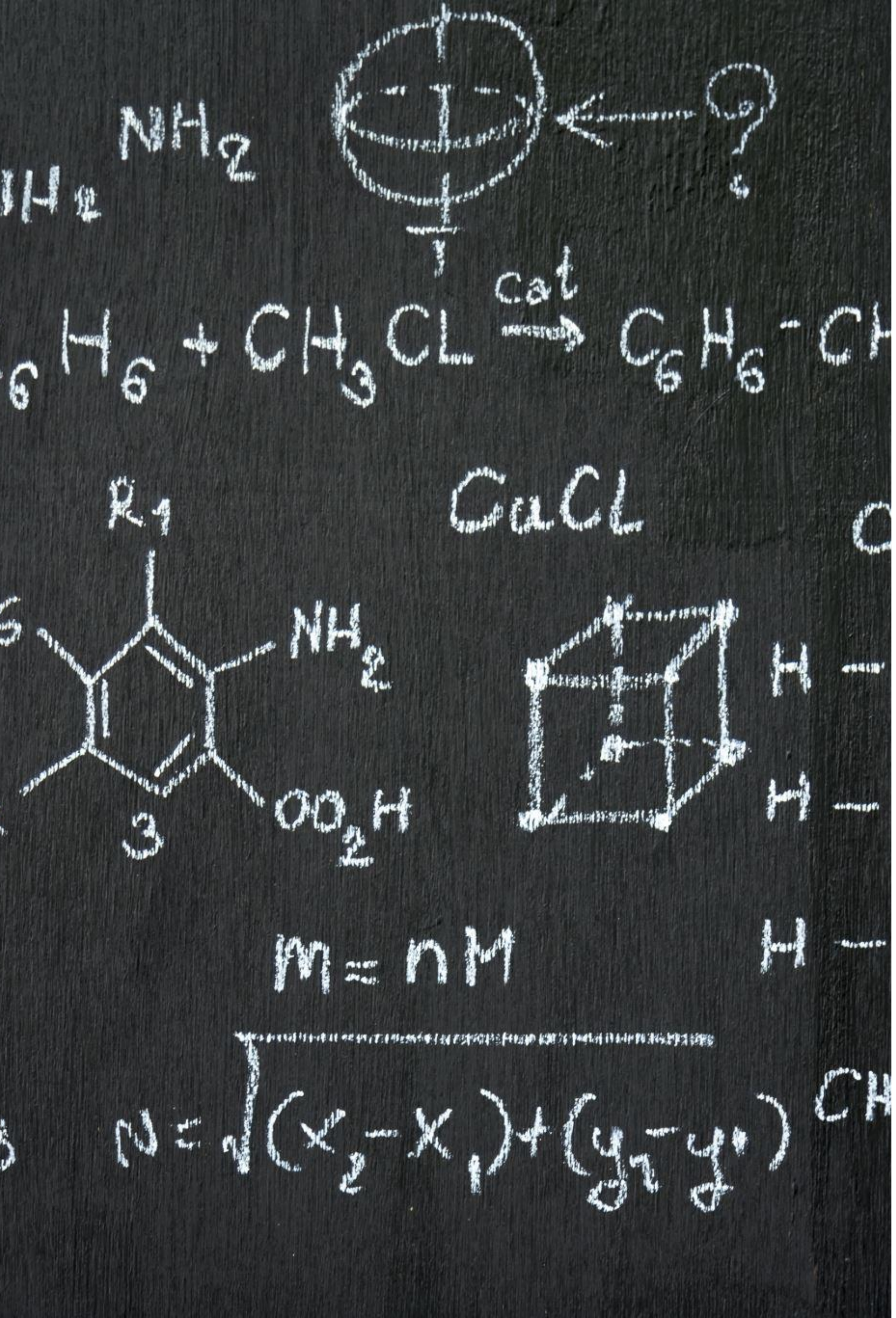
Probabilistic and Inferential Statistics



Random Variable:

- Real value of random experiment is called random variable.

- e.g.: Toss two coins
- sample space = <HH,HT,TH,TT>
- Random Variable x= no. of heads
-
- X= (HH) =2
- X=(HT)=1
- X=(TH)=1
- X=(TT)=0
-
- X may be 0,1,2 (Random Variable) with its range.

Continued …

- e.g.: Throw a dice
- sample space= {1,2,3,4,5,6}
- Random variable x= no of odd no
- X(1)=1
- X(2)=0
- X(3)=1
- X(4)=0
- X(5)=1
- X(6)=0
- -□ Discrete random variable:
- Countable
- -□ Continuous random variable:
- a<X<b
-
- x= continuous random variable greater than a and less than b but continuously change the values.

PROBABILITY

- Probability always measures between 0 to 1…(low to high)
-
- Probability Type: 1) Independent / Dependent
- 2) Mutually Exclusive (Can’t execute two events at same time)
-
- Joint Probability: A= Getting HEAD
- B = Getting Even no when we rolled a DICE
-
- P(A)=1/2 P(B)= 1/2
- Joint Probability = P(AΩB) = 1/2 * 1/2 =1/4

Continued …

- Conditional probability:
- A= Getting even no of DICE
- B= Getting a of greater than 4

- $P(A|B)= P(A\Omega B) / P(B) = (1/6) / (2/6) =1 / 2$

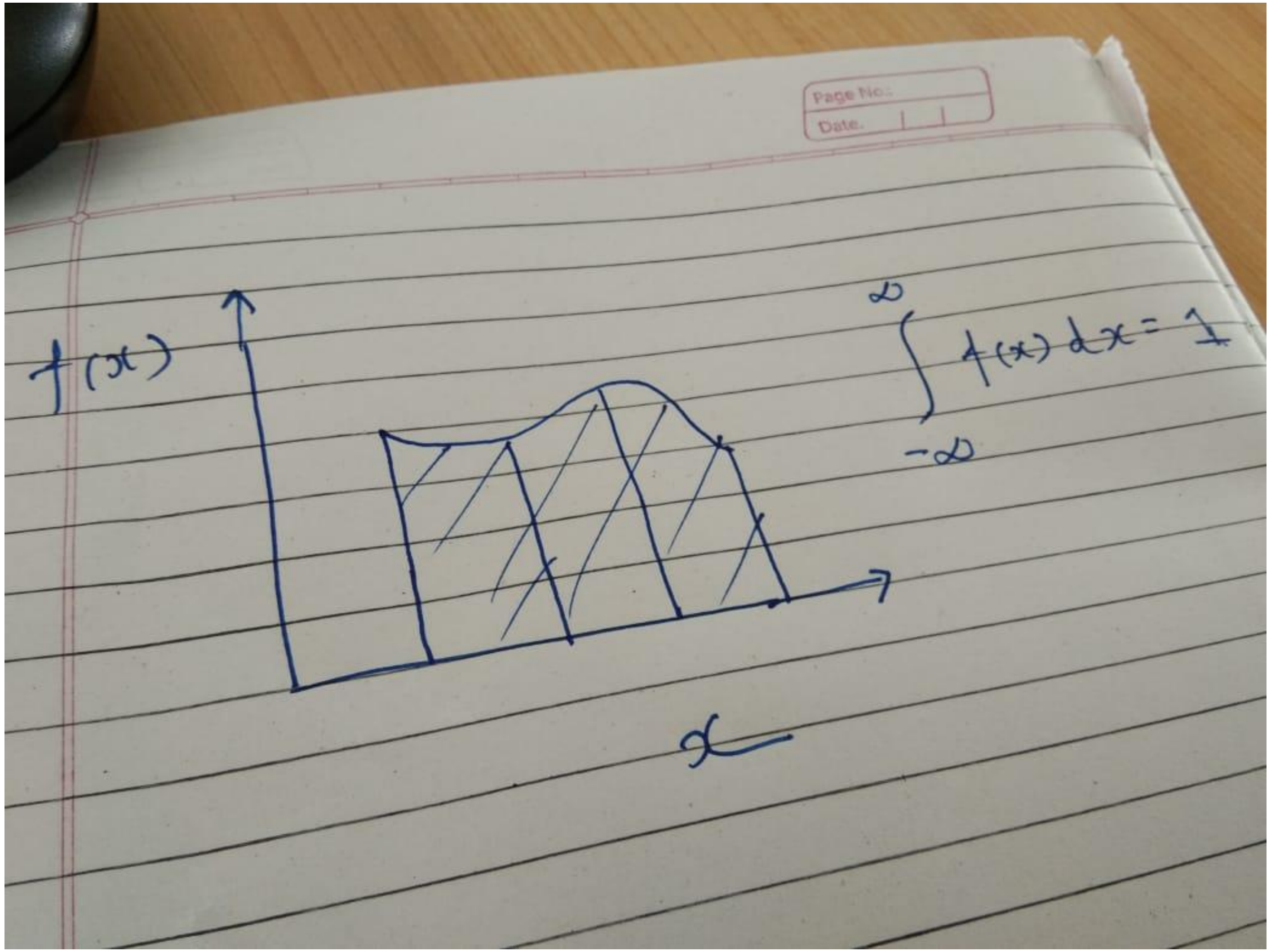
Probability mass function (Regarding to Discrete Variable)

