



Sanjivani Rural Education Society's
Sanjivani College of Engineering, Kopargaon-423 603
Department of Information Technology

Foundation of Data Science (IT305B)

Prepared by

Mr. Umesh B. Sangule

Assistant Professor

Department of Information Technology



Unit-IV

MATHEMATICAL DISTRIBUTIONS

Course Objectives : *To apply the Mathematical distributions on data for data understanding.*

Course Outcome(CO4) : *Apply various mathematical distributions for data understanding.*



Content

- Mathematical Distributions
- Sampling and Estimation



Probability Distributions

- **Random Variable:** *A random variable is a variable whose value is unknown or a function that assigns values to each of an experiment's outcomes.*
- Random variables are often designated by letters and can be classified as discrete, which are variables that have specific values, or continuous, which are variables that can have any values within a continuous range.
- The use of random variables is most common in probability and statistics, where they are used to quantify outcomes of random occurrences.
- ***Risk analysts use random variables to estimate the probability of an adverse event occurring.***



Probability Distributions

- **Probability Distribution:** A probability distribution is a statistical function that describes all the possible values and likelihoods that a *random variable* can take within a given *range*.
- This range will be bounded between the minimum and maximum possible values, but precisely where the possible value is likely to be plotted on the probability distribution depends on a number of factors.
- These factors include the distribution's *mean (average), standard deviation, skewness, and kurtosis*



Probability Distributions

- **Probability Distribution**
- A probability distribution depicts the expected outcomes of possible values for a given data-generating process.
- Probability distributions come in many shapes with different characteristics, as defined by the mean, standard deviation, skewness, and kurtosis,
- Probability distributions are often used in risk management as well to evaluate the probability and amount of losses that an investment portfolio would incur based on a distribution of historical returns.



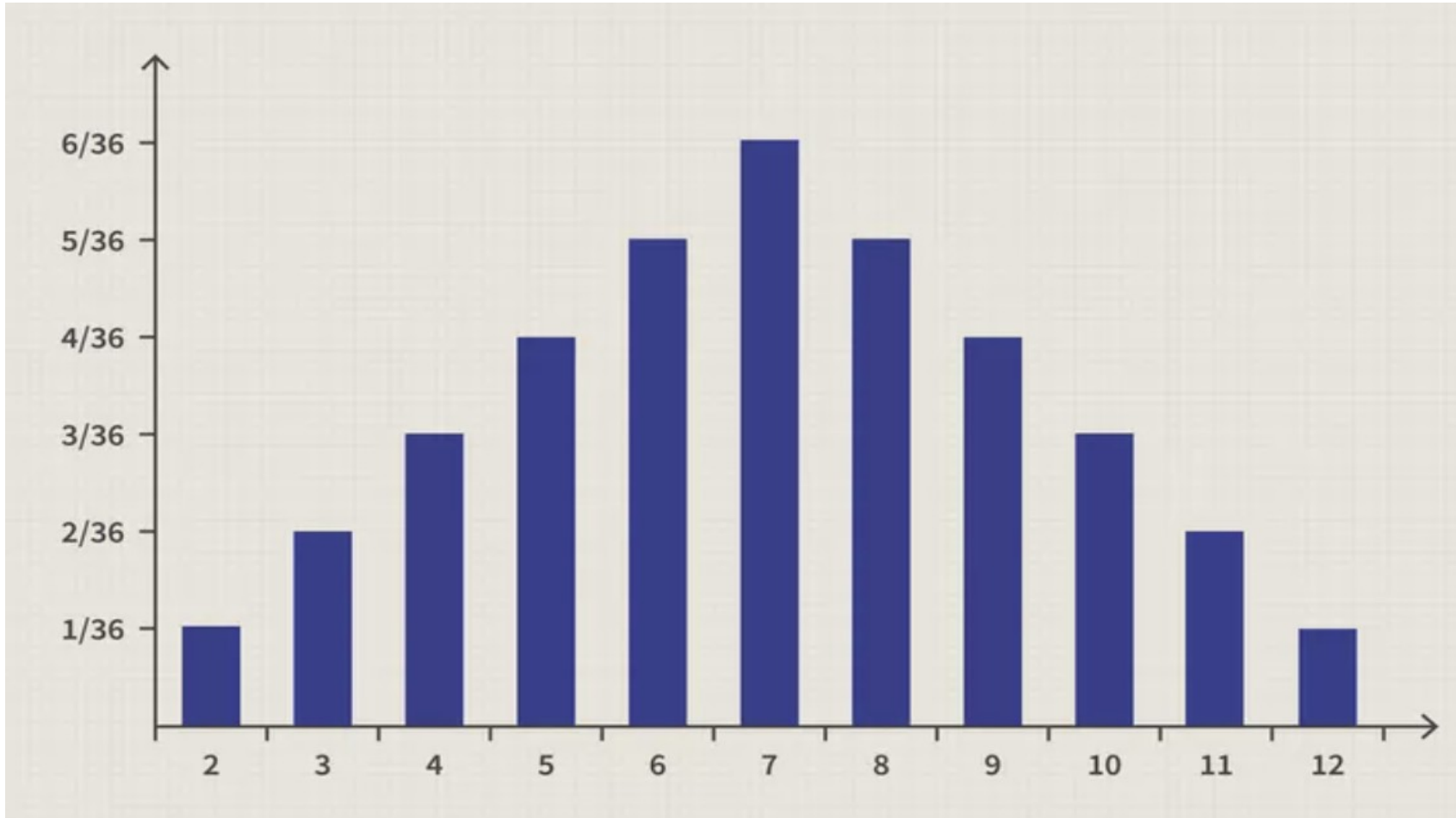
Probability Distributions

- **Example of a Probability Distribution:**
- As a simple example of a probability distribution, let us look at the number observed when rolling two standard six-sided dice.
- Each die has a $1/6$ probability of rolling any single number, one through six, but the sum of two dice will form the probability distribution.
- Seven is the most common outcome (1+6, 6+1, 5+2, 2+5, 3+4, 4+3), Two and twelve, on the other hand, are far less likely (1+1 and 6+6).



