UNIT 4

Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate to deal with them. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying, updating and information privacy. The term "big data" often refers simply to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set.

Big data can be described by the following **CHARACTERISTICS**:

Volume: The quantity of generated and stored data. The size of the data determines the value and potential insight- and whether it can actually be considered big data or not.

Variety: The type and nature of the data. This helps people who analyze it to effectively use the resulting insight.

Velocity: In this context, the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development.

Variability: Inconsistency of the data set can hamper processes to handle and manage it.

Veracity: The quality of captured data can vary greatly, affecting accurate analysis.

Challenges in big Data -

1. **Uncertainty of the Data Management Landscape** – There are many competing technologies, and within each technical area there are numerous rivals. Our first challenge is making the best choices while not introducing additional unknowns and risk to big data adoption.

2. **The Big Data Talent Gap** – The excitement around big data applications seems to imply that there is a broad community of experts available to help in implementation. However, this is not yet the case, and the talent gap poses our second challenge.

3. **Getting Data into the Big Data Platform** – The scale and variety of data to be absorbed into a big data environment can overwhelm the unprepared data practitioner, making data accessibility and integration our third challenge.

4. **Synchronization Across the Data Sources** – As more data sets from diverse sources are incorporated into an analytical platform, the potential for time lags to impact data currency and consistency becomes our fourth challenge.

5. **Getting Useful Information out of the Big Data Platform** – Lastly, using big data for different purposes ranging from storage augmentation to enabling high-performance analytics is impeded if the information cannot be adequately provisioned back within the other

components of the enterprise information architecture, making big data syndication our fifth challenge.

CHALLENGES OF CONVENTIONAL SYSTEMS -

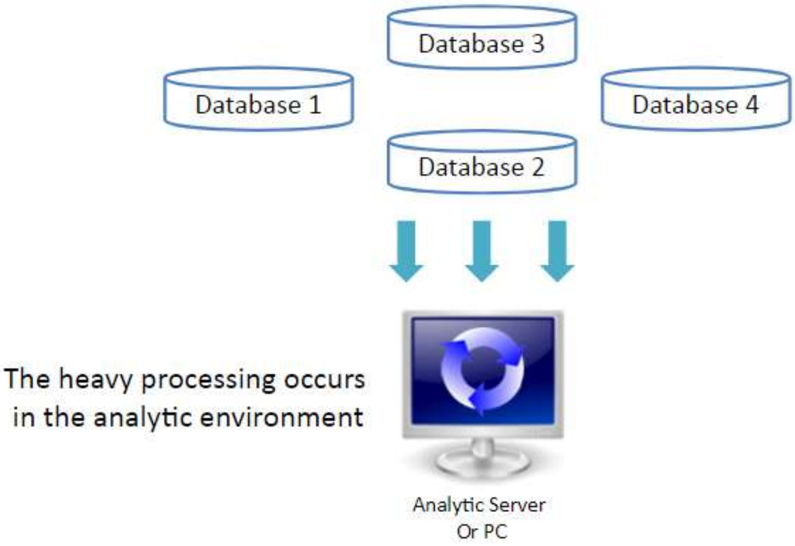
Big Data vs. Conventional Data

Big Data	**Normal or Conventional Data**
Huge data sets.	Data set size in control.
Unstructured data such as text, video, and audio.	Normally structured data such as numbers and categories, but it can take other forms as well.
Hard-to-perform queries and analysis.	Relatively easy-to-perform queries and analysis.
Needs a new methodology for analysis.	Data analysis can be achieved by using conventional methods.
Need tools such as Hadoop, Hive, Hbase, Pig, Sqoop, and so on.	Tools such as SQL, SAS, R, and Excel alone may be sufficient.
Raw transactional data.	The aggregated or sampled or filtered data.
Used for reporting, basic analysis, and text mining. Advanced analytics is only in a starting stage in big data.	Used for reporting, advanced analysis, and predictive modeling .
Big data analysis needs both programming skills (such as Java) and analytical skills to perform analysis.	Analytical skills are sufficient for conventional data; advanced analysis tools don’t require expert programing skills.
Petabytes/exabytes of data. Millions/billions of accounts. Billions/trillions of transactions.	Megabytes/gigabytes of data. Thousands/millions of accounts. Millions of transactions.
Generated by big financial institutions, Facebook, Google, Amazon, eBay, Walmart, and so on.	Generated by small enterprises and small banks.