- Example: Set of two coins ={HH,HT,TH,TT}
- X (Random Variable) : No. of Heads
-
- P(X=0) =1/4
- P(X=1)=2/4 -□ Convert this into graph is Probability mass function
- P(X=2)=1/4

- **Cumulative distribution function**

- Example: Set of two coins ={HH,HT,TH,TT}
- X (Random Variable) : No. of Heads
-
- P(X=0) =1/4
- P(X=1)=2/4 -□ (Make a cumulative like the probability of getting 0 or 1 Head)Convert this
- into graph **Cumulative** mass function (1/4 + 2/4= 3/4)
- P(X=2)=1/4

Probability Density function (Regarding to Continuous Variable)



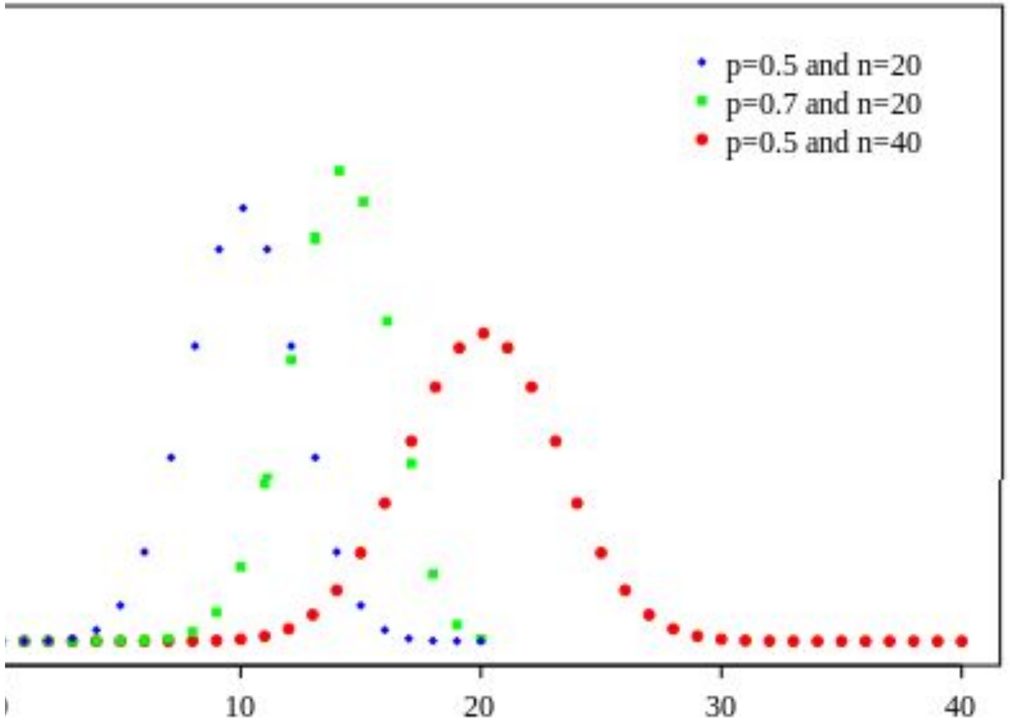
Binomial and Poisson Distribution

- A **binomial distribution** can be thought of as simply the probability of a SUCCESS or FAILURE outcome in an experiment or survey that is repeated multiple times.
- The binomial is a type of distribution that has **two possible outcomes** (the prefix “bi” means two, or twice).
- For example, a coin toss has only two possible outcomes: heads or tails and taking a test could have two possible outcomes: pass or fail.

- The first variable in the binomial formula, **n**, stands for the number of times the experiment runs.
- The second variable, **p**, represents the probability of one specific outcome.
- For example, let’s suppose you wanted to know the probability of getting a 1 on a die roll.
- If you were to roll a die 20 times, the probability of rolling a one on any throw is 1/6.
- Roll twenty times and you have a binomial distribution of (n=20, p=1/6). SUCCESS would be “roll a one” and FAILURE would be “roll anything else.” If the outcome in question was the probability of the die landing on an even number, the binomial distribution would then become (n=20, p=1/2). That’s because your probability of throwing an even number is one half.

Continued …

- *A Binomial Distribution shows either (S)uccess or (F)ailure.*



Criteria

- Binomial distributions must also meet the following three criteria:
- **The number of observations or trials is fixed.** In other words, you can only figure out the probability of something happening if you do it a certain number of times. This is common sense—if you toss a coin once, your probability of getting a tails is 50%. If you toss a coin a 20 times, your probability of getting a tails is very, very close to 100%.
- **Each observation or trial is** independent. In other words, none of your trials have an effect on the probability of thenext trial.
- The **probability of success** (tails, heads, fail or pass) is **exactly the same** from one trial to another.
- Once you know that your distribution is binomial, you can apply the binomial distribution formula to calculate the probability.

What is a Binomial Distribution?

- Many instances of binomial distributions can be found in real life.
- For example, if a new drug is introduced to cure a disease, it either cures the disease (it’s successful) or it doesn’t cure the disease (it’s a failure).
- If you purchase a lottery ticket, you’re either going to win money, or you aren’t.
- Basically, anything you can think of that can only be a success or a failure can be represented by a binomial distribution.

The binomial distribution formula is:

b(x; n, P) = nCx * Px * (1 – P)n – x

Where:

- b = binomial probability
- x = total number of “successes” (pass or fail, heads or tails etc.)
- P = probability of a success on an individual trial n = number of trials

Using the First Binomial Distribution Formula

The binomial distribution formula can calculate the probability of success for binomial distributions.

Example 1

Q. A coin is tossed 10 times. What is the probability of getting exactly 6 heads?

We can use this formula:

$b(x; n, P) = nCx * P^{x} * (1 - P)^{n - x}$

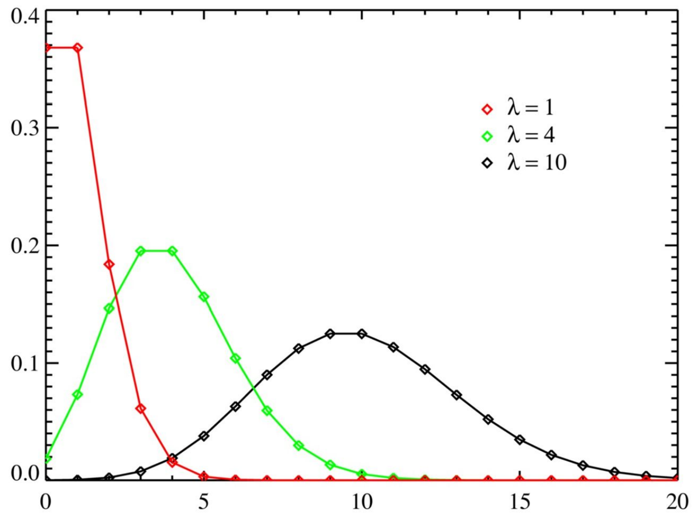
The number of trials (n) is 10

The odds of success (“tossing a heads”) is 0.5 (So 1-p = 0.5)

