Statistics for Scientists – CSC261

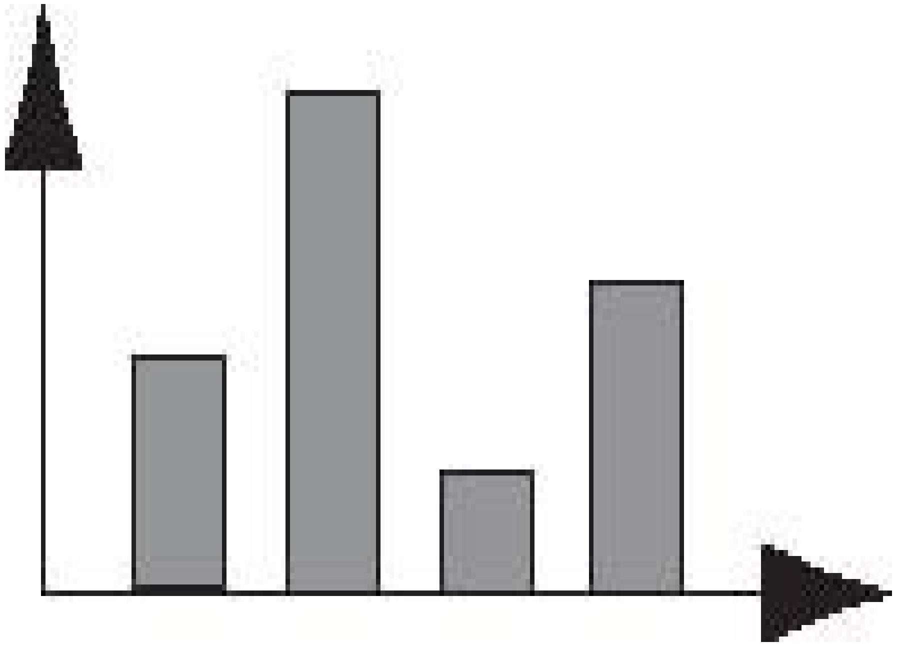
Introduction

Dr DEBAJYOTI PAL

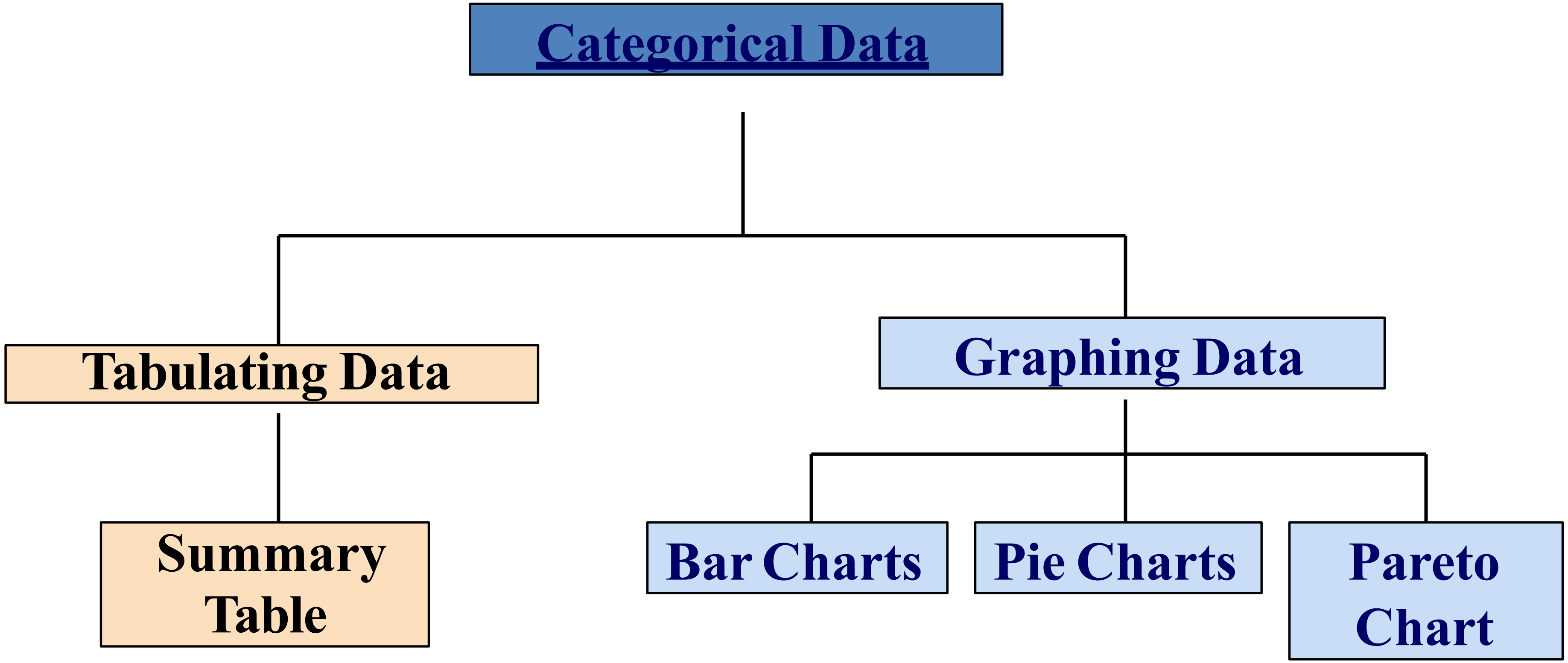
SCHOOL OF INFORMATION TECHNOLOGY,
KMUTT

Where We’re Going

- Describe Data by Using Graphs
- Describe Data by Using Numerical Measures
    - Summation Notation
    - Central Tendencies
    - Variability
    - The Standard Deviation
    - Relative Standing
    - Outliers
    - Graphing Bivariate Relationships
    - Distorting the Truth



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%

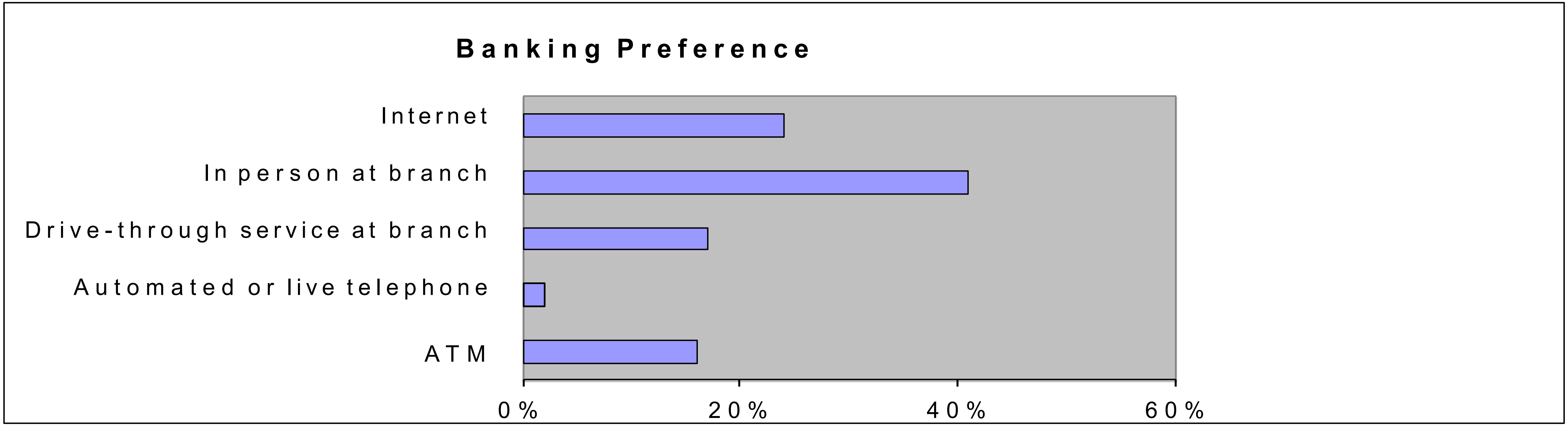
Bar and Pie Charts

- 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.

