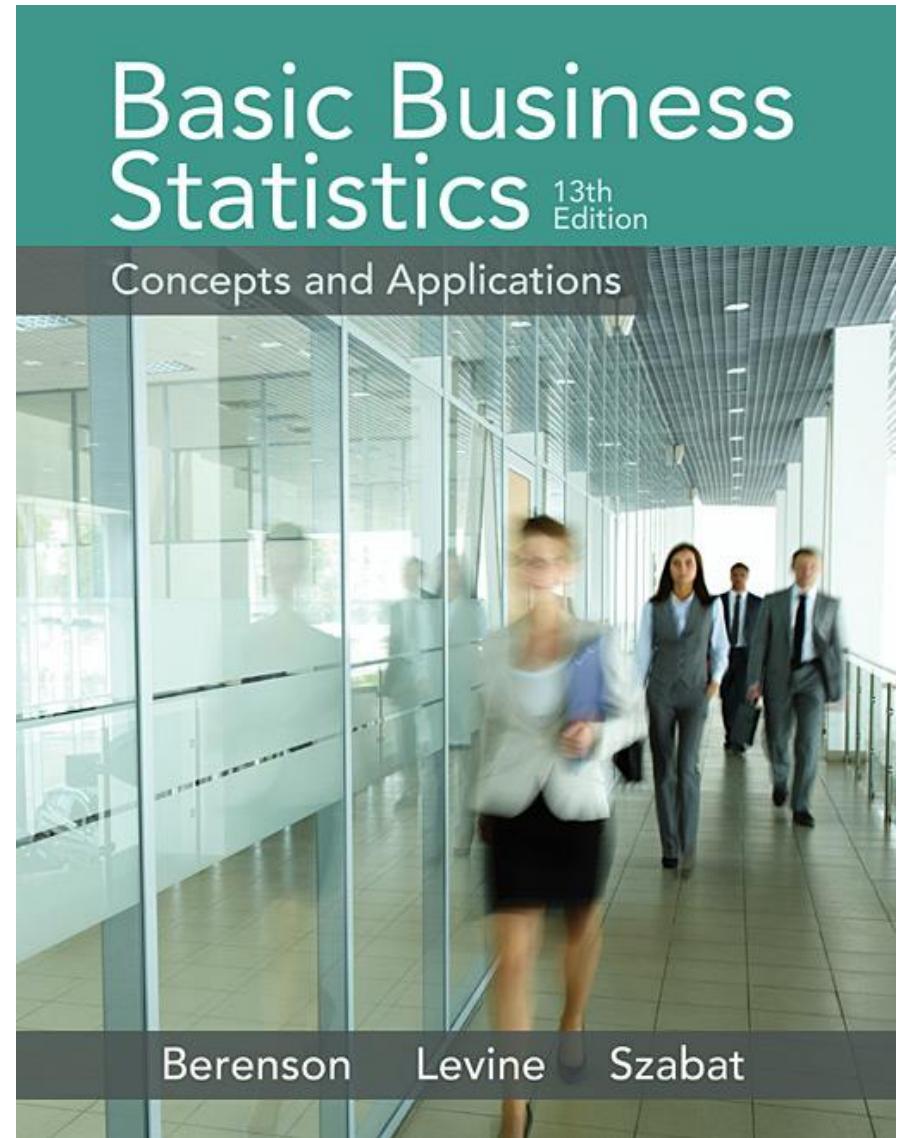


Chapter GS

Getting Started



Learning Objectives

In this chapter you learn:

- That the volume of data that exists in the world makes learning about statistics critically important
- That statistics is a way of thinking that can help you make better decisions
- How the DCOVA framework for applying statistics can help you solve business problems
- What business analytics is and how these techniques represent an opportunity for you
- How to make best use of this book
- How to prepare for using Microsoft Excel® or Minitab with this book

In Today's Business World You Cannot Escape From Data

- In today's digital world ever increasing amounts of data are gathered, stored, reported on, and available for further study.
- You hear the word data everywhere.
- Data are facts about the world and are constantly reported as numbers by an ever increasing number of sources.

Each Business Person Faces A Choice Of How To Deal With This Explosion Of Data

- They can ignore it and hope for the best.
- They can count on other people's summaries of data and hope they are correct.
- They can develop their own capability and insight into data by learning about statistics and its application to business.

Statistics Is Evolving So Businesses Can Use The Vast Amount Of Data Available

The emerging field of Business Analytics makes “extensive use of:

- Data
- Statistical and quantitative analysis
- Explanatory & predictive models
- Fact based management

to drive decisions and actions.”

To Properly Apply Statistics You Should Follow A Framework To Minimize Possible Errors

In this book we will use **DCOVA**

- **Define** the data you want to study in order to solve a problem or meet an objective
- **Collect** the data from appropriate sources
- **Organize** the data collected by developing tables
- **Visualize** the data by developing charts
- **Analyze** the data collected to reach conclusions and present results

Using The DCOVA Framework Helps You To Apply Statistics To:

- Summarize & visualize business data
- Reach conclusions from those data
- Make reliable forecasts about business activities
- Improve business processes

Definition Of Some Terms

DCOVA

VARIABLE

A characteristic of an item or individual.

DATA

The set of individual values associated with a variable.

STATISTICS

The methods that help transform data into useful information for decision makers.

Are These Numbers Useful In Making Decisions

- A survey of 1,179 adults 18 and over reported that 54% thought that 15 seconds was an acceptable online ad length before seeing free content.
- A survey of more than 3,000 full-time traditional-age students found that the students spent 51% of their time on socializing, recreation, and other activities; 9% of their time attending class/lab; and 7% of their time studying.

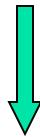
Without Statistics You Can't

- Determine if the numbers in these studies are useful information
- Validate claims of causality
- See patterns that large amounts of data sometimes reveal

Two Different Branches Of Statistics Are Used In Business

Statistics

The branch of mathematics that transforms data into useful information for decision makers.



Descriptive Statistics

Collecting, summarizing, presenting and analyzing data



Inferential Statistics

Using data collected from a small group to draw conclusions about a larger group

Business Analytics: The Changing Face Of Statistics

- Use statistical methods to analyze and explore data to uncover unforeseen relationships.
- Use management science methods to develop optimization models that impact an organization's strategy, planning, and operations.
- Use information systems' methods to collect and process data sets of all sizes, including very large data sets that would otherwise be hard to examine efficiently.

Business Analytics Has Already Been Applied In Many Business Decision-Making Contexts

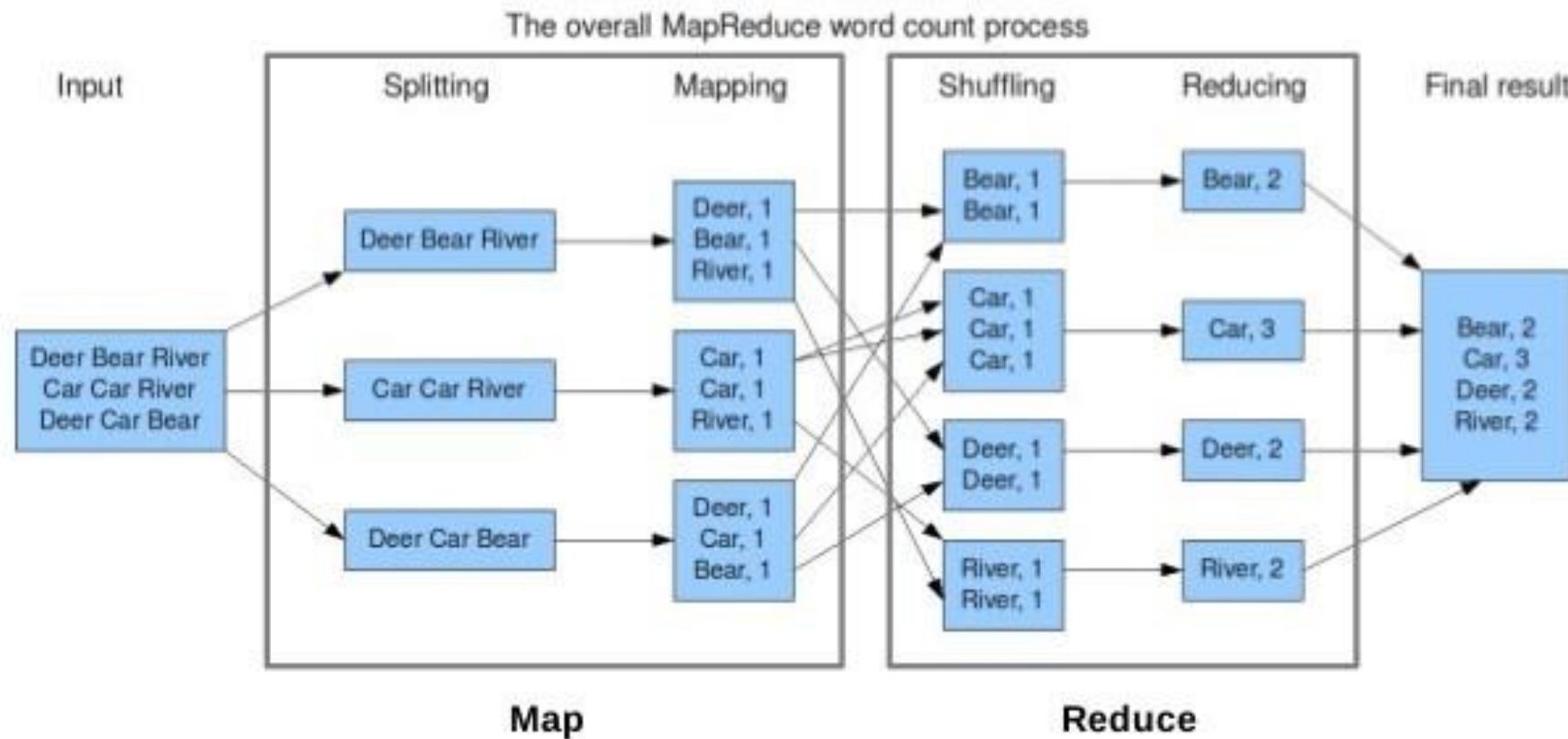
- Human resource managers (HR) understanding relationships between HR drivers, key business outcomes, employee skills, capabilities, and motivation.
- Financial analysts determining why certain trends occur to predict future financial environments.
- Marketers driving loyalty programs and customer marketing decisions to drive sales.
- Supply chain managers planning and forecasting based on product distribution and optimizing sales distribution based on key inventory measures.

The Growth Of “Big Data” Spurs The Use Of Business Analytics

- “Big Data” is still a fuzzy concept.
- Very large data sets are arising because of the automatic collection of high volumes of data at very fast rates.
- Older statistical techniques are often times impractical with big data.

Map Reduce: Google

Julius Ceasar...
“Divide and Conquer”



Statistics: An Important Part of Your Business Education

- You need analytical skills for the increasingly data-driven environment of business.
- Studies show an increase in productivity, innovation, and competitiveness for organizations that embrace business analytics.
- To quote Hal Varian, the chief economist at Google Inc., “the sexy job in the next 10 years will be statisticians. And I’m not kidding.”

How To Use This Book

- The Using Statistics scenario at the beginning and end of each chapter provide
 - A real business situation that the chapter's topics can help address
 - Context which is an important part of the learning process
- Throughout each chapter you will find Excel® and Minitab solutions to example problems.
- Numerous case studies are provided so you can
 - Apply what you have learned
 - Enhance your analytical & communication skills

Software and Statistics

- Software is used in statistics to assist you in applying statistical methods.
- This book covers the use of two software packages:
 - Microsoft Excel[®] -- Microsoft Office's data analysis application
 - Minitab -- a dedicated statistical analysis package.
- Either Excel[®] or Minitab can be used to learn and practice the statistical methods in this book.

Checklist For Preparing to Use Excel® or Minitab With This Book

- Determine which program, Excel or Minitab, you will use with this book.
- Read and review the Excel or Minitab Guide for this chapter to verify your knowledge of required basic skills.
- Read Appendix C to learn about the online resources you need to make best use of this book. Appendix C includes a complete list of the data files that are used in the examples and problems found in this book. Names of data files appear in this distinctive type face —**Retirement Funds**— throughout this book.
- Download the online resources that you will need to use this book, using the instructions in Appendix C.
- Check for updates to the program that you plan to use with this book, using the Appendix Section D.1 instructions.
- If you plan to use Excel with PHStat, the Visual Explorations add-in workbooks, or the Analysis ToolPak and you maintain your own computer system, read the special instructions in Appendix D.
- Examine Appendix G to learn answers to frequently asked questions (FAQs).

Basic Computing Skills You Need To Use Chapter Software Guides

Basic Skill

Identification of application window objects

Knowledge of mouse operations

Identification of dialog box objects

Specifics

Title bar, minimize/resize/close buttons, scroll bars, formula bar, workbook area, cell pointer, shortcut menu. For Excel only, pane and these Ribbon parts: tab, group, gallery, and launcher button

Click (also called select), check and clear, double-click, right-click, drag/drag-and-drop

Command button, list box, drop-down list, edit box, option button, check box

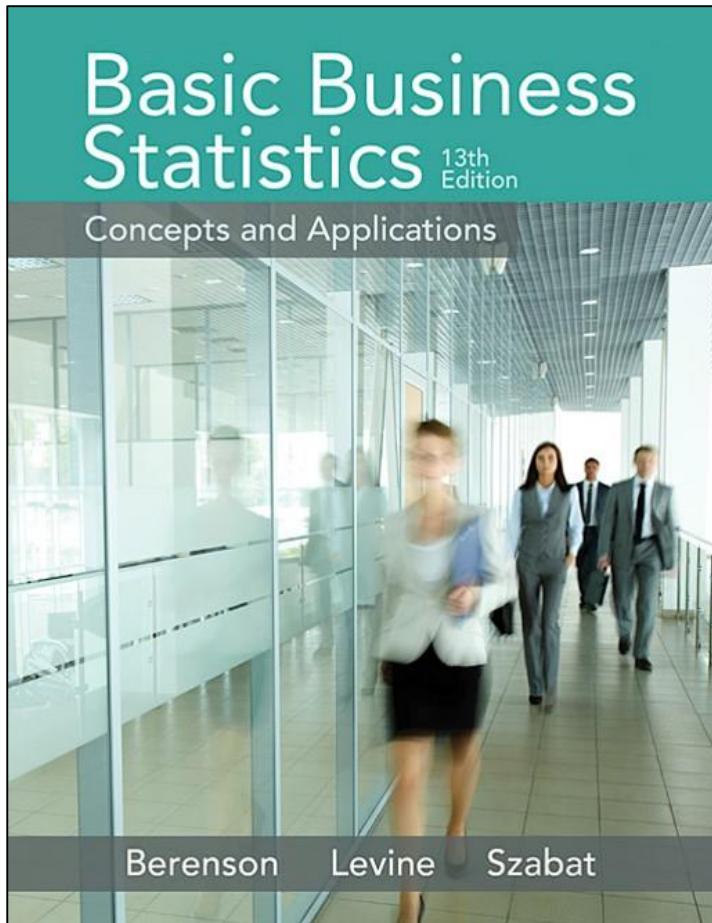
Chapter Summary

In this chapter we have seen:

- That the volume of data that exists in the world makes learning about statistics critically important
- That statistics is a way of thinking that can help you make better decisions
- What business analytics is and how these techniques represent an opportunity for you
- How the DCOVA framework for applying statistics can help you solve business problems
- How to make best use of this book
- How to prepare for using software with this book

Basic Business Statistics: Concepts and Applications

Thirteenth Edition



Chapter 1

Defining and Collecting Data

Learning Objectives

In this chapter you learn to:

1. Understand the types of variables used in statistics
2. Know the different measurement scales
3. Know how to collect data
4. Know the different ways to collect a sample
5. Understand the types of survey errors

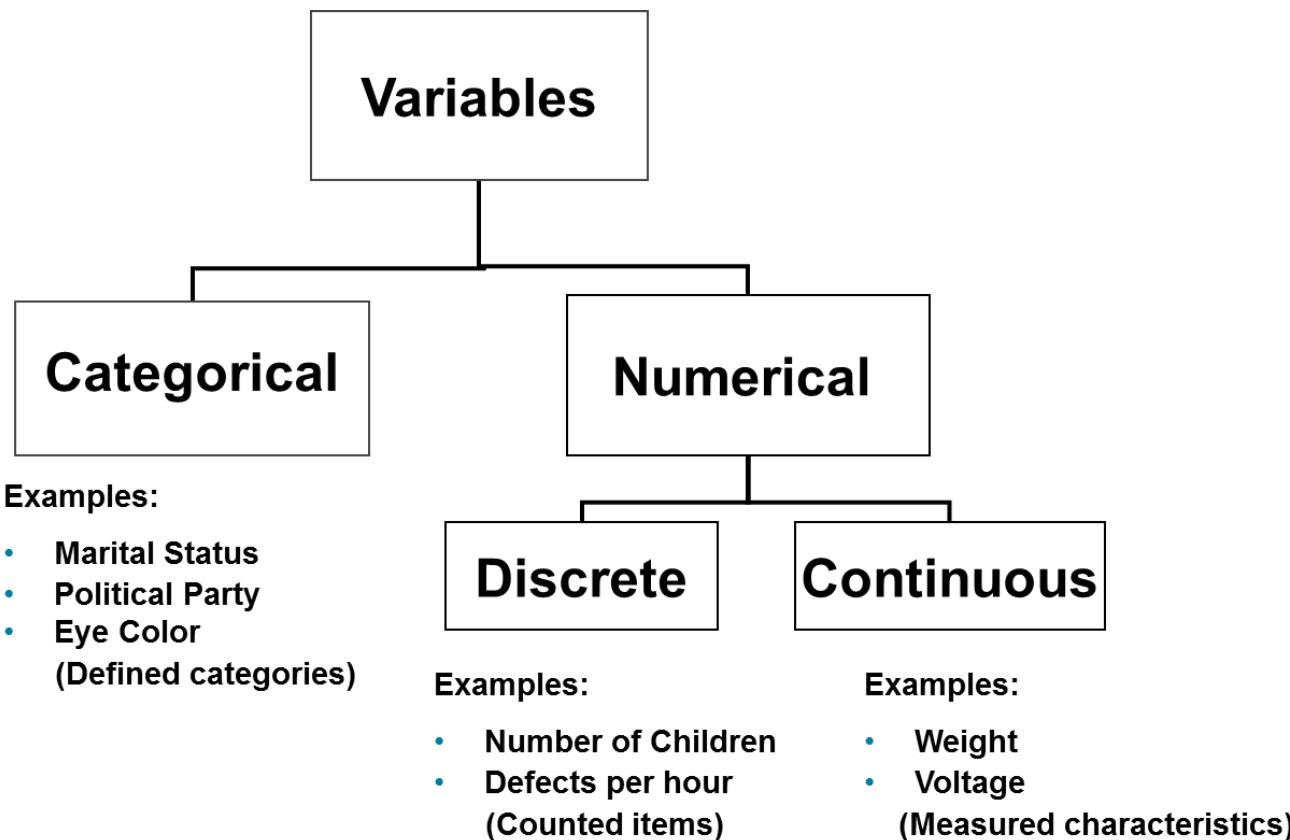
Types of Variables (1 of 2)

DCOVA

- Categorical (**qualitative**) variables have values that can only be placed into categories, such as “yes” and “no.”
- Numerical (**quantitative**) variables have values that represent a counted or measured quantity.
 - **Discrete** variables arise from a **counting process**
 - **Continuous** variables arise from a **measuring process**

Types of Variables (2 of 2)

DCOVA



Levels of Measurement

(1 of 3)

* For Categorical Variable

DCOVA

A **nominal scale** classifies data into distinct categories in which no ranking is implied.

| Categorical Variables | Categories |
|---------------------------------|------------------------------------|
| Do you have a Facebook profile? | Yes, No |
| Type of investment | Growth , Value , Other |
| Cellular Provider | AT&T, Sprint, Verizon, Other, None |

Levels of Measurement (2 of 3)

* For Categorical Variable

DCOVA

An **ordinal scale** classifies data into distinct categories in which ranking is implied

| Categorical Variable | Ordered Categories |
|--------------------------------|---|
| Student class designation | Freshman, Sophomore, Junior, Senior |
| Product satisfaction | Very unsatisfied, Fairly unsatisfied, Neutral, Fairly satisfied, Very satisfied |
| Faculty rank | Professor, Associate Professor, Assistant Professor, Instructor |
| Standard & Poor's bond ratings | AAA, AA, A, BBB, BB, B, CCC, CC, C, DDD, D, D |
| Student Grades | A, B, C, D, F |

Levels of Measurement (3 of 3)

For numerical Variables only

DCOVA

- An **interval scale** is an ordered scale in which the difference between measurements is a meaningful quantity but the measurements do not have a true zero point.
- A **ratio scale** is an ordered scale in which the difference between the measurements is a meaningful quantity and the measurements have a true zero point.

Interval and Ratio Scales

DCOVA

| <u>Numerical Variable</u> | <u>Level of Measurement</u> |
|--|-----------------------------|
| Temperature (in degrees Celsius or Fahrenheit) | → Interval |
| Standardized exam score (e.g., ACT or SAT) | → Interval |
| Height (in inches or centimeters) | → Ratio |
| Weight (in pounds or kilograms) | → Ratio |
| Age (in years or days) | → Ratio |
| Salary (in American dollars or Japanese yen) | → Ratio |

Establishing A Business Objective Focuses Data Collection

DCOVA

Examples Of Business Objectives:

- 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.

Collecting Data Correctly Is A Critical Task

DCOVA

- Need to avoid data flawed by biases, ambiguities, or other types of errors.
- Results from flawed data will be suspect or in error.
- Even the most sophisticated statistical methods are not very useful when the data is flawed.

Sources of Data

DCOVA

- **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.

Sources of data fall into five categories

DCOVA

- Data distributed by an organization or an individual
- The outcomes of a designed experiment
- The responses from a survey
- The results of conducting an observational study
- Data collected by ongoing business activities

Examples Of Data Distributed By Organizations or Individuals

DCOVA

- Financial data on a company provided by investment services.
- Industry or market data from market research firms and trade associations.
- Stock prices, weather conditions, and sports statistics in daily newspapers.

Examples of Data From A Designed Experiment

DCOVA

- Consumer testing of different versions of a product to help determine which product should be pursued further.
- Material testing to determine which supplier's material should be used in a product.
- Market testing on alternative product promotions to determine which promotion to use more broadly.

Examples Of Survey Data

DCOVA

- A survey asking people which laundry detergent has the best stain-removing abilities
- Political polls of registered voters during political campaigns.
- People being surveyed to determine their satisfaction with a recent product or service experience.

Evaluating Survey Worthiness

DCOVA

- What is the purpose of the survey?
- Is the survey based on a probability sample?
- Coverage error - appropriate frame?
- Nonresponse error - follow up
- Measurement error - good questions elicit good responses
- Sampling error - always exists

Types of Survey Errors (1 of 2)

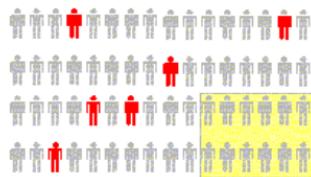
DCOVA

- Coverage error (results in selection bias)
 - Exists if some groups/items are excluded from the frame so that they have no chance of being selected in the sample or if items are included from outside the frame.
- Nonresponse error (results in nonresponse bias)
 - People who do not respond may be different from those who do respond so you need to follow up on the nonresponses, make several attempts to convince them to complete the survey.
- Sampling error
 - Variation/chance differences from sample to sample will always exist, based on the probability of particular individuals or items being selected in the particular samples. * results often have \pm
- Measurement error
 - Due to weaknesses in question design and / or respondent error

Types of Survey Errors (2 of 2)

DCOVA

- Coverage error



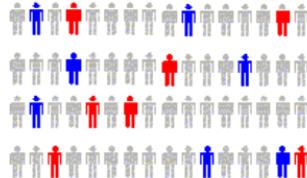
**Excluded from
frame**

- Nonresponse error



**Follow up on
nonresponses**

- Sampling error



**Random
differences from
sample to sample**

- Measurement error



**Bad or leading
question**

Examples of Data Collected From Observational Studies

DCOVA

- Market researchers utilizing focus groups to elicit unstructured responses to open-ended questions.
- Measuring the time it takes for customers to be served in a fast food establishment.
- Measuring the volume of traffic through an intersection to determine if some form of advertising at the intersection is justified.

Examples of Data Collected From Ongoing Business Activities

DCOVA

- A bank studies years of financial transactions to help them identify patterns of fraud.
- Economists utilize data on searches done via Google to help forecast future economic conditions.
- Marketing companies use tracking data to evaluate the effectiveness of a web site.

Data Is Collected From Either A Population or A Sample

DCOVA

Population

A **population** consists of all the items or individuals about which you want to draw a conclusion. The population is the “large group”

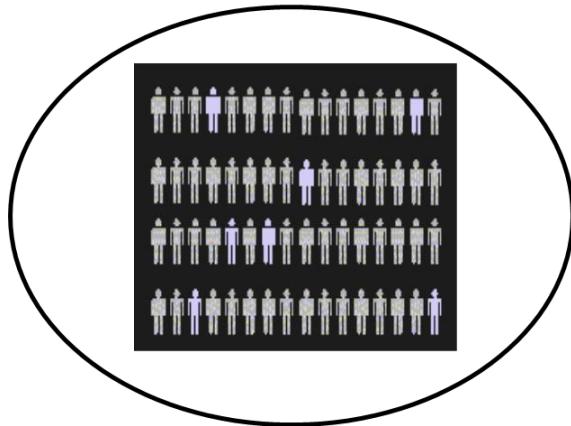
Sample

A **sample** is the portion of a population selected for analysis. The sample is the “small group”

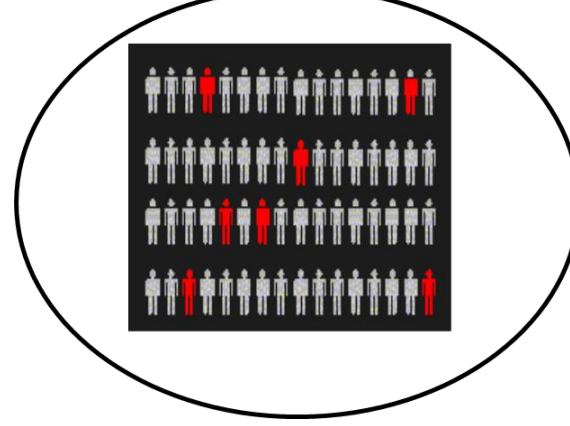
Population vs. Sample

DCOVA

Population



Sample



All the items or individuals
about which you want to
draw conclusion(s)

A portion of the population
of items or individuals

Collecting Data Via Sampling Is Used When Selecting A Sample Is

DCOVA

- Less time consuming than selecting every item in the population.
- Less costly than selecting every item in the population.
- Less cumbersome and more practical than analyzing the entire population.

Structured Data Follows An Organizing Principle & Unstructured Data Does Not

DCOVA

- A Stock Ticker Provides Structured Data:
 - The stock ticker repeatedly reports a company name, the number of shares last traded, the bid price, and the percent change in the stock price.
- Due to their inherent structure, data from tables and forms are structured data.
- E-mails from five people concerning stock trades is an example of unstructured data.
 - In these e-mails you cannot count on the information being shared in a specific order or format.
- This book will deal almost exclusively with structured data

Almost All Of The Methods In This Book Deal With Structured Data

DCOVA

- Some of the methods in Chapter 17 involve unstructured data.
- For many of the questions you might want to answer, the starting point will be tabular data.
- To deal with unstructured data, you will probably need to seek out help with more advanced methods / techniques.

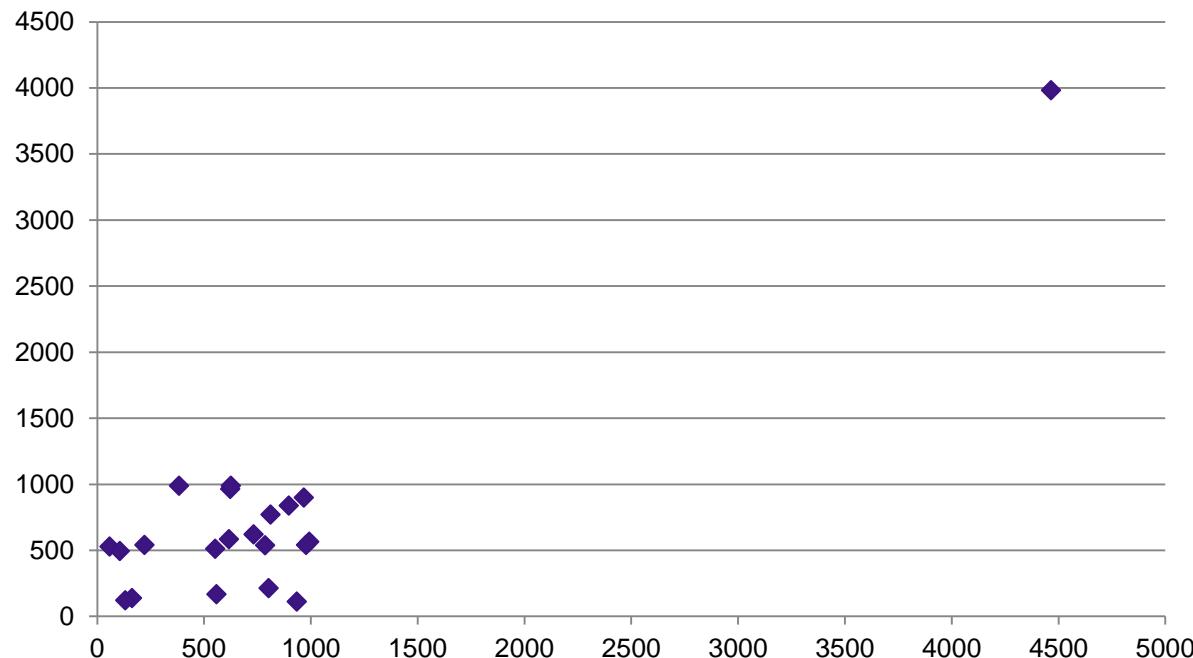
Data Cleaning Is Often A Necessary Activity When Collecting Data

DCOVA

- Often find “irregularities” in the data
 - Typographical or data entry errors
 - Values that are impossible or undefined
 - Missing values
 - Outliers
- When found these irregularities should be reviewed / addressed
- Both Excel & Minitab can be used to address irregularities

Data Cleaning

- Missing Values - can result in missstated results... Excel has no special values representing missing values, so you have to exclude manually
- Outliers – values that result in missstated relationships



After Collection It Is Often Helpful To Recode Some Variables

DCOVA

- Recoding a variable can either supplement or replace the original variable.
- Recoding a categorical variable involves redefining categories.
- Recoding a quantitative variable involves changing this variable into a categorical variable.
- When recoding be sure that the new categories are mutually exclusive (categories do not overlap) and collectively exhaustive (categories cover all possible values).

