Statistics for Business and Economics

8th Edition

Chapter 1

Describing Data: Graphical

Chapter Goals

After completing this chapter, you should be able to:

- Explain how decisions are often based on incomplete information
- Explain key definitions:
  - Population vs. Sample
  - Parameter vs. Statistic
  - Descriptive vs. Inferential Statistics
- Describe random sampling and systematic sampling
- Explain the difference between Descriptive and Inferential statistics
- Identify types of data and levels of measurement

Chapter Goals

(continued)

After completing this chapter, you should be able to:

- Create and interpret graphs to describe categorical variables:
  - frequency distribution, bar chart, pie chart, Pareto diagram
- Create a line chart to describe time-series data
- Create and interpret graphs to describe numerical variables:
  - frequency distribution, histogram, ogive, stem-and-leaf display
- Construct and interpret graphs to describe relationships between variables:
  - Scatter plot, cross table
- Describe appropriate and inappropriate ways to display data graphically

1.1

Decision Making in an Uncertain Environment

Everyday decisions are based on incomplete information

Examples:

- Will the job market be strong when I graduate?
- Will the price of Yahoo stock be higher in six months than it is now?
- Will interest rates remain low for the rest of the year if the federal budget deficit is as high as predicted?

Decision Making in an Uncertain Environment

(continued)

Data are used to assist decision making

- Statistics is a tool to help process, summarize, analyze, and interpret data

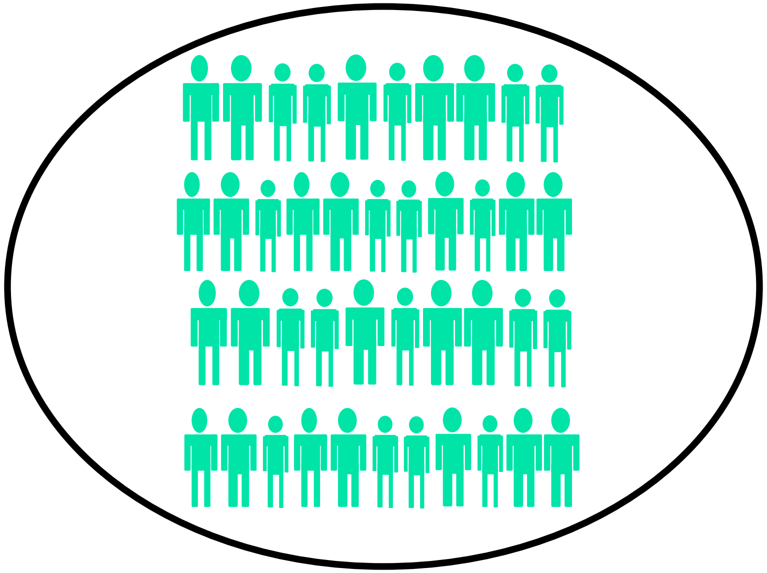
Key Definitions

- A population is the collection of all items of interest or under investigation
  - N represents the population size
- A sample is an observed subset of the population
  - n represents the sample size

- A parameter is a specific characteristic of a population
- A statistic is a specific characteristic of a sample

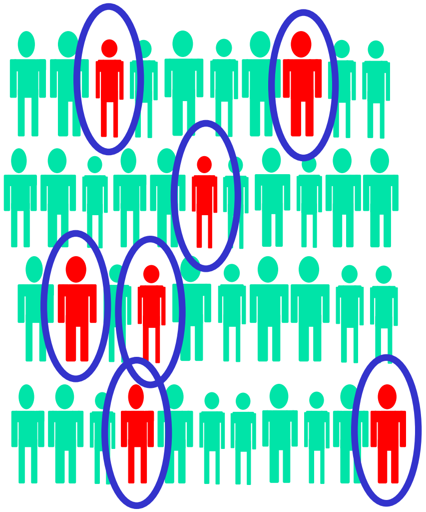
Population vs. Sample

Population



Values calculated using population data are called parameters

Sample



Values computed from sample data are called statistics

Examples of Populations

- Names of all registered voters in the United States
- Incomes of all families living in Daytona Beach
- Annual returns of all stocks traded on the New York Stock Exchange
- Grade point averages of all the students in your university

Random Sampling

Simple random sampling is a procedure in which

- each member of the population is chosen strictly by chance,
- each member of the population is equally likely to be chosen,
- every possible sample of n objects is equally likely to be chosen

The resulting sample is called a random sample

Systematic Sampling

For systematic sampling,

- Assure that the population is arranged in a way that is not related to the subject of interest
- Select every j^{th} item from the population…
- …where j is the ratio of the population size to the sample size, $j = N/n$
- Randomly select a number from 1 to j for the first item selected

The resulting sample is called a systematic sample

Systematic Sampling

(continued)

Example:

Suppose you wish to sample n = 9 items from a population of N = 72.

$$\boxed{j = N / n = 72 / 9 = 8}$$

Randomly select a number from 1 to 8 for the first item to include in the sample; suppose this is item number 3.

Then select every 8^{th} item thereafter

(items 3, 11, 19, 27, 35, 43, 51, 59, 67)

Descriptive and Inferential Statistics

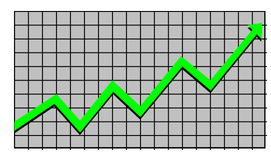
Two branches of statistics:

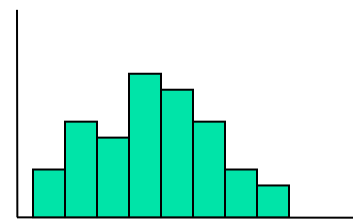
- Descriptive statistics
  - Graphical and numerical procedures to summarize and process data
- Inferential statistics
  - Using data to make predictions, forecasts, and estimates to assist decision making

Descriptive Statistics

- Collect data
  - e.g., Survey

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





- Summarize 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 140 pounds

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

1.2

Classification of Variables

Data

- Categorical
- Numerical
  - Discrete
  - Continuous

Categorical

Examples:

- Marital Status
- Are you registered to vote?
- Eye Color

(Defined categories or groups)

Discrete

Examples:

- Number of Children
- Defects per hour

(Counted items)

Continuous

Examples:

- Weight
- Voltage

(Measured characteristics)

Measurement Levels

Differences between measurements, true zero exists

Ratio Data

Quantitative Data

Differences between measurements but no true zero

Interval Data

Ordered Categories (rankings, order, or scaling)

Ordinal Data

Qualitative Data

Categories (no ordering or direction)