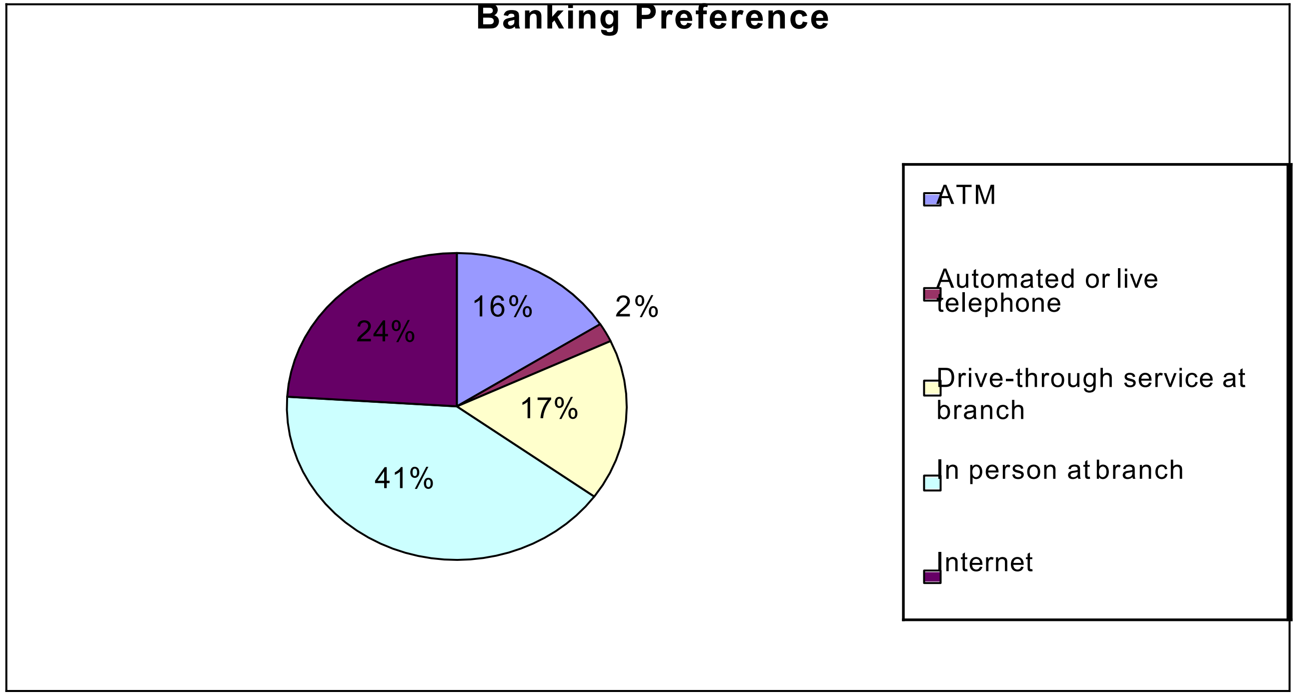
Organizing Categorical Data: Bar Chart

- 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

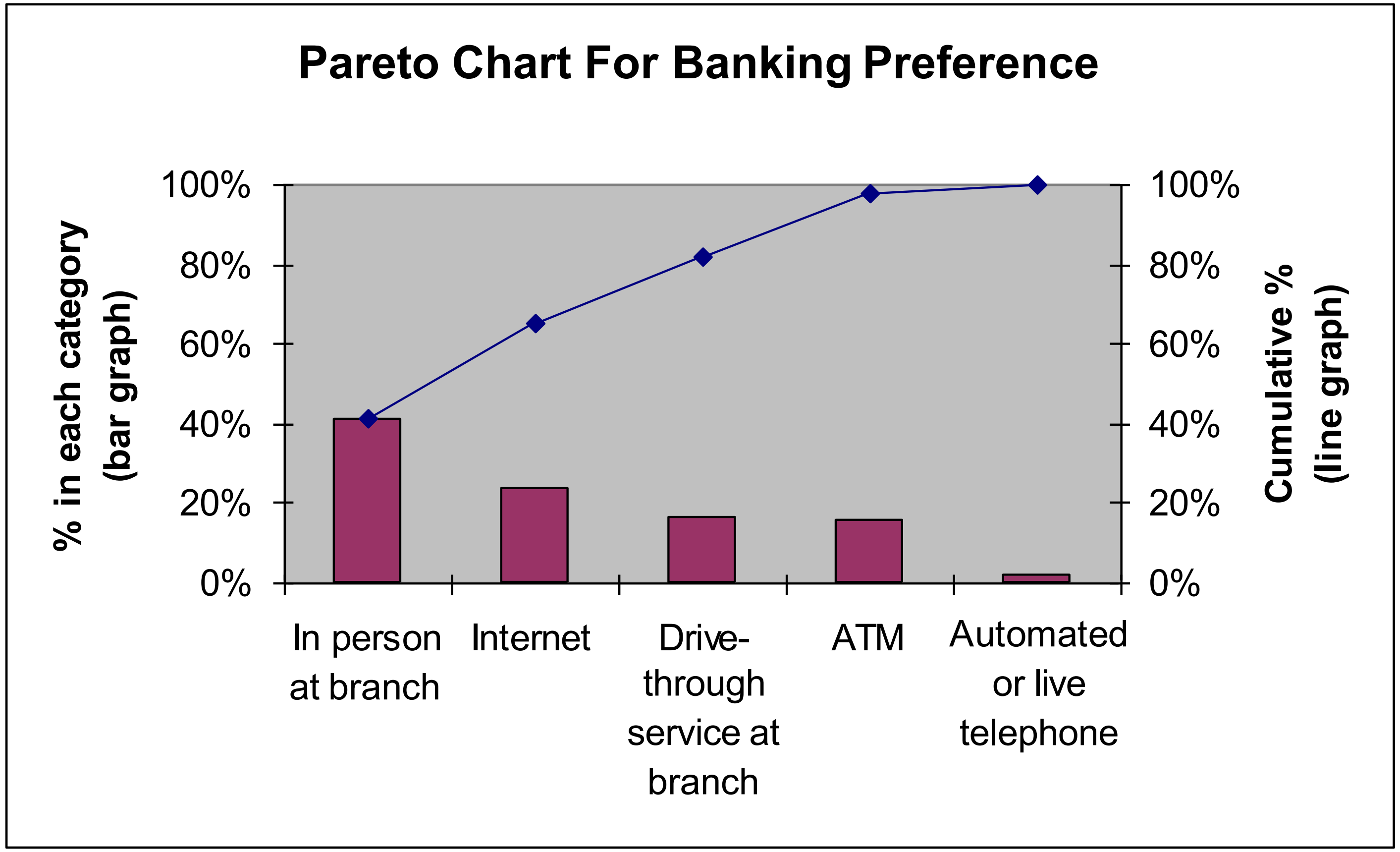
- The **pie chart** is a circle broken up into slices that represent categories. The size of each slice of the pie varies according to the percentage in each category.