Probability Distributions

➤ *Example of a Probability Distribution:*





Probability Distributions

Types of Probability Distributions:

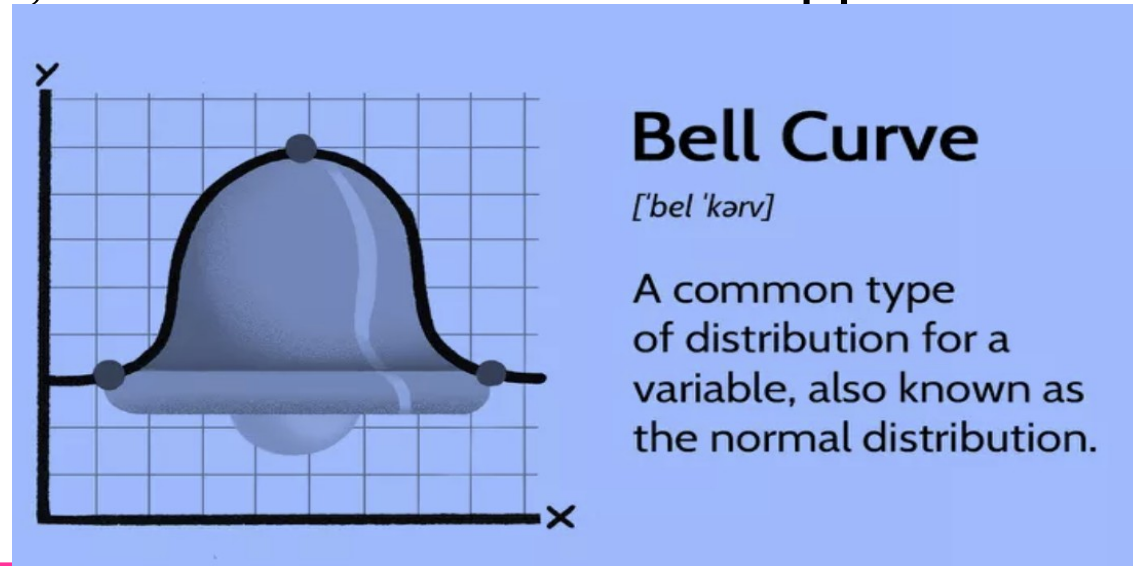
- There are many different classifications of probability distributions.
- Some of them include-
 - 1) Normal Distribution,
 - 2) Binomial Distribution,
 - 3) Poisson distribution,
 - 4) Exponential distribution.
- The different probability distributions serve different purposes and represent different data generation processes.



Probability Distributions

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.
- In graphical form, the normal distribution appears as a "bell curve".





Probability Distributions

Normal Distribution:

- 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*.
- Normal distributions are *symmetrical*, but not all symmetrical distributions are normal.
- Many naturally-occurring phenomena tend to approximate the normal distribution.



Probability Distributions

Normal Distribution:

- Regardless of its exact shape, a normal distribution bell curve is always *symmetrical* about the mean,
- A symmetrical distribution means that a vertical dividing line drawn through the maximum/mean value will produce two mirror images on either side of the line, in which half the population is less than the mean and half is greater.
- However, the reverse is not always true; that is, not all symmetrical distributions are normal. *In the bell curve, the peak is always in the middle, and the mean, mode and median are all the same.*



Probability Distributions

Importance of normal distribution:

- The normal distribution is one of the most important probability distributions for independent random variables for three main reasons.
- First, normal distribution describes the distribution of values for many natural phenomena in a wide range of areas, including biology, physical science, mathematics, finance and economics. It can also represent these random variables accurately.
- Second, the normal distribution is important because it can be used to approximate other types of probability distribution, such as binomial, Poisson distribution.



Probability Distributions

Importance of normal distribution:

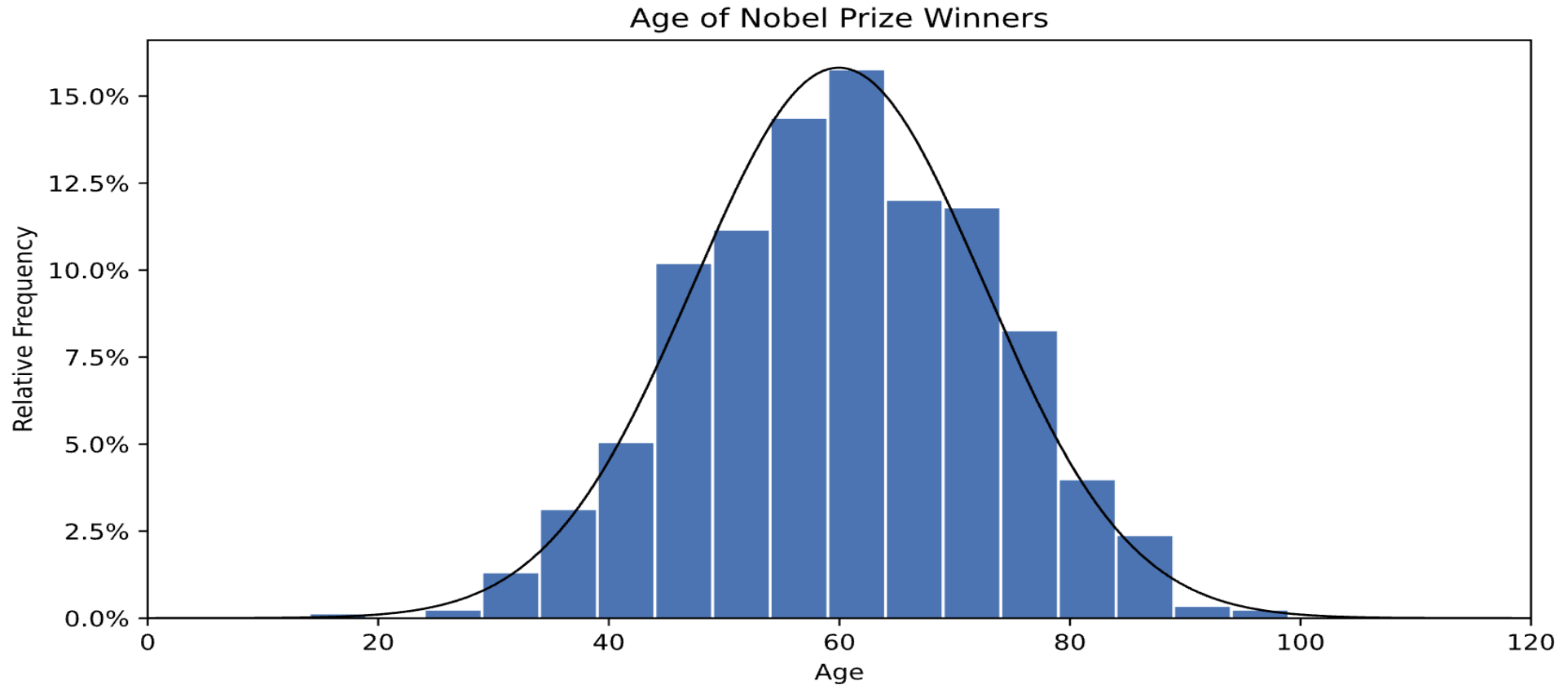
- Third, normal distribution is the key idea behind the central limit theorem, or CLT, which states that averages calculated from independent, identically distributed random variables have approximately normal distributions.
- This is true regardless of the type of distribution from which the variables are sampled, as long as it has finite variance.