Nominal Data

1.3-1.5

Graphical Presentation of Data

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

Graphical Presentation of Data

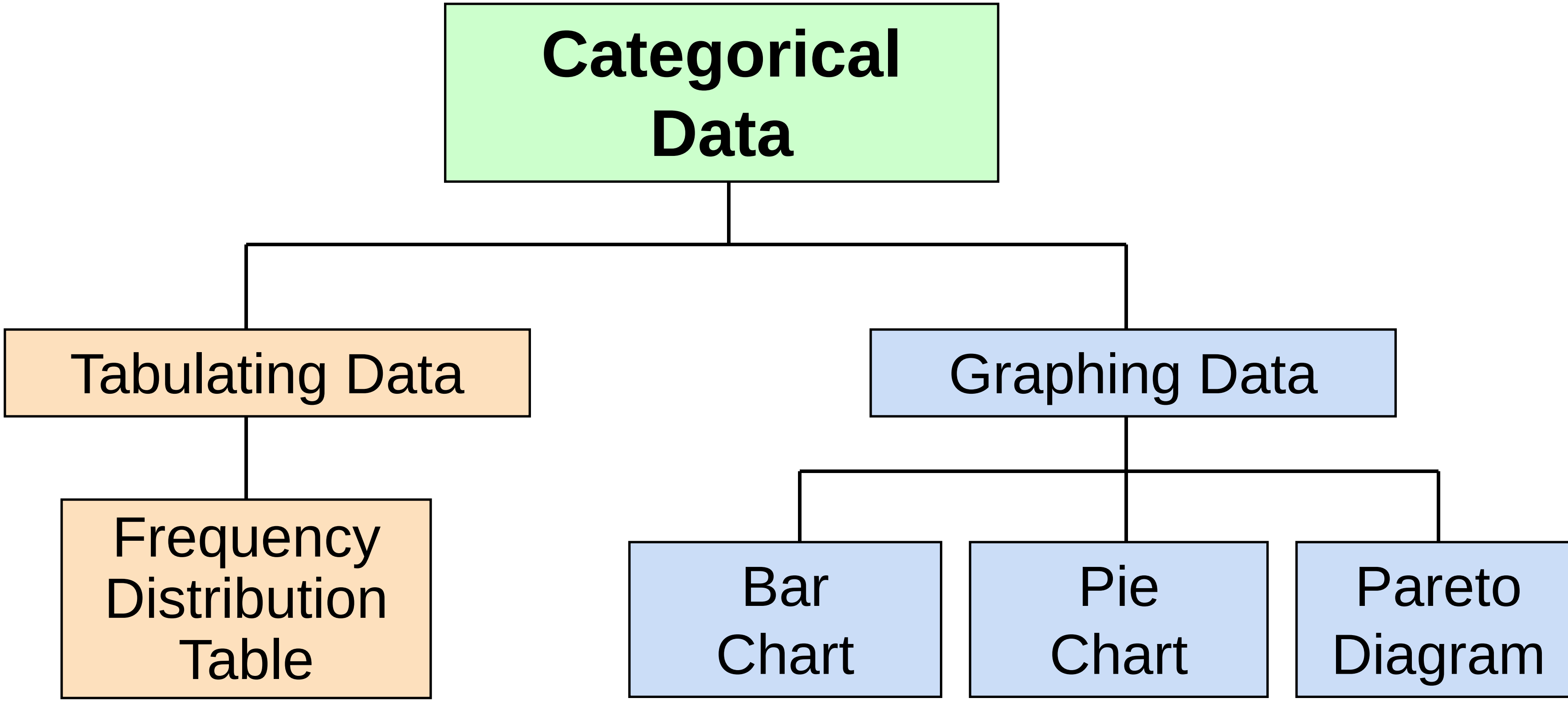
(continued)

- Techniques reviewed in this chapter:

Categorical Variables	Numerical Variables
• Frequency distribution • Cross table • Bar chart • Pie chart • Pareto diagram	• Line chart • Frequency distribution • Histogram and ogive • Stem-and-leaf display • Scatter plot

1.3

Tables and Graphs for Categorical Variables



The Frequency Distribution Table

Summarize data by category

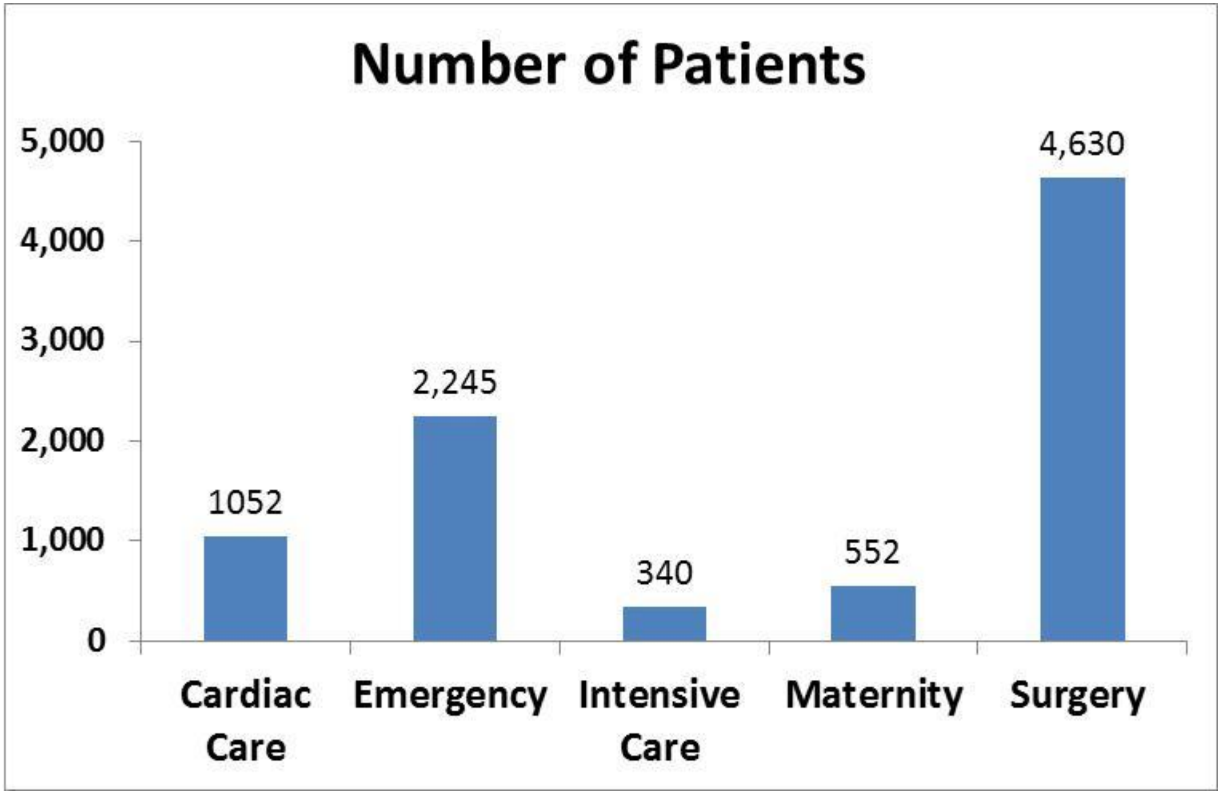
Example: Hospital Patients by Unit

Hospital Unit	Number of Patients	Percent (rounded)
Cardiac Care	1,052	11.93
Emergency	2,245	25.46
Intensive Care	340	3.86
Maternity	552	6.26
Surgery	4,630	52.50
Total:	8,819	100.0

(Variables are categorical)

Graph of Frequency Distribution

- Bar chart of patient data



Cross Tables

- Cross Tables (or contingency tables) list the number of observations for every combination of values for two categorical or ordinal variables

- If there are r categories for the first variable (rows) and c categories for the second variable (columns), the table is called an $r \times c$ cross table

Cross Table Example

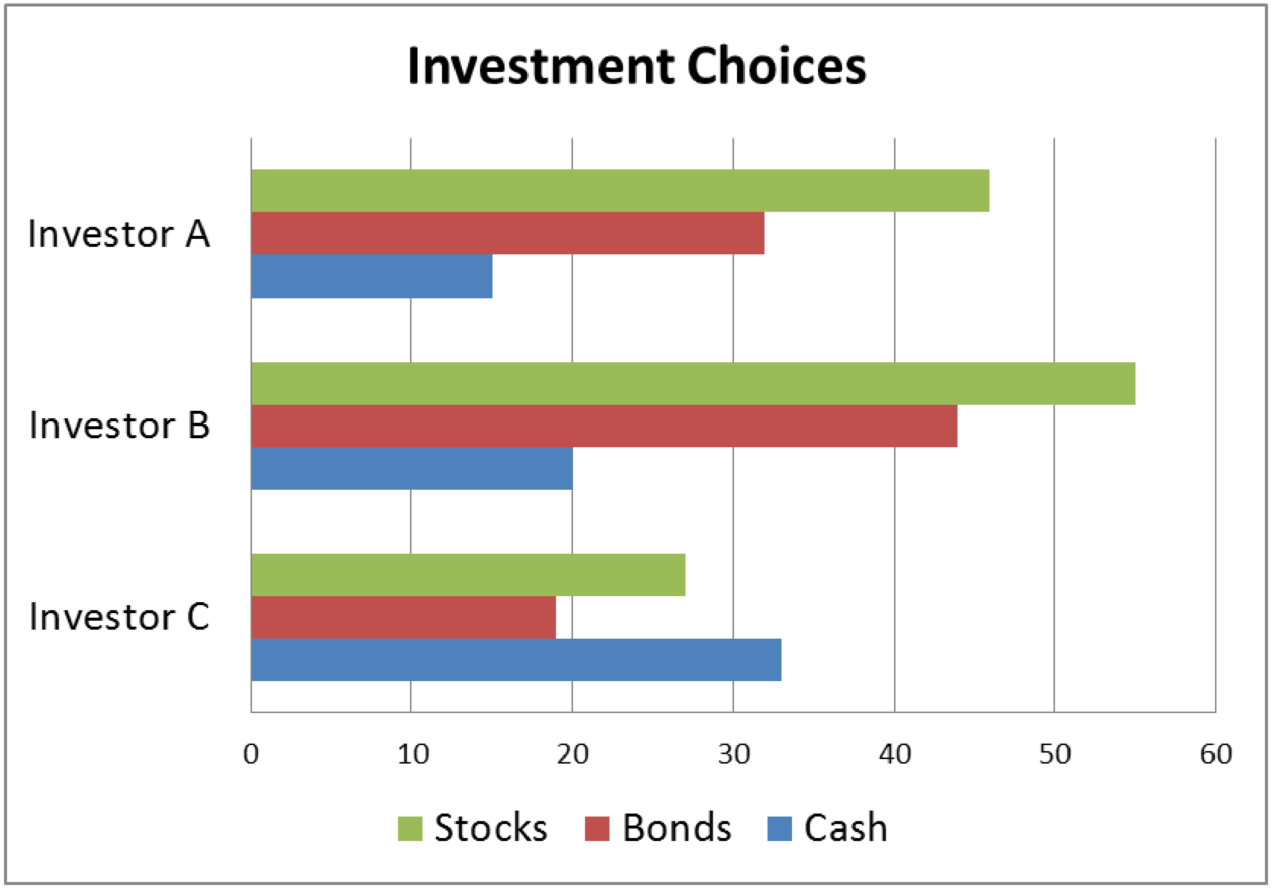
- 3 x 3 Cross Table for Investment Choices by Investor (values in $1000's)

