What is statistics?

- A branch of mathematics taking and transforming numbers into useful information for decision makers
- Methods for processing & analyzing numbers
- Methods for helping reduce the uncertainty inherent in decision making

Types of Statistics

Descriptive Statistics

Collecting, summarizing, and describing data

Inferential Statistics

Drawing conclusions and/or making decisions concerning a population based only on sample data

Why Study Statistics?

Decision Makers Use Statistics To:

- Present and describe business data and information properly
- Draw conclusions about large groups of individuals or items, using information collected from subsets of the individuals or items.
- Make reliable forecasts about a business activity??
- Improve business processes.

Why Learn Statistics?

So you are able to make better sense of the ubiquitous use of numbers:

- Business memos
- Business research
- Technical reports
- Technical journals
- Newspaper articles
- Magazine articles

Data: Are collection of any number of related **observations**

Data set: A collection of data is data set

Data point: A single observation

Raw data: Information before it arranged and analyzed

Data: Information + Noise

Example of raw data:

	HS	Colleage
High school and college CGPA	3.6	2.5
	2.6	2.7
	2.7	2.2
	3.7	3.2
	4.0	3.8

Example of raw data:

Pounds of pressure per square inch that concrete can with stand (10 batches)	2500.35	2500.30
	2500.02	2500.10
	2500.34	2500.29
	2499.00	2499.00
	2498.50	2500.00

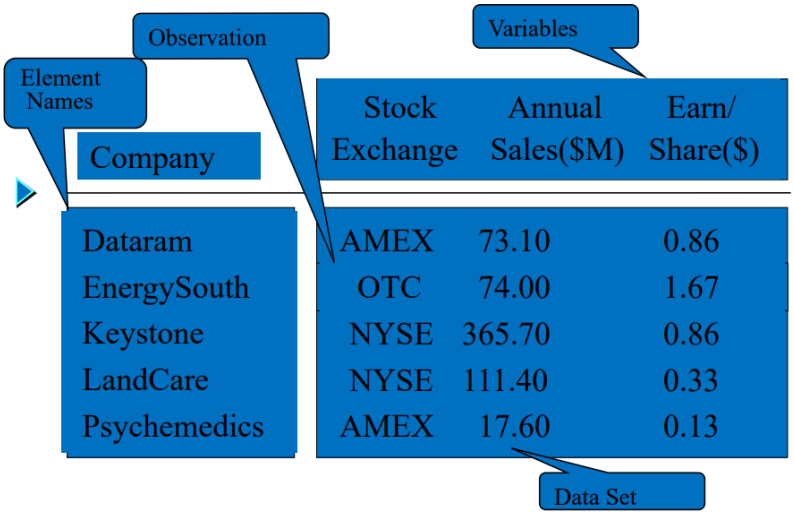
Criteria/Tests for Evaluating Data

Criteria	Issues	Remarks
Specifications & Methodology	Data collection method, response rate, quality & analysis of data, sampling technique & size, questionnaire design, fieldwork.	Data should be reliable, valid, & generalizable to the problem.
Error & Accuracy	Examine errors in approach, research design, sampling, data collection & analysis, & reporting.	Assess accuracy by comparing data from different sources.
Currency	Time lag between collection & publication, frequency of updates.	Census data are updated by syndicated firms.
Objective	Why were the data collected?	The objective determines the relevance of data.
Nature	Definition of key variables, units of measurement, categories used, relationships examined.	Reconfigure the data to increase their usefulness.
Dependability	Expertise, credibility, reputation, and trustworthiness of the source.	Data should be obtained from an original source.

Elements, Variables, and Observations

- The elements are the entities on which data are collected.
- A variable is a characteristic of interest for the elements.
- The set of measurements collected for a particular element is called an observation.
- The total number of data values in a data set is the number of elements multiplied by the number of variables.

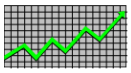
Data, Data Sets, Elements, Variables, and Observations

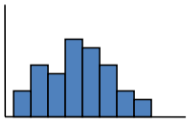


Descriptive Statistics

- Collect data
  - e.g., Survey

- Present data
  - e.g., Tables and graphs





- Characterize data
  - e.g., Sample mean = $\frac{\sum X_i}{n}$

Inferential Statistics

- Estimation
  - e.g., Estimate the population mean weight using the sample mean weight
- Hypothesis testing
  - e.g., Test the claim that the population mean weight is 120 pounds

Drawing conclusions about a large group of individuals based on a subset of the large group.

Basic Vocabulary of Statistics

VARIABLE
A **variable** is a characteristic of an item or individual.

DATA
Data are the different values associated with a variable.

Basic Vocabulary of Statistics

POPULATION
A **population** consists of all the items or individuals about which you want to draw a conclusion.
Ex: People who live within 25 kms of radius from center of the city.

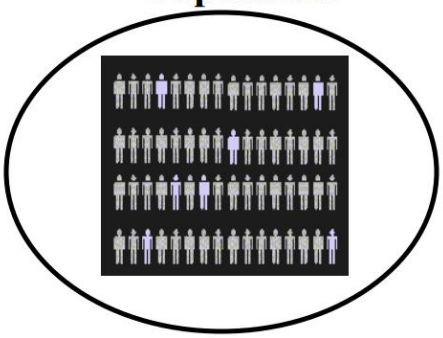
SAMPLE
A **sample** is the portion of a population selected for analysis. It has to be **representative**.

PARAMETER
A **parameter** is a numerical measure that describes a **characteristic** of a **population**.

STATISTIC
A **statistic** is a numerical measure that describes a **characteristic** of a **sample**.