Probability Distributions

Real Data Example of Normally Distributed Data

Here is a histogram of the age of Nobel Prize winners when they won the prize:





Probability Distributions


Binomial Distribution:

- Binomial distribution is a statistical distribution that summarizes the probability that a value will take one of two independent values under a given set of parameters or assumptions.
- The underlying assumptions of binomial distribution are that there is only one outcome for each trial, each trial has the same probability of success, and each trial is mutually exclusive or independent of one another.
- The “binomial” in binomial distribution means two terms the number of successes and the number of attempts. Each is useless without the other.



Probability Distributions

Binomial Distribution:



Binomial Distribution

[bī-'nō-mē-əl ,di-strə-'byü-shən]

The likelihood of observing a certain outcome when performing a series of tests for which there are only two possible outcomes, such as getting heads or tails in a coin toss.



Probability Distributions

Binomial Distribution:

- Binomial distribution is a common discrete distribution used in statistics, as opposed to a continuous distribution, such as normal distribution.
- This is because binomial distribution only counts two states, typically represented as 1 (for a success) or 0 (for a failure), given a number of trials in the data.
- Binomial distribution thus represents the probability for x successes in n trials, given a success probability p for each trial.



Probability Distributions

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 the next trial.
- The probability of success (tails, heads, fail or pass) is exactly the same from one trial to another.



Probability Distributions

Binomial Distribution:

- Binomial distribution is often used in social science statistics as a building block for models for dichotomous outcome variables, such as whether a Republican or Democrat will win an upcoming election, whether an individual will die within a specified period of time, etc.
- It also has applications in finance, banking, and insurance, among other industries.



Probability Distributions

Binomial Distribution:

- Binomial distribution is used to figure the probability of a pass or fail outcome in a survey, or experiment replicated numerous times.
- There are only two potential outcomes for this type of distribution. More broadly, distribution is an important part of analyzing data sets to estimate all the potential outcomes of the data and how frequently they occur.
- Forecasting and understanding the success or failure of outcomes is essential to business development.



Probability Distributions

Binomial Distribution:

Throwing a Die:

- Did we get a four ... ?
- ... or not?



We say the probability of a **four** is $1/6$ (one of the six faces is a four)

And the probability of **not four** is $5/6$ (five of the six faces are not a four)

Note that a die has 6 sides but here we look at only **two** cases: "**four: yes**" or "**four: no**"



Probability Distributions

Criteria of Binomial Distribution:

- Binomial distribution models the probability of occurrence of an event when specific criteria are met.

1. Fixed trials

- The process under investigation must have a fixed number of trials that cannot be altered in the course of the analysis. During the analysis, each trial must be performed in a uniform manner, although each trial may yield a different outcome.
- Example of a fixed trial may be coin flips, free throws, wheel spins, etc. If a coin is flipped 10 times, each flip of the coin is a trial.



Probability Distributions

Criteria of Binomial Distribution:

2. Independent trials

- In simple terms, the outcome of one trial should not affect the outcome of the subsequent trials.
- When using certain sampling methods, there is a possibility of having trials that are not completely independent of each other, and binomial distribution may only be used when the size of the population is large vis-a-vis the sample size.
- An example of independent trials may be tossing a coin or rolling a dice. When tossing a coin, the first event is independent of the subsequent events.



Probability Distributions

Criteria of Binomial Distribution:

3. Fixed probability of success

- In a binomial distribution, the probability of getting a success must remain the same for the trials we are investigating. For example, when tossing a coin, the probability of flipping a coin is $\frac{1}{2}$ or 0.5 for every trial we conduct, since there are only two possible outcomes.
- In some sampling techniques, such as sampling without replacement, the probability of success from each trial may vary from one trial to the other. For example, assume that there are 50 boys in a population of 1,000 students. The probability of picking a boy from that population is 0.05.



Probability Distributions

Criteria of Binomial Distribution:

3. Fixed probability of success

- In the next trial, there will be 49 boys out of 999 students. The probability of picking a boy in the next trial is 0.049. It shows that in subsequent trials, the probability from one trial to the next will vary slightly from the prior trial.
- In binomial probability, there are only two mutually exclusive outcomes, i.e., success or failure. While success is generally a positive term, it can be used to mean that the outcome of the trial agrees with what you have defined as a success, whether it is a positive or negative outcome.