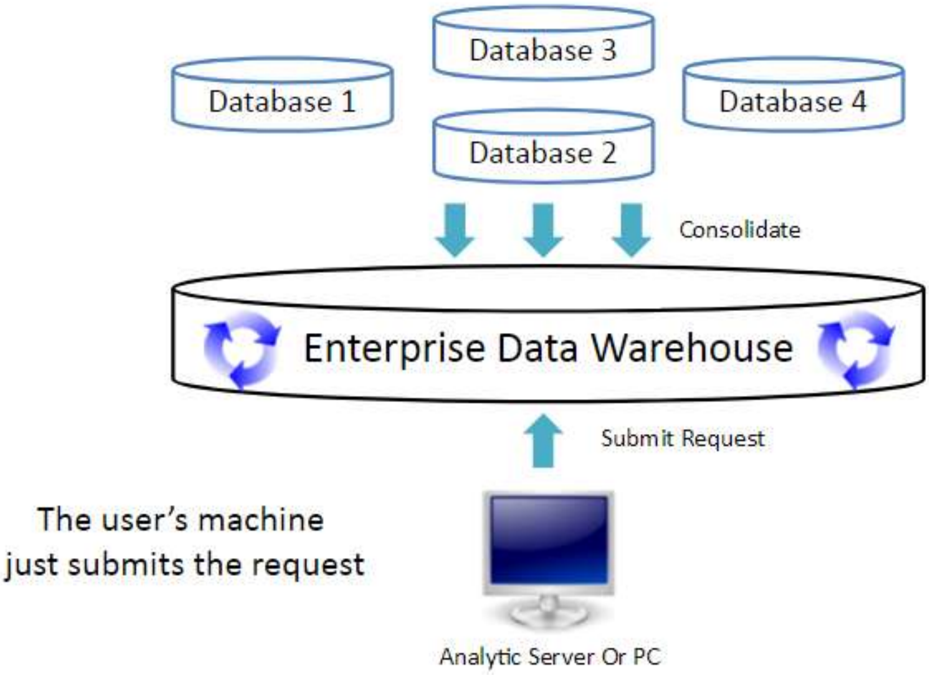
THE EVOLUTION OF ANALYTIC SCALABILITY

The amount of data organizations process continues to increase. The old methods for handling data won’t work anymore. Hence, important technologies to tame the big data tidal wave possible such as: MPP, The cloud, Grid computing and Map-Reduce.

Traditional Analytic Architecture- We had to pull all data together into a separate analytics environment to do analysis.



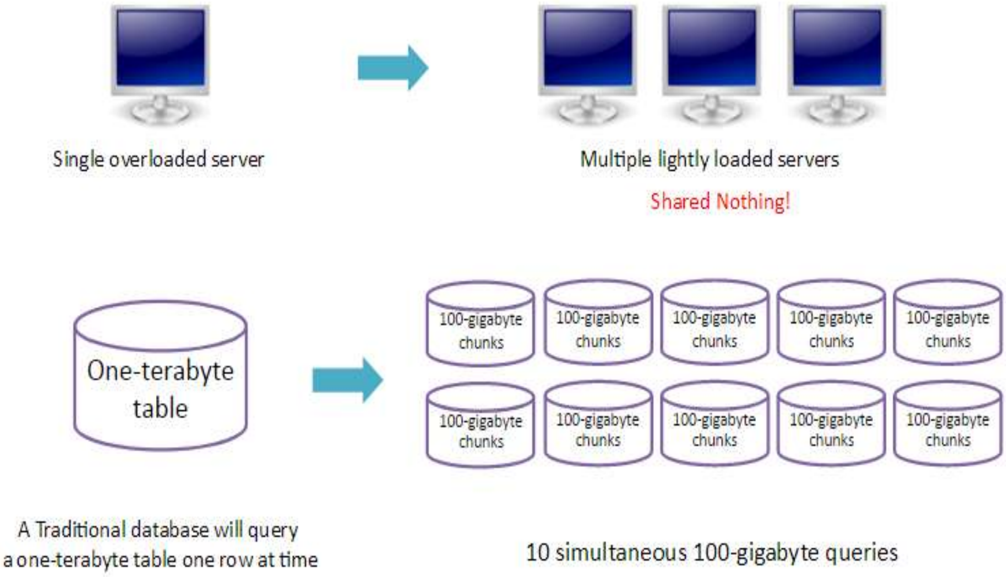
Modern In-Database Architecture- The processing stays in the database where the data has been consolidated.



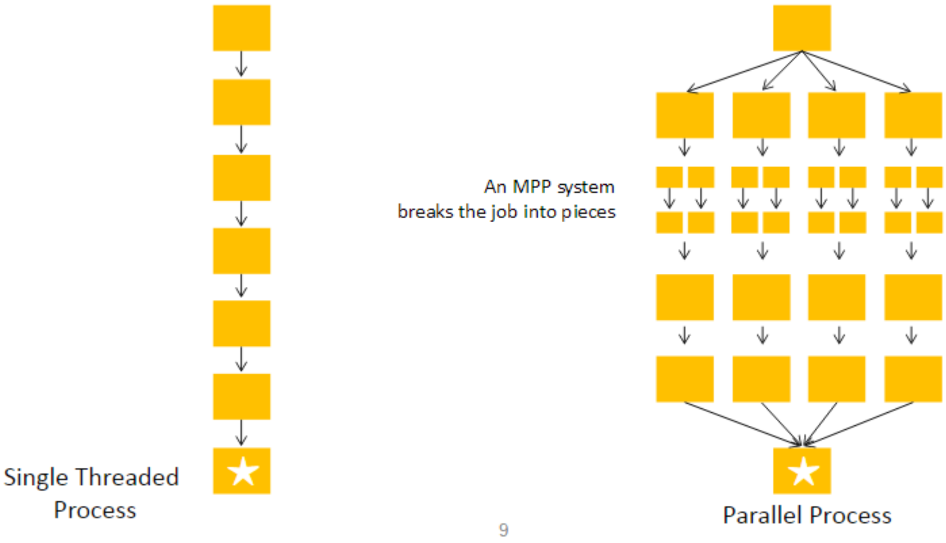
Massively Parallel Processing:

What is an MPP Database?