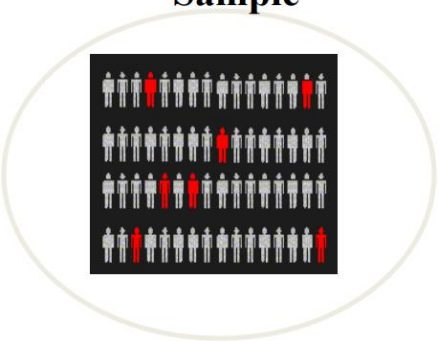
Population vs. Sample

Population



Measures used to describe the population are called **parameters**

Sample



Measures computed from sample data are called **statistics**

	Population	Sample
Definition	Complete enumeration of items is considered	Part of the population chosen for study
Characteristics	Parameters	Statistics
Symbols	Population size = N Population mean = μ Population S.d = σ	Sample size = n Sample mean = $\bar{x}$ Sample S.d = s

Benefits of sample

- Less time
- Less expensive
- Population is large
- Nature of measurement is destructive?

Why Collect Data?

- A marketing research analyst needs to assess the **effectiveness of a new television advertisement.**
- A pharmaceutical manufacturer needs to determine whether a new **drug is more effective** than those currently in use.
- An operations manager wants to monitor a manufacturing process to find out whether the **quality** of the product being manufactured is conforming to company **standards.**
- An auditor wants to review the financial transactions of a company in order to determine whether the company is in **compliance** with generally accepted **accounting principles**.

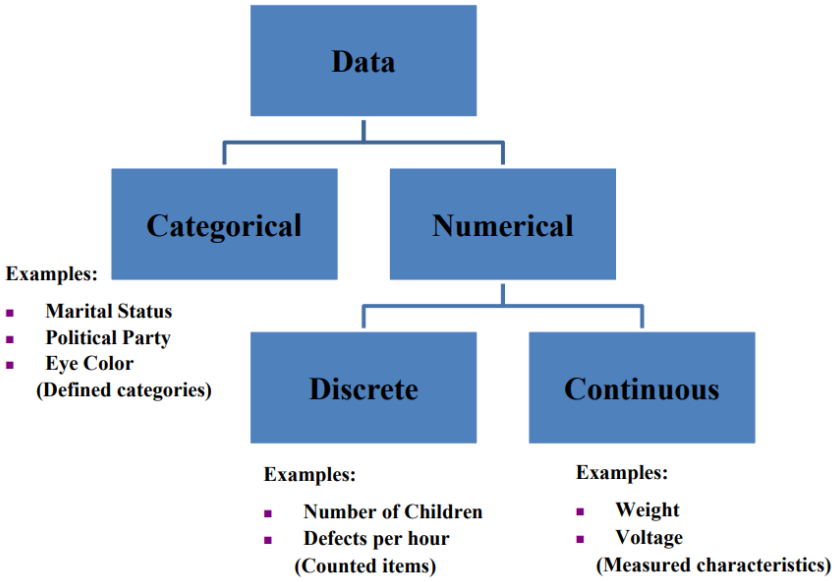
Sources of Data

- Primary Sources: The data **collector is the one using** the data for analysis
  - Data from a political survey
  - Data collected from an experiment
  - Observed data
- Secondary Sources: The **person performing data analysis is not the data collector**
  - Analyzing census data
  - Examining data from print journals or data published on the internet.

Types of Variables

- **Categorical** (qualitative) variables have values that can only be placed into categories, such as "yes" and "no."
- **Numerical** (quantitative) variables have values that represent quantities.

Types of Data



Scales of Measurement

Scales of measurement include:

Nominal	Interval
Ordinal	Ratio

The scale determines the amount of **information** contained in the data.

The scale indicates the data **summarization** and **statistical** analyses that are most appropriate.

Scales of Measurement

- Nominal

Data are labels or names used to identify an attribute of the element.

A nonnumeric label or numeric code may be used.

Example:
Students of a university are classified by the school in which they are enrolled using a nonnumeric label such as Business, Humanities, Education, and so on.

Alternatively, a numeric code could be used for the school variable (e.g. 1 denotes Business,2 denotes Humanities, 3 denotes Education, and so on).

- Ordinal
  - The data have the properties of nominal data and the order or rank of the data is meaningful.
  - A nonnumeric label or numeric code may be used.

Scales of Measurement

- Interval
  - The data have the properties of ordinal data, and the interval between observations is expressed in terms of a **fixed unit of** measure.
  - Interval data are always numeric.

Interval

Example:
Melissa has an SAT score of 1205, while Kevin has an SAT score of 1090. Melissa scored 115 points more than Kevin.

Scales of Measurement

- Ratio

The data have all the properties of interval data and the ratio of two values is meaningful.

Variables such as distance, height, weight, and time use the ratio scale.

This scale must contain a zero value that indicates that **nothing exists** for the variable at the zero point.

Scales of Measurement

Ratio

Example:
Melissa's college record shows 36 credit hours earned, while Kevin's record shows 72 credit hours earned. Kevin has twice as many credit hours earned as Melissa.

Qualitative and Quantitative Data

- Data can be further classified as being qualitative or quantitative.
- The statistical analysis that is **appropriate depends** on whether the data for the variable are qualitative or quantitative.
- In general, there are **more** alternatives for statistical analysis when the data are **quantitative**.