Probability Distributions

Applications of Binomial Distribution:

- Binomial distribution is used to figure the probability of a pass or fail outcome in a survey, or experiment replicated numerous times.
- There are only two potential outcomes for this type of distribution. More broadly, distribution is an important part of analyzing data sets to estimate all the potential outcomes of the data and how frequently they occur.
- Forecasting and understanding the success or failure of outcomes is essential to business development.



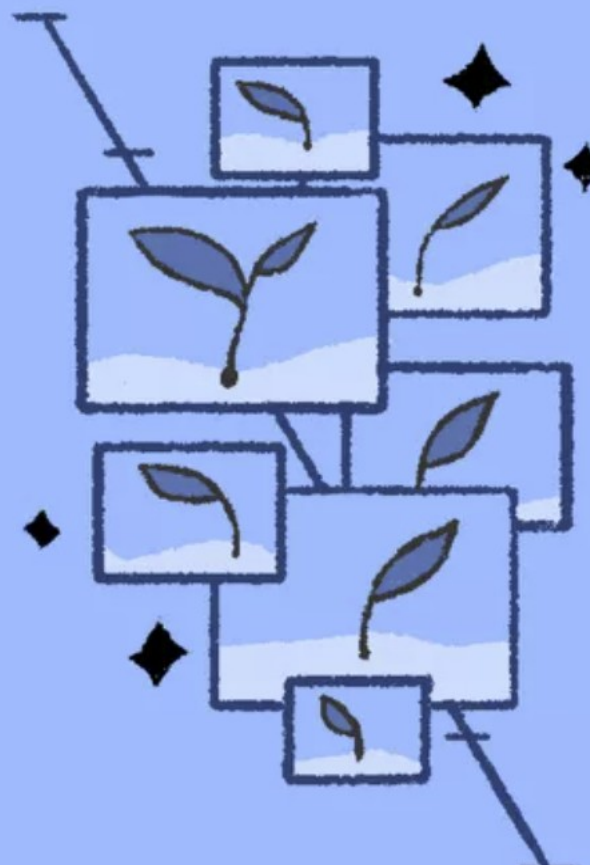
Probability Distributions

Poisson Distribution:

- In statistics, a Poisson distribution is a probability distribution that is used to show how many times an event is likely to occur over a specified period.
- In other words, it is a count distribution. Poisson distributions are often used to understand independent events that occur at a constant rate within a given interval of time.
- The Poisson distribution is a discrete function, meaning that the variable can only take specific values in a (potentially infinite) list.

Probability Distributions

Poisson Distribution:



Poisson Distribution

[pwä-'sõⁿ- ,di-strə-'byü-shən]

A probability distribution that is used to show how many times an event is likely to occur over a specified period.



Probability Distributions

Characteristics of Poisson Distribution:

- The Poisson distribution possesses the following characteristics:
- 1) ***Continuous random variable:*** The Poisson distribution deals with continuous variables, representing events occurring over a fixed interval.
 - 2) ***Measures events over a fixed interval:*** It models the number of events that occur within a specific interval of time or space.



Probability Distributions

Poisson Distribution:

- Put differently, the variable cannot take all values in any continuous range. For the Poisson distribution, the variable can only take whole number values (0, 1, 2, 3, etc.), with no fractions or decimals.
- Poisson Distribution can be used to estimate how many times an event is likely to occur within "X" periods of time,
- Many economic and financial data appear as count variables, such as how many times a person becomes unemployed in a given year, thus lending themselves to analysis with a Poisson distribution.



Probability Distributions

Poisson Distribution:

- If the mean is very large, then the Poisson distribution is approximately a normal distribution.
- The Poisson distribution is best applied to statistical analysis when the variable in question is a count variable. For instance, how many times X occurs based on one or more explanatory variables.
- For instance, to estimate how many defective products will come off an assembly line given different inputs.



Probability Distributions

Poisson Distribution:

- In order for the Poisson distribution to be accurate, all events are independent of each other, the rate of events through time is constant, and events cannot occur simultaneously.
- Moreover, the mean and the variance will be equal to one another
- Poisson distribution can be obtained as an approximation of a binomial distribution when the number of trials n of the latter distribution is large



Probability Distributions

Poisson Distribution:

Binomial Distribution	Poisson Distribution
Applicable to experiments with fixed number of trials	Applicable to events occurring within fixed interval
Requires a known number of trials	Does not require a specific number of events
Assumes independent trials	Assumes events occur randomly and independently
Deals with discrete random variables	Deals with continuous random variables
Probability of success remains constant for each trial	No explicit consideration of probability of success



Probability Distributions

Poisson Distribution:

- The Binomial and Poisson distributions are both essential tools in probability theory and statistics.
- While the Binomial distribution deals with experiments involving a fixed number of independent trials, the Poisson distribution focuses on events occurring over a fixed interval.
- Understanding their differences and knowing when to apply each distribution is crucial for accurate data analysis and modelling.



Probability Distributions

Applications 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.
- 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.



