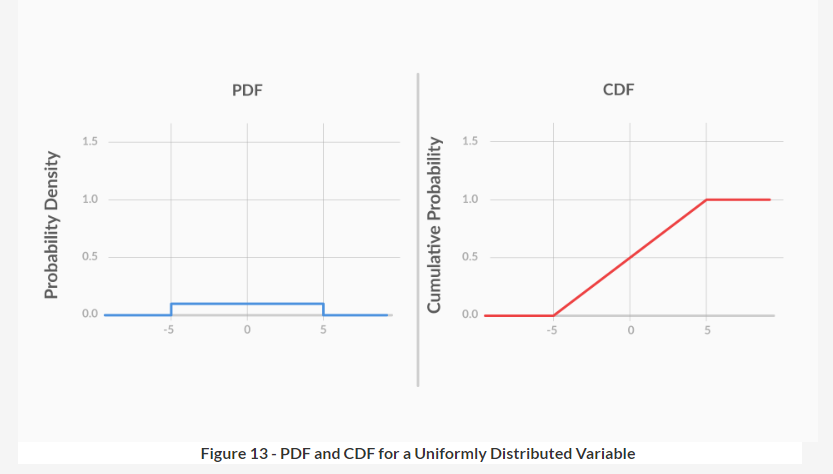
You started this session by learning that, for a **continuous random variable**, the **probability of getting an exact value is** very low, almost **zero**. Hence, when talking about the probability of continuous random variables, you can only talk **in terms of intervals**. For example, for a particular company, the probability of an employee’s commute time being exactly equal to 35 minutes was zero, but the probability of an employee having a commute time 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 the bar chart type of distribution used for the probability of discrete random variables. These functions are preferred because they talk about probability in terms of intervals.

Then, you understood that the major difference between a PDF and a CDF is that in a CDF, you can find the cumulative probability directly by checking the value at x. However, for a PDF, you need to find the area under the curve between the lowest value and x to find the cumulative probability.

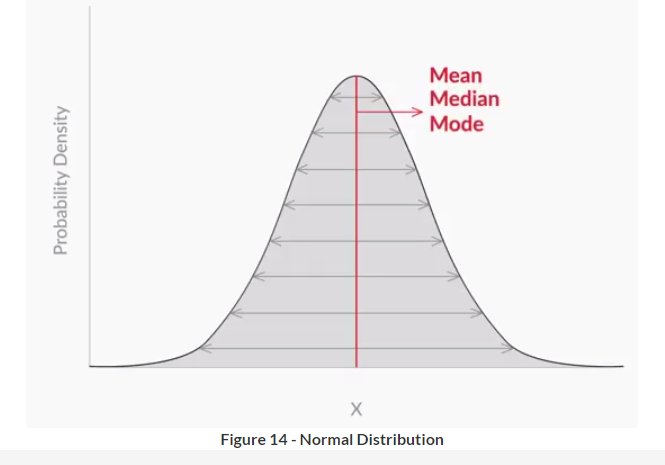


However, you also learnt that **PDFs are still more commonly used**, mainly because it is very **easy to see patterns** in them. For example, for a uniformly distributed variable, the PDF and CDF look like this:



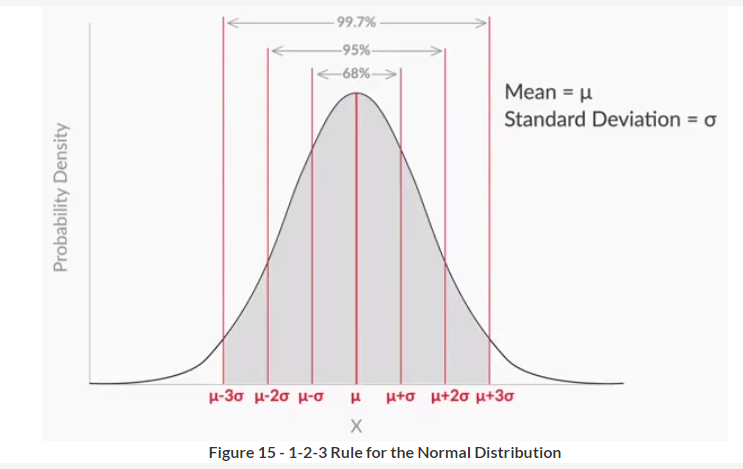
While the fact that the variable is uniformly distributed is clear from the PDF, the CDF does not offer any such quick insights.