Organizing Categorical Data: Pareto Chart

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

Organizing Categorical Data: Pareto Chart



Describing Qualitative Data

Example: Adult Aphasia

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

Describing Qualitative Data

Example: Adult Aphasia

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

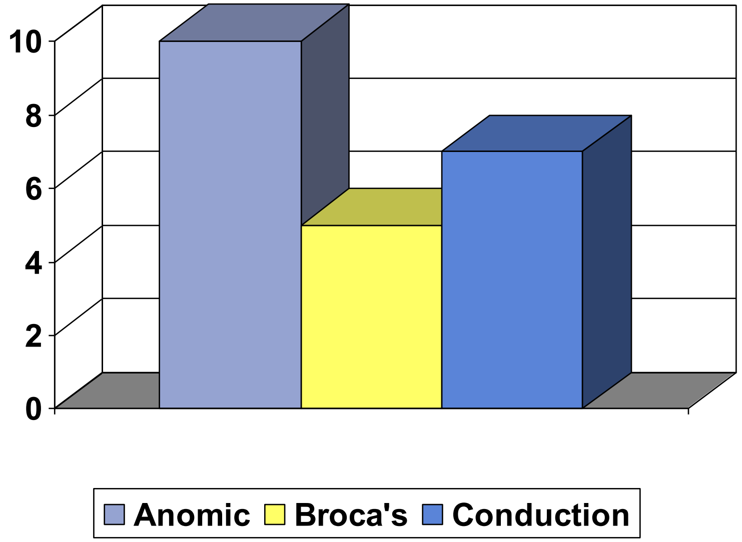
Describing Qualitative Data

Example: Adult Aphasia

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

Describing Qualitative Data

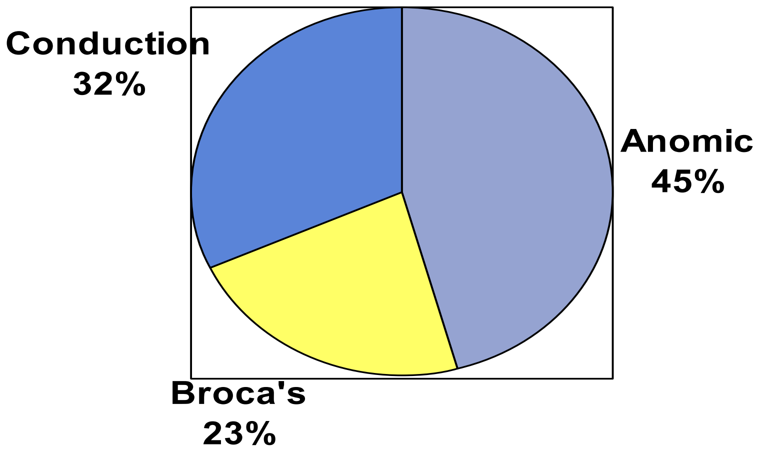
Example: Adult Aphasia



Bar Graph: The categories (classes) of the qualitative variable are represented by bars, where the height of each bar is either the class frequency, class relative frequency or class percentage.

Describing Qualitative Data

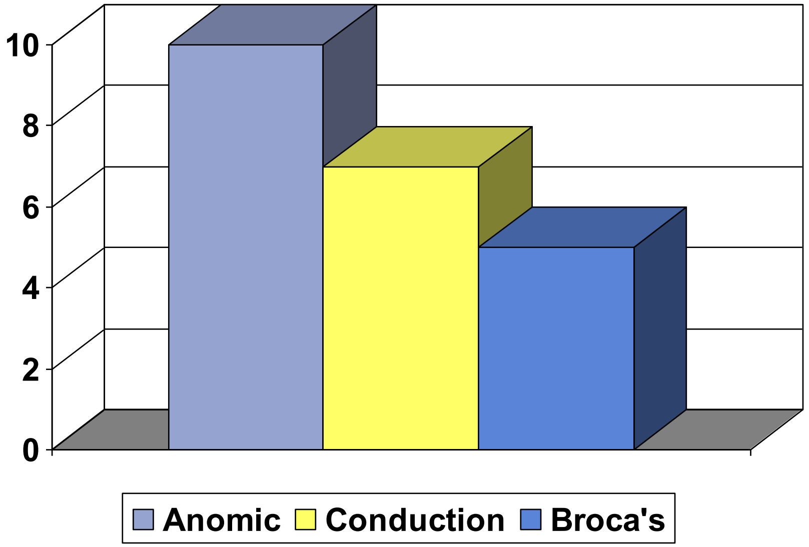
Example: Adult Aphasia



Pie Chart: The categories (classes) of the qualitative variable are represented by slices of a pie. The size of each slice is proportional to the class relative frequency.

Describing Qualitative Data

Example: Adult Aphasia



Pareto Diagram: A bar graph with the categories (classes) of the qualitative variable (i.e., the bars) arranged in height in descending order from left to right.

Organizing Numerical Data: Ordered Array

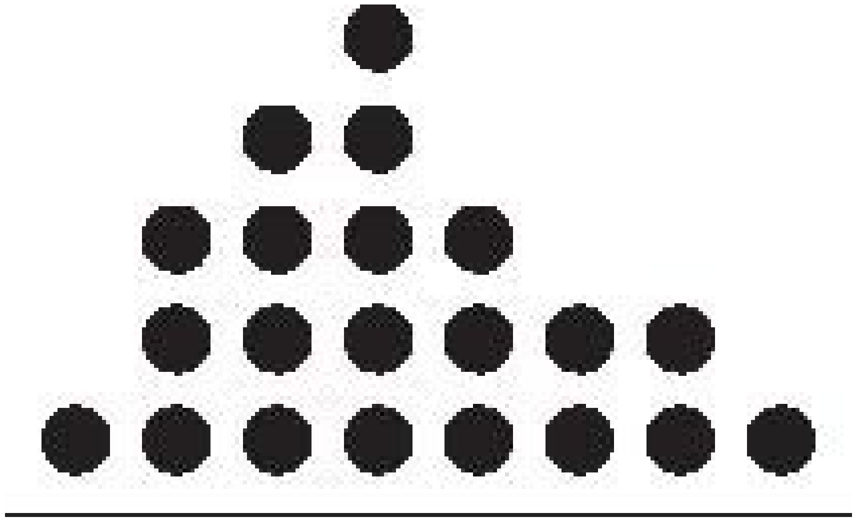
- An **ordered array** is a sequence of data, in rank order, from the **smallest** value to the **largest** value.
- Shows **range** (minimum value to maximum value)
- May help identify **outliers** (unusual observations)
- Which values appear **more than one**
- Divide data in **sections** (Day students-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