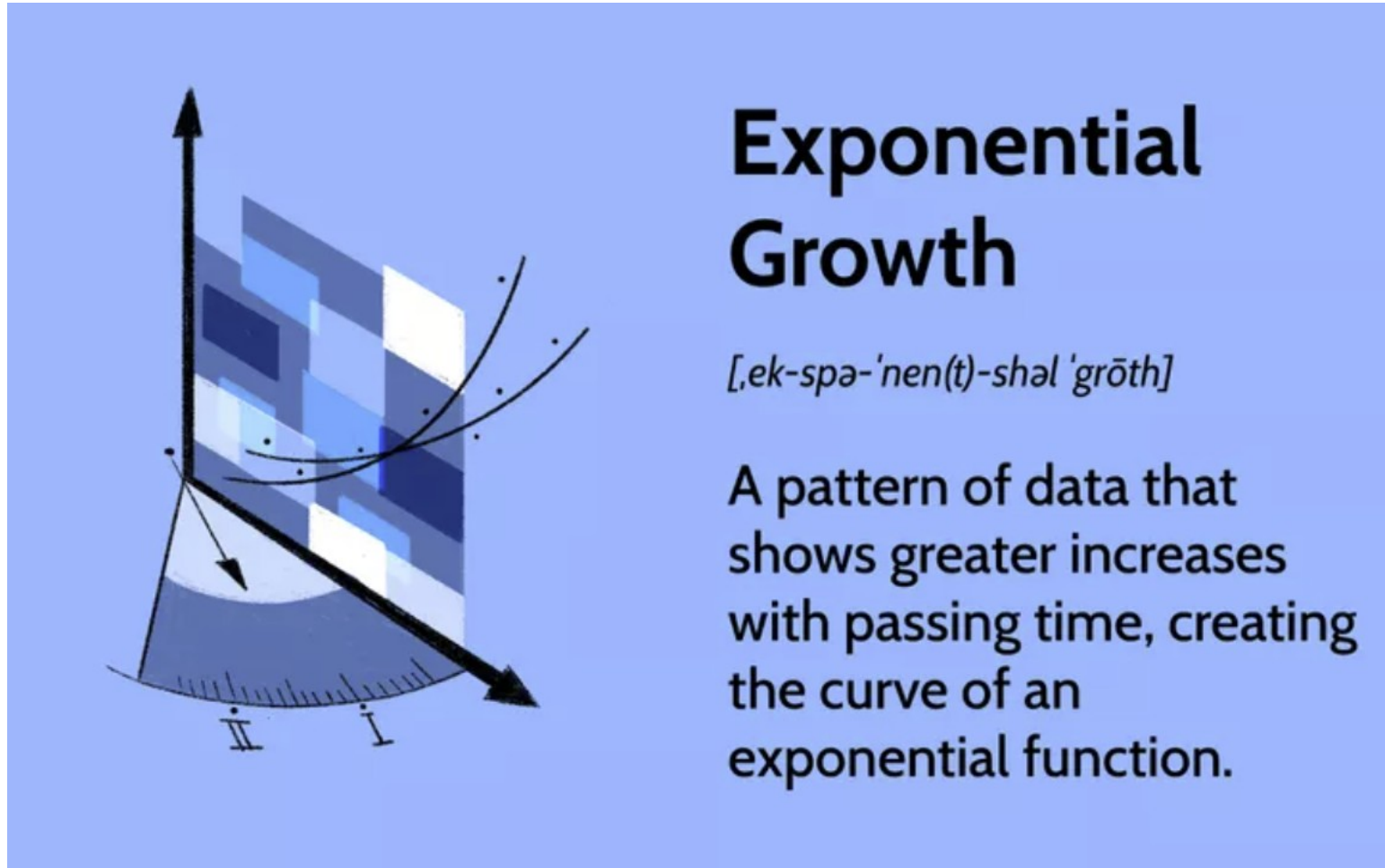
Probability Distributions

Exponential Distribution:

- The exponential distribution is often concerned with the amount of time until some specific event occurs.
- The exponential distribution is a continuous distribution that is commonly used to measure the expected time for an event to occur.
- Exponential growth (which is multiplicative) can be contrasted with linear growth (which is additive) and with geometric growth (which is raised to a power).

Probability Distributions

Exponential Distribution:





Probability Distributions

Exponential Distribution:

- The exponential distribution (also called the negative exponential distribution) is a probability distribution that describes time between events in a Poisson process.
- There is a strong relationship between the Poisson distribution and the Exponential distribution.
- For example, let's say a Poisson distribution models the number of births in a given time period. The time in between each birth can be modeled with an exponential distribution.



Probability Distributions

Exponential Distribution:

- *What is a Poisson Process?*
- The Poisson process gives you a way to find probabilities for random points in time for a process. A “process” could be almost anything:
 - 1) Accidents at an interchange.
 - 2) File requests on a server.
 - 3) Customers arriving at a store.
 - 4) Battery failure and replacement.
- The Poisson process can tell you when one of these random points in time will likely happen



Probability Distributions

Applications of Exponential Distribution:

- The exponential distribution is mostly used for testing product reliability.

1. Predict the time when an Earthquake might occur:

- The exponential distribution is prominently used by seismologists and earth scientists to predict the approximate time when an earthquake is likely to occur in a particular locality.
- For this purpose, the history of the earthquakes and other natural calamities occurring in a particular locality is recorded and monitored.



Probability Distributions

Applications of Exponential Distribution:

2. Life Span of Electronic Gadgets

- Exponential distribution finds its prime application in calculating the reliability of electronic gadgets such as a laptop, battery, processor, mobile phone, etc.
- It helps the engineers and manufacturers to know an approximate time after which the product will get ruptured.
- The engineers use this data to improve the quality of their products by replacing the low-quality components with those having comparatively high quality.



Probability Distributions

Applications of Exponential Distribution:

3. Average Time a Call Centre Employee Spends With the Customer

- If the average time that a call centre executive takes to complete his/her given task to communicate with a customer is known to be twenty minutes, then the probability that the executive will handle eight customers per hour can be estimated with the help of exponential distribution.
- This helps the managers draft an appropriate schedule to tackle all the customers approaching the firm with their concerns. This helps improve the customer satisfaction index.