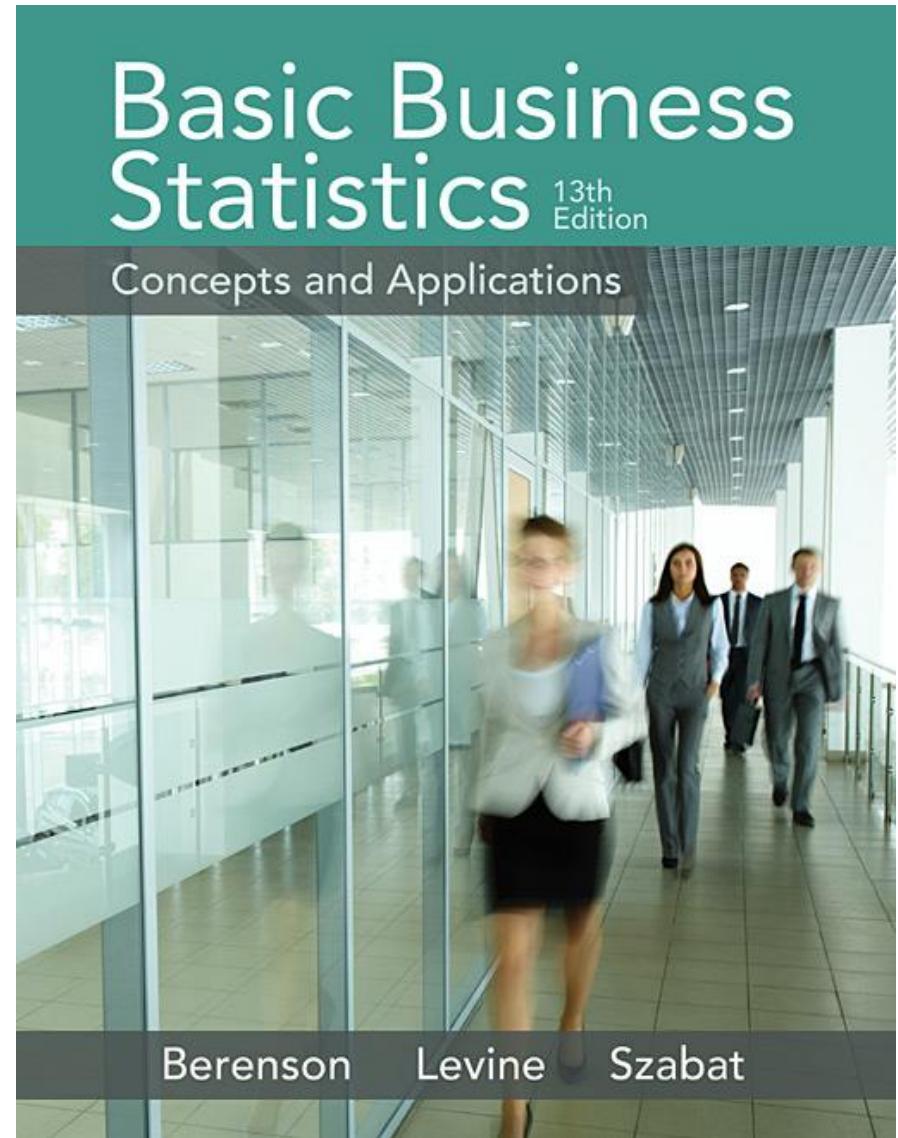
Chapter Summary

In this chapter we have discussed:

- The types of variables used in statistics
- The different measurement scales
- How to collect data
- The different ways to collect a sample
- The types of survey errors

Chapter 2

Organizing and Visualizing Variables



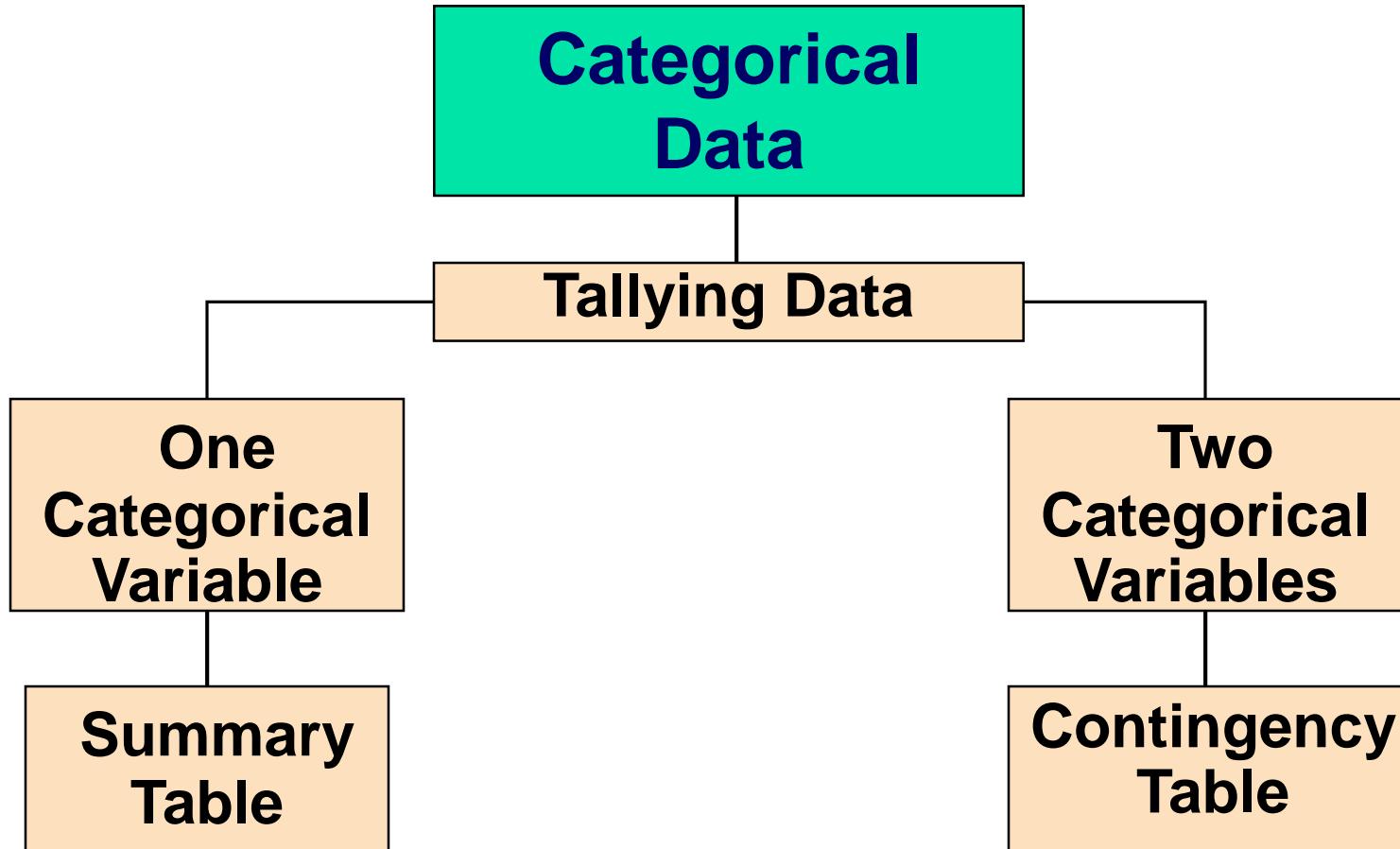
Learning Objectives

In this chapter you learn:

- To construct tables and charts for categorical data
- To construct tables and charts for numerical data
- The principles of properly presenting graphs
- To organize and analyze many variables

Categorical Data Are Organized By Utilizing Tables

DCOVA



Organizing Categorical Data: Summary Table

DCOVA

- A **summary table** tallies the frequencies or percentages of items in a set of categories so that you can see differences between categories.

Main Reason Young Adults Shop Online

| Reason For Shopping Online? | Percent |
|------------------------------------|---------|
| Better Prices | 37% |
| Avoiding holiday crowds or hassles | 29% |
| Convenience | 18% |
| Better selection | 13% |
| Ships directly | 3% |

Source: Data extracted and adapted from "Main Reason Young Adults Shop Online?"
USA Today, December 5, 2012, p. 1A.

A Contingency Table Helps Organize Two or More Categorical Variables

DCOVA

- Used to study patterns that may exist between the responses of two or more categorical variables
- Cross tabulates or tallies jointly the responses of the categorical variables
- For two variables the tallies for one variable are located in the rows and the tallies for the second variable are located in the columns

Contingency Table - Example

DCOVA

- A random sample of 400 invoices is drawn.
- Each invoice is categorized as a small, medium, or large amount.
- Each invoice is also examined to identify if there are any errors.
- This data are then organized in the contingency table to the right.

Contingency Table Showing Frequency of Invoices Categorized By Size and The Presence Of Errors

| | No Errors | Errors | Total |
|---------------|-----------|--------|-------|
| Small Amount | 170 | 20 | 190 |
| Medium Amount | 100 | 40 | 140 |
| Large Amount | 65 | 5 | 70 |
| Total | 335 | 65 | 400 |

Contingency Table Based On Percentage Of Overall Total

DCOVA

| | No Errors | Errors | Total |
|---------------|-----------|--------|-------|
| Small Amount | 170 | 20 | 190 |
| Medium Amount | 100 | 40 | 140 |
| Large Amount | 65 | 5 | 70 |
| Total | 335 | 65 | 400 |

83.75% of sampled invoices have no errors and 47.50% of sampled invoices are for small amounts.



$$42.50\% = 170 / 400$$

$$25.00\% = 100 / 400$$

$$16.25\% = 65 / 400$$



| | No Errors | Errors | Total |
|---------------|-----------|--------|--------|
| Small Amount | 42.50% | 5.00% | 47.50% |
| Medium Amount | 25.00% | 10.00% | 35.00% |
| Large Amount | 16.25% | 1.25% | 17.50% |
| Total | 83.75% | 16.25% | 100.0% |

Contingency Table Based On Percentage of Row Totals

DCOVA

| | No Errors | Errors | Total |
|---------------|-----------|--------|-------|
| Small Amount | 170 | 20 | 190 |
| Medium Amount | 100 | 40 | 140 |
| Large Amount | 65 | 5 | 70 |
| Total | 335 | 65 | 400 |



$$89.47\% = 170 / 190$$
$$71.43\% = 100 / 140$$
$$92.86\% = 65 / 70$$



| | No Errors | Errors | Total |
|---------------|-----------|--------|--------|
| Small Amount | 89.47% | 10.53% | 100.0% |
| Medium Amount | 71.43% | 28.57% | 100.0% |
| Large Amount | 92.86% | 7.14% | 100.0% |
| Total | 83.75% | 16.25% | 100.0% |

Medium invoices have a larger chance (28.57%) of having errors than small (10.53%) or large (7.14%) invoices.

Contingency Table Based On Percentage Of Column Totals

DCOVA

| | No Errors | Errors | Total |
|---------------|-----------|--------|-------|
| Small Amount | 170 | 20 | 190 |
| Medium Amount | 100 | 40 | 140 |
| Large Amount | 65 | 5 | 70 |
| Total | 335 | 65 | 400 |

$$50.75\% = 170 / 335$$
$$30.77\% = 20 / 65$$

| | No Errors | Errors | Total |
|---------------|-----------|--------|--------|
| Small Amount | 50.75% | 30.77% | 47.50% |
| Medium Amount | 29.85% | 61.54% | 35.00% |
| Large Amount | 19.40% | 7.69% | 17.50% |
| Total | 100.0% | 100.0% | 100.0% |

There is a 61.54% chance that invoices with errors are of medium size.

Tables Used For Organizing Numerical Data

DCOVA

Numerical Data

Ordered Array

Frequency
Distributions

Cumulative
Distributions

Organizing Numerical Data: Ordered Array

DCOVA

- 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)

| 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 |

Organizing Numerical Data: Frequency Distribution

DCOVA

- 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

DCOVA

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

DCOVA

- 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

Organizing Numerical Data: Frequency Distribution Example

DCOVA

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 | Midpoints | 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 |
| Total | | 20 |

Organizing Numerical Data: Relative & Percent Frequency Distribution Example

DCOVA

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 | $3/20 = .15$ | 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% |

Organizing Numerical Data: Cumulative Frequency Distribution Example

DCOVA

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 | 20 | 100% |

Why Use a Frequency Distribution?