Graphical Methods for Describing Quantitative Data

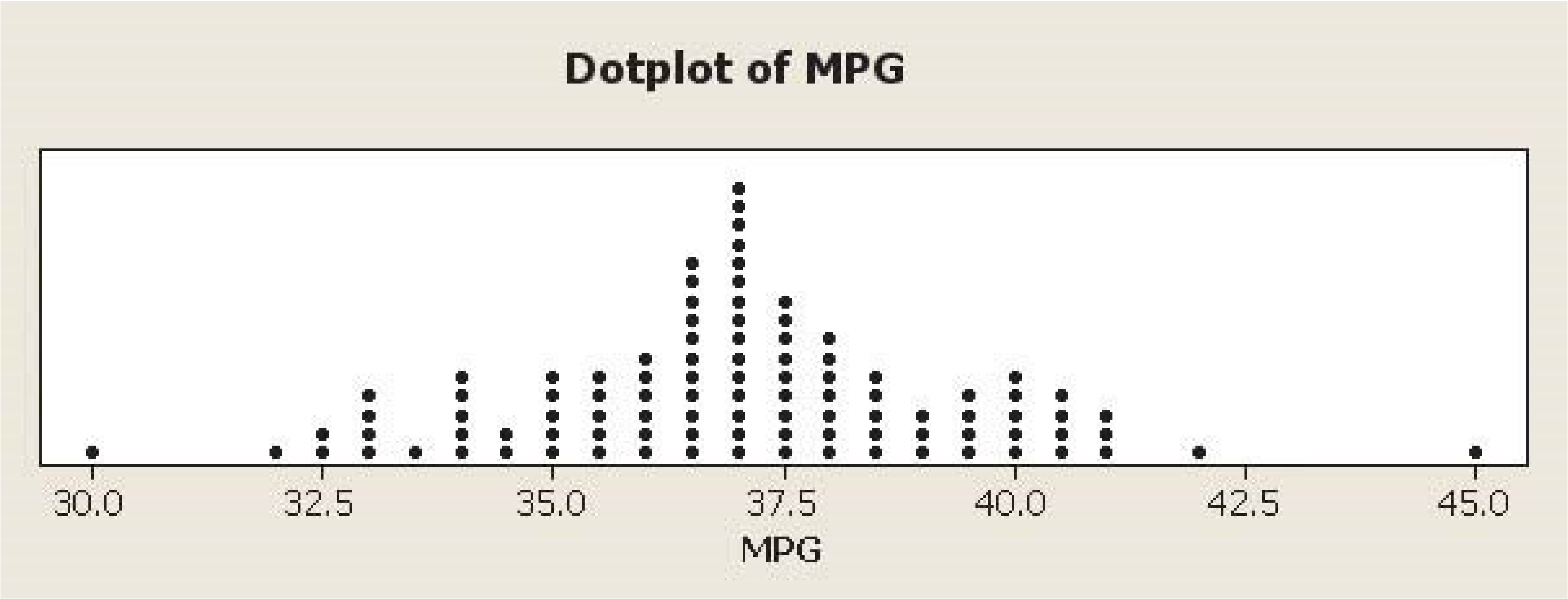
- Dot Plots
  - Dots on a horizontal scale represent the values
    - Good for small data sets
- Stem-and-Leaf Displays
  - Divides values into “stems” and “leafs.”
    - Good for small data sets

Graphical Methods for Describing Quantitative Data



- **Dot plots** display a dot for each observation along a horizontal number line
  - Duplicate values are piled on top of each other
  - The dots reflect the shape of the distribution

Graphical Methods for Describing Quantitative Data



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**)

Organizing Numerical Data: Stem and Leaf Display

- 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

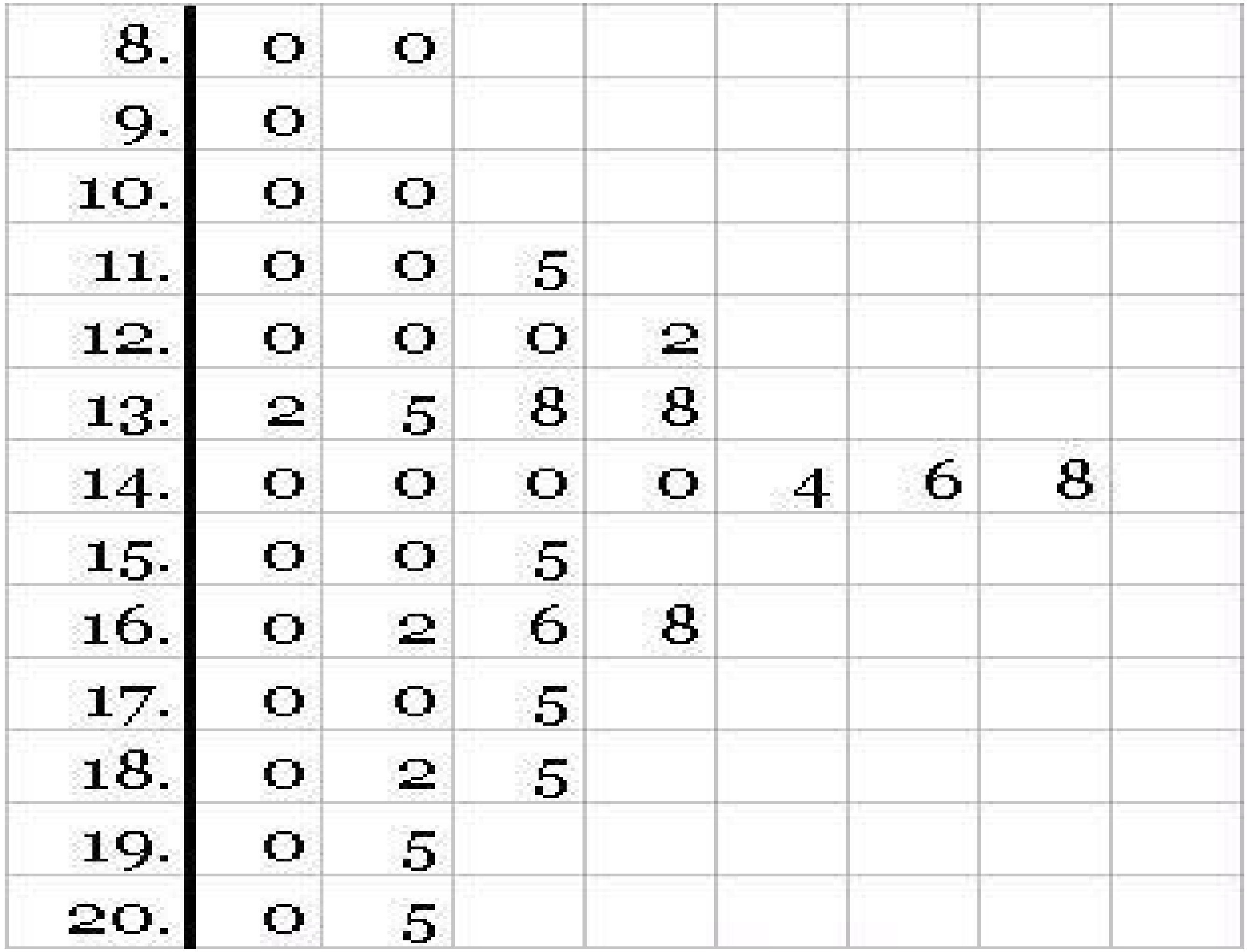
Day Students

Stem	**Leaf**
1	67788899
2	0012257
3	28
4	2

Night Students

Stem	**Leaf**
1	8899
2	0138
3	23
4	15

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						

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.

