Basic Statistical Methods

Session 1

02/11/2020

Statistics

- Every minute of the working day, decisions are made by businesses around the world that determine whether companies will be profitable and growing or whether they will stagnate and die.
- Most of these decisions are made with the assistance of information gathered about the market place, the economic and financial environment, the workforce, the competition and other factors.
- Such information comes usually in the form of data or is accompanied by data.
- Statistics provides the tool through which such data are collected, analyzed and summarized and presented to facilitate the decision making process.
- For eg. a survey of 477 executives by The Association of Executive Search Consultants determined that 48% of the men and 67% of the women say they are more likely to negotiate for less business travel compared with five years ago.

Basic Statistical Concepts

- Population: A collection of persons, objects or items of interest.
- The population can be a widely defined category such as “all automobiles” or it can be narrowly defined such as “all Ford cars produced from 2016-2018”. A population can be a group of people such as “all workers presently employed by Microsoft”, or it can be a set of objects such as “all dishwashers produced on July 31,2018” by a specific manufacturing company.
- When a researcher gather data from the whole population for a given measurement of interest, it is called Census.

- Sample: A sample is a portion of the whole and is a representative of a whole population. Researchers often prefer to work with sample instead of entire population.
- For example, to conduct a quality control experiment to determine the average life of light bulbs, a light bulb manufacturer might randomly sample say 75 light bulbs during a production run. A HR manager may take a random sample of only 10 employees instead of taking a whole employees to save time and money.
- Descriptive Statistics: if a business analyst is using data gathered on a group to describe or reach conclusions about that same group, the statistics are called Descriptive.
- Inferential Statistics: If a researcher gathers data from a sample and uses the statistics generated to reach conclusions about the population from which the sample was taken, it is inferential statistics.
- One application of inferential statistics is in pharmaceutical research.

Parameter and Statistic

- A descriptive measure of the population is called a parameter. Parameters are usually denoted by Greek letters. Examples of parameters are Population mean (μ), Population variance (σ^2), and Population standard deviation (σ).
- A descriptive measure of a sample is called a Statistic. These are usually denoted by Roman letters. Examples of Statistic are sample mean ($\bar{x}$) , sample variance (s^2) and sample standard deviation(s).
- A analyst often wants to estimate the value of a parameter or conduct test about a parameter. However the calculation of parameters usually either impossible or infeasible because of the amount of time and money required to take a census. In such cases, a business researcher can take a random sample of the population , calculate a statistic on the sample and infer by estimation the value of the parameter.

Variables and Data

- Variable : In Statistics, a variable is a characteristic of any entity being studied that is capable of taking on different values. Some eg: are Return on investment, stock price, total sales, age of worker, time spent in shopping etc.
- Data : Data are recorded measurement. A measurement is when a standard process is used to assign numbers to particular attributes or characteristics of a variable.

Concept of measurement

- *Measurement* is the process of systematically assigning numbers to objects and their properties, to facilitate the use of mathematics in studying and describing objects and their relationships.
- Some types of measurement are fairly concrete: for instance, measuring a person's weight in pounds or kilograms, or their height in feet and inches or in meters.
- Measurement is not limited to physical qualities like height and weight. Tests to measure abstractions like intelligence and aptitude are commonly used in education and psychology.

Data Measurement

- Millions of numerical data are gathered in businesses everyday. All such data should not be analyzed the same way statistically because the entities represented by the numbers are different.
- For this reason , the business researcher needs to know the level of data measurement represented by the numbers being analyzed.
- There are four common levels of data measurements, namely

- Nominal
- Ordinal
- Interval
- Ratio

Nominal level

- The lowest level of data measurement is Nominal level. Numbers representing nominal level data can be used only to classify or categorize.
- Employee Identification Numbers is an example of Nominal data.
- Many demographic questions in surveys result in data that are nominal because the questions are used to classify only. If we assign numbers 1, 2, 3 etc., these numbers should be only to use classify the respondents/items. The number 1 does not denote the top classification.

Ordinal Level

- Ordinal level data measurement is higher than nominal level. In addition to nominal level capabilities, ordinal level measurement can be used to rank or order objects.
- For example, using ordinal data, a supervisor can evaluate three employees by ranking their productivity with the numbers 1 trough 3. The supervisor could identify one employee as most productive, one as the least productive and one as somewhere between by using ordinal data.
- However the supervisor could not use ordinal data to establish that the intervals between employees ranked 1 and 2 and between the employees ranked 2 and 3 are equal.ie, she could not say the differences in the amount of productivity between workers ranked 1, 2 and 3 necessarily the same.
- Likert-type scale are considered by many researchers to be ordinal.

Levels of data measurement (contd…)

Session 2

04/11/2020

Interval Level

- Interval level data measurement is next to the ordinal level of data in which distances between consecutive numbers have meaning and the data are always numerical.
- The distances represented by the differences between consecutive numbers are equal.
- An example of interval measurement is Fahrenheit temperature. With Fahrenheit temperature numbers, the temperature can be ranked, and the amounts of heat between consecutive readings, such as 38^o, 39^o and 40^o are the same.
- In addition, with interval level data, the zero point is a matter of convention or convenience and not a natural or fixed zero point. The zero does not mean the absence of the phenomenon.

Ratio Level

- Ratio level data measurement is the highest level of data measurement. Ratio data has the same properties of interval data, but ratio data have an absolute zero, and the ratio of the two numbers is meaningful.
- The notion of absolute zero means that zero is fixed, and the zero value in the data represents the absence of the characteristic being studied.
- Examples of ratio data are height, weight, time, volume etc. Many of the data gathered by machines in industry are ratio data.
- Other examples in business world are Production cycle time, work measurement time, number of trucks sold etc.

Comparison of the four levels of data

- Because Nominal and ordinal data are often derived from imprecise measurements such as demographic questions, the categorization of people or objects, or the ranking of items, these two levels of data are known as nonmetric data or qualitative data.
- Because interval and ratio level data are usually gathered by precise instruments often used in production and engineering processes, or in standardized accounting procedures, they are called metric data or quantitative data.

Collection of Data

Introduction

- Before collection of data for a given statistical enquiry, we should examine the following points.
- Objective and scope of enquiry
- Statistical units to be used
- Sources of information
- Method of data collection
- Degree of accuracy aimed at the final results
- Types of enquiry

Objective and scope of enquiry

- The first and foremost step in organizing any statistical enquiry is to define in clear concrete terms, the objectives of the enquiry. Scope of enquiry relates to the coverage w.r.t. the type of information, subject matter and geographical data.

Statistical units to be used

- A well defined and identifiable object or a group of objects with which the measurements or counts in any statistical investigation are associated is called Statistical Unit.
- A **statistical unit** is a **unit** of observation or measurement for which data are collected or derived
- Common examples of a unit would be a single person, animal, plant, manufactured item.

Collection of Data
Session 3
06/11/2020

Introduction

- Before collection of data for a given statistical enquiry, we should examine the following points.
- Objective and scope of enquiry
- Statistical units to be used
- Sources of information
- Method of data collection
- Degree of accuracy aimed at the final results
- Types of enquiry

Objective and scope of enquiry

- The first and foremost step in organizing any statistical enquiry is to define in clear concrete terms, the objectives of the enquiry. Scope of enquiry relates to the coverage w.r.t. the type of information, subject matter and geographical data.

Statistical units to be used

- A well defined and identifiable object or a group of objects with which the measurements or counts in any statistical investigation are associated is called Statistical Unit.
- A **statistical unit** is a **unit** of observation or measurement for which data are collected or derived
- Common examples of a unit would be a single person, animal, plant, manufactured item.

Sources of information

- For any statistical enquiry, the investigator may collect the data first hand or he may use the data from other published sources such as publications of the govt. and semi govt. organizations, magazines, Journals etc.

Method of data collection

- The problem does not arise if secondary data are used. However, if primary data are to be collected, a decision has to be taken whether (i) census method or (ii) sample technique.
- The census method can be applied in a situation where the separate data for every unit in the population is to be collected. For eg. ,the preparation of the voter's list for election purposes, income tax assessment, recruitment of personnel, etc. are some of the areas where the census method is adopted. This method can be used where the population is comprised of heterogeneous items, i.e. different characteristics.

Degree of accuracy aimed at the final results

- A decision regarding precision of results very much depends upon the objectives and scope of enquiry. In any statistical enquiry, perfect accuracy with the final result is impossible to achieve because of the errors in measurement, collection of data, its analysis and interpretation.

Types of enquiry

- The statistical enquiries may be of different types as given below.
- Official, Semi-official or Unofficial
- Initial or repetitive(Initial : Original and there is freedom for adopting any method of data collection) or (continuation : the old method is usually continued. It can only be modified to suit the new situation)
- Confidential or non-confidential
- Regular or Ad-hoc : If the enquiry is conducted periodically at equal intervals of time(monthly, weekly etc.), it is called regular. On the other hand, if an enquiry is conducted as and when necessary without any regularity or periodicity, it is termed as ad-hoc.
- Census or Sample
- Primary or Secondary

Types of Data

- 1) Primary and Secondary data
- 2) Internal and external data : Internal data of an organization are those which are collected by the organization from its own internal operations like production, sales, profit, imports and exports etc. are used by its own purpose. On the other hand, external data are those which are obtained from the publications of some other agencies like governments(central/state) , private research institutions etc. for use by the given organization.

- 3) Qualitative and Quantitative data
- Qualitative data : Data in which classification of objects is based on attributes and properties like Social status, Gender, Nationality, Occupation etc.
- For eg: population of whole country can be classified into married, unmarried, widowed, divorced etc.
- Quantitative data: Data which can be measured and expressed numerically like weight in kilogram, Height in cm etc.
- For eg: population of whole country may be classified according to different variables like age, income etc.

Classification of Quantitative data

- Quantitative data can again be classified into Discrete and continuous.

➢ (a) : Discrete : Data which can take up only exact values and not any fractional values are called discrete data.

- Eg:- Number of workmen in a factory

  Number of telephone calls during a specific time.

➢ (b) : Continuous data: These are data which can take up any numerical value within a certain range.

- Eg:- Height in cm, Weight in kg, Rainfall in mm, Time, temperature etc.

Chronological(Time series) Data

- Chronological data : Data collected in a chronological manner(time based) are called time series data.
- Eg: Population of country in several decades

  Monthly production of a company

  Yearly rainfall in India

Methods of collecting Primary Data

- Direct Personal Investigation
- Indirect Oral Interviews: Under this method of collecting data, the investigator contacts third parties called witnesses capable of supplying the necessary information. The method is generally adopted in those cases where the information to be obtained is of a complex nature and the informants are not inclined to respond if approached directly. For example, in an enquiry regarding addiction to drugs, alcohol, etc., people may be reluctant to supply information about their own habits.
- Information received through local agencies

Methods of collecting Primary Data (contd…)

- Mailed questionnaire method

A questionnaire is a list of questions which are answered by the respondent himself in his own hand writing

- Schedules sent through enumerators

Schedule is a device of obtaining answers to the given questions in a form which is filled by the interviewers or enumerators in a face to face situation with respondents.

Drafting/ Framing the questionnaire

- The size of the questionnaire should be as small as possible.
- The questions should be clear and brief.
- The questions should be arranged in natural logical order
- The usage of vague and multiple meaning words should be avoided
- Questions of sensitive and personal nature should be avoided.
- Types of questions: Shut and Open
- Leading questions should be avoided
- Pre-testing the questionnaire
- Covering letter.

Organization of data: Classification and Tabulation

Session 4

25/11/2020

Introduction

- After collecting the data, the next important step is to organize it, i.e., to present it in a readily comprehensible and condensed form which will highlight the important characteristics of the data.

- The presentation of data is broadly classified into two

➢ Tabular presentation

➢ Diagrammatic and Graphical presentation

Classification

- The process of arranging data into groups or classes according to resemblances and similarities is called classification.
- Classification of data is preliminary to tabulation.
- Functions of classification

➢ It condenses the data

➢ It facilitates comparison

➢ It helps us to study the relationships.

➢ It facilitates the statistical treatment of the data

Types(Bases) of Classification

- Geographical Classification

If the classification is based on the geographical/ locational differences between the various items in the data like state, cities, regions, zones, area etc.

Eg: Density of population (per sq km) in different cities of India.

Cities	Density of population
Kolkata	685
Mumbai	684
Delhi	423
Chennai	205
Chandigarh	48

- Chronological Classification

Chronological classification is one in which the data are classified on the basis of differences in time. The time series data which are quite frequent in Economics and Business Statistics are generally classified chronologically.

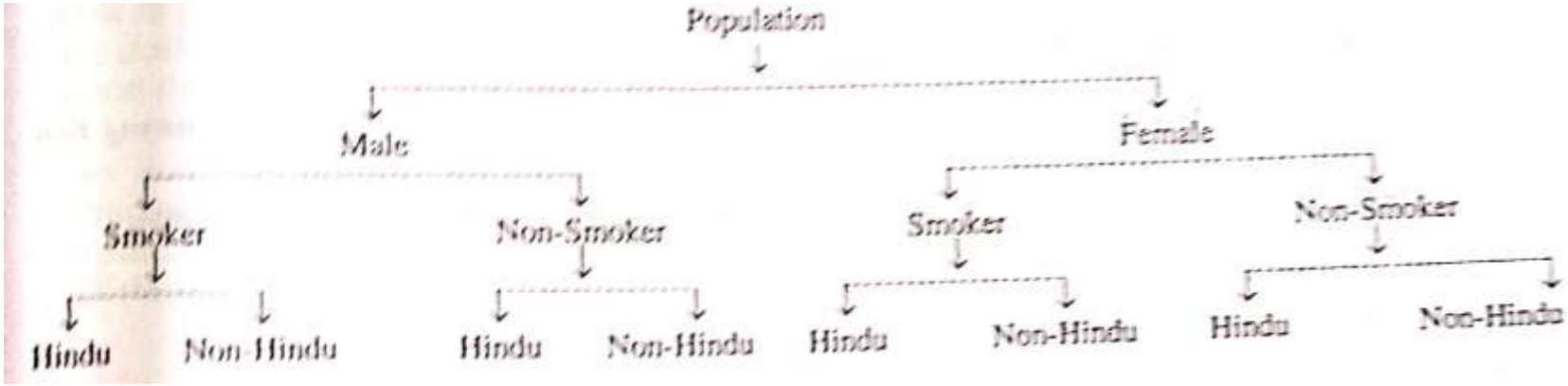
Eg: Population in India

Year	Population
1981	-
1991	-
2001	-
2011	-

- Qualitative Classification: When the data are classified according to qualitative phenomena, which are not capable of quantitative measurement like honesty, beauty, employment etc., the classification is termed as Qualitative.
- In classification if the data are classified according to presence or absence of an attribute, it is termed as simple or dichotomous classification.
- On the other hand, if the population is classified more than two classes w.r.t. a given attribute, it is said to be manifold classification.
- Eg: attribute intelligence in different classes

Genius Highly Intelligent Average Intelligent etc

Another example of manifold classification



- Quantitative Classification : If the data are classified on the basis of phenomenon which is capable of quantitative measurement like age, height, price, production etc., it is termed as quantitative classification. The quantitative phenomena under study is called variable and hence it is also knows as classification by variable.
- For eg: Daily earnings (in ’00 Rs.)of 60 departmental stores

Daily earnings	Number of stores
Upto 100	20
101-200	18
201-300	14
301-400	8

Tabulation

- Tabulation is one of the most important device of presenting the data in a condensed form to furnish maximum information in minimum possible space.
- Tabulation is the final stage in collection and compilation of data and opens the gateway for further statistical analysis and interpretations.