x = 6

P(x=6) = 10C6 * 0.5^6 * 0.5^4 = 210 * 0.015625 * 0.0625 = 0.205078125

Poisson Distribution

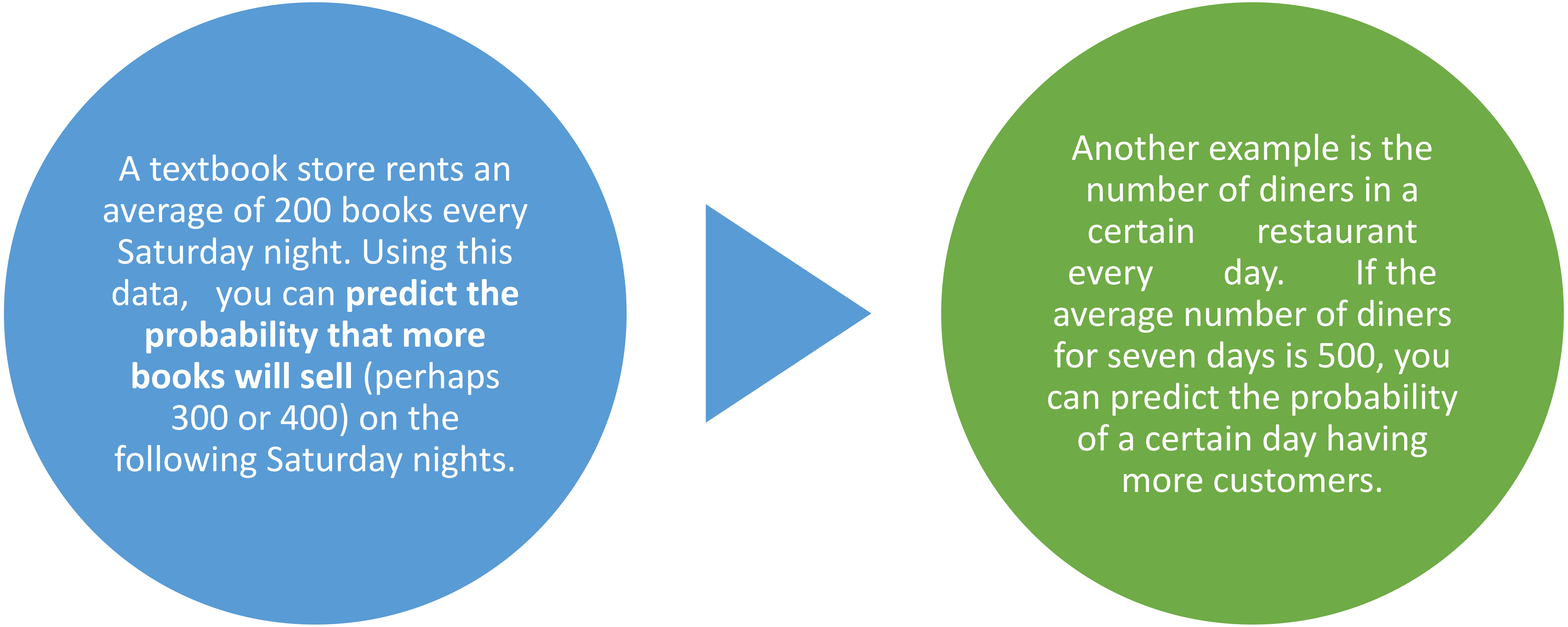


A Poisson distribution is a tool that helps to predict the probability of certain events from happening when you know how often the event has occurred.

It gives us the **probability of a given number of events happening in a fixed interval of time**.

Poisson distributions, valid only for integers on the horizontal axis. λ (also written as μ) is the expected number of event occurrences.

Practical Uses of the Poisson Distribution



A textbook store rents an average of 200 books every Saturday night. Using this data, you can **predict the probability that more books will sell** (perhaps 300 or 400) on the following Saturday nights.

Another example is the number of diners in a certain restaurant every day. If the average number of diners for seven days is 500, you can predict the probability of a certain day having more customers.

Continued …

- Because of this application, Poisson distributions are used by businessmen to make **forecasts** about the number of customers or sales on certain days or seasons of the year.
- In business, overstocking will sometimes mean losses if the goods are not sold. Likewise, having too few stocks would still mean a lost business opportunity because you were not able to maximize your sales due to a shortage of stock.
- By using this tool, businessmen are able to estimate the time when demand is unusually higher, so they can purchase more stock.
- Hotels and restaurants could prepare for an influx of customers, they could hire extra temporary workers in advance, purchase more supplies, or make contingency plans just in case they cannot accommodate their guests coming to the area.
- With the Poisson distribution, companies can adjust supply to demand in order to keep their business earning good profit. In addition, waste of resources is prevented.

Calculating the Poisson Distribution

- *The Poisson Distribution pmf is: P(x; μ) = $(e^{-\mu} * \mu^{x}) / x!$*
- Where:
- The symbol “!” is a factorial.
- μ (the expected number of occurrences) is

sometimes written as λ. Sometimes called the **event rate** or rate parameter.

Example question

The average number of major storms in your city is 2 per year. What is the probability that exactly 3 storms will hit your city next year?

Step 1: Figure out the components you need to put into the equation.

$\mu = 2$ (average number of storms per year, historically)

$x = 3$ (the number of storms we think might hit next year)

$e = 2.71828$ (e is Euler's number, a constant)

Step 2: Plug the values from Step 1 into the Poisson distribution formula:

$P(x; \mu) = (e^{-\mu}) (\mu^{x}) / x!$

$= (2.71828^{-2}) (2^3) / 3!$

$= (0.13534) (8) / 6$

$= 0.180$

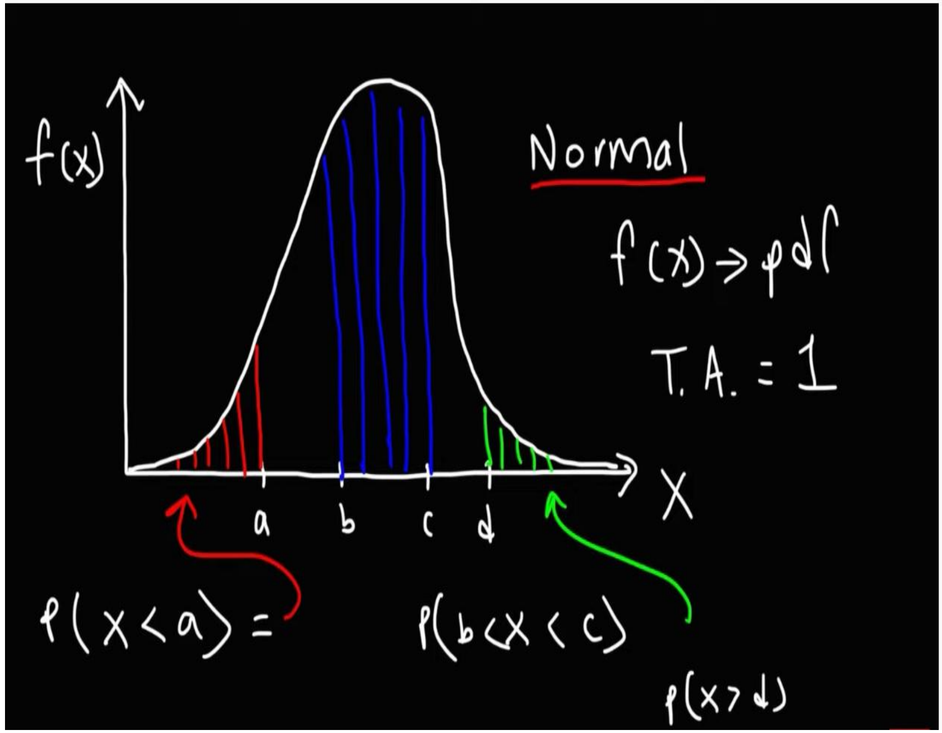
The probability of 3 storms happening next year is 0.180, or 18%

Poisson distribution vs. Binomial

If your question has an **average probability** of an event happening per unit (i.e. per unit of time, cycle, event) **and** you want to find probability of a certain number of events happening in a period of time (or number of events), then use the Poisson Distribution.

If you are given an **exact probability** and you want to find the probability of the event happening a certain number out times out of x (i.e. 10 times out of 100, or 99 times out of 1000), use the **Binomial Distribution formula**.