Investment Category	Investor A	Investor B	Investor C	Total
Stocks	46	55	27	**128**
Bonds	32	44	19	**95**
Cash	15	20	33	**68**
Total	**93**	**119**	**79**	**291**

Graphing Multivariate Categorical Data

(continued)

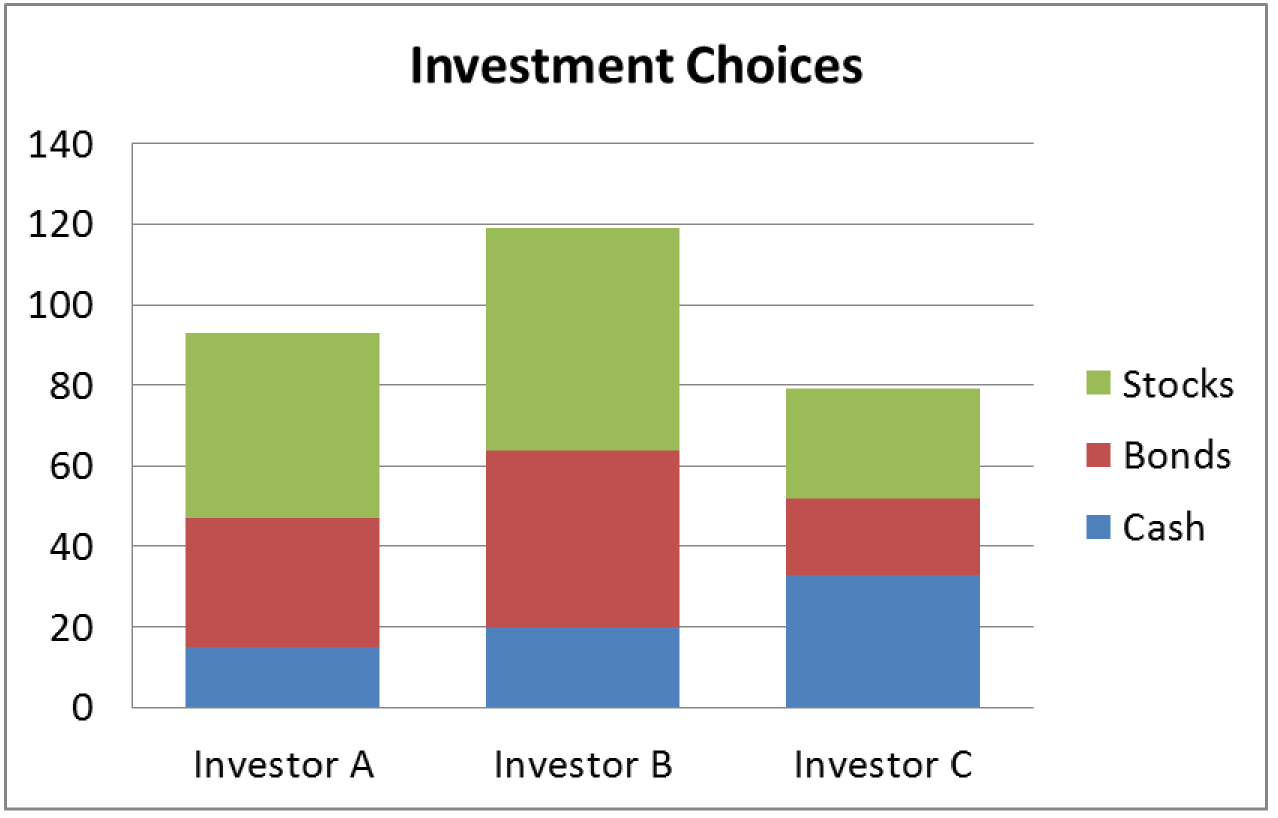
- Side by side horizontal bar chart



Graphing Multivariate Categorical Data

(continued)

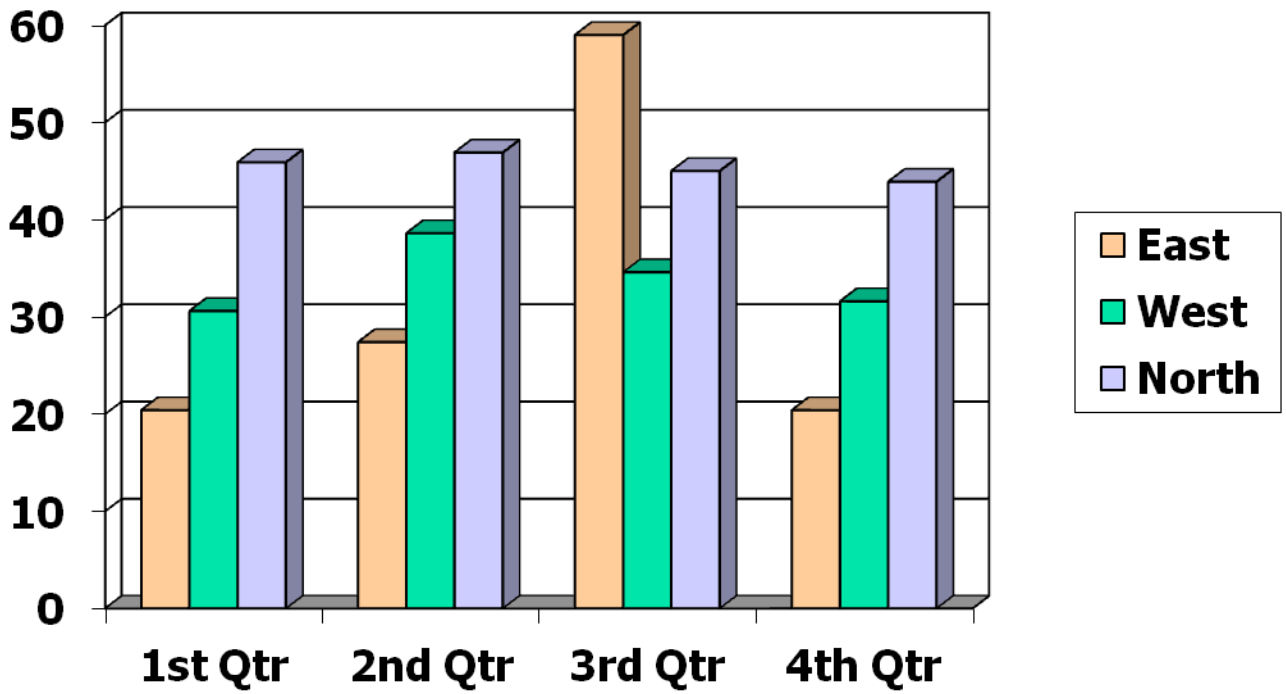
- Stacked bar chart



Vertical
Side-by-Side Chart Example

- Sales by quarter for three sales territories:

	1st Qtr	2nd Qtr	3rd Qtr	4th Qtr
East	20.4	27.4	59	20.4
West	30.6	38.6	34.6	31.6
North	45.9	46.9	45	43.9



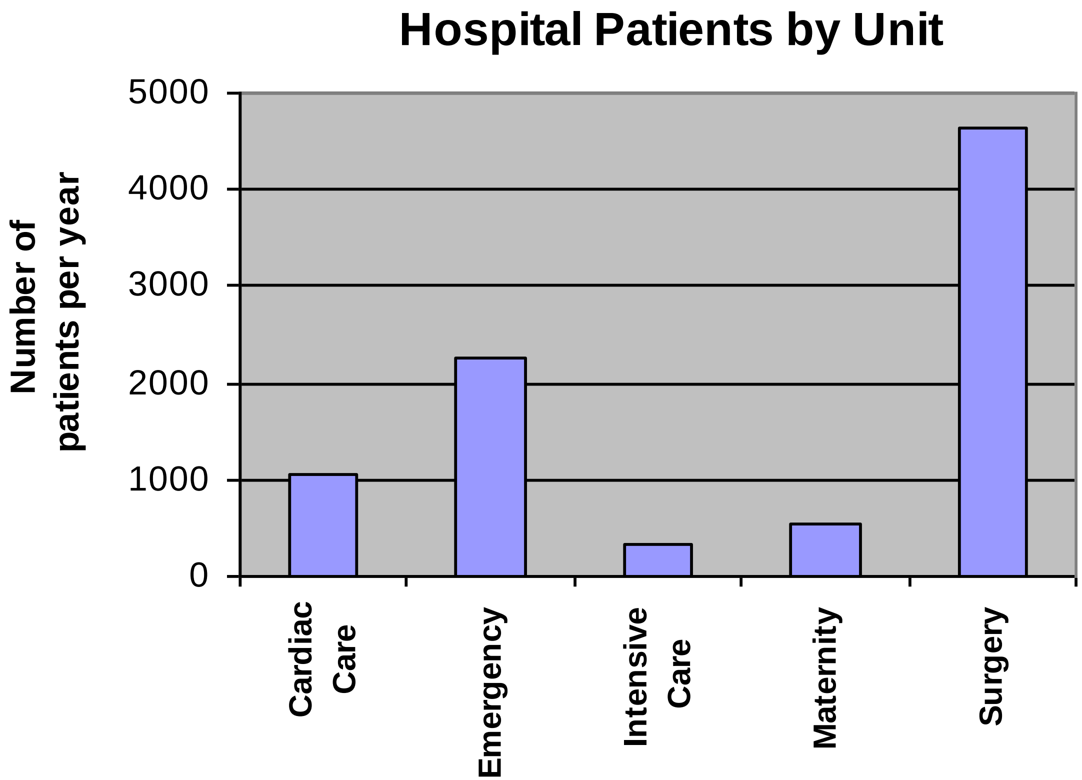
Bar and Pie Charts

- Bar charts and Pie charts are often used for qualitative (categorical) data

- Height of bar or size of pie slice shows the frequency or percentage for each category

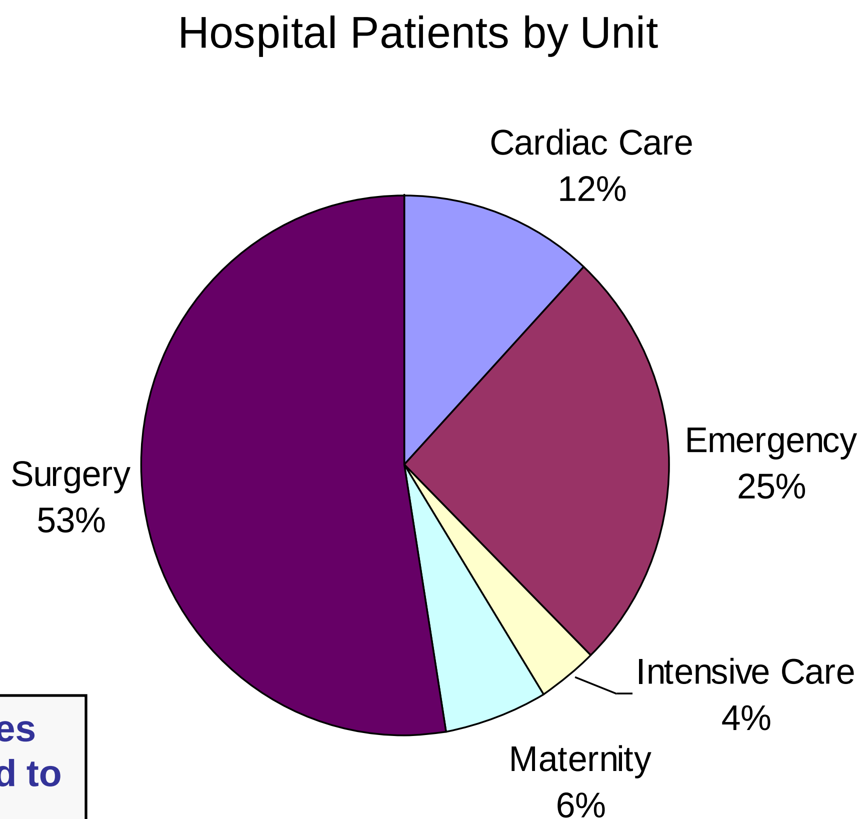
Bar Chart Example

Hospital Unit	Number of Patients
Cardiac Care	1,052
Emergency	2,245
Intensive Care	340
Maternity	552
Surgery	4,630



Pie Chart Example

Hospital Unit	Number of Patients	% of Total
Cardiac Care	1,052	11.93
Emergency	2,245	25.46
Intensive Care	340	3.86
Maternity	552	6.26
Surgery	4,630	52.50



(Percentages are rounded to the nearest percent)

Pareto Diagram

- Used to portray categorical data
- A bar chart, where categories are shown in descending order of frequency
- A cumulative polygon is often shown in the same graph
- Used to separate the "vital few" from the "trivial many"