Format of a blank table

Table Number: 1

Title

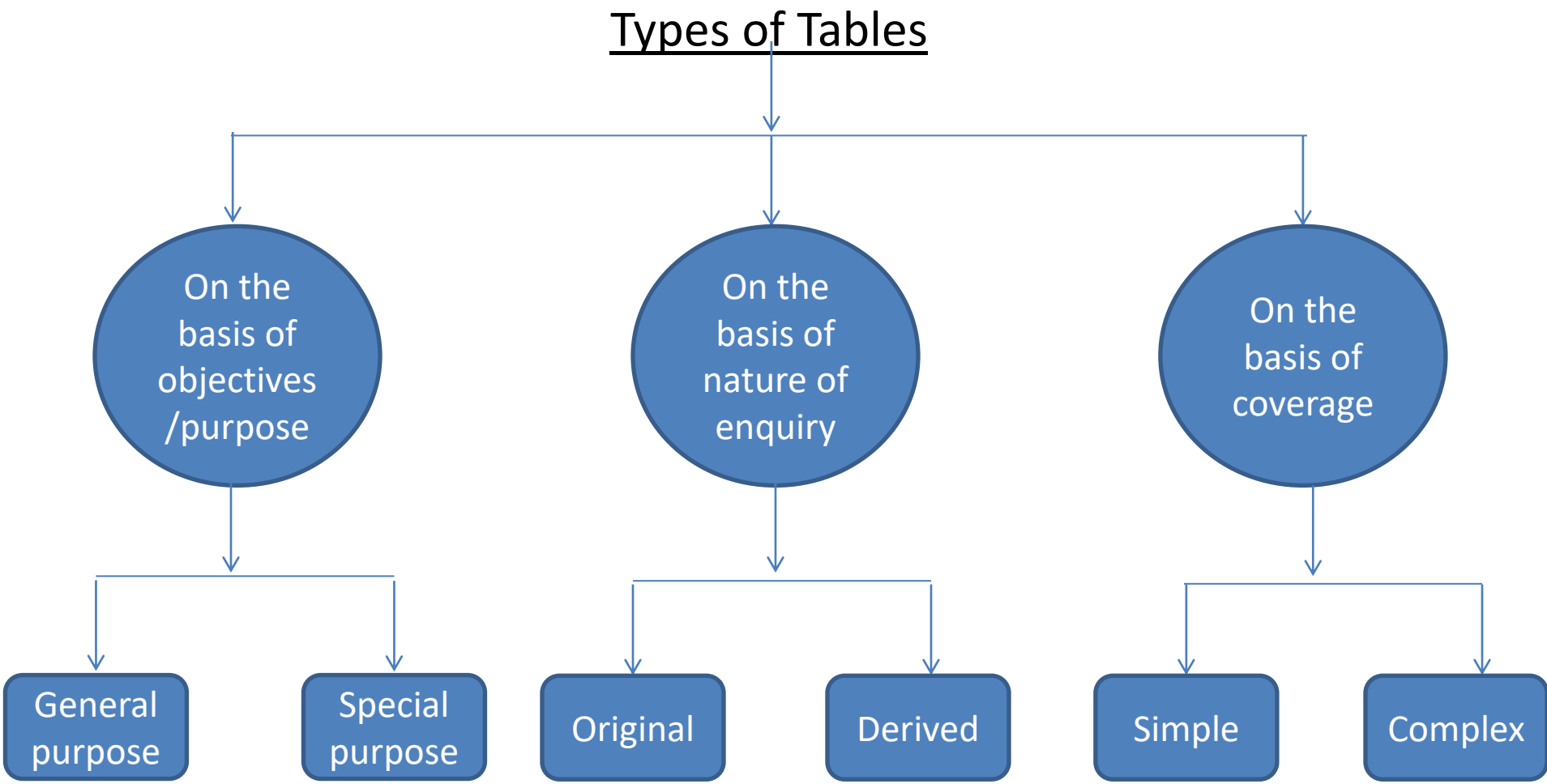
(Head Note, if any)

Stub	Caption					Total
	Sub-heads		Sub-Heads			
	Column head	Column head	Column head	Column head	Column head	
	Body of the table					
Total						

Footnote:

Source note:

Types of Tables



- General Purpose (reference) table and Special purpose (Summary) table

General purpose tables are systematically arranged data usually in chronological order in a form which is suitable for ready reference and record with out any intension of comparative studies, relationship or significance of figures.

Eg: Most of the tables prepared by government agencies.

Special Purpose tables are of analytical nature and are prepared with the idea of making comparative studies, studying relationship and significance of figures provided by the data. In such tables, interpretive figures like ratios, percentages etc. are used in order to facilitate comparisons.

Special Purpose/Summary table

An example of special purpose or summary tables.

Relationship Between the Total Number of Persons Died in Industrial Accidents and Persons Died in Coal Mines

Year	Persons Died in Industrial Accidents	Persons Died in Coal Mines	Persons Died in Coal Mines as a % in Total Deaths in Industrial Accidents
1976	930	150	16.1
1977	1,154	285	24.7
1978	1,250	115	9.2
1979	930	108	12.0
1980	1,350	270	20.0

- Original (Primary) and Derived(Derivative) table

In a primary table, the statistical facts are expressed in the original form. It therefore contains absolute and actual figures and are not rounded numbers or percentages.

On the other hand, derived table is one which contains figures and results derived from the original or primary data. It express the information in terms of aggregates or statistical measures like average, dispersion, skewness etc.

For instance, time series data is expressed in a primary table but a table expressing trend values, seasonal and cyclic variations is a derived table.

- Simple and complex tables

In a simple table, data are classified with respect to a single characteristic and accordingly it is also termed as one-way table.

Eg: Agricultural output of different countries(in kg per hectare)

If the data are grouped into different classes with respect to two or more characteristic, or criterion simultaneously, we get a complex(manifold) table.

Eg: Distribution of number of students in a college w.r.t.

age and gender

Distribution of a given population w.r.t. age, gender and literacy

A simple illustration of table

A table showing classification of 50,000 students of a university according to their faculty and gender

Faculty	Gender of student		Total
	Boys	Girls	
Commerce	10,000	7500	17500
Arts	5000	10,000	15,000
Science	6000	4000	10,000
Engineering	3500	1500	5000
Medical	1250	1250	2500
Total	**25,750**	**24,250**	**50,000**

Data Quality and Issues

Session 5

27/11/2020

Data Quality Issues

- Success of ML largely depends on the quality of the data.
- A data which has the right quality helps to achieve better prediction accuracy in case of supervised learning.
- There are multiple factors which lead to the data quality issues.

Data Quality Issues

- Incorrect sample set selection
- Random responding and motivated mis- responding
- Who are the true respondents?
- Errors in data collection
- Missing values
- Extreme values

Incorrect sample set selection

- The data may not reflect normal or regular quality due to incorrect selection of sample set.
- Eg: sales transaction from a festive period to predict for future period.
- If you are interested in studying the effects of smoking on a particular outcome, you must define what it means to be a smoker which helps make the research more precise.
- For example do you include people who just smoke occasionally? What about those smokers who previously used tobacco but have stopped?

Can data cleaning fix sampling problems?

- Unfortunately in many cases, poor sampling cannot be corrected by data cleaning.
- The goal of sampling is to gather data on whatever phenomenon is being studied in such a way as to make the best case possible for withdrawing inferences about the population of interest.

Random Responding, Motivated Misresponding

- Let me ask you “How are you doing today”. Ever notice that some people tend to flip between extremes(e.g. wonderful or horrible) while others seem to be more stable (e.g. OK or fine) no matter what is going on. This is an example of one sort of Response set, where individual tend to vary in a narrow band around the average or vary around extremes.

Response Set

- A response set is a strategy people use (consciously or otherwise)when responding to educational tests, questionnaires, or things like psychological tests (or even questions paused in casual conversation in the above example).
- Researchers should pay more attention to response sets and effects of random responding which can substantially increase probability of Type II errors.

Common types of Response Sets

- Random Responding
- Malingering and Dissimulation
- Social desirability
- Other response styles
- Response styles peculiar to educational testing

Common types of Response Sets

- Random responding : Random responding is a response set in which individuals respond with little pattern or thought. The behaviour adds substantial error variance to analysis which completely make ineffective the usefulness of responses.
- This may be motivated by lack of preparation, reactivity to observation, lack of motivation to cooperate with testing or disinterest.
- If we are not careful, participants with lower motivation to perform at their maximum level may increase the odds of Type II errors masking real effects of our research through response sets such as random responding.

- Malingering is a response set where individuals falsify and exaggerate answers to appear weaker or more medically or psychologically symptomatic, often motivated by the goal of receiving services they would not otherwise be entitled to.
- Someone might pretend to be injured so they can collect an insurance settlement or obtain prescription medication.
- Dissimulation refers to a response set in which respondents falsify answers in an attempt to be seen in more negative or more positive side than honest answers can provide.
- These response sets are common in psychological test like “Do you have suicidal thought?”

- Social Desirability : Social desirability is related to malingering and dissimulation in that it involves altering responses in systematic ways to achieve a desired goal. In this case to behave in the same way as most other people in a group or society, the respondent want to “look good” to the examiner.

- Other response styles : Other response styles such as acquiescence and criticality are response patterns wherein individuals are more likely to agree with (acquiescence) or disagree with (criticality) questionnaire items in general, regardless of the nature of the item.

- Response styles peculiar to educational testing: The biases peculiar to tests of academic mastery (often multiple choice) include :

  (a) response bias for particular column on multiple choice items

  (b)bias for or against guessing when uncertain of the correct answer.

Detecting random responding in your research

- An important issue is whether we can be confident that what we call random responding truly is random?
- There is a large and well developed literature on how to detect many different types of response sets.
- Examples include addition of particular types of items to detect social desirability , altering instructions to respondents in particular ways, creating equally desirable items worded positively and negatively, use item response theory (ITR) to explicitly estimate a guessing parameter.

Item Response Theory

- One application of IRT has implications for identifying random responders using the theory to create person-fit indices.
- The idea behind this approach is to quantitatively group individuals by their pattern of responding and then use these groupings to identify individuals who deviate from an expected pattern of responding. Also it is possible to estimate the guessing parameter.
- One drawback of ITR is that it generally requires large samples.

Data Quality Issues and Data Cleaning: Identifying outliers

Session 6

02/12/2020

Data Cleaning

To illustrate the need for cleaning up of data, consider the data given below

Customer Id	Zip	Gender	Income	Age	Marital Status	Transaction amount
1001	10048	M	75,000	C	M	5000
1002	J2S7K7	F	-40,000	40	W	4000
1003	90210		10,000,000	45	S	7000
1004	6269	M	50,000	0	S	1000
1005	55101	F	99,999	30	D	3000

Introduction

- Authors spend a great deal of time describing the importance of the study, the research methods, analysis etc., but rarely mention having screened their data for outliers or extreme observations.
- Jumping from data collection to data analysis is a wrong practice.
- Some techniques such as “robust” procedures and non-parametric tests (which do not require an assumption of normally distributed data) exists in the literature.
- Parametric tests are rarely robust to violations of distributional assumptions.

What are extreme scores

- An outlier(extreme score) is generally considered to be a data point that is far outside the norm for a variable or population.
- It is an observation that deviates so much from other observations in the data.
- It is not to say that extreme scores are not of value, but they should be examined more closely in depth.
- The literature on extreme scores reveals two broad categories: outliers and fringeliers

Outliers and Fringeliers

- Outliers are clearly problematic that they are far from the rest of the distribution.
- Fringeliers are those scores around +/-3.0 standard deviations from the mean, which represents a good rule of thumb for identifying scores that merit further examination.
- Why we are concerned with scores around +/-3.0 SD from the mean?

What cause extreme scores

- It can arise from different mechanisms or causes.

➢ Extreme scores from data errors like human errors

➢ Extreme scores from Intentional or motivated misreporting

(eg: responses to income, study time, educational attainment etc.)

➢ Extreme scores from sampling error or bias: A **sampling error** is a statistical **error** that occurs when an analyst does not select a **sample** that represents the entire population of data and the results found in the **sample** do not represent the results that would be obtained from the entire population

➢ Extreme scores from faulty distributional assumptions.

How extreme values affect statistical analysis?

- Increase error rates and decrease the quality and precision of your results
- Decrease power of the test by altering skew or kurtosis of a variable
- Seriously bias or influence estimates that may be of substantive interest such as mean, SD etc.

Identification of extreme scores

- Empirical rule
- Box and whisker plot
- Standardized residuals in case of bivariate and multivariate extreme scores.
- Mahalanobis distance or cook’s distance

Empirical Rule

- Empirical Rule : This is an important rule of thump used to state the approximate percentage of values that lie within a given number of standard deviations from the mean of a set of data if the data are normally distributed.

Distance from the mean	Values within distance
$\mu \pm 1\sigma$	68%
$\mu \pm 2\sigma$	95%
$\mu \pm 3\sigma$	99.7%

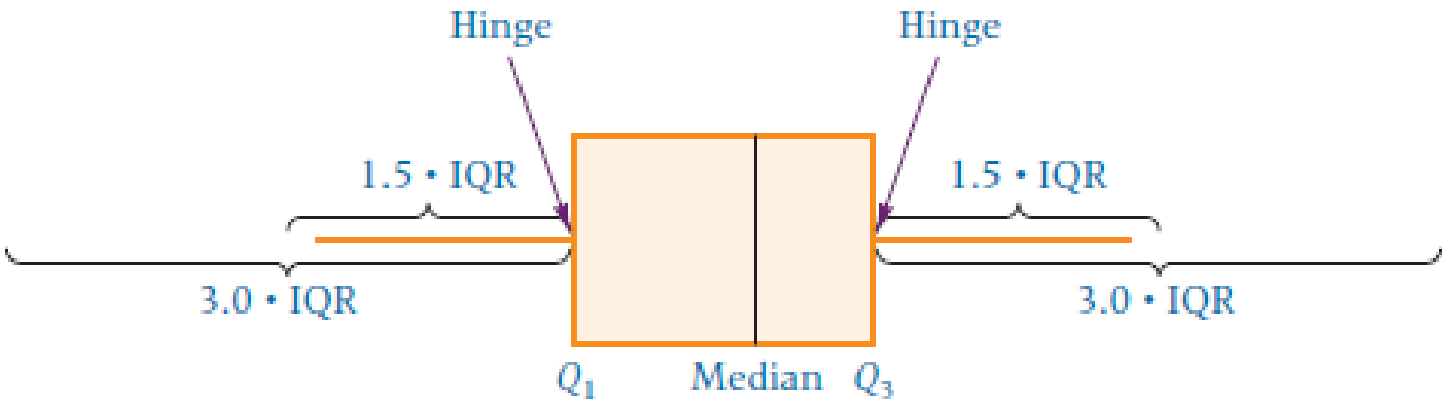
(Based on the assumption that the data are approximately normally distributed)

Box-and-Whisker Plots

- A box-and-whisker plot (Box-plot) is a diagram that utilizes the upper and lower quartiles along with the median and the two most extreme values to depict a distribution graphically.
- The plot is constructed by using a box to enclose the median along a continuum to the lower and upper quartiles.
- Quartiles : Quartiles are measures of central tendency that divide a group of data in to four parts. The three quartiles are denoted as Q_1,Q_2 and Q_3.

Box-and-Whisker Plots (contd…)

- From the lower and upper quartiles, lines referred to as whiskers are extended out from the box toward the outermost data values.
- The box end points (Q_1 and Q_3) are referred to as the hinges of the box.



Five Number summary

- Box and Whisker plot is determined by five specific numbers referred to as five number summary

  1.The median (Q_2)

  2. The lower quartile (Q_1)

  3. The upper quartile (Q_3)

  4. The smallest value (Min)

  5. The largest value (Max)

Box-and-Whisker Plots (contd…)

- Next the value of Inter Quartile Range (IQR= Q_3-Q_1) is to be computed. IQR includes the middle 50% of the data and should equal to the length of the box.
- A whisker, a line segment is drawn from the lower hinge of the box outward to the smallest data value and a second whisker is drawn from the upper hinge of the box outward to the largest data value.
- At a distance of 1.5IQR outward from the lower and upper quartiles are what are referred to as inner fences.

- The inner fences are established as follows.

$$Q_1\text{-}1.5\ IQR$$

$$Q_3\text{+}1.5\ IQR$$

- If data falls beyond the inner fences, then outer fences can be constructed.

$$Q_1\text{-}3\ IQR$$

$$Q_3\text{+}3\ IQR$$

- Values in the data distribution that are outside the inner fences are referred to as mild outliers.
- Values that are outside the outer fences are called extreme outliers.
- Thus one of the main uses of Box-and-Whisker plot is to identify outliers.

Data Quality and Issues (contd…) Identifying outliers and handling missing data

Session 7

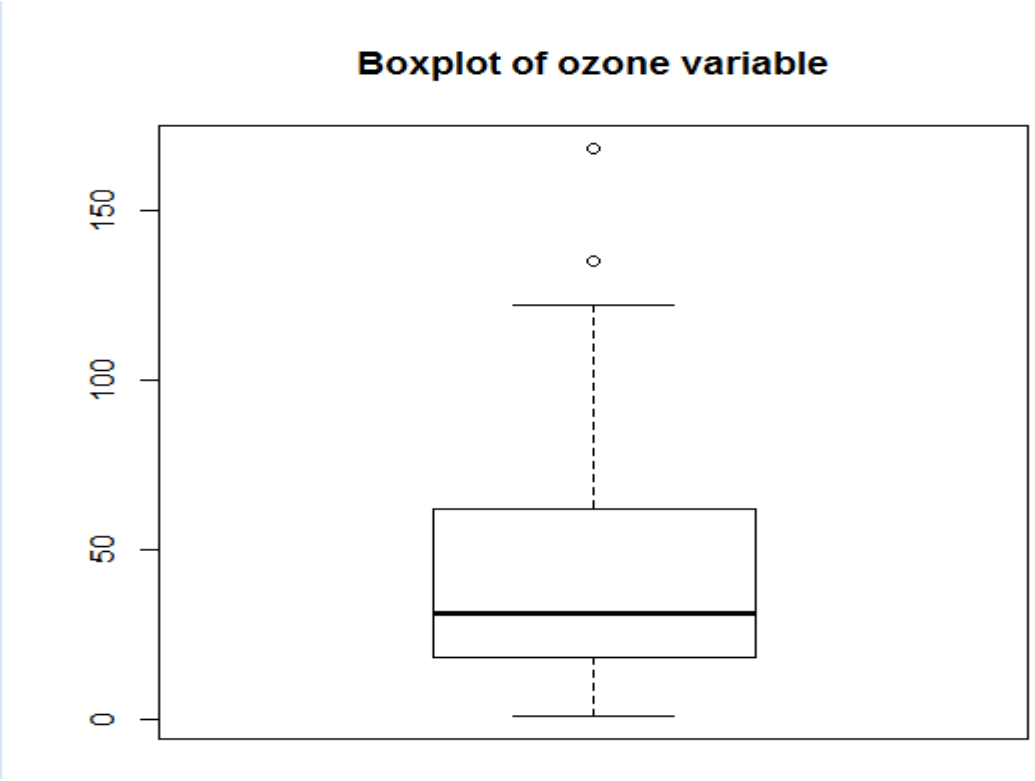
04/12/2020

The data “ozone” contains variables ozone(ozone concentration), wind(wind speed), temp(air temperature) and rad(intensity of solar radiation). The objective is to know how is ozone concentration related to wind speed, air temperature and the intensity of solar radiation.