Organizing Numerical Data: Frequency Distribution Example

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

Organizing Numerical Data: 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: **5** (usually between 5 and 15)
- Compute class interval (width): **10** (46/5 then round up)
- Determine class boundaries (limits):
  - **Class 1: 10 to <u>less than</u> 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

Organizing Numerical Data: 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	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

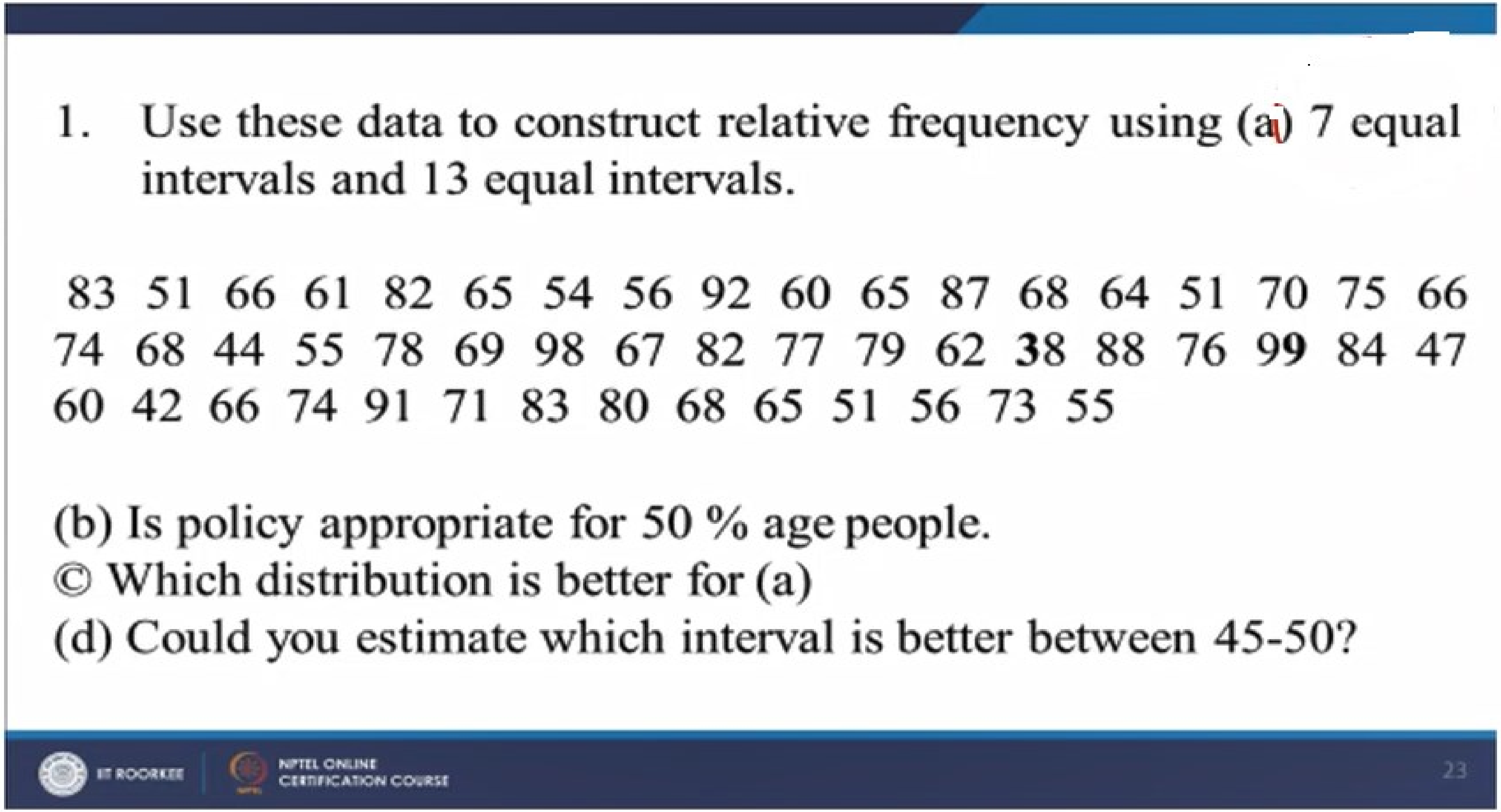
- Different **class boundaries** may provide **different pictures** for the same data (especially for smaller data sets)

- **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 r**elative frequency or a percentage distribution**

Activity



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?

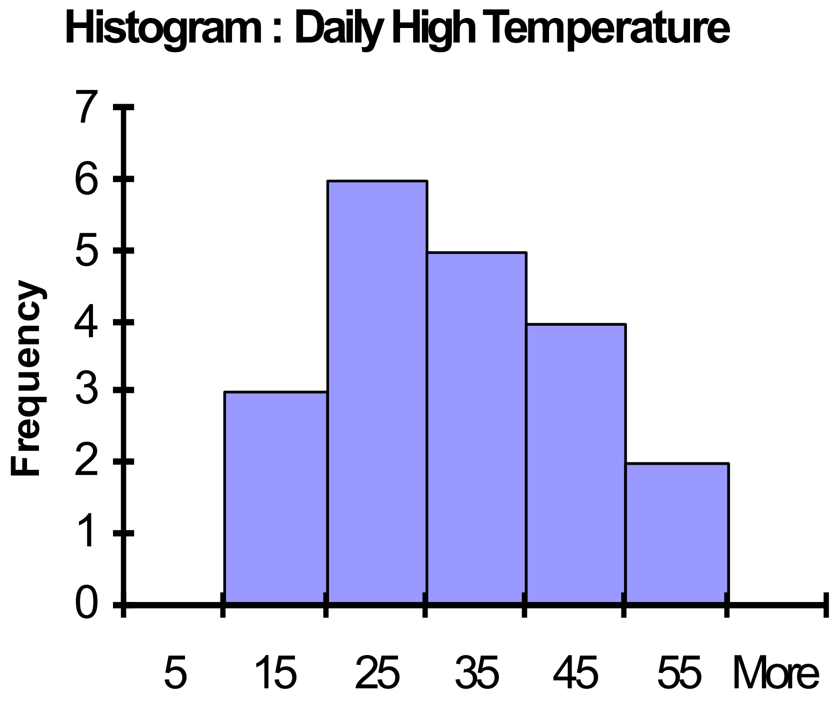
Organizing Numerical Data: The Histogram

- 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.**

Organizing Numerical Data: The Histogram

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 <u>percentage histogram</u> the vertical axis would be defined to show the percentage of observations per class)