An MPP database breaks the data into independent chunks with independent disk and CPU.


Single overloaded server
Multiple lightly loaded servers
Shared Nothing!
One-terabyte table
100-gigabyte chunks
A Traditional database will query a one-terabyte table one row at time
10 simultaneous 100-gigabyte queries

Concurrent Processing: An MPP system allows the different sets of CPU and disk to run the process concurrently.



MPP systems build in redundancy to make recovery easy. MPP systems have resource management tools.

- Manage the CPU and disk space
- Query optimizer

Cloud computing is the new computing paradigm which provides large pool of dynamical scalable and virtual resources as a service on demand. The main principle behind cloud computing model is to offer computing, storage, and software as a service or as a utility. We just need internet to use these utilities. Cloud Computing refers to both the applications delivered as services over the Internet and the hardware and systems software in the datacenters that provide those services. The services themselves have long been referred to as Software as a Service (SaaS). The datacenter hardware and software is what we will call a Cloud.

Two Types of Cloud Environment-

Public Cloud:

- The services and infrastructure are provided off-site over the internet
- Greatest level of efficiency in shared resources
- Less secured and more vulnerable than private clouds

Private Cloud

- Infrastructure operated solely for a single organization
- The same features of a public cloud
- Offer the greatest level of security and control
- Necessary to purchase and own the entire cloud infrastructure

5 Essential Cloud Characteristics

- On-demand self-service
- Broad network access
- Resource pooling
  - Location independence
  - Rapid elasticity
- Measured service
  - Pay as you go.

Grid is a shared collection of reliable (cluster-tightly coupled) & unreliable resources (loosely coupled machines) and interactively communicating researchers of different virtual organisations (doctors, biologists, physicists).
Grid System controls and coordinates the integrity of the Grid by balancing the usage of reliable and unreliable resources among its participants providing better quality of service.

How Grid computing works?

In general, a grid computing system requires:

- **At least one computer, usually a server, which handles all the administrative duties for the System**
- **A network of computers running special grid computing network software.**

- **A collection of computer software called middleware**

Benefits of Grid Computing-

- Federate data and distribute it globally.

Support large multi-disciplinary collaboration across organizations and business.

- Enable recovery and failure
- Ability to run large-scale applications comprising thousands of computes, for wide range of applications.
- Reduces signal latency – the delay that builds up as data are transmitted over the Internet.

What is Map-reduce? A Parallel programming framework.

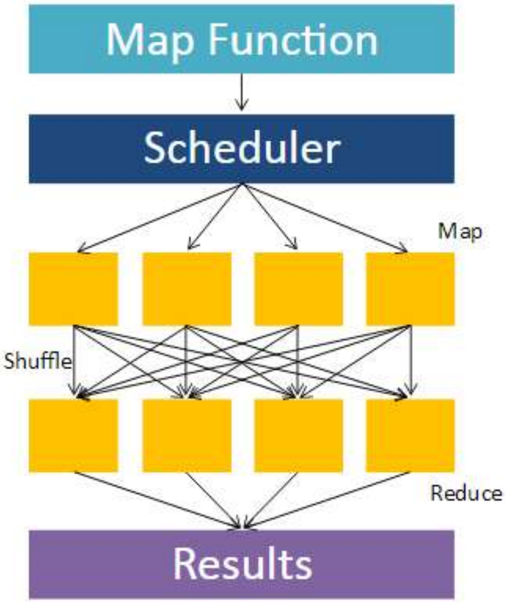
Map function- Processing a key/value pairs to generate a set of intermediate key/value pairs.

Reduce function- Merging all intermediate values associated with the same intermediate key.

How Map-Reduce Works

Let's assume there are 20 terabytes of data and 20 Map-Reduce server nodes for a project.

- Distribute a terabyte to each of the 20 nodes using a simple file copy process.
- Submit two programs (Map, Reduce) to the scheduler.
- The map program finds the data on disk and executes the logic it contains.
- The results of the map step are then passed to the reduce process to summarize and aggregate the final answers.



Strengths and Weaknesses:

Good for-

- Lots of input, intermediate, and output data
- Batch oriented datasets (ETL: Extract, Load, Transform)
- Cheap to get up and running because of running on commodity hardware

Bad for-

- Fast response time
- Large amounts of shared data
- CPU intensive operations (as opposed to data intensive)
- NOT a database!- No built-in security; No indexing; No query or process optimizer; No knowledge of other data that exists

ANALYTIC PROCESSES AND TOOLS -

Big data analytics examines large amounts of data to uncover hidden patterns, correlations and other insights.

Web analytics is the measurement, collection, analysis and reporting of **web** data for purposes of understanding and optimizing **web** usage.

Data Analytics is the science of analysis where statistics, data mining, data processing and even computer technology is utilized to break down data and come up with conclusive insights and information.

Big data analytics helps organizations harness their data and use it to identify new opportunities. That, in turn, leads to smarter business moves, more efficient operations, higher profits and happier customers.

Why big data analytics is important?