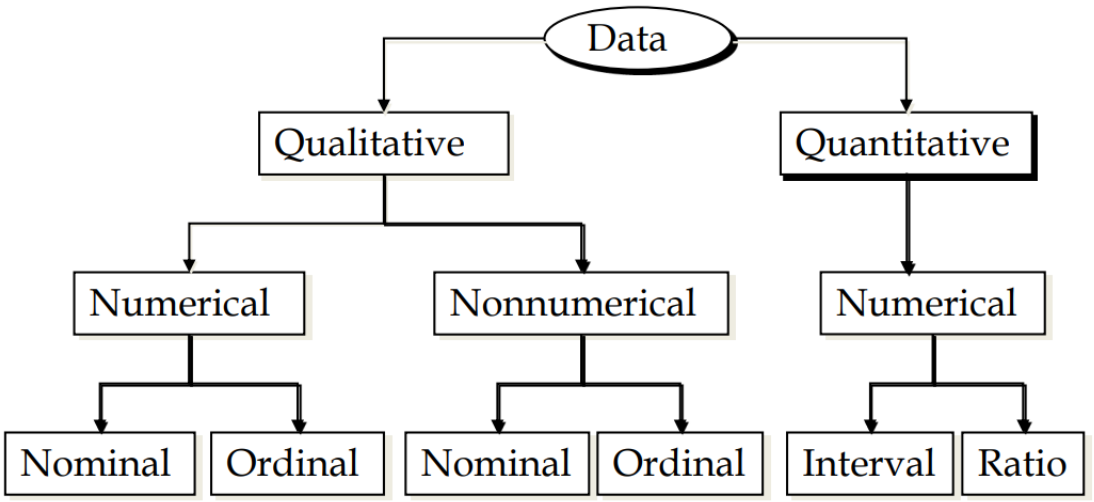
Qualitative Data

- Labels or names used to identify an attribute of each element
- Often referred to as categorical data
- Use either the **nominal or ordinal** scale of measurement
- Can be either numeric or nonnumeric
- Appropriate statistical analyses are **rather limited**

Quantitative Data

- Quantitative data indicate how many or how much:
  - discrete, if measuring how many
  - continuous, if measuring how much
- Quantitative data are always numeric.
- Ordinary arithmetic operations are meaningful for quantitative data.

Scales of Measurement



Cross-Sectional Data

- Cross-sectional data are collected at the same or approximately the **same point in time**.
- Example: data detailing the number of building permits issued in June 2017 in each of the District of India

Time Series Data

- Time series data are collected over several time periods.
- Example: data detailing the number of building permits issued in a city in the last 36 months

Data Sources

- **Existing Sources**
  - Within a firm – almost any department
  - Business database services – NSE.
  - Government agencies – Ministries
  - Industry associations – FICCI,CII,etc
  - Special-interest organizations – AICTE,MCI,Graduate Management Admission Council
  - Internet – more and more firms

Data Sources

- **Statistical Studies**

In experimental studies the **variables** of interest are first identified. Then one or more factors are **controlled** so that data can be obtained about how the **factors influence the variables**.

In observational (nonexperimental) studies no attempt is made to control or influence the variables of interest.

a survey is a good example

Data Acquisition Considerations

Time Requirement

- Searching for information can be time consuming.
- Information may no longer be useful by the time it is available.

Cost of Acquisition

- Organizations often charge for information even when it is not their primary business activity.

Data Errors

- Using any data that happens to be available or that were acquired with little care can lead to poor and misleading information.

Descriptive Statistics

- Descriptive statistics are the tabular, graphical, and numerical methods used to summarize data.

Example: Hudson Auto Repair

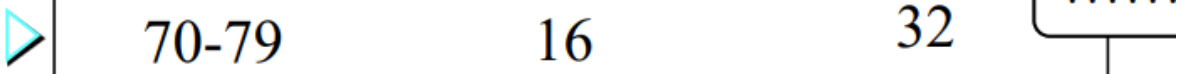
The manager of Hudson Auto would like to have a better understanding of **the cost** of parts used in the engine tune-ups performed in the shop. He examines 50 customer invoices for tune-ups. The costs of parts, rounded to the nearest Rs, are listed on the next slide.