Histogram : Daily High Temperature

Organizing Numerical Data: The Polygon

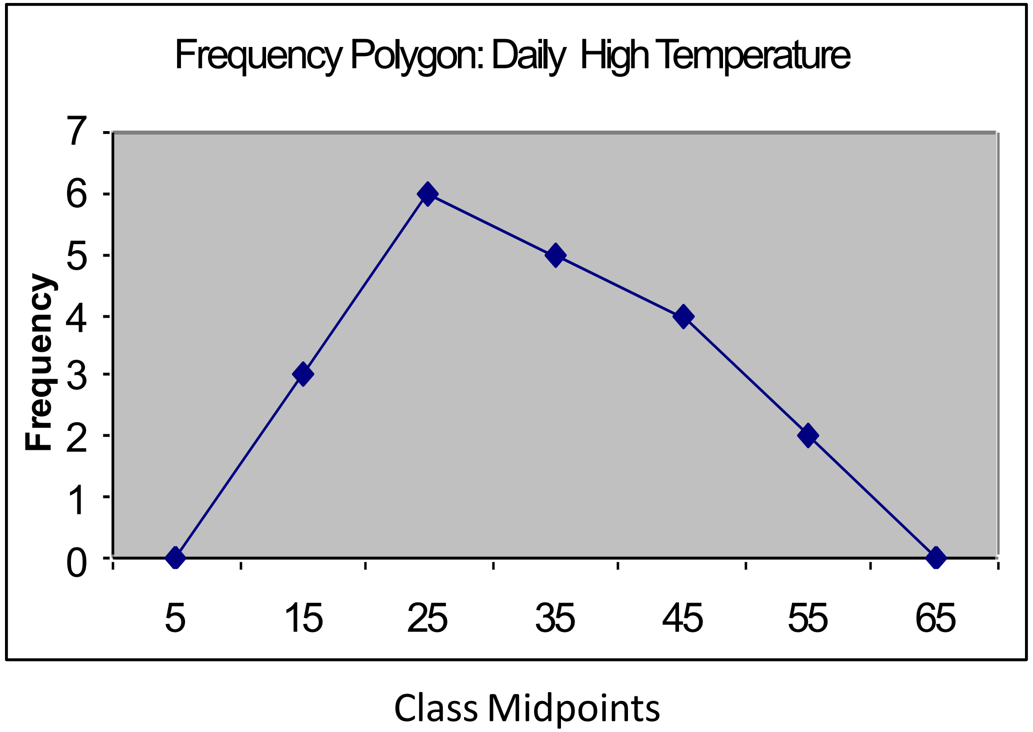
- A **percentage polygon** is formed by having the **midpoint of each class represent the data in that class and then connecting the sequence of midpoints** at their respective class percentages.

- 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**.

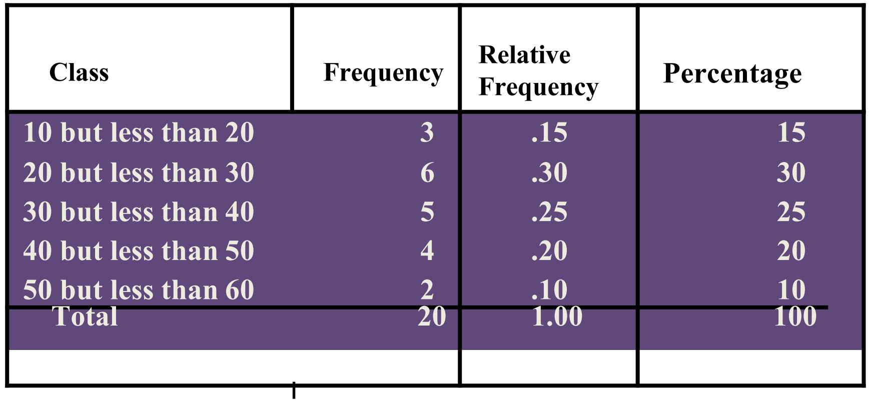
Graphing Numerical Data:The Frequency Polygon

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**)

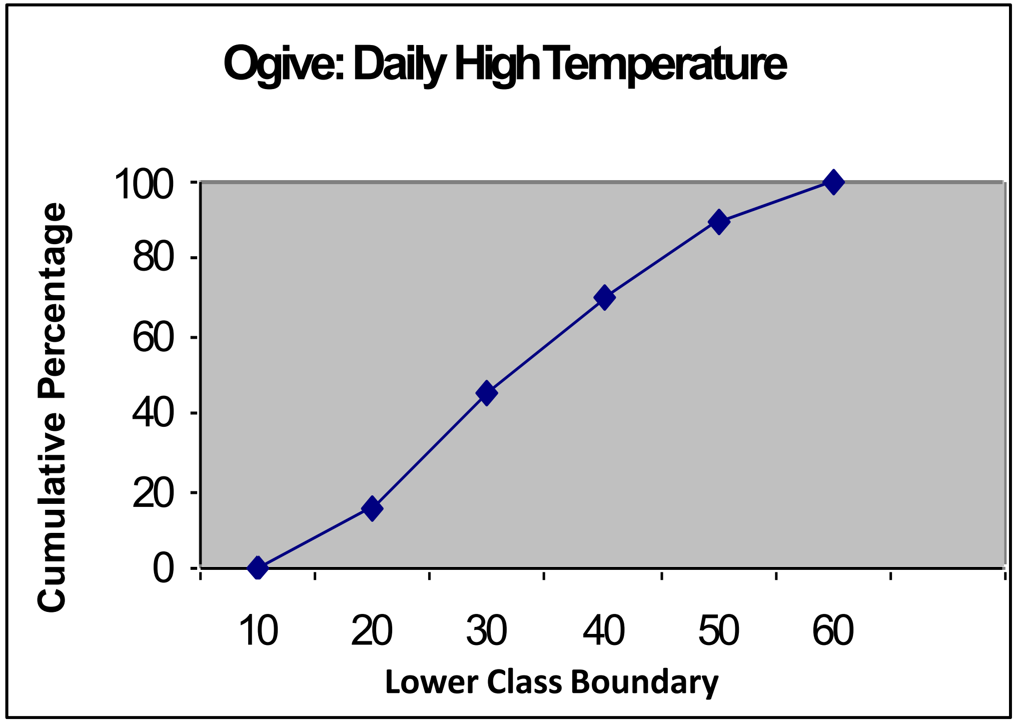
Graphing Cumulative Frequencies: The Ogive (Cumulative % Polygon)