Continuous probability distributions (Normal and Exponential).



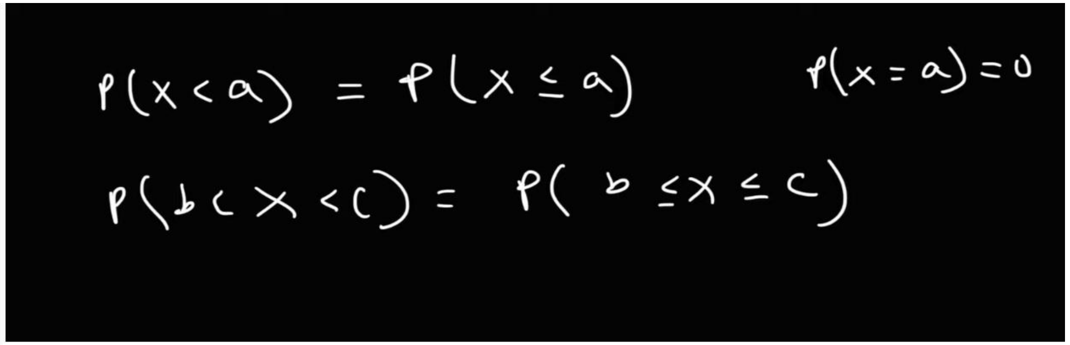
$f(x) \rightarrow pdf$

$T.A. = 1$

$P(x<a) =$

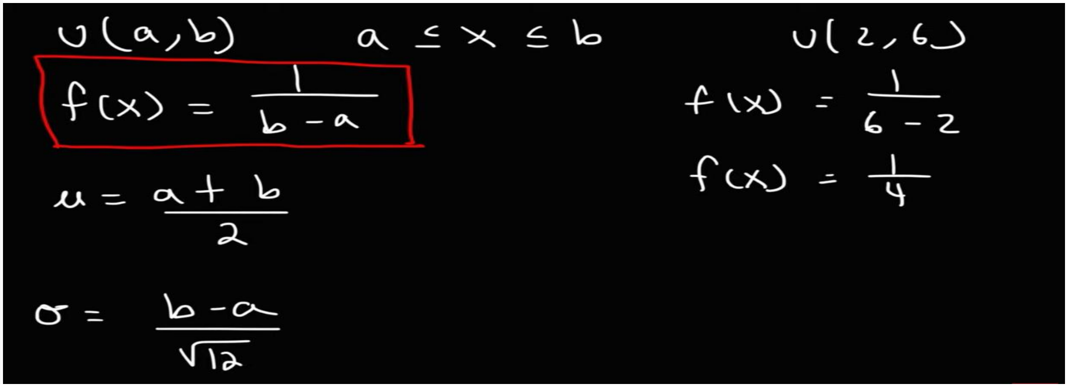
$P(b<X<c)$

$P(x>d)$



$P(x<a) = P(x \le a)$ $\quad P(x=a)=0$

$P(b<x<c) = P(b \le x \le c)$



$U(a,b)$ $\quad a \le x \le b$ $\quad U(2,6)$

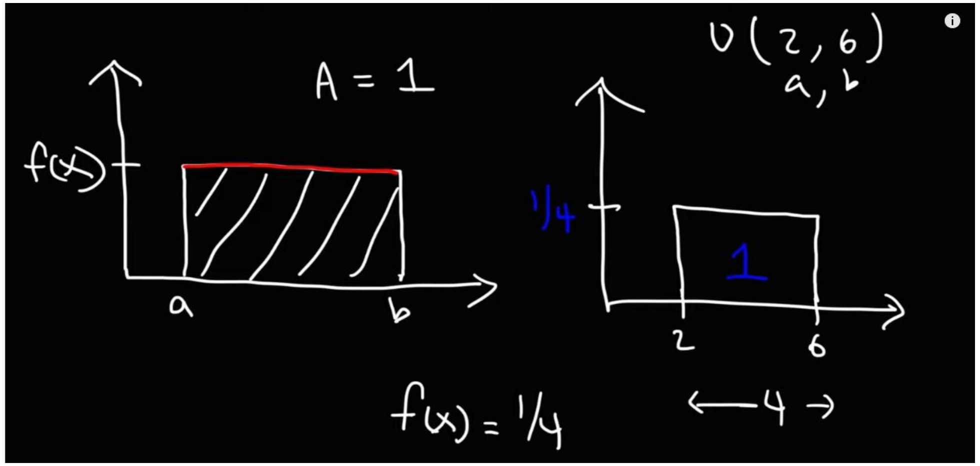
$$\boxed{f(x) = \frac{1}{b-a}}$$

$$f(x) = \frac{1}{6-2}$$

$$f(x) = \frac{1}{4}$$

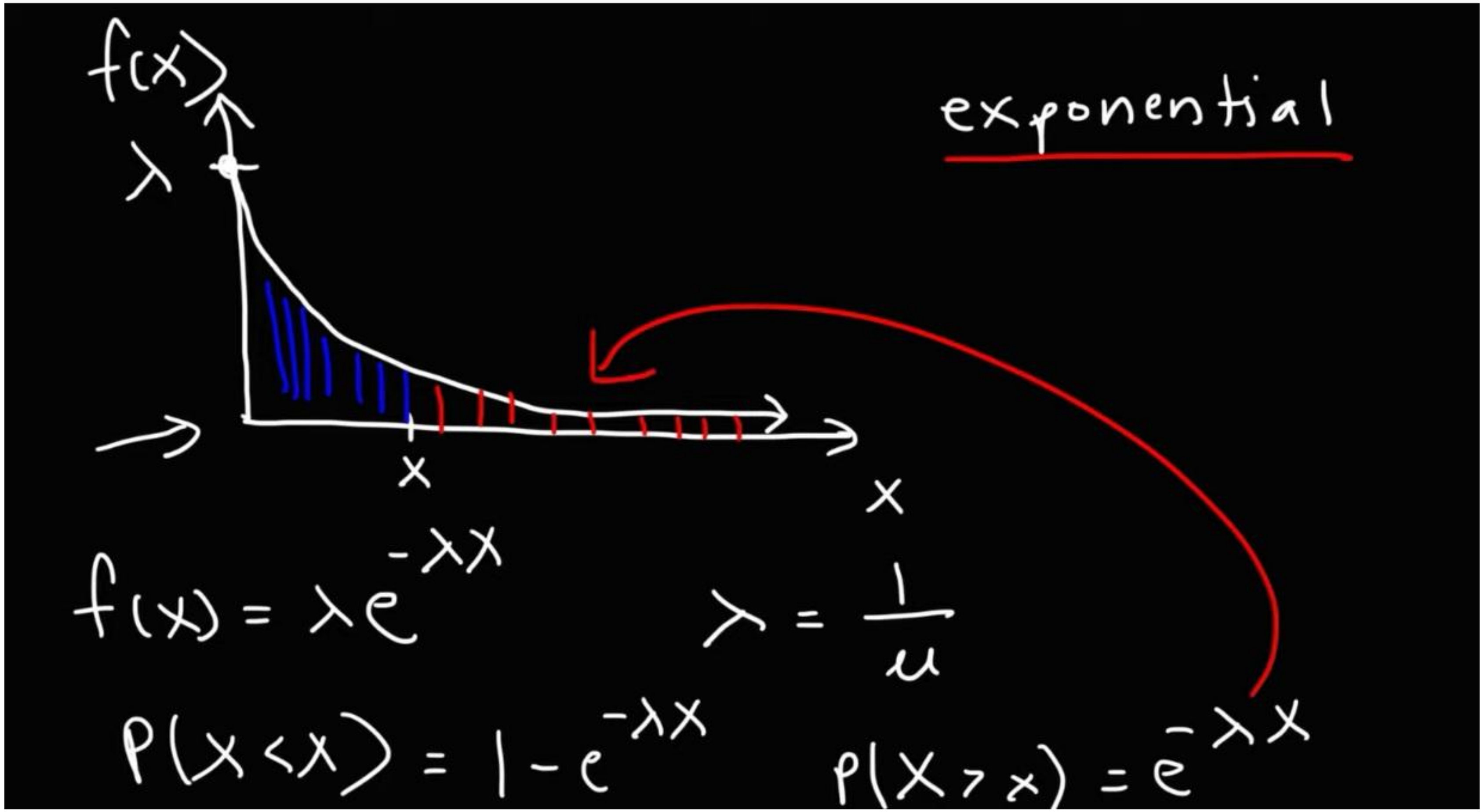
$$\mu = \frac{a+b}{2}$$

$$\sigma = \frac{b-a}{\sqrt{12}}$$



$f(x) = 1/4$

Exponential Distribution



$$f(x) = \lambda e^{-\lambda x}$$

$$\lambda = \frac{1}{u}$$

$$P(X<x) = 1 - e^{-\lambda x}$$

$$P(X>x) = e^{-\lambda x}$$

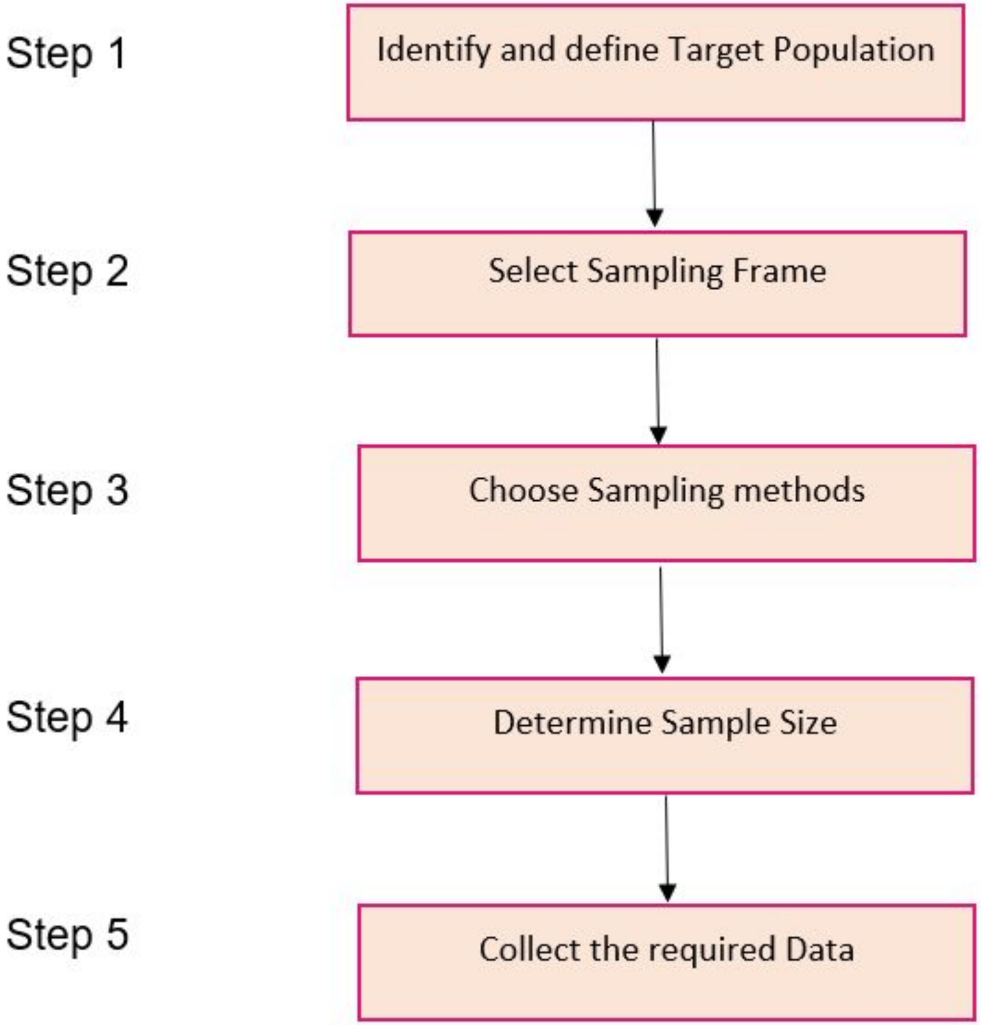
Sampling and its various techniques

- **What is Sampling?**
- Sampling is a method that allows us to get information about the population based on the statistics from a subset of the population (sample), without having to investigate every individual.
- We want to find the average height of all adult males in Delhi. The population of Delhi is around 3 crore and males would be roughly around 1.5 crores (these are general assumptions for this example so don't take them at face value!). As you can imagine, it is nearly impossible to find the average height of all males in Delhi.

Why do we need Sampling?

- I’m sure you have a solid intuition at this point regarding the question.
- Sampling is done to draw conclusions about populations from samples, and it enables us to determine a population’s characteristics by directly observing only a portion (or sample) of the population.
- Selecting a sample requires less time than selecting every item in a population
- Sample selection is a cost-efficient method
- Analysis of the sample is less cumbersome and more practical than an analysis of the entire population