DCOVA

- 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

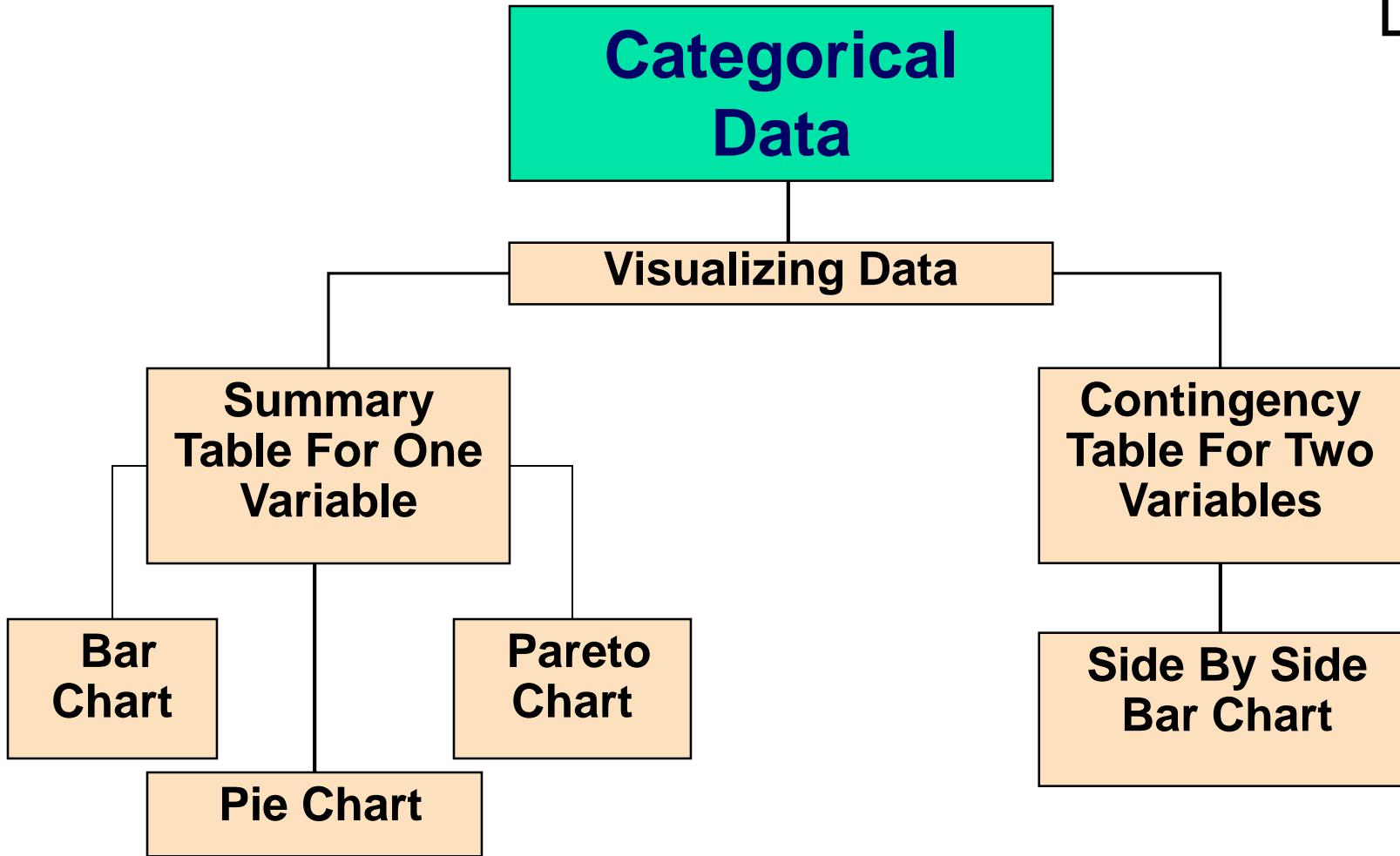
Frequency Distributions: Some Tips

DCOVA

- 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

Visualizing Categorical Data Through Graphical Displays

DCOVA



Visualizing Categorical Data: The Bar Chart

DCOVA

- The **bar chart** visualizes a categorical variable as a series of bars. The length of each bar represents either the frequency or percentage of values for each category. Each bar is separated by a space called a gap.

| Reason For Shopping Online? | Percent |
|------------------------------------|---------|
| Better Prices | 37% |
| Avoiding holiday crowds or hassles | 29% |
| Convenience | 18% |
| Better selection | 13% |
| Ships directly | 3% |

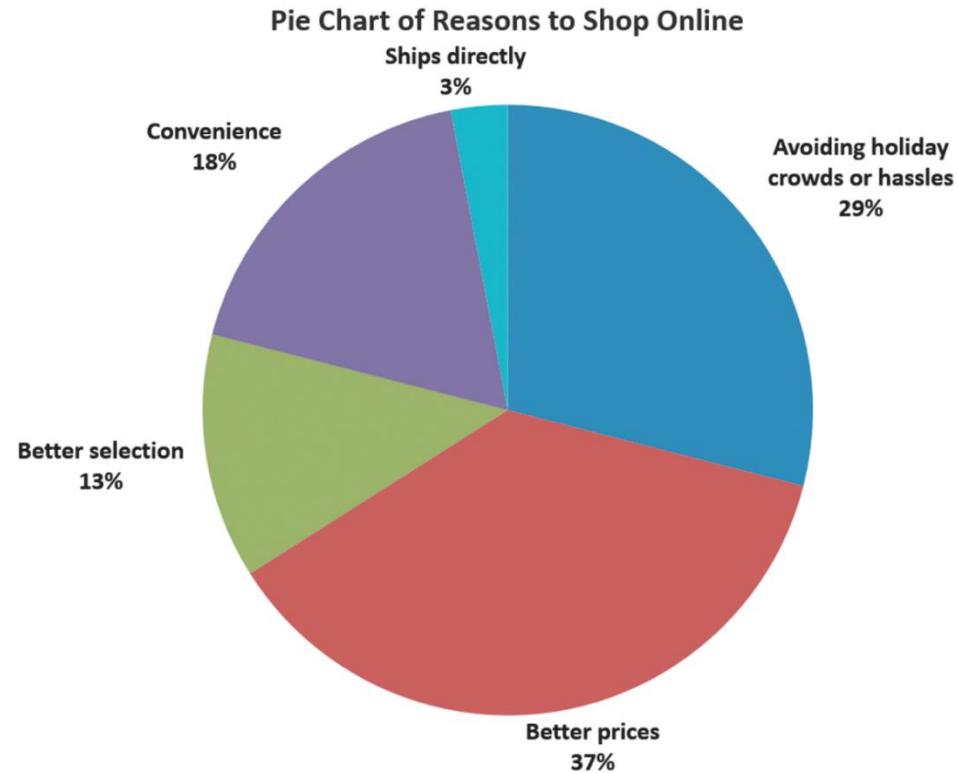


Visualizing Categorical Data: The Pie Chart

DCOVA

- 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.

| Reason For Shopping Online? | Percent |
|------------------------------------|---------|
| Better Prices | 37% |
| Avoiding holiday crowds or hassles | 29% |
| Convenience | 18% |
| Better selection | 13% |
| Ships directly | 3% |



Visualizing Categorical Data: The Pareto Chart

DCOVA

- 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”

80/20 Rule

Visualizing Categorical Data:

The Pareto Chart (con't)

DCOVA

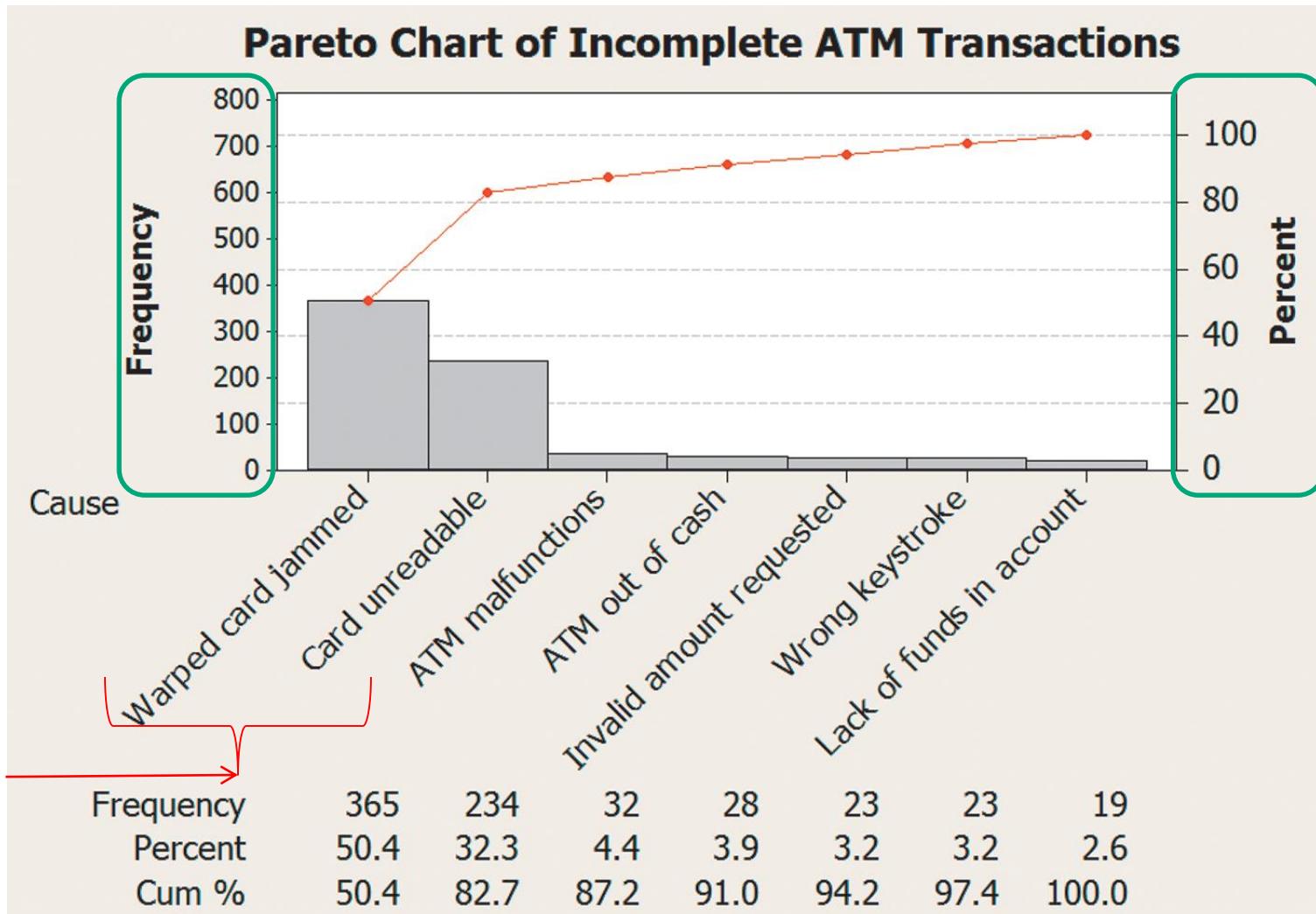
**Ordered Summary Table For Causes
Of Incomplete ATM Transactions**

| Cause | Frequency | Percent | Cumulative Percent |
|--------------------------|------------|----------------|--------------------|
| Warped card jammed | 365 | 50.41% | 50.41% |
| Card unreadable | 234 | 32.32% | 82.73% |
| ATM malfunctions | 32 | 4.42% | 87.15% |
| ATM out of cash | 28 | 3.87% | 91.02% |
| Invalid amount requested | 23 | 3.18% | 94.20% |
| Wrong keystroke | 23 | 3.18% | 97.38% |
| Lack of funds in account | 19 | 2.62% | 100.00% |
| Total | 724 | 100.00% | |

Source: Data extracted from A. Bhalla, "Don't Misuse the Pareto Principle," *Six Sigma Forum Magazine*, May 2009, pp. 15–18.

Visualizing Categorical Data: The Pareto Chart (con't)

DCOVA



The “Vital Few”

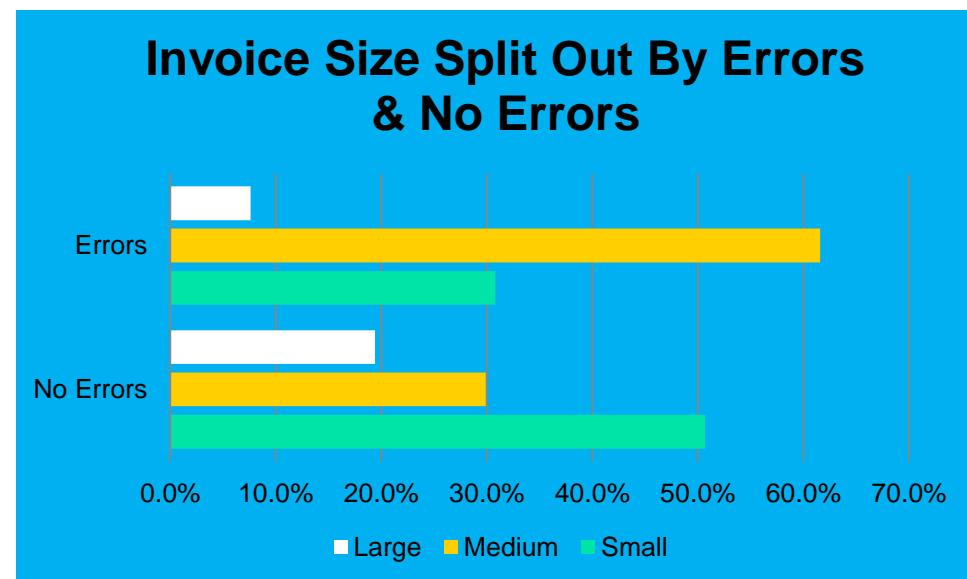
Notice the cumulative red dot(what we want) in the red line that show cum %

Visualizing Categorical Data: Side By Side Bar Charts

DCOVA

- The side by side bar chart represents the data from a contingency table.

| | No Errors | Errors | Total |
|---------------|-----------|--------|--------|
| Small Amount | 50.75% | 30.77% | 47.50% |
| Medium Amount | 29.85% | 61.54% | 35.00% |
| Large Amount | 19.40% | 7.69% | 17.50% |
| Total | 100.0% | 100.0% | 100.0% |



Invoices with errors are much more likely to be of medium size (61.54% vs 30.77% and 7.69%)

Organizing Many Categorical Variables: The Multidimensional Contingency Table

DCOVA

- A **multidimensional contingency table** is constructed by tallying the responses of three or more categorical variables.
- In Excel creating a Pivot Table yields an interactive display of this type.
- While Minitab will not create an interactive table, it has many specialized statistical & graphical procedures (not covered in this book) to analyze & visualize multidimensional data.

Using Excel Pivot Tables To Organize & Visualize Many Variables

DCOVA

A pivot table:

- Summarizes variables as a multidimensional summary table
- Allows interactive changing of the level of summarization and formatting of the variables
- Allows you to interactively “slice” your data to summarize subsets of data that meet specified criteria
- Can be used to discover possible patterns and relationships in multidimensional data that simpler tables and charts would fail to make apparent.

A Two Variable Contingency Table For The Retirement Funds Data

DCOVA

There are many more growth funds of average risk than of low or high risk

| | A | B | C | D | E |
|---|---|--------|---------|--------|-------------|
| 1 | Contingency Table of Fund Type and Risk | | | | |
| 2 | | | | | |
| 3 | RISK <input type="button" value="▼"/> | | | | |
| 4 | TYPE <input type="button" value="▼"/> | Low | Average | High | Grand Total |
| 5 | +Growth | 19.50% | 35.53% | 15.09% | 70.13% |
| 6 | +Value | 11.64% | 10.06% | 8.18% | 29.87% |
| 7 | Grand Total | 31.13% | 45.60% | 23.27% | 100.00% |

A Multidimensional Contingency Table

Tallies Responses Of Three or More Categorical Variables

DCOVA

| | A | B | C | D | E |
|----|--|--------|---------|--------|-------------|
| 1 | Contingency Table of Fund Type, Market Cap, and Risk | | | | |
| 2 | | | | | |
| 3 | RISK ▾ | | | | |
| 4 | TYPE ▾ | Low | Average | High | Grand Total |
| 5 | Growth | 19.50% | 35.53% | 15.09% | 70.13% |
| 6 | Large | 15.09% | 14.78% | 2.52% | 32.39% |
| 7 | Mid-Cap | 3.77% | 13.84% | 3.14% | 20.75% |
| 8 | Small | 0.63% | 6.92% | 9.43% | 16.98% |
| 9 | Value | 11.64% | 10.06% | 8.18% | 29.87% |
| 10 | Large | 9.43% | 7.86% | 0.00% | 17.30% |
| 11 | Mid-Cap | 1.57% | 1.57% | 2.83% | 5.97% |
| 12 | Small | 0.63% | 0.63% | 5.35% | 6.60% |
| 13 | Grand Total | 31.13% | 45.60% | 23.27% | 100.00% |

- Growth funds risk pattern depends on market

- Value funds risk pattern is different from that of growth funds.

