Regents Park Publishers

Data Analytics

T1LM 4

Data Representation

Learning Module Goals

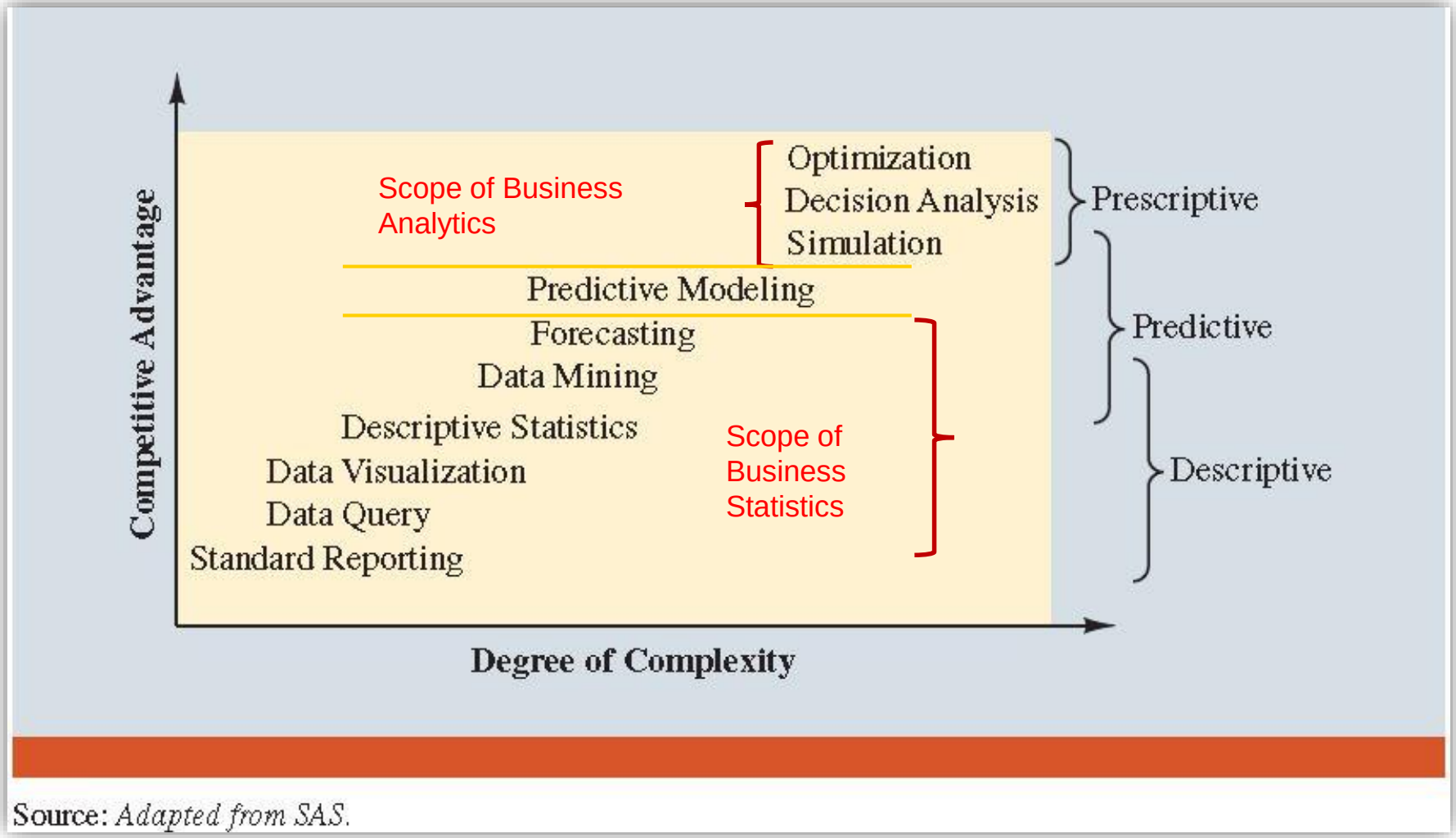
After completing this Learning Module, you should be able to:

- Describe key data collection methods
- Know key definitions:
  - ♦ Population vs. Sample
  - ♦ Primary vs. Secondary data types
  - ♦ Qualitative vs. Qualitative data
  - ♦ Time Series vs. Cross-Sectional data
- Explain the difference between descriptive and inferential statistics
- Describe different sampling methods

Tools of Business Statistics

- **Descriptive statistics**
  - Collecting, presenting, and describing data
- **Inferential statistics**
  - Drawing conclusions and/or making decisions concerning a population based only on sample data

Tools of Business Statistics



Source: *Adapted from SAS.*

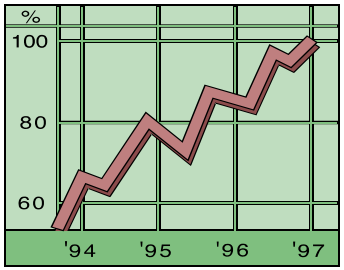
Market Research Process

Burke and Associates

Descriptive Statistics

- **Collect data**
  - e.g. Survey, Observation, Experiments

- **Present data**
  - e.g. Charts and graphs



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

Data Sources

Primary

Data Collection

Observation

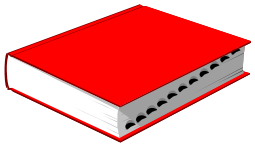
Survey

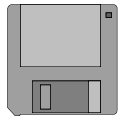
Experimentation

Secondary

Data Compilation

Print or Electronic and WEB





Populations and Samples

- A **Population** is the set of all items or individuals of interest

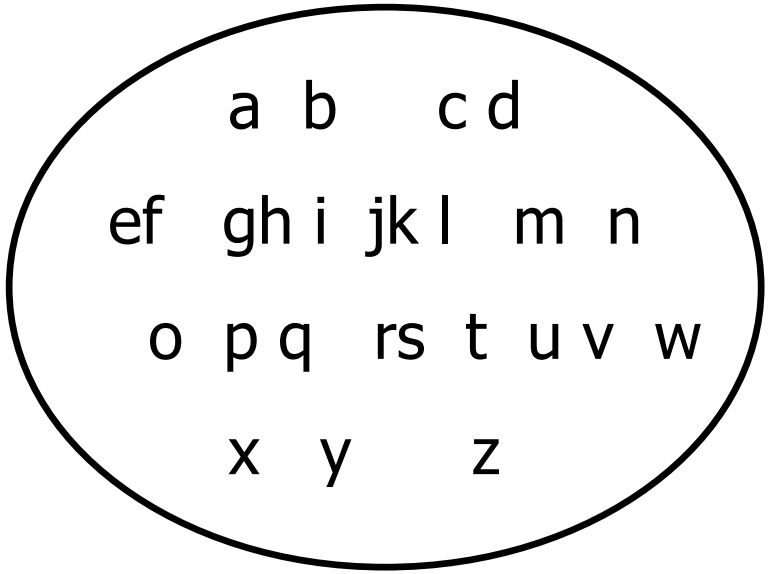
Examples:	All likely voters in the next election
	All parts produced today
	All sales receipts for November

- A **Sample** is a subset of the population

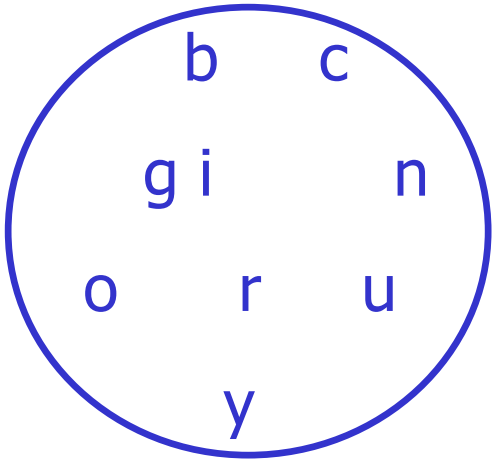
Examples:	1000 voters selected at random for interview
	A few parts selected for destructive testing
	Every 100th receipt selected for audit

Population vs. Sample

Population



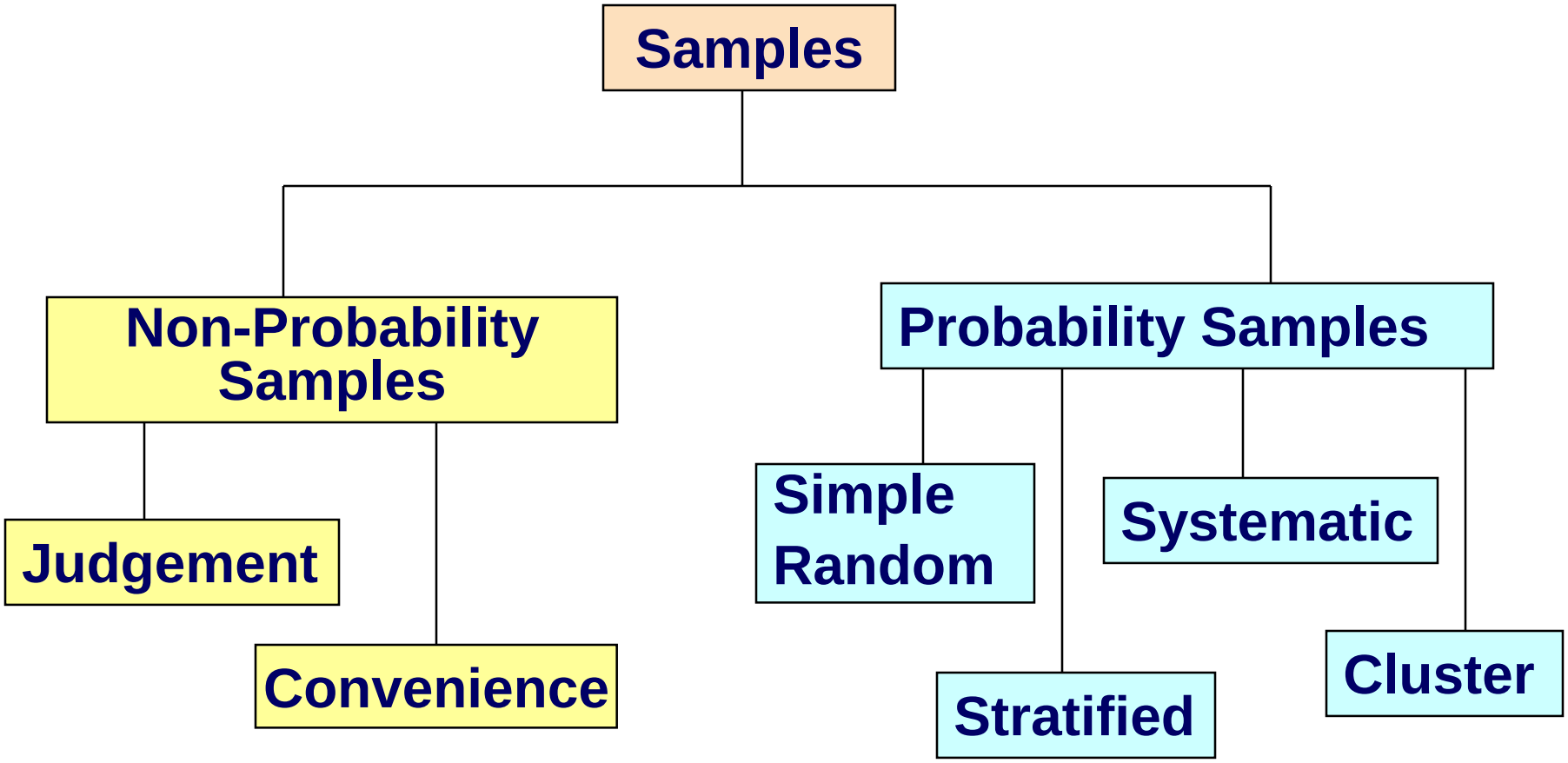
Sample



Why Sample?

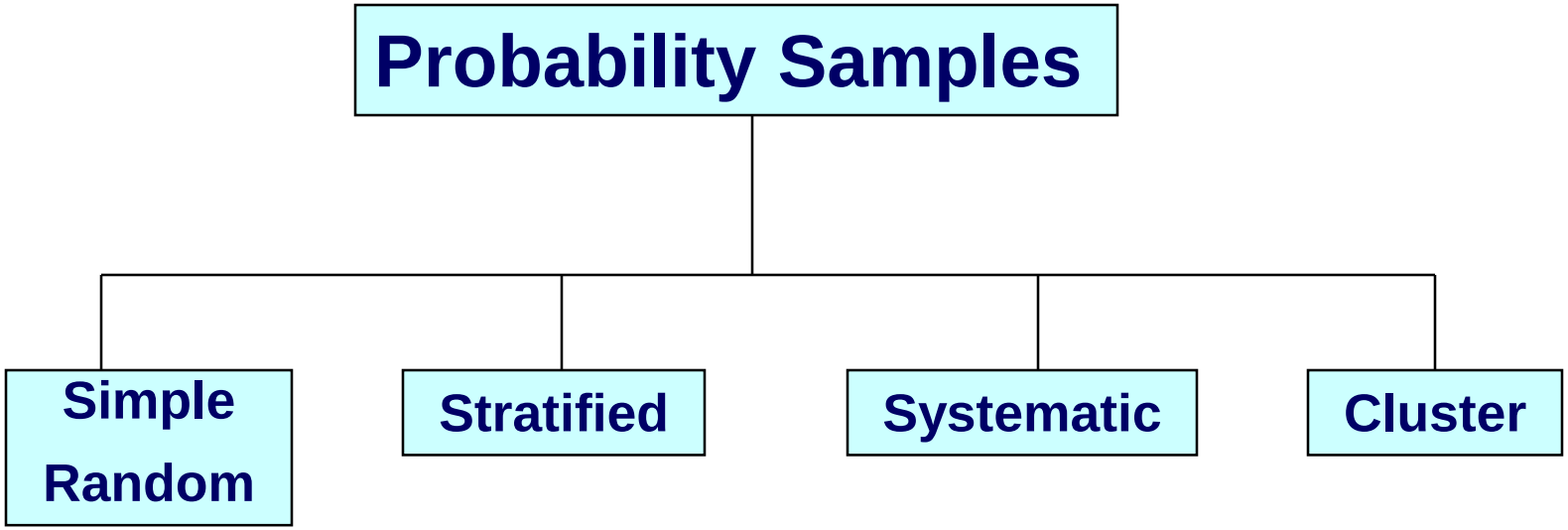
- Less time consuming than a census
- Less costly to administer than a census
- It is possible to obtain statistical results of a sufficiently high precision based on samples.

Sampling Techniques



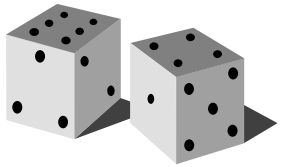
Statistical Sampling

- Items of the sample are chosen based on known or calculable probabilities



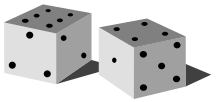
Simple Random Samples

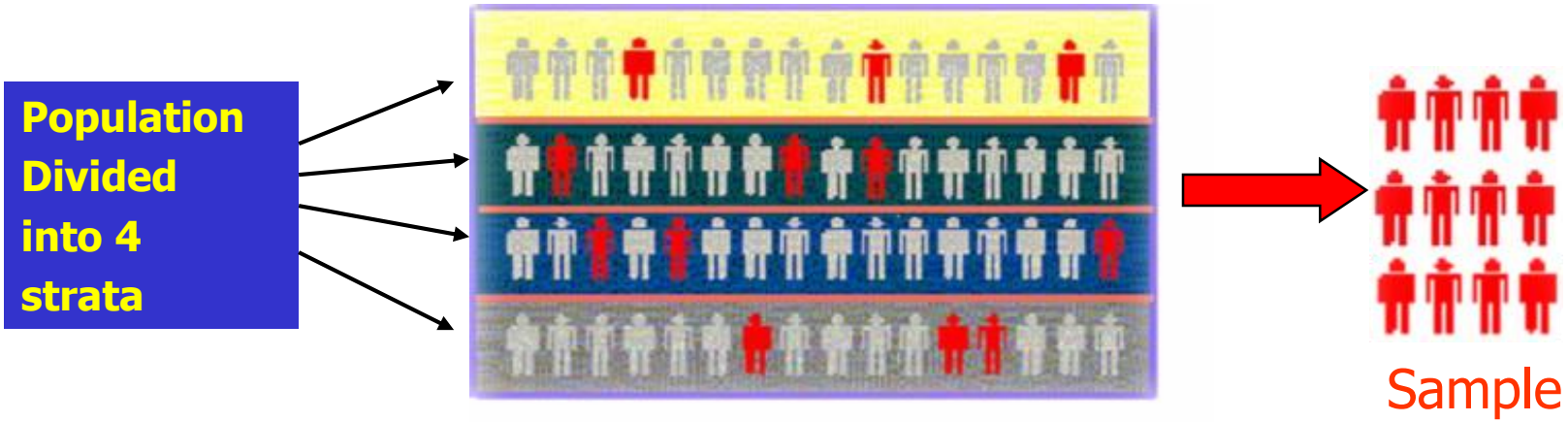
- Every individual or item from the population has an equal chance of being selected
- Selection may be with replacement or without replacement
- Samples can be obtained from a table of random numbers or computer random number generators



Stratified Samples

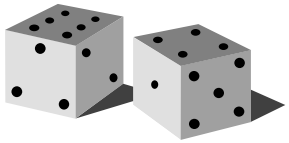
- Population divided into subgroups (called *strata*) according to some common characteristic
- Simple random sample selected from each subgroup
- Samples from subgroups are combined into one





Systematic Samples

- Decide on sample size: n
- Divide frame of N individuals into groups of k individuals: k=N/n
- Randomly select one individual from the 1st group



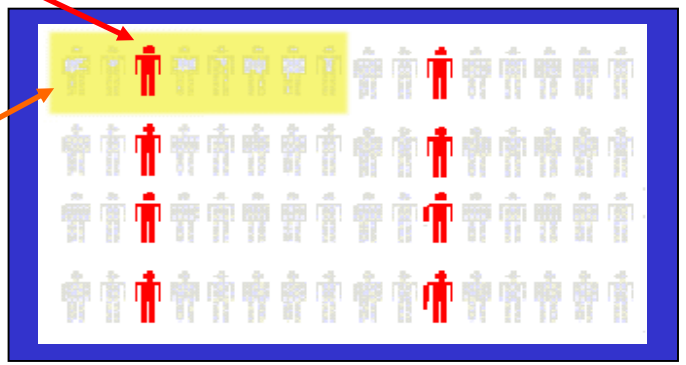
- Select every kth individual thereafter

N = 64

n = 8

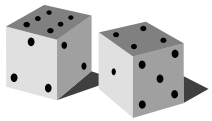
k = 8

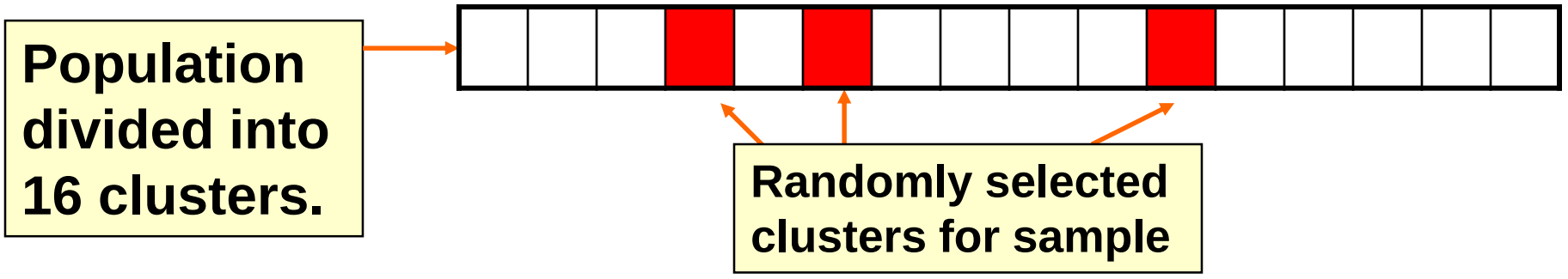
First Group



Cluster Samples

- Population is divided into several “clusters,” each representative of the population
- A simple random sample of clusters is selected
  - All items in the selected clusters can be used, or items can be chosen from a cluster using another probability sampling technique





Key Definitions

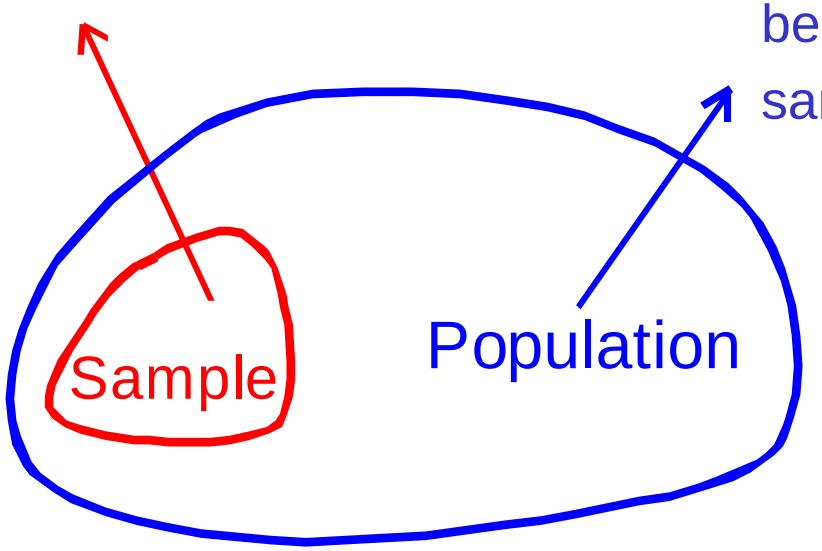
- A **population** is the entire collection of things under consideration
  - A **parameter** is a summary measure computed to describe a characteristic of the population
- A **sample** is a portion of the population selected for analysis
  - A **statistic** is a summary measure computed to describe a characteristic of the sample

Inferential Statistics

- Making statements about a population by examining sample results

Sample statistics → Population parameters

(known) **Inference** (unknown, but can be estimated from sample evidence)

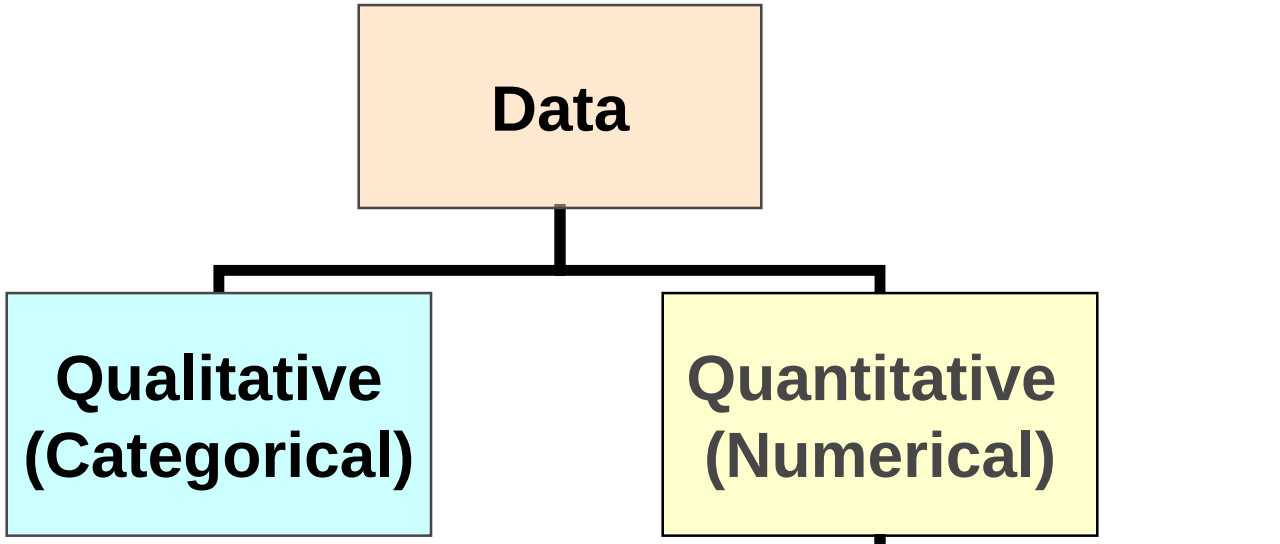


Inferential Statistics

Drawing conclusions and/or making decisions concerning a population based on sample results.

- **Estimation**
  - e.g.: Estimate the population mean weight using the sample mean weight
- **Hypothesis Testing**
  - e.g.: Use sample evidence to test the claim that the population mean weight is 120 pounds

Data Types



Examples:

- Marital Status
- Political Party
- Eye Color

(Defined categories)

Examples:

- Number of Children
- Defects per hour

(Counted items)

Examples:

- Weight
- Voltage

(Measured characteristics)

Data Types

- **Time Series Data**
  - Ordered data values observed over time
- **Cross Section Data**
  - Data values observed at a fixed point in time

Data Types

	Sales (in $1000's)			
	2003	2004	2005	2006
Atlanta	435	460	475	490
Boston	320	345	375	395
Cleveland	405	390	410	395
Denver	260	270	285	280

Time Series Data

Cross Section Data

Frequency Distributions

What is a Frequency Distribution?

- A frequency distribution is a list or a table …
- containing the values of a variable (or a set of ranges within which the data falls) ...
- and the corresponding frequencies with which each value occurs (or frequencies with which data falls within each range)

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

Frequency Distribution: Discrete Data

- Discrete data: possible values are countable

Example: An advertiser asks 200 customers how many days per week they read the daily newspaper.

Number of days read	Frequency
0	44
1	24
2	18
3	16
4	20
5	22
6	26
7	30
Total	**200**

Relative Frequency

Relative Frequency: What proportion is in each category?

Number of days read	Frequency	Relative Frequency
0	44	.22
1	24	.12
2	18	.09
3	16	.08
4	20	.10
5	22	.11
6	26	.13
7	30	.15
Total	**200**	**1.00**

$$\frac{44}{200} = .22$$

22% of the people in the sample report that they read the newspaper 0 days per week

Frequency Distribution: Continuous Data

- Continuous Data: may take on any value in some interval

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

(Temperature is a continuous variable because it could be measured to any degree of precision desired)

Grouping Data by Classes

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 20)
- Compute class width: 10 (46/5 then round off)
- Determine class boundaries:10, 20, 30, 40, 50
- Compute class midpoints: 15, 25, 35, 45, 55
- 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

Frequency Distribution

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

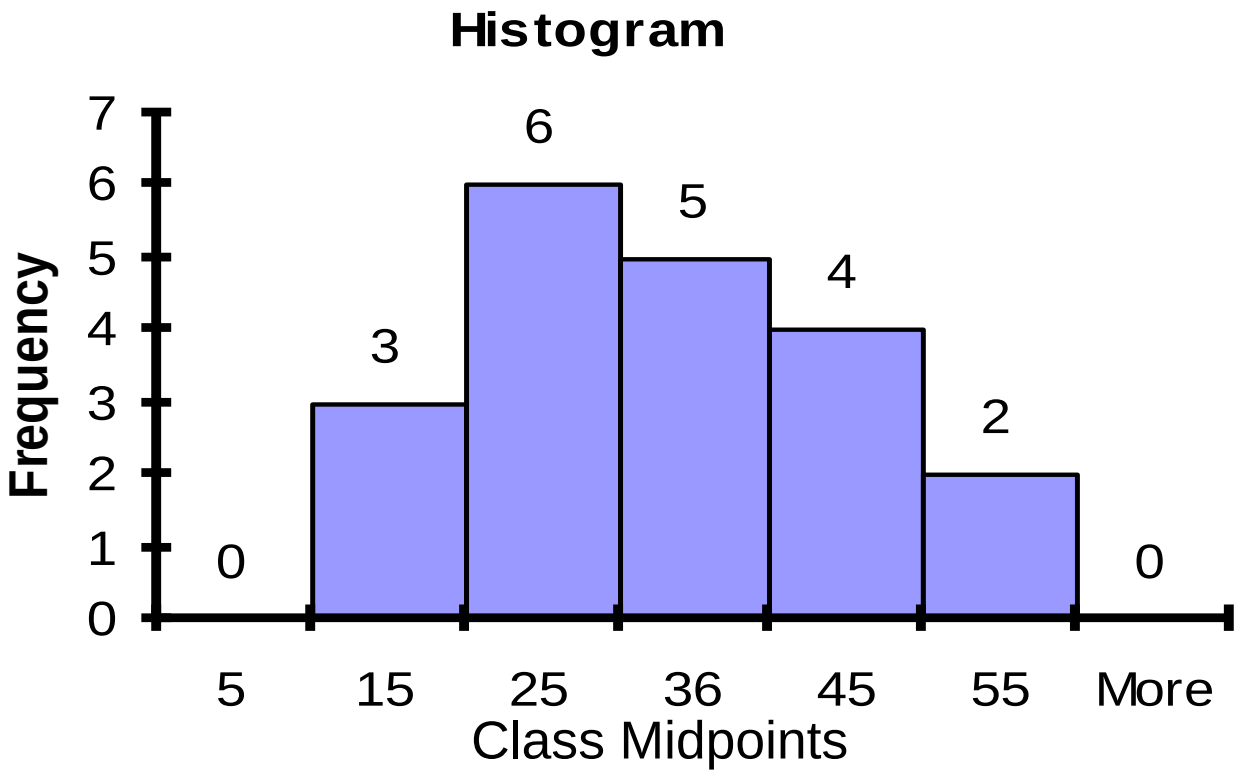
Histograms

- The **classes** or **intervals** are shown on the horizontal axis
- **frequency** is measured on the vertical axis
- Bars of the appropriate heights can be used to represent the number of observations within each class
- Such a graph is called a **histogram**

Histogram Example

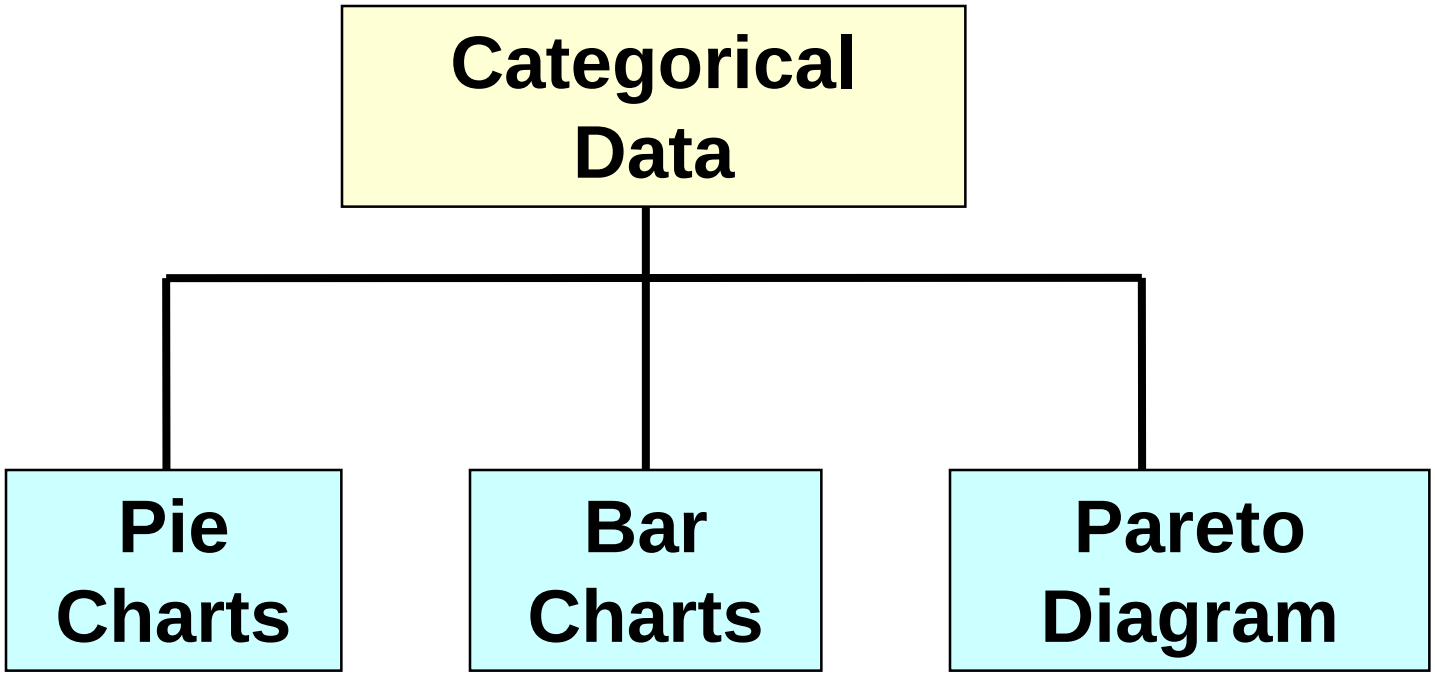
Data in ordered array:

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



No gaps between bars, since continuous data

Graphing Categorical Data



Bar and Pie Charts

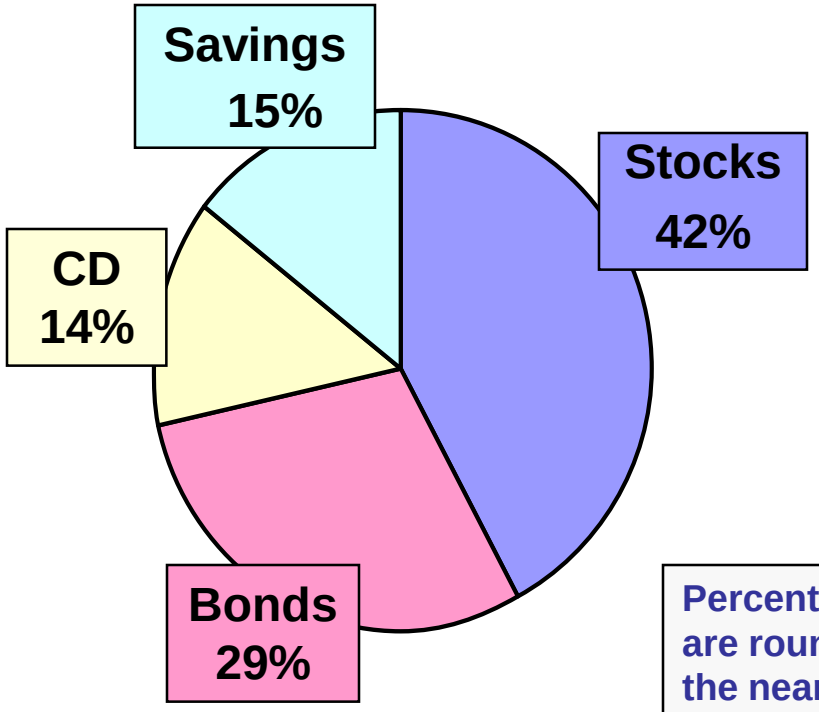
- Bar charts and Pie charts are often used for qualitative (category) data
- Height of bar or size of pie slice shows the frequency or percentage for each category

Pie Chart Example

Current Investment Portfolio

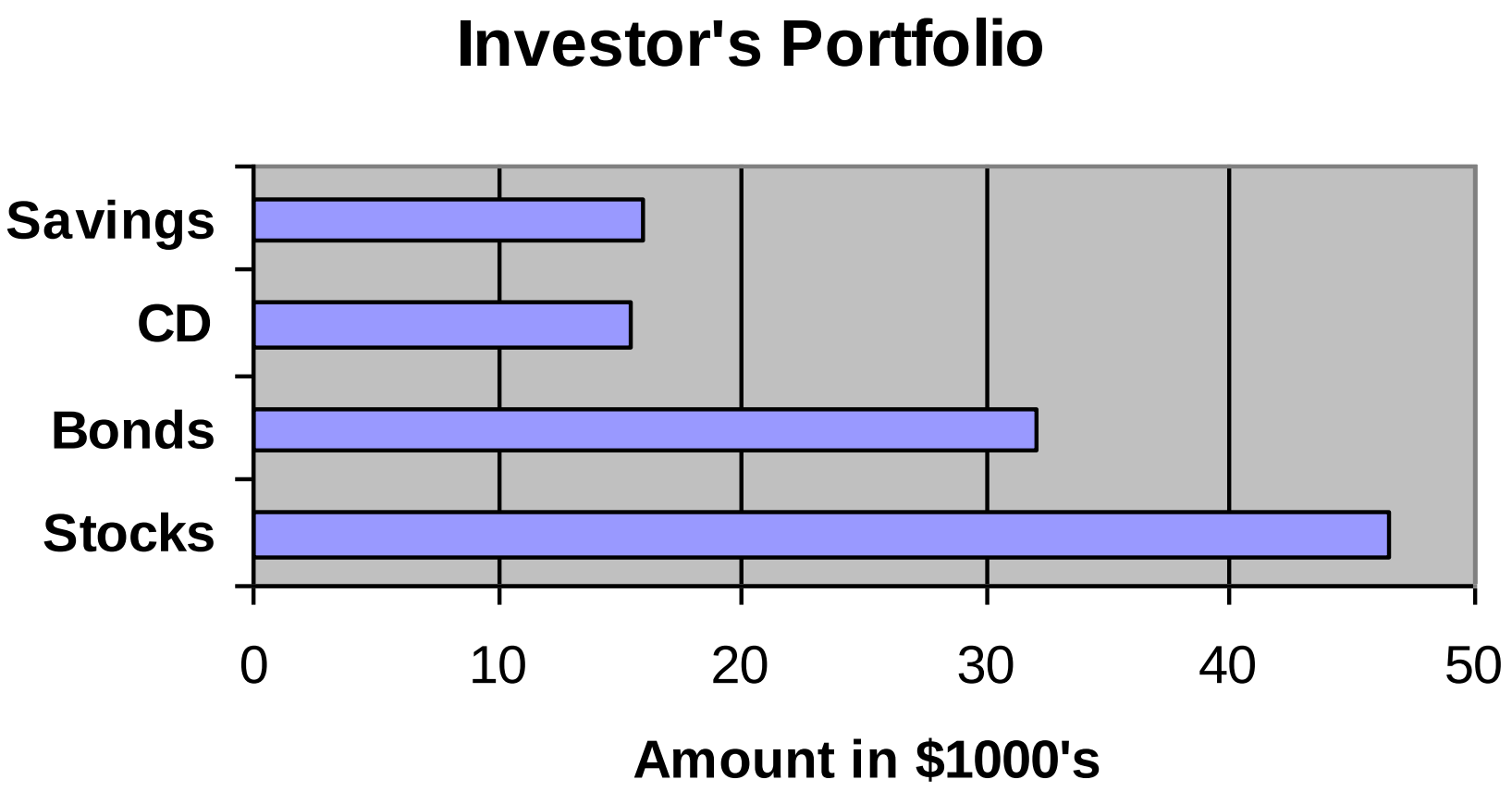
Investment Type	Amount (in thousands $)	Percentage
Stocks	46.5	42.27
Bonds	32.0	29.09
CD	15.5	14.09
Savings	16.0	14.55
Total	**110**	**100**

(Variables are Qualitative)

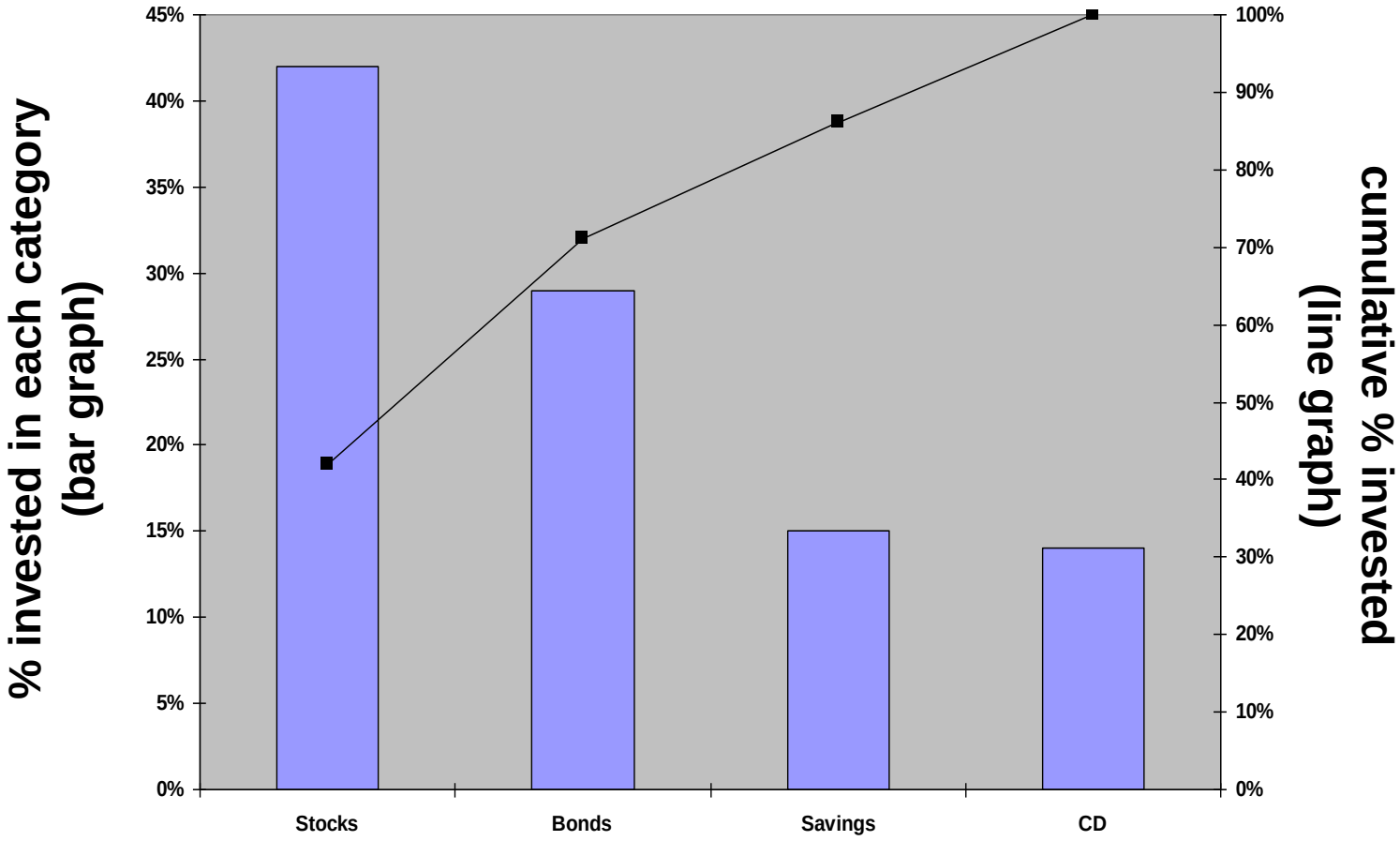


Percentages are rounded to the nearest percent

Bar Chart Example

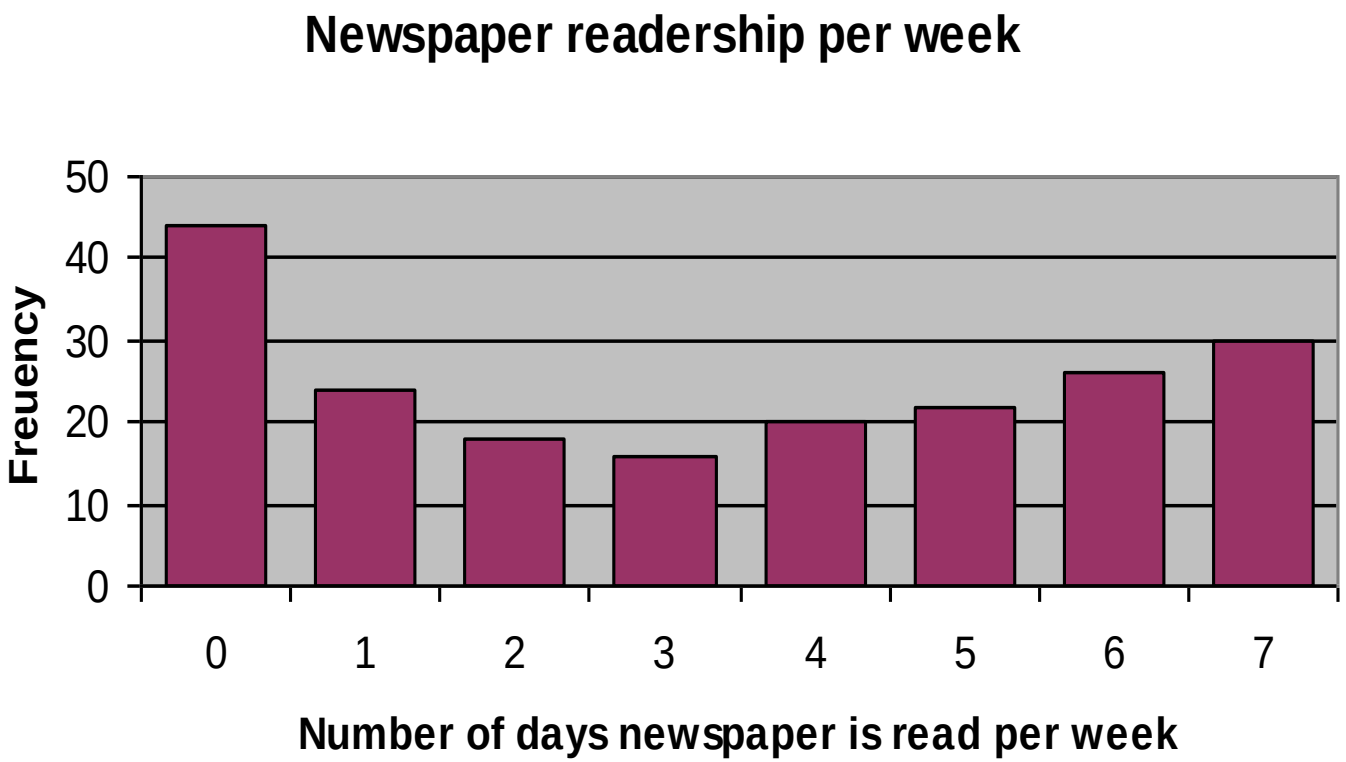


Pareto Diagram Example



Bar Chart Example

Number of days read	Frequency
0	44
1	24
2	18
3	16
4	20
5	22
6	26
7	30
Total	**200**



Tabulating and Graphing Multivariate Categorical Data

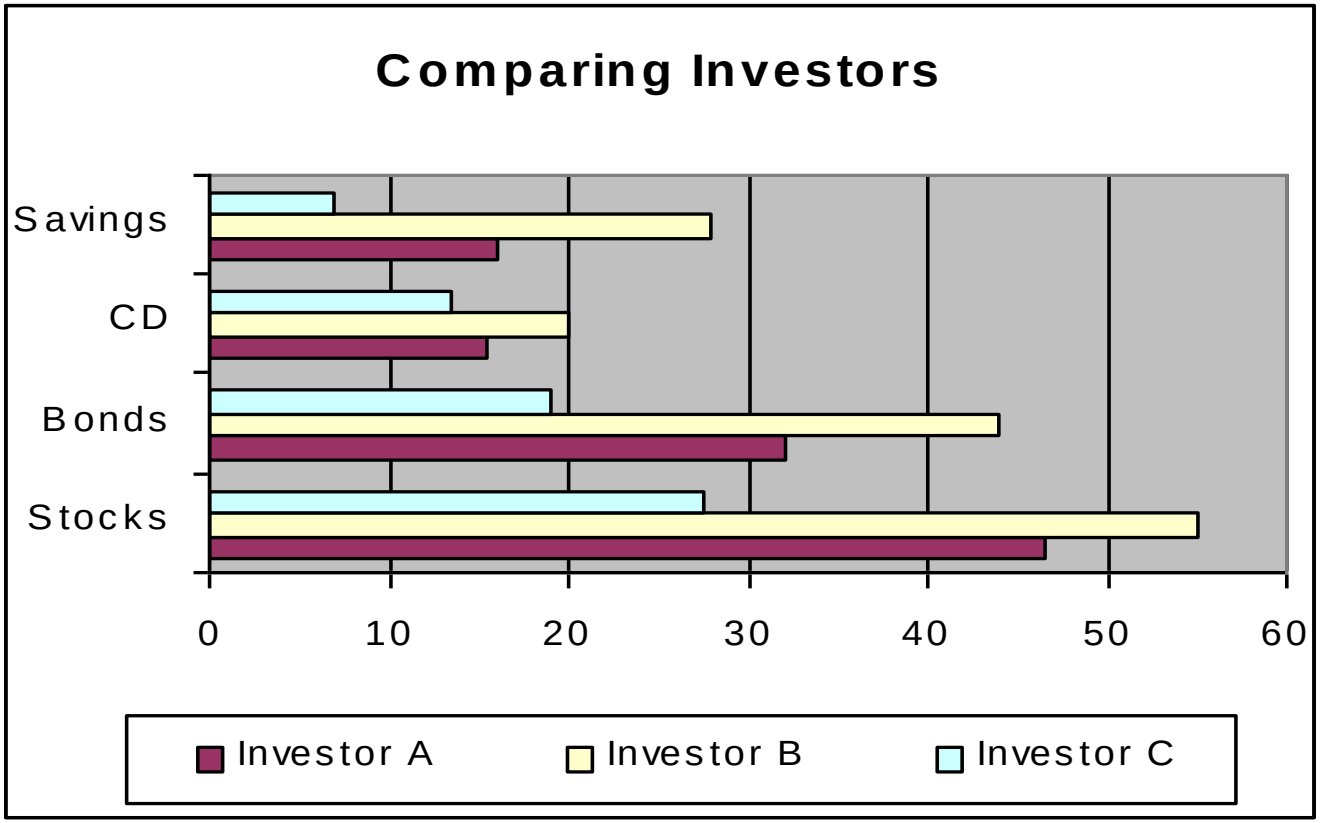
- Investment in thousands of dollars

Investment Category	Investor A	Investor B	Investor C	Total
Stocks	46.5	55	27.5	**129**
Bonds	32.0	44	19.0	**95**
CD	15.5	20	13.5	**49**
Savings	16.0	28	7.0	**51**
Total	**110.0**	**147**	**67.0**	**324**

Tabulating and Graphing Multivariate Categorical Data

(continued)

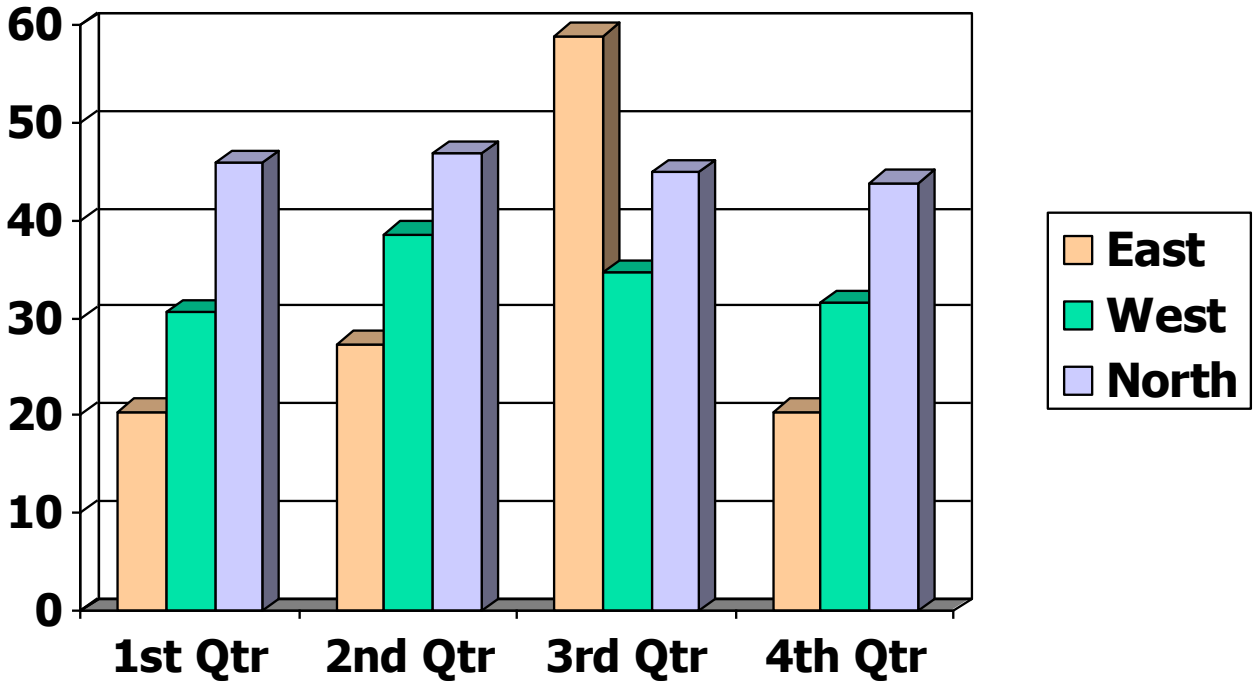
- Side by side charts



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

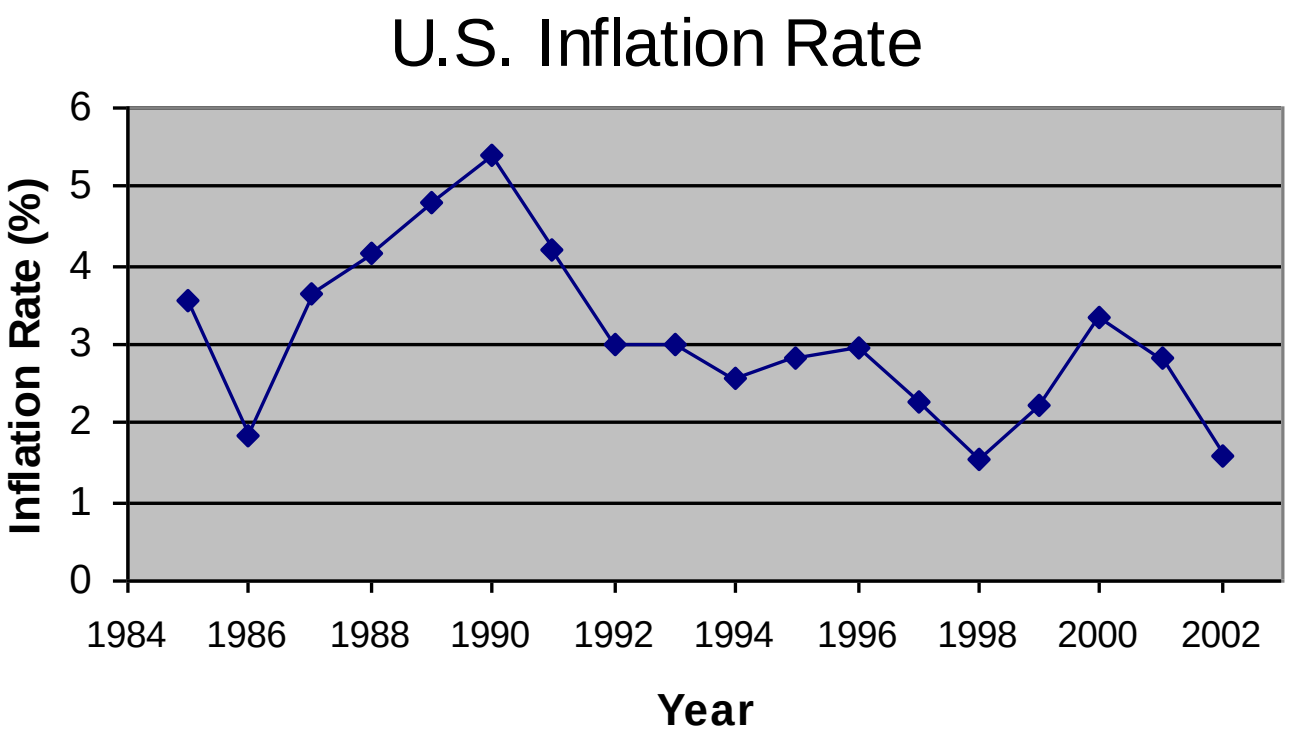


Line Charts and Scatter Diagrams

- Line charts show values of one variable vs. time
  - Time is traditionally shown on the horizontal axis
- Scatter Diagrams show points for bivariate data
  - one variable is measured on the vertical axis and the other variable is measured on the horizontal axis

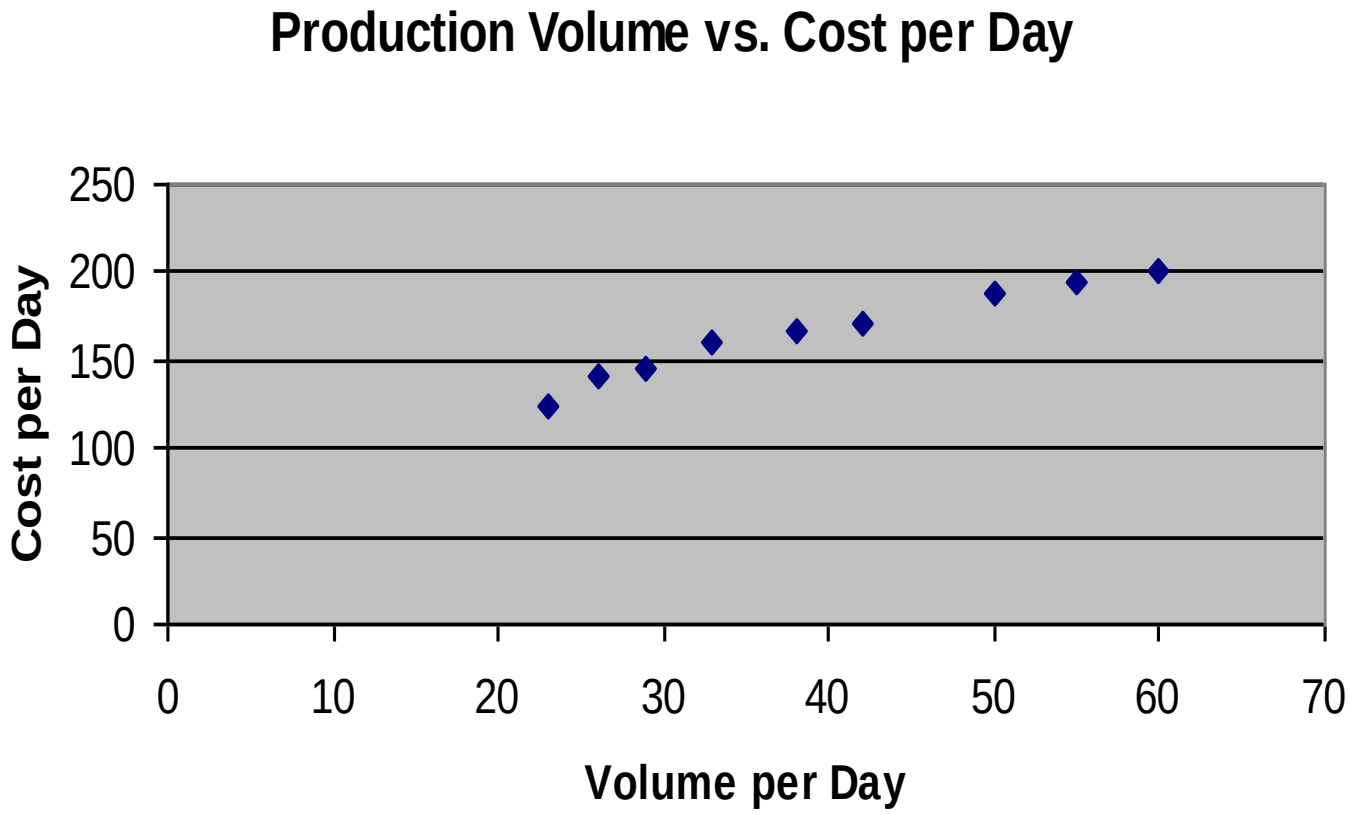
Line Chart Example

Year	Inflation Rate
1985	3.56
1986	1.86
1987	3.65
1988	4.14
1989	4.82
1990	5.40
1991	4.21
1992	3.01
1993	2.99
1994	2.56
1995	2.83
1996	2.95
1997	2.29
1998	1.56
1999	2.21
2000	3.36
2001	2.85
2002	1.58



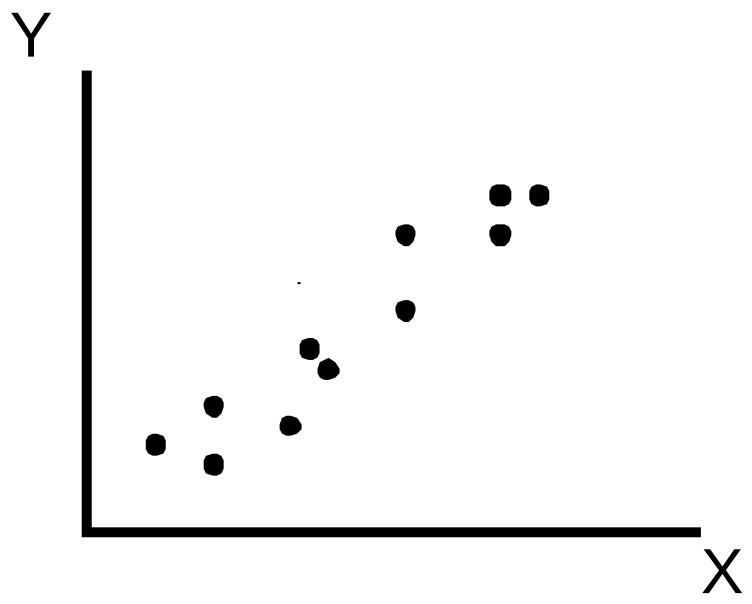
Scatter Diagram Example

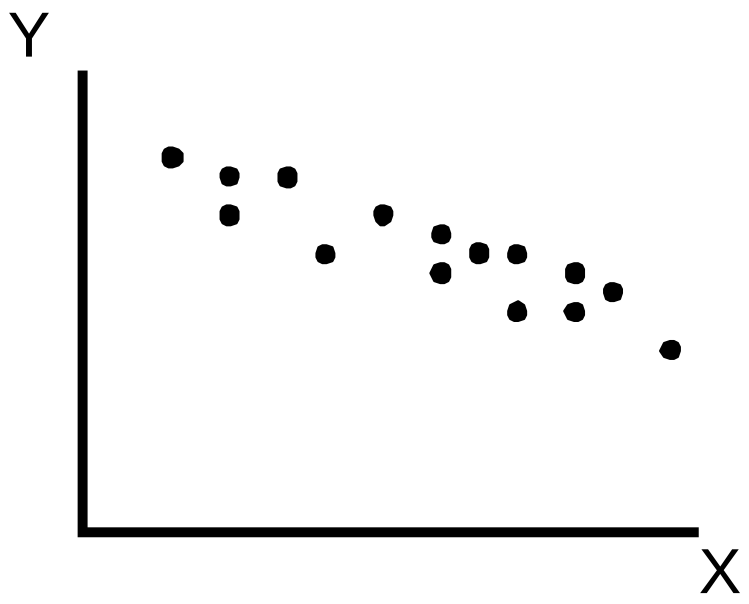
Volume per day	Cost per day
23	125
26	140
29	146
33	160
38	167
42	170
50	188
55	195
60	200



Types of Relationships

- Linear Relationships





Learning Module Summary

- Reviewed key data collection methods
- Introduced key definitions:
  - ♦ Population vs. Sample
  - ♦ Qualitative vs. Qualitative data
  - ♦ Primary vs. Secondary data types
  - ♦ Time Series vs. Cross-Sectional data
- Examined descriptive vs. inferential statistics
- Described different sampling techniques
- Reviewed data types and measurement levels

Data Collections Examples

(continued)

License Plate Reader

Smart TV

Facial Recognition

Facebook

Google Tracking

Stores

Palantir

Alexa

Regents Park Publishers

T1LM4

End