Probability Distributions

Applications of Exponential Distribution:

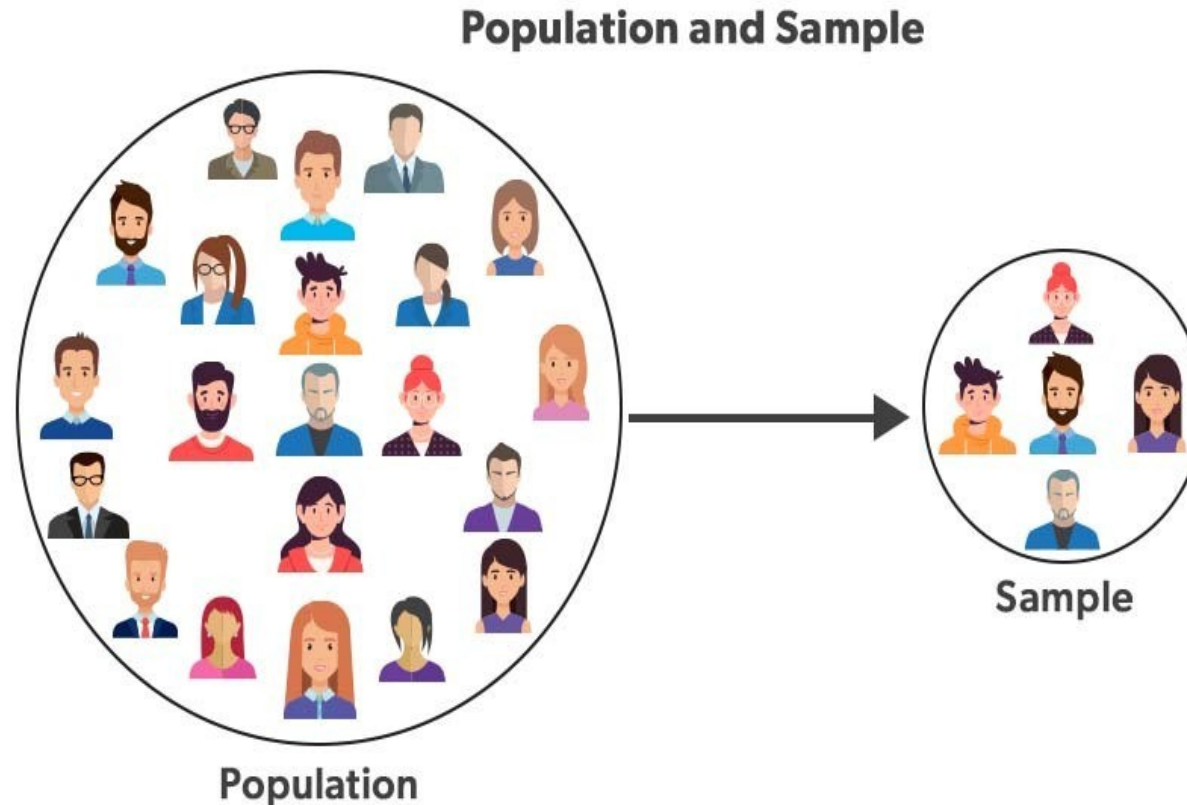
4. Cars Passing per Minute

- If it is known that on average approximately ten cars cross a particular highway every minute, then the probability that seven cars will pass the same highway the following minute can be easily estimated with the help of exponential distribution.
- It also helps to calculate the time duration between the passing of two consecutive cars, thereby helping the traffic in charge to reduce the traffic problem and to avoid collisions.

Sampling and Estimation

Sampling:

- A sample is a subset of a population.





Sampling and Estimation

Sampling:

- We can study a sample to infer conclusions about the population itself.
- For example, if all the stocks trading in the US are considered a population, then indices such as the S&P 500 are samples.
- We can look at the performance of the S&P 500 and draw conclusions about how all stocks in the US are performing. This process is known as sampling and estimation.



Sampling and Estimation

Sampling Methods:

- There are various methods for obtaining information on a population through samples.
- The information we obtain usually concerns a parameter, a quantity used to describe a population. To estimate a parameter, we use sample statistics. A statistic is a quantity used to describe a sample.
- There are two reasons why sampling is used:
 - 1) **Time saving**: In many cases it will be very time consuming to examine every member of the population.
 - 2) **Monetary saving**: In some cases, examining every member of the population becomes economically inefficient.



Sampling and Estimation

There are two types of sampling methods:

- *Probability sampling*: Every member of the population has an equal chance of being selected. Therefore, the sample created is representative of the population.
 - *Non-probability sampling*: Every member of the population may not have an equal chance of being selected. This is because sampling depends on factors such as the sampler's judgement or the convenience to access data. Therefore, the sample created may not be representative of the population.
- All else equal, the probability sampling method is more accurate and reliable as compared to the non- probability sampling method.



Sampling and Estimation

Probability Sampling:

Based on how each case in sample is selected forms the different sampling methods:


- 1) Simple Random Sampling
- 2) Stratified Random Sampling
- 3) Cluster Sampling

1) Simple Random Sampling:

- Simple random sampling is the process of selecting a sample from a larger population in such a way that each member of the population has the same probability of being included in the sample.

Sampling and Estimation

1) Simple Random Sampling:



Simple Random Sample

[sim-pəl 'ran-dəm 'sam-pəl]

A subset of a statistical population in which each member of the subset has an equal probability of being chosen and is meant to be an unbiased representation of a group.



Sampling and Estimation

Probability Sampling:

1) Simple Random Sampling:

- Random sampling is usually carried out without replacement, that is, an observation which is selected in the sample is removed from the population for subsequent selection.
- However, random samples can also be created with replacement, that is, an observation which is selected for inclusion in the sample can again be considered since it is replaced (not removed) in the population.
- ***Random sampling is ideal when the population is homogeneous. In random sampling, every subject in the population has equal probability of selection in the sample.***