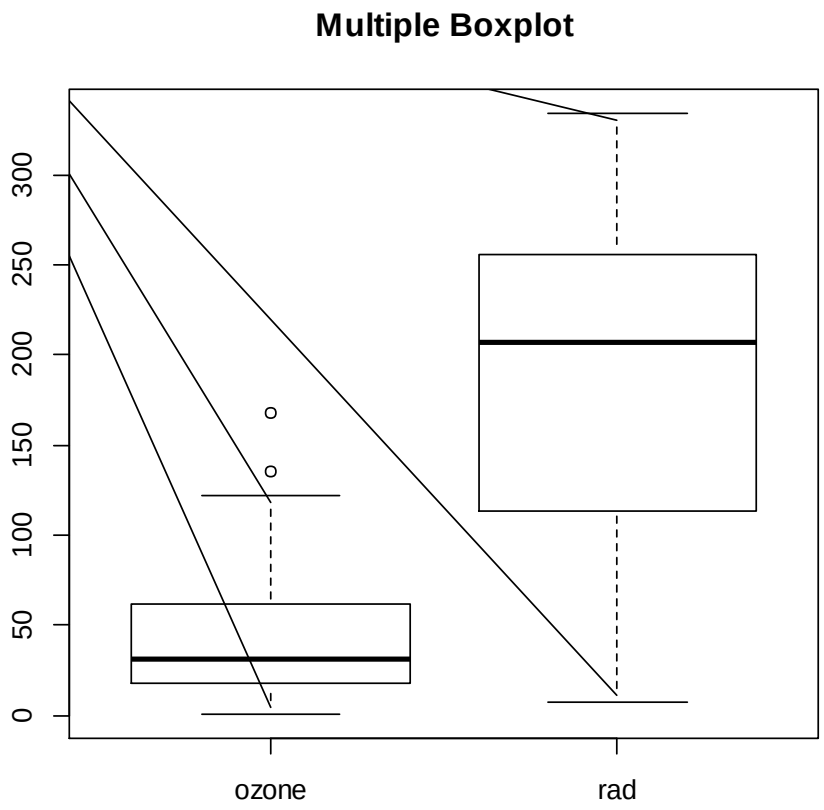
	rad	temp	wind	ozone
1	190	67	7.4	41
2	118	72	8.0	36
3	149	74	12.6	12
4	313	62	11.5	18
5	299	65	8.6	23
6	99	59	13.8	19

rad	temp	wind	ozone
Min. : 7.0	Min. :57.00	Min. : 2.300	Min. : 1.0
1st Qu.:113.5	1st Qu.:71.00	1st Qu.: 7.400	1st Qu.: 18.0
Median :207.0	Median :79.00	Median : 9.700	Median : 31.0
Mean :184.8	Mean :77.79	Mean : 9.939	Mean : 42.1
3rd Qu.:255.5	3rd Qu.:84.50	3rd Qu.:11.500	3rd Qu.: 62.0
Max. :334.0	Max. :97.00	Max. :20.700	Max. :168.0

Boxplot:R zone



Multiple Box-plot : R zone



Identifying outlier value and corresponding row number and printing

```
> boxplot.stats(Ozone$ozone)$out
[1] 135 168
> out <- boxplot.stats(Ozone$ozone)$out
> out_row <- which(Ozone$ozone %in% c(out))
> out_row
[1] 34 77
> Ozone[out_row, ]
   rad temp wind ozone
34 269   84  4.0   135
77 238   81  3.4   168
>
```

Missing Data

- In a dataset, one or more data elements may have missing values in multiple records.
- It may be due to a person who is collecting the sample data or by the responder, primarily due to his/her unwillingness to respond or lack of understanding needed to provide a response.
- It is rare that a database contains no missing values at all. How the analyst deals with the missing data may change the outcome of the analysis.
- There are multiple strategies to handle missing values.

Need for Imputation of Missing data

- It is important to learn methods for handling missing data that will not bias the results.
- Delete Records Containing Missing Values?

  Not necessarily a best approach

  Deleting records creates biased subset

  Valuable information in other fields lost

Methods to Handle Missing Data

- Replace Missing Values with User-defined Constant

  Missing numeric values replaced with 0.0.

  Missing categorical values replaced with "Missing".

- Replace Missing Values with Mode or Mean

  Mode of categorical variables.

  Mean for numeric variables.

Methods to Handle Missing Data

- Replace Missing Values with Random Values from the observed distribution of the variable

  Values randomly taken from underlying distribution.

  Method superior compared to mean substitution.

  Measures of location and spread remain closer to original.

Methods to Handle Missing Data

- Replace Missing values with imputed values.

  Estimate the missing value using regression given that all other attributes for a particular record.

  Estimate the missing value using multivariate imputation by chained equation.

The Imputation Method

- For continuous variables, the imputation is done with regression.
  - Estimate the missing value by taking the variable as response given that all other values.
  - It uses stepwise regression (forward elimination) to include significant variables.
- For categorical variables, classification algorithm such as CART will be used.
- The general question asked is “What would be the most likely value for this missing record, given all the other attributes for a particular records?”.

Imputation by Stepwise Regression

- First prepare the data for multiple regression, specially categorical variables must be converted to dummy variables.
- Let Y is the response variable and (X_1, X_2, …., X_p) are *p*-predictors.
- Let the predictor X_2 has a missing value corresponding to a record say R7.
- Consider the predictor X_2 as response variable and all the original predictors (minus X_2) represent predictors.
- Do not include the original response variable Y as predictor for imputation.
- Apply stepwise variable selection method on X_2.

- The model starts with no predictors, then the most significant predictor is entered into the model, followed by next most significant predictor and so on till all significant predictors have been entered into the model and no further predictors have been dropped.

- Once the model is built, use the predictors to estimate the value of X_2.

Imputation by MICE-Procedure

- How do we impute if there are missing values in different predictors?-MICE.
- MICE assumes the data are *Missing at Random* (MAR) which means that the probability of a missing value depends only on the observed values and can be predicted using them.
- Procedure (Azur, et.al., 2011)

Step-1: Replace (impute) the missing values in each variable with temporary "place holder" values derived from the observed values of the variables. This can be done by mean imputation.

Step-2: The "place holder" mean imputations for one variable (say X_1) are set back as missing.

Step-3: Regress X_1 based on other variables using the observed values. Drop all records where X_1 is missing during the model fitting.

Step-4: The missing values for X_1 are then replaced with predictions from the regression model.

Step-5: Repeat steps 2-4 for each variable that has missing values.

MICE Cont'd.

- Applying Steps 2-5 once for each variable constitute a "cycle".
- MICE requires a number of such cycles and at each cycle the missing values are updated.
- Generally, the number of cycles is taken as 5 which is specified in advance.
- Retain the imputed values corresponding to the final cycle.

- Note that assumptions on regression model must be taken care off.

Reference

- Azur, M.J., Stuart, E.A., Frangakis, C., & Leaf, P.J. (2011), *Multiple Imputation by Chained Equations: What is it and how does it work?*, International Journal of Methods in Psychiatric Research, 20(1), 40-49.

Missing value imputation MICE: How does it work?

Session 8

07/12/2020

A simple illustration

Age	Experience	Income
25		49
45	19	34
39	15	11
35	9	100
35	8	45
37	13	29
53	27	
	24	22
35	10	81

- Impute column mean values.

	Age	Experience	Income
	25		49
	45	19	34
	39	15	11
	35	9	100
	35	8	45
	37	13	29
	53	27	
		24	22
	35	10	81
Mean	38	15.625	46.375

Impute mean values as place holders:
Base data
Are the values reliable?

Age	Experience	Income
25	15.625	49
45	19	34
39	15	11
35	9	100
35	8	45
37	13	29
53	27	46.375
38	24	22
35	10	81

Remove place holder from left most column and fit a model with age as dependent and experience and income as predictors but with 8 observations.

Age	Experience	Income
25	15.625	49
45	19	34
39	15	11
35	9	100
35	8	45
37	13	29
53	27	46.375
35	10	81
	24	22

- The estimated regression equation is
- Age= 23.43820 +0.95885(Exp)+0.01201(Inc)
- Age= 46.70

```
Call:
lm(formula = Age ~ Exp + Inc)

Residuals:
      1       2       3       4       5       6       7       8
-13.9987  2.9452  1.0568  1.7414  3.3605  0.7584  3.1258  1.0106

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.42840    9.81298   2.387   0.0626 .
Exp          0.95885    0.44737   2.143   0.0850 .
Inc          0.01201    0.09777   0.123   0.9071
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Impute the estimated Age value in the first column and remove the place holder of second column ‘Experience’. Now repeat the procedure by taking ‘Experience’ as dependent variable and age and Income as independent variables.

Age	Experience	Income
25		49
45	19	34
39	15	11
35	9	100
35	8	45
37	13	29
53	27	46.375
46.705	24	22
35	10	81

- The estimated regression equation is
- Exp = -23.30727 + 0.98402(age) -0.02455(Inc)
- Exp. = 0.09028

```
Call:
lm(formula = Exp ~ Age + Inc)

Residuals:
     1      2      3      4      5      6      7      8
-1.1391  0.2004  0.3214 -2.0288  0.6103 -0.7075  1.8885  0.8549

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -23.30727    4.12552  -5.650 0.002413 **
Age           0.98402    0.08933  11.016 0.000107 ***
Inc          -0.02455    0.02001  -1.227 0.274450
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Impute the estimated Experience value in the second column and remove the place holder of third column ‘Income’. Now repeat the procedure by taking ‘Income’ as dependent variable and age and Experience as independent variables.

Age	Experience	Income
25	0.09208	49
45	19	34
39	15	11
35	9	100
35	8	45
37	13	29
53	27	
46.705	24	22
35	10	81

- The estimated regression equation is
- Inc. = -128.005+7.642(age) -8.971(Exp)
- Inc = 34.804

```
Call:
lm(formula = Inc ~ Age + Exp)

Residuals:
      1       2       3       4       5       6       7       8
-13.215 -11.431 -24.463  41.279 -22.692  -9.121   8.394  31.250

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -128.005    240.301  -0.533    0.617
Age            7.642      9.178   0.833    0.443
Exp           -8.971      8.469  -1.059    0.338
```

Now 1 cycle completed the following imputed values

Age	Experience	Income
25	0.09208	49
45	19	34
39	15	11
35	9	100
35	8	45
37	13	29
53	27	34.804
46.705	24	22
35	10	81

Now find the difference between base data and the first cycle data

Age	Experience	Income
25	15.625	49
45	19	34
39	15	11
35	9	100
35	8	45
37	13	29
53	27	46.375
38	24	22
35	10	81

Age	Experience	Income
25	0.09208	49
45	19	34
39	15	11
35	9	100
35	8	45
37	13	29
53	27	34.804
46.705	24	22
35	10	81

Age	Experience	Income
0	15.53292	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	11.571
8.70502	0	0
0	0	0

MICE procedure (contd…)

- The whole procedure described here repeats for a number of cycles and at each cycle the missing values are updated.
- Retain the imputed values corresponding to the final cycle.
- How do we know final cycle?
- Note that assumptions on regression model must be taken care off.

Pre-processing of data

Data transformation

- Variables tend to have ranges that vary greatly from each other. For example, if we are interested in major league baseball, players' batting averages will range from 0 to less that 0.400, while the number of homeruns hit in a season will range from 0 to around 70. Such case, greater variability in homeruns will dominate the lesser variability in batting averages.
- One should normalize their numerical variables in order to standardize the scale of effect each variable has on the results.

➢ Min-Max Normalization: X_{mm} = X-min(x)/Range(x).

Z-score

- Z-score is obtained by taking the difference between field value and the field mean value and scaling this difference by S.D. of the field values.
- Z Score Standardization: Z score = $\frac{X - mean(X)}{S.D(X)}$
- Cars data : variable weight

Mean	3005.490
Min	1613
Max	4997
Range	3384
S.D	852.646

Z-score

- For the vehicle weighing only 1613 pounds, the Z-score is
- Z score = $\frac{X-mean(X)}{S.D(X)}$= $\frac{1613-3005.490}{852.646}$ =~ -1.63
- For ‘average’ vehicle (if any)
- Z score = $\frac{X-mean(X)}{S.D(X)}$= $\frac{3005.490-3005.490}{852.646}$ =0
- For the heaviest car, the z-score is
- Z score = $\frac{X-mean(X)}{S.D(X)}$= $\frac{4997-3005.490}{852.646}$ =~ 2.34

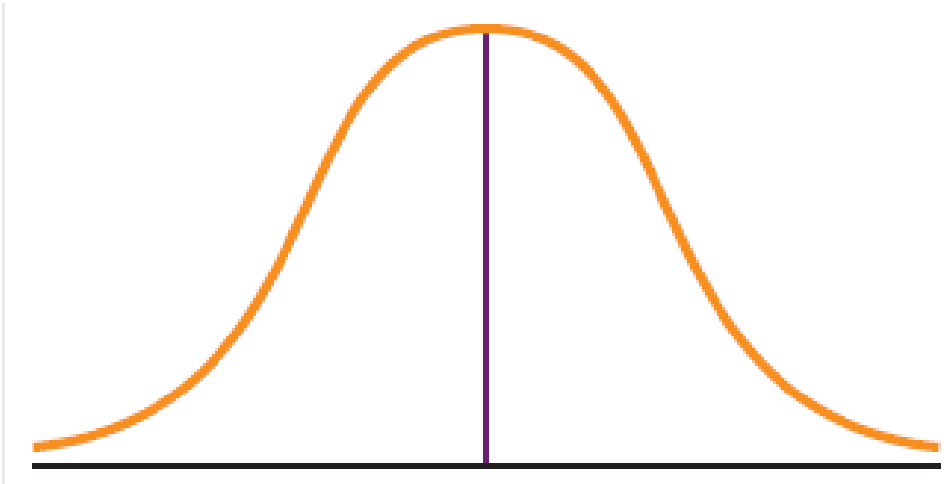
Data pre-processing

Session 9

09/12/2020

Transformations to achieve Normality

- Some data mining algorithms and statistical methods require that the variables be normally distributed.



Transformations to achieve Normality

- To make our data more normally distributed, we must first make it symmetric, which means eliminating the skewness.
- To eliminate skewness, one can apply a transformation to the data.
- Common transformations are log transformation (ln), the square root transformation and the inverse square root transformation

Contd…

- After achieving symmetry, to check for normality, we construct a normal probability plot.
- Don’t forget to de-transform the data
- $y=1/\sqrt{x}$, then $x=1/y^{2}$.

Flag (Indicator/dummy) variables

- Some analytical methods such as regression require predictors to be numeric. Thus analysts wishing to use categorical predictors in regression need to recode the categorical variable into one or more flag variables.
- For example the categorical predictor ‘gender’ taking values for female and male could be recorded into the flag variable as follows

  If gender= female then gender_flag=0;

  If gender=male then gender_flag=1

Contd…

- When a categorical predictor takes k≥ 3 possible values, then define k-1 dummy variables and use the unassigned category as the reference category. For example if a categorical predictor region has k=4 possible categories, {north, east, south, west}, then the analyst could define the following

  k-1=3 flag variables.

north_flag: If region=north then north_flag=1; otherwise

north_flag=0

east_flag: If region=east then east_flag=1; otherwise

east_flag=0

south_flag: If region=south then south_flag=1; otherwise

south_flag=0

- The flag variable for the west is not needed, instead the unassigned category becomes the reference category

A simple example how flag variables will work

	A	B	C	D	E
1	ID	Region	Age	Income	Personal Loan
2	1	South	25	49	No
3	2	West	45	34	No
4	3	North	39	11	Yes
5	4	East	35	100	No
6	5	East	35	45	No

A	B	C	D	E	F	G	H	I
ID	Region	north_flag	East_flag	South_flag	west_flag	Age	Income	Personal Loan
1	South	0	0	1	0	25	49	No
2	West	0	0	0	1	45	34	No
3	North	1	0	0	0	39	11	Yes
4	East	0	1	0	0	35	100	No
5	East	0	1	0	0	35	45	No

Transforming Categorical variables into Numerical variables

- Common error

Region	Region number
North	1
East	2
South	3
West	4

- The algorithm now erroneously thinks the following.

➢ The four regions are ordered.

➢ West>South>East>North

➢ West is 3 times closer to south compared to north and so on.

