BIG DATA ANALYTICS

INTRODUCTION TO BIG DATA

INTRODUCTION TO BIG DATA

Syllabus:

UNIT I INTRODUCTION TO BIG DATA 9

Introduction to BigData Platform – Challenges of Conventional Systems - Intelligent data analysis – Nature of Data - Analytic Processes and Tools - Analysis vs Reporting - Modern Data Analytic Tools - Statistical Concepts: Sampling Distributions - Re-Sampling - Statistical Inference - Prediction Error.

Table of Contents

Total Pages: 106

1. INTRODUCTION TO BIGDATA PLATFORM

1.1 Introduction Data

1.1.1 Data and Information

- Data are plain facts.
- The word "data" is plural for "datum."
- Data is nothing but facts and statistics stored or free flowing over a network, generally it's raw and unprocessed.
- When data are processed, organized, structured or presented in a given context so as to make them useful, they are called Information.
- It is not enough to have data (such as statistics on the economy).
- Data themselves are fairly useless, but when these data are interpreted and processed to determine its true meaning, they becomes useful and can be called Information.

  For example: When you visit any website, they might store you IP address, that is data, in return they might add a cookie in your browser, marking you that you visited the website, that is data, your name, it's data, your age, it's data.

- **What is Data?**
  - The quantities, characters, or symbols on which operations are performed by a computer,
  - which may be stored and transmitted in the form of electrical signals and
  - recorded on magnetic, optical, or mechanical recording media.

- **3 Actions on Data**
  - Capture
  - Transform
  - Store

BigData

- Big Data may well be the Next Big Thing in the IT world.
- Big data burst upon the scene in the first decade of the 21st century.

- The first organizations to embrace it were online and startup firms.
- Firms like Google, eBay, LinkedIn, and Facebook were built around big data from the beginning.
- Like many new information technologies,
    - big data can bring about dramatic cost reductions,
    - substantial improvements in the time required to perform a computing task, or
    - new product and service offerings.

- Walmart handles more than 1 million customer transactions every hour.
- Facebook handles 40 billion photos from its user base.
- Decoding the human genome originally took 10years to process; now it can be achieved in one week.

- **What is Big Data?**
    - Big Data is also **data** but with a **huge size**.
    - Big Data is a term used to describe a collection of data that is huge in size and yet growing exponentially with time.
    - In short such data is so large and complex that none of the traditional data management tools are able to store it or process it efficiently.

No single definition; here is from Wikipedia:

- **Big data** is the term for
    - a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.

Examples of Bigdata

- Following are some the examples of Big Data-
    - The **New York Stock Exchange** generates about ***one terabyte*** of new trade data per day.
    - Other examples of Big Data generation includes
        - stock exchanges,
        - social media sites,
        - jet engines,
        - etc.

Types of Big Data

- BigData could be found in three forms:

1. **Structured**
2. **Unstructured**
3. **Semi-structured**

What is Structured Data?

- Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data.
- Developed techniques for working with such kind of data (where the format is well known in advance) and also deriving value out of it.
- Foreseeing issues of today :
  - when a size of such data grows to a huge extent, typical sizes are being in the rage of multiple zetta bytes.
- ***Do you know?***
- ***10^{21} bytes*** equal to ***1 zettabyte*** or ***one billion terabytes*** forms ***a zettabyte***.
  - That is why the name Big Data is given and imagine the challenges involved in its storage and processing?
- ***Do you know?***
  - Data stored in a relational database management system is one example of a **'structured'** data.

- An 'Employee' table in a database is an example of Structured Data:

Employee_ID	**Employee_Name**	**Gender**	**Department**	**Salary_In_lacs**
2365	**Rajesh Kulkarni**	**Male**	**Finance**	**650000**
3398	**Pratibha Joshi**	**Female**	**Admin**	**650000**
7465	**Shushil Roy**	**Male**	**Admin**	**500000**
7500	**Shubhojit Das**	**Male**	**Finance**	**500000**
7699	**Priya Sane**	**Female**	**Finance**	**550000**

Unstructured Data

- Any data with unknown form or the structure is classified as unstructured data.
- In addition to the size being huge,
  - un-structured data poses multiple challenges in terms of its processing for deriving value out of it.
  - A typical example of unstructured data is
    - a heterogeneous data source containing a combination of simple text files, images, videos etc.
- Now day organizations have wealth of data available with them but unfortunately,
  - they don't know how to derive value out of it since this data is in its raw form or unstructured format.

- Example of Unstructured data
  - The output returned by 'Google Search'

Semi-structured Data

- Semi-structured data can contain both the forms of data.
- Semi-structured data as a structured in form
  - but it is actually not defined with e.g. a table definition in relational DBMS.
- Example of semi-structured data is
  - a data represented in an XML file.

- **Personal data stored in an XML file.**

```
<rec>
<name>Prashant Rao</name>
<sex>Male</sex>
<age>35</age>
</rec>
<rec>
<name>Seema R.</name>
<sex>Female</sex>
<age>41</age>
```

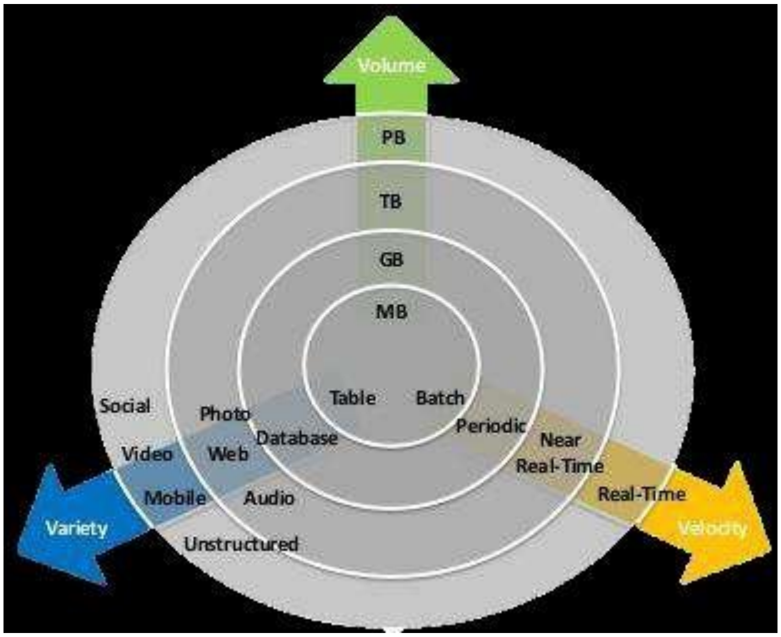
```
</rec>
<rec>
<name>Satish Mane</name>
<sex>Male</sex>
<age>29</age>
</rec>
<rec>
<name>Subrato Roy</name>
<sex>Male</sex>
<age>26</age>
</rec>
<rec>
<name>Jeremiah J.</name>
<sex>Male</sex>
<age>35</age></rec>
```

Characteristics of BD OR 3Vs of Big Data

- **Three Characteristics of Big Data V3s:**

1) Volume
   - Data quantity
2) Velocity
   - Data Speed
3) Variety
   - Data Types

Growth of Big Data



Storing Big Data

- **Analyzing your data characteristics**
  - Selecting data sources for analysis
  - Eliminating redundant data
  - Establishing the role of NoSQL
- **Overview of Big Data stores**
  - **Data models**: key value, graph, document, column-family
  - Hadoop Distributed File System (HDFS)
  - Hbase
  - Hive

Processing Big Data

- **Integrating disparate data stores**
  - Mapping data to the programming framework
  - Connecting and extracting data from storage
  - Transforming data for processing
  - Subdividing data in preparation for Hadoop MapReduce
- **Employing Hadoop MapReduce**

  - Creating the components of Hadoop MapReduce jobs
  - Distributing data processing across server farms
  - Executing Hadoop MapReduce jobs
  - Monitoring the progress of job flows

Why Big Data?

- Growth of Big Data is needed
  - Increase of storage capacities
  - Increase of processing power
  - Availability of data(different data types)
  - Every day we create 2.5 quintillion bytes of data; 90% of the data in the world today has been created in the last two years alone

- Huge storage need in Real Time Applications
  - FB generates 10TB daily
  - Twitter generates 7TB of data Daily
  - IBM claims 90% of today's stored data was generated in just the last two years.

How Is Big Data Different?

1) Automatically generated by a machine (e.g. Sensor embedded in an engine)
2) Typically an entirely new source of data(e.g. Use of the internet)
3) Not designed to be friendly(e.g. Text streams)
4) May not have much values
   - Need to focus on the important part

Big Data sources

- Users
- Application
- Systems
- Sensors

Moved to

- Large and growing files (Big data files)

Risks of Big Data

- Will be so overwhelmed
  - Need the right people and solve the right problems
- Costs escalate too fast
  - Isn't necessary to capture 100%
- Many sources of big data is privacy
  - self-regulation
  - Legal regulation

Leading Technology Vendors

Example Vendors	***Commonality***
IBM – Netezza	• MPP architectures
EMC – Greenplum	• Commodity Hardware
Oracle – Exadata	• RDBMS based
	• Full SQL compliance

1.1.2 Basics of Bigdata Platform

- Big Data platform is IT solution which combines several Big Data tools and utilities into one packaged solution for managing and analyzing Big Data.

- **Big data platform** is a type of IT solution that combines the features and capabilities of several **big data** application and utilities within a single solution.
- It is an enterprise class IT **platform** that enables organization in developing, deploying, operating and managing a **big data** infrastructure /environment.

What is Big Data Platform?

- Big Data Platform is integrated IT solution for Big Data management which combines several software system, software tools and hardware to provide easy to use tools system to enterprises.
- It is a single one-stop solution for all Big Data needs of an enterprise irrespective of size and data volume. Big Data Platform is enterprise class IT solution for developing, deploying and managing Big Data.
- There are several Open source and commercial Big Data Platform in the market with varied features which can be used in Big Data environment.
- Big data platform is a type of IT solution that combines the features and capabilities of several big data application and utilities within a single solution.
- It is an enterprise class ITplatformthat enables organization in developing, deploying, operating and managing abig datainfrastructure /environment.
- Big data platform generally consists of big data storage, servers, database, big data management, business intelligence and other big data management utilities
- It also supports custom development, querying and integration with other systems.
- The primary benefit behind a big data platform is to reduce the complexity of multiple vendors/ solutions into a one cohesive solution.
- Big data platform are also delivered through cloud where the provider provides an all inclusive big data solutions and services.

1.1.2.2 Features of Big Data Platform

Here are most important features of any good Big Data Analytics Platform:

a) Big Data platform should be able to accommodate new platforms and tool based on the business requirement. Because business needs can change due to new technologies or due to change in business process.

b) It should support linear scale-out

c) It should have capability for rapid deployment

d) It should support variety of data format

e) Platform should provide data analysis and reporting tools

f) It should provide real-time data analysis software

g) It should have tools for searching the data through large data sets

Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate.

Challenges include

- Analysis,
- Capture,
- Data Curation,
- Search,
- Sharing,
- Storage,
- Transfer,
- Visualization,
- Querying,
- Updating

Information Privacy.

- The term often refers simply to the use of predictive analytics or certain other ***advancedmethods*** to extract value from data, and seldom to a particular size of data set.

- **ACCURACY** in big data may lead to more confident decision making, and better decisions can result in greater operational efficiency, cost reduction and reduced risk.
- Big data usually includes data sets with sizes beyond the ability of commonly used
- software tools to capture, curate, manage, and process data within a tolerable elapsed
- time. Big data "size" is a constantly moving target.
- Big data requires a set of techniques and technologies with new forms of integration to
- reveal insights from datasets that are diverse, complex, and of a massive scale

1.1.2.3 List of BigData Platforms

a) Hadoop
b) Cloudera
c) Amazon Web Services
d) Hortonworks
e) MapR
f) IBM Open Platform
g) Microsoft HDInsight
h) Intel Distribution for Apache Hadoop
i) Datastax Enterprise Analytics
j) Teradata Enterprise Access for Hadoop
k) Pivotal HD

a) Hadoop

What is Hadoop?

- Hadoop is open-source, Java based programming framework and server software which is used to save and analyze data with the help of 100s or even 1000s of commodity servers in a clustered environment.
- Hadoop is designed to storage and process large datasets extremely fast and in fault tolerant way.
- Hadoop uses HDFS (Hadoop File System) for storing data on cluster of commodity computers. If any server goes down it know how to replicate the data and there is no loss of data even in hardware failure.
- Hadoop is Apache sponsored project and it consists of many software packages which runs on the top of the Apache Hadoop system.
- Top Hadoop based Commercial Big Data Analytics Platform
- Hadoop provides set of tools and software for making the backbone of the Big Data analytics system.
- Hadoop ecosystem provides necessary tools and software for handling and analyzing Big Data.
- On the top of the Hadoop system many applications can be developed and plugged-in to provide ideal solution for Big Data needs.

b) Cloudera

- Cloudra is one of the first commercial Hadoop based Big Data Analytics Platform offering Big Data solution.
- Its product range includes Cloudera Analytic DB, Cloudera Operational DB, Cloudera Data Science & Engineering and Cloudera Essentials.
- All these products are based on the Apache Hadoop and provides real-time processing and analytics of massive data sets.

Website: https://www.cloudera.com

c) Amazon Web Services

- Amazon is offering Hadoop environment in cloud as part of its Amazon Web Services package.
- AWS Hadoop solution is hosted solution which runs on Amazon's Elastic Cloud Compute and Simple Storage Service (S3).
- Enterprises can use the Amazon AWS to run their Big Data processing analytics in the cloud environment.
- Amazon EMR allows companies to setup and easily scale Apache Hadoop, Spark, HBase, Presto, Hive, and other Big Data Frameworks using its cloud hosting environment.

  Website: https://aws.amazon.com/emr/

d) Hortonworks

- Hortonworks is using 100% open-source software without any propriety software. Hortonworks were the one who first integrated support for Apache HCatalog.
- The Hortonworks is a Big Data company based in California.
- This company is developing and supports application for Apache Hadoop.

Hortonworks Hadoop distribution is 100% open source and its enterprise ready with following features:

- Centralized management and configuration of clusters
- Security and data governance are built in feature of the system
- Centralized security administration across the system

Website: https://hortonworks.com/

e) MapR

- MapR is another Big Data platform which us using the Unix file system for handling data.
- It is not using HDFS and this system is easy to learn anyone familiar with the Unix system.
- This solution integrates Hadoop, Spark, and Apache Drill with a real-time data processing feature.
- Website: https://mapr.com

f) IBM Open Platform

- IBM also offers Big Data Platform which is based on the Hadoop eco-system software.
- IBM well knows company in software and data computing.

It uses the latest Hadoop software and provides following features (IBM Open Platform Features):

- Based on 100% Open source software
- Native support for rolling Hadoop upgrades
- Support for long running applications within YEARN.
- Support for heterogeneous storage which includes HDFS for in-memory and SSD in addition to HDD
- Native support for Spark, developers can use Java, Python and Scala to written program
- Platform includes Ambari, which is a best tool for provisioning, managing & monitoring Apache Hadoop clusters
- IBM Open Platform includes all the software of Hadoop ecosystem e.g. HDFS, YARN, MapReduce, Ambari, Hbase, Hive, Oozie, Parquet, Parquet Format, Pig, Snappy, Solr, Spark, Sqoop, Zookeeper, Open JDK, Knox, Slider
- Developer can download the trial Docker Image or Native installer for testing and learning the system
- Application is well supported by IBM technology team

Website: https://www.ibm.com/analytics/us/en/technology/hadoop/

g) Microsoft HDInsight

- The Microsoft HDInsight is also based on the Hadoop distribution and it's a commercial Big Data platform from Microsoft.
- Microsoft is software giant which is into development of windows operating system for Desktop users and Server users.
- This is the big Hadoop distribution offering which runs on the Windows and Azure environment.
- It offer customized, optimized open source Hadoop based analytics clusters which uses Spark, Hive, MapReduce, HBase, Strom, Kafka and R Server which runs on the Hadoop system on windows/Azure environment.

Website: https://azure.microsoft.com/en-in/services/hdinsight/

h) Distribution for Apache Hadoop

- Intel also offers its package distribution of Hadoop software which includes company's Graph builder and Analytics toolkit.
- This distribution can be purchased with various channel partners and come with support and yearly subscription.

Website: http://www.intel.com/content/www/us/en/software/intel-distribution-for-apache-hadoop-software-solutions.html

i) Datastax Enterprise Analytics

- Datastax Enterprise Analytics is another play in the Big Data Analytics platform which offers its own distribution which is based on Apache Cassandra database management system which runs on the top of Apache Hadoop installation.
- It also included propriety system with a dashboard which is used for security management, searching data, dashboard for viewing various details and visualization engine.
- It can handle analysis of 10 million data points every second, so it's a powerful system.

Features:

- It provides powerful indexing, search, analytics and graph functionality into the Big Data system
- It supports advanced indexing and searching features

- It comes with powerful integrated analytics system
- It provides multi-model support into the platform. It supports key-value, tabular, JSON/Document and graph data formats. Powerful search features enables the users to get required data in real-time

Website: http://www.datastax.com/

j) Teradata Enterprise Access for Hadoop

- Teradata Enterprise Access for Hadoop is another player into Big Data Platform and it offers package Hadoop distribution which again based on Hortonworks distribution.
- Teradata Enterprise Access for Hadoop offers Hardware and software in its Big Data solution which can be used by enterprise to process its data sets.

Company offers:

- Teradata
- Teradata Aster and
- Hadoop

as part of its package solution.

Website: http://www.teradata.com

k) Pivotal HD

Pivotal HD offers is another Hadoop distribution with includes includes database tools Greenplum and analytics platform Gemfire.

Features:

- It can be installed on-premise and in public clouds
- This system is based on the open source software
- It supports data evolution within the 3 years subscription period.

Indian railways, BMW, China Citic Bank and many other big players are using this distribution of Big Data Platform.

Website: https://pivotal.io/

1.1.3 Open Source Big Data Platform

There are various open-source Big Data Platform which can be used for Big Data handling and data analytics in real-time environment.

Both small and Big Enterprise can use these tools for managing their enterprise data for getting best value from their enterprise data.

i) Apache Hadoop

- Apache Hadoop is Big Data platform and software package which is Apache sponsored project.
- Under Apache Hadoop project various other software is being developed which runs on the top of Hadoop system to provide enterprise grade data management and analytics solutions to enterprise.
- Apache Hadoop is open-source, distributed file system which provides data processing and analysis engine for analyzing large set of data.
- Hadoop can run on Windows, Linux and OS X operating systems, but it is mostly used on Ubunut and other Linux variants.

ii) MapReduce

- The MapReduce engine was originally written by Google and this is the system which enables the developers to write program which can run in parallel on 100 or even 1000s of computer nodes to process vast data sets.
- After processing all the job on the different nodes it comes the results and return it to the program which executed the MapReduce job.
- This software is platform independent and runs on the top of Hadoop ecosystem. It can process tremendous data at very high speed in Big Data environment.

iii) GridGain

- GridGain is another software system for parallel processing of data just like MapRedue. GridGain is an alternative of Apache MapReduce.