Guidelines For Avoiding The Obscuring Of Data

DCOVA

- Avoid chartjunk
- Use the simplest possible visualization
- Include a title
- Label all axes
- Include a scale for each axis if the chart contains axis
- Begin the scale for a vertical axis at zero
- Use a constant scale

Graphical Errors: Chart Junk

DCOVA



Bad Presentation



Good Presentation

Minimum Wage



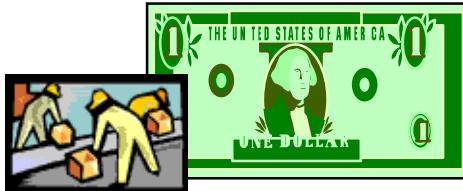
1960: \$1.00



1970: \$1.60

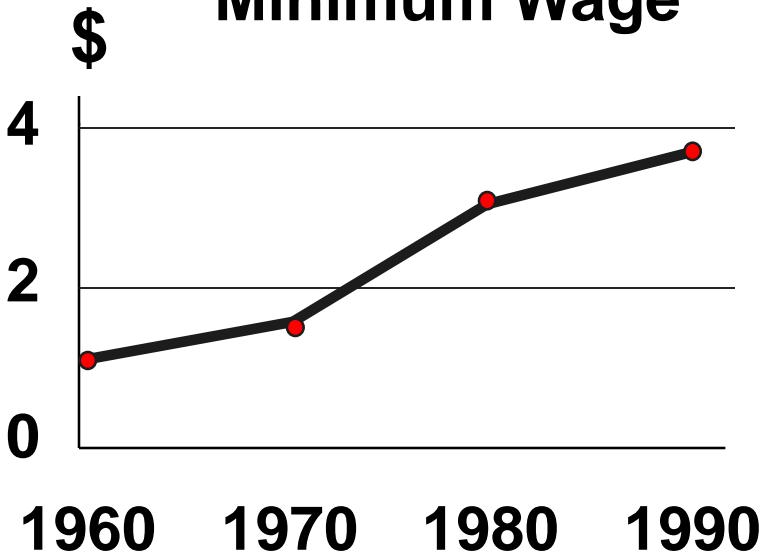


1980: \$3.10



1990: \$3.80

Minimum Wage



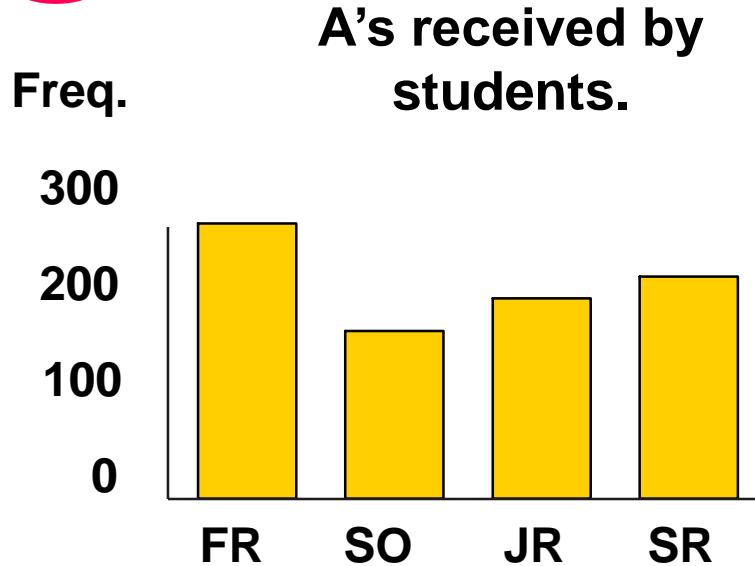
Graphical Errors:

No Relative Basis

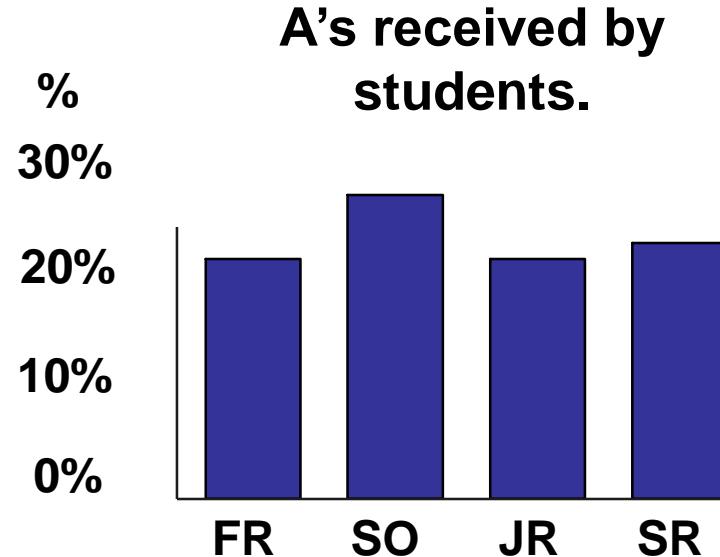
DCOVA



Bad Presentation



Good Presentation



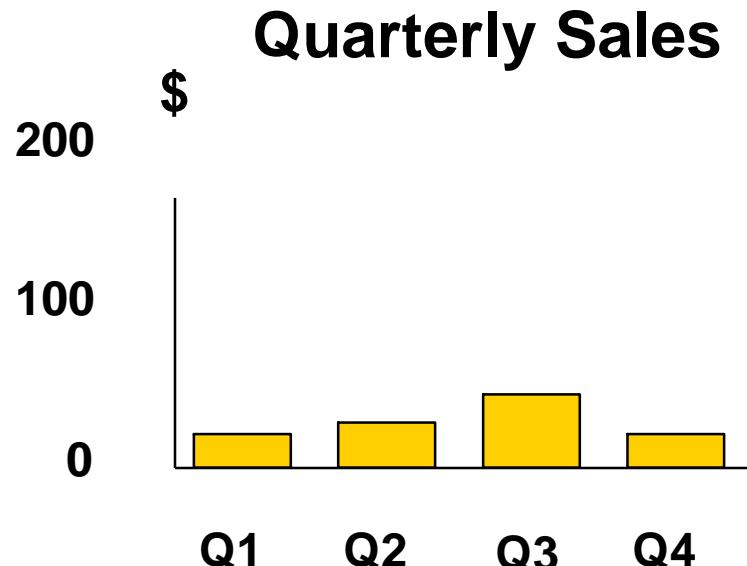
FR = Freshmen, SO = Sophomore, JR = Junior, SR = Senior

Graphical Errors: Compressing the Vertical Axis

DCOVA



Bad Presentation



Good Presentation

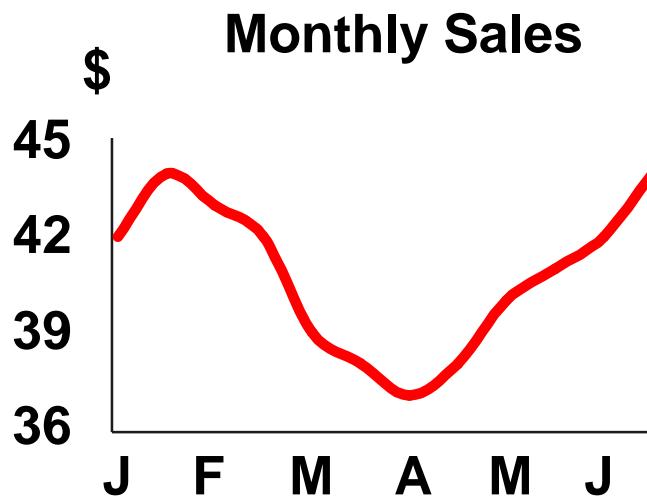


Graphical Errors: No Zero Point on the Vertical Axis

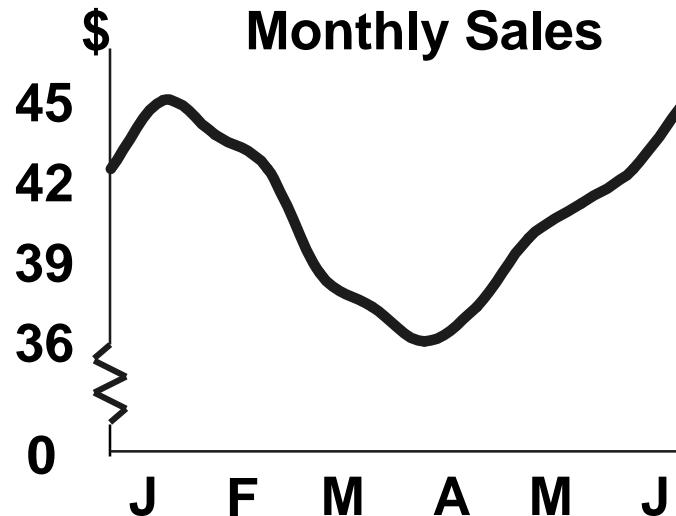
DCOVA



Bad Presentation



Good Presentations



Graphing the first six months of sales

In Excel It Is Easy To Inadvertently (unintentionally) Create Distortions

- Excel often will create a graph where the vertical axis does not start at 0
- Excel offers the opportunity to turn simple charts into 3-D charts and in the process can create distorted images
- Unusual charts offered as choices by excel will most often create distorted images

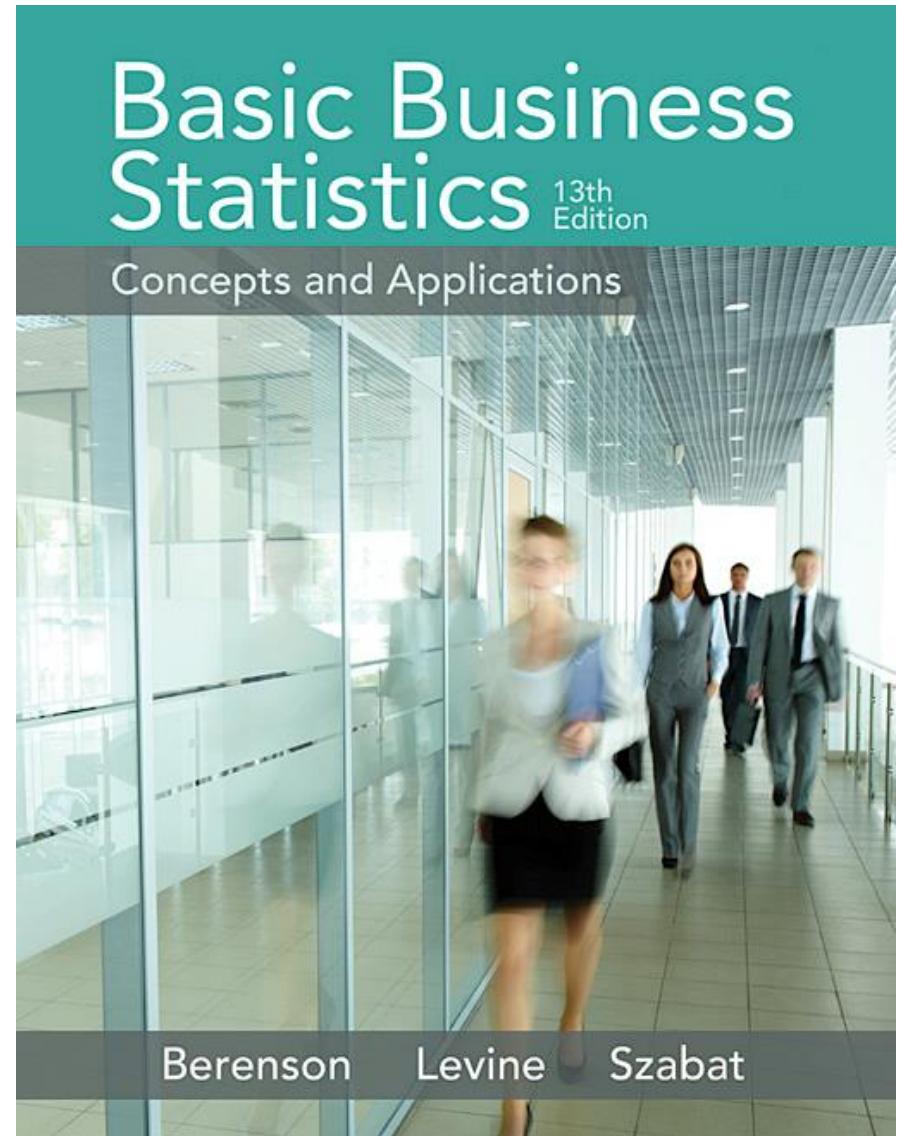
Chapter Summary

In this chapter we have:

- Constructed tables and charts for categorical data
- Constructed tables and charts for numerical data
- Examined the principles of properly presenting graphs
- Examined methods to organize and analyze many variables in Excel

Chapter 3

Numerical Descriptive Measures



Learning Objectives

In this chapter, you learn:

- To describe the properties of central tendency, variation, and shape in numerical data
- To construct and interpret a boxplot
- To compute descriptive summary measures for a population
- To calculate the covariance and the coefficient of correlation

Summary Definitions

DCOVA

- 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
- The **shape** is the pattern of the distribution of values from the lowest value to the highest value.