Steps involved in Sampling



Step 1 Identify and define Target Population

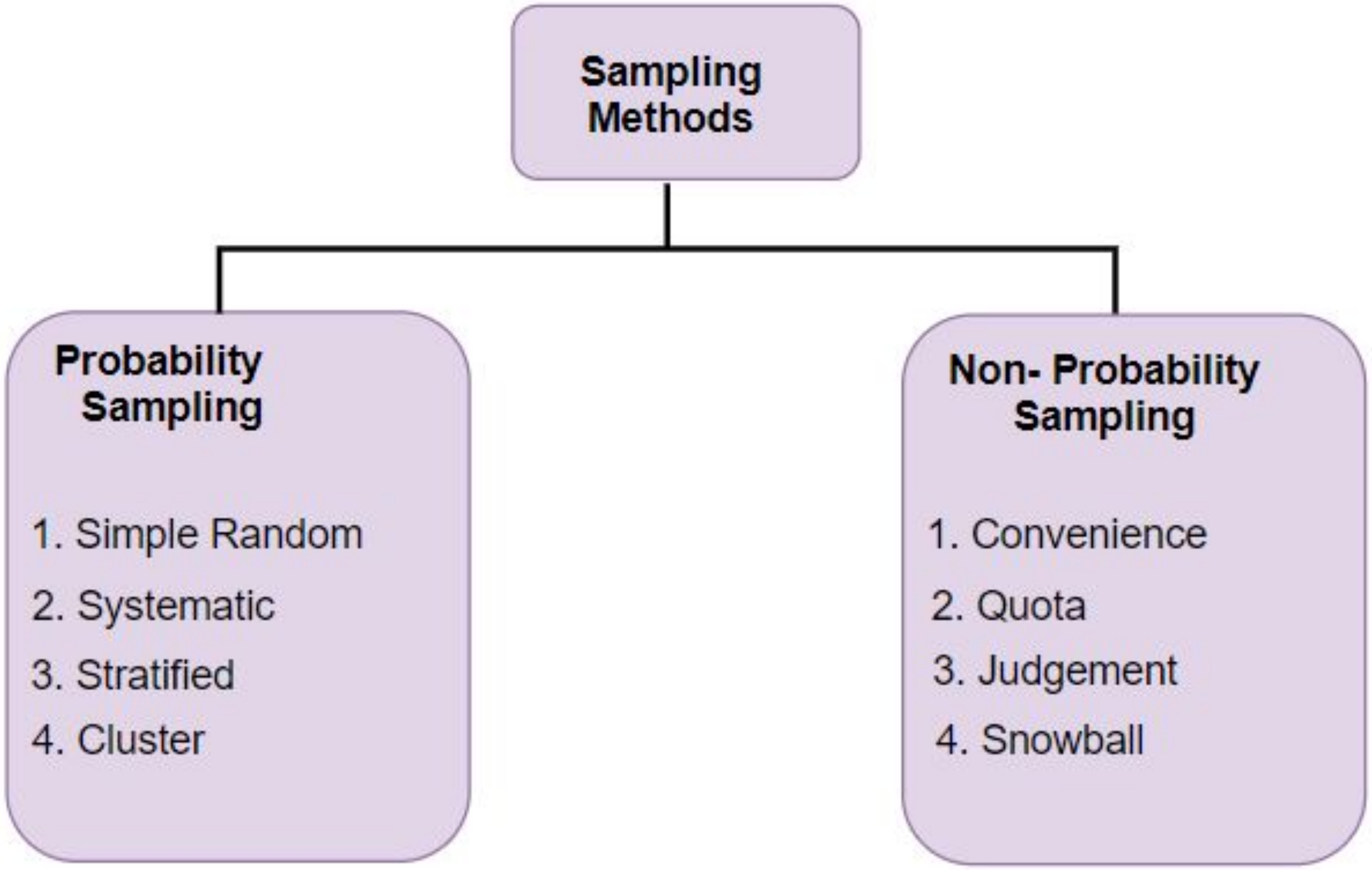
Step 2 Select Sampling Frame

Step 3 Choose Sampling methods

Step 4 Determine Sample Size

Step 5 Collect the required Data

Different Types of Sampling Techniques



Sampling Methods

Probability Sampling

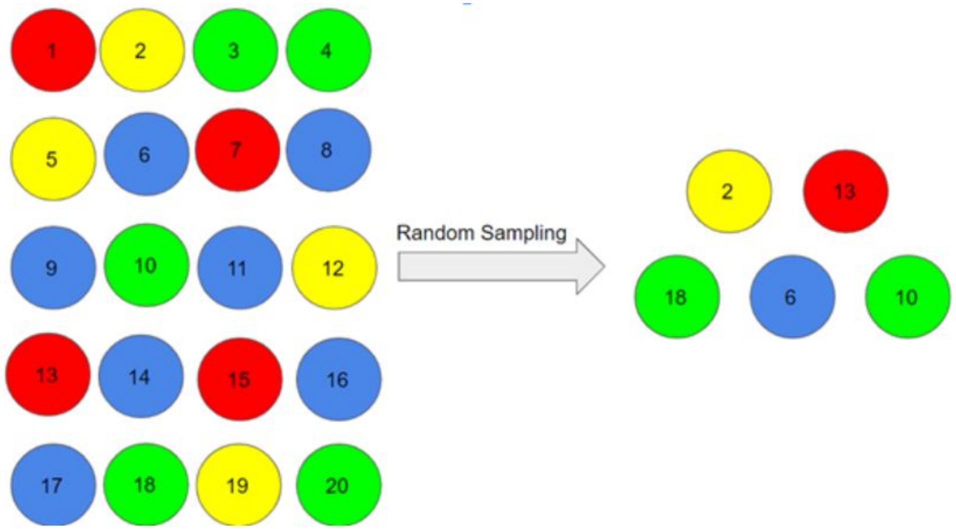
1. Simple Random
2. Systematic
3. Stratified
4. Cluster

Non- Probability Sampling

1. Convenience
2. Quota
3. Judgement
4. Snowball

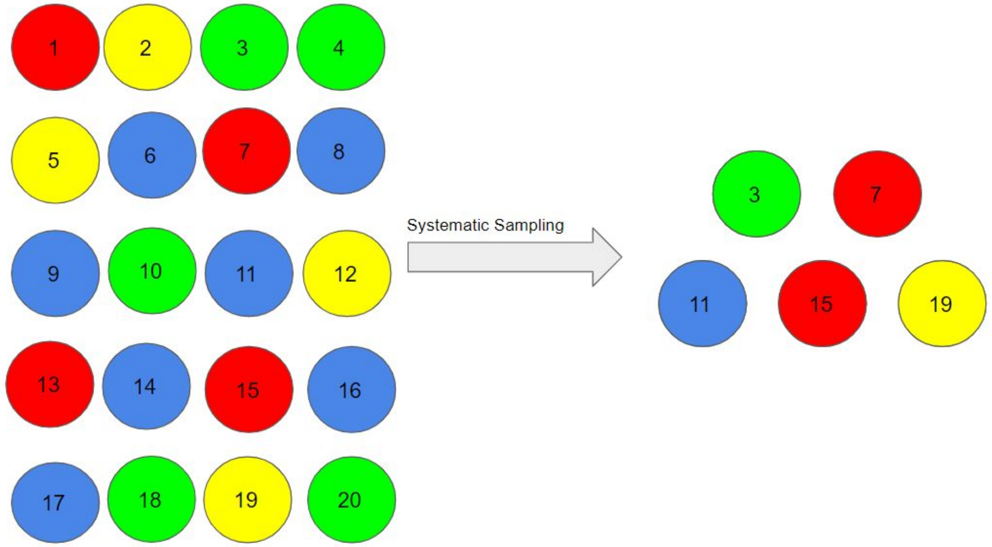
Simple Random Sampling

- This is a type of sampling technique you must have come across at some point. Here, every individual is chosen entirely by chance and each member of the population has an equal chance of being selected.
- **Simple random sampling reduces selection bias**



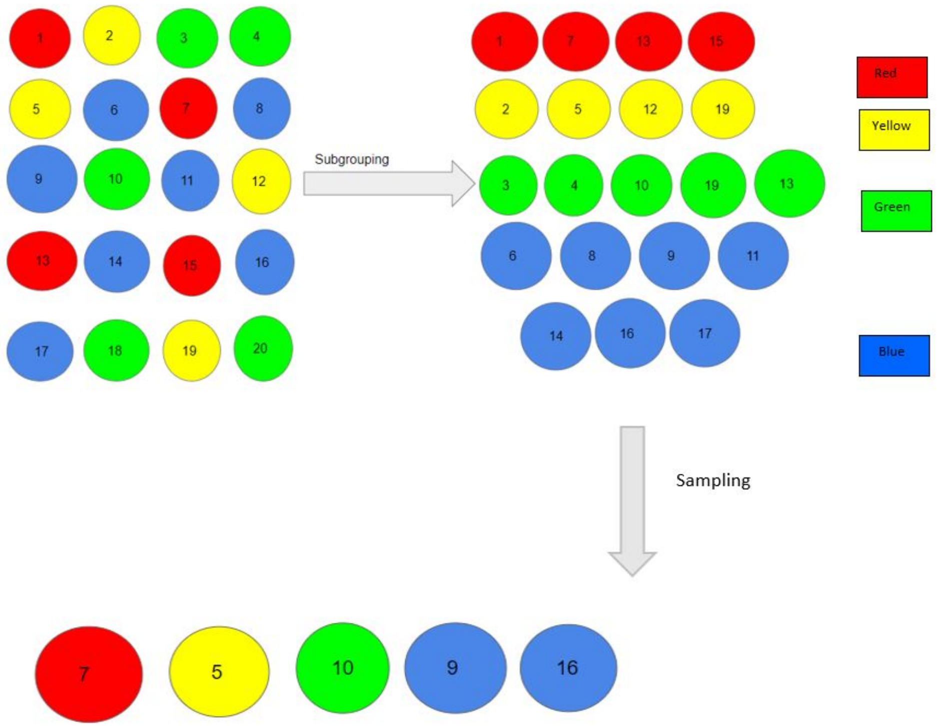
Systematic Sampling

- In this type of sampling, the first individual is selected randomly and others are selected using a fixed 'sampling interval'. Let's take a simple example to understand this.
- Say our population size is x and we have to select a sample size of n. Then, the next individual that we will select would be x/nth intervals away from the first individual. We can select the rest in the same way