Contd…

- Data analyst should avoid transforming categorical variables to numerical variables except for that are clearly ordered. For example, *survey response* taking values *always, usually, sometimes, never*. In this case one could assign numerical values to the responses.

Survey response	Number
Always	4
Usually	3
Sometimes	2
Never	1

Reclassifying categorical variables

- Often Categorical variable will contain too many easily analyzable field values. For example the predictor ‘state’ could contain 50 different field values. In such a case the analyst should reclassify the field values. i.e., 50 states could each be reclassified as a variable region containing field values *Northeast, Southeast, North Central, Southwest* and *West.* Thus instead of 50 different field values, the analyst will face with only 5.

Getting to know the dataset

- With unknown large databases, the analyst often prefer to use Exploratory Data Analysis (EDA) or Graphical Data Analysis
- EDA allows the analyst to

➢ Examine the inter relationships among the attributes

➢ Identifying interesting subsets of the observations

➢ Develop an initial idea of possible associations amongst the predictors as well as between predictors and target variables.

Dimension Reduction Methods

- The use of too many predictor variables to model a relationship with a response variable can unnecessarily complicate the interpretation of the analysis. Also retaining too many variables may lead to overfitting.
- Dimension reduction methods have the role of using the correlation structure among the predictor variables to accomplish the following

➢ To reduce the number of predictor items

➢ To help ensure that these predictor items are independent

➢ To provide a framework for interpretability of the results.

Principal Components Analysis (PCA)

- PCA seeks to explain the correlation structure of a set of predictor variables, using a smaller set of linear combinations of these variables. These linear combinations are called *components.* The total variability of a dataset produced by the complete set of *m* variables can often be mostly accounted for by a smaller set of *k* linear combinations of these variables, which would mean that there is almost as much information in the *k* components as there is in the original m variables.

Factor Analysis (FA)

- FA is related to Principal Components but the two methods have different goals. Principal Components seek to identify orthogonal linear combinations of the variables, to be used either for descriptive purposes or to substitute a smaller number of uncorrelated components for the original variables. In contrast , FA represents a model for the data.

Presentation of Data

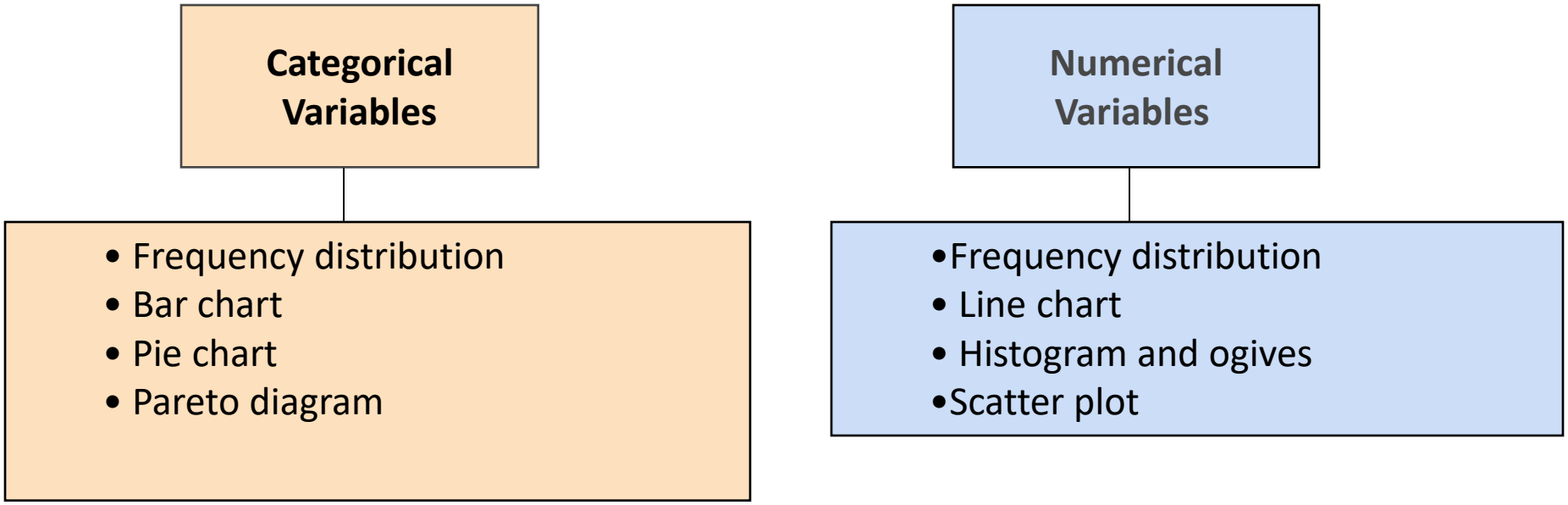
Session 10

11/12/2020

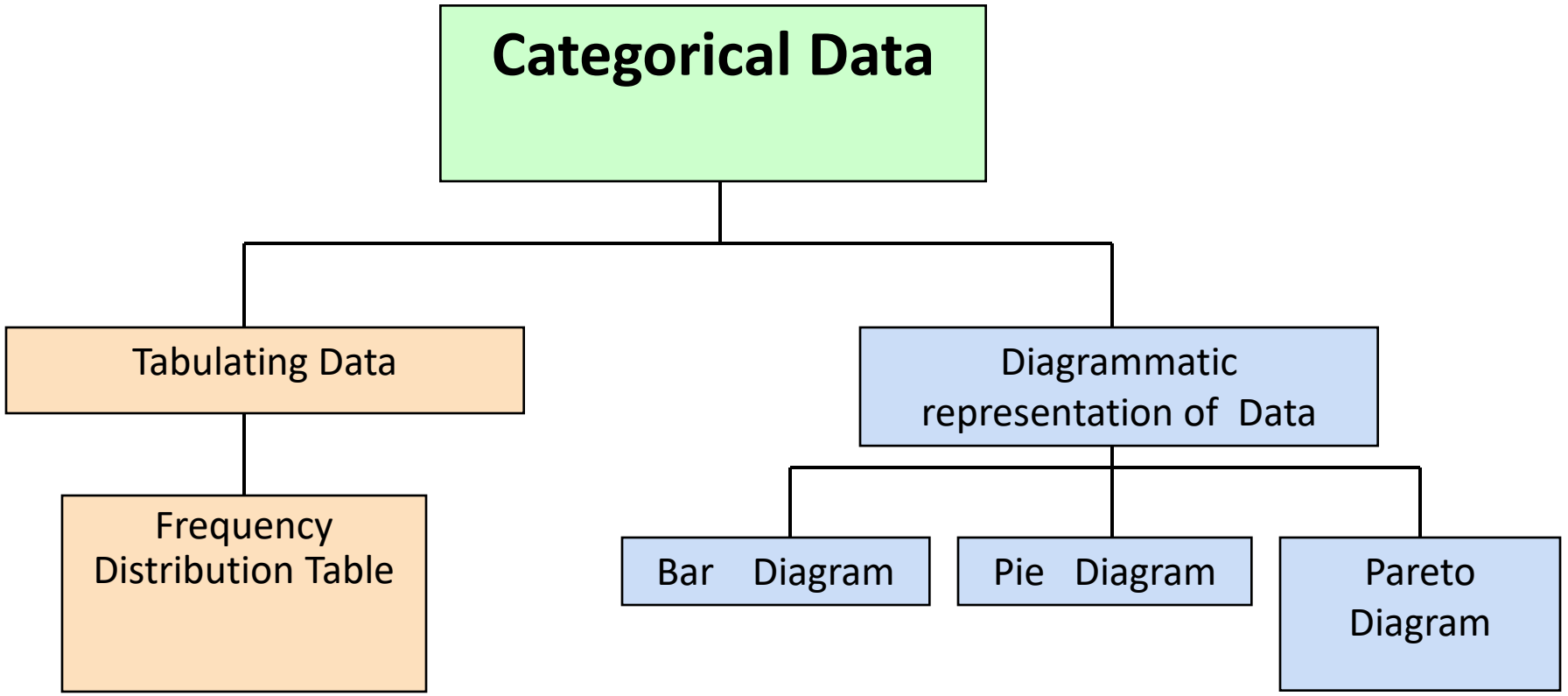
Presentation of Data

- Data in raw form are usually not easy to use for decision making
- Some type of organization is needed
  - Table
  - Diagram and Graph
- Depends on the variable being summarized

Presentation of Data



Presentation for Categorical Variables



Describing Data Sets

- Qualitative Data are nonnumerical
  - Major Discipline
  - Political Party
  - Gender
  - Eye color
- Summarized in two ways:
  - Class Frequency
  - Class Relative Frequency

- Class
  - A class is one of the categories into which qualitative data can be classified

- Class Frequency
  - Class frequency is the number of observations in the data set that fall into a particular class

- Frequency Distribution
  - It is a summary technique that organizes data into classes and provides in tabular form a list of the classes along with the number of observations in each class.

Example: Adult Aphasia

Table: Data on 22 Adult Aphasias

Subject	Type of Aphasia	Subject	Type of Aphasia
1	Broca's	12	Broca's
2	Anomic	13	Anomic
3	Anomic	14	Broca's
4	Conduction	15	Anomic
5	Broca's	16	Anomic
6	Conduction	17	Anomic
7	Conduction	18	Conduction
8	Anomic	19	Broca's
9	Conduction	20	Anomic
10	Anomic	21	Conduction
11	Conduction	22	Anomic

Table: Frequency Distribution of Data on 22 Adult Aphasias

Type of Aphasia	Frequency
Anomic	10
Broca's	5
Conduction	7
Total	22

- Class Relative Frequency
  - Class frequency divided by the total number of observations in the data set

$$\text{class relative frequency} = \frac{\text{class frequency}}{n}$$

- Class Percentage
  - Class relative frequency multiplied by 100

$$\text{class percentage} = (\text{class relative frequency}) \times 100$$

Table: Frequency, relative frequency, and class percentage on 22 Adult Aphasias

Type of Aphasia	Frequency	Relative Frequency	Class Percentage
Anomic	10	10/22 = .455	45.5%
Broca's	5	5/22 = .227	22.7%
Conduction	7	7/22 = .318	31.8%
Total	22	22/22 = 1.00	100%

Diagrammatic representation of Qualitative variables

1. Bar Chart
2. Pie chart
3. Pareto diagram

One dimensional diagrams; Bar diagrams

- Bar diagrams are one of the easiest and the most commonly used devices of presenting most of the business and economic data. These are one dimensional diagrams because in such diagrams only one dimension, say height or length of the bar is taken into account to present the given values.

➢ Types of bar diagrams.

a) Simple bar diagram

b) Subdivided or component bar diagram

c) Percentage bar diagram

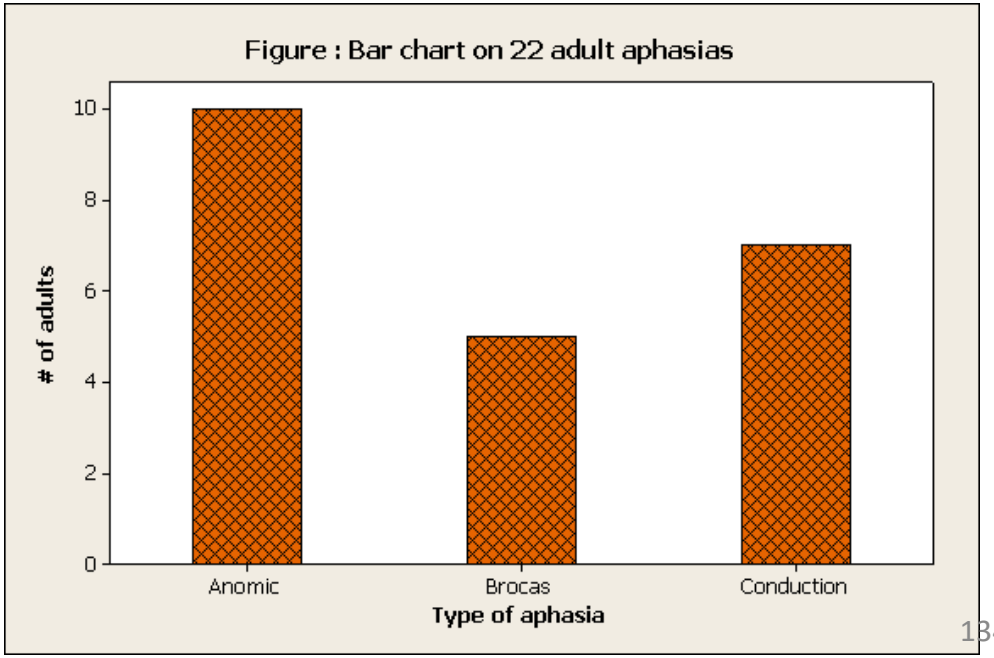
d) Multiple bar diagram.

a) Simple bar diagram

- It is the simplest of the bar diagrams and is used for comparative study of values of a single variable or single classification of data.
- Here the magnitudes of the observations are represented by the height of the bars.
- Remark: If there are large number of items or values of the variable under study then instead of bar diagram, line diagram may be drawn.

- Bar Chart Example: Data on 22 Adult Aphasias

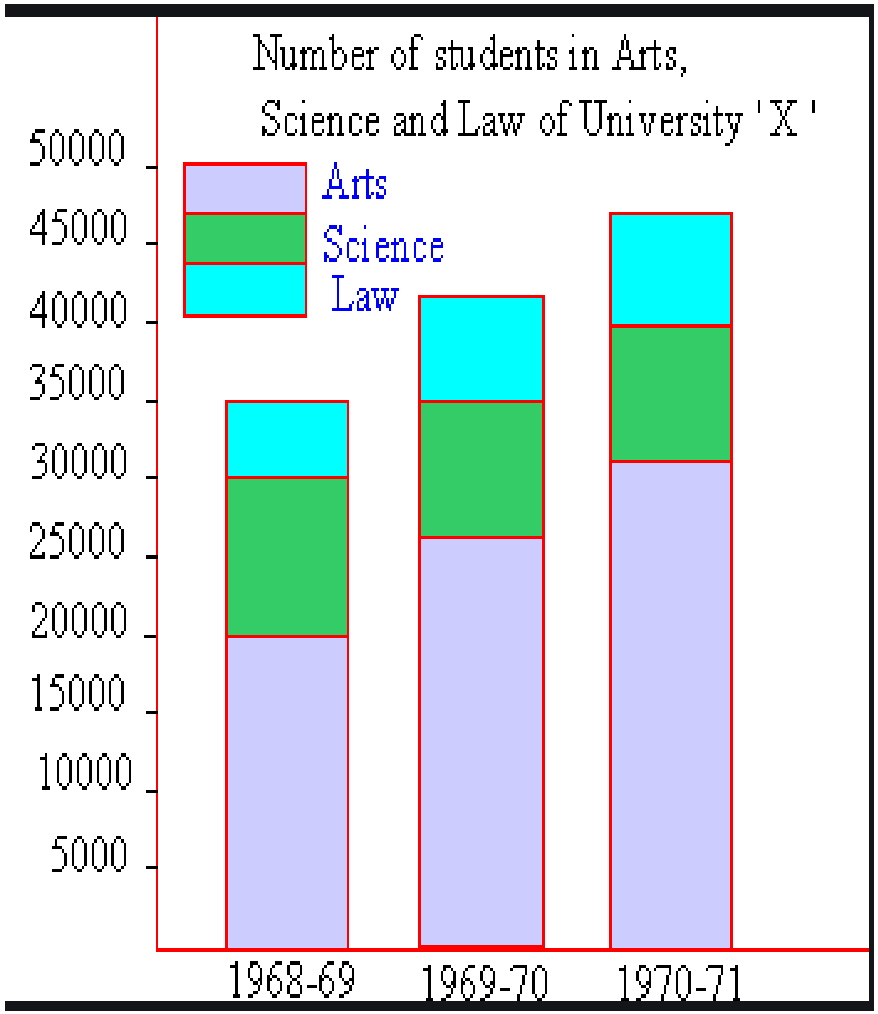
  - A diagramatic representation of information in the form of bars.
  - Bars of equal width are drawn to represent different categories, with the length of each bar being proportional to the number or frequency of occurrence of each category.



b)Subdivided or component bar diagram

- A limitation of the simple bar diagram is that it study only one item of a characteristic or classification at a time.
- Subdivided bar diagram is useful not only for presenting several items of a variable diagramatically but also enable us to make a comparative study of different parts or components among themselves and also to study the relationship between each component and the whole.
- In general subdivided bar diagrams are used if total magnitude of the given variable has to be divided into various parts or components. First bar representing total is drawn . Then it is divided into various segments, each segment representing a given component of the total . Different shades, colours or designs are used to distinguish various components .
- Remark : subdivided bar diagram is not suggested if the number of components exceed 10 .