Measures of Central Tendency:

The Mean

DCOVA

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

- For a sample of size n:

Pronounced x-bar

The ith value

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

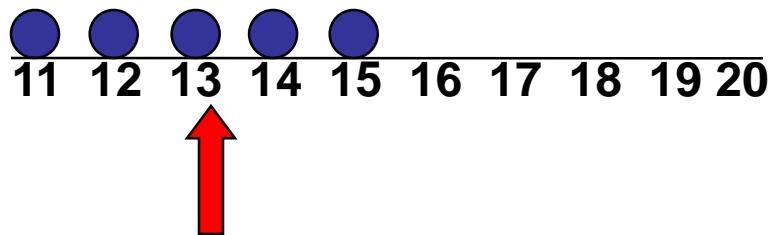
Sample size

Observed values

Measures of Central Tendency: The Mean (con't)

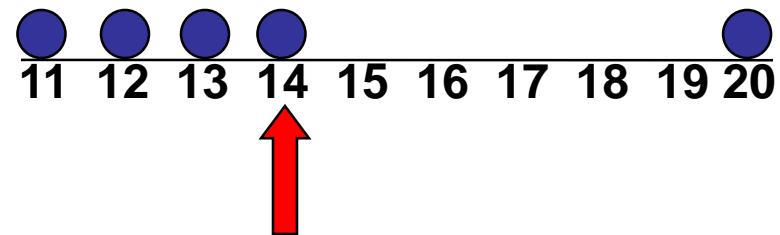
DCOV A

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



Mean = 13

$$\frac{11+12+13+14+15}{5} = \frac{65}{5} = 13$$



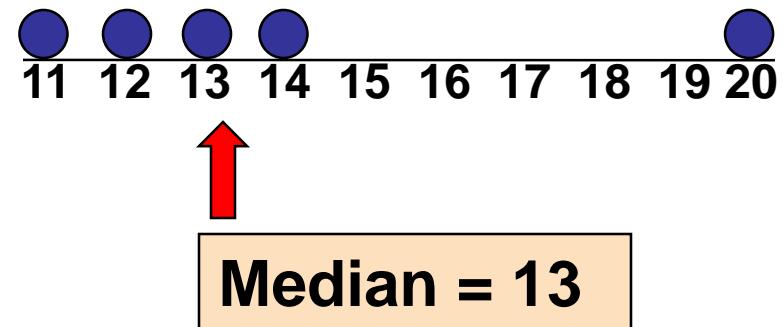
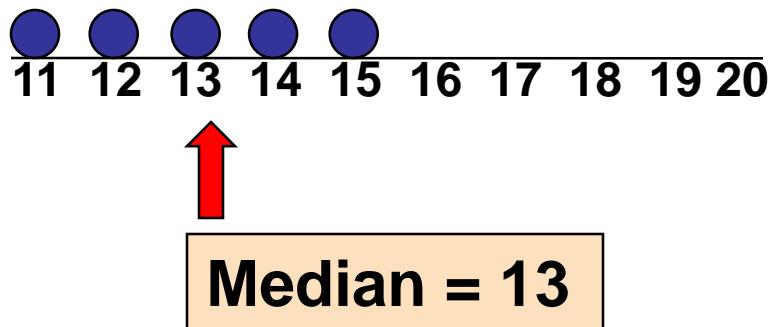
Mean = 14

$$\frac{11+12+13+14+20}{5} = \frac{70}{5} = 14$$

Measures of Central Tendency: The Median

DCOV A

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



- Less sensitive than the mean to extreme values

Measures of Central Tendency: Locating the Median

DCOV A

- 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

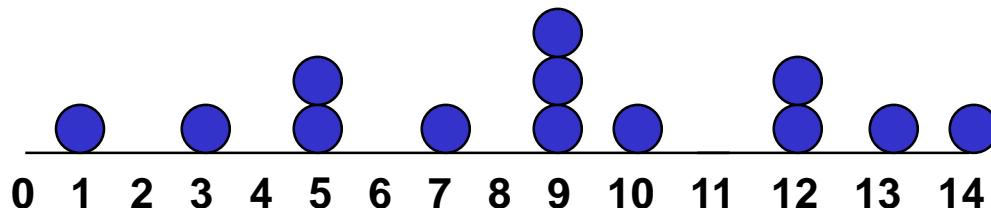
Note that $\frac{n+1}{2}$ is not the *value* of the median, only the *position* of the median in the ranked data

Measures of Central Tendency:

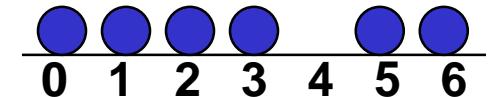
The Mode

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

DCOV A



Mode = 9



No Mode

Measures of Central Tendency: Review Example

DCOVA

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?

DCOVA

- 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.
- Use **mode** for categorical data (i.e. what make of car is most used)

Measure of Central Tendency For The Rate Of Change Of A Variable Over Time: The Geometric Mean & The Geometric Rate of Return

DCOVA

- Geometric mean
 - Used to measure the rate of change of a variable over time. X_n is the nth value.

$$\bar{X}_G = (X_1 \times X_2 \times \cdots \times X_n)^{1/n}$$

- Geometric mean rate of return
 - Measures the status of an investment over time

$$\bar{R}_G = [(1 + R_1) \times (1 + R_2) \times \cdots \times (1 + R_n)]^{1/n} - 1$$

- Where R_i is the rate of return in time period i

The Geometric Mean & The Mean Rate of Return: Example

DCOVA

An investment of \$100,000 declined to \$50,000 at the end of year one and rebounded to \$100,000 at end of year two:

$$X_1 = \$100,000 \quad X_2 = \$50,000 \quad X_3 = \$100,000$$



50% decrease

100% increase

The overall two-year return is zero, since it started and ended at the same level.

The Geometric Mean & The Mean Rate of Return: Example (con't)

$$R_1 = (50000 - 100000) / 100000 = -50\% \text{ or } -.5$$

$$R_2 = (100000 - 50000) / 50000 = 1.00 \text{ or } 100\%$$

DCOVA

Use the 1-year returns to compute the arithmetic mean and the geometric mean:

Arithmetic mean rate of return:

$$\bar{X} = \frac{(-.5) + (1)}{2} = .25 = 25\%$$

Misleading result

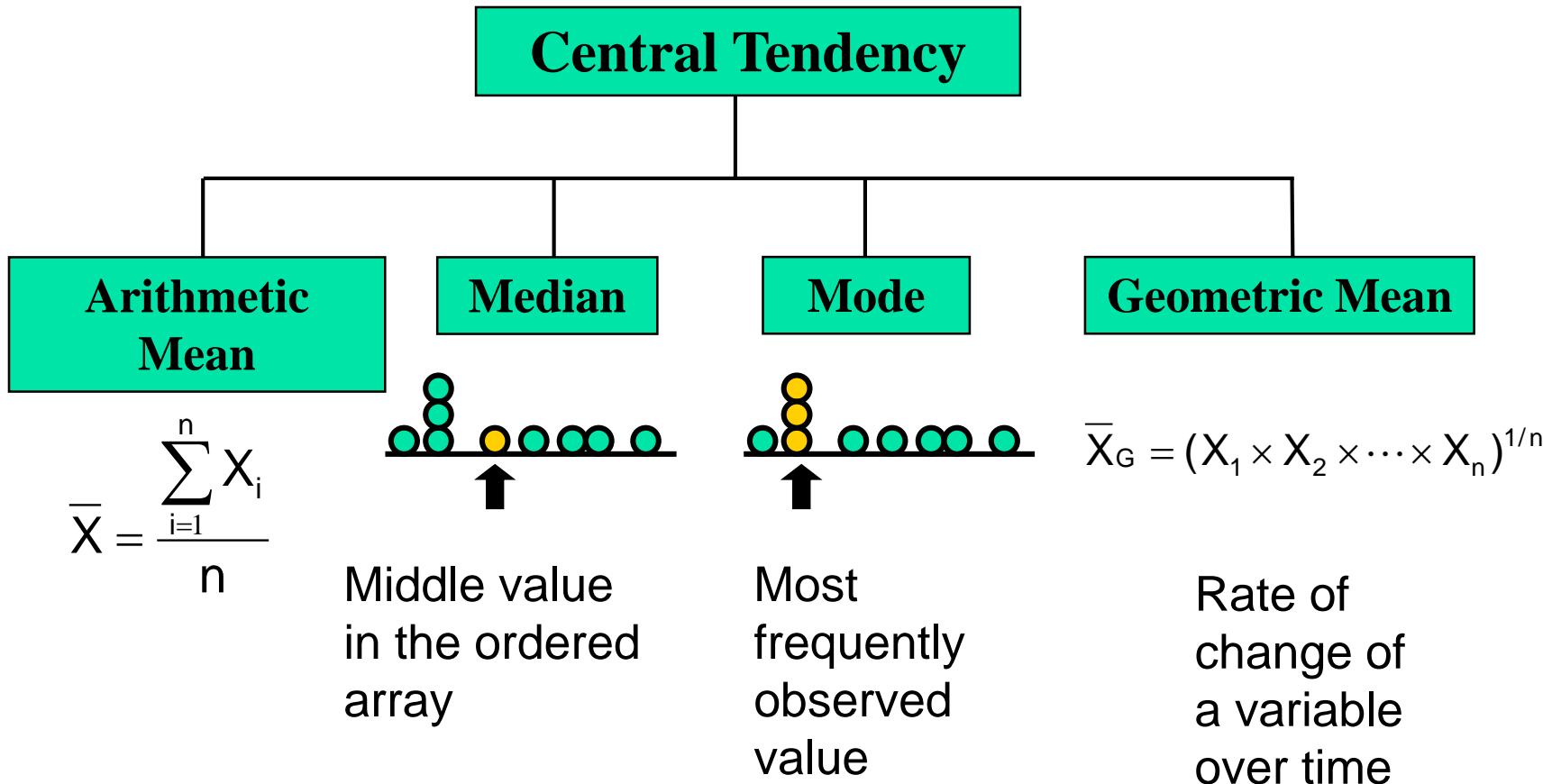
Geometric mean rate of return for the first 2 years:

$$\begin{aligned}\bar{R}_G &= [(1 + R_1) \times (1 + R_2) \times \cdots \times (1 + R_n)]^{1/n} - 1 \\ &= [(1 + (-.5)) \times (1 + (1))]^{1/2} - 1 \\ &= [(.50) \times (2)]^{1/2} - 1 = 1^{1/2} - 1 = 0\%\end{aligned}$$

More representative result

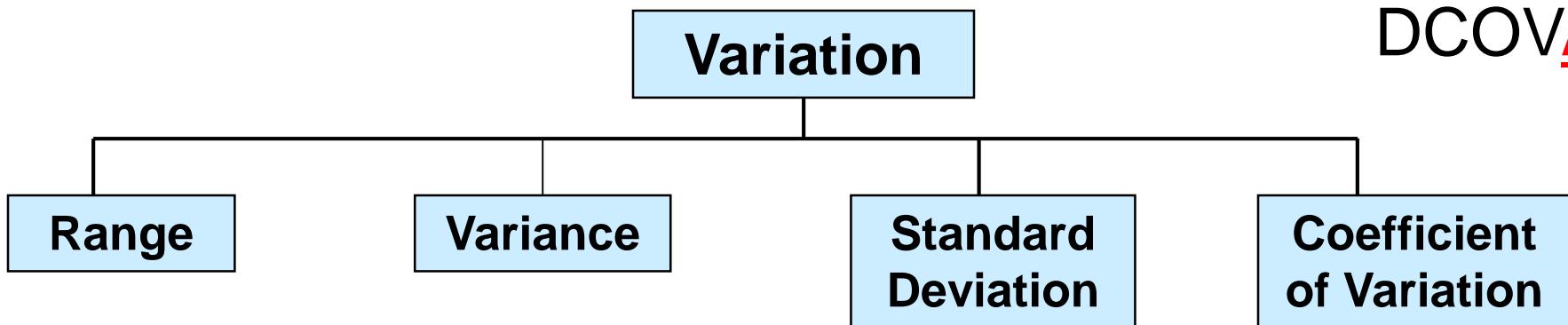
Measures of Central Tendency: Summary

DCOVA

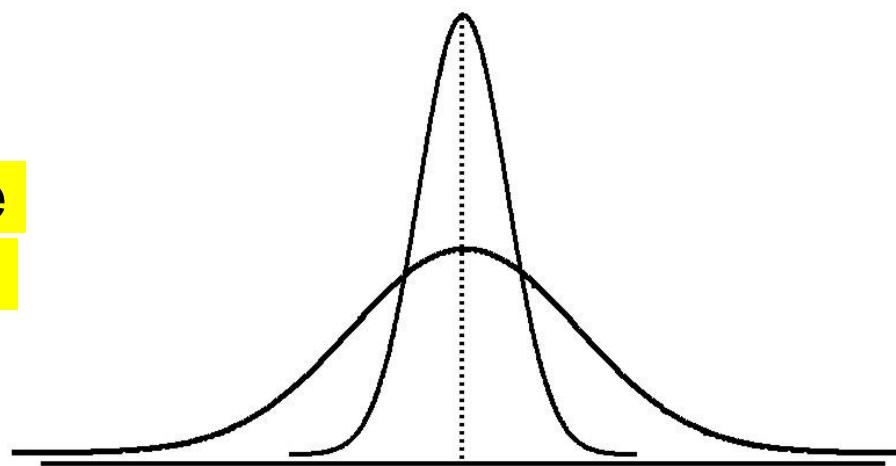


Measures of Variation

DCOV_A



- Measures of variation give information on the **spread** or **variability** or **dispersion** of the data values.



Same center,
different variation

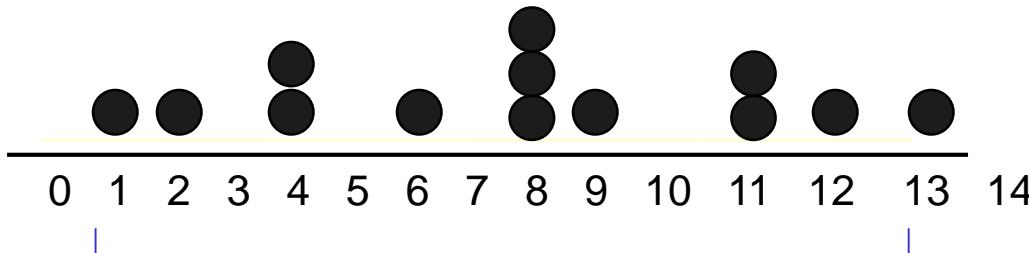
Measures of Variation: The Range

DCOVA

- Simplest measure of variation
- Difference between the largest and the smallest values:

$$\text{Range} = X_{\text{largest}} - X_{\text{smallest}}$$

Example:

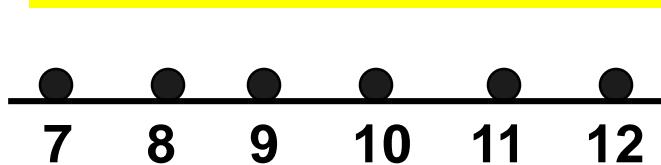


$$\text{Range} = 13 - 1 = 12$$

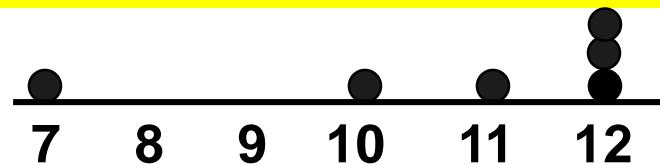
Measures of Variation: Why The Range Can Be Misleading

DCOVA

- Does not account for how the data are distributed



$$\text{Range} = 12 - 7 = 5$$



$$\text{Range} = 12 - 7 = 5$$

- Sensitive to outliers

1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,3,3,3,3,4,5

$$\text{Range} = 5 - 1 = 4$$

1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,3,3,3,3,4,120

$$\text{Range} = 120 - 1 = 119$$

Measures of Variation: The Sample Variance

DCOVA

- Average (approximately) of squared deviations of values from the mean (**Sum of Squares**)

- Sample variance:

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

Where

\bar{X} = arithmetic mean

n = sample size

X_i = ith value of the variable X

Measures of Variation: The Sample Standard Deviation

DCOV A

- Most commonly used measure of variation
- Shows variation about the mean
- Is the square root of the variance
- Has the **same units as the original data**

SE is standard error.