- GridGain is used for the processing of in-memory data and its is based on Apache Iginte framework.
- GridGain is compatable with the Hadoop HDFS and runs on the top of Hadoop ecosystem.
- Then enterprise version of GridGain can be purchased from official website of GridGain. While free version can be downloaded from GitHub repository.

Website: https://www.gridgain.com/

iv) HPCC Systems

- HPCC Systems stands for "high performance computing cluster" and this system is developed by LexisNexis Risk Solutions.
- According to the company this software is much faster than Hadoop and can be used in the cloud environment.
- HPCC Systems is developed in C++ and compiled into binary code for distribution.
- HPCC Systems is open-source, massive parallel processing system which is installed in cluster to process data in real-time.
- It requires Linux operating system and runs on the commodity servers connected with high-speed network.
- It is scalable from one node to 1000s of nodes to provide performance and scalability.
- Website: https://hpccsystems.com/

v) Apache Storm

- Apache Storm is a software for real-time computing and distributed processing.
- Its free and open-source software developed at Apache Software foundation. It's a real-time, parallel processing engine.
- Apache Storm is highly scalable, fault-tolerant which supports almost all the programming language.

vi) Apache Strom can be used in:

- Realtime analytics
- Online machine learning

- Continuous computation
- Distributed RPC
- ETL
- And all other places where real-time processing is required.

Apache Strom is used by Yahoo, Twitter, Spotify, Yelp, Flipboard and many other data giants.

Website: http://storm.apache.org/

vii) Apache Spark

- Apache Spark is software that runs on the top of Hadoop and provides API for real-time, in-memory processing and analysis of large set of stored in the HDFS.
- It stores the data into memory for faster processing.
- Apache Spark runs program 100 times faster in-memory and 10 times faster on disk as compared to the MapRedue.
- Apache Spark is here to faster the processing and analysis of big data sets in Big Data environment.
- Apache Spark is being adopted very fast by the business to analyze their data set to get real value of their data.
- Website: http://spark.apache.org/

viii) SAMOA

- SAMOA stands for Scalable Advanced Massive Online Analysis,
- It's a system for mining the Big Data streams.
- SAMOA is open-source software distributed at GitHub, which can be used as distributed machine learning framework also.
- Website: https://github.com/yahoo/samoa

Thus, the Big Data industry is growing very fast in 2017 and companies are fast moving their data to Big Data Platform. There is huge requirement of Big Data in the job market; many companies are providing training and certifications in Big Data technologies.

****

1.2. CHALLENGES OF CONVENTIONAL SYSTEMS

1.2.1 Introduction to Conventional Systems

What is Conventional System?

Conventional Systems.

- The **system** consists of one or more zones each having either manually operated call points or automatic detection devices, or a combination of both.
- **Big data** is **huge** amount of **data** which is beyond the processing capacity of**conventional data** base **systems** to manage and analyze the **data** in a specific time interval.

Difference between conventional computing and intelligent computing

- The conventional computing functions logically with a set of rules and calculations while the neural computing can function via images, pictures, and concepts.
- Conventional computing is often unable to manage the variability of data obtained in the real world.
- On the other hand, neural computing, like our own brains, is well suited to situations that have no clear algorithmic solutions and are able to manage noisy imprecise data. This allows them to excel in those areas that conventional computing often finds difficult.

1.2.2 Comparison of Big Data with Conventional Data

Big Data	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.

The aggregated or sampled or filtered data.	Raw transactional 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.

1.2.2 List of challenges of Conventional Systems

The following list of challenges has been dominating in the case Conventional systems in real time scenarios:

1) **Uncertainty of Data Management Landscape**
2) **The Big Data Talent Gap**
3) **The talent gap that exists in the industry Getting data into the big data platform**
4) **Need for synchronization across data sources**
5) **Getting important insights through the use of Big data analytics**

1) **Uncertainty of Data Management Landscape:**

- Because big data is continuously expanding, there are new companies and technologies that are being developed everyday.

- A big challenge for companies is to find out which technology works bests for them without the introduction of new risks and problems.

2) **The Big Data Talent Gap:**

- While Big Data is a growing field, there are very few experts available in this field.
- This is because Big data is a complex field and people who understand the complexity and intricate nature of this field are far few and between.

3) **The talent gap that exists in the industry Getting data into the big data platform:**

- Data is increasing every single day. This means that companies have to tackle limitless amount of data on a regular basis.
- The scale and variety of data that is available today can overwhelm any data practitioner and that is why it is important to make data accessibility simple and convenient for brand mangers and owners.

4) **Need for synchronization across data sources:**

- As data sets become more diverse, there is a need to incorporate them into an analytical platform.
- If this is ignored, it can create gaps and lead to wrong insights and messages.

5) **Getting important insights through the use of Big data analytics:**

- It is important that companies gain proper insights from big data analytics and it is important that the correct department has access to this information.
- A major challenge in the big data analytics is bridging this gap in an effective fashion.

Other Three challenges of Conventional systems

Three Challenges That big data face.

1. Data
2. Process
3. Management

1. Data Challenges

Volume

1. The volume of data, especially machine-generated data, is exploding,
2. how fast that data is growing every year, withnew sources of data that are emerging.
3. For example, in the year 2000, 800,000petabytes (PB) of data were stored in the world, and it is expected to reach 35 zetta bytes (ZB) by2020 (according to IBM).

Social media plays a key role: Twitter generates 7+ terabytes (TB) of data every day. Facebook, 10 TB.
•Mobile devices play a key role as well, as there were estimated 6 billion mobile phones in 2011.
•The challenge is how to deal with the size of Big Data.

Variety, Combining Multiple Data Sets
•More than 80% of today's information is unstructured and it is typically too big to manage effectively.
•Today, companies are looking to leverage a lot more•data from a wider variety of sources both inside and outside the organization.
•Things like documents, contracts, machine data, sensor data, social media, health records, emails, etc. The list is endless really.

Variety•A lot of this data is unstructured, or has a complex structure that's hard to represent in rows and columns.

2. Processing

- More than 80% of today's information isunstructured and it is typically too big to manage effectively.
- Today, companies are looking to leverage a lot more data from a wider variety of sources both inside and outside the organization.
- Things like documents, contracts, machine data, sensor data, social media, health records, emails, etc. The list is endless really.

3. Management

- A lot of this data is unstructured, or has acomplex structure that’s hard to represent in rows and columns.

Big Data Challenges

  - The challenges include capture, duration, storage, search, sharing, transfer,
  - analysis, and visualization.

- Big Data is trend to larger data sets
- due to the additional information derivable from analysis of a single large set of related data,
  - as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to
    - "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions.”

Challenges of Big Data

The following are the five most important challenges of the Big Data

a) Meeting the need for speed

In today’s hypercompetitive business environment, companies not only have to find and analyze the relevant data they need, they must find it quickly.

b) Visualization helps organizations perform analyses and make decisions much more rapidly, but the challenge is going through the sheer volumes of data and accessing the level of detail needed, all at a high speed.

c) The challenge only ***grows as the degree of granularity increases***. One possible solution is hardware. Some vendors are using increased memory and powerful parallel processing to crunch large volumes of data extremely quickly

d) Understanding the data

- It takes a lot of understanding to get data in the **RIGHT SHAPE** so that you can use
- visualization as part of data analysis.

d) Addressing data quality

- Even if you can find and analyze data quickly and put it in the proper context for the
- audience that will be consuming the information, the value of data for **DECISION-MAKING PURPOSES** will be jeopardized if the data is not accurate or timely.

This is a challenge with any data analysis.

e) Displaying meaningful results

- Plotting points on a graph for analysis becomes difficult when dealing with extremely
- large amounts of information or a variety of categories of information.
- For example, imagine you have 10 billion rows of retail SKU data that you're trying to
- compare. The user trying to view 10 billion plots on the screen will have a hard time
- seeing so many data points.
- . By grouping the data together, or "binning," you can more effectively visualize the data.

f) Dealing with outliers

- The graphical representations of data made possible by visualization can communicate
- trends and outliers much faster than tables containing numbers and text.
- Users can easily spot issues that need attention simply by glancing at a chart. Outliers typically represent about 1 to 5 percent of data, but when you're working with massive amounts of data, viewing 1 to 5 percent of the data is rather difficult
- We can also bin the results to both view the distribution of data and see the outliers.
- While outliers may not be representative of the data, they may also reveal previously
- unseen and potentially valuable insights.

- Visual analytics enables organizations to take raw data and present it in a meaningful way that generates the most value. However, when used with big data, visualization is bound to lead to some challenges.

****

1.3. INTELLIGENT DATA ANALYSIS

1.3.1 INTRODUCTION TO INTELLIGENT DATA ANALYSIS (IDA)

Intelligent Data Analysis (IDA) is one of the hot issues in the field of artificial **intelligence** and information.

What is **Intelligent Data Analysis** (IDA)?

IDA is

… an interdisciplinary study concerned with the effective analysis of data;

… used for extracting useful information from large quantities of online data; extractingdesirable knowledge or interesting patterns from existing databases;

- *the distillation of information that has been collected, classified, organized, integrated, abstracted and value-added;*
- *at a level of abstraction higher than the data, and information on which it is based and can be used to deduce new information and new knowledge;*
- *usually in the context of human expertise used in solving problems.*
- *the distillation of information that has been collected, classified, organized, integrated, abstracted and value-added;*
- *at a level of abstraction higher than the data, and information on which it is based and can be used to deduce new information and new knowledge;*
- *usually in the context of human expertise used in solving problems.*

Goal:

Goal of Intelligent data analysis is to extract useful knowledge, the process demands a combination of extraction, **analysis**, conversion, classification, organization, reasoning, and so on.

1,3,2 Uses / Benefits of IDA

Intelligent Data Analysis provides a forum for the examination of issues related to the research and applications of Artificial Intelligence techniques in data analysis across a variety of disciplines and the techniques include (but are not limited to):

The benefit areas are:

- Data Visualization
- Data pre-processing (fusion, editing, transformation, filtering, sampling)
- Data Engineering
- Database mining techniques, tools and applications
- Use of domain knowledge in data analysis
- Big Data applications
- Evolutionary algorithms
- Machine Learning(ML)
- Neural nets
- Fuzzy logic
- Statistical pattern recognition
- Knowledge Filtering and
- Post-processing

Intelligent Data Analysis (IDA)

Why IDA?

- ➢ Decision making is asking for information and knowledge
- ➢ Data processing can give them
- ➢ Multidimensionality of problems is looking for methods for adequate and deep data processing and analysis
- ➢ Epidemiological study (1970-1990)

- Sample of examinees died from cardiovascular diseases during the period
- **Question:** Did they know they were ill?

  1 – they were healthy

  2 – they were ill (drug treatment, positive clinical and laboratory findings)

1.3.4 Intelligent Data Analysis

Knowledge Acquisition

- *The process of eliciting, analyzing, transforming, classifying, organizing and integrating knowledge and representing that knowledge in a form that can be used in a computer system.*

Knowledge in a domain can be expressed as a number of rules

A Rule :

A formal way of specifying a recommendation, directive, or strategy, expressed as "IF premise THEN conclusion" or "IF condition THEN action".

How to discover rules hidden in the data?

1.3.4 Intelligent Data Examples:

Example of IDA

- **Epidemiological study (1970-1990)**
- **Sample of examinees died from cardiovascular diseases during the period**

Question: Did they know they were ill?

1 – they were healthy

2 – they were ill (drug treatment, positive clinical and laboratory findings)

Illustration of IDA by using See5

- **application.*names* - lists the *classes* to which cases may belong and the *attributes* used to describe each case.**
- **Attributes are of two types: *discrete* attributes have a value drawn from a set of possibilities, and *continuous* attributes have numeric values.**

- **application.*data* - provides information on the *training* cases from which See5 will extract patterns.**
- **The entry for each case consists of one or more lines that give the values for all attributes.**

- **application.*data* - provides information on the *training* cases from which See5 will extract patterns.**
- **The entry for each case consists of one or more lines that give the values for all attributes.**

- **application.*test* - provides information on the *test* cases (used for evaluation of results).**
- **The entry for each case consists of one or more lines that give the values for all attributes.**

Goal 1.1 :

application.*names* – example

gender:M,F
activity:1,2,3
age: continuous
smoking: No, Yes
…

Goal:1,2 :

application.*data* – example

M,1,59,Yes,0,0,0,0,119,73,103,86,247,87,15979,?,?,?,1,73,2.5

M,1,66,Yes,0,0,0,0,132,81,183,239,?,783,14403,27221,19153,23187,1,73,2.6

M,1,61,No,0,0,0,0,130,79,148,86,209,115,21719,12324,10593,11458,1,74,2.5

… …

Result:

Results – example

Rule 1: (cover 26)
gender = M
SBP > 111
oil_fat > 2.9
-> class 1 [0.929]

Rule 1: (cover 26)
gender = M
SBP > 111
oil_fat > 2.9
-> class 1 [0.929]

Rule 4: (cover 14)
smoking = Yes
SBP > 131
glucose > 93
glucose <= 118
oil_fat <= 2.9
-> class 2 [0.938]

Rule 15: (cover 2)
SBP <= 111
oil_fat > 2.9

```
-> class 2 [0.750]
```

Evaluation on training data

```
(199 cases):
  (a)   (b)      <-classified as
 ----  ----
  107     3      (a): class 1
   17    72      (b): class 2
```

Results on (training set):

Sensitivity=0.97
Specificity=0.81

Sensitivity=0.97
Specificity=0.81

Sensitivity=0.98
Specificity=0.90

Evaluation of IDA results

- **Absolute & relative accuracy**
- **Sensitivity & specificity**
- **False positive & false negative**
- **Error rate**
- **Reliability of rules**
- **Etc.**

****

1. 4 NATURE OF DATA

1.4.1 INTRODUCTION

Data

- **Data** is a set of values of qualitative or quantitative variables; restated, pieces of **data** are individual pieces of information.
- **Data** is measured, collected and reported, and analyzed, whereupon it can be visualized using graphs or images.

Properties of Data
For examining the properties of data, reference to the various definitions of data.

Reference to these definitions reveals that following are the properties of data:

a) Amenability of use
b) Clarity
c) Accuracy
d) Essence
e) Aggregation
f) Compression
g) Refinement

.

a) **Amenability of use:** From the dictionary meaning of data it is learnt that data are facts used in deciding something. In short, data are meant to be used as a base for arriving at definitive conclusions.

b) **Clarity:** Data are a crystallized presentation. Without clarity, the meaning desired to be communicated will remain hidden.

c) **Accuracy:** Data should be real, complete and accurate. Accuracy is thus, an essential property of data.

d) **Essence:** A large quantities of data are collected and they have to be Compressed and refined. Data so refined can present the essence or derived qualitative value, of the matter.

e) **Aggregation:** Aggregation is cumulating or adding up.

f) **Compression:** Large amounts of data are always compressed to make them more meaningful. Compress data to a manageable size.Graphs and charts are some examples of compressed data.

g) **Refinement:** Data require processing or refinement. When refined, they are capable of leading to conclusions or even generalizations. Conclusions can be drawn only when data are processed or refined.

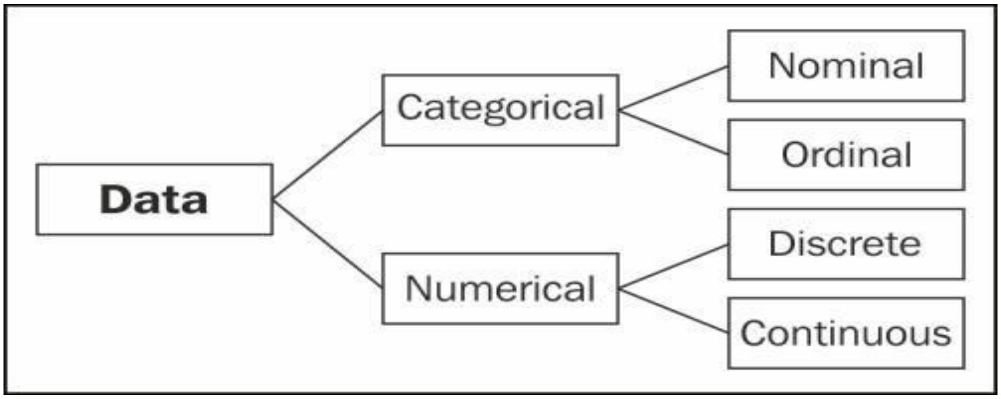
1.4.2 TYPES OF DATA

- In order to understand the nature of data it is necessary to categorize them into various types.
- Different categorizations of data are possible.
- The first such categorization may be on the basis of disciplines, e.g., Sciences, Social Sciences, etc. in which they are generated.
- Within each of these fields, there may be several ways in which data can be categorized into types.

There are four types of data:

- Nominal
- Ordinal
- Interval
- Ratio

Each offers a unique set of characteristics, which impacts the type of analysis that can be performed.



The distinction between the four types of scales center on three different characteristics:

1. The **order** of responses – whether it matters or not
2. The **distance between observations** – whether it matters or is interpretable
3. The presence or inclusion of a **true zero**

1.4.2.1 Nominal Scales

Nominal scales measure categories and have the following characteristics:

- **Order:** The order of the responses or observations does not matter.
- **Distance:** Nominal scales do not hold distance. The distance between a 1 and a 2 is not the same as a 2 and 3.
- **True Zero:** There is no true or real zero. In a nominal scale, zero is uninterruptable.

Appropriate statistics for nominal scales: mode, count, frequencies

Displays: histograms or bar charts

1.4.2.2 Ordinal Scales

At the risk of providing a tautological definition, ordinal scales measure, well, order. So, our characteristics for ordinal scales are:

- **Order:** The order of the responses or observations matters.
- **Distance:** Ordinal scales do not hold distance. The distance between first and second is unknown as is the distance between first and third along with all observations.

- **True Zero:** There is no true or real zero. An item, observation, or category cannot finish zero.

Appropriate statistics for ordinal scales: count, frequencies, mode

Displays: histograms or bar charts

1.4.2.3 Interval Scales

Interval scales provide insight into the variability of the observations or data.

Classic interval scales are Likert scales (e.g., 1 - strongly agree and 9 - strongly disagree) and Semantic Differential scales (e.g., 1 - dark and 9 - light).

In an interval scale, users could respond to "I enjoy opening links to thwebsite from a company email" with a response ranging on a scale of values.

The characteristics of interval scales are:

- **Order:** The order of the responses or observations does matter.
- **Distance:** Interval scales do offer distance. That is, the distance from 1 to 2 appears the same as 4 to 5. Also, six is twice as much as three and two is half of four. Hence, we can perform arithmetic operations on the data.
- **True Zero:** There is no zero with interval scales. However, data can be rescaled in a manner that contains zero. An interval scales measure from 1 to 9 remains the same as 11 to 19 because we added 10 to all values. Similarly, a 1 to 9 interval scale is the same a -4 to 4 scale because we subtracted 5 from all values. Although the new scale contains zero, zero remains uninterruptable because it only appears in the scale from the transformation.