Example: Hudson Auto Repair

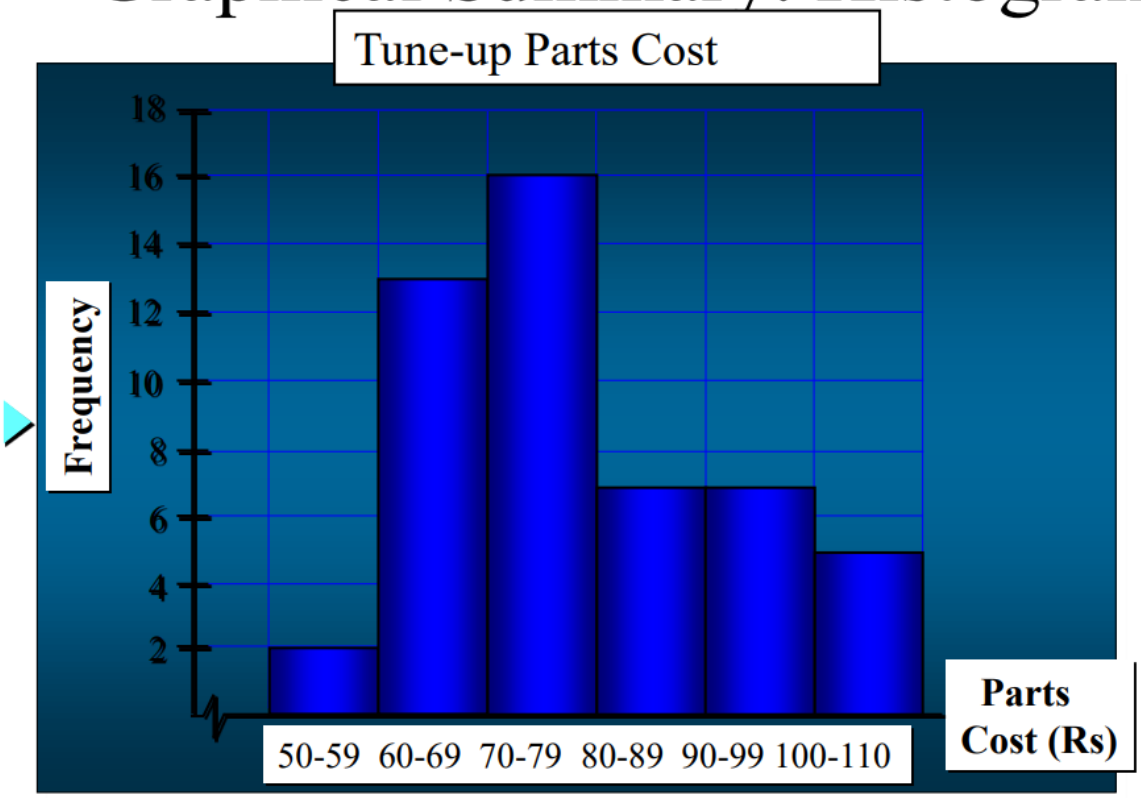
91	78	93	57	75	52	99	80	97	62
71	69	72	89	66	75	79	75	72	76
104	74	62	68	97	105	77	65	80	109
85	97	88	68	83	68	71	69	67	74
62	82	98	101	79	105	79	69	62	73

Tabular Summary:
Frequency and Percent Frequency

Parts Cost (Rs)	Parts Frequency	Percent Frequency
50-59	2	4
60-69	13	26
70-79	16	32
80-89	7	14
90-99	7	14
100-109	5	10
	50	100



Graphical Summary: Histogram



Numerical Descriptive Statistics

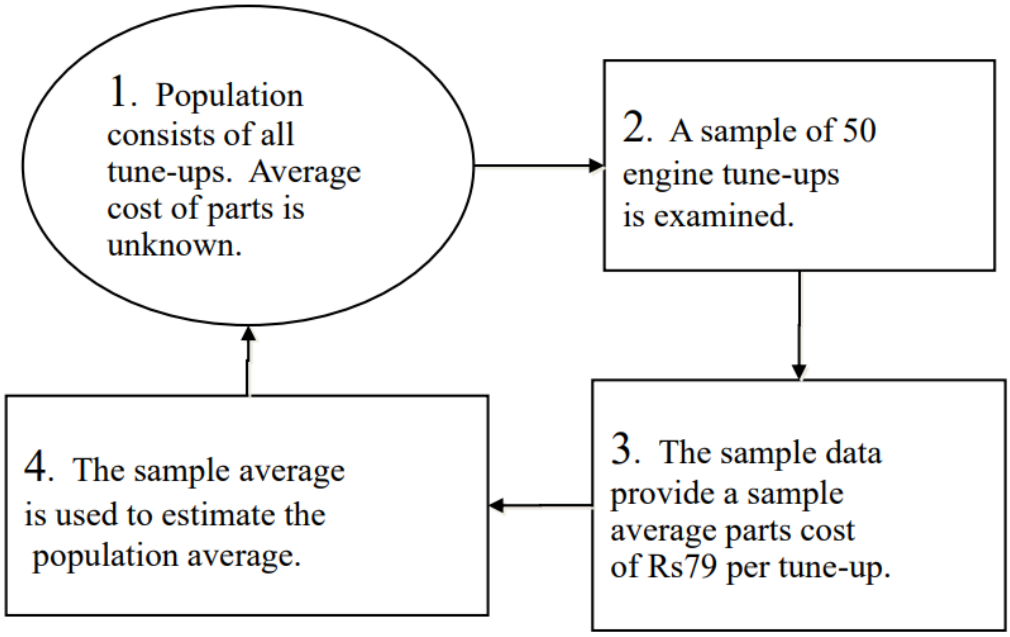
- The most common numerical descriptive statistic is the average (or mean).

- Hudson's average cost of parts, based on the 50 tune-ups studied, is Rs79 (found by summing the 50 cost values and then dividing by 50).