- **Sample standard deviation:**

σ ← Standard deviation

$$SE = \frac{\sigma}{\sqrt{n}}$$

← Number of samples

$$S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}}$$

Measures of Variation: The Standard Deviation

DCOVA

Steps for Computing Standard Deviation

1. Compute the difference between each value and the mean.
2. Square each difference.
3. Add the squared differences.
4. Divide this total by $n-1$ to get the sample variance.
5. Take the square root of the sample variance to get the sample standard deviation.

Measures of Variation: Sample Standard Deviation: Calculation Example

DCOVA

Sample
Data (X_i) :

10 12 14 15 17 18 18 24

$$n = 8$$

$$\text{Mean} = \bar{X} = 16$$

$$s = \sqrt{\frac{(10 - \bar{X})^2 + (12 - \bar{X})^2 + (14 - \bar{X})^2 + \dots + (24 - \bar{X})^2}{n - 1}}$$

$$= \sqrt{\frac{(10 - 16)^2 + (12 - 16)^2 + (14 - 16)^2 + \dots + (24 - 16)^2}{8 - 1}}$$

$$= \sqrt{\frac{130}{7}}$$

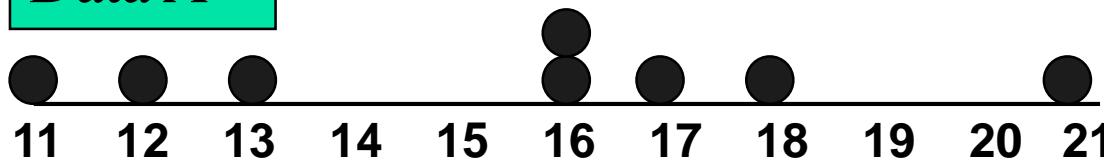
$$= 4.3095$$

A measure of the “average”
scatter around the mean

Measures of Variation: Comparing Standard Deviations

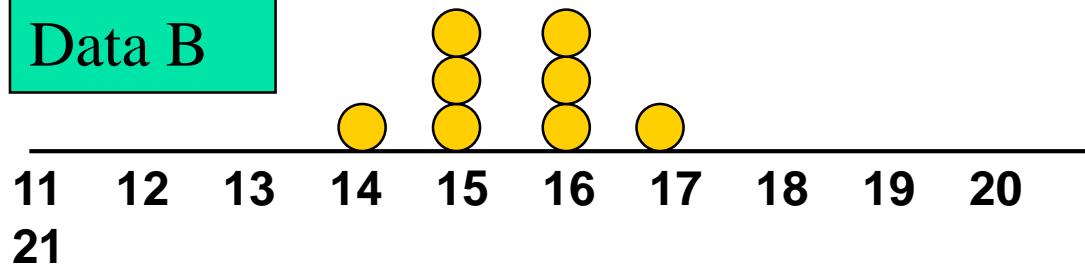
DCOVA

Data A



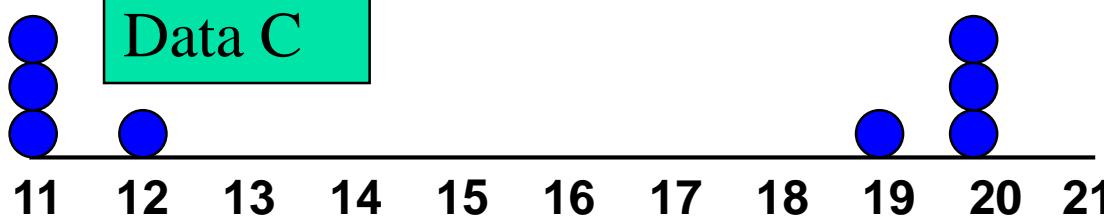
Mean = 15.5
 $S = 3.338$

Data B



Mean = 15.5
 $S = 0.926$

Data C



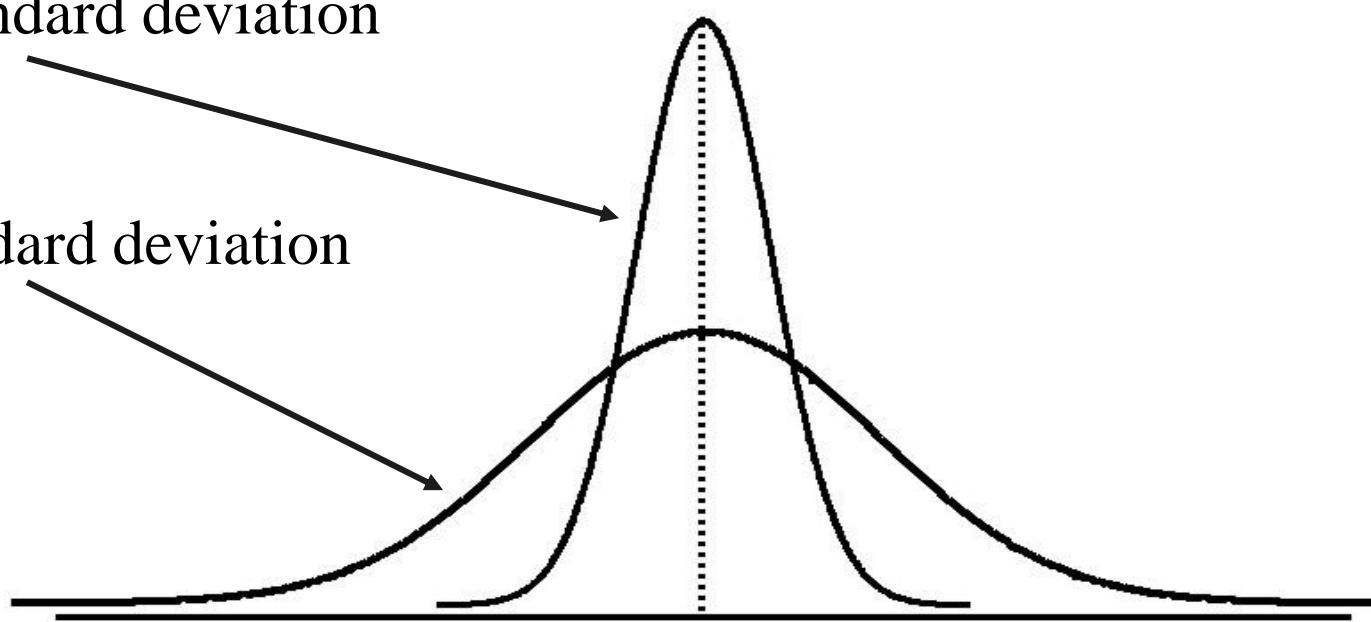
Mean = 15.5
 $S = 4.567$

Measures of Variation: Comparing Standard Deviations

DCOVA

Smaller standard deviation

Larger standard deviation



Measures of Variation: Summary Characteristics

DCOVA

- The more the data are spread out, the greater the range, variance, and standard deviation.
- The more the data are concentrated, the smaller the range, variance, and standard deviation.
- If the values are all the same (no variation), all these measures will be zero.
- None of these measures are ever negative.

Measures of Variation: The Coefficient of Variation

DCOVA

- Measures relative variation
- Always in percentage (%)
- Shows variation relative to mean
- Can be used to compare the variability of two or more sets of data measured in different units

$$CV = \left(\frac{S}{\bar{X}} \right) \cdot 100\%$$

Measures of Variation: Comparing Coefficients of Variation

DCOV_A

Stock A:

- Average price last year = \$50
- Standard deviation = \$5

$$CV_A = \left(\frac{S}{\bar{X}} \right) \cdot 100\% = \frac{\$5}{\$50} \cdot 100\% = 10\%$$

Stock B:

- Average price last year = \$100
- Standard deviation = \$5

$$CV_B = \left(\frac{S}{\bar{X}} \right) \cdot 100\% = \frac{\$5}{\$100} \cdot 100\% = 5\%$$

Both stocks have the same standard deviation, but stock B is less variable relative to its price.

B is less risky with less growth than A.

Measures of Variation: Comparing Coefficients of Variation (con't)

- Stock A:

- Average price last year = \$50
- Standard deviation = \$5

$$CV_A = \left(\frac{S}{\bar{X}} \right) \cdot 100\% = \frac{\$5}{\$50} \cdot 100\% = 10\%$$

DCOVA

- Stock C:

- Average price last year = \$8
- Standard deviation = \$2

$$CV_C = \left(\frac{S}{\bar{X}} \right) \cdot 100\% = \frac{\$2}{\$8} \cdot 100\% = 25\%$$

Stock C has a much smaller standard deviation but a much higher coefficient of variation

Locating Extreme Outliers: Z-Score

DCOV A

- To compute the **Z-score** of a data value, subtract the mean and divide by the standard deviation.
- The Z-score is the number of standard deviations a data value is from the mean.
- A data value is considered an extreme outlier if its Z-score is less than -3.0 or greater than +3.0.
- The larger the absolute value of the Z-score, the farther the data value is from the mean.

Locating Extreme Outliers: Z-Score Formula

DCOVA

$$Z = \frac{X - \bar{X}}{S}$$

where X represents the data value

\bar{X} is the sample mean

S is the sample standard deviation

Locating Extreme Outliers: Z-Score Example

DCOVA

- Suppose the mean math SAT score is 490, with a standard deviation of 100.
- Compute the Z-score for a test score of 620.

$$Z = \frac{X - \bar{X}}{S} = \frac{620 - 490}{100} = \frac{130}{100} = 1.3$$

A score of 620 is 1.3 standard deviations above the mean (above b/c $620 > 490$) and would not be considered an outlier (b/c it's < 3.0).

Shape of a Distribution

DCOVA

- Describes how data are distributed
- Two useful shape related statistics are:
 - Skewness
 - Measures the extent to which data values are not symmetrical
 - Kurtosis
 - Kurtosis affects the peakedness of the curve of the distribution—that is, how sharply the curve rises approaching the center of the distribution

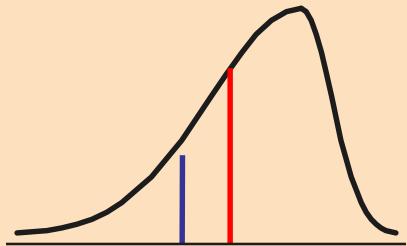
Shape of a Distribution (Left, Right Skewness)

DCOVA

- Measures the extent to which data is not symmetrical

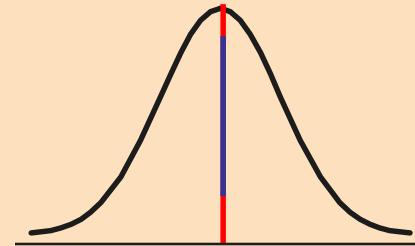
Left-Skewed

Mean < Median



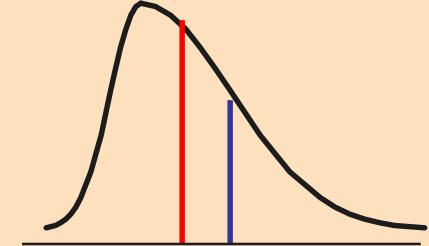
Symmetric

Mean = Median



Right-Skewed

Median < Mean



Skewness
Statistic

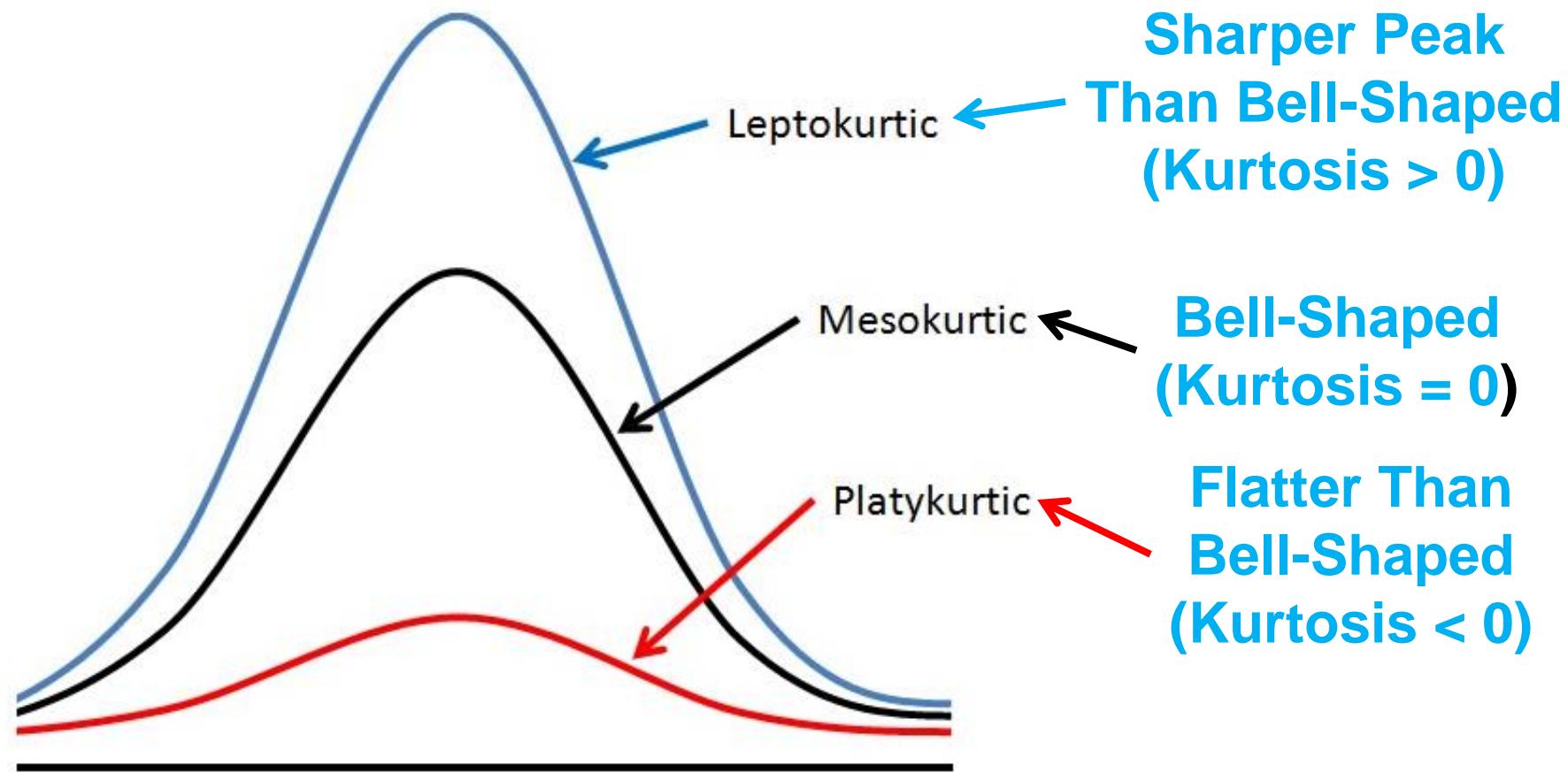
< 0

0

>0

Shape of a Distribution -- Kurtosis
measures how sharply the curve rises
approaching the center of the distribution)

DCOVA



General Descriptive Stats Using

Microsoft Excel Functions

DCOV_A

| House Prices | | <u>Descriptive Statistics</u> | | |
|--------------|--|-------------------------------|-----------------|-----------------|
| \$ 2,000,000 | | Mean | \$ 600,000 | =AVERAGE(A2:A6) |
| \$ 500,000 | | Standard Error | \$ 357,770.88 | =D6/SQRT(D14) |
| \$ 300,000 | | Median | \$ 300,000 | =MEDIAN(A2:A6) |
| \$ 100,000 | | Mode | \$ 100,000.00 | =MODE(A2:A6) |
| \$ 100,000 | | Standard Deviation | \$ 800,000 | =STDEV(A2:A6) |
| | | Sample Variance | 640,000,000,000 | =VAR(A2:A6) |
| | | Kurtosis | 4.1301 | =KURT(A2:A6) |
| | | Skewness | 2.0068 | =SKEW(A2:A6) |
| | | Range | \$ 1,900,000 | =D12 - D11 |
| | | Minimum | \$ 100,000 | =MIN(A2:A6) |
| | | Maximum | \$ 2,000,000 | =MAX(A2:A6) |
| | | Sum | \$ 3,000,000 | =SUM(A2:A6) |
| | | Count | 5 | =COUNT(A2:A6) |
| | | | | |

General Descriptive Stats Using Microsoft Excel Data Analysis Tool

DCOVA

A screenshot of the Microsoft Excel ribbon. The 'Data' tab is highlighted in blue, indicating it is selected. Below the ribbon, a portion of the worksheet is visible, showing data in columns A through E. The first row contains labels: '1 House Prices', '2 \$ 2,000,000', '3 \$ 500,000', '4 \$ 300,000', '5 \$ 100,000', and '6 \$ 100,000'. The second column (B) has a yellow background color.