1. **Cost reduction.**

2. **Faster, better decision making.**

3. **New products and services.**

Types of Analytics Process-

- ***<u>Descriptive</u>***: A set of techniques for reviewing and examining the data set(s) to understand the data and analyze business performance.

- ***<u>Diagnostic</u>***: A set of techniques for determine what has happened and why

- ***Predictive***: A set of techniques that analyze current and historical data to determine what is most likely to (not) happen
- ***Prescriptive***: A set of techniques for computationally developing and analyzing alternatives that can become courses of action – either tactical or strategic – that may discover the unexpected
- ***Decisive***: A set of techniques for visualizing information and recommending courses of action to facilitate human decision-making when presented with a set of alternatives.

ANALYSIS VS. REPORTING,

While both areas are part of web analytics (note that analytics isn't similar to analysis), there's a vast difference between them-

Here are **five differences between reporting and analysis:**

1. Purpose

Reporting helps companies monitor their data even before digital technology boomed. Various organizations have been dependent on the information it brings to their business, as reporting extracts that and makes it easier to understand.

Analysis interprets data at a deeper level. While reporting can link between cross-channels of data, provide comparison, and make understand information easier (think of a dashboard, charts, and graphs, which are reporting *tools* and not analysis reports), analysis interprets this information and provides recommendations on actions.

2. Tasks

Here's a great differentiator to keep in mind if what you're doing is reporting or analysis:

Reporting includes building, configuring, consolidating, organizing, formatting, and summarizing. It's very similar to the abovementioned like turning data into charts, graphs, and linking data across multiple channels.

Analysis consists of questioning, examining, interpreting, comparing, and confirming. With big data, predicting is possible as well.

3. Outputs

Reporting and analysis have the push and pull effect from its users through their outputs. Reporting has a push approach, as it pushes information to users and outputs come in the forms of canned reports, dashboards, and alerts.

Analysis has a pull approach, where a data analyst draws information to further probe and to answer business questions. Outputs from such can be in the form of ad hoc responses and analysis presentations. Analysis presentations are comprised of insights, recommended actions, and a forecast of its impact on the company—all in a language that's easy to understand at the level of the user who'll be reading and deciding on it.

This is important for organizations to realize truly the value of data, such that a standard report is not similar to a meaningful analytics.

4. Delivery

Considering that reporting involves repetitive tasks—often with truckloads of data, automation has been a lifesaver, especially now with big data.

Analysis requires a more custom approach, with human minds doing superior reasoning and analytical thinking to extract insights, and technical skills to provide efficient steps towards accomplishing a specific goal. This is why data analysts and scientists are demanded these days, as organizations depend on them to come up with recommendations for leaders or business executives make decisions about their businesses.

5. Value

This isn't about identifying which one brings more value, rather understanding that both are indispensable when looking at the big picture. It should help businesses grow, expand, move forward, and make more profit or increase their value.

This Path to Value diagram illustrates how data converts into value by reporting and analysis such that it's not achievable without the other.

Data — Reporting — Analysis — Decision-making — Action — VALUE

Data alone is useless, and action without data is baseless. Both reporting and analysis are vital to bringing value to your data and operations.

MODERN DATA ANALYTIC TOOLS

Data analysis is about breaking that data down and assessing the impact of those patterns over time.

Qubole

Qubole simplifies speeds and scales big data analytics workloads against data stored on AWS, Google, or Azure clouds. This cloud-based data platform self-manages, self-optimizes and learns to improve automatically and as a result delivers unbeatable agility, flexibility, and TCO.

BigML

BigML is attempting to simplify machine learning. They offer a powerful Machine Learning service with an easy-to-use interface for you to import your data and get predictions out of it. You can even use their models for predictive analytics.

Statwing

Statwing takes data analysis to a new level, providing everything from beautiful visuals to complex analysis. Statwing selects statistical tests with the goal of making statistical testing intuitive and error-free.

Domo

Domo helps you put your data in one place, so you have access to the numbers you need to generate analysis, visualize changes, and make decisions. With plenty of visualization and collaboration tools, Domo is trying to make spreadsheets a thing of the past.

ThoughtSpot

By using the power of relational search – like the AI behind Google – anyone can search for terms and instantly find the data they need. ThoughtSpot will even help a user visualize and share that data to inform decision making.

STATISTICAL CONCEPTS: SAMPLING DISTRIBUTIONS, RE-SAMPLING, STATISTICAL INFERENCE, PREDICTION ERROR

SAMPLING DISTRIBUTIONS

Population and sample: population can be of two class- finite population and infinite population

Population- A set or collection of all the objects, actual or conceptual and mainly the set of numbers, measurements or observations which are under investigation.

Finite Population : All students in a College

Infinite Population : Total water in the sea or all the sand particle in sea shore.

Populations are often described by the distributions of their values, and it is common practice to refer to a population in terms of its distribution.