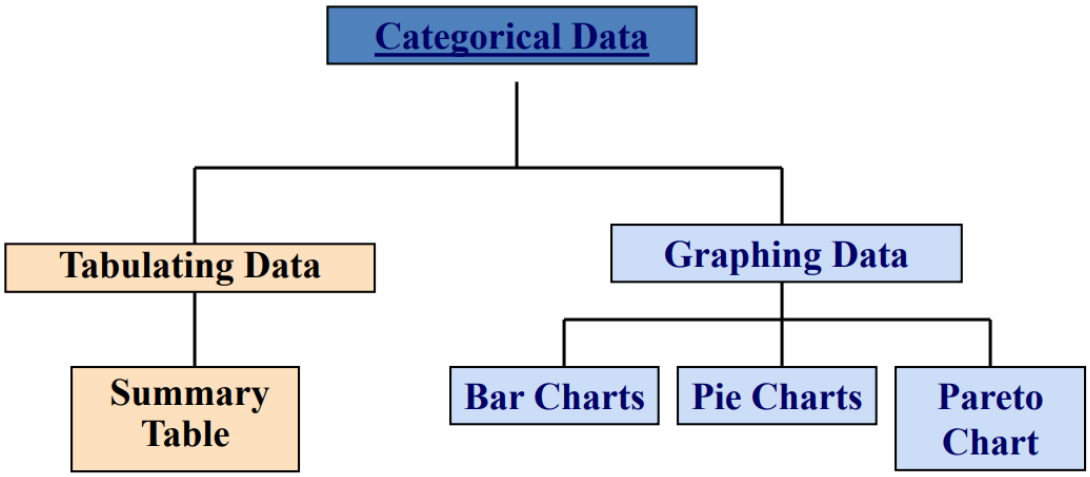
Statistical Inference

- Population – the set of all elements of interest in a particular study
- Sample – a subset of the population
- **Statistical inference** – the process of using data obtained from a sample to make estimates and test hypotheses about the characteristics of a population
- Census – collecting data for a population
- Sample survey – collecting data for a sample

Process of Statistical Inference



Categorical Data Are Summarized By Tables & Graphs



Organizing Categorical Data: Summary Table

- A **summary table** indicates the frequency, amount, or percentage of items in a set of categories so that you can see differences between categories.

Banking Preference?	**Percent**
ATM	16%
Automated or live telephone	2%
Drive-through service at branch	17%
In person at branch	41%
Internet	24%

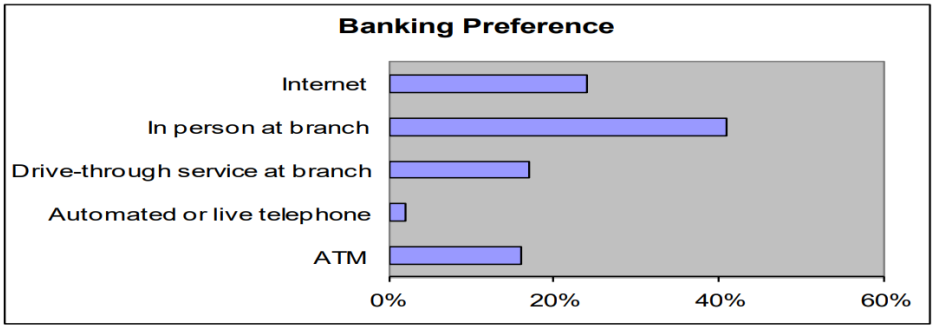
Bar and Pie Charts

- Bar charts and Pie charts are often used for categorical data.

- **Length** of bar or **size** of pie slice shows the **frequency** or **percentage** for each category.

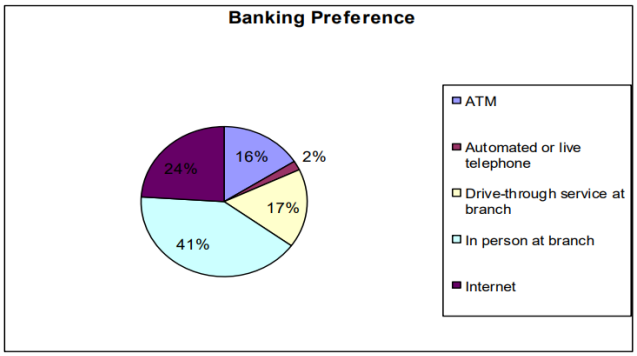
Organizing Categorical Data: Bar Chart

- In a **bar chart,** a bar shows each category, the length of which represents the amount, frequency or percentage of values falling into a category.



Organizing Categorical Data: Pie Chart

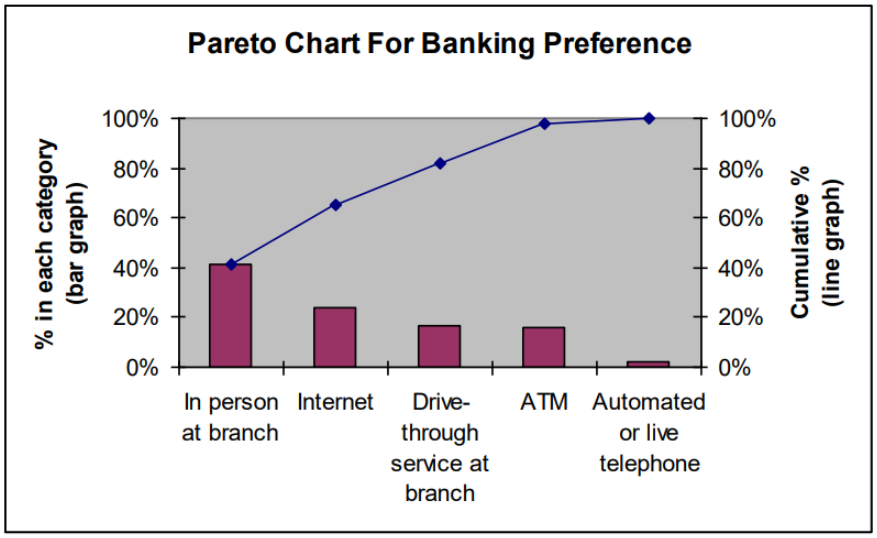
- The **pie chart** is a circle broken up into slices that represent categories. The size of each slice of the pie varies according to the percentage in each category.