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

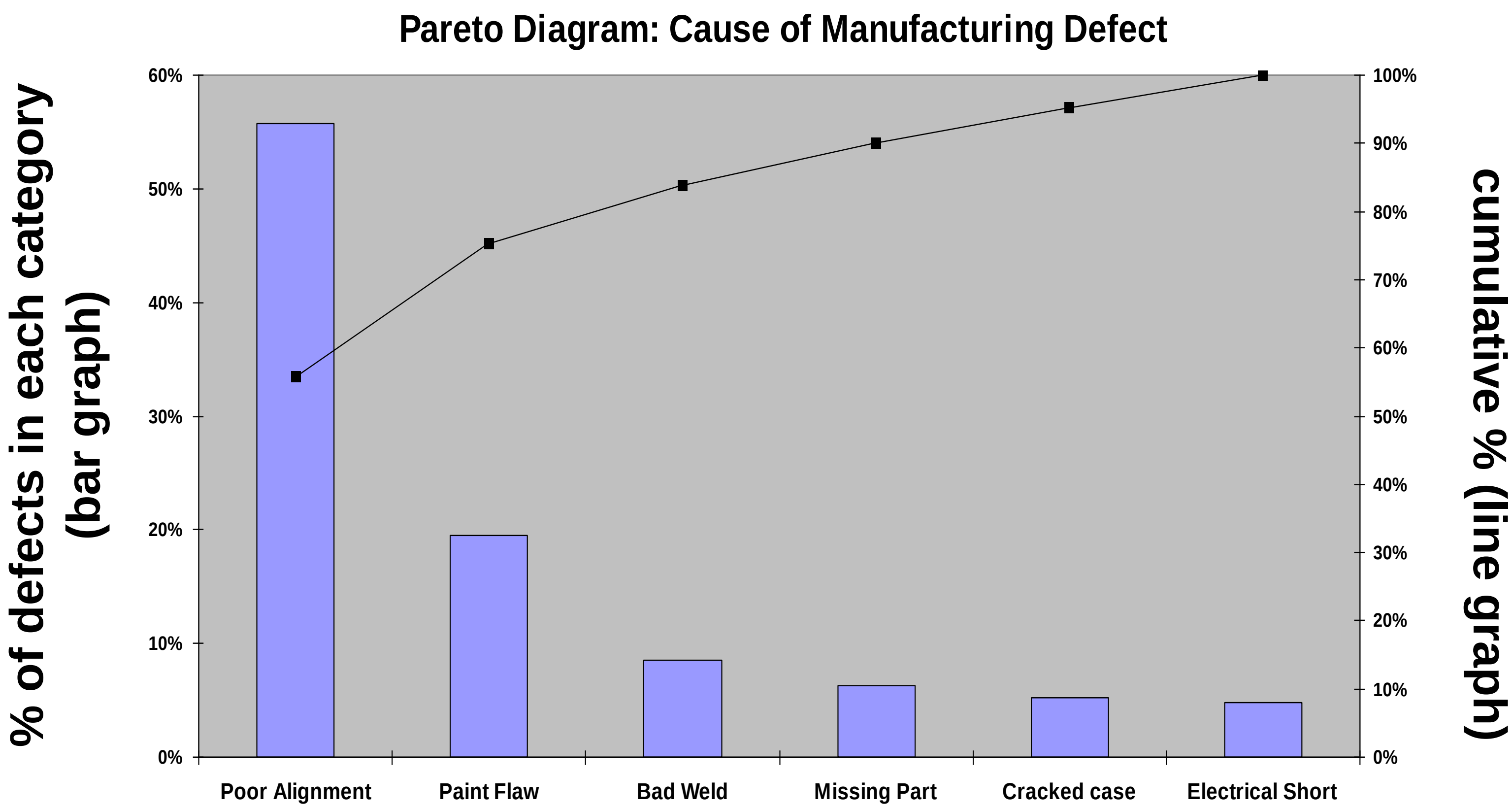
Step 2: Determine % in each category

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

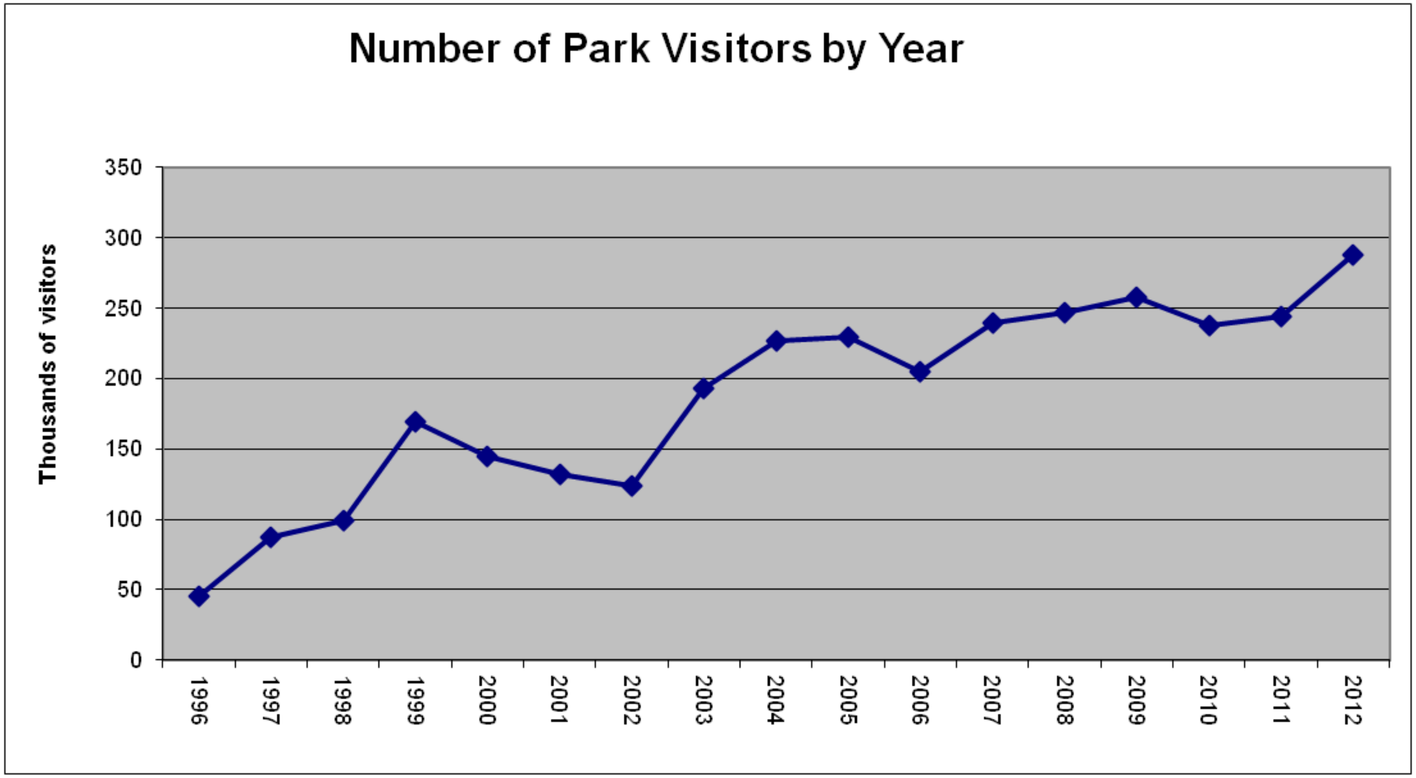


1.4

Graphs to Describe Time-Series Data

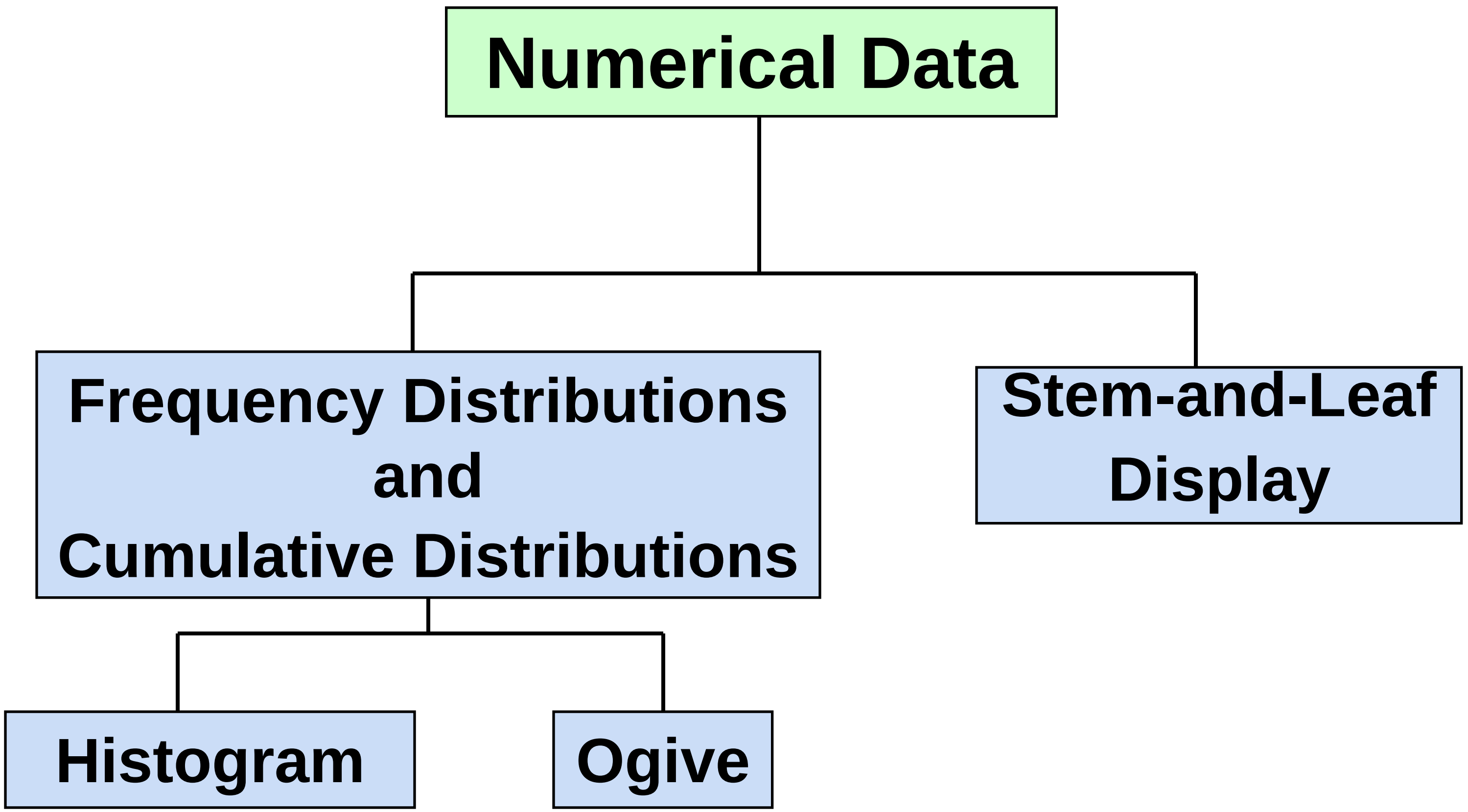
- A line chart (time-series plot) is used to show the values of a variable over time
- Time is measured on the horizontal axis
- The variable of interest is measured on the vertical axis