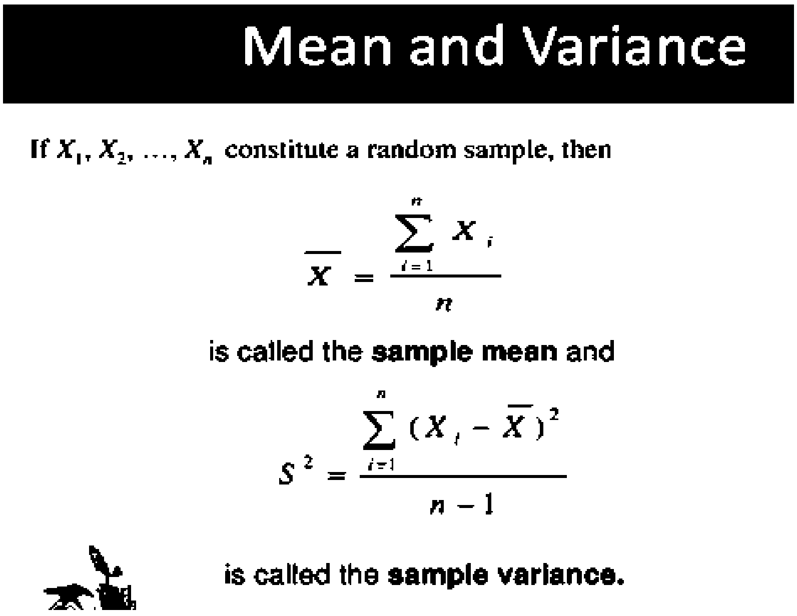
“Population f(x)” means a population is described by a frequency distribution, a probability distribution or a density f(x).

If a population is infinite it is impossible to observe all its values, and even if it is finite it may be impractical or uneconomical to observe it in its entirety. Thus it is necessary to use a sample.

Sample: A part of population collected for investigation which needed to be representative of population and to be large enough to contain all information about population.

1. Random Sample (finite population): • A set of observations X1, X2, …,Xn constitutes a random sample of size n from a finite population of size N, if its values are chosen so that each subset of n of the N elements of the population has the same probability of being selected.

2. Random Sample (infinite Population): A set of observations X1, X2, …, Xn constitutes a random sample of size n from the infinite population f(x) if: (1). Each Xi is a random variable whose distribution is given by f(x) (2). These n random variables are independent.



Mean and Variance

If $X_1, X_2, \ldots, X_n$ constitute a random sample, then

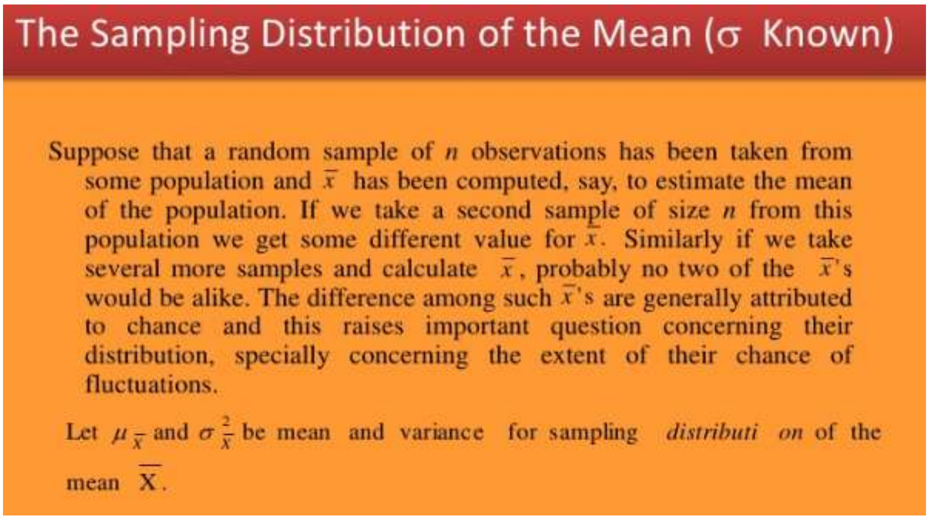
$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

is called the **sample mean** and

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$

is called the **sample variance.**

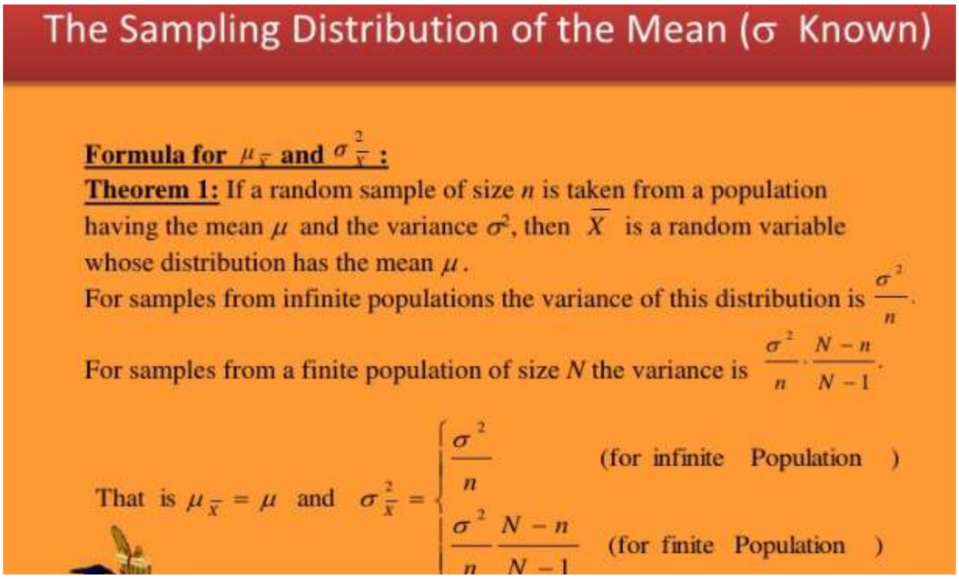
Sampling distribution: • The probability distribution of a random variable defined on a space of random samples is called a sampling distribution.