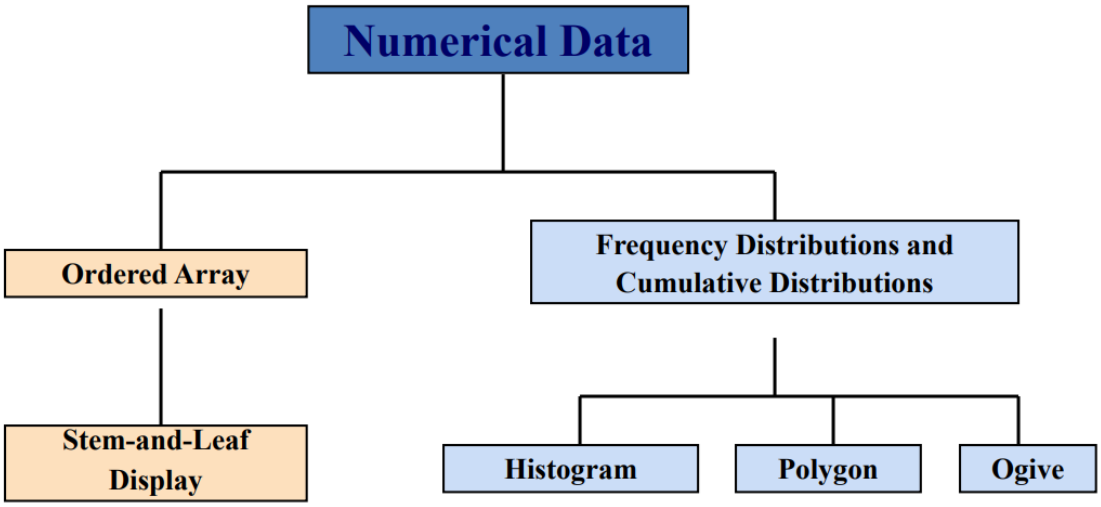
Organizing Categorical Data: Pareto Chart

- Used to portray categorical data (nominal scale)
- **A vertical bar chart, where categories are shown in <u>descending order of frequency</u>**
- A **cumulative polygon** is shown in the same graph
- **Used to separate the "vital few" from the "trivial many"**

Organizing Categorical Data: Pareto Chart



Tables and Charts for Numerical Data



Organizing Numerical Data: Ordered Array

- An **ordered array** is a sequence of data, in rank order, from the **smallest** value to the **largest** value.
- Shows **range** (minimum value to maximum value)
- May help identify **outliers** (unusual observations)
- Which values appear **more than one**
- Divide data in **sections** (Day students- 1/3rd of data below 18, 2/3rd below 22,etc)

Age of Surveyed College Students	**Day Students**					
	16	17	17	18	18	18
	19	19	20	20	21	22
	22	25	27	32	38	42
	Night Students					
	18	18	19	19	20	21
	23	28	32	33	41	45

Stem-and-Leaf Display

- A simple way to see how the data are **distributed and where concentrations** of data exist

METHOD: Separate the sorted data series into **leading** digits (the **stems**) and the **trailing** digits (the **leaves**)

- A **stem-and-leaf display** organizes data into groups (called stems) so that the values within each group (the leaves) branch out to the right on each row.

Age of Surveyed College Students	Day Students					
	16	17	17	18	18	18
	19	19	20	20	21	22
	22	25	27	32	38	42
	Night Students					
	18	18	19	19	20	21
	23	28	32	33	41	45

Age of College Students

Day Students

Stem	Leaf
1	67788899
2	0012257
3	28
4	2

Night Students

Stem	Leaf
1	8899
2	0138
3	23
4	15

Girls		Boys
7, 8, 2, 2, 1	1	5, 8
3, 3, 3, 2	2	2, 2, 3, 6
5, 4, 3	3	4, 5, 5, 5
7, 5, 4	4	0, 0, 2, 7, 9
1, 1, 0	5	0, 0, 1

Stems	Leaves
10	4 7
11	2 5 5 6
12	3
13	0 4
14	5 7

Means 145

Stem and Leaf plot for decimal numbers

8.	0	0						
9.	0							
10.	0	0						
11.	0	0	5					
12.	0	0	0	2				
13.	2	5	8	8				
14.	0	0	0	0	4	6	8	
15.	0	0	5					
16.	0	2	6	8				
17.	0	0	5					
18.	0	2	5					
19.	0	5						
20.	0	5						

Decimal Between Stem and Leaf	Decimal in the Stem
12.3, 12.5, 13.0	1.23, 1.25, 1.30
Becomes	Becomes
12 | 3, 5 13 | 0	1.2 | 3, 5 1.3 | 0
Key: 12 | 3 = 12.3 units	Key: 1.2 | 3 = 1.23 units

Organizing Numerical Data: Frequency Distribution

- The **frequency distribution** is a summary table in which **the data are arranged into numerically ordered classes.**
- You must give attention to selecting the appropriate *number* of **class groupings** for the table, determining a suitable *width* of a class grouping, and establishing the *boundaries* of each class grouping to avoid overlapping.
- The number of classes depends on the number of values in the data. With a **larger** number of values, typically there are **more classes**. In general, a frequency distribution should have at **least 5 but no more than 15 classes**.
- To determine the **width of a class interval,** you divide the **range** (Highest value–Lowest value) of the data by the number of class groupings desired.

Example: A manufacturer of insulation 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

- 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: **5** (usually between 5 and 15)
- Compute class interval (width): **10** (46/5 then round up)
- Determine class boundaries (limits):
  - **Class 1: 10 to less than 20**
  - **Class 2: 20 to less than 30**
  - **Class 3: 30 to less than 40**
  - **Class 4: 40 to less than 50**
  - **Class 5: 50 to less than 60**
- Compute class midpoints: **15, 25, 35, 45, 55**
- Count observations & assign to classes

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	Frequency	Relative Frequency	Percentage
10 but less than 20	3	.15	15
20 but less than 30	6	.30	30
30 but less than 40	5	.25	25
40 but less than 50	4	.20	20
50 but less than 60	2	.10	10
Total	20	1.00	100

Tabulating Numerical Data: Cumulative Frequency

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	Frequency	Percentage	Cumulative Frequency	Cumulative Percentage
10 but less than 20	3	15	3	15
20 but less than 30	6	30	9	45
30 but less than 40	5	25	14	70
40 but less than 50	4	20	18	90
50 but less than 60	2	10	20	100
Total	20	100		

Why Use a Frequency Distribution?