Line Chart Example



Graphs to Describe Numerical Variables

1.5



Frequency Distributions

What is a Frequency Distribution?

- A frequency distribution is a list or a table …
- containing class groupings (categories or ranges within which the data fall) ...
- and the corresponding frequencies with which data fall within each class or category

Why Use Frequency Distributions?

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

Class Intervals and Class Boundaries

- Each class grouping has the same width
- Determine the width of each interval by

$$w = \text{interval width} = \frac{\text{largest number} - \text{smallest number}}{\text{number of desired intervals}}$$

- Use at least 5 but no more than 15-20 intervals
- Intervals never overlap
- Round up the interval width to get desirable interval endpoints

Frequency Distribution Example

Example: A manufacturer of insulation randomly selects 20 winter days and records the daily high temperature

data:

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

Frequency Distribution Example

(continued)

- 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 interval width: **10 (46/5 then round up)**
- Determine interval boundaries: **10 but less than 20, 20 but less than 30, . . . , 60 but less than 70**
- Count observations & assign to classes

Frequency Distribution Example

(continued)

Data in ordered array:

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

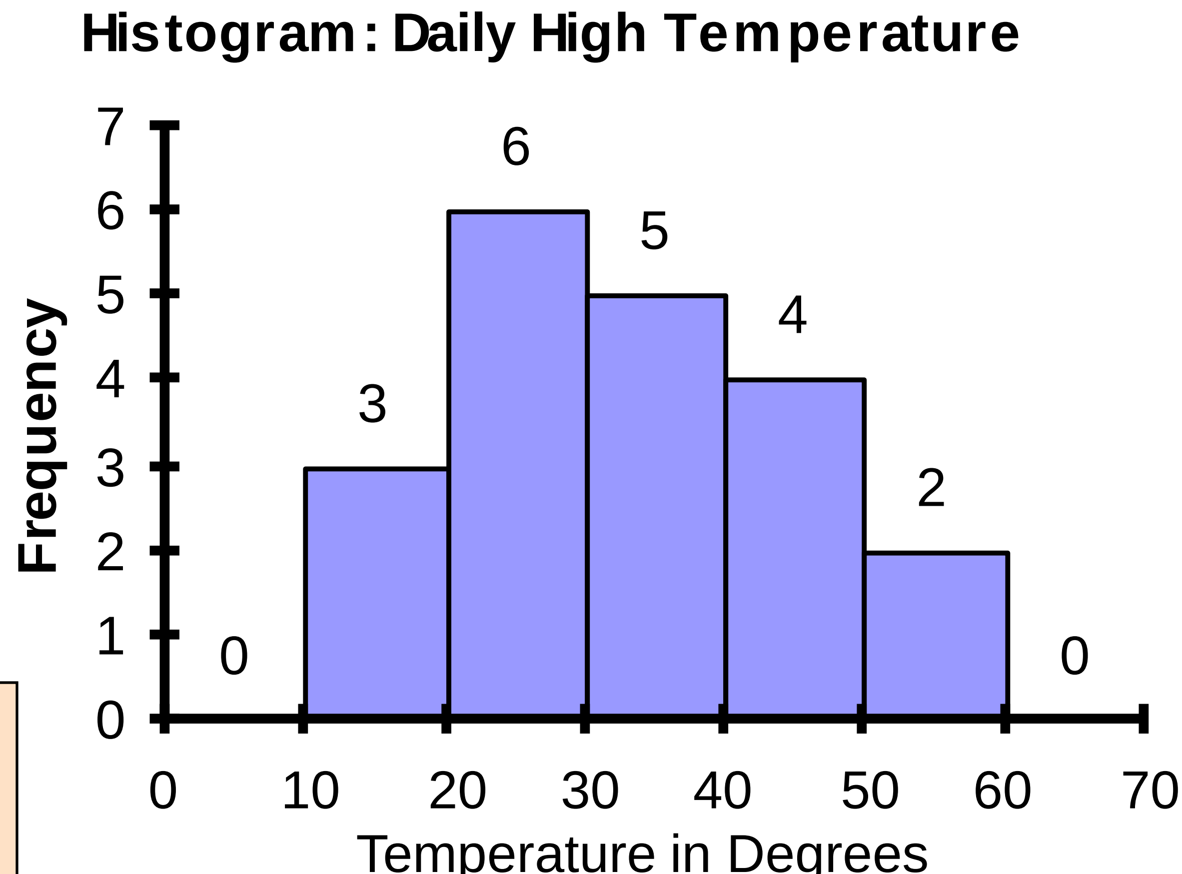
Interval	Frequency	Relative Frequency	Percentage
10 but less than 20	3	.15	15
20 but less than 30	6	.30	30
30 but less than 40	5	.25	25
40 but less than 50	4	.20	20
50 but less than 60	2	.10	10
Total	20	1.00	100

Histogram

- A graph of the data in a frequency distribution is called a **histogram**
- The **interval endpoints** are shown on the horizontal axis
- the vertical axis is either **frequency, relative frequency,** or **percentage**
- Bars of the appropriate heights are used to represent the number of observations within each class

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)