The Sampling Distribution of the Mean (σ Known)

Suppose that a random sample of n observations has been taken from some population and $\bar{x}$ has been computed, say, to estimate the mean of the population. If we take a second sample of size n from this population we get some different value for $\bar{x}$. Similarly if we take several more samples and calculate $\bar{x}$, probably no two of the $\bar{x}$'s would be alike. The difference among such $\bar{x}$'s are generally attributed to chance and this raises important question concerning their distribution, specially concerning the extent of their chance of fluctuations.

Let $\mu_{\bar{X}}$ and $\sigma^2_{\bar{X}}$ be mean and variance for sampling *distributi on* of the mean $\overline{X}$.



The Sampling Distribution of the Mean (σ Known)

Formula for $\mu_{\overline{X}}$ and $\sigma^2_{\overline{X}}$:

Theorem 1: If a random sample of size n is taken from a population having the mean μ and the variance σ^2, then $\overline{X}$ is a random variable whose distribution has the mean μ.

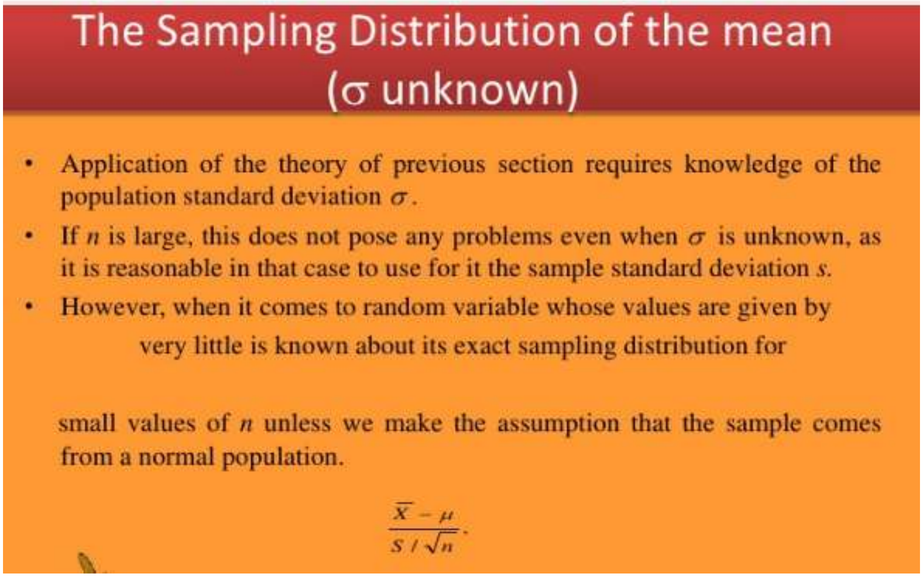
For samples from infinite populations the variance of this distribution is $\frac{\sigma^2}{n}$.

For samples from a finite population of size N the variance is $\frac{\sigma^2}{n}\cdot\frac{N-n}{N-1}$.

$$\text{That is } \mu_{\overline{X}} = \mu \text{ and } \sigma^2_{\overline{X}} = \begin{cases} \dfrac{\sigma^2}{n} & \text{(for infinite Population)} \\ \dfrac{\sigma^2}{n}\cdot\dfrac{N-n}{N-1} & \text{(for finite Population)} \end{cases}$$

If the random samples come from a normal population, the X sampling distribution of the mean is normal regardless of the size of the sample.

The Sampling Distribution of the Mean (*σ* unknown)