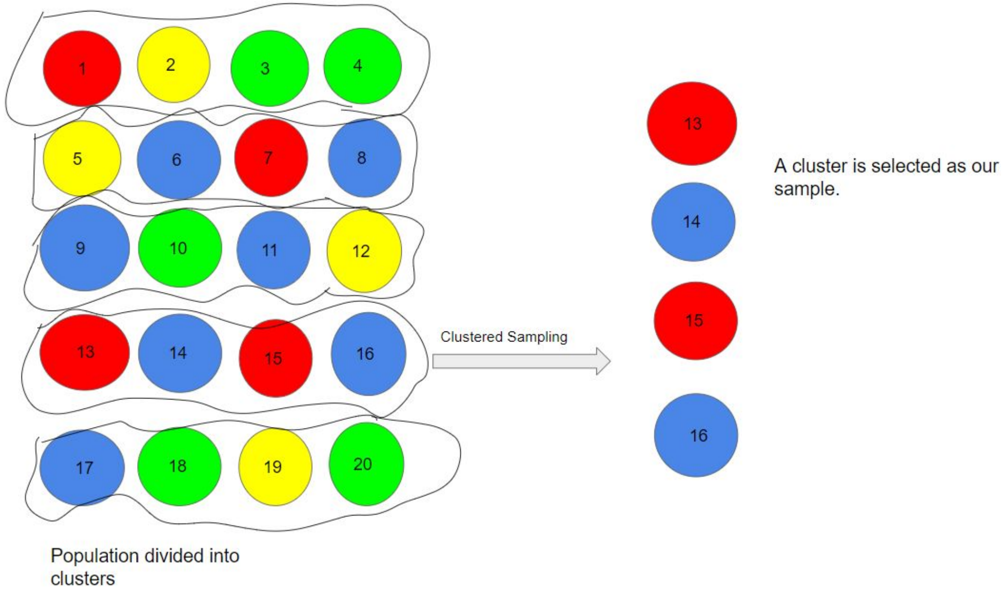
Stratified Sampling

- In this type of sampling, we divide the population into subgroups (called strata) based on different traits like gender, category, etc. And then we select the sample(s) from these subgroups:



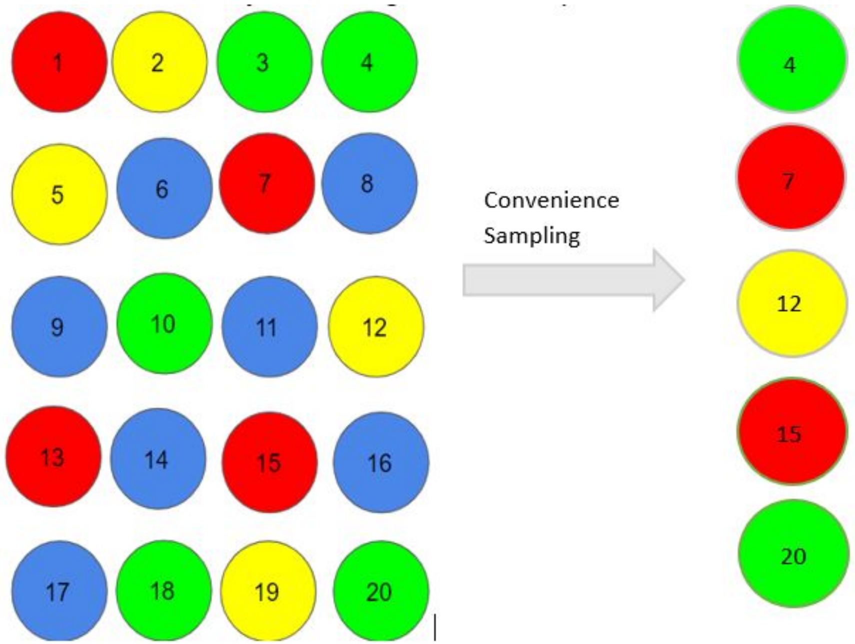
Cluster Sampling

- In a clustered sample, we use the subgroups of the population as the sampling unit rather than individuals. The population is divided into subgroups, known as clusters, and a whole cluster is randomly selected to be included in the study:


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Population divided into clusters
Clustered Sampling
13
14
15
16
A cluster is selected as our sample.

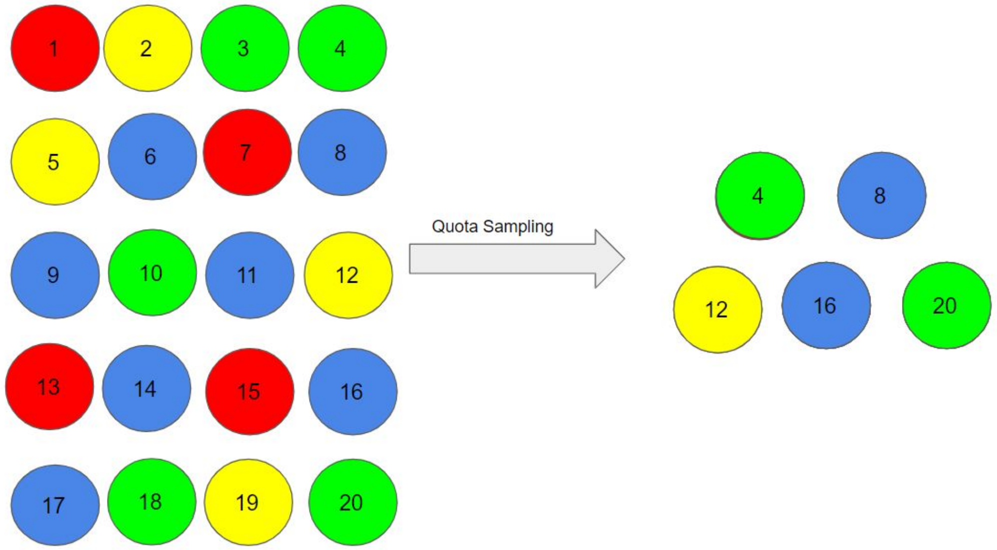
Types of Non-Probability Sampling

- **Convenience Sampling**
- This is perhaps the easiest method of sampling because individuals are selected based on their availability and willingness to take part.
- Here, let's say individuals numbered 4, 7, 12, 15 and 20 want to be part of our sample, and hence, we will include them in the sample.



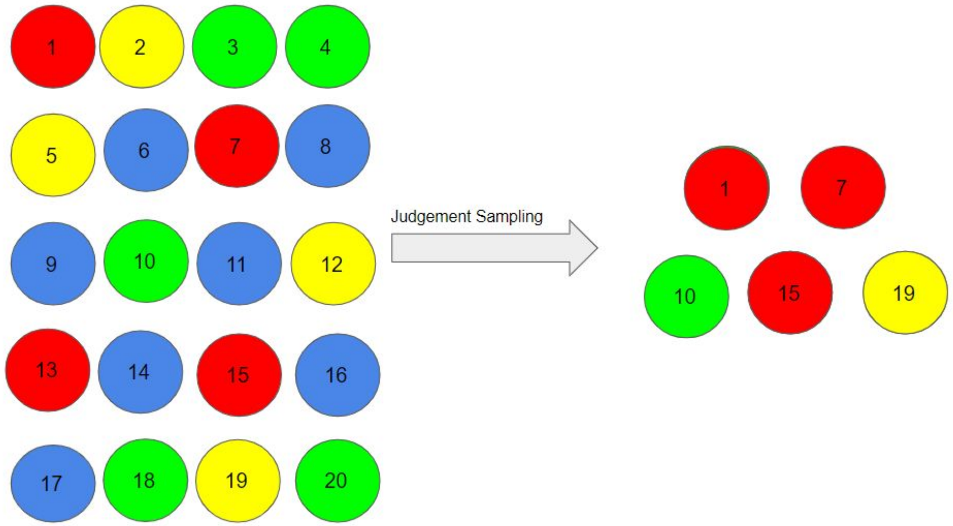
Quota Sampling

- In this type of sampling, we choose items based on predetermined characteristics of the population. Consider that we have to select individuals having a number in multiples of four for our sample:



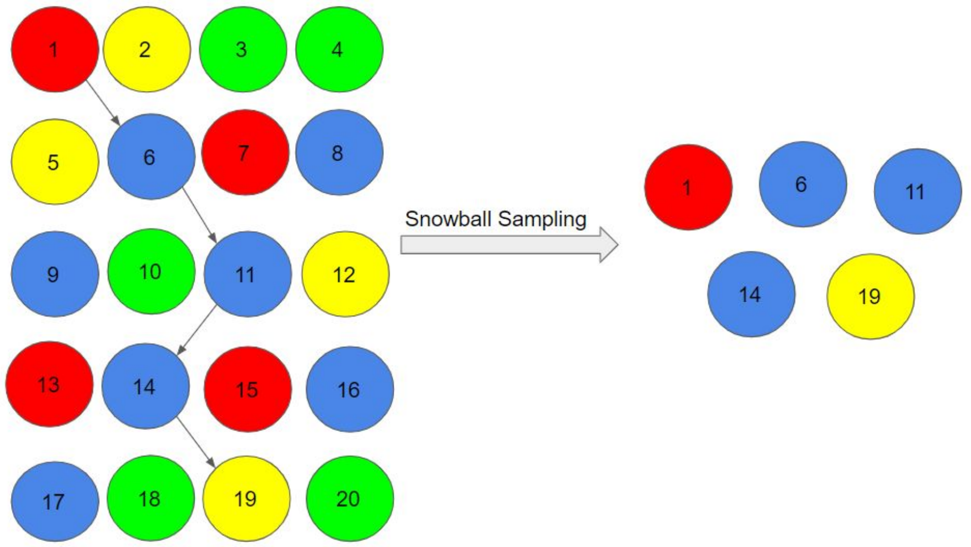
Judgment Sampling

- It is also known as selective sampling. It depends on the judgment of the experts when choosing whom to ask to participate.