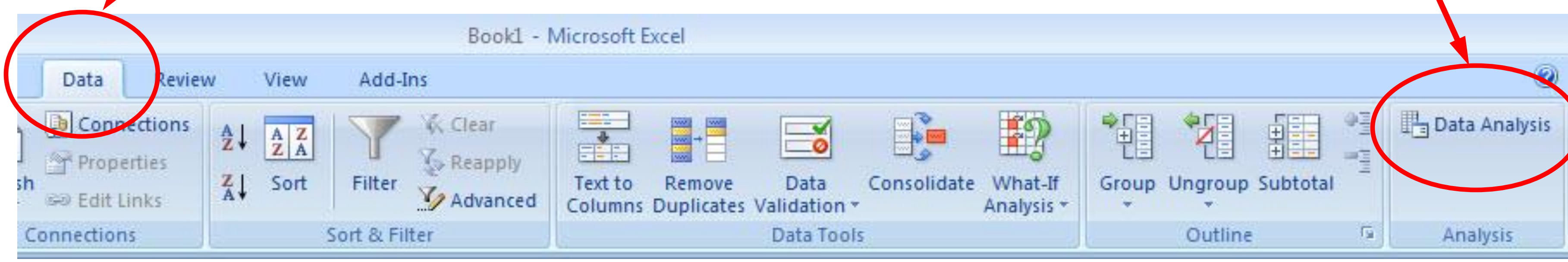
Histograms in Excel

1

Select Data Tab

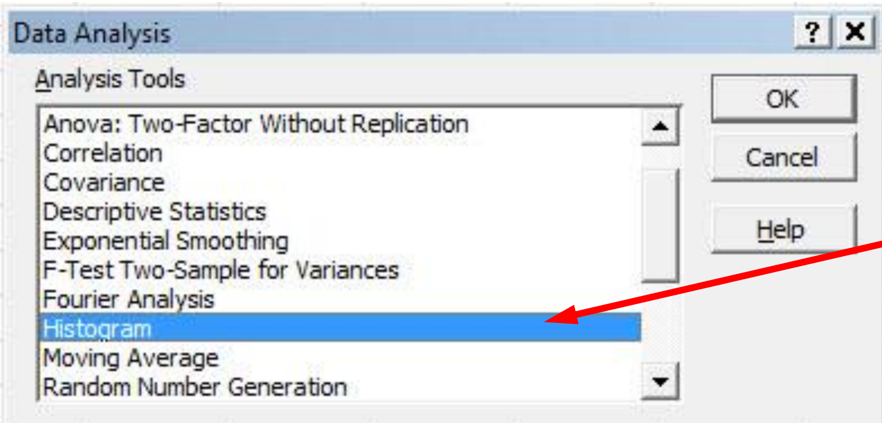
2

Click on Data Analysis



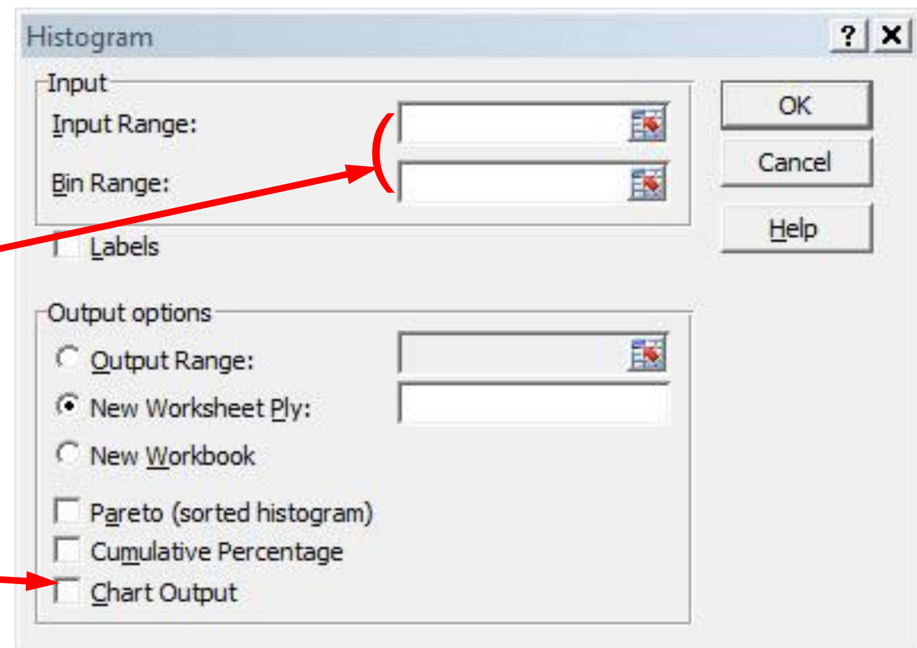
Histograms in Excel

(continued)



3

Choose Histogram



4

Input data range and bin range (bin range is a cell range containing the upper interval endpoints for each class grouping)

Select Chart Output and click "OK"

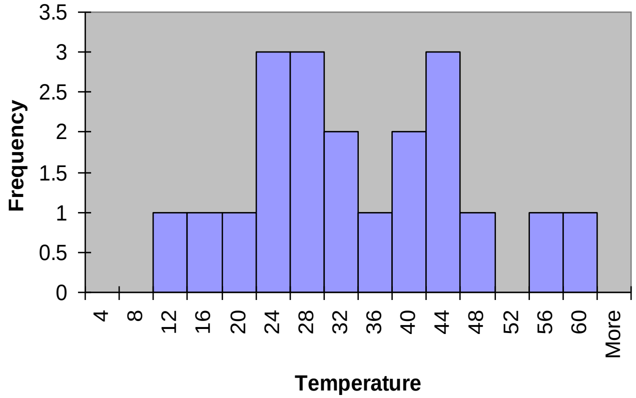
Questions for Grouping Data into Intervals

- 1. How wide should each interval be?
  (How many classes should be used?)

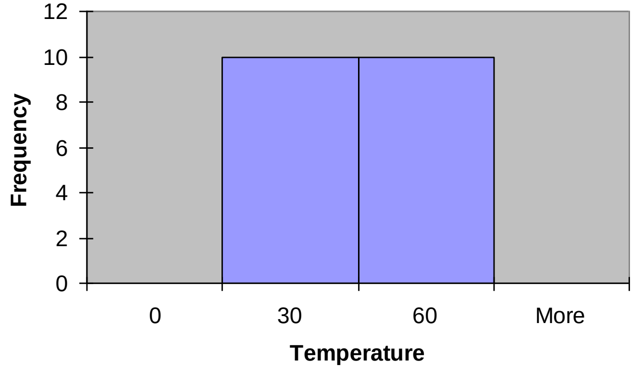
- 2. How should the endpoints of the intervals be determined?
  - Often answered by trial and error, subject to user judgment
  - The goal is to create a distribution that is neither too "jagged" nor too "blocky”
  - Goal is to appropriately show the pattern of variation in the data

How Many Class Intervals?

- **Many (Narrow class intervals)**
  - may yield a very jagged 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)

The Cumulative Frequency Distribuiton

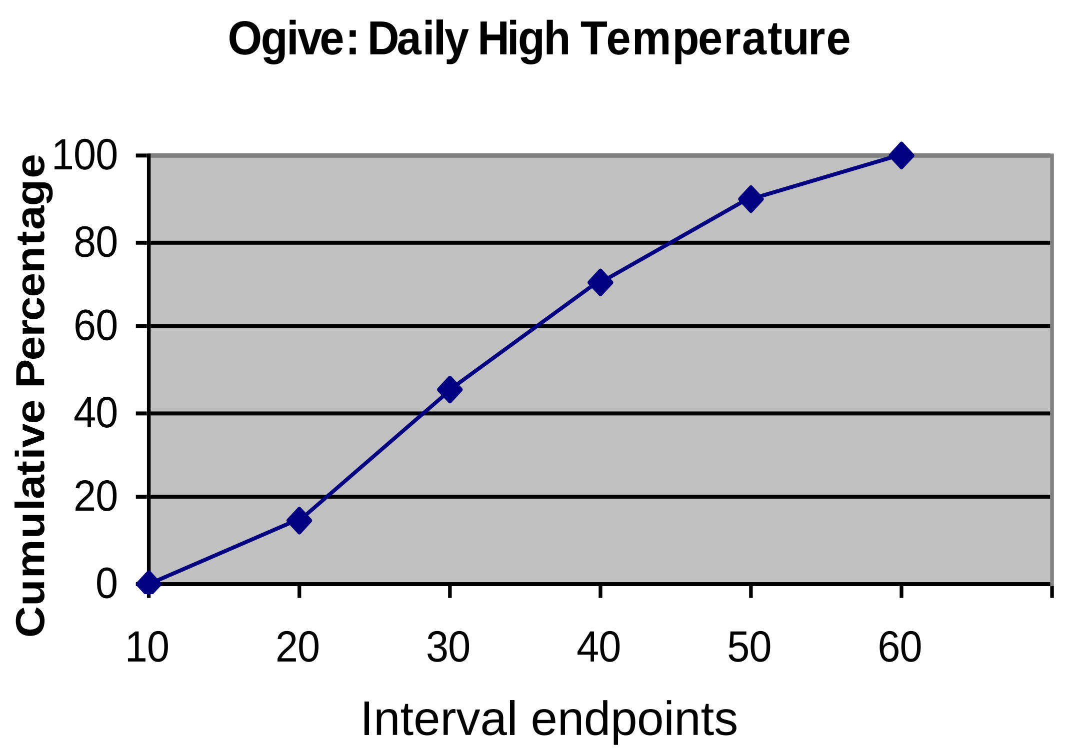
Data in ordered array:

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		

The Ogive
Graphing Cumulative Frequencies

Interval	Upper interval endpoint	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



Stem-and-Leaf Diagram

- A simple way to see distribution details in a data set

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

Example

Data in ordered array:

21, 24, 24, 26, 27, 27, 30, 32, 38, 41

- Here, use the 10's digit for the stem unit:
  - 21 is shown as
  - 38 is shown as