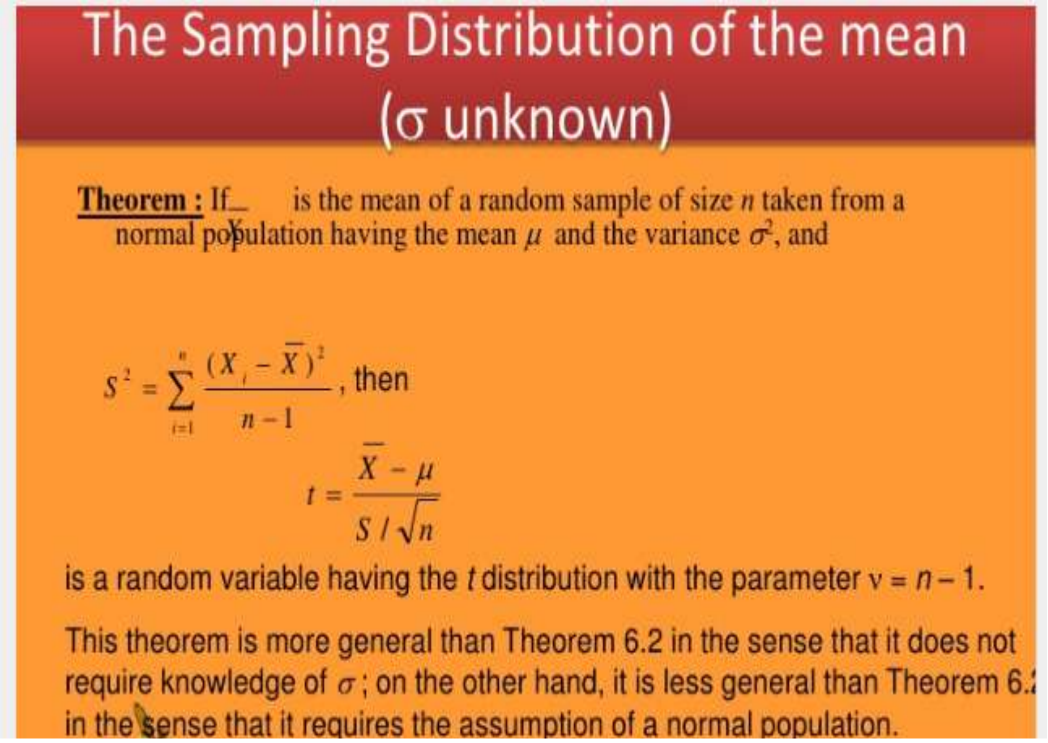


The Sampling Distribution of the mean (σ unknown)

- Application of the theory of previous section requires knowledge of the population standard deviation σ.
- If n is large, this does not pose any problems even when σ is unknown, as it is reasonable in that case to use for it the sample standard deviation s.
- However, when it comes to random variable whose values are given by very little is known about its exact sampling distribution for

  small values of n unless we make the assumption that the sample comes from a normal population.

$$\frac{\overline{X}-\mu}{S/\sqrt{n}}.$$



The Sampling Distribution of the mean (σ unknown)

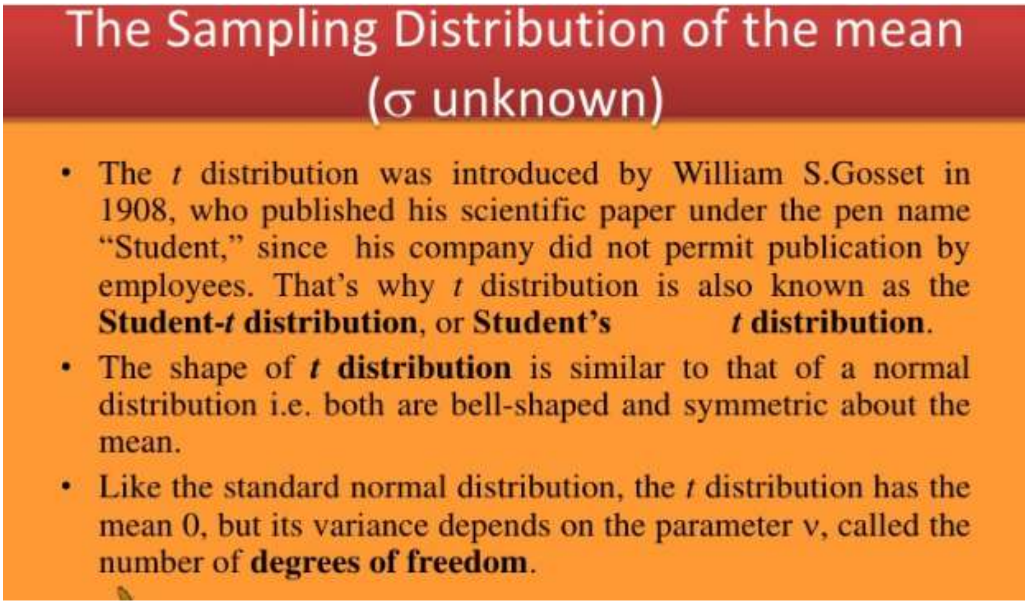
Theorem : If $\overline{X}$ is the mean of a random sample of size n taken from a normal population having the mean μ and the variance σ^2, and

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n-1} \text{, then}$$

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

is a random variable having the t distribution with the parameter $\nu = n - 1$.

This theorem is more general than Theorem 6.2 in the sense that it does not require knowledge of σ; on the other hand, it is less general than Theorem 6. in the sense that it requires the assumption of a normal population.



The Sampling Distribution of the mean (σ unknown)

- The *t* distribution was introduced by William S.Gosset in 1908, who published his scientific paper under the pen name "Student," since his company did not permit publication by employees. That's why *t* distribution is also known as the **Student-*t* distribution**, or **Student's *t* distribution**.
- The shape of ***t* distribution** is similar to that of a normal distribution i.e. both are bell-shaped and symmetric about the mean.
- Like the standard normal distribution, the *t* distribution has the mean 0, but its variance depends on the parameter ν, called the number of **degrees of freedom**.

RESAMPLING

In statistics, resampling is any of a variety of methods for doing one of the following:

1. Estimating the precision of sample statistics (medians, variances, percentiles) by using subsets of available data (jackknifing) or drawing randomly with replacement from a set of data points (bootstrapping)

2. Exchanging labels on data points when performing significance tests (permutation tests, also called exact tests, randomization tests, or re-randomization tests)

3. Validating models by using random subsets (bootstrapping, cross validation)

Common resampling techniques include bootstrapping, jackknifing and permutation tests.