Next, you learnt about a very famous probability density function — the **normal distribution**. You saw that it is **symmetric** and its **mean, median and mode** lie at the **centre**.



You also learnt the **1-2-3 rule**, which states that there is a -

1. **68%** probability of the variable lying **within 1 standard deviation** of the mean
2. **95%** probability of the variable lying **within 2 standard deviations** of the mean
3. **99.7%** probability of the variable lying **within 3 standard deviations** of the mean

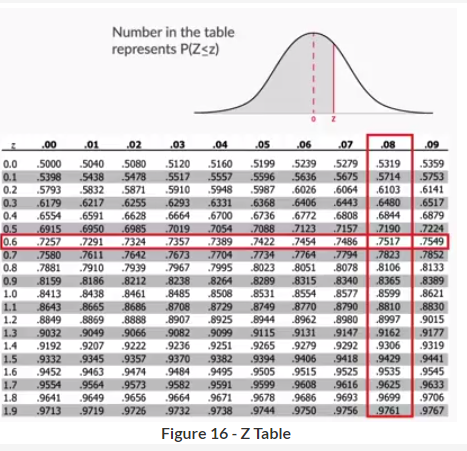


However, you saw that, to find the probability, you do not need to know the value of the mean or the standard deviation — it is enough to know **the number of standard deviations away from the mean**your random variable is. That is given by:

Z=(X−μσ)Z=(X−μσ)

This is called the **Z score**, or the **standard normal variable**.

Finally, you learnt how to find the cumulative probability for various values of Z, using the **Z table**. For example, you found the cumulative probability for Z = 0.68 by using the Z table.



The intersection of row “0.6” and column “0.08”, i.e. 0.7517, is your answer.

Also, you learnt how to use Excel or R to find this probability. For example, cumulative probability for Z =1.5 can be found using Excel by typing:

= NORM.DIST(1.5, TRUE)

In R, you would type:

= pnorm(1.5)

Also, you can find the probability without standardising. The syntax for that is:

= NORM.DIST(x, mean, standard\_dev, TRUE)

The normal distribution finds use in many statistical analyses. In our module, it finds use in the next session, central limit theorem, which is then useful for understanding the next module on hypothesis testing.

# Introduction: Central Limit Theorem

Welcome to the session on 'Central Limit Theorem'. In the last session, you learnt about probability density functions, specifically the normal and standard normal distributions.

## In this session

In this session, you will learn what a sample is and why it is so error prone. You will then learn how to quantify this error made in sampling, using a very popular theorem in statistics, called the central limit theorem.

## Prerequisites

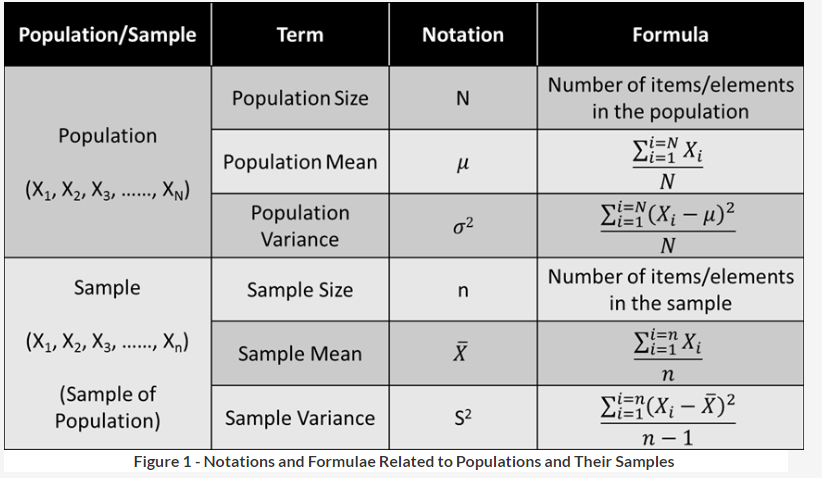
There are no prerequisites for this session, other than of course, knowledge of the previous three sessions!