Sampling and Estimation

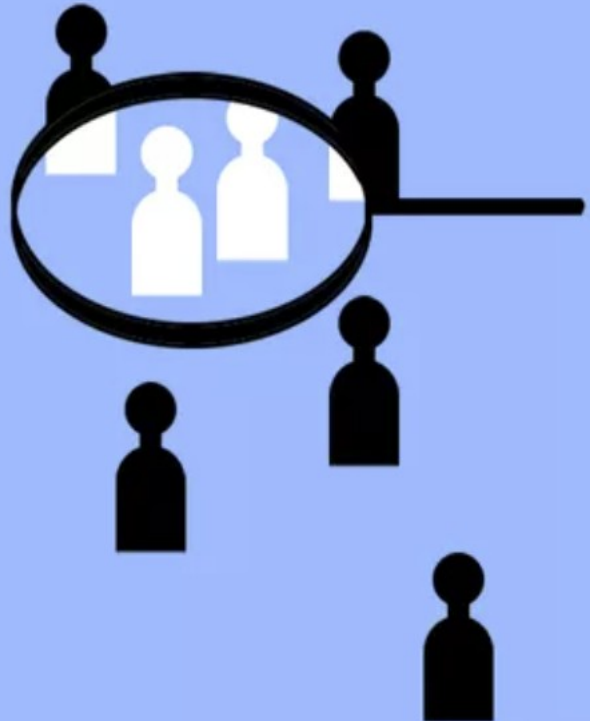
Probability Sampling:

2) Stratified Random Sampling:

- The population can be divided into mutually exclusive groups using some factor (for example, age, gender, marital status, income, geographical regions, etc.). The groups, thus, formed are called ***Stratum***.
- It is important that the groups are mutually exclusive and exhaustive of the population.
- Random sampling method can be used within each stratum samples in each group. Size of the sample in each strata should be proportional to the proportion of the strata in the population.

Sampling and Estimation

2) Stratified Random Sampling:



Stratified Random Sampling

[ˈstra-tə-,fɪd ˈran-dəm ˈsam-plɪŋ]

A method of sampling that involves the division of a population into smaller sub-groups known as strata.



Sampling and Estimation

Probability Sampling:

2) Stratified Random Sampling:

- Samples from each stratum can be combined to create the final sample.
- Stratified sampling is necessary when the population is heterogeneous and creating homogeneous stratum before sampling is recommended for precise estimation of population parameters.



Sampling and Estimation

Probability Sampling:

3) Cluster Sampling:

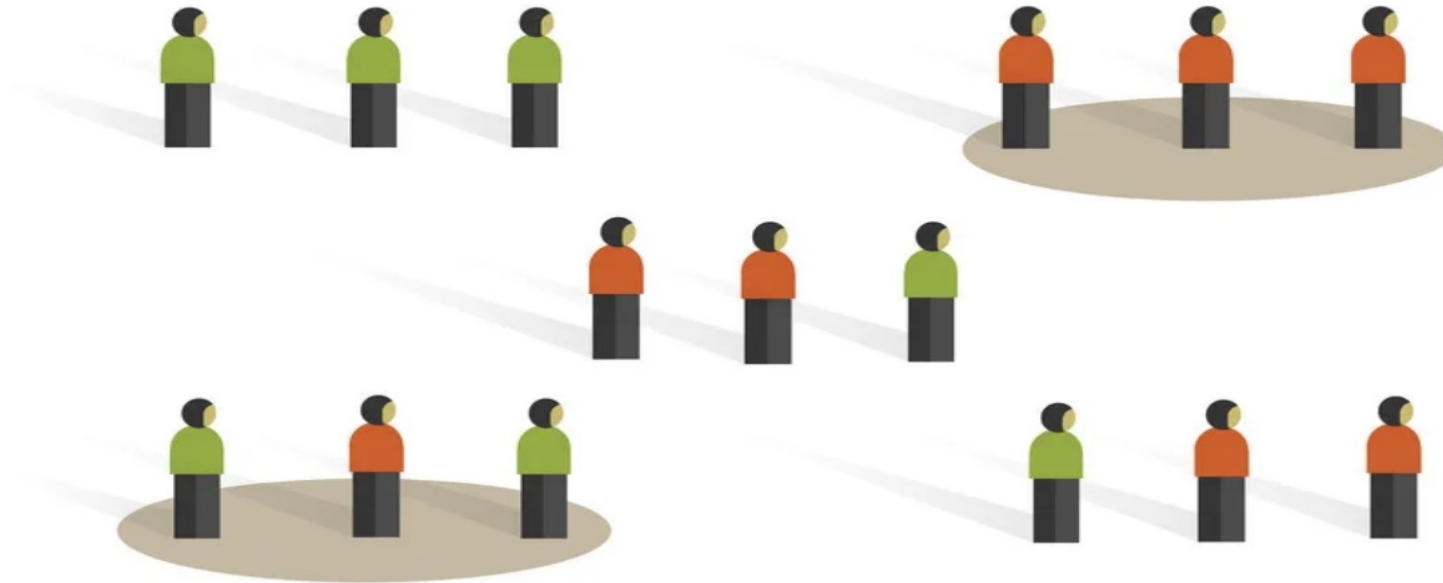
- Cluster sampling is similar to stratified random sampling as it also requires the population to be divided into subpopulation groups, called ***clusters***.
- Each cluster is essentially a mini-representation of the entire population. Then some random clusters are chosen as a whole for sampling.
- Clusters are generally based on natural groups separating the population. For example, you might be able to divide your data into natural groupings like city blocks, voting districts, or school districts.



Sampling and Estimation

3) Cluster Sampling:

Cluster sampling





Sampling and Estimation

Probability Sampling:

3) Cluster Sampling:

- The main difference between cluster sampling and stratified random sampling is that in cluster sampling, the whole cluster is selected; and not all clusters are included in the sample.
- In stratified random sampling, however, only a few members from each stratum are selected; but all strata are included in the sample.
- Cluster sampling is less accurate because the chosen sample may be less representative of the entire population.



Sampling and Estimation

Non-Probability Sampling:

The two major types of non-probability sampling methods are:

1) Convenience sampling:

- In this method, the researcher selects members from a population based on how easy it is to access the member i.e., data is collected from a conveniently available pool of respondents.
- The disadvantage of this method is that the sample selected may not be representative of the entire population.
- The advantage is that data can be collected quickly and at a low cost.
Hence this method is particularly suitable for small-scale pilot studies.