- It **condenses** the raw data into a more useful form
- It allows for a quick **visual interpretation** of the data
- It enables the determination of the major characteristics of the data set including **where the data are concentrated / clustered**

Frequency Distributions: Some Tips

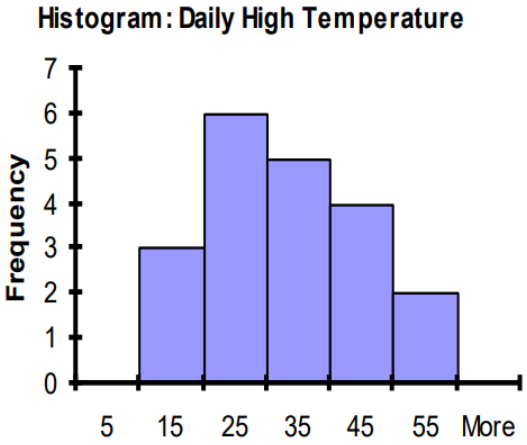
- Different **class boundaries** may provide **different pictures** for the same data (especially for smaller data sets)
- **Shifts in data concentration** may show up when **different class** boundaries are chosen
- As the **size of the data set increases**, the impact of alterations in the **selection of class boundaries is greatly reduced**
- When comparing two or more groups with **different sample sizes,** you must use either a **relative frequency or a percentage distribution**

Organizing Numerical Data: The Histogram

- A **vertical bar chart** of the data in a frequency distribution is called a **histogram.**
- In a histogram there are **no gaps** between adjacent bars.
- The **class boundaries** (or **class midpoints**) are shown on the horizontal axis.
- The vertical axis is either **frequency, relative frequency,** or **percentage**.
- The **height** of the bars represent the **frequency, relative frequency, or percentage.**

Class	Frequency	Relative Frequency	Percentage
10 but less than 20	3	.15	15
20 but less than 30	6	.30	30
30 but less than 40	5	.25	25
40 but less than 50	4	.20	20
50 but less than 60	2	.10	10
Total	20	1.00	100

(In a percentage histogram the vertical axis would be defined to show the percentage of observations per class)



Organizing Numerical Data: The Polygon

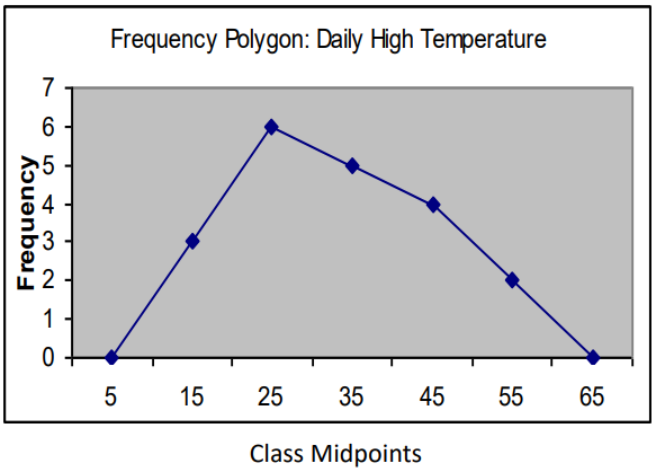
- A **percentage polygon** is formed by having the **midpoint of each class represent the data in that class and then connecting the sequence of midpoints** at their respective class percentages.

- The **cumulative percentage polygon,** or **ogive,** displays the variable of interest along the *X* axis, and the cumulative percentages along the *Y* axis.

- **Useful when there are two or more groups to compare**.

Class	Class Midpoint	Frequency
10 but less than 20	15	3
20 but less than 30	25	6
30 but less than 40	35	5
40 but less than 50	45	4
50 but less than 60	55	2

(In a percentage polygon the **vertical axis** would be defined to show the **percentage of observations per class**)

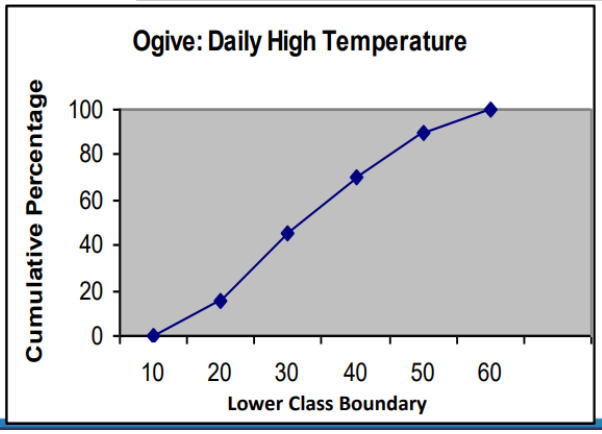


Graphing Cumulative Frequencies: The Ogive (Cumulative % Polygon)

Class	Frequency	Relative Frequency	Percentage
10 but less than 20	3	.15	15
20 but less than 30	6	.30	30
30 but less than 40	5	.25	25
40 but less than 50	4	.20	20
50 but less than 60	2	.10	10
Total	20	1.00	100

Class	Lower class boundary	% less than lower boundary
10 but less than 20	10	15
20 but less than 30	20	45
30 but less than 40	30	70
40 but less than 50	40	90
50 but less than 60	50	100

(In an ogive the percentage of the observations less than each lower class boundary are plotted versus the lower class boundaries.