Snowball Sampling

- **Existing people are asked to nominate further people known to them so that the sample increases in size like a rolling snowball.** This method of sampling is effective when a sampling frame is difficult to identify.



Types of Estimates

1. Point Estimate: If the estimate of the population parameter is described by a single numeric value, then such single numerical value is called point estimation (like x=5.8)

2. Interval Estimate: If the estimate of the population parameter is described by an interval, constituted by two limits is called interval estimation (like x lies between 5.75 to 5.85)

Standard normal distribution

•Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.

- The normal distribution is the proper term for a probability bell curve.
- In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.
- Normal distributions are symmetrical, but not all symmetrical distributions are normal.
- Many naturally occurring phenomena tend to approximate the normal distribution.
- In finance, most pricing distributions are not, however, perfectly normal.

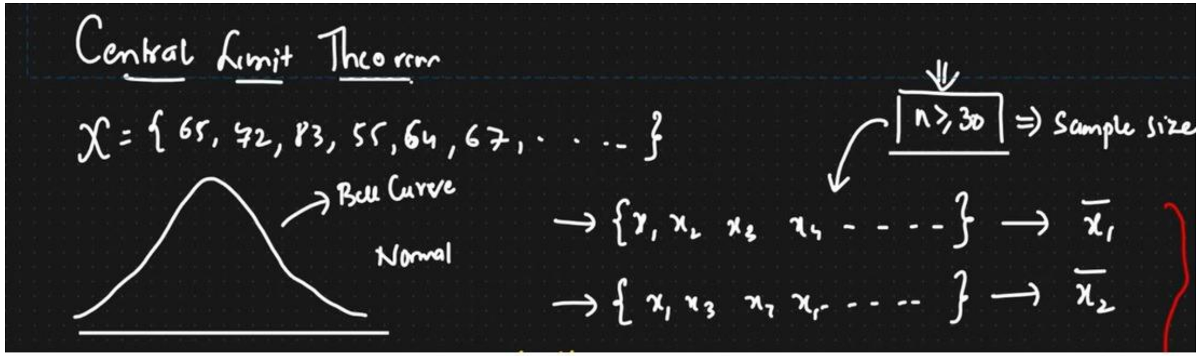
Central limit theorem

- Central Limit Theorem is a statistical concept that states that the sample means the distribution of a random variable will approach a normal or Gaussian distribution if the sample sizes are large irrespective of the shape of the original population distribution.
-

- Normal Distribution means that a set of numbers should look like a bell curve when mapped on a graph.
- Sample Mean (μ) is the average of the subsets of a large dataset or population where subsets are chosen randomly.

- The rule of the thumb or something which is considered safe is that the sample size should be greater than 30 (n>=30).
- In simpler terms, CLT states that for 30 or more data points in your sample, the mean of that sample will be a part of a bell-shaped curve closer to the Centre with few averages lying on either extreme and will represent the mean of the entire population. The same applies to standard deviation 'σ' as well. The average standard deviation of all the samples is representative of the standard deviation of the entire population.

Continued ...



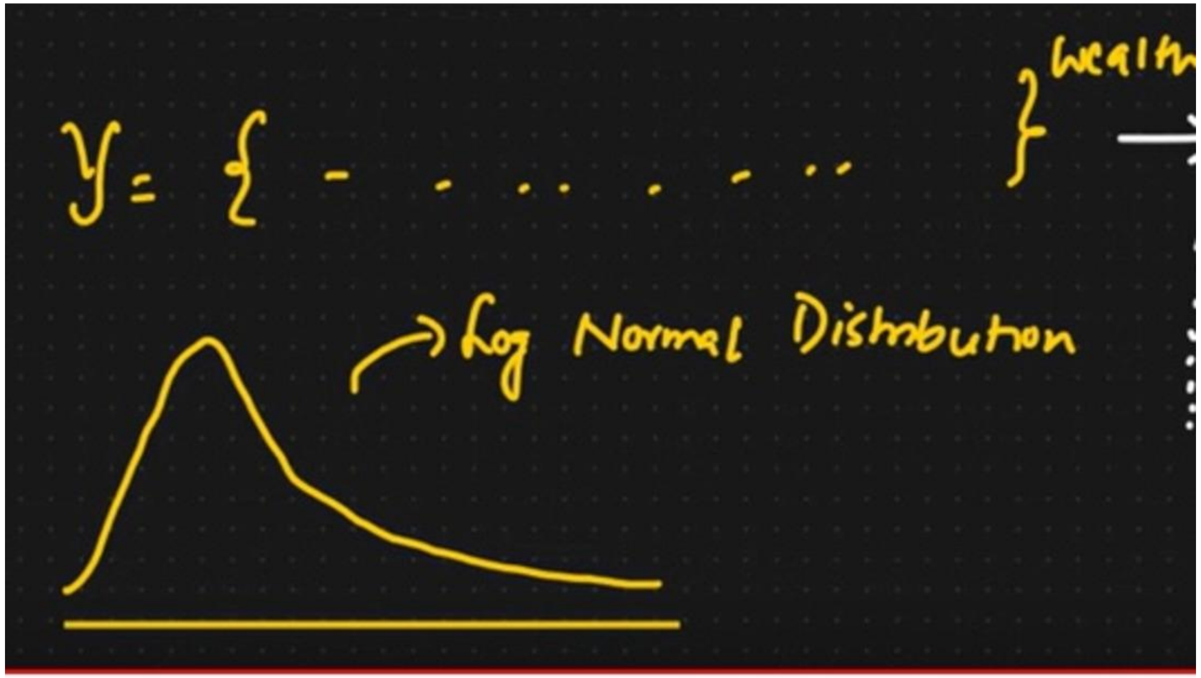
Central Limit Theorem

$X = \{65, 72, 83, 55, 64, 67, \cdots\}$

$\boxed{n \geq 30} \Rightarrow$ Sample size

$\rightarrow \{x_1, x_2\ x_3\ x_4 \cdots\} \rightarrow \bar{x}_1$

$\rightarrow \{x_1, x_3\ x_7\ x_5 \cdots\} \rightarrow \bar{x}_2$



$Y = \{ - \cdots \}$ wealth

Hypothesis testing through one sample test.

- The **null hypothesis** is a kind of hypothesis which explains the population parameter whose purpose is to test the validity of the given experimental data. This hypothesis is either rejected or not rejected based on the viability of the given population or sample. In other words, the null hypothesis is a hypothesis in which the sample observations results from the chance. It is said to be a statement in which the surveyors wants to examine the data. It is denoted by H_0.

- The alternative hypothesis is a statement used in statistical inference experiment. It is contradictory to the null hypothesis and **denoted by H_a or H_1**. We can also say that it is simply an alternative to the null. In hypothesis testing, an alternative theory is a statement which a researcher is testing. This statement is true from the researcher's point of view and ultimately proves to reject the null to replace it with an alternative assumption. In this hypothesis, the difference between two or more variables is predicted by the researchers, such that the pattern of data observed in the test is not due to chance.

Continue …

Null Hypothesis	Alternative Hypothesis
It denotes there is no relationship between two measured phenomena.	It's a hypothesis that a random cause may influence the observed data or sample.
It is represented by H_0	It is represented by H_a or H_1
Example: Federer will win at least 25 grand slams in tennis	Example: Federer will win less than 25 grand slams in tennis

Thank you