# Samples

So far, you have conducted analysis for data on 75 people, 3,000 people, and so on. But what if you need to analyse a very large amount of data, e.g. data on 300,000 people? Or what if you need to do this for, say, the entire Indian population?

Let’s say that, for a business application, you want to find out the average number of times people in urban India visited malls last year. That’s 400 million (40 crore) people! You can't possibly go and ask every single person how many times they visited the mall. That’s a costly and time-consuming process. How can you reduce the time and money spent on finding this number?

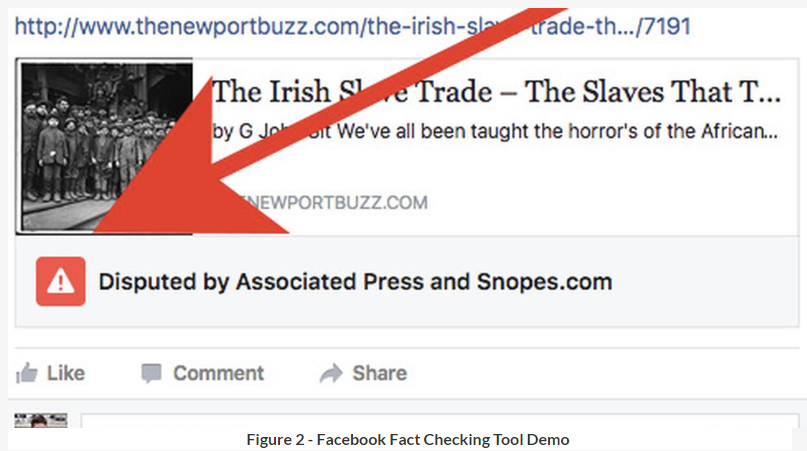
To reiterate, these are the notations and formulae related to populations and their samples:



The reason for dividing by n-1 and not n is not covered in this course. If you are still curious as to why that is the case, please refer to this [link](https://www.ma.utexas.edu/users/mks/M358KInstr/SampleSDPf.pdf).

Let's go through another example of sampling.

In order to counter fake news, let’s say Facebook is planning to include a new feature in its timeline. Below each post, a fact-checking warning will be provided, like this:



**Figure 2 - Facebook Fact Checking Tool Demo**

In case you want to read more about this feature, you can do so by going through this [link](http://fortune.com/2017/03/22/facebook-fact-checking-tool/).

Before changing the timelines of all Facebook users to include this feature, Facebook first wants to evaluate how its users would react to this new feature.

So, it lets a small sample (~10,000 users) try out the new timeline. Then, it asks the 10,000 users whether they prefer the new timeline (Feature B) or the old timeline (Feature A).

Let’s say that one such survey shows that 50.5% of the people prefer feature B over feature A. Based on this, Facebook can say that feature B is preferred by more people than feature A, and hence should replace it.

Hence, by sampling, Facebook found that feature B is preferred by more people than feature A. By conducting the exercise on a sample and not the population, it **saved time**, **money** and **avoided risks**that would come if it would have rolled out an untested feature.

But hold on! How can you be sure that the insights inferred for the sample hold true for the population as well? In other words, just because 50.5% of the people in the sample preferred feature B, is it fair to infer that 50.5% of the people in the population (1.86 billion Facebook users) will also prefer feature B over A?

You cannot answer this question with the information you have right now. However, after the next few lectures, which will cover sampling distributions, central limit theorem and confidence intervals, you will be in a situation to answer this question.