Stem	Leaf
2	1
3	8

Example

(continued)

Data in ordered array:

21, 24, 24, 26, 27, 27, 30, 32, 38, 41

- Completed stem-and-leaf diagram:

Stem	Leaves
2	1 4 4 6 7 7
3	0 2 8
4	1

Using other stem units

- Using the 100's digit as the stem:
  - Round off the 10's digit to form the leaves
    - 613 would become →
    - 776 would become →
    - . . .
    - 1224 becomes →

Stem	Leaf
6	1
7	8
12	2

Using other stem units

(continued)

- Using the 100's digit as the stem:
  - The completed stem-and-leaf display:

Data:

613, 632, 658, 717, 722, 750, 776, 827, 841, 859, 863, 891, 894, 906, 928, 933, 955, 982, 1034, 1047,1056, 1140, 1169, 1224

Stem	Leaves
6	1 3 6
7	2 2 5 8
8	3 4 6 6 9 9
9	1 3 3 6 8
10	3 5 6
11	4 7
12	2

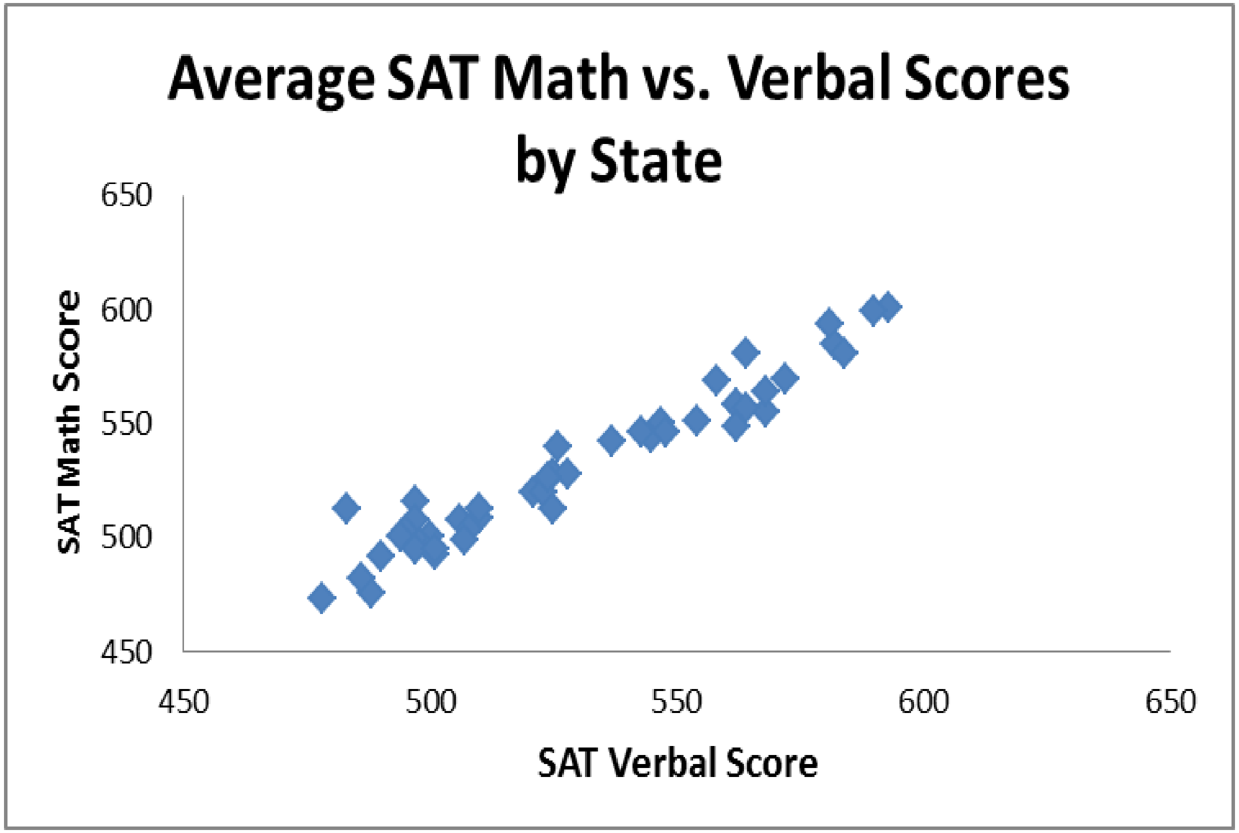
Scatter Diagrams

- Scatter Diagrams are used for paired observations taken from two numerical variables

- The Scatter Diagram:
  - one variable is measured on the vertical axis and the other variable is measured on the horizontal axis

Scatter Diagram Example

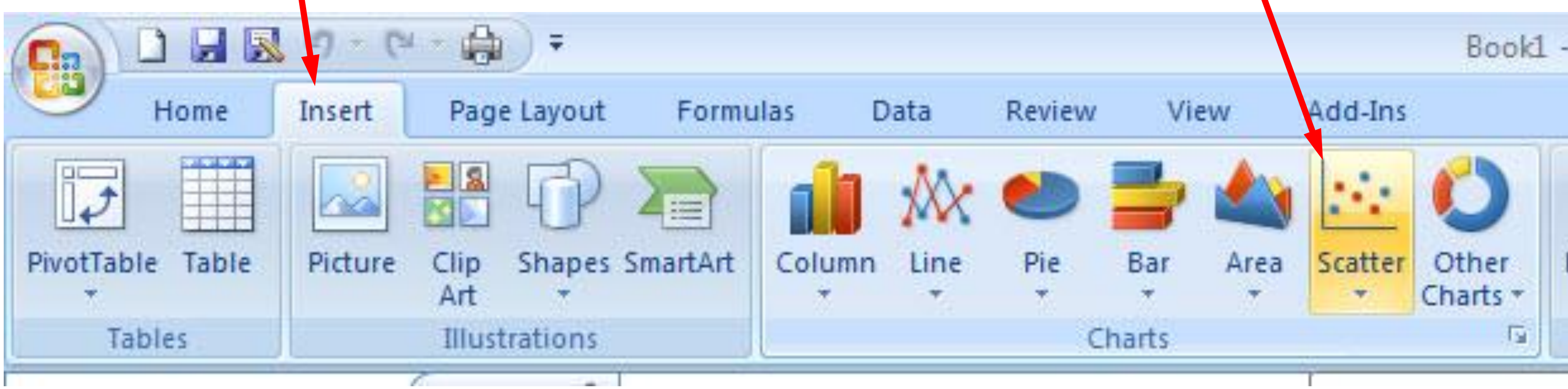
Average SAT scores by state: 1998		
	Verbal	Math
Alabama	562	558
Alaska	521	520
Arizona	525	528
Arkansas	568	555
California	497	516
Colorado	537	542
Connecticut	510	509
Delaware	501	493
D.C.	488	476
Florida	500	501
Georgia	486	482
Hawaii	483	513
. . .		
W.Va.	525	513
Wis.	581	594
Wyo.	548	546



Scatter Diagrams in Excel

1 Select the Insert tab

2 Select Scatter type from the Charts section



3 When prompted, enter the data range, desired legend, and desired destination to complete the scatter diagram

1.6

Data Presentation Errors

Goals for effective data presentation:

- Present data to display essential information
- Communicate complex ideas clearly and accurately
- Avoid distortion that might convey the wrong message

Data Presentation Errors

(continued)

- Unequal histogram interval widths
- Compressing or distorting the vertical axis
- Providing no zero point on the vertical axis
- Failing to provide a relative basis in comparing data between groups

Chapter Summary

- Reviewed incomplete information in decision making
- Introduced key definitions:
  - Population vs. Sample
  - Parameter vs. Statistic
  - Descriptive vs. Inferential statistics
- Described random sampling
- Examined the decision making process

Chapter Summary

(continued)

- Reviewed types of data and measurement levels
- Data in raw form are usually not easy to use for decision making -- Some type of organization is needed:
  - Table
  - Graph
- Techniques reviewed in this chapter:
  - Frequency distribution
  - Cross tables
  - Bar chart
  - Pie chart
  - Pareto diagram
  - Line chart
  - Frequency distribution
  - Histogram and ogive
  - Stem-and-leaf display
  - Scatter plot