Sampling and Estimation

Non-Probability Sampling:

2) Judgmental sampling:

- In this method, the researcher uses his knowledge and professional judgment to selectively handpick members from the population.
- The disadvantage of this method is that the sampling may be impacted by the researcher's bias and the results may be skewed.
- The advantage of this method is that it allows the researcher to directly go to the target population of interest. For example, when auditing financial statements, experienced auditors can use their professional judgment to select important accounts or transactions that can provide sufficient data.



Sampling and Estimation

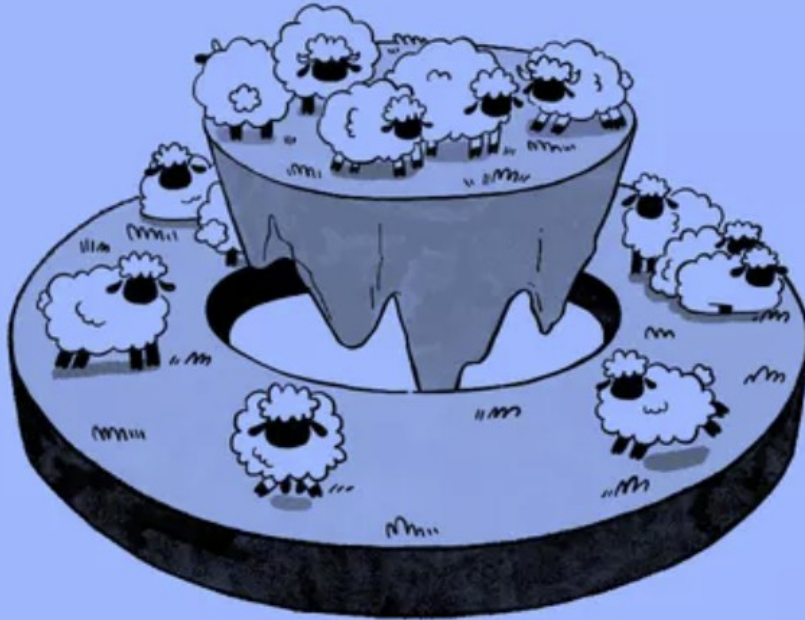
Central Limit Theorem(CLT):

- CLT is a statistical premise that, given a sufficiently large sample size from a population with a finite level of variance, the mean of all sampled variables from the same population will be approximately equal to the mean of the whole population.
- Furthermore, these samples approximate a normal distribution, with their variances being approximately equal to the variance of the population as the sample size gets larger,
- A sufficiently large sample size can predict the characteristics of a population more accurately.



Sampling and Estimation

Central Limit Theorem (CLT):



Central Limit Theorem (CLT)

[ˈsen-trəl ˈli-mət ˈthē-ə-rəm]

The principle that the distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the population's distribution.



Sampling and Estimation

Central Limit Theorem(CLT):

- The central limit theorem (CLT) states that the distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the population's distribution.
- Sample sizes equal to or greater than 30 are often considered sufficient for the CLT to hold.
- A key aspect of CLT is that the average of the sample means and standard deviations will equal the population mean and standard deviation.



Sampling and Estimation

Key Components of the Central Limit Theorem:

The central limit theorem is comprised of several key characteristics. These characteristics largely revolve around samples, sample sizes, and the population of data.

- 1) *Sampling is successive:*** This means some sample units are common with sample units selected on previous occasions.
- 2) *Sampling is random:*** All samples must be selected at random so that they have the same statistical possibility of being selected.



Sampling and Estimation

Key Components of the Central Limit Theorem:

- 3) Samples should be independent:** The selections or results from one sample should have no bearing on future samples or other sample results.
- 4) Samples should be limited:** It's often cited that a sample should be no more than 10% of a population if sampling is done without replacement. In general, larger population sizes warrant the use of larger sample sizes.
- 5) Sample size is increasing:** The central limit theorem is relevant as more samples are selected.



Sampling and Estimation

Central Limit Theorem Applications:

➤ The central limit theorem is useful when analyzing large data sets because it allows one to assume that the sampling distribution of the mean will be normally-distributed in most cases.

➤ This allows for easier statistical analysis and inference.

For example, investors can use central limit theorem to aggregate individual security performance data and generate distribution of sample means that represent a larger population distribution for security returns over a period of time.



Sampling and Estimation

Central Limit Theorem Applications:

1) Economics:

- Economists often use the central limit theorem when using sample data to draw conclusions about a population.
- For example, an economist may collect a simple random sample of 50 individuals in a town and use the average annual income of the individuals in the sample to estimate the average annual income of individuals in the entire town.
- If the economist finds that the average annual income of the individuals in the sample is \$58,000, then her best guess for the true average annual income of individuals in the entire town will be \$58,000.



Sampling and Estimation

Central Limit Theorem Applications:

2) Biology:

- Biologists use the central limit theorem whenever they use data from a sample of organisms to draw conclusions about the overall population of organisms.
- For example, a biologist may measure the height of 30 randomly selected plants and then use the sample mean height to estimate the population mean height.
- If the biologist finds that the sample mean height of the 30 plants is 10.3 inches, then her best guess for the population mean height will also be 10.3 inches..



Sampling and Estimation

Central Limit Theorem Applications:

3) Manufacturing:

- Manufacturing plants often use the central limit theorem to estimate how many products produced by the plant are defective.
- For example, the manager of the plant may randomly select 60 products produced by the plant in a given day and count how many of the products are defective.
- He can use the proportion of defective products in the sample to estimate the proportion of all products that are defective that are produced by the entire plant.



Sampling and Estimation

Central Limit Theorem Applications:

4) Surveys:

- Human Resources departments often use the central limit theorem when using surveys to draw conclusions about overall employee satisfaction at companies.
- For example, the HR department of some company may randomly select 50 employees to take a survey that assesses their overall satisfaction on a scale of 1 to 10.
- If it's found that the average satisfaction among employees in the survey is 8.5 then the best guess for the average satisfaction rating of all employees at the company is also 8.5.



THANK YOU