Appropriate statistics for interval scales: count, frequencies, mode, median, mean, standard deviation (and variance), skewness, and kurtosis.

Displays: histograms or bar charts, line charts, and scatter plots.

1.4.2.4 Ratio Scales

Ratio scales appear as nominal scales with a true zero.

They have the following characteristics:

- **Order:** The order of the responses or observations matters.
- **Distance:** Ratio scales do do have an interpretable distance.

- **True Zero:** There is a true zero.

Income is a classic example of a ratio scale:

- Order is established. We would all prefer $100 to $1!
- Zero dollars means we have no income (or, in accounting terms, our revenue exactly equals our expenses!)
- Distance is interpretable, in that $20 appears as twice $10 and $50 is half of a $100.

For the web analyst, the statistics for ratio scales are the same as for interval scales.

Appropriate statistics for ratio scales: count, frequencies, mode, median, mean, standard deviation (and variance), skewness, and kurtosis.

Displays: histograms or bar charts, line charts, and scatter plots.

The table below summarizes the characteristics of all four types of scales.

	Nominal	Ordinal	Interval	Ratio
Order Matters	No	Yes	Yes	Yes
Distance Is Interpretable	No	No	Yes	Yes
Zero Exists	No	No	No	Yes

1.4.3 DATA CONVERSION

- We can convert or transform our data from **ratio** to **interval** to **ordinal** to **nominal**. However, we *cannot* convert or transform our data from **nominal** to **ordinal** to **interva**l to **ratio.**

- ***Scaled data** can be measured in **exact amounts.***
  For example, 60 degrees , 12.5 feet, 80 Miles per hour

- ***Scaled data** can be measured w**ith equal intervals.***
  For example, Between **0** and **1** is **1 inch,** Between **13** and **14** is also **1 inch**

- ***.Ordinal or ranked data*** *provides* ***comparative Amounts***

  Example:

  1st Place **2nd Place** **3rd Place**

- ***Not equal intervals***

1st Place	**2nd Place**	**3rd Place**
19.6 feet	**18.2 feet**	**12.4 feet**

1.4.4 DATA SELECTION

Another Example that handle the question as :

What is the average driving **speed** of teenagers on the freeway?

a) Scaled
b) Ordinal

Scaled – Speed:- Speed *can be measured in* ***exact amounts withequal intervals.***

Example :

60 degrees **12.5 feet** **80 Miles per hour**

- ***Ordinal or ranked data*** *provides* ***comparative amounts.***

For example, **1st Place** **2nd Place** **3rd Place**

- ***Percentiles*** *provide* ***comparative amounts.***

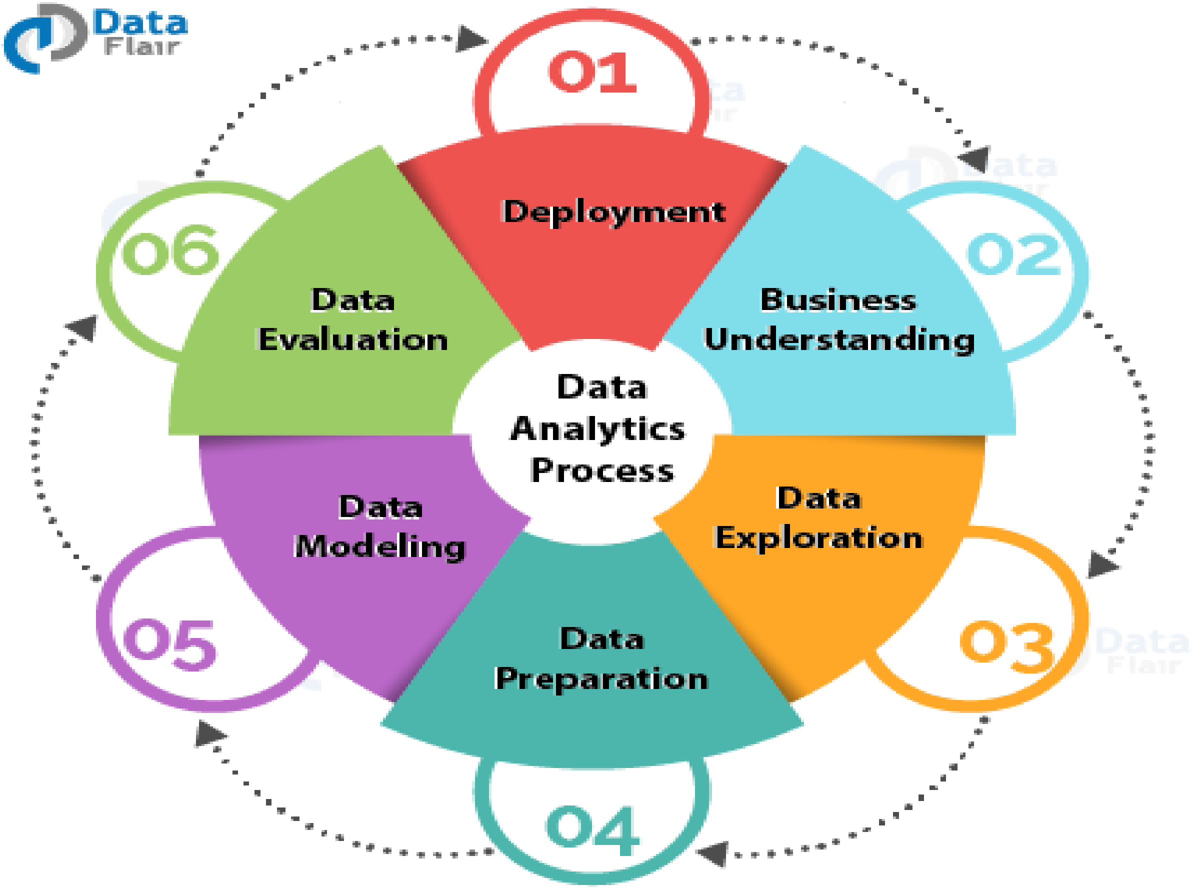
In this case, 93% of all hospital have lower patient satisfaction scores than Eastridge hospital. 31% have lower satisfaction scores than Westridge Hospital.

Thus the nature of data and its value have great influence on data insight in it.

****

5. ANALYTIC PROCESS AND TOOLS

- **There are 6 analytic processes:**
  1. Deployment
  2. Business Understanding
  3. Data Exploration
  4. Data Preparation
  5. Data Modeling
  6. Data Evaluation



Step 1: Deployment

- Here we need to:
  - plan the deployment and monitoring and maintenance,
  - we need to produce a final report and review the project.
  - In this phase,
    - we deploy the results of the analysis.
    - This is also known as reviewing the project**.**

Step 2: Business Understanding

- **Business Understanding**
  - The very first step consists of business understanding.
  - Whenever any requirement occurs, firstly we need to determine the business objective,
  - assess the situation,
  - determine data mining goals and then
  - produce the project plan as per the requirement.
- Business objectives are defined in this phase.

Step 3: Data Exploration

- The second step consists of Data understanding.
  - For the further process, we need to gather initial data, describe and explore the data and verify data quality to ensure it contains the data we *require*.

  - Data collected from the various sources is described in terms of its application and the need for the project in this phase.
  - This is also known as data exploration.
- This is necessary to verify the quality of data collected.

Step 4: Data Preparation

- From the *data collected in the last step,*
  - we need to *select data as per the need, clean it, construct it to get useful information* and
  - then *integrate it all*.
- Finally, we need to *format the data to get the appropriate data.*
- Data is selected, cleaned, and integrated into the format finalized for the analysis in this phase.

Step 5: Data Modeling

- we need to
  - select a modeling technique, generate test design, build a model and assess the model built.
- The data model is build to
  - analyze relationships between various selected objects in the data,
  - test cases are built for assessing the model and model is tested and implemented on the data in this phase.

- Where processing is hosted?
  - Distributed Servers / Cloud (e.g. Amazon EC2)
- Where data is stored?
  - Distributed Storage (e.g. Amazon S3)
- What is the programming model?
  - Distributed Processing (e.g. MapReduce)

- How data is stored & indexed?
  - High-performance schema-free databases (e.g. MongoDB)
- What operations are performed on data?
  - Analytic / Semantic Processing

- Big data tools for HPC and supercomputing
  - MPI
- Big data tools on clouds
  - MapReduce model
  - Iterative MapReduce model
  - DAG model
  - Graph model
  - Collective model
- Other BDA tools
  - SaS
  - R
  - Hadoop

Thus the BDA tools are used through out the BDA applications development.

****

1.6 ANALYSIS AND REPORTING

1.6.1 INTRODUCTION TO ANALYSIS AND REPORTING

What is Analysis?

- *The process of exploring data and reports*
  - *in order **to extract meaningful insights**,*
  - *which can be **used to better understand and improve business performance.***

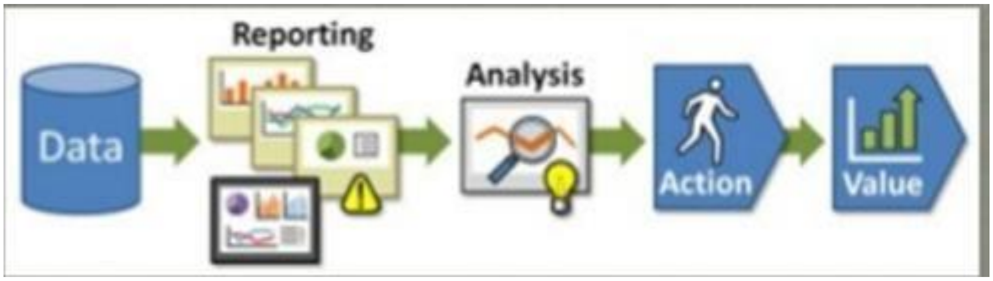
- **What is Reporting ?**
- **Reporting** is

- “the **process of organizing data**
- **into informational summaries**
- in order **to monitor how different areas of a business are performing**.”

1.6.2 COMPARING ANALYSIS WITH REPORTING

- **Reporting** is “the process of organizing data into informational summaries in order to monitor how different areas of a business are performing.”
    - Measuring core metrics and presenting them — whether in an email, a slidedeck, or online dashboard — falls under this category.
- **Analytics** is “the process of exploring data and reports in order to extract meaningful insights, which can be used to better understand and improve business performance.”
- Reporting helps companies to monitor their online business and be alerted to when data falls outside of expected ranges.
- **Good reporting**
    - should **raise questions** about the business from its end users.
- The **goal of analysis** is
    - to **answer questions** by interpreting the data at a deeper level and providing actionable recommendations.

- A firm may be **focused on the general area of analytics** (strategy, implementation, reporting, etc.)
    - but not necessarily on the specific aspect of analysis.
- It’s almost like **some organizations run out of gas after the initial set-up-related activities and don’t make it to the analysis stage**



A reporting activity deliberately proposes Analysis activity.

1.6.3 CONTRAST BETWEEN ANALYSIS AND REPORTING

The basis differences between Analysis and Reporting are as follows:

Analysis	**Reporting**
Provides what is needed	Provides what is asked for
Is typically customized	Is Typically standardized
Involves a person	Does not involve a person
Is extremely flexible	Is fairly Inflexible

- Reporting translates raw data into **information**.
- Analysis transforms data and information into **insights**.
- reporting shows you ***what is happening***
- while analysis focuses on explaining ***why it is happening*** and ***what you can do about it*.**

- Reports are like Robots n monitor and alter you and where as analysis is like parents - c an figure out what is going on (hungry, dirty diaper, no pacifier, , teething, tired, ear infection, etc).
- Reporting and analysis can go hand-in-hand:
- Reporting provides no limited context about what is happening in the data. Context is critical to good analysis.
- Reporting translate a raw data into information
- Reporting usually raises a question – **What is happening ?**
- Analysis transforms the data into insights **- Why is it happening ? What you can do about it?**

Thus, Analysis and Reporting is synonym to each other with respect their need and utilizing in the needy context.

****

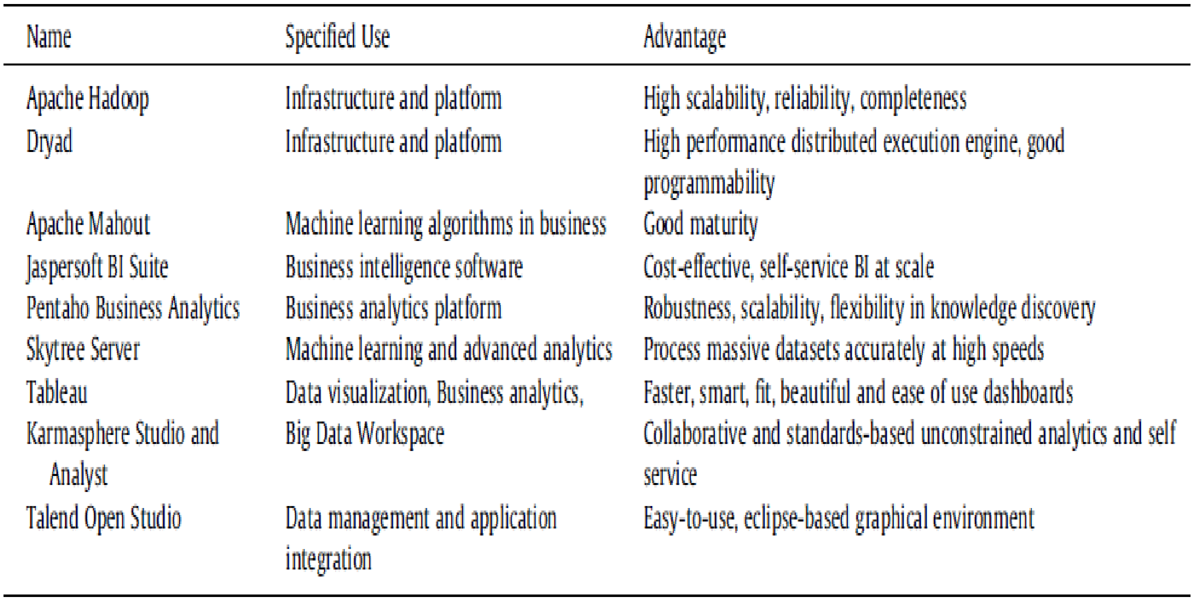
1.7 MODERN ANALYTIC TOOLS

1.7.1 Introduction to Modern Analytic Tools

- **Modern Analytic Tools:** Current Analytic tools concentrate on three classes:
  a) Batch processing tools
  b) Stream Processing tools and
  c) Interactive Analysis tools.

a) **Big Data Tools Based on Batch Processing:**
Batch processing system :-

- Batch Processing System involves
  – collecting a series of processing jobs and carrying them out periodically as a group (or batch) of jobs.
- It allows a large volume of jobs to be processed at the same time.
- An organization can schedule batch processing for a time when there is little activity on their computer systems, for example overnight or at weekends.
- One of the **most famous and powerful batch process-based Big Data tools is Apache Hadoop**.
  - It provides infrastructures and platforms for other specific Big Data applications.



Name	Specified Use	Advantage
Apache Hadoop	Infrastructure and platform	High scalability, reliability, completeness
Dryad	Infrastructure and platform	High performance distributed execution engine, good programmability
Apache Mahout	Machine learning algorithms in business	Good maturity
Jaspersoft BI Suite	Business intelligence software	Cost-effective, self-service BI at scale
Pentaho Business Analytics	Business analytics platform	Robustness, scalability, flexibility in knowledge discovery
Skytree Server	Machine learning and advanced analytics	Process massive datasets accurately at high speeds
Tableau	Data visualization, Business analytics,	Faster, smart, fit, beautiful and ease of use dashboards
Karmasphere Studio and Analyst	Big Data Workspace	Collaborative and standards-based unconstrained analytics and self service
Talend Open Studio	Data management and application integration	Easy-to-use, eclipse-based graphical environment

b) Stream Processing tools

- Stream processing – Envisioning (predicting) the life in data as and when it transpires.
- The key strength of stream processing is that **it can provide insights faster, often within milliseconds to seconds**.
    - It **helps understanding the hidden patterns in millions of data records in real time**.
    - It **translates into processing of data from single or multiple sources**
    - in real or near-real time applying the desired business logic and emitting the processed information to the sink.
- Stream processing serves
    - multiple
    - resolves in today's business arena.

Real time data streaming tools are:

a) Storm

- Storm is a ***stream processing engine without batch support***,
- a true ***real-time processing framework***,
- taking in a stream as an entire 'event' instead of series of small batches.
- *Apache Storm is a distributed real-time computation system.*
- It's *applications are designed as directed acyclic graphs.*

b) Apache flink

- Apache flink is
    - an open source platform
    - which **is a streaming data flow engine** that **provides communication fault tolerance** and
    - **data distribution computation over data stream .**
    - flink is a top level project of Apache flink is scalable data analytics framework that is fully compatible to hadoop .
    - flink can **execute both stream processing and batch processing easily.**
    - flink was **designed as an alternative to map-reduce**.

c) Kinesis

- Kinesis as an out of the box **streaming data tool.**
- Kinesis **comprises of shards which Kafka calls partitions.**

- For organizations that take **advantage of real-time or near real-time access to large stores of data**,
    - Amazon Kinesis is great.
- Kinesis Streams solves a variety of streaming data problems.
- One common use is **the real-time aggregation of data which is followed by loading the aggregate data into a data warehouse.**
- Data is put into Kinesis streams.
    - This ensures durability and elasticity.

c) Interactive Analysis -Big Data Tools

- The **interactive analysis presents**
    - the **data in an interactive environment**,
    - allowing users to undertake their own analysis of information.
- **Users are directly connected to**
    - the computer and hence can interact with it in real time.
- **The data can be** :
    - reviewed,
    - compared and
    - analyzed
- ***in tabular or graphic format or both at the same time***.

IA -Big Data Tools -

a) **Google's Dremel is the g**oogle proposed an interactive analysis system in 2010. And named named Dremel.
  - which is **scalable for processing nested data**.
  - Dremel provides
    - **a very fast SQL** like interface to the data by using a different technique than MapReduce.
- Dremel has a **very different architecture**:
    - **compared with well-known Apache Hadoop**, and
    - **acts as a successful complement of Map/Reduce-**based computations.

- **Dremel** has capability to:
    - ***run aggregation queries over trillion-row tables in seconds***
    - by means of:

- combining multi-level execution trees and
- columnar data layout.

b) Apache drill

- **Apache drill** is:
  - Drill is **an Apache open-source SQL query engine for Big Data exploration.**
  - It is similar to Google's Dremel.
- For Drill, there is:
  - **more flexibility to support**
    - **a various different query languages,**
    - **data formats and**
    - **data sources**.
- Drill is designed from the **ground up to:**
  - **support high-performance analysis** on the semi-structured and
  - **rapidly evolving data coming from modern Big Data applications.**
- Drill **provides plug-and-play integration with existing Apache Hive and Apache HBase deployments.**

7.1.2 Categories of Modern Analytic Tools

a) Big data tools for HPC and supercomputing

- MPI

b) Big Data Tools for HPC and Supercomputing

- MPI(Message Passing Interface, 1992)
  - Provide standardized function interfaces for communication between parallel processes.
- **Collective communication operations**
  - Broadcast, Scatter, Gather, Reduce, Allgather, Allreduce, Reduce-sc**Popular implementations**
  - atter.
  - MPICH (2001)
  - OpenMPI (2004)

c) Big data tools on clouds

i. MapReduce model
ii. Iterative MapReduce model
iii. DAG model

iv. Graph model
v. Collective model

a) **MapReduce Model**

Jeffrey Dean et al. MapReduce: Simplified Data Processing on Large Clusters. OSDI 2004.

b) **Apache Hadoop (2005)**

Apache Hadoop YARN: Yet Another Resource Negotiator, SOCC 2013.

Key Features of MapReduce Model

Designed for clouds

Large clusters of commodity machines

Designed for big data

Support from local disks based distributed file system (GFS / HDFS)

Disk based intermediate data transfer in Shuffling

MapReduce programming model

Computation pattern: Map tasks and Reduce tasks

Data abstraction: KeyValue pairs

Iterative MapReduce Model

- **Twister:**

  A runtime for iterative MapReduce.

  Have simple collectives: Boradcasting and aggregation.

- **HaLoop**
- An efficient Data Processing on Large clusters
- **Have features**:
  - Loop-Aware Task Scheduling
  - Caching and Indexing for Loop-Invariant Data on local disk.

Resilient Distributed Datasets(RDD):

- A Fault-Tolerant Abstraction for In-Memory Cluster Computing

– **RDD operations**

- MapReduce-like parallel operations

– DAG of execution stages and pipelined transformations

– **Simple collectives**: broadcasting and aggregation

DAG (Directed Acyclic Graph) Model

  - **A Distributed Data-Parallel Programs from Sequential Building Blocks,**
- **Apache Spark**
  - **Cluster Computing with Working Sets**

Graph Model

- **Graph Processing with BSP model**
- **Pregel (2010)**
  - A System for Large-Scale Graph Processing. SIGMOD 2010.
  - Apache Hama (2010)
- **Apache Giraph (2012)**
  - Scaling Apache Giraph to a trillion edges

Pregel & Apache Giraph

- **Computation Model**
  - Superstep as iteration
  - Vertex state machine:
    Active and Inactive, vote to halt
  - Message passing between vertices
  - Combiners
  - Aggregators
  - Topology mutation
- Master/worker model
- Graph partition: hashing
- Fault tolerance: checkpointing and confined recovery

GraphLab (2010)

- **GraphLab:** A New Parallel Framework for Machine Learning. UAI 2010.
- Distributed GraphLab: A Framework for Machine Learning and Data Mining in the Cloud.
- Data graph
- Update functions and the scope

- **PowerGraph (2012)**
  - **PowerGraph: Distributed Graph-Parallel Computation on Natural Graphs.**
  - **Gather, apply, Scatter (GAS) model**
- **GraphX (2013)**
  - **A Resilient Distributed Graph System on Spark. GRADES**

Collective Model

- **Harp (2013)**
  - **A Hadoop Plugin (on Hadoop 1.2.1 and Hadoop 2.2.0)**
  - Hierarchical data abstraction on arrays, key-values and graphs for easy programming expressiveness.
  - Collective communication model to support various communication operations on the data abstractions.
  - Caching with buffer management for memory allocation required from computation and communication
  - BSP style parallelism
  - Fault tolerance with check-pointing.

Other major Tools

a) AWS
b) BigData
c) Cassandra
d) Data Warehousing
e) DevOps
f) HBase
g) Hive
h) MongoDB
i) NiFi
j) Tableau
k) Talend
l) ZooKeeper

Thus the modern analytical tools play an important role in the modern data world.

****

1.8. STATISTICAL CONEPTS: SAMPLING DISTRIBUTIONS

1.8.1 Fundamental Statistics

- Statistics is a very broad subject, with applications in a vast number of different fields.
- In generally one can say that statistics is the methodology for collecting, analyzing, interpreting and drawing conclusions from information.
- Putting it in other words, statistics is the methodology which scientists and mathematicians have developed for interpreting and drawing conclusions from collected data.
- Everything that deals even remotely with the collection, processing, interpretation and presentation of data belongs to the domain of statistics, and so does the detailed planning of that precedes all these activities.

1.8.1.1 Data and Statistics

- **Data consists of**
  - **information coming from observations, counts, measurements, or responses.**

Statistics is :

-science of collecting, organizing, analyzing, and interpreting data in order to make decisions.

A population is:

- the collection of *all* outcomes, responses, measurement, or counts that are of interest.

A sample is:

- a subset of a population.

- **All data can be classified as either,**
  - 1) Categorical or
  - 2) Quantitative.

- **there exist deferent ways to summaries each type of data;**
  - **for example an average measure of eye colour is nonsensical.**

What is statistics?

Definition 1.1 (Statistics). Statistics consists of a body of methods for collecting and analyzing data. (Agresti & Finlay, 1997)

- **Statistics** is the science of data.
- **A bunch of numbers looking for a fight.**
- It involves:
  - collecting,
  - classifying,
  - summarizing,
  - organizing,
  - analyzing, and
  - interpreting numerical information.

- Statistics is much more than just the tabulation of numbers and the graphical presentation of these tabulated numbers.
- Statistics is the science of gaining information from numerical and categorical data.

- Statistical methods can be used to find answers to the questions like:
  - What kind and how much data need to be collected?
  - How should we organize and summarize the data?
  - How can we analyse the data and draw conclusions from it?
  - How can we assess the strength of the conclusions and evaluate their uncertainty?

- Categorical data (or qualitative data) results from descriptions, e.g. the blood type of person, marital status or religious affiliation.

Why Statistics ?

- Statistics is the science of dealing with uncertain phenomenon and events.

- Statistics in practice is applied successfully to study the effectiveness of medical treatments, the reaction of consumers to television advertising, the attitudes of young people toward sex and marriage, and much more.
- It's safe to say that nowadays statistics is used in every field of science.
- A statistic is a piece of data from a **portion of a population.**
- It's the opposite of a parameter.
- A parameter is data from a census.
- A census surveys *everyone*.
- If you have a bit of information, it's a statistic.
- If you look at part of a data set, it's a statistic.
- If you know something about 10% of people, that's a statistic too.
- Parameters are **all** the information.
- And all the information is rarely known.
- **That's why we need stats!**

Statistics provides methods for:
1. Design: Planning and carrying out research studies.
2. Description: Summarizing and exploring data.
3. Inference: Making predictions and generalizing about phenomena represented by the data.

Example 1.1 (Statistics in practice).
Consider the following problems:

- Agricultural problem: Is new grain seed or fertilizer more productive?
- Medical problem: What is the right amount of dosage of drug to treatment?
- Political science: How accurate are the gallups and opinion polls?
- Eeconomics: What will be the unemployment rate next year?
- Technical problem: How to improve quality of product?

1.8.2 Statistical Concepts

Basic statistical operations

- **Mean** A measure of central tendency for Quantitative data i.e. the long term average value.
- **Median** A measure of central tendency for Quantitative data i.e. the half-way point.
- **Mode** The most frequently occurring (discrete), or where the probability density

function peaks (contin- ious).

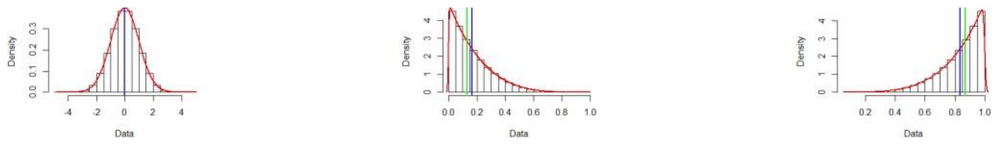
- **Minimum** The smallest value.
- **Maximum** The largest value.
- **Inter quartile range** Can be thought or as the *middle 50* of the (Quantitative) data, used as a measure of spread.
- **Variance -** Used as a measure of spread, may be thought of as the *moment of inertia*.
- **Standard deviation -** A measure of spread, the square root of the variance.

Data shape

Continuous data can be;

- **Symmetrical** Perfectly symmetrical about the mean, i.e. **Normally distributed** data.
- **Skewed** Right/Positively skewed i.e. long tail to the right of the *peak*, Left/Negitively skewed i.e. long tail to the left of the *peak* .
- **Multimodal** Multiple *peaks* in the data.

In addition when dealing with continuous data the shape of the data effects which summary statistic are most appropriate to use i.e heavily skewed data will pull the mean toward the tail of the data.



(A) Symmetrical (B) Rigt Skewed (c) Left Skewed

FIGURE 1. Displaying different data shapes with median, mean, density respectively.

1.8.3 Population and Sample

- Population and sample are two basic concepts of statistics.
- Population can be characterized as the set of individual persons or objects in which an investigator is primarily interested during his or her research problem.
- Sometimes wanted measurements for all individuals in the population are obtained, but often only a set of individuals of that population are observed; such a set of individuals constitutes a sample.
- This gives us the following definitions of population and sample.

Definition Population: Population is the collection of all individuals or items under consideration in a statistical study. (Weiss, 1999)

Definition Sample: Sample is that part of the population from which information is collected. (Weiss, 1999)

A population is:

- the collection of *all* outcomes, responses, measurement, or counts that are of interest.

A sample is:

- a subset of a population.

Example of Populations & Samples

Question: In a recent survey,

- ***250 college students at Union College*** were asked:
  - **if they smoked cigarettes regularly.**
- **Answer:**
  - 35 of the students said yes.
- ***Identify the population and the sample***.

- Responses of all students at Union College **(population)**
- Responses of students in survey **(sample)**

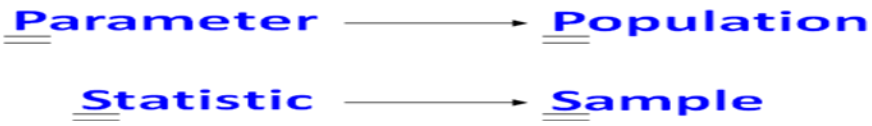
Parameters & Statistics

A parameter is

-a numerical description of a *population* characteristic.

A statistic is:

- a numerical description of a *sample* characteristic.



Statistical process

Statistics (*in action have*) Two Processes

1) **Describing** sets of data

2) Drawing conclusions

(making estimates, decisions, predictions, etc. about sets of data based on sampling)

What is a Statistic used for?

- **Statistics is**
  - a way to understand the data that is collected about us and the world.
- For example,
  - every time you send a package through the mail, that package is tracked in a huge database.
- All of that data is meaningless without a way to interpret it,
  - which is where statistics comes in.
- **Statistics is:**
  - about **data** and **variables**.
- It's also about:
  - **analyzing** that data and
  - producing some meaningful information about that data.

Other Statistical Definitions

Variable: -

- Any characteristic of an individual or entity.
- A variable can take different values for different individuals.

Variables can be *categorical* or *quantitative*.

- Per S. S. Stevens…

- Variables are influenced by their nature of values that it holds.
- **Nominal** - Categorical variables with no inherent order or ranking sequence such as names or classes (e.g., gender). Value may be a numerical, but without numerical value (e.g., I, II, III). The only operation that can be applied to Nominal variables is enumeration.
- **Ordinal** - Variables with an inherent rank or order, e.g. mild, moderate, severe. Can be compared for equality, or greater or less, but not *how much* greater or less.
- **Interval** - Values of the variable are ordered as in Ordinal, and additionally, differences between values are meaningful, however, the scale is not absolutely anchored. Calendar dates and temperatures on the Fahrenheit scale are examples. Addition and subtraction, but not multiplication and division are meaningful operations.
- **Ratio** - Variables with all properties of Interval plus an absolute, non-arbitrary zero point, e.g. age, weight, temperature (Kelvin). Addition, subtraction, multiplication, and division are all meaningful operations.

Why study statistics?

1. Data are everywhere.

2. Statistical techniques are used to make many decisions that affect our lives.

3. No matter what your career, you will make professional decisions that involve data.

4. An understanding of statistical methods will help you make these decisions efectively.

Applications of statistical concepts in the business world

- **Finance** – correlation and regression, index numbers, time series analysis
- **Marketing** – hypothesis testing, chi-square tests, nonparametric statistics
- **Personel** – hypothesis testing, chi-square tests, nonparametric tests
- **Operating management** – hypothesis testing, estimation, analysis of variance, time series analysis

Application Areas

- **Economics**
  - Forecasting
  - Demographics

- **Sports**
  - Individual & Team Performance

- **Engineering**
  - Construction
  - Materials
- **Business**
  - Consumer Preferences
  - Financial Trends

Statistics Vs Statistical analysis

- **Statistics :-** The science of
  - collectiong,
  - organizing,
  - presenting,
  - analyzing, and
  - interpreting data
- to assist in making more effective decisions.
- **Statistical analysis**: – used to
  - manipulate summarize, and
  - investigate data,
- so that useful decision-making information results.

Fundamental Elements of Statistics

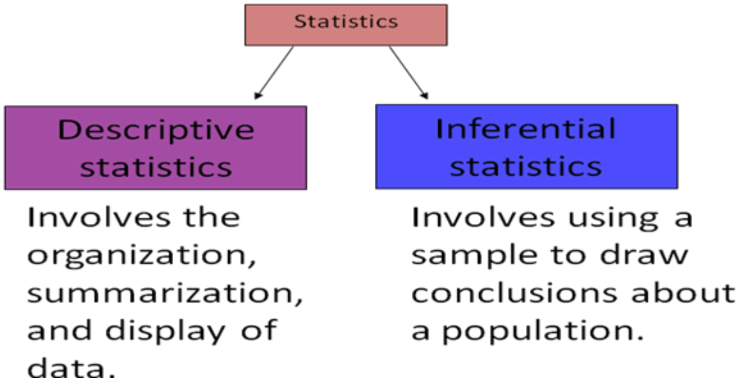
1. **Experimental unit**
   - Object upon which we collect data
2. **Population**
   - All items of interest
3. **Variable**
   - Characteristic of an individual experimental unit
4. **Sample**
   - Subset of the units of a population

- **P** in **P**opulation & **P**arameter
- **S** in **S**ample & **S**tatistic

5. **Statistical Inference**
   - Estimate or prediction or generalization about a population based on information contained in a sample
6. **Measure of Reliability**
   - Statement (usually qualified) about the degree of uncertainty associated with a statistical inference

1.8.4 Types or Branches of Statistics

The study of statistics has two major branches: **descriptive statistics** and **inferential statistics**.


Statistics
Descriptive statistics
Inferential statistics
Involves the organization, summarization, and display of data.
Involves using a sample to draw conclusions about a population.

- **Descriptive statistics:** –
  - Methods of organizing, summarizing, and presenting data in an informative way.
  - Involves: Collecting Data
    - Presenting Data
    - Characterizing Data
    - Purpose
    - Describe Data

- **Inferential statistics:** –
  - The methods used to determine something about a population on the basis of a sample:

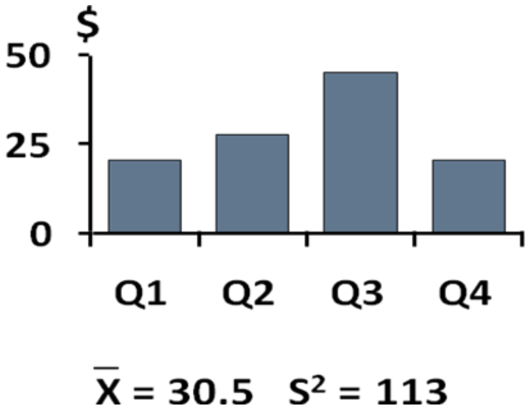
- **Population** –The entire set of individuals or objects of interest or the measurements obtained from all individuals or objects of interest
- **Sample** – A portion, or part, of the population of interest

Descriptive and Inferential Statistics

- **Example**:
- In a recent study, volunteers who had less than 6 hours of sleep were four times more likely to answer incorrectly on a science test than were participants who had at least 8 hours of sleep.
- Decide which part is the descriptive statistic and what conclusion might be drawn using inferential statistics?

Answer:-
The statement ***"four times more likely to answer incorrectly" is a descriptive statistic.***
An inference drawn from the sample is that

all individuals sleeping less than 6 hours are more likely to answer science question incorrectly than individuals who sleep at least 8 hours.



Inferential Statistics & Its' Techniques

- **Inference is the process of drawing conclusions or making decisions about a population based on sample results**

Involves

- **Estimation**
- **Hypothesis Testing**

Purpose

- **Make decisions about population characteristics**

- **Estimation:-**
  - e.g., Estimate the population mean weight using the sample mean weight
- **Hypothesis testing:-**
  - e.g., Test the claim that the population mean weight is 70 kg

Four Elements of Descriptive Statistical Problems

1. The population or sample of interest
2. One or more variables (characteristics of the population or sample units) that are to be investigated
3. Tables, graphs, or numerical summary tools
4. Identification of patterns in the data

Five Elements of Inferential Statistical Problems

1. The population of interest
2. One or more variables (characteristics of the population units) that are to be investigated
3. The sample of population units
4. The inference about the population based on information contained in the sample
5. A measure of reliability for the inference

1.8.5 SAMPLING DISTRIBUTION

1.8.5.1. Introduction to Sampling

- **What is sample?**
- A sample is “a smaller (but hopefully representative) collection of units from a population used to determine truths about that population” (Field, 2005)

Why sample?

  - Resources (time, money) and workload
  - Gives results with known accuracy that can be calculated mathematically
- The sampling frame is the list from which the potential respondents are drawn
  - Registrar’s office
  - Class rosters
  - Must assess sampling frame errors

Major Types of Samples

- **Probability (Random) Samples**
- Simple random sample
  - Systematic random sample
  - Stratified random sample
  - Multistage sample
  - Multiphase sample
  - Cluster sample
- **Non-Probability Samples**
  - Convenience sample
  - Purposive sample
  - Quota

Specific Types of Samples

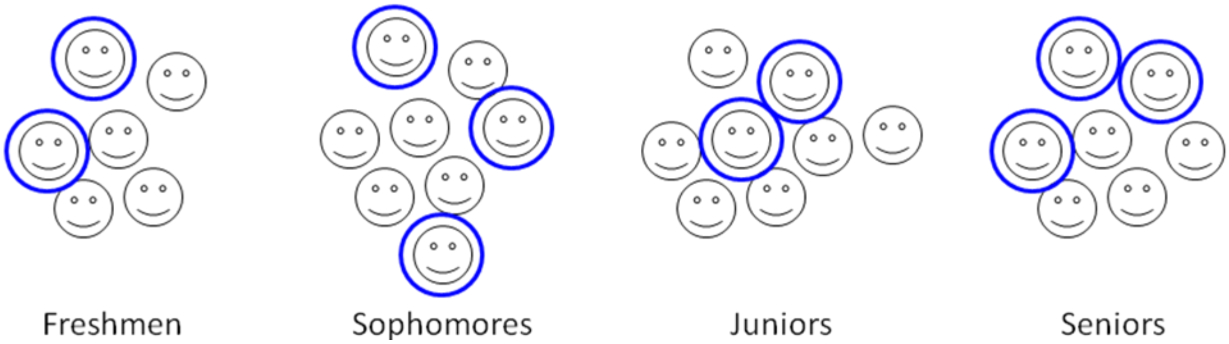
1. **Stratified Samples**
2. **Cluster Samples**
3. **Systematic Samples**
4. **Convenience Samples**

1. **Stratified Samples**

   A **stratified sample** has:

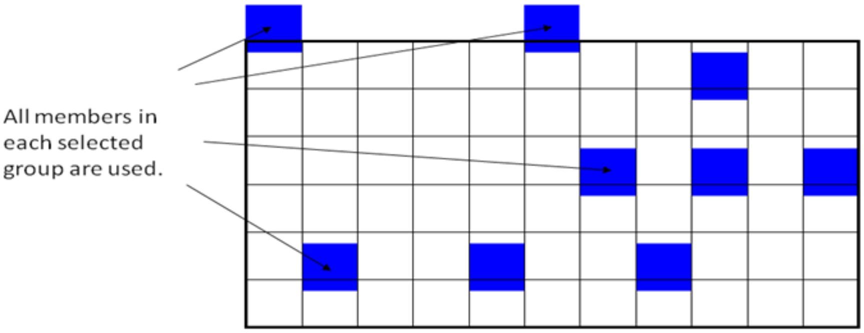
   i. *members from each segment of a population.*

   ii. This ensures that *each segment from the population is represented.*



2. **Cluster Samples**

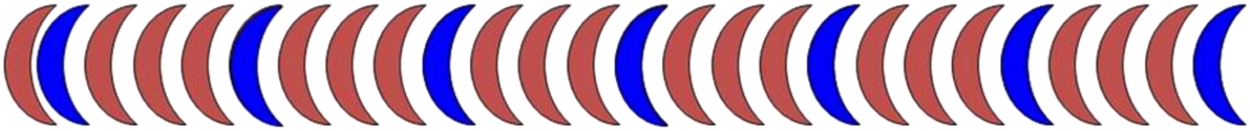
- A **cluster sample** has
  - *all members from randomly selected segments of a population.*
  - This is used when the population falls into naturally occurring subgroups.



The city of Clarksville divided into city blocks.

3. **Systematic Samples**

- A **systematic sample** is
  - a sample **<u>in which each member of the population is assigned a number.</u>**
- A starting number is
  - **<u>randomly selected and sample members are selected at regular intervals.</u>**



Every fourth member is chosen.

4. **Convenience Samples**

   A convenience sample <u>consists only of available members of the population.</u>

   Example:

   You are doing a study to determine the number of years of education each teacher at your college has.

   Identify the sampling technique used if you select the samples listed.

<u>1.8.5.2 Sampling Distribution</u>

What is Sampling Distribution?

- **<u>The way our means would be distributed</u>**
  - if we (1) ***<u>collected a sample</u>***,

- recorded the mean and
- threw it back, and
- (2) ***collected another***,
- recorded the mean and
- threw it back, and
- did this again and again….

- *ad nauseam*!

- From Vogt:

  A theoretical frequency distribution of the scores for or values of a statistic, such as a mean.

  - Any statistic that can be computed for a sample has a sampling distribution.

- A sampling distribution is:
  - the distribution of statistics
  - that *would be* produced
  - in repeated random sampling (with replacement) from the same population.

Glimpses of Sampling Distribution

- **Sampling distributions is**
  - *all possible values of a statistic* and
  - *their probabilities of occurring for a sample of a particular size.*
- **Sampling distributions are used to**
  - calculate the probability that sample statistics
  - could have occurred by chance and
  - thus to decide whether something that is true of a sample statistic is
    - also likely to be true of a population parameter.

A Positive move of Sampling Distribution

- We are moving **from descriptive statistics to inferential statistics.**
- Inferential statistics allow the researcher:
  - **to come to conclusions about a population**
  - **on the basis of descriptive statistics about a sample.**

Examples of Sampling Distribution

1) Your sample says that

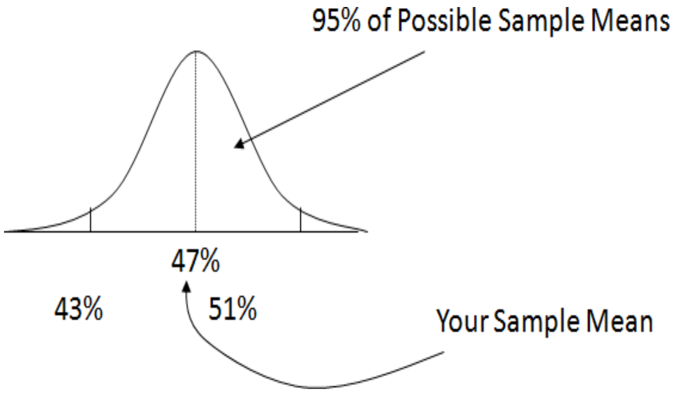
a candidate gets support from 47%.

2) **Inferential statistics allow you** to say that
   - (a) the candidate gets support from 47% of the population
   - (b) with a margin of error of +/- 4%.
   - This means that the support in the population is

likely somewhere between 43% and 51%.

Error on Sampling Distribution

- Margin of error is taken directly from a sampling distribution.
- It looks like this:



A Real Time Scenario on Sampling Distribution

Let's create a sampling distribution of means…

a) Take a sample of size 1,500 from the US.
b) Record the mean income.
c) Our census said the mean is $30K.

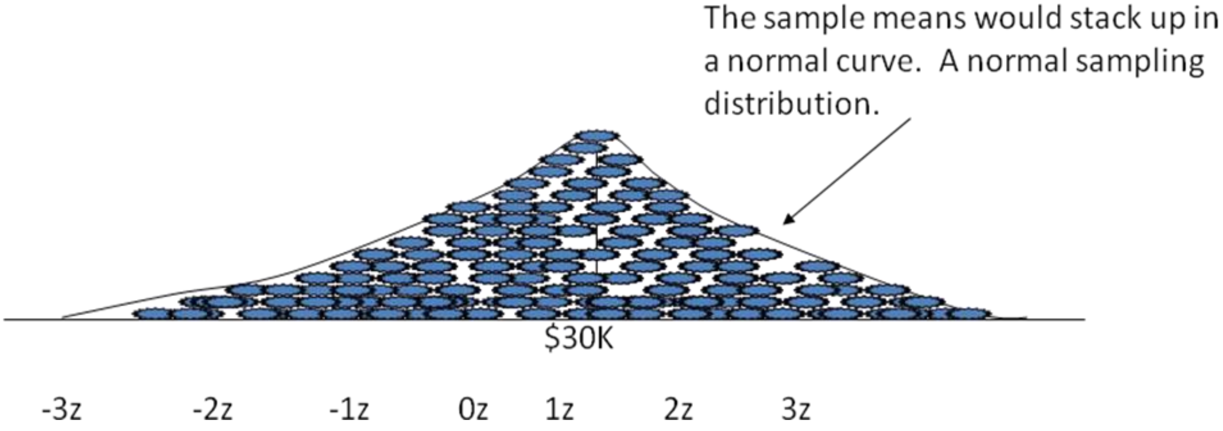
- Let's create a sampling distribution of means…

- Take another sample of size 1,500 from the US. Record the mean income. Our census said the mean is $30K.

- Take another sample of size 1,500 from the US. Record the mean income. Our census said the mean is $30K.

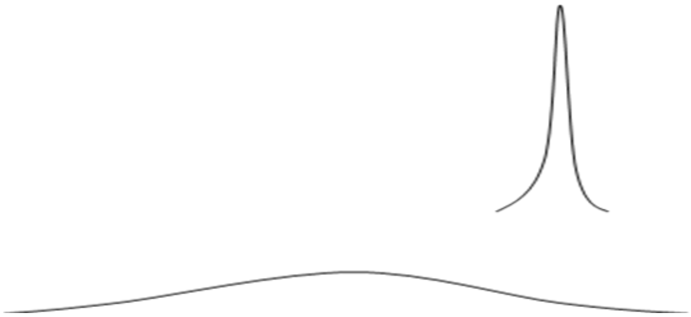
- Take another sample of size 1,500 from the US. Record the mean income. Our census said the mean is $30K.

- ….
- Say that the standard deviation of this distribution is $10K.
- Think back to the empirical rule. What are the odds you would get a sample mean that is more than $20K off.



The sample means would stack up in a normal curve. A normal sampling distribution.

Knowing the likely variability of the sample means from repeated sampling gives us a context within which to judge how much we can trust the number we got from our sample.
For example, if the variability is low, , we can trust our number more than if the variability is high, .

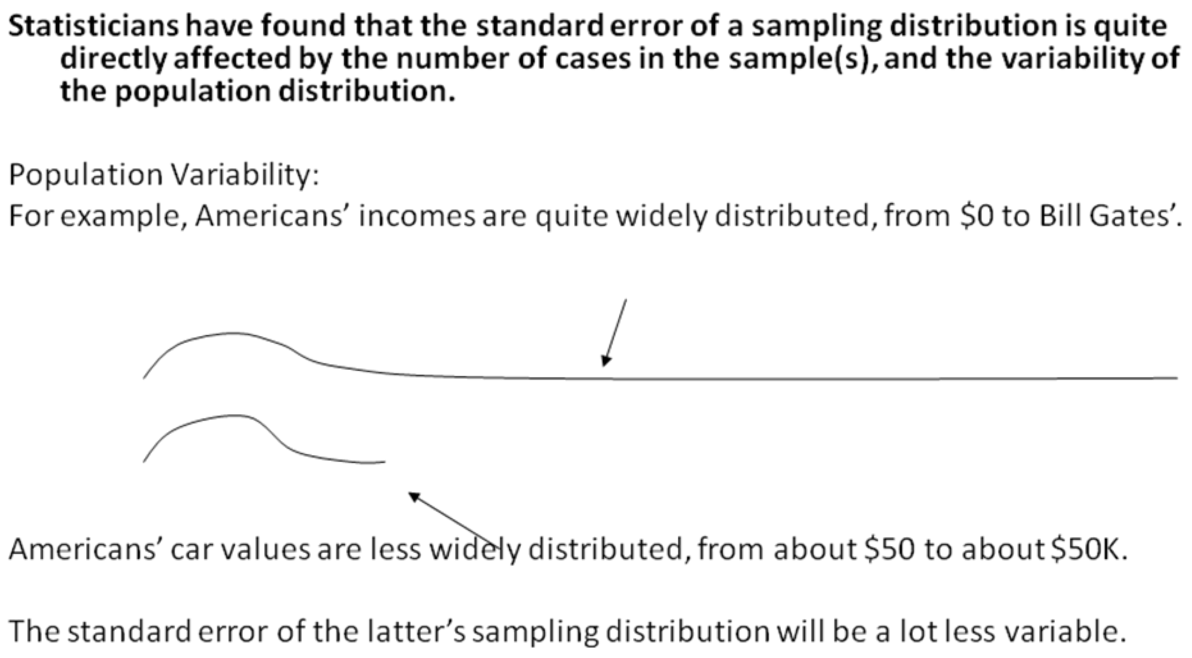


- **The first sampling distribution above, a, has a lower standard error.**
  Now a definition!
  The standard deviation of a normal sampling distribution is called the *standard error*.

- **Statisticians have found that**
  - the *standard error of a sampling distribution is :*
    - **quite directly affected by**
    - the number of cases in the sample(s), and
    - the variability of the population distribution.

Statisticians have found that the standard error of a sampling distribution is quite directly affected by the number of cases in the sample(s), and the variability of the population distribution.

Population Variability:
For example, Americans' incomes are quite widely distributed, from $0 to Bill Gates'.
Americans' car values are less widely distributed, from about $50 to about $50K.
The standard error of the latter's sampling distribution will be a lot less variable.



Statisticians have found that the standard error of a sampling distribution is quite directly affected by the number of cases in the sample(s), and the variability of the population distribution.

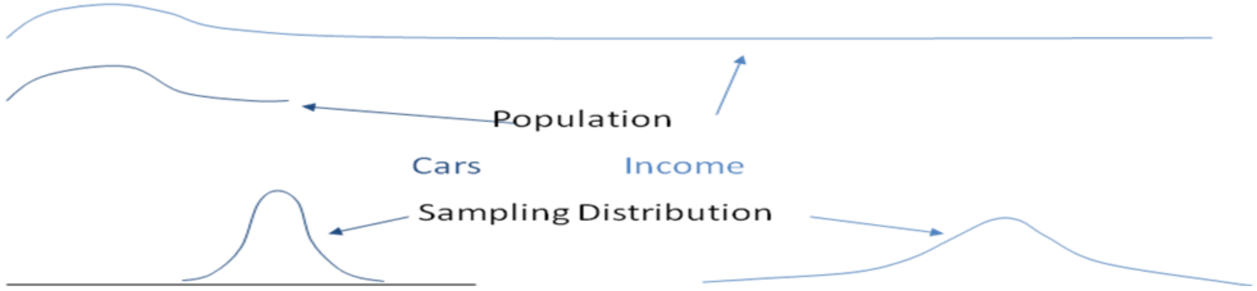
Population Variability:

For example, Americans' incomes are quite widely distributed, from $0 to Bill Gates'.

Americans' car values are less widely distributed, from about $50 to about $50K.

The standard error of the latter's sampling distribution will be a lot less variable.

Population Variability:



The standard error of income's sampling distribution will be a lot higher than car price's.

What decides the Sampling Distribution?

Standard error :

<u>The sample size affects the sampling distribution too</u>:

Standard error = population standard deviation / square root of sample size

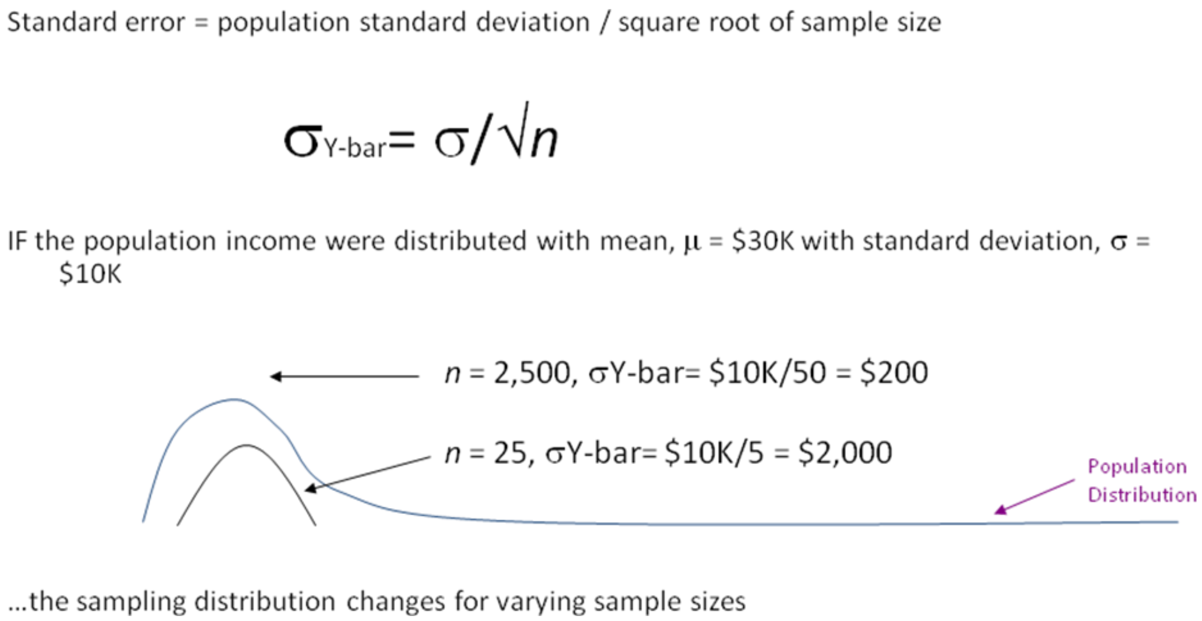
σY-bar= $\sigma/\sqrt{n}$

Example for Standard error

Standard error = population standard deviation / square root of sample size

σY-bar= σ/√*n*

IF the population income were distributed with mean, μ = $30K with standard deviation, σ = $10K



Standard error = population standard deviation / square root of sample size

$$\sigma_{\text{Y-bar}} = \sigma/\sqrt{n}$$

IF the population income were distributed with mean, μ = $30K with standard deviation, σ = $10K

n = 2,500, σY-bar= $10K/50 = $200

n = 25, σY-bar= $10K/5 = $2,000

…the sampling distribution changes for varying sample sizes

So why are sampling distributions less variable when sample size is larger?

Example 1:

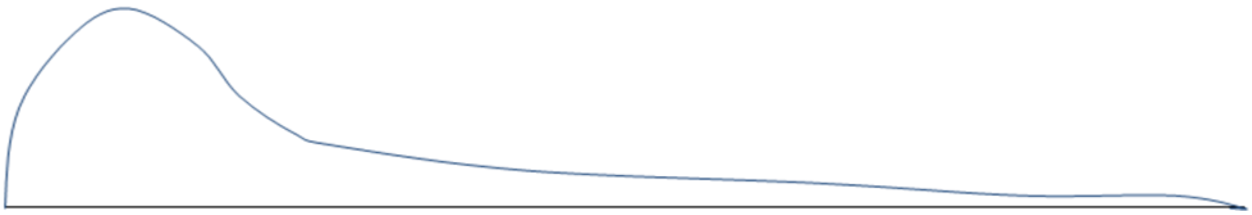
- Think about what kind of variability you would get
  - *if you collected income through repeated samples of size 1 each.*
- Contrast that with the variability you would get:
  - i*f you collected income through repeated samples of size N – 1 (or 300 million minus one) each.*

Example 2:

- Think about drawing the population distribution and playing "darts" where the mean is the bull's-eye. Record each one of your attempts.
- Contrast that with playing "darts" but doing it in rounds of 30 and recording the average of each round.

- What kind of variability will you see in the first versus the second way of recording your scores.

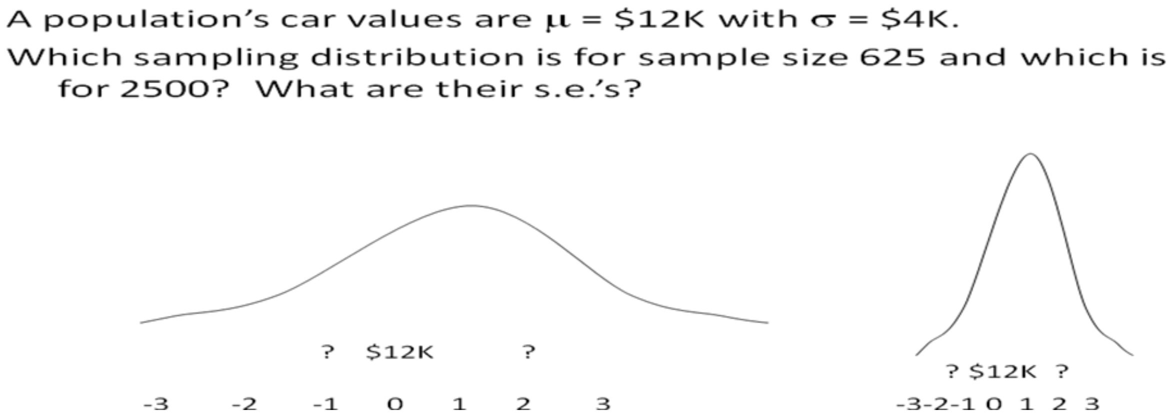
…Now, do you trust larger samples to be more accurate?



An Example:

A population's car values are μ = $12K with σ = $4K.

Which sampling distribution is for sample size 625 and which is for 2500? What are their s.e.'s?



A population's car values are μ = $12K with σ = $4K.

Which sampling distribution is for sample size 625 and which is for 2500? What are their s.e.'s?

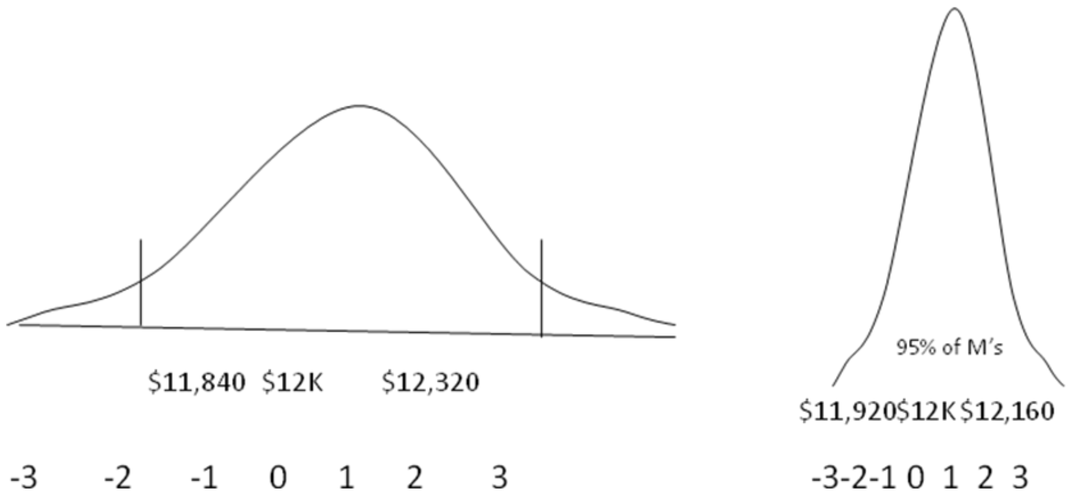
An Example:

A population's car values are μ = $12K with σ = $4K.

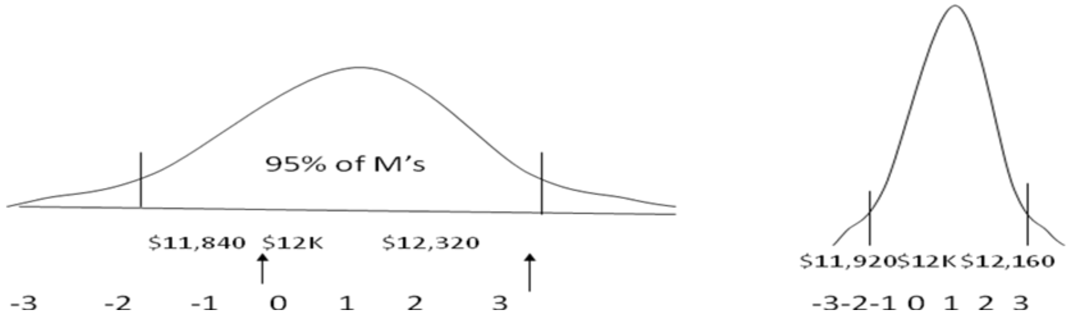
Which sampling distribution is for sample size 625 and which is for 2500? What are their s.e.'s?

s.e. = $4K/25 = $160 s.e. = $4K/50 = $80

(√625 = 25) (√2500 = 50)



A population’s car values are μ = \$12K with σ = \$4K.
Which sampling distribution is for sample size 625 and which is for 2500?
Which sample will be more precise? If you get a particularly bad sample, which sample size will help you be sure that you are closer to the true mean?



Some rules about the sampling distribution of mean

1. For a random sample of size *n* from a population having mean μ and standard deviation σ, the sampling distribution of Y-bar (glitter-bar?) has mean μ and standard error $\sigma_{\text{Y-bar}} = \sigma/\sqrt{n}$
2. The Central Limit Theorem says that for random sampling, as the sample size *n* grows, the sampling distribution of Y-bar approaches a normal distribution.
3. The sampling distribution will be normal *no matter what* the population distribution’s shape as long as $n > 30$.

4. If *n* < 30, the sampling distribution is likely normal only if the underlying population's distribution is normal.
5. As *n* increases, the standard error (remember that this word means standard deviation of the sampling distribution) gets smaller.
6. Precision provided by any given sample increases as sample size *n* increases.

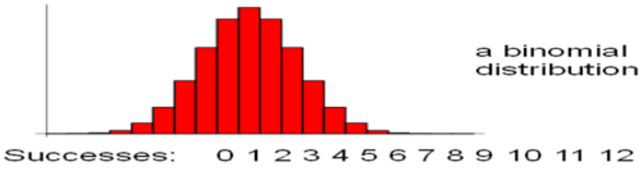
Other Rules of Sampling Distribution

So we know in advance of ever collecting a sample, that if sample size is sufficiently large:

- Repeated samples would pile up in a normal distribution
- The sample means will center on the true population mean
- The standard error will be a function of the population variability and sample size
- The larger the sample size, the more precise, or efficient, a particular sample is
- 95% of all sample means will fall between +/- 2 s.e. from the population mean

Probability Distributions

- **A Note:**
  - Not all theoretical probability distributions are Normal. One example of many is the binomial distribution.
- **The binomial distribution gives**
  - the discrete probability distribution of obtaining exactly n successes out of N trials
    - where **the result of each trial is true with known probability** of **success and false with the inverse probability**.
- The binomial distribution has
  - a formula and
  - changes shape with each probability of success and number of trials.



- However, in this class the normal probability distribution is the most useful!

- However, in this class the normal probability distribution is the most useful!

POPULATION AND SAMPLING DISTRIBUTIONS

- Population Distribution
- Sampling Distribution

Population Distribution

Definition

The ***population distribution*** is the probability distribution of the population data.

- Suppose there are only five students in an advanced statistics class and the midterm scores of these five students are:

  70 78 80 80 95
- Let *x* denote the score of a student

Table 7.1 Population Frequency and Relative Frequency Distributions

x	*f*	Relative Frequency
70	1	1/5 = .20
78	1	1/5 = .20
80	2	2/5 = .40
95	1	1/5 = .20

$N = 5$	Sum = 1.00

Table 7.2 Population Probability Distribution

x	P (x)
70	.20
	$\Sigma P\ (x) = 1.00$

Definition

The probability distribution of $\overline{x}$ is called its sampling distribution.
It lists the various values that can assume and the probability of each value of .
In general, the probability distribution of a sample statistic is called its ***<u>sampling distribution</u>***.

****

1.9 RE-SAMPLING

1.9.1 Introduction to Re-sampling

What is re-sampling?

- **Re-sampling** is:
  - the method that consists of drawing repeated samples from the original **data** samples.
- The method of **Resampling** is
  - a nonparametric method of statistical inference. ...
- The method of **resampling** uses:
  - experimental methods, rather than analytical methods, to generate the unique sampling distribution.

Re-sampling in statistics

- In **statistics**, **re-sampling** is any of a variety of methods for doing one of the following:
    - 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)

1.9.2 Need for Re-sampling

- Re-sampling involves:
    - the selection of randomized cases with replacement from the original data sample
        - in such a manner that each number of the sample drawn has a number of cases that are similar to the original data sample.
- Due to replacement:
    - the drawn number of samples that are used by the method of re-sampling consists of repetitive cases.

- Re-sampling generates a unique sampling distribution on the basis of the actual data.
- The method of re-sampling uses
    - experimental methods, rather than analytical methods, to generate the unique sampling distribution.
- The method of re-sampling yields
    - unbiased estimates as it is based on the unbiased samples of all the possible results of the data studied by the researcher.

- *Re-sampling methods* are:
    - processes of *repeatedly drawing samples* from a data set and *refitting a given model* on each sample with the *goal of learning more* about the fitted model.
- *Re-sampling methods* can be expensive since they require repeatedly performing the *same statistical methods* on *N different subsets* of the data.
- *Re-sampling methods* refit a *model of interest* to samples formed from the training set,
    - in order to obtain additional information about the fitted model.
- For example, they provide *estimates of test-set prediction error*, and the *standard deviation* and *bias* of our parameter estimates.

- The re-sampling method :

- "addresses a key problem in statistics:
- how to infer the 'truth' from a sample of data that may be incomplete or drawn from an ill-defined population."

Peterson, I. (July 27, 1991). Pick a sample. Science News, 140, 56-58.

- Using re-sampling methods,
    - “you're trying to get something for nothing.
    - You use the same numbers over and over again until you get an answer that you can't get any other way.
- In order to do that,
    - you have to assume something, and you may live to regret that hidden assumption later on”

Statement by Stephen Feinberg, cited in:
Peterson, I. (July 27, 1991). Pick a sample. Science News, 140, 56-58.

The re-sampling method frees researchers from **two limitations of conventional statistics:** “the assumption that the data conform to a bell-shaped curve and the need to focus on statistical measures whose theoretical properties can be analyzed mathematically.”
Diaconis, P., and B. Efron. (1983). Computer-intensive methods in statistics. Scientific American, May, 116-130.

1.9.3 Re-sampling methods

There are four major re-sampling methods available and are:

1. Permutation
2. Bootstrap
3. Jackknife
4. Cross validation

Resampling Method	**Application**	**Sampling procedure used**

Bootstrap	Standard deviation, confidence interval,	Samples drawn at random, with replacement
Jackknife	Standard deviation, confidence interval, bias	Samples consist of full data set with one observation left out
Permutation	Hypothesis testing	Samples drawn at random, without replacement.
Cross-validation	Model validation	Data is randomly divided into two or more subsets, with results validated across sub-samples.

1. Permutation

Permutation Origin

Re-sampling procedures date back to 1930s, when permutation tests were introduced by R.A. Fisher and E.J.G. Pitman.

They were not feasible until the computer era.

Permutation Example: Fisher’s Tea Taster

- 8 cups of tea are prepared
  - four with tea poured first
  - four with milk poured first
- The cups are presented to her in random order.
- Mark a strip of paper with eight guesses about the order of the
  - "tea-first" cups and
  - "milk-first" cups
  - let's say T T T T M M M M.

Permutation solution

- Make a deck of eight cards, four marked "T" and four marked "M.“
- Deal out these eight cards successively in all possible orderings (permutations)
- Record how many of those permutations show >= 6 matches.

Approximate Permutation

- Shuffle the deck and
- deal it out along the strip of paper with the marked guesses, record the number of matches.
- Repeat many times.

Permutation Re-sampling Processes

Step 1: Collect Data from Control & Treatment Groups

Step 2: Merge samples to form a pseudo population

Step 3: Sample without replacement from pseudo population to simulate control Treatment groups

Step 4: Compute target statistic for each example

Compute “different statistic” , save result in table and repeat resampling process 1000+ iterations.

Extension to multiple samples

- Fisher went on to apply the same idea to
  - agricultural experiments involving two or more samples.

The question became:

"How likely is it that random arrangements of the observed data would produce samples differing as much as the observed samples differ?"

Permutation Tests

- In classical hypothesis testing,
  - we start with assumptions about the underlying distribution and
    - then derive the sampling distribution of the test statistic under H_0.
- In Permutation testing,
  - the initial assumptions are not needed (except exchangeability), and
    - the sampling distribution of the test statistic under H_0 is computed by using permutations of the data.

Permutation Tests (example)

- The Permutation test is a technique that bases inference on “experiments” within the observed dataset.
- Consider the following example:
- In a medical experiment, rats are randomly assigned to a treatment (Tx) or control (C) group.
- The outcome X_i is measured in the i^{th} rat.
- Under H_0,
    - the outcome does not depend on whether a rat carries the label Tx or C.
- Under H_1,
    - the outcome tends to different, say larger for rats labeled Tx.
- A test statistic T measures
    - the difference in observed outcomes for the two groups.
    - T may be the difference in the two group means (or medians), denoted as t for the observed data.
- Under H_0,
    - the individual labels of Tx and C are unimportant, since they have no impact on the outcome. Since they are unimportant, the label can be randomly shuffled among the rats without changing the joint null distribution of the data.
- Shuffling the data creates a:
    - “new” dataset.
    - It has the same rats, but with the group labels changed so as to appear as there were different group assignments.

- Let t be the value of the test statistic from the original dataset.
- Let t_1 be the value of the test statistic computed from a one dataset with permuted labels.
- Consider all M possible permutations of the labels, obtaining the test statistics,

$t_1, \ldots, t_M$.

- Under H_0, $t_1, \ldots, t_M$ are all generated from the same underlying distribution that generated t.
- Thus, t can be compared to the permuted data test statistics, $t_1, \ldots, t_M$, to test the hypothesis and obtain a p-value or to construct confidence limits for the statistic.

- Survival times
- Treated mice 94, 38, 23, 197, 99, 16, 141
- Mean: 86.8

- Untreated mice 52, 10, 40, 104, 51, 27, 146, 30, 46
- Mean: 56.2

(Efron & Tibshirani)

- Calculate the difference between the means of the two observed samples – it's 30.6 days in favor of the treated mice.
- Consider the two samples combined (16 observations) as the relevant universe to resample from.
- Draw 7 hypothetical observations and designate them "Treatment"; draw 9 hypothetical observations and designate them "Control".
- Compute and record the difference between the means of the two samples.
- Repeat steps 3 and 4 perhaps 1000 times.
- Determine how often the resampled difference exceeds the observed difference of 30.6
- If the group means are truly equal,
- then shifting the group labels will not have a big impact the sum of the two groups (or mean with equal sample sizes).
- Some group sums will be larger than in the original data set and some will be smaller.

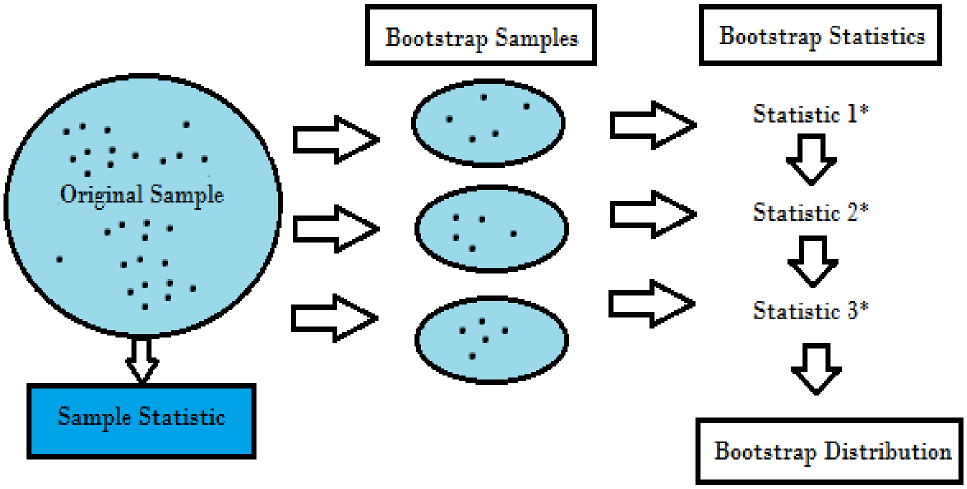
Permutation Test Example 1

- 16!/(16-7)!= 57657600
- Dataset is too large to enumerate all permutations, a large number of random permutations are selected.
- When permutations are enumerated, this is an exact permutation test.

2. <u>Bootstrap</u>

- The *bootstrap* is
  - a widely applicable tool that
  - can be used to *quantify the uncertainty* associated with a given *estimator or statistical learning approach*, including those for which it is difficult to obtain a *measure of variability*.
- The *bootstrap* generates:
  - distinct data sets by *repeatedly sampling* observations from the original data set.
  - These generated data sets can be used to *estimate variability* in lieu of sampling independent data sets from the full population.

- 1969 Simon publishes the bootstrap as an example in *Basic Research Methods in Social Science* (the earlier pigfood example)
- 1979 Efron names and publishes first paper on the bootstrap
- Coincides with advent of personal computer.



- The sampling employed by the *bootstrap* involves randomly selecting *n* observations with replacement,
  - which means some observations can be selected multiple times while other observations are not included at all.
- This process is repeated *B* times to yield *B* bootstrap data sets,
  - $Z{*}1, Z{*}2, \ldots, Z{*}B$,which can be used to *estimate* other quantities such as *standard error*.

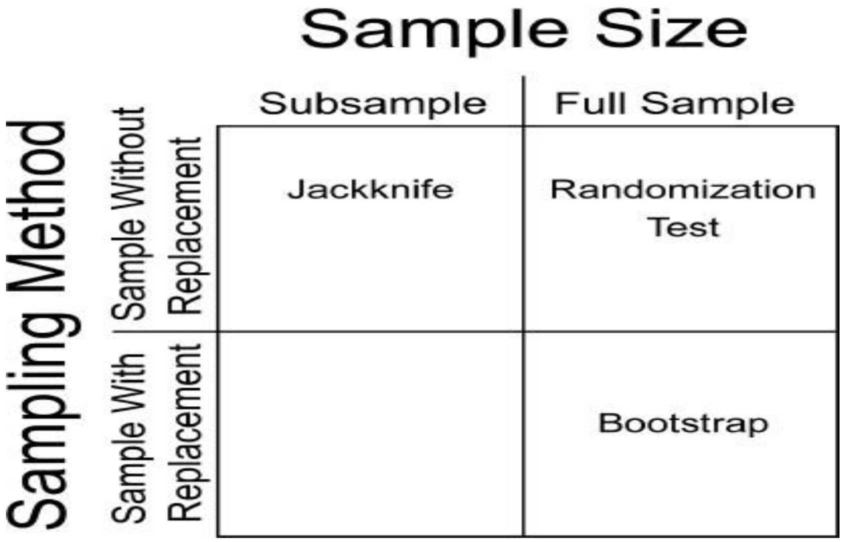
How Bootstrap Works?

Bootstrapping is a method for estimating the sampling distribution of an estimator by resampling with replacement from the original sample.

- The bootstrap procedure is a means of estimating the statistical accuracy . . . from the data in a single sample.
- Bootstrapping is used to mimic the process of selecting many samples when the population is too small to do otherwise
- The samples are generated from the data in the original sample by copying it many number of times (Monte Carlo Simulation)

- Samples can then selected at random and descriptive statistics calculated or regressions run for each sample
- The results generated from the bootstrap samples can be treated as if it they were the result of actual sampling from the original population

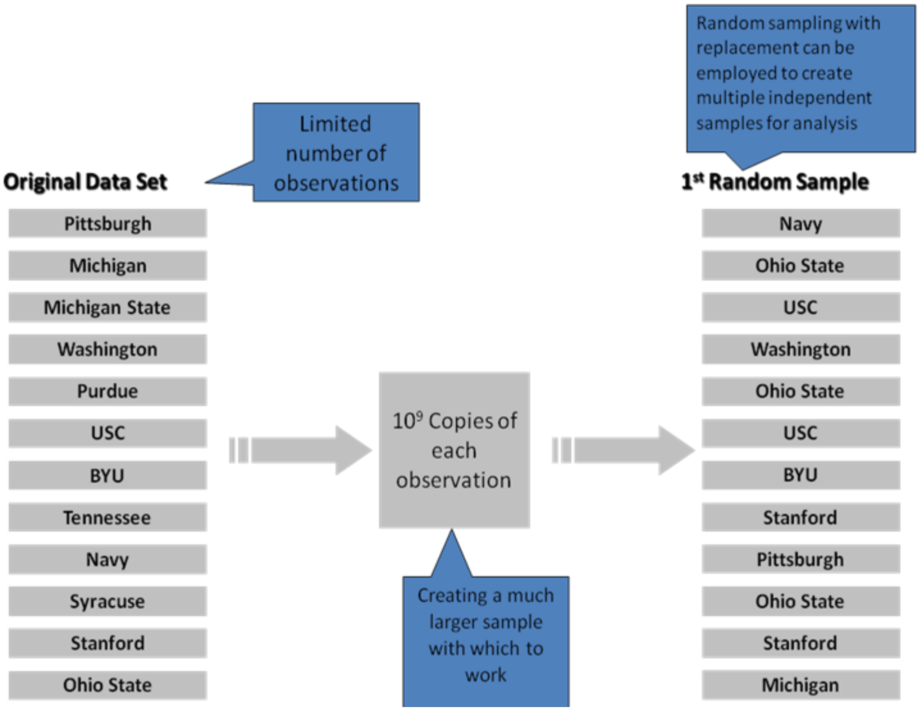
Characteristics of Bootstrapping



Sample Size

Sampling Method	Subsample	Full Sample
Sample Without Replacement	Jackknife	Randomization Test
Sample With Replacement		Bootstrap

Bootstrapping Example



When Bootstrapping should be used?

Bootstrapping is especially useful in situations when no analytic formula for the sampling distribution is available.

- Traditional forecasting methods, like exponential smoothing, work well when demand is constant – patterns easily recognized by software
- In contrast, when demand is irregular, patterns may be difficult to recognize.
- Therefore, when faced with irregular demand, bootstrapping may be used to provide more accurate forecasts, making some important assumptions…

Assumptions and Methodology

- Bootstrapping makes no assumption regarding the population
- No normality of error terms
- No equal variance
- Allows for accurate forecasts of intermittent demand
- If the sample is a good approximation of the population, the sampling distribution may be estimated by generating a large number of new samples
- For small data sets, taking a small representative sample of the data and replicating it will yield superior results

Applications and Uses

a) Criminology:- Statistical significance testing is important in criminology and criminal justice.

b) Actuarial Practice:- Process of developing an actuarial model begins with the creation of probability distributions of input variables. Input variables are generally asset-side generated cash flows (financial) or cash flows generated from the liabilities side (underwriting)

c) Classifications Used by Ecologists:- Ecologists often use cluster analysis as a tool in the classification and mapping of entities such as communities or landscapes

d) Human Nutrition:- Inverse regression used to estimate vitamin B-6 requirement of young women & Standard statistical methods were used to estimate the mean vitamin B-6 requirement.

e) Outsourcing:- Agilent Technologies determined it was time to transfer manufacturing of its 3070 in-circuit test systems from Colorado to Singapore & Major concern was the change in environmental test conditions (dry vs humid).

Bootstrap Types

a) **Parametric Bootstrap**
b) **Non-parametric Bootstrap**

a) Parametric Bootstrap

- Re-sampling makes no assumptions about the population distribution.
- The bootstrap covered thus far is a nonparametric bootstrap.
- If we have information about the population distr., this can be used in resampling.
- In this case, when we draw randomly from the sample we can use population distr.
- For example, if we know that the population distr. is normal then estimate its parameters using the sample mean and variance.
- Then approximate the population distr. with the sample distr. and use it to draw new samples.
- As expected, if the assumption about population distribution is correct then the parametric bootstrap will perform better than the nonparametric bootstrap.
- If not correct, then the nonparametric bootstrap will perform better.

Bootstrap Example

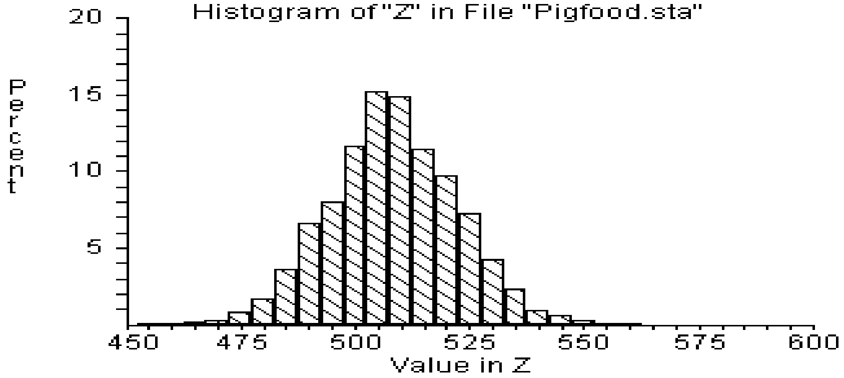
- A new pigfood ration is tested on twelve pigs, with six-week weight gains as follows:
- 496 544 464 416 512 560 608 544 480 466 512 496
- Mean: 508 ounces (establish a confidence interval)

Draw simulated samples from a hypothetical universe that embodies all we know about the universe that this sample came from – our sample, replicated an infinite number of times.

The Bootstrap process steps

1. Put the observed weight gains in a hat
2. Sample 12 with replacement
3. Record the mean
4. Repeat steps 2-3, say, 1000 times
5. Record the 5th and 95th percentiles (for a 90% confidence interval)

Bootstrapped sample means



b) Nonparametric bootstrap

- nonparametric bootstrap method which relies on the empirical distribution function.
- The idea of the nonparametric bootstrap is to simulate data from the empirical cdf Fn.
- Since the bootstrap samples are generated from Fn, this method is called the nonparametric bootstrap.
- Here Fn is a discrete probability distribution that gives probability 1/n to each observed value x1, · · · , xn.
- A sample of size n from Fn is thus a sample of size n drawn with replacement from the collection x1, · · · , xn.
- The standard deviation of ˆθ is then estimated by
- sθˆ = SQRT(1 B X B i=1 (θ ∗ i − ¯θ ∗)^ 2)
- where θ ∗ 1 , . . . , θ∗ B are produced from B sample of size n from the collection x1, · · · , xn.

Example of Bootstrap (Nonparametric)

- Have test scores (out of 100) for two consecutive years for each of 60 subjects.
- Want to obtain the correlation between the test scores and the variance of the correlation estimate.

3. Jackknife Method

- Jackknife method was introduced by Quenouille (1949)
  - to estimate the bias of an estimator.
- The method is later shown to be useful in reducing the bias as well as in estimating the variance of an estimator.

- Let ˆθn be an estimator of θ based on n i.i.d. random vectors X1, . . . , Xn, i.e., ˆθn = fn(X1, . . . , Xn), for some function fn. Let

- A statistical method for estimating and removing bias***** and for deriving robust estimates of standard errors and confidence intervals.
- Created by systematically dropping out subsets of data one at a time and assessing the resulting variation.

- θn,−i = fn−1(X1, . . . , Xi−1, Xi+1, . . . , Xn) be the corresponding recomputed statistic based on all but the i-th observation. The jackknife estimator of bias E(ˆθn) − θ is given by
- biasJ = (n − 1) n Xn i=1 ¡ ˆθn,−i − ˆθn ¢ .
- Jackknife estimator θJ of θ is given by:
  - θJ = ˆθn − biasJ = 1 n Xn i=1 ¡ nˆθn − (n − 1)ˆθn,−i ¢ .
    (2)

- Such a bias corrected estimator hopefully reduces the over all bias.
- The summands above θn,i = ˆθn − (n − 1)ˆθn,−i , i = 1, . . . , n are called pseudo-values.

A comparison of the Bootstrap & Jackknife

- Bootstrap
  - Yields slightly different results when repeated on the same data (when estimating the standard error)
  - Not bound to theoretical distributions

- Jackknife
  - Less general technique
  - Explores sample variation differently
  - Yields the same result each time
  - Similar data requirements

4. Cross validation

- **Cross-validation is**
  - a technique used to protect against overfitting in a predictive model,
  - particularly in a case where the amount of data may be limited.

- In cross-validation, you make a fixed number of folds (or partitions) of the data, run the analysis on each fold, and then average the overall error estimate.

- ***Cross validation* is a *re-sampling method* that**
    - can be used to estimate a given statistical methods *test error* or to determine the appropriate amount of *flexibility*.
    - *Model assessment* is the process of evaluating a *model's performance*.
    - *Model selection* is the process of selecting the *appropriate level of flexibility* for a model.
    - *Bootstrap* is used in a number of contexts,
    - but *most commonly* it is used to provide *a measure of accuracy* of a given *statistical learning method* or *parameter estimate*.

Need of Cross validation

- Use the entire data set when training a learner.
- Some of the data is removed before training begins.
- Then when training is done, the data that was removed can be used to test the performance of the learned model on ``new'' data.
- This is the basic idea for a whole class of model evaluation methods called *cross validation*.

Bootstrap vs. Cross-Validation

- **Bootstrap**
    - Requires a small of data
    - More complex technique – time consuming

- **Cross-Validation**
    - Not a resampling technique
    - Requires large amounts of data
    - Extremely useful in data mining and artificial intelligence

Cross Validation Methods

1. **holdout method**
2. **K-fold cross validation**
3. **Leave-one-out cross validation**

1. holdout method

- The holdout method is the simplest kind of cross validation.
- The data set is separated into
  - two sets, called the training set and the testing set.
- The function approximator fits a function using the training set only.
- Then the function approximator is asked to
  - predict the output values for the data in the testing set (it has never seen these output values before).
- The errors it makes are accumulated as before to
  - give the mean absolute test set error, which is used to evaluate the model.

- The advantage of this method is that
  - it is usually preferable to the residual method and takes no longer to compute.
  - However, its evaluation can have a high variance.
- The evaluation may depend heavily on
  - which data points end up in the training set and which end up in the test set, and
  - thus the evaluation may be significantly different depending on how the division is made.

2. K-fold cross validation

- K-fold cross validation is one way to improve over the holdout method.
  - The data set is divided into *k* subsets, and the holdout method is repeated *k* times.
- Each time, one of the *k* subsets is used as
  - the test set and
  - the other *k-1* subsets are put together to form a training set.
  - Then the average error across all *k* trials is computed.
- The advantage of this method is that
  - it matters less how the data gets divided.

- Every data point gets to be in a test set exactly once, and gets to be in a training set *k-1* times.
- The variance of the resulting estimate is reduced as *k* is increased.

- The disadvantage of this method is that the training algorithm has to be rerun from scratch *k* times, which means it takes *k* times as much computation to make an evaluation.
- A variant of this method is to randomly divide the data into a test and training set *k*different times.
- The advantage of doing this is that
  - you can independently choose how large each test set is and how many trials you average over.

Leave-one-out cross validation

- Leave-one-out cross validation is K-fold cross validation taken to its :
  - logical extreme, with K equal to N, the number of data points in the set.
  - That means that N separate times, the function approximator is trained on all the data except for one point and a prediction is made for that point.
  - As before the average error is computed and used to evaluate the model.

  - The evaluation given by leave-one-out cross validation error (LOO-XVE) is good, but at first pass it seems very expensive to compute.
  - Fortunately, locally weighted learners can make LOO predictions just as easily as they make regular predictions.
  - That means computing the LOO-XVE takes no more time than computing the residual error and it is a much better way to evaluate models.

****

1.10 STATISTICAL INFERENCE

1.10.1 Introduction to Statistical Inference

Inference : **-** Use a random sample to learn something about a larger population

- **Two ways to make inference**
  1) **Estimation of parameters**
     -Point Estimation ($\bar{X}$ or p)
     -Intervals Estimation

  2) **Hypothesis Testing**

Statistical Inference

The process of making guesses about the truth from a sample.

- Statistical inference is
  - the process through which inferences
  - about a population are made based on certain statistics calculated from a sample of data drawn from that population.
  - **What is the practical example of statistical inference?**

  - Use a statistical approach to make an inference about the **distribution of a sample of data** we collect.
- **What distribution(s) are the data from?**

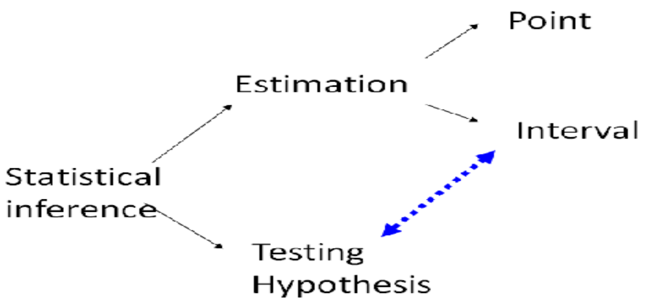
Normal distribution? Poisson distribution?

or other distributions we have not seen before?

Suppose that they are from the normal distribution.

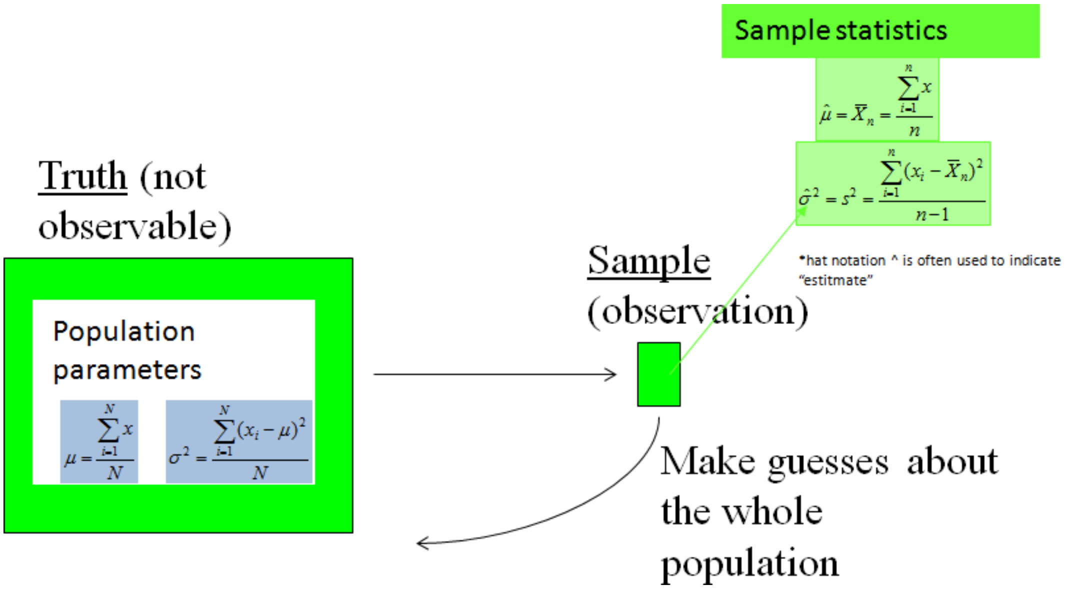
- **What normal distribution(s) are the data from?**

N(0,1)? N(0,5)? N(-3, 5)? or other normal distributions?



Use a statistical approach to **make an inference about the distribution of a sample of data we collect.**

The population or macroscopic phenomenon is always unknown itself, because some, but not all, of the data of our interest can be taken.


Sample statistics
$\hat{\mu} = \bar{X}_n = \frac{\sum_{i=1}^{n} x}{n}$
$\hat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X}_n)^2}{n-1}$
Truth (not observable)
Population parameters
$\mu = \frac{\sum_{i=1}^{N} x}{N}$
$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$
Sample (observation)
*hat notation ^ is often used to indicate “estitmate”
Make guesses about the whole population

1.10.2 Types of statistical inference

There are T**wo most common types** of Statistical Inference and they are:

1. Confidence intervals and
2. Tests of significance.

1. Confidence Intervals

- <u>Range</u> of values that *m* is expected to lie within
- 95% confidence interval
  - .95 probability that *m* will fall within range
  - probability is the *level of confidence*
    - e.g., .75 (uncommon), or .99 or .999
- Which level of confidence to use?
  - Cost vs. benefits judgement ~

Finding Confidence Intervals

- Method depends on whether s is known
- If s known

$$\underbrace{X - z_{CV}(\sigma_X)}_{\text{Lower limit}} < \mu < \underbrace{X + z_{CV}(\sigma_X)}_{\text{Upper limit}}$$

$$\text{or} \qquad X \pm z_{CV}\,(\sigma_X)$$

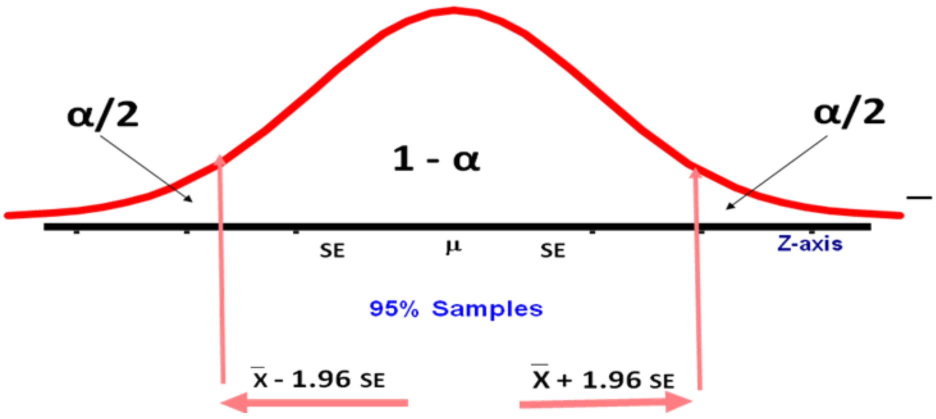
Meaning of Confidence Interval

- 95% confident that *m* lies between lower & upper limit
  - NOT absolutely certain
  - .95 probability
- If computed C.I. 100 times
  - using same methods
  - *m* within range about 95 times
- Never know *m* for certain
  - 95% confident within interval ~

Example

- Compute 95% C.I.
- IQ scores
  - *s* = 15
- Sample: 114, 118, 122, 126
  - SX_i = 480, X = 120, s_X = 7.5
  - 120 ± 1.96(7.5)
  - 120 $\underline{+}$ 14.7
  - 105.3 < m < 134.7

- We are 95% confident that population means lies between 105.3 and 134.7 ~



Changing the Level of Confidence

- We want to be 99% confident
  - using same data
  - *z* for area = .005
  - $z_{CV..01}$ = 2.57
- 120 ± 2.57(7.5)
  - 100.7 < *m* < 139.3
- Wider than 95% confidence interval
  - wider interval ---> more confident ~

When *s* Is Unknown?

- Usually do not know *s*
- Use different formula
- "Best"(unbiased) point-estimator of *s* = *s*
  - standard error of mean for sample

$$s_{\overline{X}} = \frac{s}{\sqrt{n}}$$

- Cannot use z distribution
  - 2 uncertain values: *m* and *s*
  - need wider interval to be confident
- Student's *t* distribution
  - also normal distribution
  - width depends on how well *s* approximates *s* ~

Student's *t* Distribution

- if *s* = *s*, then *t* and *z* identical
  - if *s* ¹ *s*, then *t* wider
- Accuracy of *s* as point-estimate
  - depends on sample size
  - larger *n* ---> more accurate
- *n* > 120
  - *s* » *s*
  - *t* and *z* distributions almost identical ~

Degrees of Freedom

- Width of *t* depends on *n*
- Degrees of Freedom
  - related to sample size
  - larger sample ---> better estimate
  - *n* - 1 to compute *s* ~

Critical Values of *t*

- Table A.2: "Critical Values of *t*"
- df = n - 1
- level of significance for <u>two-tailed</u> test
  - *a*
  - area in <u>both</u> tails for critical value
- level of confidence for CI ~
  - 1 - *a* ~

Confidence Intervals: *s* unknown

- Same as known but use *t*
  - Use <u>sample</u> standard error of mean
  - df = n-1

$$\underbrace{\bar{X} - t_{CV}\,(s_{\bar{X}})}_{\text{Lower limit}} < \mu < \underbrace{\bar{X} + t_{CV}(s_{\bar{X}})}_{\text{Upper limit}} \qquad [df = n - 1]$$

$$\text{or} \qquad \bar{X} \pm t_{CV}\,(s_{\bar{X}}) \qquad [df = n - 1]$$

4 factors that affect CI width

- Would like to be narrow as possible
  - usually reflects less uncertainty
- Narrower CI by...

1.* Increasing *n

- decreases standard error

2.* Decreasing *s* or *s

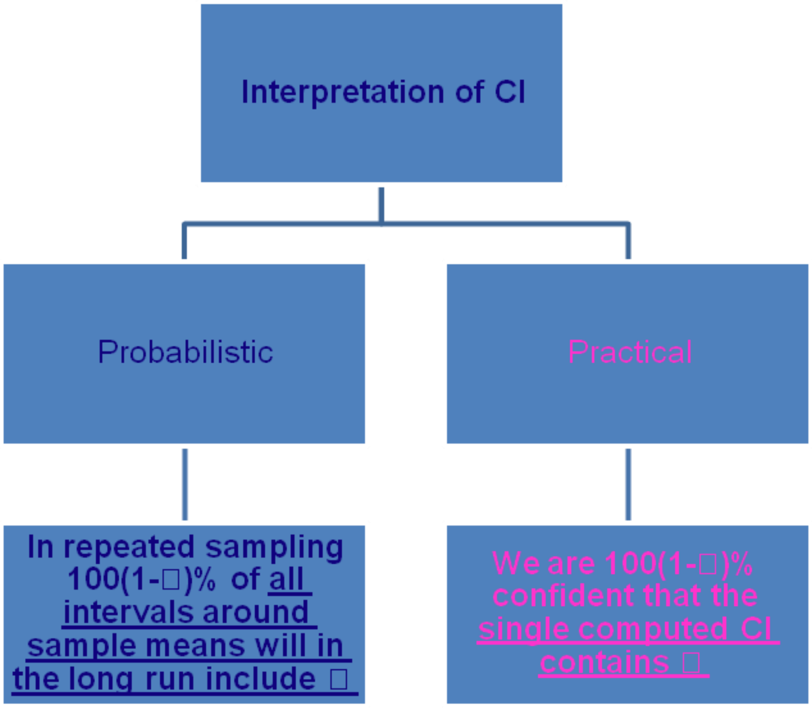
- little control over this ~

3. *s* known

- use *z* distribution, critical values

4. Decreasing level of confidence

- increases uncertainty that m lies within interval
- costs / benefits ~



Interpretation of CI

Probabilistic

Practical

In repeated sampling 100(1-□)% of <u>all intervals around sample means will in the long run include □</u>

We are 100(1-□)% confident that the <u>single computed CI contains □</u>

<u>2. Test of Significance (Hypothesis testing)</u>

- **A statistical method that uses:**
  - sample data to evaluate a hypothesis about a population parameter.

- It is intended to help researchers differentiate between real and random patterns in the data.

What is a Hypothesis?

- **A hypothesis is an assumption about the population parameter.**
  - A parameter is a Population mean or proportion
  - The parameter must be identified before analysis.

Hypothesis Testing

- Is also called *significance testing*
- Tests a claim about a parameter using evidence (data in a sample
- The technique is introduced by considering a one-sample z test
- The procedure is broken into four steps
- *Each* element of the procedure must be understood

Hypothesis Testing Steps

A. Null and alternative hypotheses
B. Test statistic
C. P-value and interpretation
D. Significance level (optional)

A. The Null Hypothesis, H_0

- States the Assumption (numerical) to be tested
- e.g. The average # TV sets in US homes is at least 3 (H_0: m ³ 3)
- Begin with the assumption that the null hypothesis is TRUE.
  (Similar to the notion of innocent until proven guilty)
- The Null Hypothesis may or may not be rejected.

The Alternative Hypothesis, H_1

- Is the opposite of the null hypothesis e.g. The average # TV sets in US homes is less than 3 (H_1: m < 3)
- Challenges the Status Quo
- Never contains the '=' sign

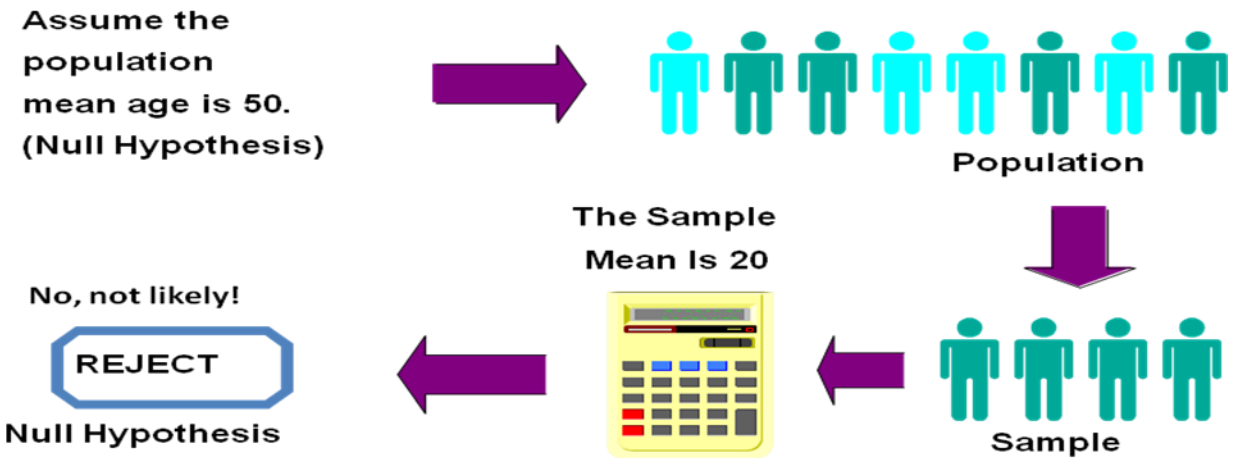
- The Alternative Hypothesis may or may not be accepted

Identify the Problem

Steps:

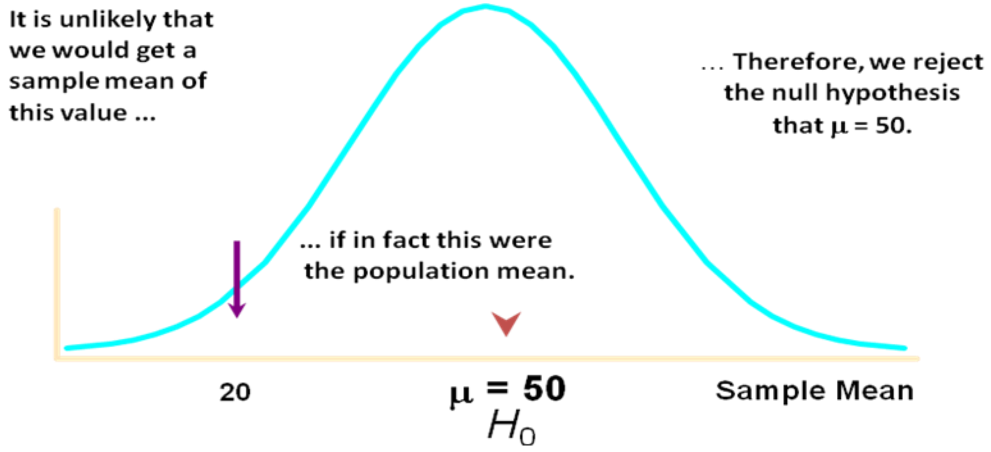
- – State the Null Hypothesis (H_0: m ³ 3)
- – State its opposite, the Alternative Hypothesis (H_1: m < 3)
  - Hypotheses are mutually exclusive & exhaustive
  - Sometimes it is easier to form the alternative hypothesis first.

Hypothesis Testing Process



Reason for Rejecting H_0

Sampling Distribution :



Level of Significance, *a*

- Defines Unlikely Values of Sample Statistic if Null Hypothesis Is True
    - Called Rejection Region of Sampling Distribution
- Designated *a* (alpha)
    - Typical values are 0.01, 0.05, 0.10
- Selected by the Researcher at the Start
- Provides the Critical Value(s) of the Test

Errors in Making Decisions

- **Type I Error**
    - Reject True Null Hypothesis
    - Has Serious Consequences
    - Probability of Type I Error Is *a*
        - Called Level of Significance
- **Type II Error**
    - Do Not Reject False Null Hypothesis
    - Probability of Type II Error Is *b* (Beta)

Hypothesis Testing: Steps

Test the Assumption that the true mean SBP of participants is 120 mmHg.

State H_0	H_0 **:** $m = 120$
State H_1	H_1 **:** $m \neq 120$
Choose a	**a = 0.05**
Choose *n*	**n = 100**
Choose Test:	**Z, *t*, X^2 *Test (or p Value)***

<u>One sample-mean Test</u>

- **Assumptions**
    - **Population is normally distributed**
- **t test statistic**

$$t = \frac{\text{sample mean} - \text{null value}}{\text{standard error}} = \frac{\bar{x} - \mu_0}{s \Big/ \sqrt{n}}$$

Example 2:

Normal Body Temperature

What is **normal body temperature**? Is it actually 37.6ºC (on average)?

State the null and alternative hypotheses

H_0: m = 37.6ºC

H_a: $m \neq$ 37.6ºC

Data: random sample of n = 18 normal body temps

37.2	**36.8**	**38.0**	**37.6**	**37.2**	**36.8**	**37.4**	**38.7**	**37.2**
36.4	**36.6**	**37.4**	**37.0**	**38.2**	**37.6**	**36.1**	**36.2**	**37.5**

Summarize data with a test statistic

Variable	**n**	**Mean**	**SD**	**SE**	**t**	**P**
Temperature	**18**	**37.22**	**0.68**	**0.161**	**2.38**	**0.029**

$$t = \frac{\text{sample mean} - \text{null value}}{\text{standard error}} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Find the p-value

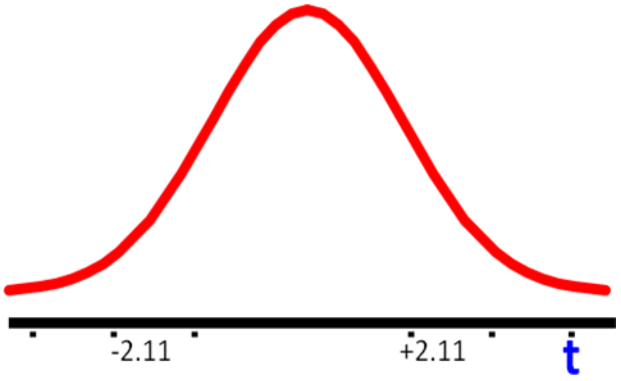
Df = n – 1 = 18 – 1 = 17

From SPSS: p-value = 0.029

From t Table: p-value is between 0.05 and 0.01.

Area to left of t = -2.11 equals area to right of t = +2.11.

The value t = 2.38 is between column headings 2.110& 2.898 in table, and for df =17, the p-values are 0.05 and 0.01.



Decide whether or not the result is statistically significant based on the p-value

Using *a* = 0.05 as the level of significance criterion, the results are **statistically significant** because 0.029 is less than 0.05. In other words, we can reject the null hypothesis.

Report the Conclusion

We can conclude, based on these data, that the mean temperature in the human population does not equal 37.6.

Example 2:

- **In a survey of diabetics in a large city, it was found that 100 out of 400 have diabetic foot. Can we conclude that 20 percent of diabetics in the sampled population have diabetic foot.**
- **Test at the a =0.05 significance level.**

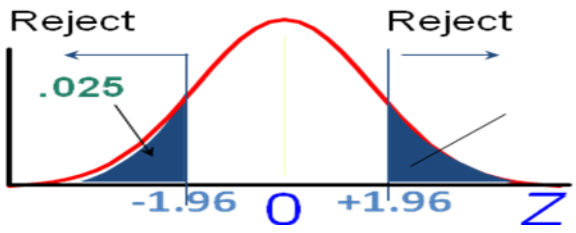
Solution

$H_o: \pi = 0.20$

$H_1: \pi \neq 0.20$

$$Z = \frac{0.25 - 0.20}{\frac{0.20\,(1 - 0.20)}{400}} = 2.50$$

Critical Value: 1.96



Decision:

We have sufficient evidence to reject the Ho value of 20%

We conclude that in the population of diabetic the proportion who have diabetic foot does not equal 0.20

****

1.11 PREDICTION ERROR

1.11.1 Introduction to Prediction Error

- A **prediction error** is the failure of some expected event to occur.
- **Errors** are an inescapable element of **predictive** analytics that should also be quantified and presented along with any model, often in the form of a confidence interval that indicates how accurate its **predictions** are expected to be.
- A prediction error is the failure of some expected event to occur.
- When predictions fail, humans can use metacognitive functions, examining prior predictions and failures.
- For example, whether there are correlations and trends, such as consistently being unable to foresee outcomes accurately in particular situations.
- Applying that type of knowledge can inform decisions and improve the quality of future predictions.

1.11.2 Error in Predictive Analysis

- Errors are an inescapable element of predictive analytics that should also be quantified and presented along with any model, often in the form of a confidence interval that indicates how accurate its predictions are expected to be.
- Analysis of prediction errors from similar or previous models can help determine confidence intervals.

Predictions always contain errors

- Predictive analytics has many applications, the above mentioned examples are just the tip of the iceberg.
- Many of them will add value, but it remains important to stress that the outcome of a prediction model will always contain an error. Decision makers need to know how big that error is.
- To illustrate, in using historic data to predict the future you assume that the future will have the same dynamics as the past, an assumption which history has proven to be dangerous.
- In artificial intelligence (AI), the analysis of prediction errors can help guide machine learning (ML), similarly to the way it does for human learning.

- In reinforcement learning, for example, an agent might use the goal of minimizing error feedback as a way to improve.
- Prediction errors, in that case, might be assigned a negative value and predicted outcomes a positive value, in which case the AI would be programmed to attempt to maximize its score.
- That approach to ML, sometimes known as error-driven learning, seeks to stimulate learning by approximating the human drive for mastery.

1.11.3 Prediction Error in statistics

1.11.3.1 Standard Error of the Estimate

- The standard error of the estimate is a measure of the accuracy of predictions.
- Recall that the regression line is the line that minimizes the sum of squared deviations of prediction (also called the *sum of squares error*).

- The standard error of the estimate is closely related to this quantity and is defined below:

$$\sigma_{est} = \sqrt{\frac{\sum (Y - Y')^2}{N}}$$

where σ_{est} is the standard error of the estimate, Y is an actual score, Y' is a predicted score, and N is the number of pairs of scores.

- In statistics, the **mean squared error** (**MSE**) or **mean squared deviation** (**MSD**) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors—that is, the average squared difference between the estimated values and what is estimated.
- MSE is a risk function, corresponding to the expected value of the squared error loss.

- The fact that MSE is almost always strictly positive (and not zero) is because of randomness or because the estimator does not account for information that could produce a more accurate estimate.

1.11.4 Mean squared prediction error

- In **statistics** the mean squared **prediction error** or mean squared **error** of the **predictions** of a smoothing or curve fitting procedure is the expected value of the squared difference between the fitted values implied by the **predictive** function and the values of the (unobservable) function g.
- The MSE is a measure of the quality of an estimator—it is always non-negative, and values closer to zero are better.
- Root-Mean-Square error or Root-Mean-Square Deviation (RMSE or RMSD)
- In an analogy to standard deviation, taking the square root of MSE yields the root-mean-square error or root-mean-square deviation (RMSE or RMSD), which has the same units as the quantity .being estimated; for an unbiased estimator, the RMSE is the square root of the variance, known as the standard error.
- The RMSD represents the square root of the second sample moment of the differences between predicted values and observed values or the quadratic mean of these differences.
- These deviations are called *residuals* when the calculations are performed over the data sample that was used for estimation and are called *errors* (or prediction errors) when computed out-of-sample.
- The RMSD serves to aggregate the magnitudes of the errors in predictions for various times into a single measure of predictive power.
- RMSD is a measure of accuracy, to compare forecasting errors of different models for a particular dataset and not between datasets, as it is scale-dependent.
- RMSD is always non-negative, and a value of 0 (almost never achieved in practice) would indicate a perfect fit to the data.
- In general, a lower RMSD is better than a higher one. However, comparisons across different types of data would be invalid because the measure is dependent on the scale of the numbers used.
- RMSD is the square root of the average of squared errors.
- The effect of each error on RMSD is proportional to the size of the squared error; thus larger errors have a disproportionately large effect on RMSD.

- Consequently, RMSD is sensitive to outliers.

What is prediction error in regression?

- **Regressions** differing in accuracy of **prediction**.
- The standard **error** of the estimate is a measure of the accuracy of **predictions**.
- Recall that the **regression** line is the line that minimizes the sum of squared deviations of **prediction** (also called the sum of squares **error**).

Thus, the prediction error influence the analytics functionalities and its applications areas.

****