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

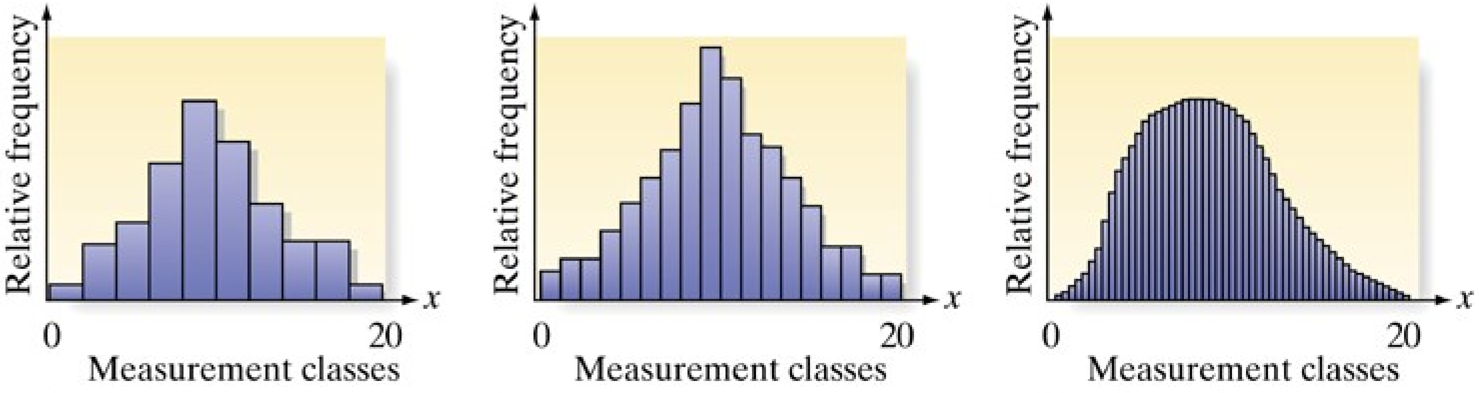
Class	Lower class boundary	% less than lower boundary
10 but less than 20	10	15
20 but less than 30	20	45
30 but less than 40	30	70
40 but less than 50	40	90
50 but less than 60	50	100

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



Graphical Methods for Describing Quantitative Data

- More on Histograms



a. Small data set

b. Larger data set

c. Very large data set

Number of Observations in Data Set	Number of Classes
Less than 25	5-6
25-50	7-14
More than 50	15-20

Scatter Plots

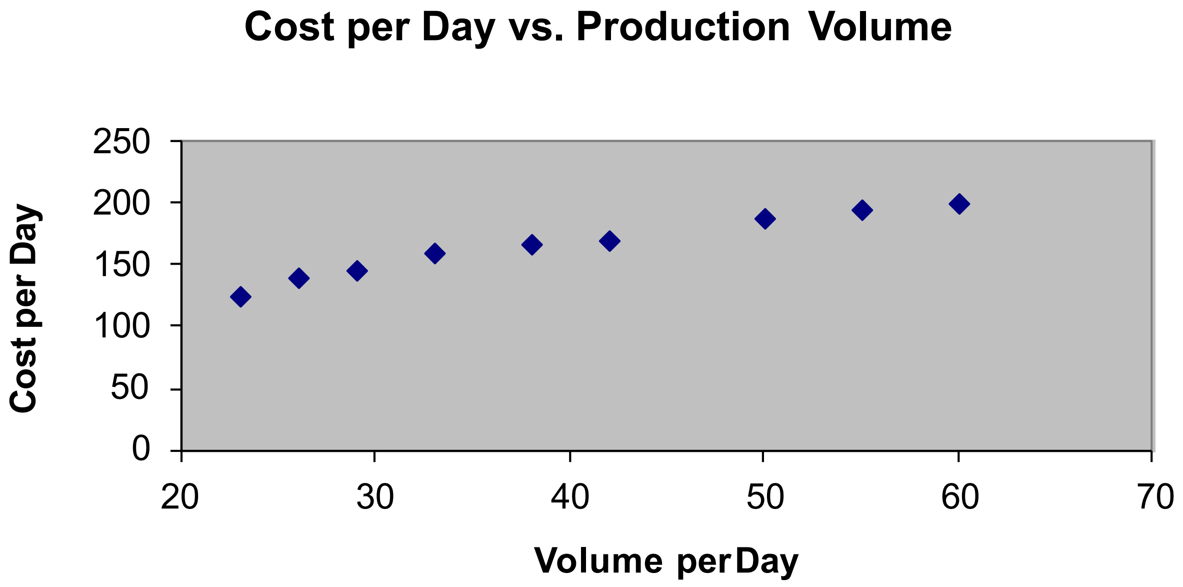
- **Scatter plots** are used for numerical data consisting of paired observations taken **from two *numerical variables***

- 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



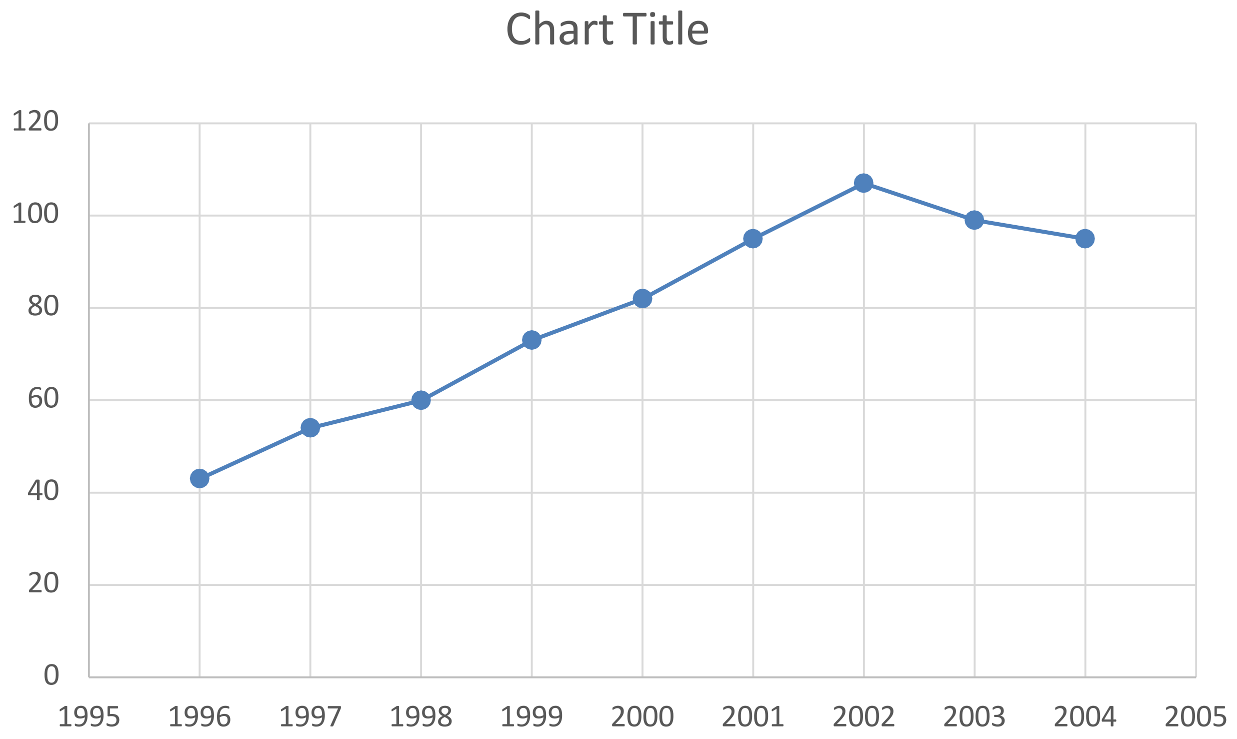
Time Series Plot

- 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.