Subdivided bar diagram



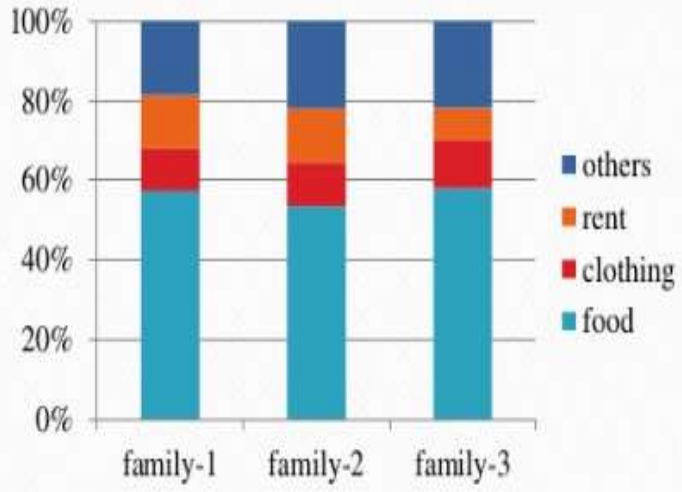
c)Percentage bar diagram

- Subdivided or component bar diagram presented on a percentage basis gives percentage bar diagram.
- These are useful for diagrammatic portrayal of the relative change in the data and hence highlight the relative importance of the various component parts to the whole.
- The total for each bar is taken as 100 and the value of each component is expressed as percentage of the respective totals. It is quite convenient and useful for comparing two or more sets of data .

Percentage bar diagram

Eg: percentage expenditure on various items of the 3 families

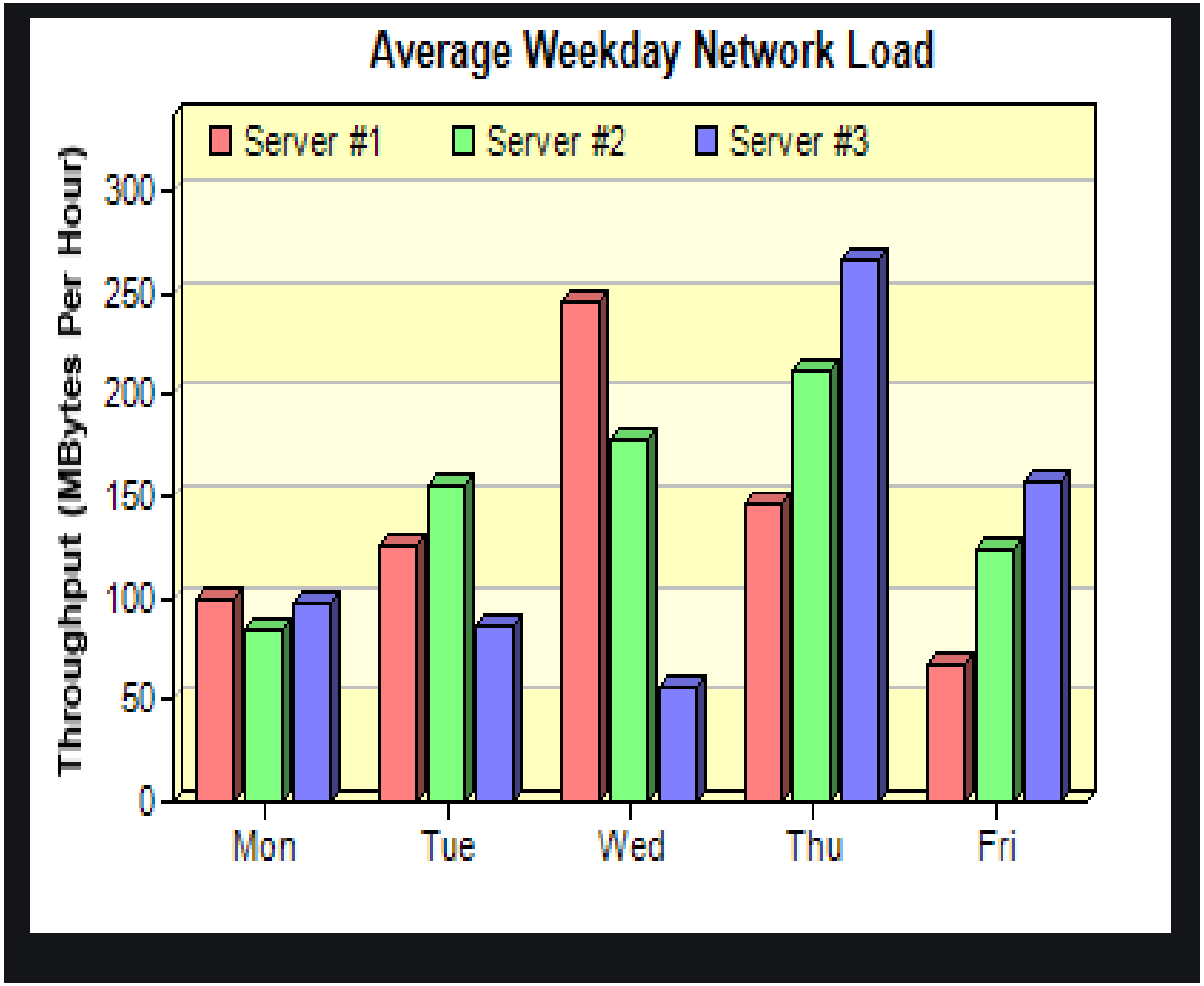
Components	Family-1	Family-2	Family-3
Food	57.3	53.5	58.0
Clothing	10.7	11.0	12.0
Rent	13.3	13.5	8.2
Others	18.7	22.0	21.8



d)Multiple bar diagram

- As discussed earlier one limitation of simple bar diagram is the representation of a single characteristic. If two or more sets of interrelated variables are to be presented, multiple bar diagrams are used.
- The technique of drawing multiple bar diagrams are same as that of simple bar diagram except that here a set of adjacent bars (one for each variable) is drawn.
- Proper and equal spacing is given between different sets of the bars. To distinguish different bars, a set different shades may be used.

Multiple bar diagram



Presentation of Data(contd…)

Session 11

14/12/2020

Two-dimensional diagrams

Introduction

- We know that in one dimensional diagrams magnitude of the observations are represented by only one of the dimensions say height or length of the bar, while the width of the bar is arbitrary and uniform. However in two dimensional diagrams magnitude of the given observations are represented by the area of the diagram.
- Thus in two dimensional diagrams the length as well as width of the bars will have to be considered.
- Two dimensional diagrams are also known as area diagram or surface diagram.
- Some of the commonly used two dimensional diagrams are rectangles and pie diagrams .

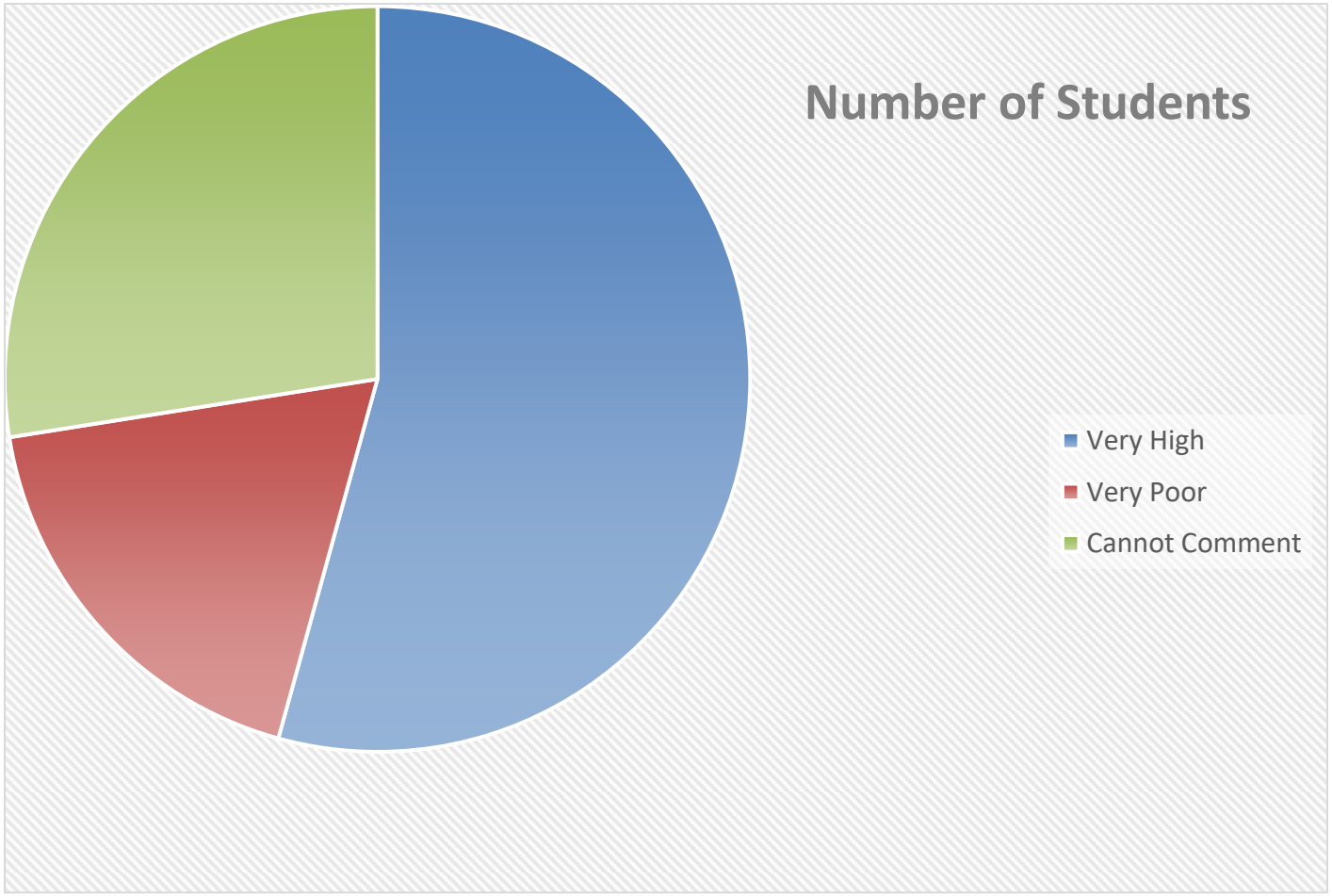
Pie diagrams

- Just as subdivided or percentage bar or rectangles are used to represent the total magnitude and its various components, the circle may be divided into various sections or segments say sectors representing certain proportion or percentage of various component parts.
- Sub divided circle is known as pie diagram or angular or circular diagram.

Steps of construction of pie diagram

- 1.Express each of the component values as a percentage of the respective total.
- 2. The percentage of component parts obtained in step 1 can be converted to degrees by multiplying each of them by 3.6.
- 3. Draw a circle of appropriate radius and represent the components sector wise.
- 4. Different sectors to be distinguished with different shades.

Pie chart representing students' opinion



Pareto Diagram

- Used to portray categorical data
- A bar chart, where categories are shown in descending order of frequency.
- One of the important aspects of total quality management is the constant search for causes of problems in products and processes.
- A graphical technique for displaying problem causes is Pareto Analysis. Pareto analysis is a quantitative tallying of the number and types of defects that occur with a product or service. Analysts use this tally to produce a vertical bar chart that display the most common types of defects ,ranked in order of occurrence from left to right. This bar chart is called a Pareto Chart.

Pareto Diagram Example

Example: 400 defective items are examined for cause of defect:

Source of Manufacturing Error	Number of defects
Bad Weld	34
Poor Alignment	223
Missing Part	25
Paint Flaw	78
Electrical Short	19
Cracked case	21
Total	**400**

Pareto Diagram Example

(continued)

Step 1: Sort by defect cause, in descending order

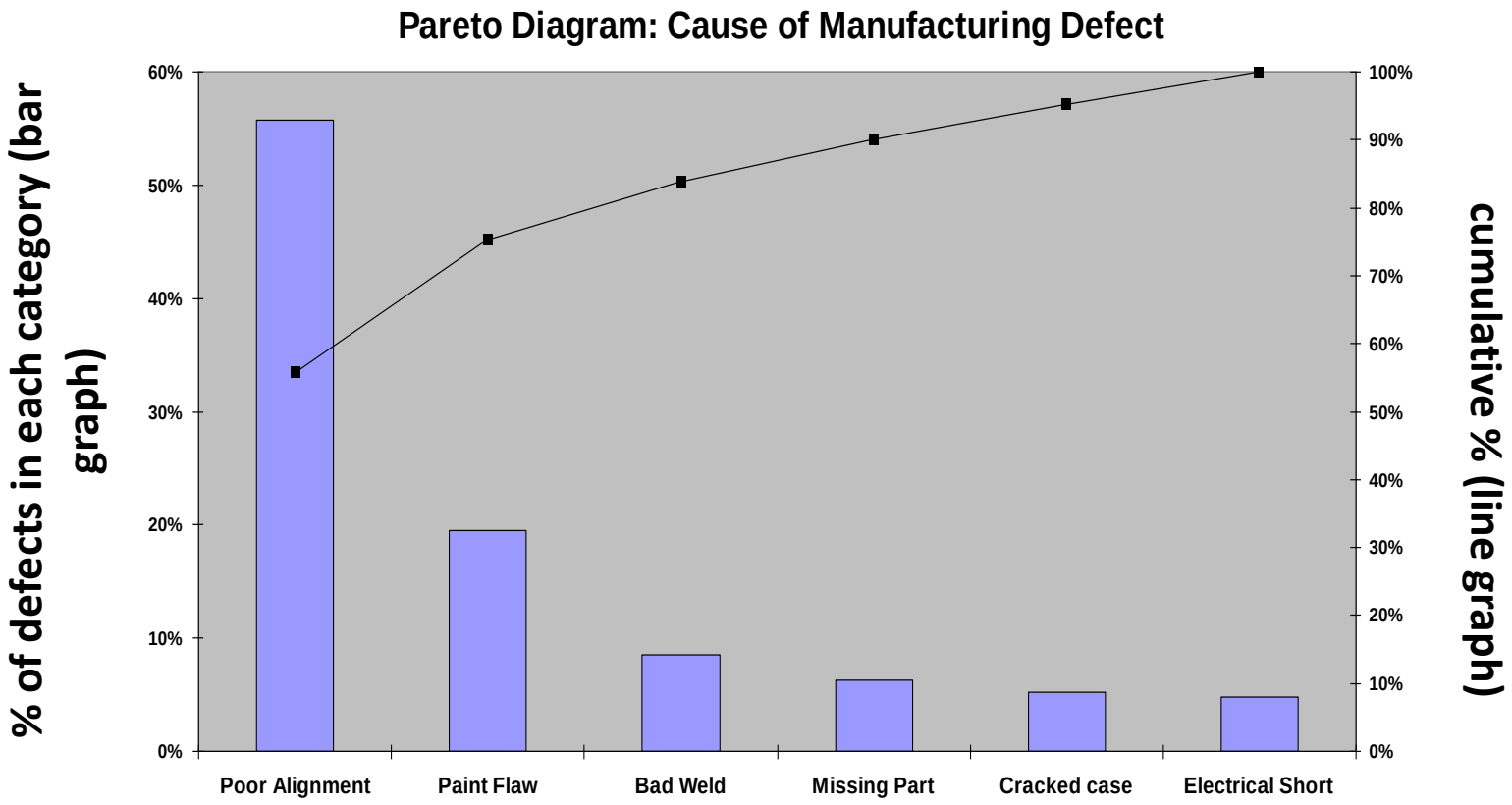
Step 2: Determine % in each category

Source of Manufacturing Error	Number of defects	% of Total Defects
Poor Alignment	223	55.75
Paint Flaw	78	19.50
Bad Weld	34	8.50
Missing Part	25	6.25
Cracked case	21	5.25
Electrical Short	19	4.75
Total	**400**	**100%**

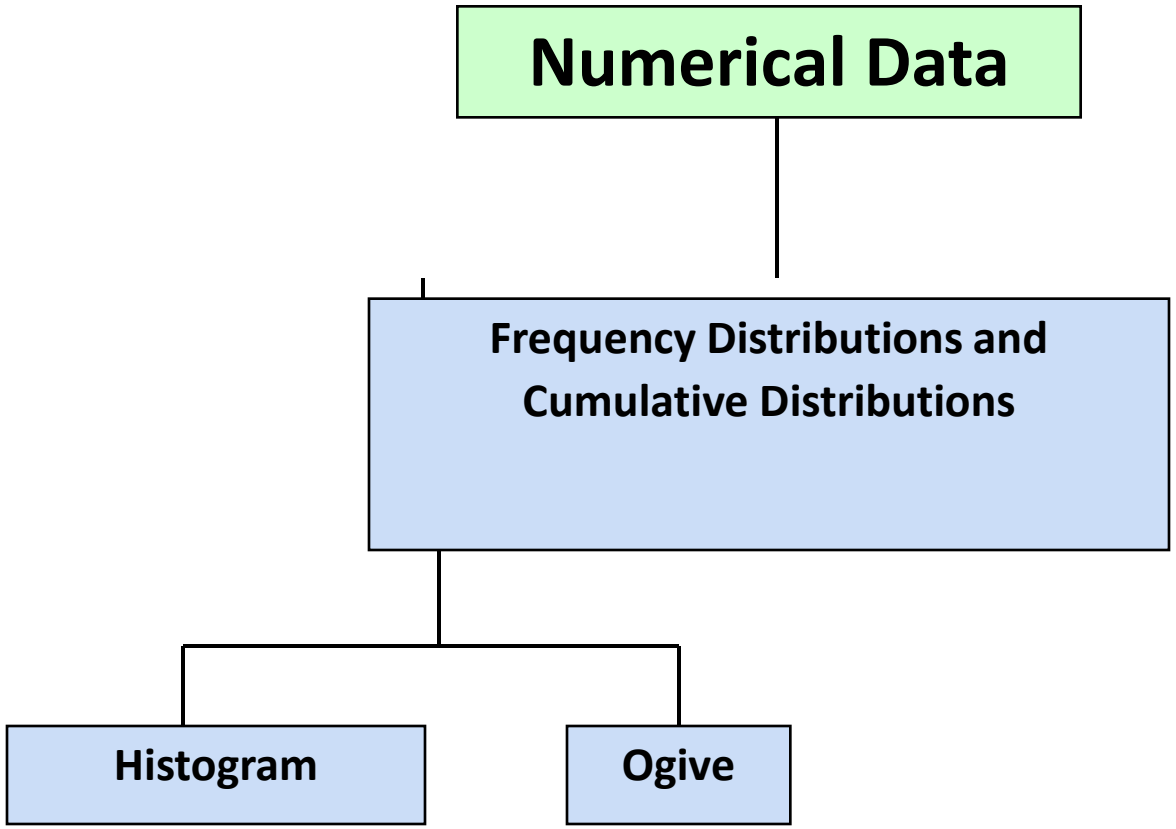
Pareto Diagram Example

(continued)

Step 3: Show results graphically



Graphs to Describe Numerical Variables



Frequency Distribution

- One particularly useful tool for grouping data is the frequency distribution, which is a summary of data presented in the form of class intervals and frequencies.
- The way of tabulating the data of a variable and their respective frequencies side by side is called a frequency distribution of that data.
- We discuss the following cases

➢ Discrete (ungrouped) frequency distribution

➢ Grouped frequency distribution

➢ Continuous frequency distribution

Discrete (ungrouped) frequency distribution

- Here first we arrange the data in ascending/descending order.
- In the first column, we place the possible values of the variable
- Then we count the number of times each value of the variable occurs. In the second column a vertical bar(called Tally mark) is put against the number whenever it occurs.
- After putting the tally mark for all the values in the data, we count the number of times each value is repeating and write it against the corresponding number in the third column.