Cross Tabulations

- Used to study patterns that may exist between **two or more** categorical variables.

- Cross tabulations can be presented in **Contingency Tables**

Cross Tabulations: The Contingency Table

- A **cross-classification** (or **contingency) table** presents the results of **two categorical variables**. The joint responses are classified so that the **categories of one variable are located in the rows** and the categories of the **other variable are located in the columns**.

- The cell is the **intersection** of the row and column and the value in the cell represents the data corresponding to that specific **pairing** of row and column categories.

A survey was conducted to study the **importance of brand name to consumers as compared to a few years ago**. The results, classified by gender, were as follows:

Importance of Brand Name	Male	Female	Total
More	450	300	750
Equal or Less	3300	3450	6750
Total	3750	3750	7500

Scatter Plots

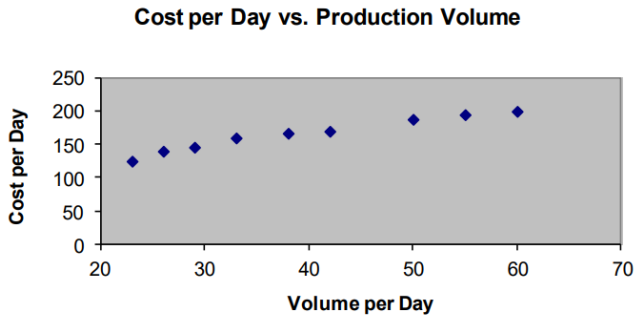
Scatter plots are used for numerical data consisting of paired observations taken **from two *numerical variables***

One variable is measured on the **vertical** axis and the other variable is measured on the **horizontal** axis

Scatter plots are used to examine possible **relationships** between two numerical variables

Scatter Plot Example

Volume per day	Cost per day
23	125
26	140
29	146
33	160
38	167
42	170
50	188
55	195
60	200



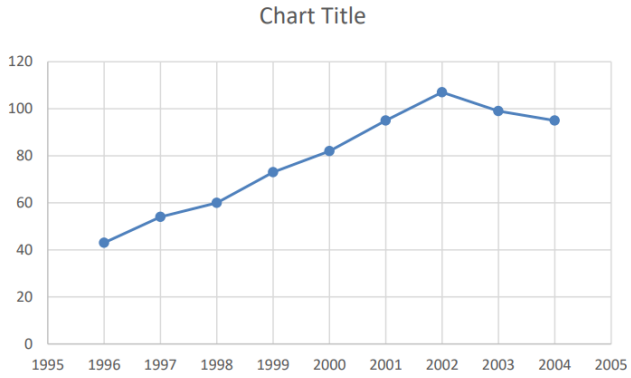
Time Series Plot

- A Time Series Plot is used to study **patterns** in the values of a numeric variable over time

- The Time Series Plot:
  - Numeric variable is measured on the vertical axis and the **time** period is measured on the **horizontal** axis

Time Series Plot Example

Year	Number of Franchises
1996	43
1997	54
1998	60
1999	73
2000	82
2001	95
2002	107
2003	99
2004	95



Principles of Excellent Graphs

The graph should not **distort** the data.

The graph should not contain **unnecessary** adornments (sometimes referred to as chart junk**).**

The scale on the vertical axis should **begin at zero.**

All axes should be properly **labeled**.

The graph should contain a **title**.

The simplest possible graph should be used for a given set of data.

1. Use these data to construct relative frequency using (a) 7 equal intervals and 13 equal intervals.

83 51 66 61 82 65 54 56 92 60 65 87 68 64 51 70 75 66
74 68 44 55 78 69 98 67 82 77 79 62 **38** 88 76 **99** 84 47
60 42 66 74 91 71 83 80 68 65 51 56 73 55

(b) Is policy appropriate for 50 % age people.
© Which distribution is better for (a)
(d) Could you estimate which interval is better between 45-50?

7 intervals

Class	Relative Frequency
30-39	0.02
40-49	0.06
50-59	0.16
60-69	**0.32**
70-79	0.20
80-89	0.16
90-99	0.08
	1.00

13 Intervals

Class	Relative Frequency	Class	Relative Frequency
35-39	0.02	70-74	0.10
40-44	0.04	75-79	0.10
45-49	**0.02**	80-84	0.12
50-54	0.08	85-89	0.04
55-59	0.08	90-94	0.04
60-64	0.10	95-99	0.04
65-69	0.22		**1.00**

The 13-interval distribution **gives a better estimate** because it has a class for 45-49, whereas the 7-interval distribution lumps together all observations between 40 and 49.

2. Construct a frequency distribution for these given data and a relative frequency distribution. Use intervals of 6 days.

4 12 8 14 11 6 7 13 13 11 11 20 5 19 10 15 24 7 29 6

Class	1-6	7-12	13-18	19-24	25-30
Frequency	4	8	4	3	1
Relative Frequency	0.20	0.40	0.20	0.15	0.05

3. The frequency distribution of 150 people who use ski lift. Construct a histogram for these data.

Class (weight-pounds)	Relative Frequency	Class	Relative Frequency
75-89	10	150-164	23
90-104	11	165-179	9
105-119	23	180-194	9
120-134	26	195-209	6
135-149	31	210-224	2

(a) What can you see from histogram which you cannot infer from the frequency distribution.