Summary

In this class, we have

- Organized **categorical** data using the **summary table, bar chart, pie chart, and Pareto chart.**

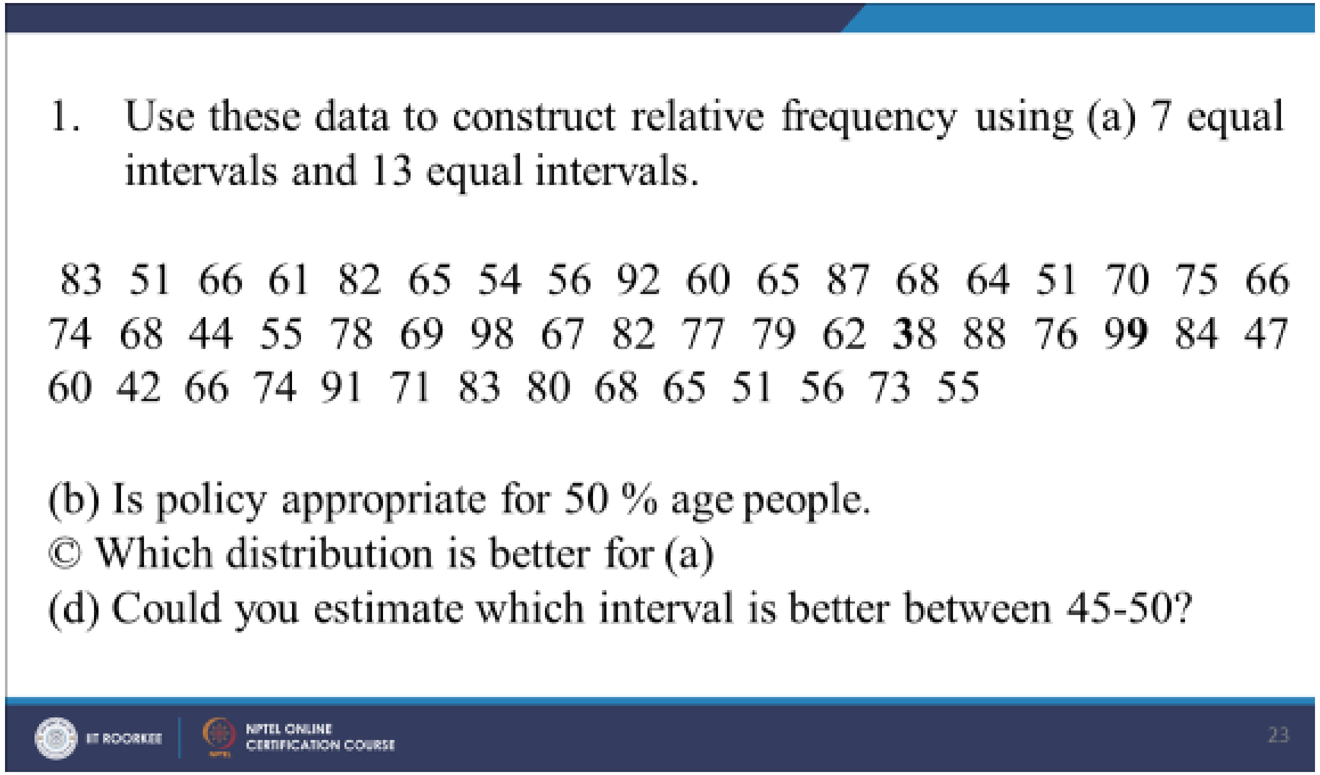
- Organized **numerical** data using the ordered array, **stem-and-leaf display, frequency distribution, histogram, polygon, and ogive.**

- Examined cross tabulated data using the contingency table.

- Developed **scatter plots and time series** graphs.

- Examined the **do’s and don'ts of graphically** displaying data.

Problem 1

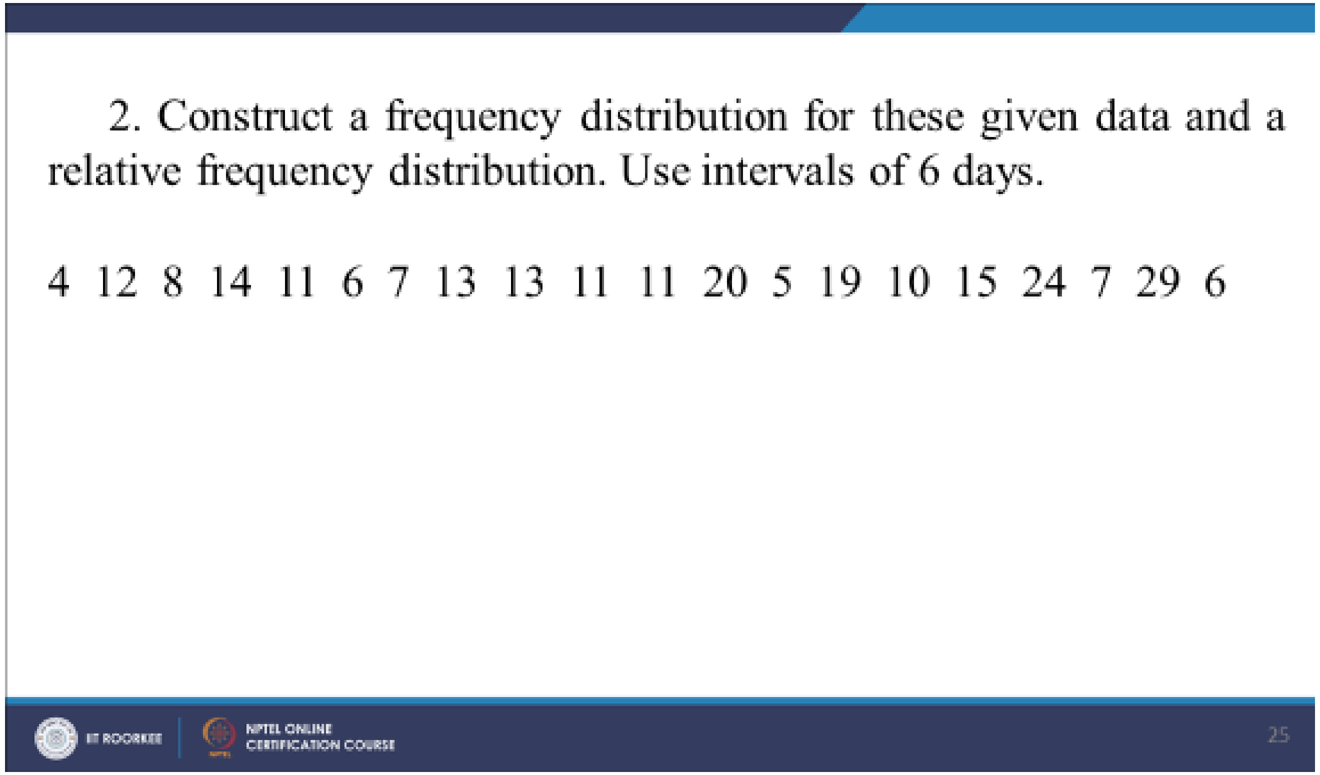


1. Use these data to construct relative frequency using (a) 7 equal intervals and 13 equal intervals.

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?

Problem 2



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