Discrete (ungrouped) frequency distribution

- Marks of 10 students are given below. Construct a discrete frequency distribution.

70	45	33	64	45
25	64	45	30	20

- Arrange the data : 20,25,30,33,45,45,45,64,64,70.

Marks	Tally	Frequency
20	|	1
25	|	1
30	|	1
33	|	1
45	|||	3
64	||	2
70	|	1
Total	10	10

Grouped frequency distribution

- Here classify the data into different classes(or class intervals) by dividing the entire range of the values of the variable into a suitable number of groups which is called classes.
- Then we record(count) the number of observations in each group or class.
- The grouped frequency distribution of earlier example of marks are done below.

Marks	No.of students
20-35	4
36-50	3
51-65	2
66-80	1
Total	10

Continuous frequency distribution

- While dealing with a continuous variable, it is not desirable to present the data in to a form which we have seen earlier.
- For example, if we consider the ages of a group of students in a college, then grouped frequency distribution into the classes 18-20, 21-23, 24-26 etc. will not be correct. Why?
- In such situation, we form continuous class intervals like Below 20, 20 or more but less that 23 etc.
- The presentation of data into continuous classes of above type along with corresponding frequencies is known as continuous frequency distribution.

Unit-2
Basic Statistics

Frequency Distribution

Session 12

16/12/2020

Why Use Frequency Distributions?

- A frequency distribution is a way to summarize data
- The distribution condenses the raw data into a more useful form.
- Allows for a quick visual interpretation of the data

Basic principles for forming a grouped frequency distribution

- The following are the general guidelines
- 1.Types of classes: The classes should be clearly defined and should be non-overlapping.
- 2. Number of classes : Usually number of classes should not be greater than 15 and should not be less than 5 based on the total frequency, nature of the data etc.
- There is one rule exists which is known as Sturge's rule which is as follows.

Sturge's Rule

- A rule for determining number of classes to use in a histogram or frequency distribution table.
- Sturge's Rule: $k = 1 + 3.322(\log_{10} n)$,
  k is the number of classes,
  n is the size of the data.
- $k = 1 + 3.322(\log_{10} 21) = 5.4$

Basic principles for forming a grouped frequency distribution

- 3. Size of class intervals :

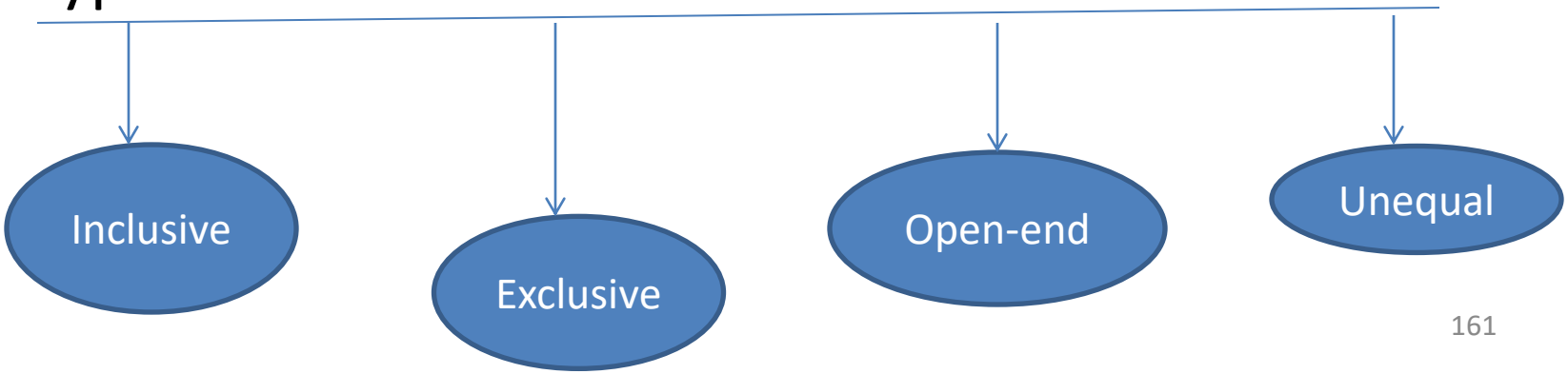
$$h = \frac{Range}{Number\ of\ classes}$$

where

Range = X_{max}-X_{min}

= Largest observation-Smallest observation.

- 4. Types of class intervals



4. Types of class intervals

- (a) Inclusive type classes : The classes of the type 30-39,40-49,50-59 etc. in which both the upper and lower limits are included in the class are called Inclusive classes.
- (b) Exclusive type classes : The classes of the type 15-20, 20-25,25-30 etc. in which the upper limit of the classes are excluded from the respective classes are called Exclusive classes.
- (c)Open-end classes: The classification is termed as ‘open end classification’ if the lower limit of the first class or the upper limit of the last class are not specified . For example classes like marks less than 20, age above 60 are open-end classes.

4. Types of class intervals (contd…)

- (d) Unequal Class interval : When the width of the classes are NOT equal in a particular frequency distribution, such type of classes are unequal classes.
- Examples are 0-10, 10-30, 30-80,80-100 etc.

Converting inclusive class to exclusive class(continuous class)

- The upper and lower class limits of new exclusive type classes are called class boundaries.
- If d is the gap between the upper limit of any class and lower limit of succeeding class, the class boundaries for any classes are given by
- Upper class boundary = Upper class limit +[1(d)/2]
- Lower class boundary =Lower class limit –[1(d)/2]

Marks (Inclusive)	Class boundary (Exclusive)
20-24	19.5-24.5
25-29	24.5-29.5
30-34	29.5-34.5

Mid-value (class mark)

- As the name suggests, the mid-value or class mark is the value of the variable which is exactly at the middle of the class.
- The mid-value of any class is obtained on dividing the sum of upper and lower class limits (class boundaries) by 2.
- Mid-value of a class = ½[Lower class limit+ Upper class limit] = ½[Lower class boundary +Upper class boundary].

Marks	Class boundary	Mid value
20-24	19.5-24.5	22
25-29	24.5-29.5	27
30-34	29.5-34.5	32

Frequency Distribution Example

Example: An analyst randomly selects 20 winter days and records the daily high temperature

24, 35, 17, 21, 24, 37, 26, 46, 58, 30,
32, 13, 12, 38, 41, 43, 44, 27, 53, 27

Frequency Distribution Example

- Sort raw data in ascending order:
  12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58
- Find range: **58 - 12 = 46**
- Select number of classes:

  k =1+3.322log10(20)=5(apprx)
- Compute interval width: **9.2 (round up) ~ 10**
- Determine interval boundaries: **10 but less than 20, 20 but less than 30, . . . , 50 but less than 60**
- Count observations & assign to classes

Frequency Distribution Example

Data in ordered array:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Class interval	frequency	Relative frequency	Percentage
10-20	3	0.15	15
20-30	6	0.3	30
30-40	5	0.25	25
40-50	4	0.2	20
50-60	2	0.1	10
Total	20	1	100

Cumulative frequency distribution and graphical presentation of data

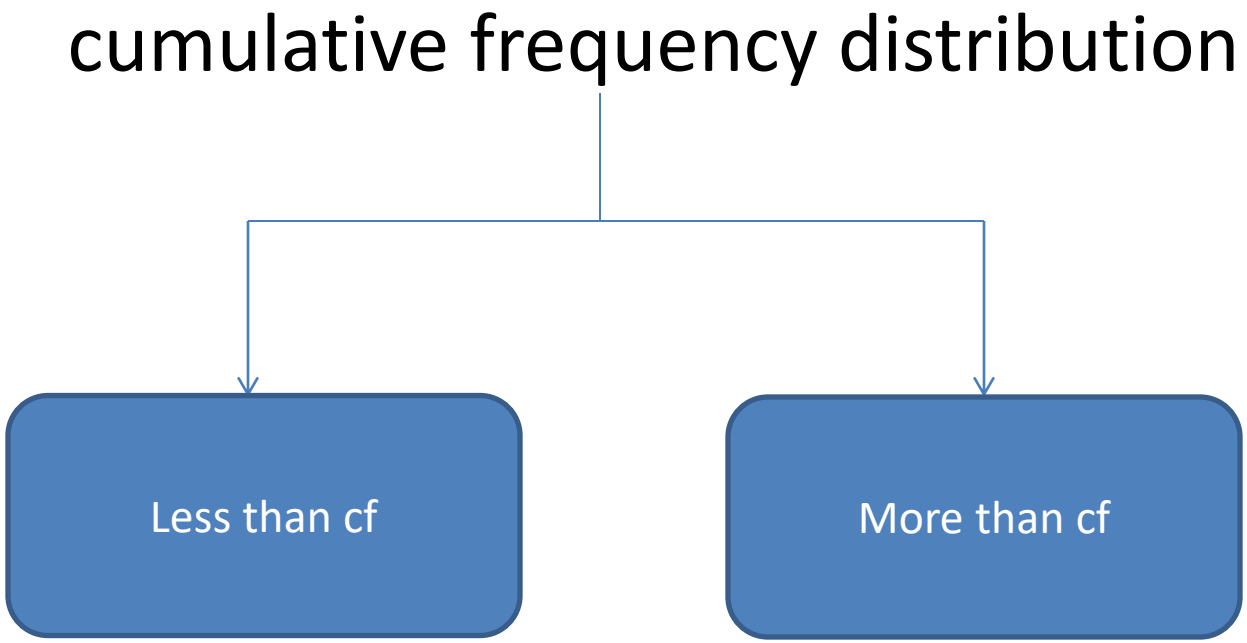
Session 13

18/12/2020

Cumulative frequency distribution

- A frequency distribution simply tells us how frequently a particular value of the variable is occurring.
- However if we want to know the total number of observations getting a value ‘less than’ or ‘more than’ a particular value of the variable, the normal frequency table fails.
- Cumulative frequency distribution is a modification of the given frequency distribution, and is obtained on successively adding the frequencies of the values of the variable according to a certain law.
- The frequencies so obtained are called ‘cumulative frequencies’ (c.f).

Types of cumulative frequency distribution



Less than c.f.

- Less than c.f. for any value of the variable (class) is obtained on adding successively the frequencies of all the previous values (classes), including the frequency of the variable against which the totals are written, provided the values(classes) are arranged in ascending order of magnitude.
- In a less than cumulative frequency distribution, the cumulative frequencies are in ascending order.

More than c.f.

- The more than c.f. is obtained similarly by finding the cumulative totals of frequencies starting from the highest value of the variable (class) to the lowest value (class).
- In more than cumulative frequency distribution, cumulative frequencies are in descending order.

- Remarks
- 1: In fact ‘less than’ and ‘more than’ words also include ‘equality sign’.
- 2: Cumulative frequency is of particular importance in computation of median, quartiles etc.

The Cumulative Frequency Distribution

Example: 20 winter days daily high temperature

Data in ordered array:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58.

Class interval	frequency	Percentage	Less than c.f	Greater than c.f
10-20	3	15	3	20
20-30	6	30	9	17
30-40	5	25	14	11
40-50	4	20	18	6
50-60	2	10	20	2
Total	20	100		

Graphic representation of data

- The most commonly used graphs for charting a frequency distribution are
- Histogram
- Frequency polygon
- Frequency curve
- Ogive/Cumulative frequency curves

Histogram

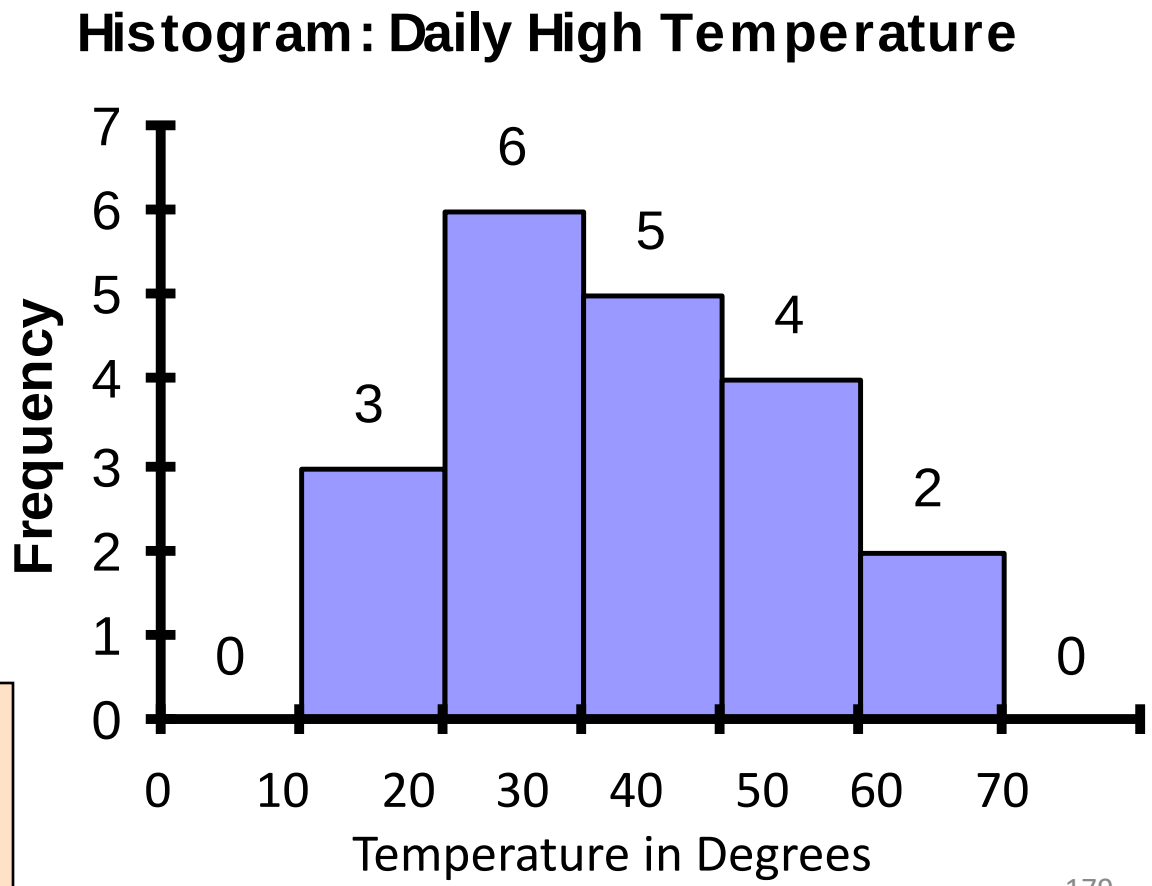
- It is one of the most commonly used devices for continuous frequency distribution. Usually the variate takes values (class) are taken along the X axis and the frequencies along the Y axis.
- It is possible to construct histogram with equal classes intervals and unequal classes.
- If the classes are of equal interval, erect a rectangle with height proportional to the corresponding frequency of the class. The area under the histogram is the total frequency of the distribution as distributed through out different classes.