# Sampling Distributions

Now, let's move on to sampling distributions, whose properties, as we said earlier, will help you estimate the population mean from the sample mean.

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

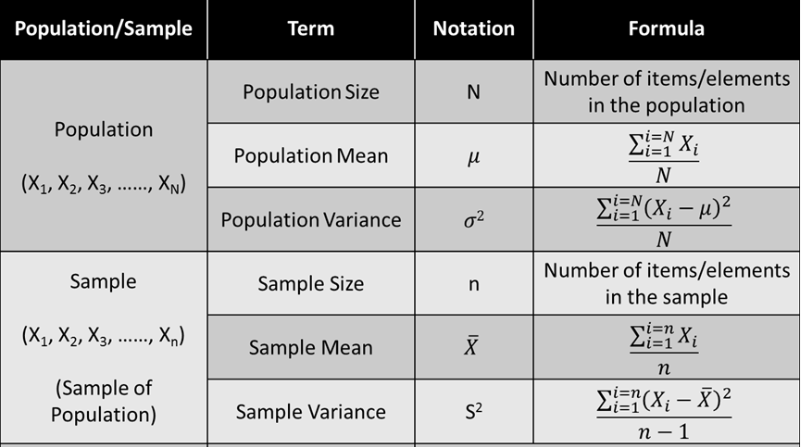
This distribution has some very interesting properties, which will later help you estimate the sampling error. Let's take a look at these properties.

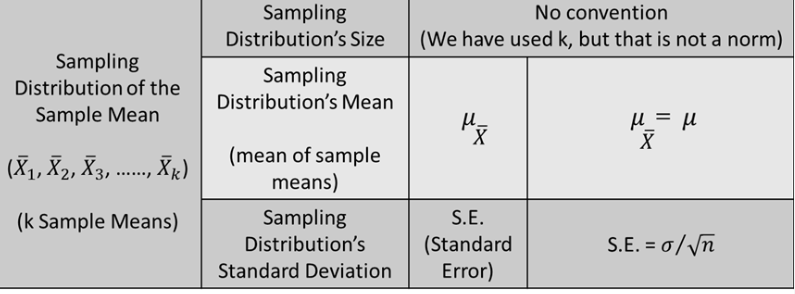
The **sampling distribution’s mean is denoted by**μ¯XμX¯**,** as it is the mean of the sampling means. Let’s see what it is for this sampling distribution.

# Properties of Sampling Distributions

We’ve been saying that the sampling distribution has some interesting properties that will later help you estimate the error in your samples. Let’s finally see what these properties are.

Again, to recap, let’s see what the notations and formulae are for populations, samples and sampling distributions.





So, there are two important properties for a sampling distribution of the mean:

1. **Sampling distribution’s mean** (μ¯XμX¯) = **Population mean**(μ)
2. Sampling distribution’s standard deviation (**Standard error**) = σ√nσn, where σ is the population’s standard deviation and n is the sample size

Recall that, in the last video, we created a sampling distribution and found the value of its mean and standard deviation, which turned out to be 2.348 and 0.4248, respectively. These values were very close to the values suggested by the formula, i.e. 2.385 and 0.44.

However, you may be thinking “Well, I want to try that out too.” Unfortunately, that would take a lot of time. But that’s not a problem, as you can simulate the whole process in R. In the next lecture, you will see how this can be done.

# Sampling Distributions - R Simulation

So, how can you simulate the process of making a sampling distribution in R, in order to confirm what theory tells you about them? Let’s hear from Kshitij on how that can be done.

(This whole process is just being shown for your understanding, you will not be tested in this module on how to simulate sampling distributions in R.)

Notice that, at the end of the video, Kshitij says that **for n > 30**, the sampling distribution is expected to be a **normal curve**. This is a very important property of sampling distributions.

Actually, even for n < 30, the sampling distribution is usually approximated to a normal curve. That is not a good approximation, but it is used when necessary. You will also learn later how a different curve, called the t distribution, is used for n < 30.