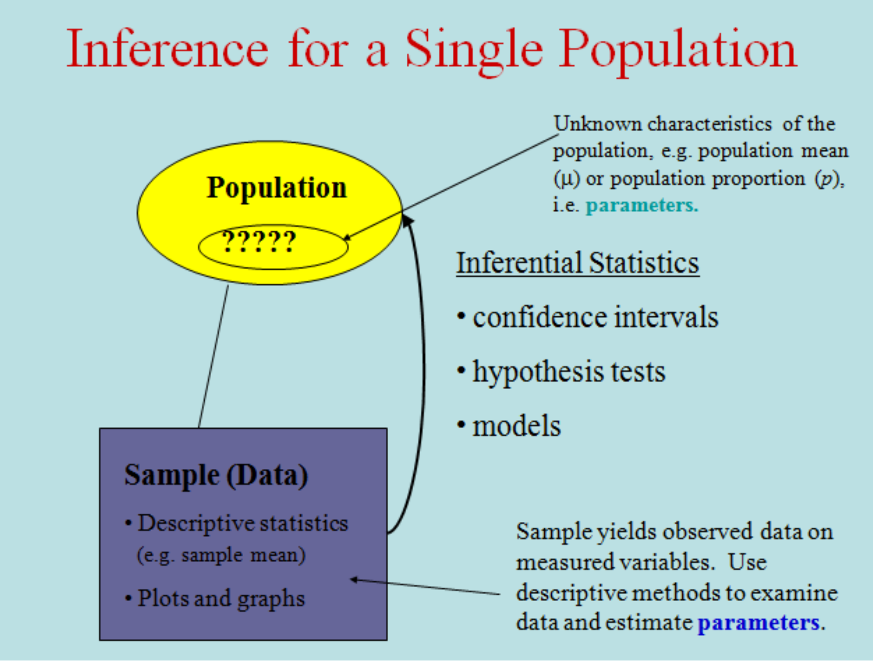
Bootstrapping is a statistical method for estimating the sampling distribution of an estimator by sampling with replacement from the original sample, most often with the purpose of deriving robust estimates of standard errors and confidence intervals of a population parameter like a mean, median, proportion, odds ratio, correlation coefficient or regression coefficient. It may also be used for constructing hypothesis tests.

Jackknifing, which is similar to bootstrapping, is used in statistical inference to estimate the bias and standard error (variance) of a statistic, when a random sample of observations is used to calculate it.

A **permutation test** (also called a randomization test, re-randomization test, or an exact test) is a type of statistical significance test in which the distribution of the test statistic under the null hypothesis is obtained by calculating all possible values of the test statistic under rearrangements of the labels on the observed data points.

STATISTICAL INFERENCE

Use a random sample to learn something about a larger population.



Inference for a Single Population

Population

?????

Unknown characteristics of the population, e.g. population mean (μ) or population proportion (*p*), i.e. **parameters.**

Inferential Statistics

- confidence intervals
- hypothesis tests
- models

Sample (Data)

- Descriptive statistics (e.g. sample mean)
- Plots and graphs

Sample yields observed data on measured variables. Use descriptive methods to examine data and estimate **parameters**.

Two main ways to learn about a population-

a. Confidence intervals

b. Hypothesis testing

Confidence intervals

- Allow us to use sample data to **estimate** a population value, like the true mean or the true proportion, i.e. estimate parameters.

- *Example:* What is the current mean GPA of U.S. college & university students?

Hypothesis testing

- Allows us to use sample data to **test a claim** about a population, such as testing whether a population proportion or population mean equals some number.

- ***Example:*** The mean GPA of U.S. college & university students today is larger than 2.70 which was the mean GPA in 1990?

General Idea of Hypothesis Testing-

- Make an initial assumption.
- Collect evidence (data).
- Based on the available evidence, decide whether or not the initial assumption is reasonable.

PREDICTION ERROR

Prediction attempts to form patterns that permit it to predict the next event(s) given the available input data.

- Deterministic predictions: If Bob leaves the bedroom before 7:00 am on a workday, then he will make coffee in the kitchen.

- Probabilistic sequence models: If Bob turns on the TV in the evening then he will 80% of the time go to the kitchen to make popcorn.

The **Prediction Error** tries to represent the noise through the concept of training error versus test error. We fit our model to the training set. We take our model, and then we apply it to new data that the model hasn't seen. In general, because the more data, the bigger the sample size, the more information you have, the lower the error is.

Types of prediction error-

Training: Training error is the error we get applying the model to the same data from which we trained.

Test: Test error is the error that we incur on new data. The test error is actually how well we'll do on future data the model hasn't seen.

Training vs. Test

Training error almost always Underestimates test error, sometimes dramatically.

Training error usually Underestimates test error when the model is very complex (compared to the training set size), and is a pretty good estimate when the model is not very complex. However, it's always possible we just get too few hard-to-predict points in the test set, or too many in the training set. Then the test error can be LESS than training error, when by chance the test set has easier cases than the training set.