Histogram (contd…)

- If the classes are of unequal width, erect a rectangle with height proportional to the frequency density where
- Frequency density = $\frac{\text{frequency of the class}}{\text{Class width}}$
- Remark:
- The **interval endpoints** are shown on the **horizontal axis**
- The vertical axis is either **frequency, relative frequency, or percentage**

Histogram Example

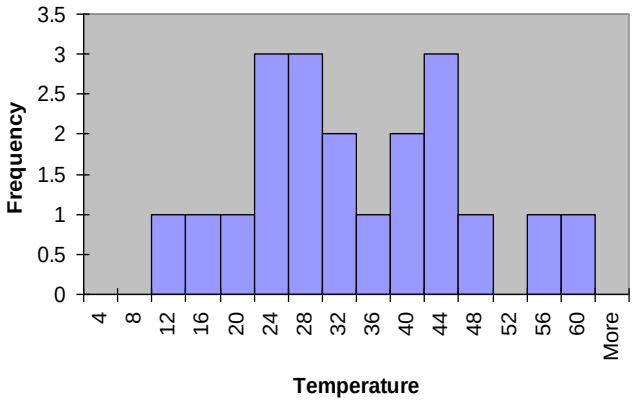
Interval	Frequency
10 but less than 20	3
20 but less than 30	6
30 but less than 40	5
40 but less than 50	4
50 but less than 60	2



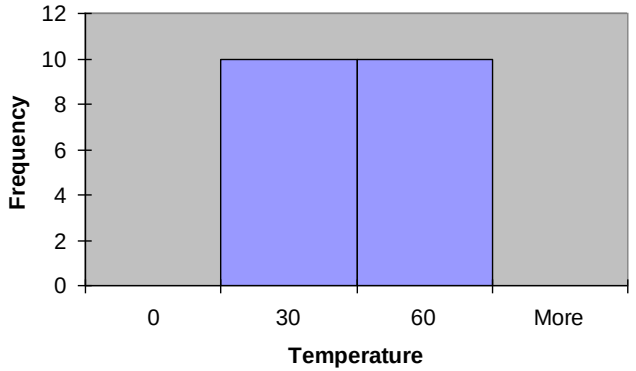
(No gaps between bars)

How Many Class Intervals?

- **Many (Narrow class intervals)**
  - may yield a very rough with sharp distribution
  - with gaps from empty classes
  - Can give a poor indication of how frequency varies across classes



- **Few (Wide class intervals)**
  - may compress variation too much and yield a blocky distribution
  - can obscure important patterns of variation.



(X axis labels are upper class endpoints)

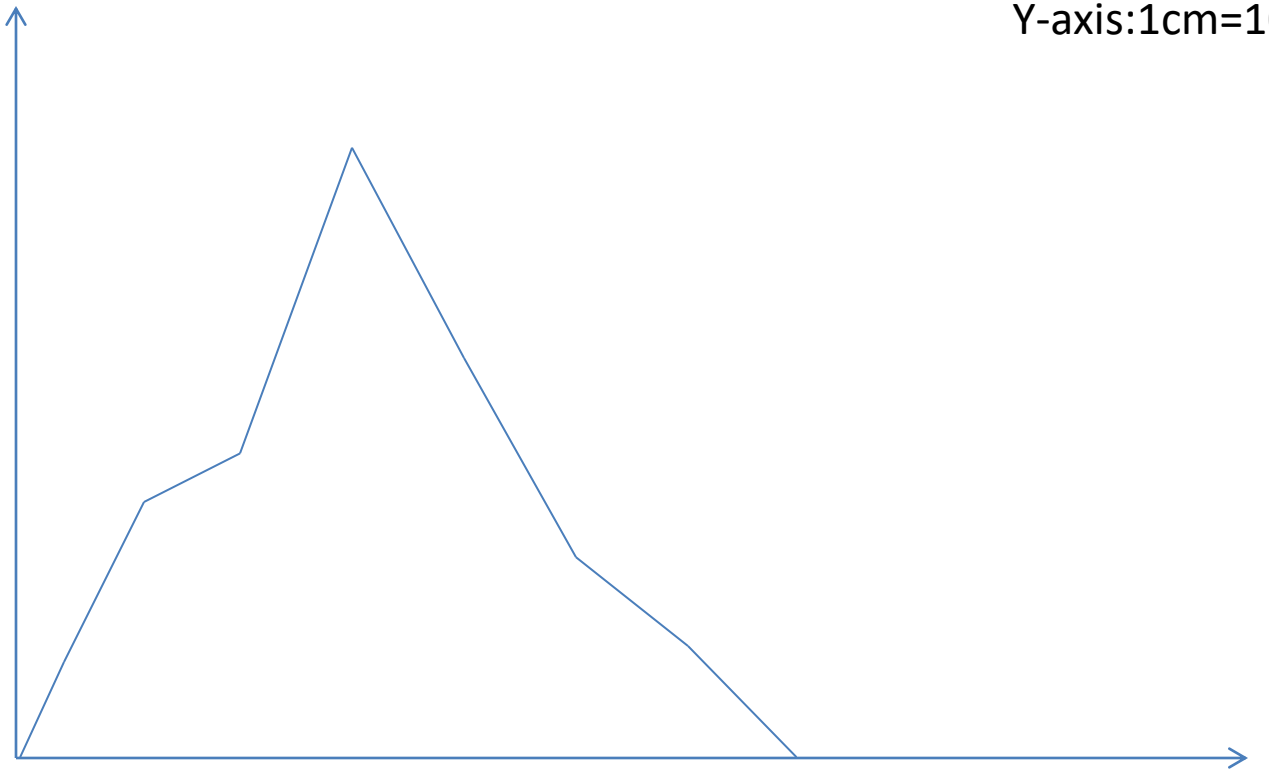
Frequency Polygon

- It is another device of graphic presentation of a frequency distribution (continuous, grouped or discrete) . In case of discrete frequency distribution, frequency polygon is obtained on plotting the frequencies on the vertical axis (Y-axis) against the corresponding values of the variable on the horizontal axis (X-axis) and joining the points so obtained by straight lines.
- In case of grouped/continuous frequency distribution, frequency polygon may be drawn two ways.

Frequency polygon without constructing Histogram

X-axis: 1 cm =10 units,

Y-axis:1cm=10 units



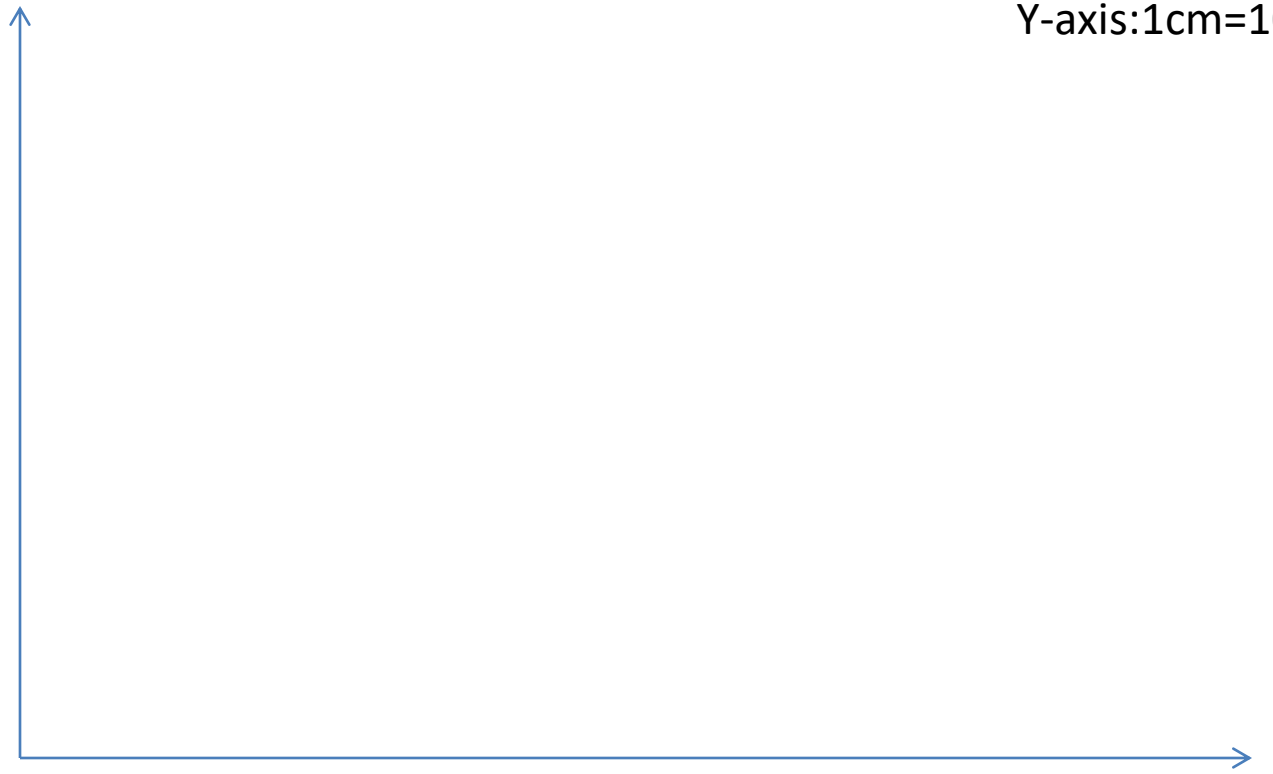
Frequency Curve

- A frequency curve is a smooth free hand curve drawn through the vertices of a frequency polygon. The area enclosed by the frequency curve is same as that of histogram or polygon but its shape is smooth one without sharp edges.

Frequency curve

X-axis: 1 cm =10 units,

Y-axis:1cm=10 units



Comparison of Histogram and frequency polygon

Histogram	Polygon
➢ A two dimensional figure with a collection of adjacent rectangles	➢ A line graph
➢ For comparative studies, we need to draw different histograms	➢ Two or more frequency polygon can be drawn on the same graph for comparative studies.
➢ Suitable adjustment should be done for drawing histogram with unequal class intervals	➢ A continuous curve and hence may be used to determine rate of change, slope etc.

Graphical presentation of Data(contd..)

Session 14

21/12/2020

Ogive(Cumulative frequency curves)

- This is a graphical presentation of cumulative frequency distribution of a continuous variable. It consists in plotting the cumulative frequencies(along the y axis) against the class boundaries (along x axis) .
- Since there are two types of c .f. distribution, we have accordingly two types of ogives.
- (i) Less than ogive
- (ii) More than(greater than) ogive

Less than ogive

- This consist in plotting the less than cumulative frequencies against the upper class boundaries of the respective classes. The points so obtained are joined by a smooth free hand curve to give less than ogive.
- It is an increasing curve slopping upwards from left to right .
- Less than ogive should start on the left with cumulative frequency zero at the lower boundary of the first class .

More than ogive

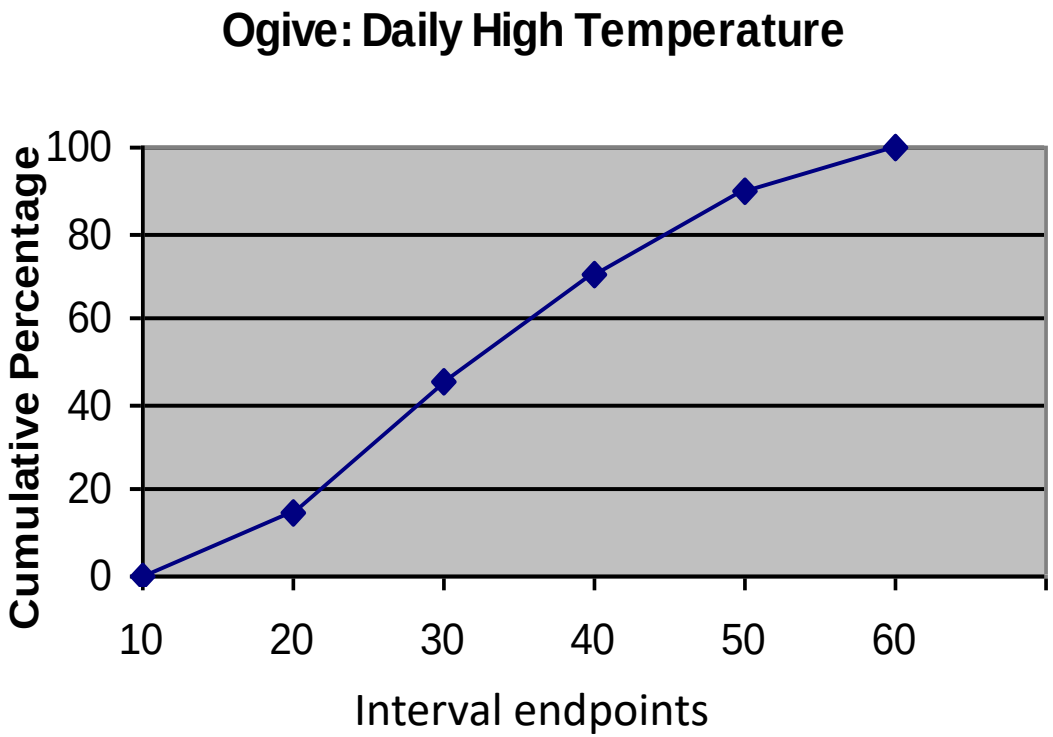
- In more than ogive the more than cumulative frequencies are plotted against the lower class boundaries of the respective classes. The points so obtained are joined by a smooth free hand curve to give more than ogive.
- It is a decreasing curve and slopes downward from left to right.

Note:

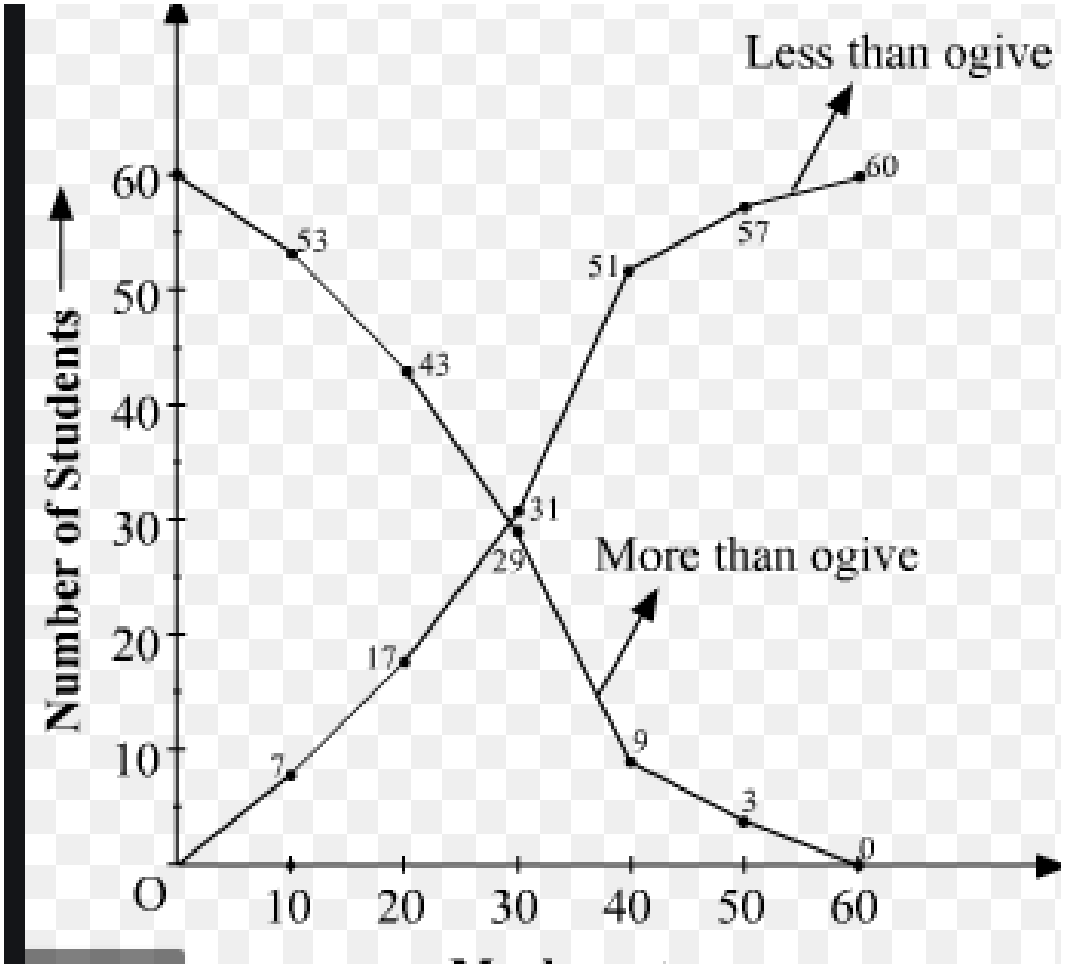
- 1. it is possible to draw less than and more than ogive in a single graph.
- 2. These are useful for graphic computation of partition values say median, quartiles, deciles, percentiles etc .