Actually, even for n < 30, the sampling distribution is usually approximated to a normal curve. That is not a good approximation, but it is used when necessary. You will also learn later how a different curve, called the t distribution, is used for n < 30.

# Central Limit Theorem

Now, you understand the third property for sampling distributions, which talks about their shape. Basically, it says that for n > 30, the sampling distributions become normally distributed. Let's recall all the three properties you have learnt so far for sampling distributions.

So, the central limit theorem says that, for any kind of data, provided a high number of samples has been taken, the following properties hold true:

1. **Sampling distribution’s mean** (μ¯XμX¯) = **Population mean** (μ)
2. Sampling distribution’s standard deviation (**Standard error**) = σ√nσn
3. **For n > 30**, the sampling distribution becomes a **normal distribution**

We made two sampling distributions (UpGrad game and Banking data set) and saw that they follow these three properties.

Now, let’s listen to Prof. Tricha as she verifies the central limit theorem for some more population distributions.

Prof. Tricha verified CLT by performing simulations on different kinds of data. In case you want to try out these simulations yourself, you can go to this [link](http://onlinestatbook.com/stat_sim/sampling_dist/index.html) and try them out. Press the "Begin" button in the top-right corner to get started.

Recall that, in the first lecture on samples, we found the mean commute time of 30,000 employees of an office by taking a small sample of 100 employees and finding their mean commute time. This sample’s mean was \bar{X} = 36.6 minutes and its standard deviation was S = 10 minutes.

We then said that this sample mean cannot be taken as the population mean, as there might be some errors in the sampling process. However, we can say that the population mean, i.e. daily commute time of all 30,000 employees \bar{X} = 36.6 (sample mean) + some margin of error.

Now, you may be thinking that you can use the standard error for the margin of error. But, although the standard error provides a good estimate of this margin of error, you cannot use it in place of the margin of error. To understand why, and how you would find the margin of error in that case, let's move on to the next lecture, where we will use CLT (central limit theorem) to find the aforementioned margin of error.

# Summary: Central Limit Theorem - Part I

This is a really intense session! Let’s summarise everything that's been taught so far and then you can move on the rest of the session.

First, you saw how, instead of finding the mean and standard deviation for the entire population, it is sometimes 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, for an office of 30,000 employees, we wanted to find the average commute time. So, instead of asking all employees, we asked only 100 of them and collected the data. The mean = 36.6 minutes and the standard deviation = 10 minutes.

However, 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 sample mean’s value has to be reported with some **margin of error**.

For example, the mean commute time for the office of 30,000 employees would be equal to 36.6 + 3 minutes, 36.6 + 1 minutes, or 36.6 + 10 minutes, i.e. 36.6  minutes + some margin of error.

However, at this point in time, you do not exactly know how to find what this margin of error is.

Then, we moved on to sampling distributions, which have some properties that would help you find this margin of error.

We created a sampling distribution, which was a probability density function for 100 sample means with sample size 5.

The sampling distribution, which is basically the distribution of sample means of a population, has some interesting properties which are collectively called the **central limit theorem**, which states that no matter how the original population is distributed, the sampling distribution will follow these three properties:

1. **Sampling distribution’s mean** (\mu_{\bar{X}}) = **Population mean** (\mu)
2. Sampling distribution’s standard deviation (**Standard error**) = \frac{\sigma}{\sqrt{n}}, where \sigma is the population’s standard deviation and n is the sample size
3. **For n > 30**, the sampling distribution becomes a **normal distribution**

To verify these properties, we performed sampling using the data collected for our UpGrad game from the first session on inferential statistics. The values for the sampling distribution thus created (\mu_{\bar{X}} = 2.348, S.E. = 0.4248) were pretty close to the values predicted by theory (\mu_{\bar{X}} = 2.385, S.E. = 0.44).

To summarise, the notations and formulae for populations, samples and sampling distributions are as follows: