Statistics &Data Analysis Concepts for Data Science and ML

1

Basic Statistical Concepts

Learning Objectives

- Learn basic vocabulary of statistics and different ways of defining statistics
- Understand data and different classifications of data
- Learn the two broad categories of statistics: Descriptive and Inferential Statistics
- Define and understand basic statistical terms including population, sample, parameters, and statistics
- Understand the tools of Descriptive and Inferential Statistics
- Understand different levels and classification of data: nominal, ordinal, interval and ratio
- Understand the importance of Sampling and Sampling Techniques
- Understand the uses, applications and importance of statistics

Statistics Defined

- Statistics is the science and art of making decisions from data.
- Statistics is a science that deals with:

  Collection, tabulation, analysis, interpretation, and presentation of data (in order to make decisions).
- Statistics is concerned with making decisions from data influencing chance variations.
- Statistics deals with making inferences or predictions about a population based on sample data.

Why study statistics?

- Statistics is the only tool that deals with variation. Most of the data we collect show variation.

  Statistical thinking and variation reduction are major goals of data analysis, decision making, and quality improvement.

- Statistical methods enable us to draw conclusions from limited data.

Importance of Statistics

- In fields such as business, research, health care, and engineering, a vast amount of raw data are collected and warehoused rapidly
- The data must be analyzed to be meaningful.
- This course is about learning how to make efficient decisions from data.

Some Statistics on Education

In the latest figures available :

- US had a total of 4,599 Title IV-eligible, degree-granting institutions:

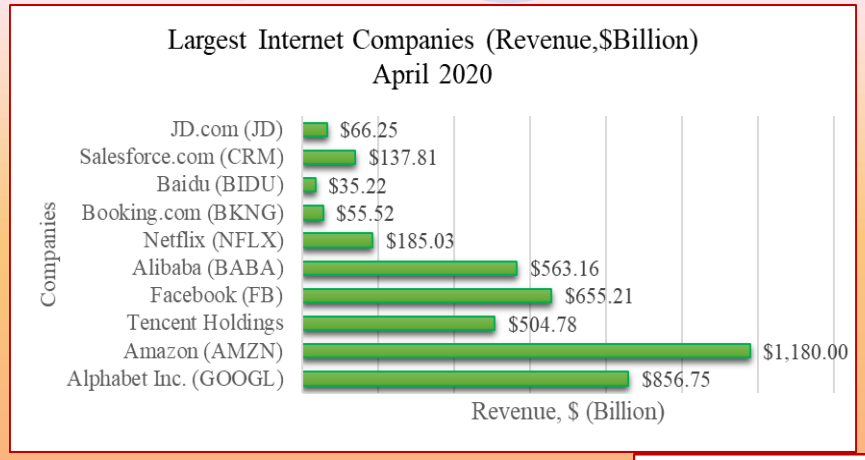
  2,870 4-year institutions and
  1,729 2-year institutions.

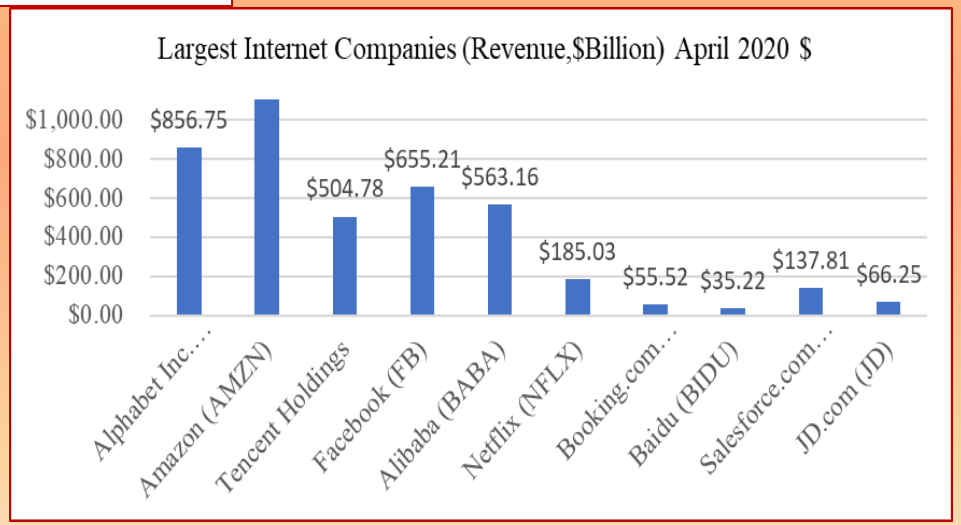
- The US had 21 million students in higher education, roughly 5.7% of the total population.
- About 13 million of these students were enrolled full-time which were 81,000 students lower than 2010 [source: NCES (2014). "Fast Facts Enrollment" U.S. Department of Education Retrieved 21 May 2014.]

More Statistics: Manufacturing in the U.S.

- In the most recent data, manufacturers contributed $2.17 trillion to the U.S. economy.
- For every $1.00 spent in manufacturing, another $1.40 is added to the economy.
- There are 12.33 million manufacturing workers in the U.S., accounting for 9 percent of the workforce.
- In 2014, the average manufacturing worker in the United States earned $79,553 annually.
- Manufacturers have experienced tremendous growth over the past few decades, making them more "lean" and helping them become more competitive globally.
- Over the next decade, nearly 3.5 million manufacturing jobs will likely be needed, and 2 million are expected to go unfilled due to the skills gap.
- Taken alone, manufacturing in the United States would be the ninth-largest economy in the world with $2.1 trillion in value added from manufacturing in 2014.
- Manufacturers in the United States perform more than three-quarters of all private-sector research and development (R&D) in the nation, driving more innovation than any other sector. [Source: National Association of Manufacturer Report, www.nam.org.]

Data Visualization_Recent Statistics & Applications





A Gallup Poll Result on Job Outlook – How Polls are Conducted

Forty-three percent of Americans say it is a good time to find a quality job. This reading is similar to the 40% to 45% recorded in 2016

Results for this Gallup poll are based on telephone interviews conducted during June 1-5, 2016, on the Gallup U.S. Daily survey, with a random sample of 1,027 adults, aged 18 and older, living in all 50 U.S. states and the District of Columbia. For results based on the total sample of national adults, the margin of sampling error is ±4 percentage points at a 95% confidence level.

What you will learn in this course?

This course is about learning how to make efficient decisions from data.

Statistics and Statistical tools will aid in gaining skills such as:

1) collecting, describing, analyzing and interpreting data for intelligent decision-making,
2) realizing that variation is an integral part of data,
3) understanding the nature and pattern of variability of a phenomenon in the data, and
4) measuring reliability of the population parameters from which the sample data are collected to draw valid inferences.

What you will learn in this course?

In this course you will learn that:

- Statistics is the science of data and variation
- statistics and statistical methods are used in collecting, presenting, and analyzing data.
- Statistics has wide applications in areas including:

  stock market activities, unemployment rates, medical research findings, opinion poll results, weather forecasts, sports data, etc. all use some form of statistics.
- Almost every field uses data and statistics to learn about systems and their processes. In fields such as business, research, health care, and engineering, a vast amount of raw data is collected and warehoused rapidly; this data must be analyzed to be meaningful.

Data and Classification of Data

- Data are any number of related observations
- Data are some form of measurements
- Data contain the information to make decision
- A single data or observation is known as a ***data point***. A collection of data is a ***data set***.

Entities or the specific items on which we collect data are known as ***variables***.

Data collected on sales, profit, the number of customers served by a bank, the diameter of a shaft produced by a manufacturing company, the number of housing starts; all show variation therefore, these are *variables*.

Classification of Data

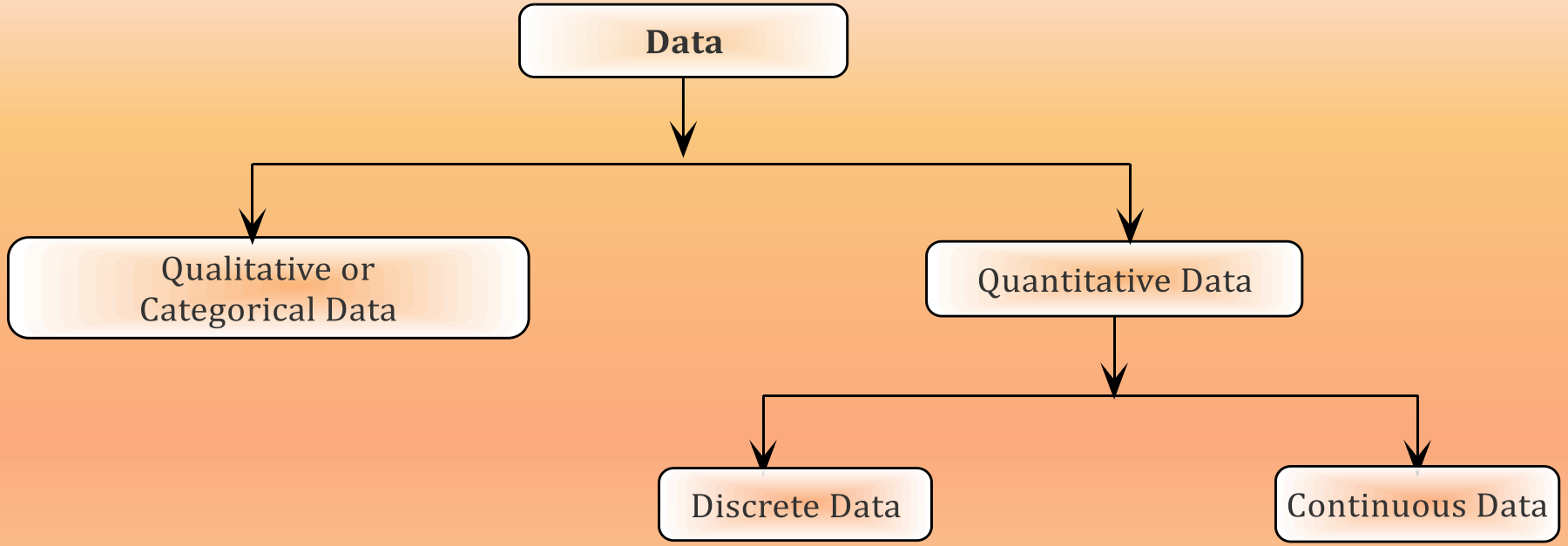
- Qualitative or Categorical Data

  Examples of qualitative data include the color of your car, response to a yes/no question, or the product rating using a Likert scale of 1 to 5 where the numbers correspond to a category (excellent, good, etc.).

- Quantitative data

  Numerical data that can be expressed in numbers. For example, data collected on temperature, sales and demand, length, height, and volume are all examples of quantitative data.

Classification of Data



Classification of Data

Time Series Data

Time series data are data recorded over time Examples: weekly sales, monthly demand for a product, or the number of orders received by an online shopping department of a department store

Cross-sectional data

The values observed at the same point in time. For example, the closing value of the stock market on the 5^{th} of each month for the past twelve months

Classification of Data

Discrete data

are the result of a counting process and are expressed as whole numbers or integers. Examples: cars sold by Toyota in the last quarter, the number of houses sold last year, or the number of defective parts produced by a company.

Continuous data

can take any value within a given range - are measured on a continuum or a scale that can be divided infinitely. Examples: measurements of length, height, diameter, temperature, stock value, sales, etc.

More powerful statistical tools are available to deal with continuous data as compared to discrete data; therefore, continuous data are preferred wherever possible.

Limitations of Statistics

- Quantitative measurements are better suited for statistical analysis.
- Statistical decisions are only approximations and are not exact. On the basis of statistical analysis, we can talk only in terms of probability and chance, and not in terms of certainty. Statistical conclusions are not universally true. They are true only on an average.
- Statistics is likely to be misused and must be used properly if it is to be of use.
- Statistics does not study individual items since individual items taken separately do not constitute statistical data.

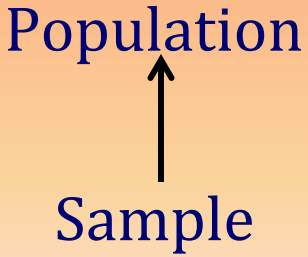
Two Broad Categories of Statistics

- Descriptive statistics,

  Describing the data using: (1) charts and graphs, and (2)numerical methods.

- Inferential statistics

  process of using ***sample*** statistics to draw conclusions about the ***population*** parameters.



Important Terms

- Population

  denotes the entire measurements that are theoretically possible. It is also known as the universe and is the totality of items or things under consideration. Example: total number of products manufactured by a company in a given period of time, or number of people who can vote in a country, etc.

- Sample

  is the portion of the population that is selected for analysis (a subset of population).

Important Terms

- A population is described by its ***parameters***

  A parameter is a summary measure that is computed to describe the characteristics of a population.

- A sample is described by its ***statistics***.

  A statistic is a summary measure that is computed to describe the characteristics of a sample.

Important Terms

- ***Population parameters***

  population mean (μ - read as "mu")

  population variance (σ^2), ,

  population standard deviation (σ, read as "sigma"), a

  population proportion, p

- ***Sample statistics***

  sample mean ($\bar{x}$ read as "x-bar"),

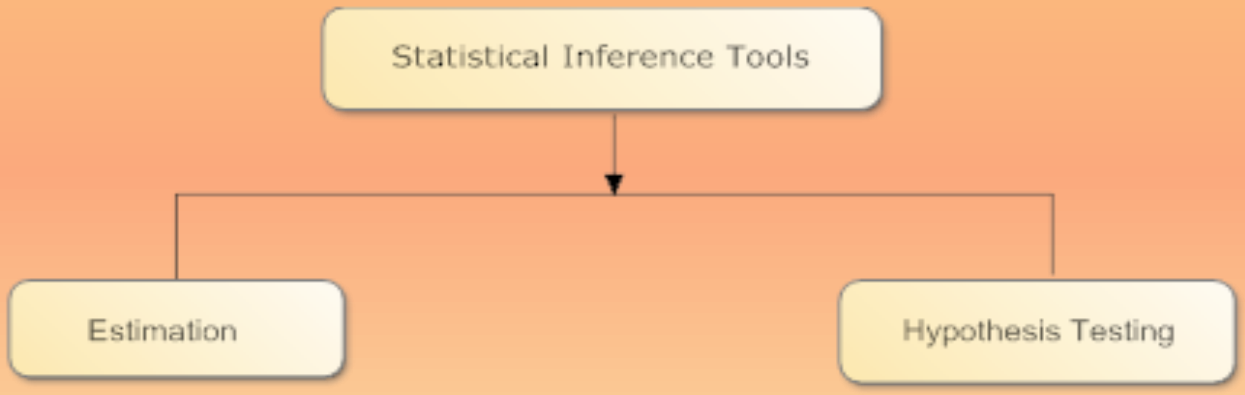
  sample variance (s^2),

  sample standard deviation (s),

  sample median, sample proportion, $\bar{p}$ read as p-bar).

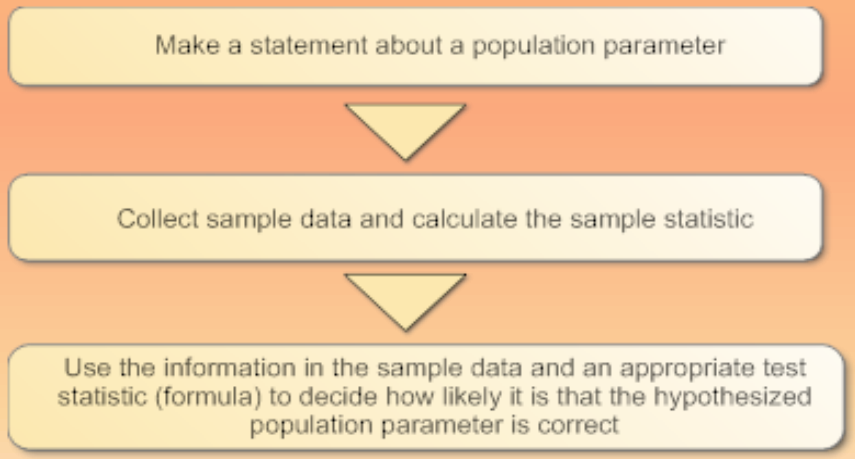
Tools of Inferential Statistics

- Inferential tools allow a decision maker to draw conclusions about the population data using the information from the sample data.
- Two major tools of inferential statistics:



Estimation and Hypothesis Testing

- ***Estimation*** is the simplest form of inferential statistics in which a sample statistic is used to draw conclusion about an unknown ***population parameter***.
- Many problems require us to decide whether or not a statement about some parameter is true or false. This statement about the population parameter is called a ***hypothesis***.



Data Collection, Presentation, and Analysis

- How data are collected (obtained) and prepared for statistical analysis
- Tabular presentation of data
- Graphical presentation of data
- Analysis, and interpretation of data

Analysis of data involves both descriptive and inferential tools.

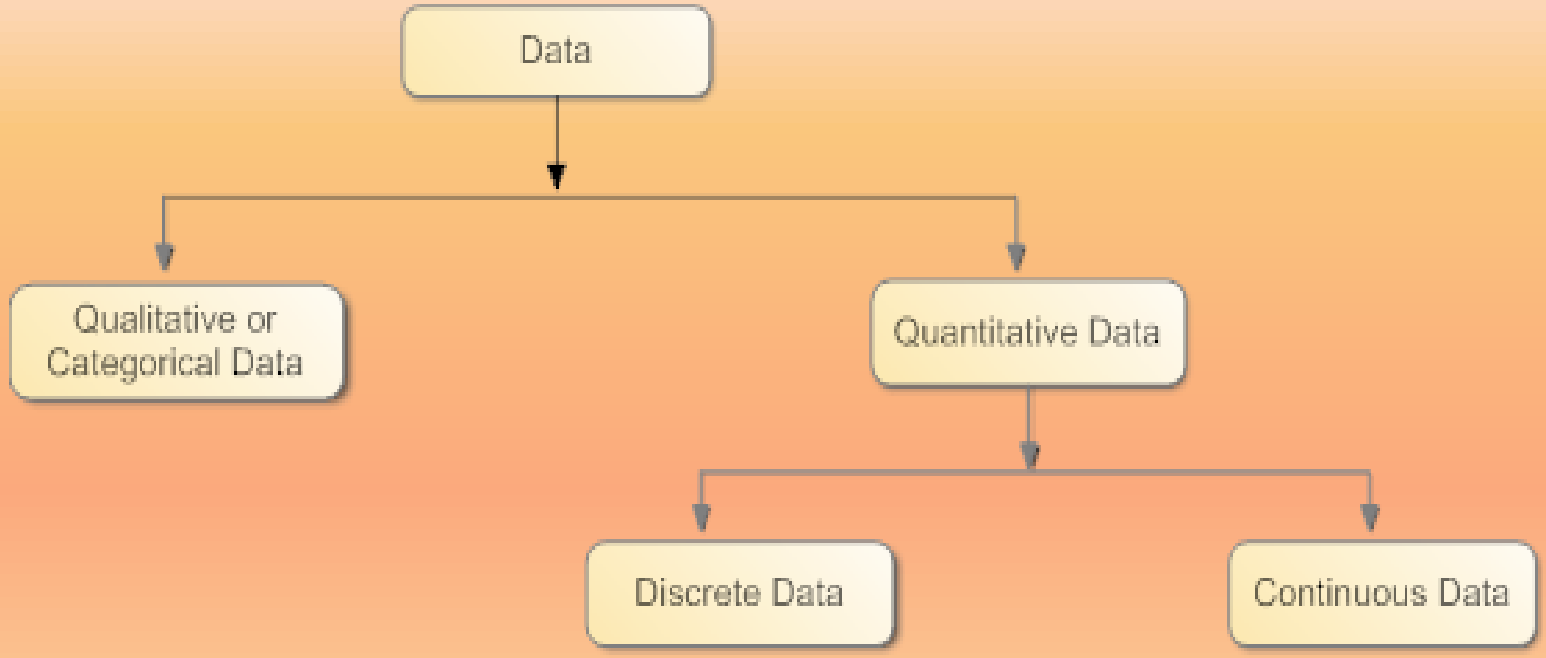
Sources of Data

- Government agencies
- Experimental design
- Telephone /mail surveys
- Processes

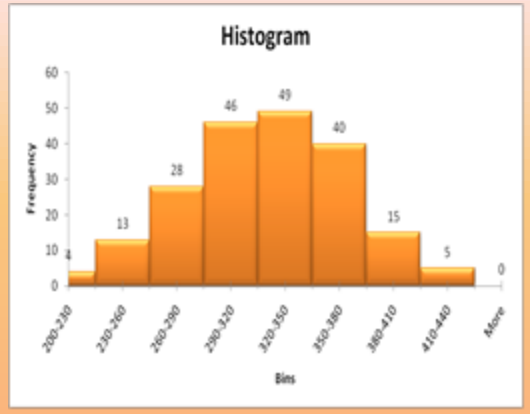
Sources of Data: Web as a Major Source of Data

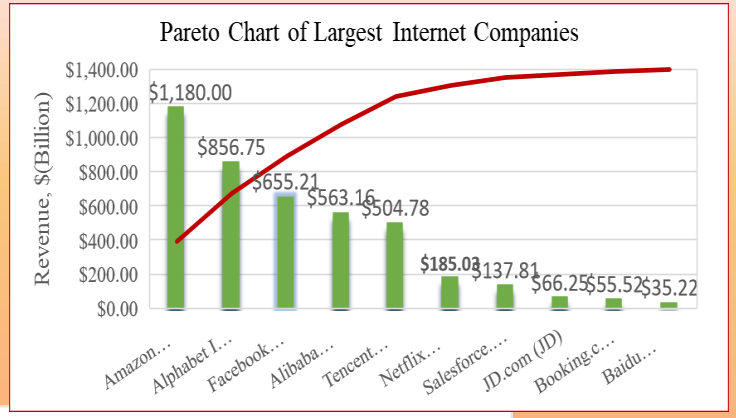
Internet Site	Available Data
Bureau of Labor Statistics (*www.bls.gov)*	Principal fact-finding agency for the Federal Government in the broad field of labor economics and statistics. Information on jobs, salary, economic analysis, unemployment rates, occupational outlook, inflation rates, and more.
U.S. Census Bureau (*www.census.gov)*	Principal agency of the U.S. Federal Statistical System, responsible for producing data about the American people and economy.
Zillow.com	Real Estate data, data on home sales, rent, buy, home values, mortgage rates.
www.cnbc.com	Latest business news on stock markets, financial & earnings on *CNBC*. View world markets streaming charts & video; check stock tickers and quotes.
Finance.yahoo.com	Stock quotes, up to date news, stock prices, portfolio management resources, international market data, message boards, and mortgage rates.
The Wall Street Journal, The New York Times, Money Magazine, Fortune	Data on finance, taxes, mortgage rate, retirement and other general interest.
Bureau of Economic Analysis (BEA)	Source of US economic statistics including national income and product accounts (NIPAs), consumer price index, gross domestic product (GDP), international data on trade.

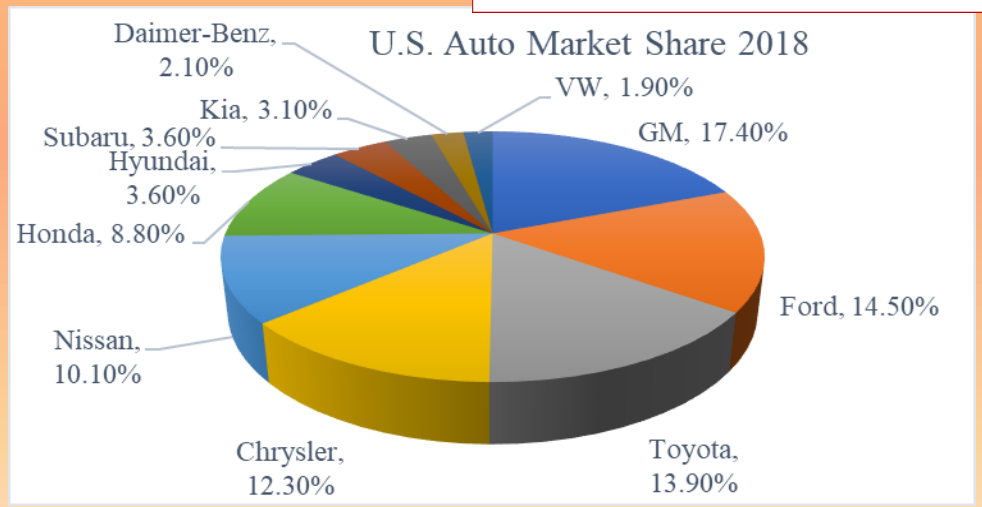
Data Types


Data
Qualitative or Categorical Data
Quantitative Data
Discrete Data
Continuous Data

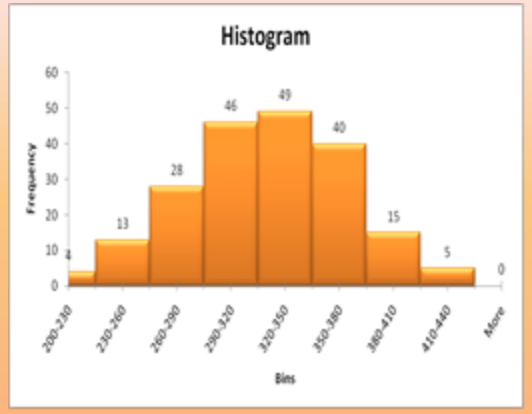
Presenting Data using Charts and Graphs: Some Examples_ Preview of Chapter 2

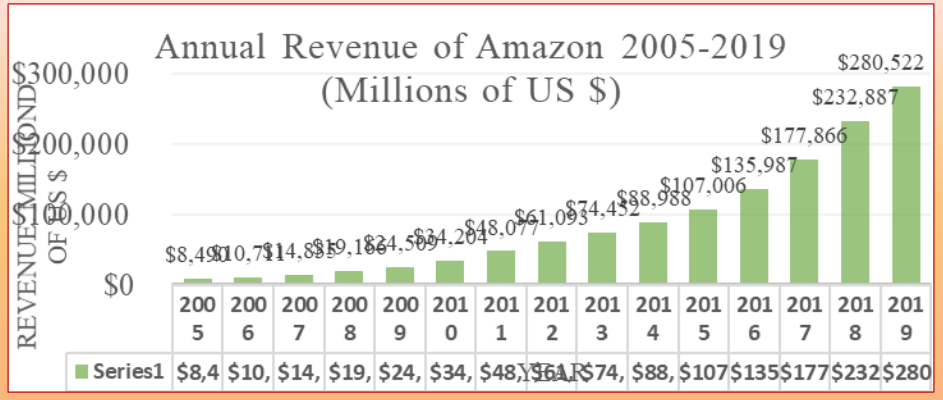


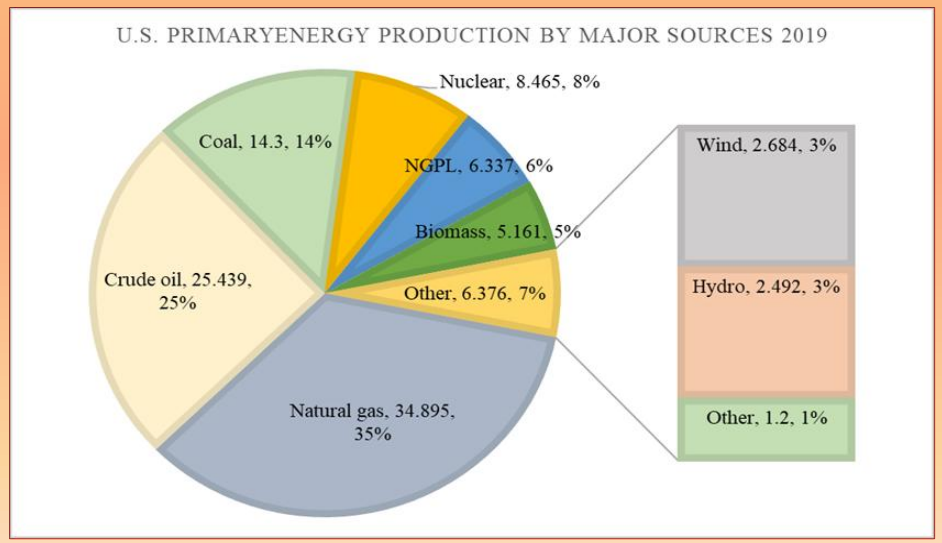




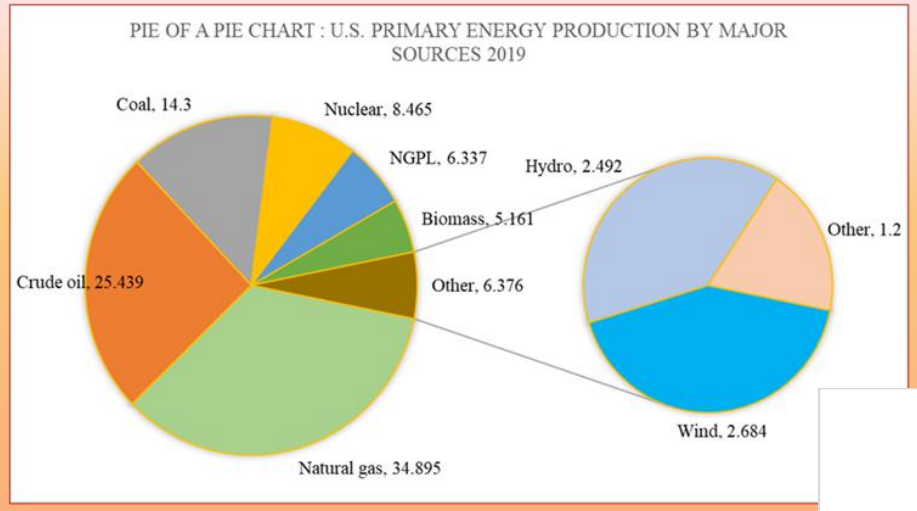
Presenting Data using Charts and Graphs: Some Examples

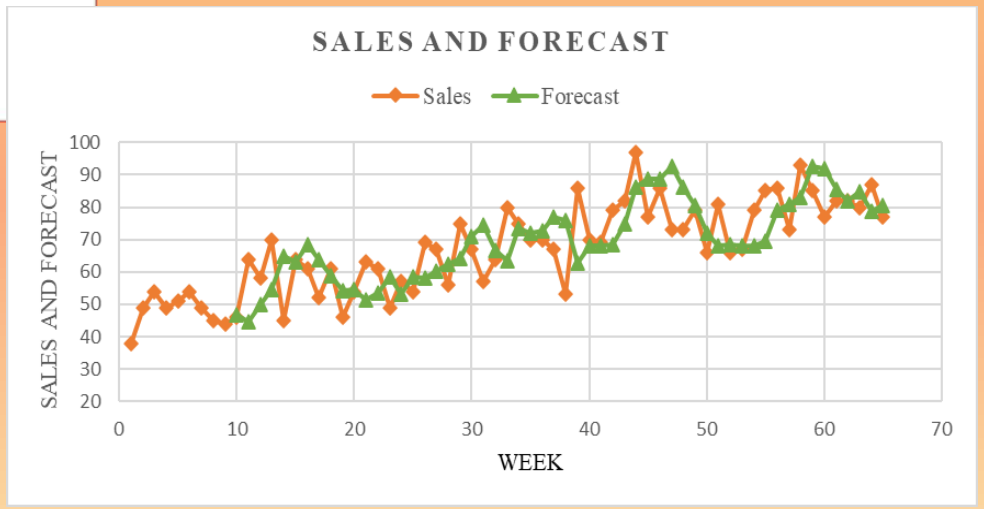






Presenting Data using Charts and Graphs: Some Examples





Levels of Measurements and Measurement Scales

Data are also described according to the *levels of measurement* used. All collected data are *measured* in some form.

There are four levels of measurements:

1. Nominal Scale
2. Ordinal Scale
3. Interval Scale, and
4. Ratio scale

The nominal is the weakest and ratio is the strongest form of measurement.

Nominal and Ordinal Scales

- ***Nominal Scale:*** If the observed data are classified into various distinct categories in which *no ordering* is implied, a ***nominal level*** of measurement is achieved.

Qualitative variable	***Category***		
Do you smoke?	Yes	No	
Stock ownership	Yes	No	
Political party affiliation	Democrat	Republican	Independent

- ***Ordinal Scale:*** If the observed data are classified into distinct categories in which ordering is implied, an ***ordinal level*** of measurement is obtained.

Qualitative variable	***Ordered categories***
Student grades	A B C D F
Rank of employees	Senior Engineer, Engineer, and Engineer trainee
Product quality	Excellent, good, poor (highest to lowest)

Nominal and Ordinal cont.…

- Nominal scale is the weakest form of measurement. Ordinal scale is also a weak form of measurement because no meaningful numerical statements can be made about the different categories. For example: the ordinal scale only tells which category is greater, but does not tell how much greater.

- For the data to be meaningful, we should be able to tell which category is greater and by how much

Interval and Ratio Scales

- Interval Scale: These measurements are made on a quantitative scale. It is an ordered scale in which the difference between any two measurements is a meaningful quantity.

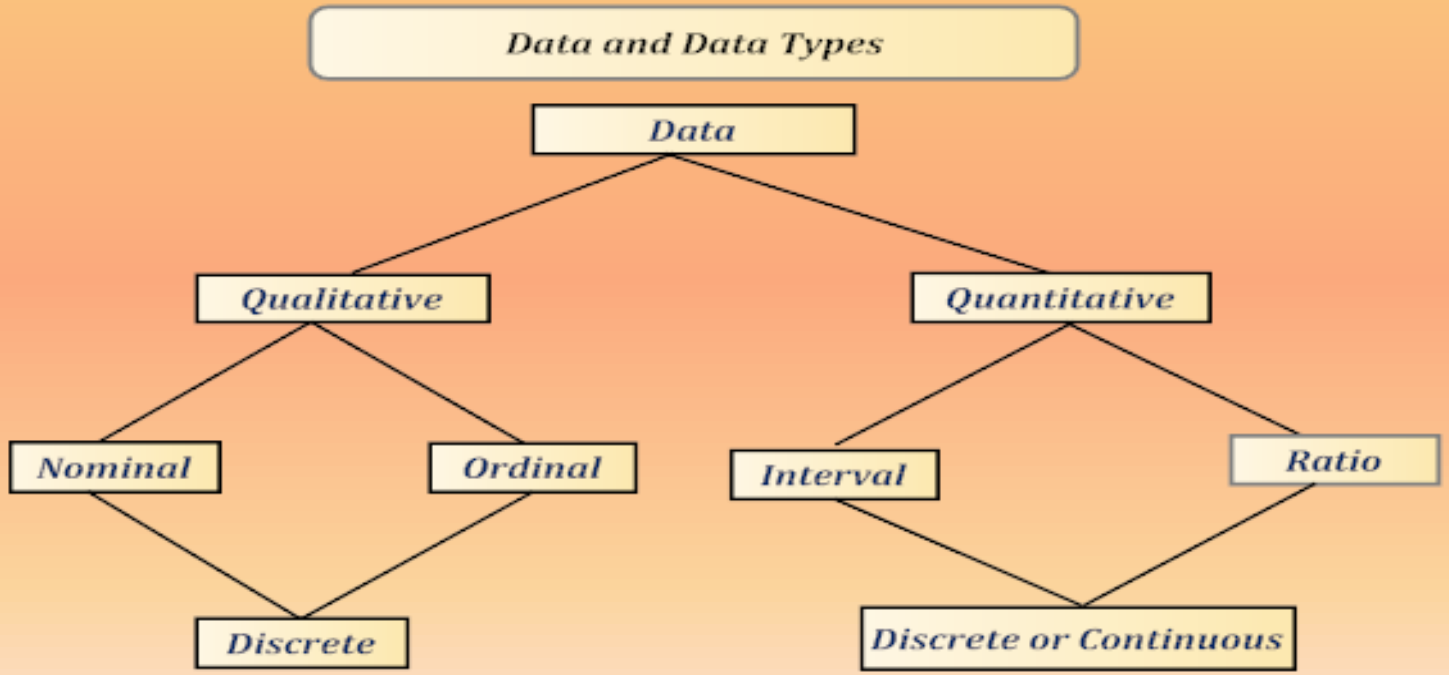
Quantitative variable	*Level of measurement*
Temperature	Interval
Time	Interval

- *Ratio Scale*: If in addition to difference being meaningful and equal at all points on a scale, there is also a "true zero" point in which the ratio of measurements are sensible to consider, then the scale is a ratio scale. The measurements are made from the same reference point.

Quantitative variable	*Level of measurements*
Height (in feet, inches)	Ratio
Weight (in pounds, kilograms)	Ratio
Age (in years, days)	Ratio

Classification data according to the levels of measurements

Data obtained from quantitative variable are measured on an interval or a ratio scale. *Ratio is the highest level of measurement*. It tells which observed value is largest, and by how much.



Making Sense from Data

Starting Salary for Management Majors: Raw Data

Starting Salary ($000)													
40.0	41.6	41.1	39.7	43.5	35.3	38.4	38.3	37.5	40.8	36.2	38.5	37.3	38.6
34.8	39.2	40.2	41.0	39.8	44.2	42.9	40.0	36.1	36.3	37.3	37.7	39.3	36.4
42.1	37.5	36.2	44.3	36.7	37.0	34.3	33.3	35.7	38.3	37.1	41.8	41.9	39.0
38.6	40.0	35.5	38.7	36.7	42.0	40.0	41.4	36.4	37.9	46.0	44.3	45.1	45.5
45.3	47.5	46.0	48.5										

Data from Table 1.4 Arranged in Increasing Order_ Data Array

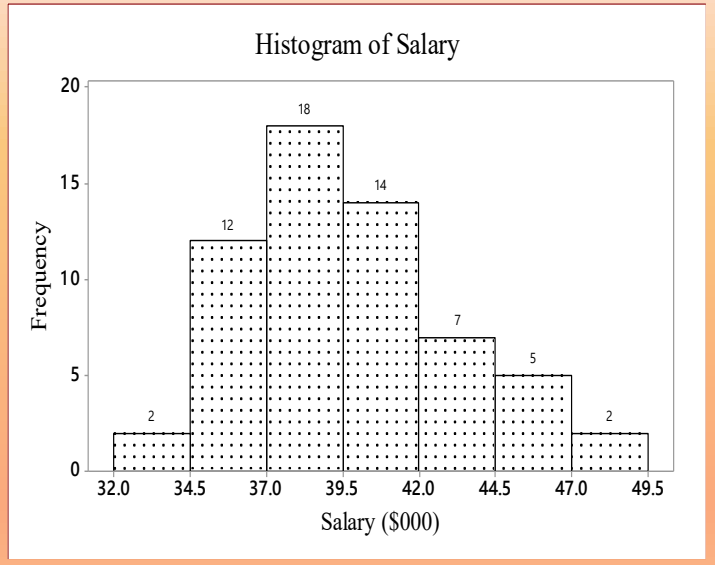
Sorted Data or Data Array (read row-wise)													
33.3	34.3	34.8	35.3	35.5	35.7	36.1	36.2	36.2	36.3	36.4	36.4	36.7	36.7
37.0	37.1	37.3	37.3	37.5	37.5	37.7	37.9	38.3	38.3	38.4	38.5	38.6	38.6
38.7	39.0	39.2	39.3	39.7	39.8	40.0	40.0	40.0	40.0	40.2	40.8	41.0	41.1
41.4	41.6	41.8	41.9	42.0	42.1	42.9	43.5	44.2	44.3	44.3	45.1	45.3	45.5
46.0	46.0	47.5	48.5										

Making Sense from Data

Frequency Distribution

Class Interval ($)	Frequency (f)
32.0- 34.5	2
34.5 – 37.0	12
37.0 – 39.5	18
39.5 – 42.0	14
42.0 – 44.5	7
44.5 – 47.0	5
47.0 – 49.5	2

$\sum f = 60$

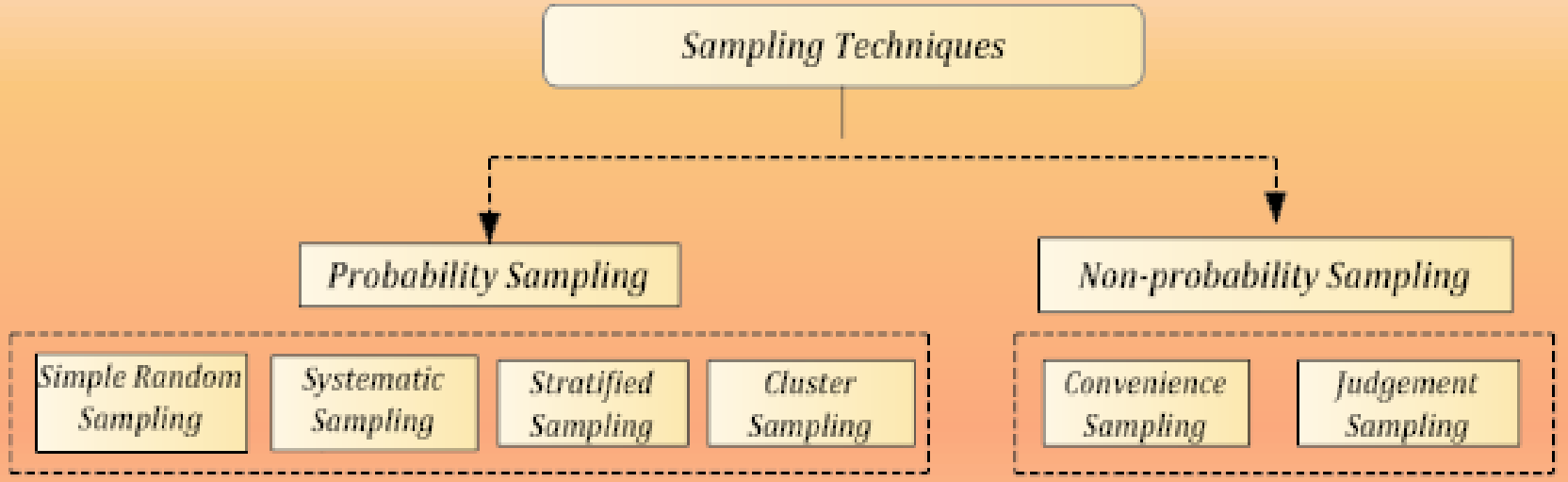


Histogram of the Salary Data

Sampling and Sampling Techniques

- A ***sample*** is a *part or a subset* of the population.
- ***Sampling*** is a systematic way of selecting a few items from the population.
- The *purpose of sampling* is to draw a conclusion or make a decision about the population parameters using the information contained in the sample statistics.

Sampling Techniques



Recent Trends

- Because of the advancement in technology, it is now possible to collect massive amounts of data.
- Data such as:

  web data, e-commerce, purchase transactions at retail stores, and bank and credit card transaction data, among more, is collected and warehoused by businesses.

- There has been an increasing amount of pressure on businesses to provide high quality products and services to improve their market share in this highly competitive market.

Recent Trends…*continued*

It is important for businesses to process and analyze a large amount of data efficiently in order to seek hidden patterns in the data. The processing and analysis of large data sets comes under the emerging field known as ***Data Mining*** and ***Big Data.***

Other Emerging Areas:

Business Analytics

Business Intelligence

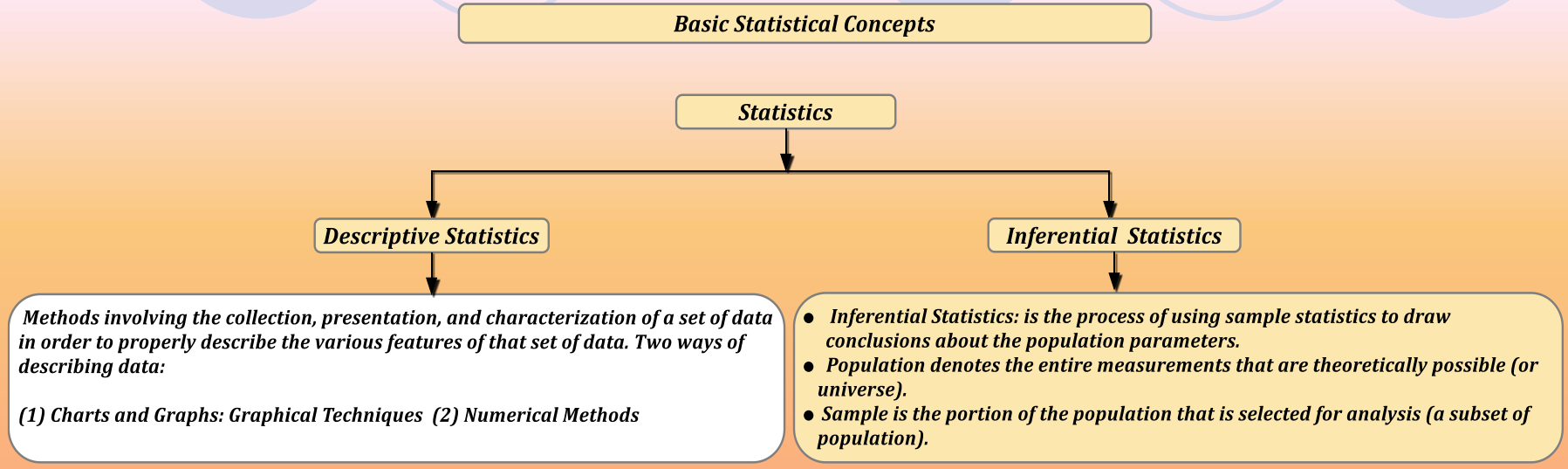
Summary and Review

Know the following concepts:

- **Various ways statistics is defined**
- **Importance of statistics in data analysis**
- **Reasons behind studying statistics**
- **Applications of statistics in different areas**
- **Limitations of statistics**
- **Concept of data and why and how data are collected**
- **Difference between a data set and a data point**
- **Types of data: qualitative vs. quantitative and cross-sectional data**
- **Discrete vs. continuous data**
- **Time series data**
- **Data elements**
- **Concept of variables**
- **Qualitative vs. quantitative variable**
- **Recent Trends in Statistics and Data Analysis**

Summary and Review

- **Descriptive and inferential statistics**
- **Statistical inference and the tools of inferential statistics**
- **Concepts of inferential tools: estimation and hypothesis testing**
- **Difference between a population and a sample**
- **Describing a population and a sample**
- **Parameter vs. statistic**
- **Levels of measurement: nominal, ordinal, interval, and ratio scales examples**
- **Operational definition**
- **Sampling and sampling techniques used in statistical analysis**
- **Symbols used to denote population parameters and sample statistics**



Population proportion

N=population size

μ = population mean

p = population proportion

σ^2 = population variance

σ =population standard deviation

Sample Statistic

n=sample size

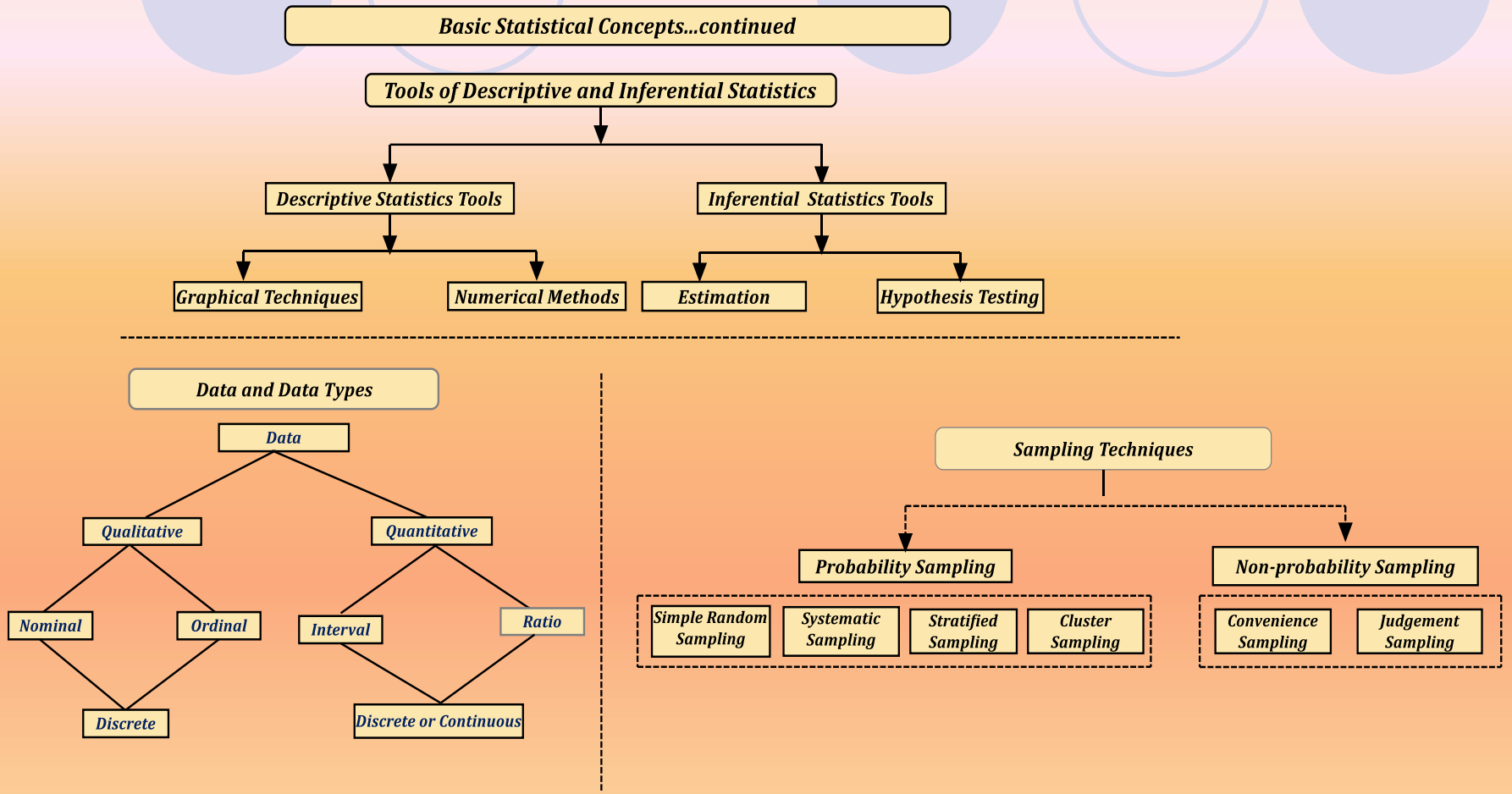
$\bar{x}$ = sample mean

$\bar{p}$ = sample proportion

s^2 = population variance

s =population standard deviation

Chapter 1: Basic Statistical Concepts: Flow Diagram (1)



Chapter 1 : Basic Statistical Concepts: Flow Diagram (2)