The Ogive
Graphing Cumulative Frequencies

Interval	Upper boundary point	Cumulative Percentage
Less than 10	10	0
10 but less than 20	20	15
20 but less than 30	30	45
30 but less than 40	40	70
40 but less than 50	50	90
50 but less than 60	60	100

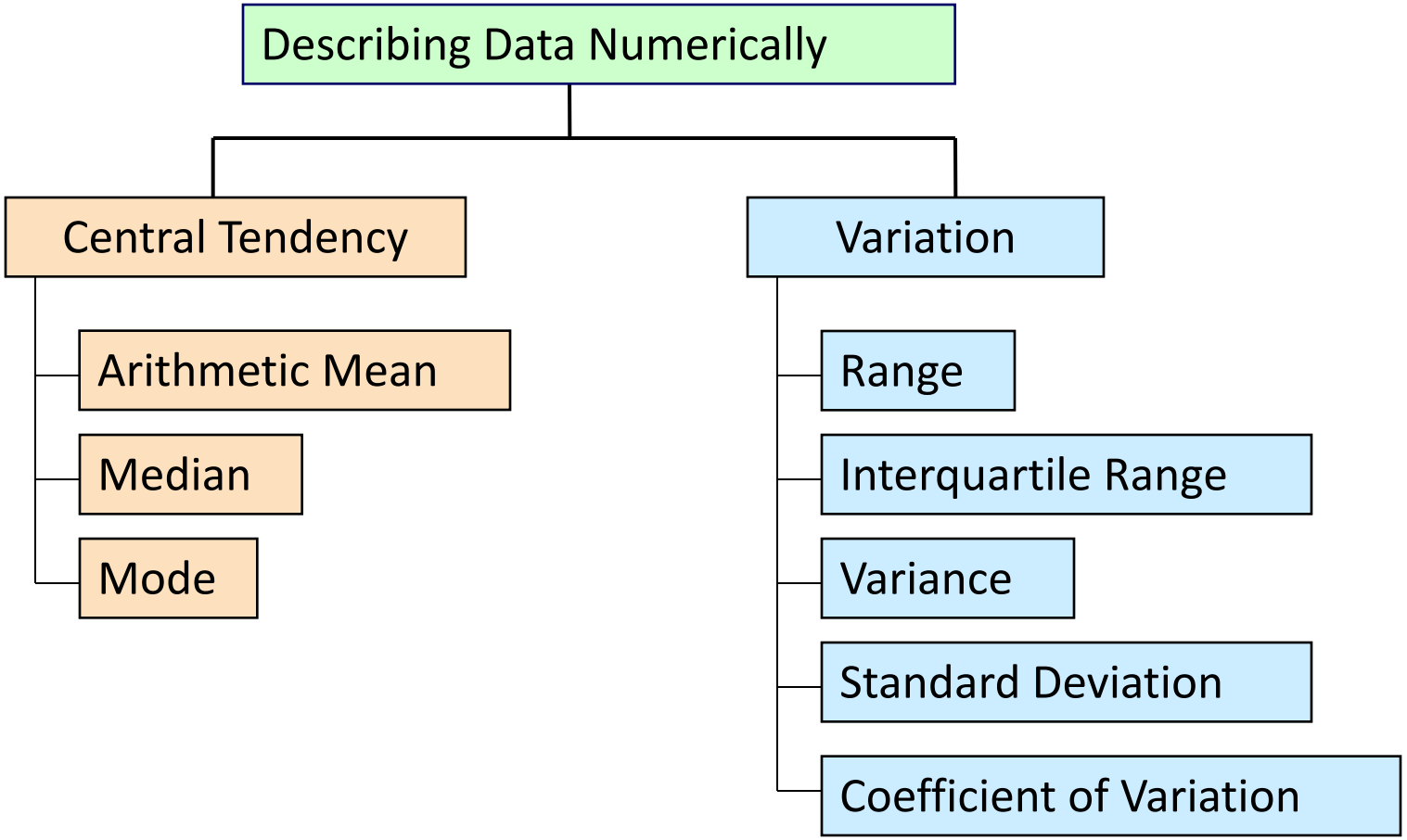


less than and more than ogive in a single graph.



MEASURES OF CENTRAL TENDENCY

Describing Data Numerically



Summary Definitions

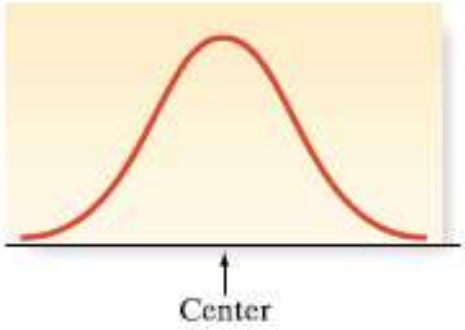
- The **central tendency** is the extent to which all the data values group around a typical or central value.

- The **variation** is the amount of dispersion, or scattering, of values

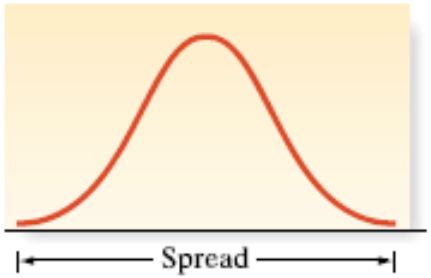
Summarizing Data Sets

Numerical Measures of Central Tendency

- **Central tendency** is the value or values around which the data tend to cluster



- **Variability** shows how strongly the data cluster around that (those) value(s)



MEASURES OF CENTRAL TENDENCY (UNGROUPED DATA)

Measures of Central Tendency: The Mean

- The arithmetic mean (often just called "mean") is the most common measure of central tendency

Pronounced x-bar

The i^{th} value

– For a sample of size n:

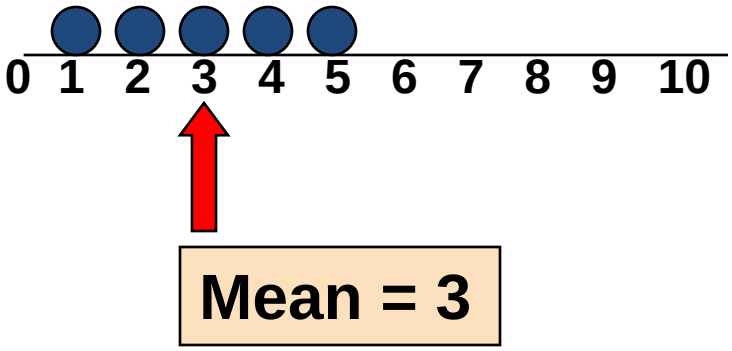
$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Sample size

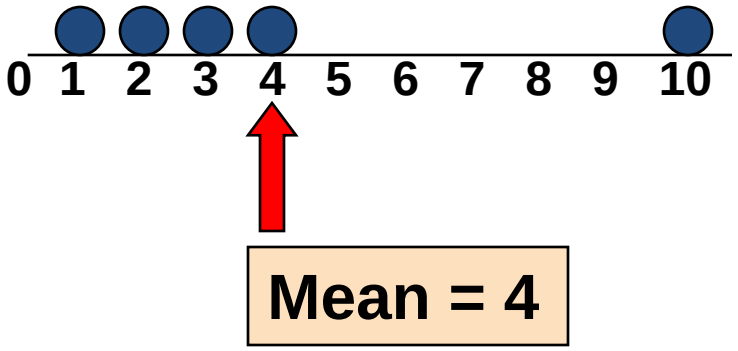
Observed values

Measures of Central Tendency: The Mean

- The most common measure of central tendency
- Mean = sum of values divided by the number of values
- Affected by extreme values (outliers)



$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$



$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

Example: During a two week period 10 houses were sold(in $) in Fancytown.

House Price in Fancytown x
231,000
313,000
299,000
312,000
285,000
317,000
294,000
297,000
315,000
287,000
$\sum x =$ 2,950,000

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{n} = \frac{2{,}950{,}000}{10} = 295{,}000$$

The "average" or mean price for this sample of 10 houses in Fancytown is $295,000

Example: During a two week period 10 houses were sold in (in $)Lowtown.

House Price in Lowtown x
97,000
93,000
110,000
121,000
113,000
95,000
100,000
122,000
99,000
2,000,000
$\sum x =$ 2,950,000

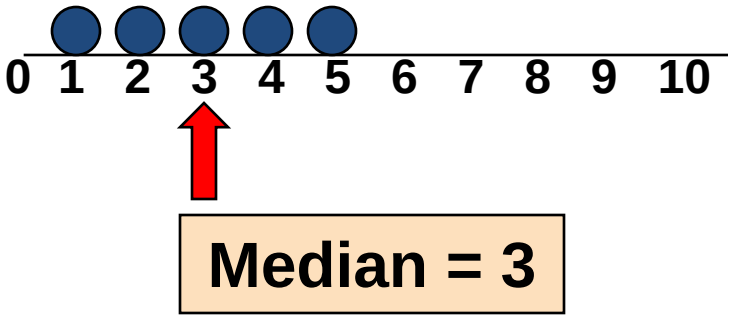
$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{n} = \frac{2{,}950{,}000}{10} = 295{,}000$$

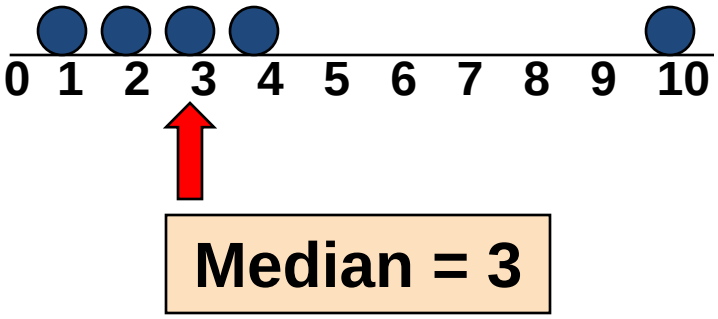
The "average" or mean price for this sample of 10 houses in Lowtown is $295,000

Outlier

Measures of Central Tendency: The Median

- In an ordered array, the median is the “middle” number (50% above, 50% below)





- Not affected by extreme values

Measures of Central Tendency: Locating the Median

- The location of the median when the values are in numerical order (smallest to largest):

$$\text{Median position} = \frac{n+1}{2} \text{ position in the ordered data}$$

- If the number of values is odd, the median is the middle number

- If the number of values is even, the median is the average of the two middle numbers

Note that $\frac{n+1}{2}$ is not the *value* of the median, only the *position* of the median in the ranked data

Example: Consider the Fancytown data. First, we put the data in numerical increasing order to get

231,000 285,000 287,000 294,000 297,000
299,000 312,000 313,000 315,000 317,000

Since there are 10 (even) data values, the median is the mean of the two values in the middle.

$$Median,\ M = \frac{297{,}000 + 299{,}000}{2} = \$298{,}000$$

Example: Consider the Lowtown data. We put the data in numerical increasing order to get

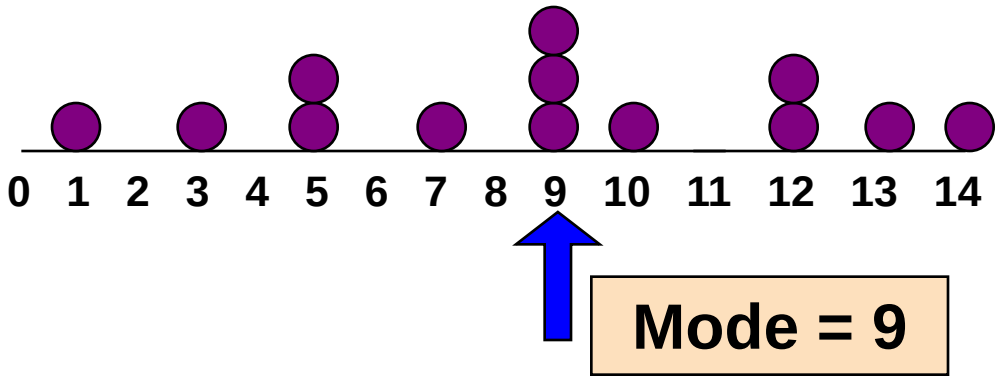
93,000 95,000 97,000 99,000 100,000
110,000 113,000 121,000 122,000, 2,000,000

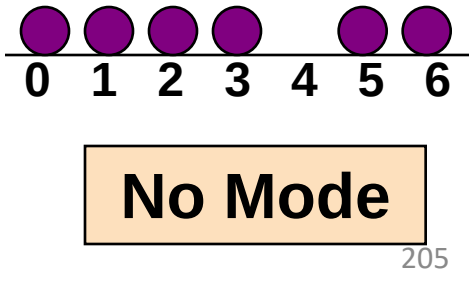
Since there are 10 (even) data values, the median is the mean of the two values in the middle.

$$Median,\ M = \frac{100{,}000 + 110{,}000}{2} = 105{,}000$$

Measures of Central Tendency: The Mode

- Value that occurs most often
- Not affected by extreme values
- Used for either numerical or categorical data
- There may be no mode
- There may be several modes



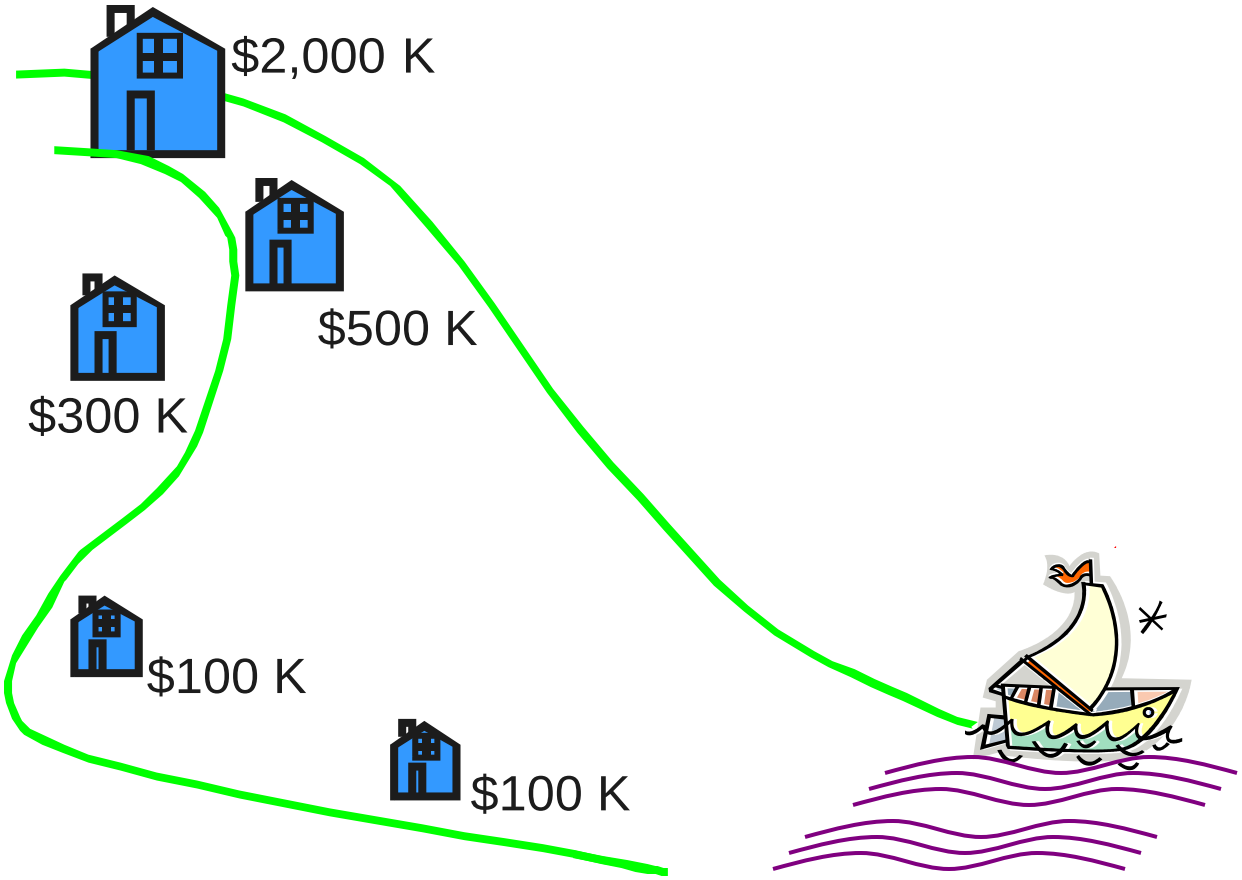


Review Example

- Five houses on a hill by the beach

House Prices:

$2,000,000
500,000
300,000
100,000
100,000



Measures of Central Tendency: Review Example

House Prices:

$2,000,000
$500,000
$300,000
$100,000
$100,000

Sum $**3,000,000**

- **Mean:** ($3,000,000/5)
  = **$600,000**
- **Median:** middle value of ranked data
  = **$300,000**
- **Mode:** most frequent value
  = **$100,000**

Measures of Central Tendency: Which Measure to Choose?

- The **mean** is generally used, unless extreme values (outliers) exist.
- The **median** is often used, since the median is not sensitive to extreme values. For example, median home prices may be reported for a region; it is less sensitive to outliers.
- In some situations it makes sense to report both the **mean** and the **median**.