1. Select Data.

2. Select Data Analysis.

3. Select Descriptive Statistics and click OK.

A screenshot of the Microsoft Excel ribbon with the 'Data' tab selected. In the bottom right corner of the ribbon, there is a 'Data Analysis' button. A callout arrow points from the text '3. Select Descriptive Statistics and click OK.' to this button. Below the ribbon, a screenshot of the worksheet shows the same data as before. A callout arrow also points from the 'Data Analysis' button to a 'Data Analysis' dialog box that is open over the worksheet. The dialog box lists various statistical tools, with 'Descriptive Statistics' highlighted with a blue selection bar.

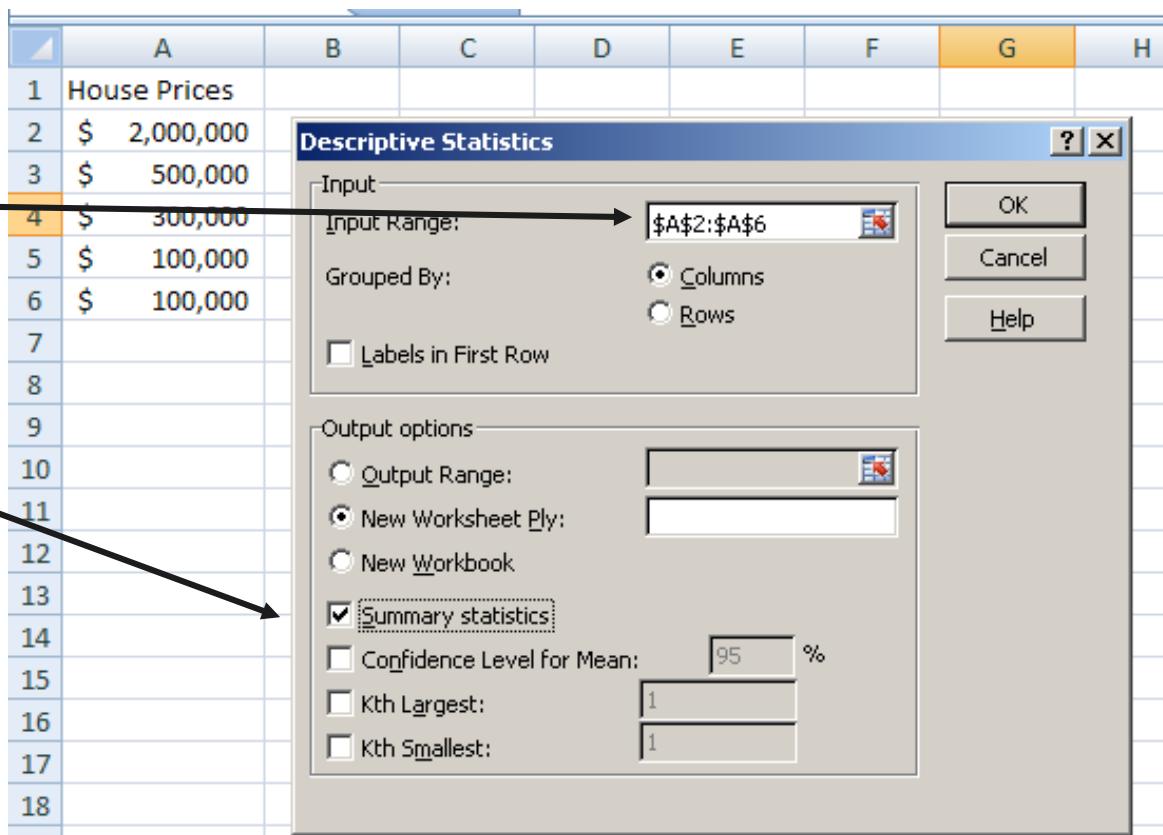
General Descriptive Stats Using Excel Data Analysis tool

DCOV A

4. Enter the cell range.

5. Check the Summary Statistics box.

6. Click OK



Excel output

DCOVA

Microsoft Excel
descriptive statistics output,
using the house price data:

House Prices:

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

| <i>House Prices</i> | |
|---------------------|-----------------|
| Mean | 600000 |
| Standard Error | 357770.8764 |
| Median | 300000 |
| Mode | 100000 |
| Standard Deviation | 800000 |
| Sample Variance | 640,000,000,000 |
| Kurtosis | 4.1301 |
| Skewness | 2.0068 |
| Range | 1900000 |
| Minimum | 100000 |
| Maximum | 2000000 |
| Sum | 3000000 |
| Count | 5 |

Minitab Output

DCOVA

Minitab descriptive statistics output using the house price data:

House Prices:

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

Descriptive Statistics: House Price

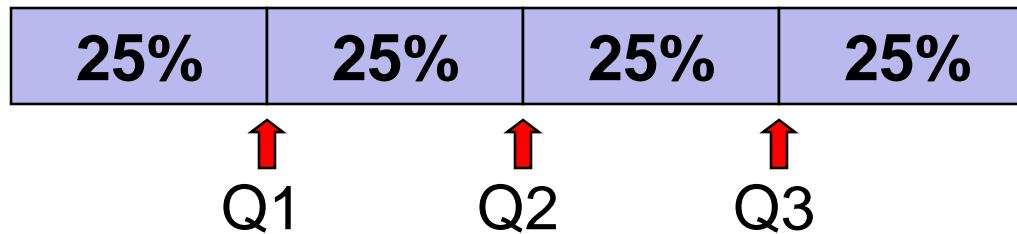
| | Total | | | | | | |
|-------------|-------|--------|---------|--------|-------------|---------|---------|
| Variable | Count | Mean | SE Mean | StDev | Variance | Sum | Minimum |
| House Price | 5 | 600000 | 357771 | 800000 | 6.40000E+11 | 3000000 | 100000 |

| Variable | Median | Maximum | Range | Mode | N for Mode | Skewness | Kurtosis |
|-------------|--------|---------|---------|--------|------------|----------|----------|
| House Price | 300000 | 2000000 | 1900000 | 100000 | 2 | 2.01 | 4.13 |

Quartile Measures

DCOV A

- Quartiles split the ranked data into 4 segments with an equal number of values per segment



- The first quartile, Q_1 , is the value for which 25% of the observations are smaller and 75% are larger
- Q_2 is the same as the median (50% of the observations are smaller and 50% are larger)
- Only 25% of the observations are greater than the third quartile

Quartile Measures: Locating Quartiles

DCOVA

Find a quartile by determining the value in the appropriate position in the ranked data, where

First quartile position: $Q_1 = (n+1)/4$ ranked value

Second quartile position: $Q_2 = (n+1)/2$ ranked value

Third quartile position: $Q_3 = 3(n+1)/4$ ranked value

where n is the number of observed values

Quartile Measures: Calculation Rules

DCOVA

- When calculating the ranked position use the following rules
 - If the result is a whole number then it is the ranked position to use
 - If the result is a fractional half (e.g. 2.5, 7.5, 8.5, etc.) then average the two corresponding data values.
 - If the result is not a whole number or a fractional half then round the result to the nearest integer to find the ranked position.

Quartile Measures: Locating Quartiles Example

DCOVA

Sample Data in Ordered Array: 11 12 13 16 16 17 18 21 22

(n = 9)



Q_1 is in the $(9+1)/4 = 2.5$ position of the ranked data
so use the value half way between the 2nd and 3rd values,

so
$$Q_1 = 12.5 = (12+13)/2$$

Q_1 and Q_3 are measures of non-central location
 Q_2 = median, is a measure of central tendency

Quartile Measures

Calculating The Quartiles: Example

DCOVA

Sample Data in Ordered Array: 11 12 13 16 16 17 18 21 22

(n = 9)

Q_1 is in the $(9+1)/4 = 2.5$ position of the ranked data,

$$\text{so } Q_1 = (12+13)/2 = 12.5$$

Q_2 is in the $(9+1)/2 = 5^{\text{th}}$ position of the ranked data,

$$\text{so } Q_2 = \text{median} = 16$$

Q_3 is in the $3(9+1)/4 = 7.5$ position of the ranked data,

$$\text{so } Q_3 = (18+21)/2 = 19.5$$

Q_1 and Q_3 are measures of non-central location

Q_2 = median, is a measure of central tendency

Quartile Measures: The Interquartile Range (IQR)

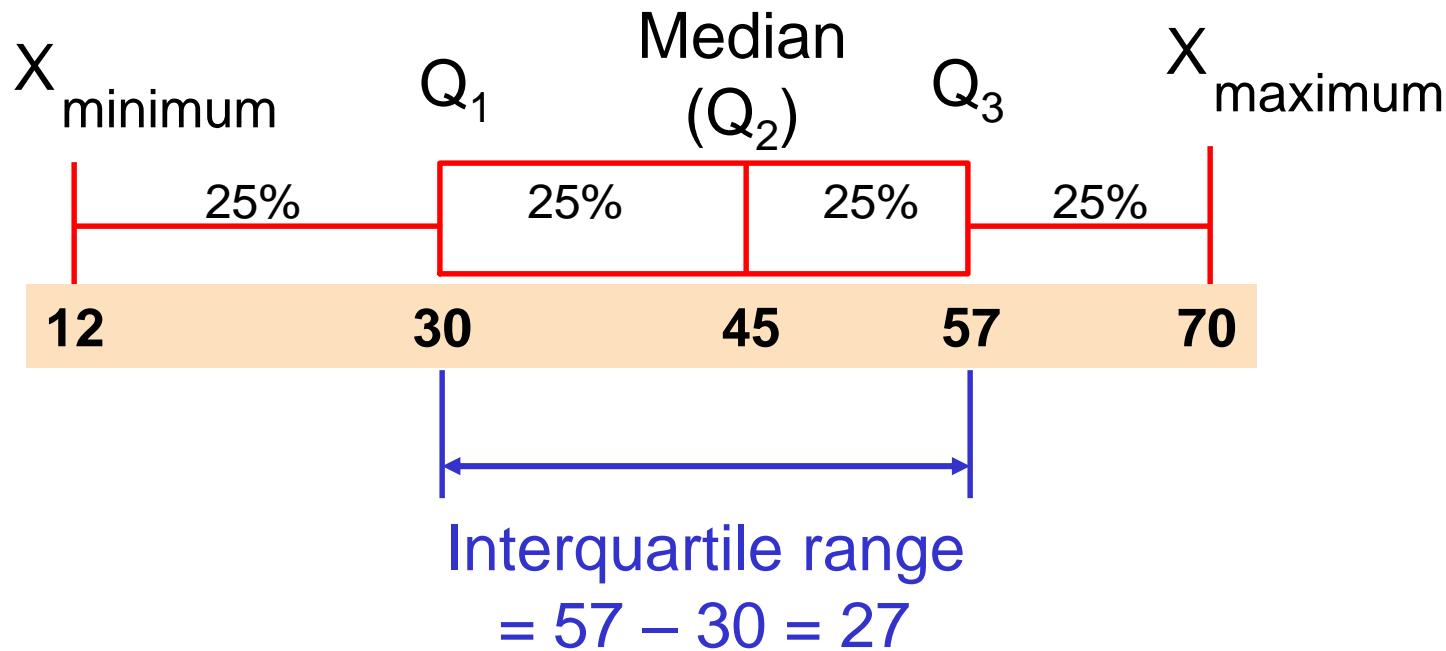
DCOVA

- The IQR is $Q_3 - Q_1$ and measures the spread in the middle 50% of the data
- The IQR is also called the midspread because it covers the middle 50% of the data
- The IQR is a measure of variability that is not influenced by outliers or extreme values
- Measures like Q_1 , Q_3 , and IQR that are not influenced by outliers are called resistant measures

Calculating The Interquartile Range

DCOVA

Example:



The Five Number Summary

DCOVA

The five numbers that help describe the center, spread and shape of data are:

- X_{smallest}
- First Quartile (Q_1)
- Median (Q_2)
- Third Quartile (Q_3)
- X_{largest}

Relationships among the five-number summary and distribution shape

DCOVA

| Left-Skewed | Symmetric | Right-Skewed |
|--|--|--|
| $\text{Median} - X_{\text{smallest}}$ > $X_{\text{largest}} - \text{Median}$ | $\text{Median} - X_{\text{smallest}}$ ≈ $X_{\text{largest}} - \text{Median}$ | $\text{Median} - X_{\text{smallest}}$ < $X_{\text{largest}} - \text{Median}$ |
| $Q_1 - X_{\text{smallest}}$ > $X_{\text{largest}} - Q_3$ | $Q_1 - X_{\text{smallest}}$ ≈ $X_{\text{largest}} - Q_3$ | $Q_1 - X_{\text{smallest}}$ < $X_{\text{largest}} - Q_3$ |
| $\text{Median} - Q_1$ > $Q_3 - \text{Median}$ | $\text{Median} - Q_1$ ≈ $Q_3 - \text{Median}$ | $\text{Median} - Q_1$ < $Q_3 - \text{Median}$ |

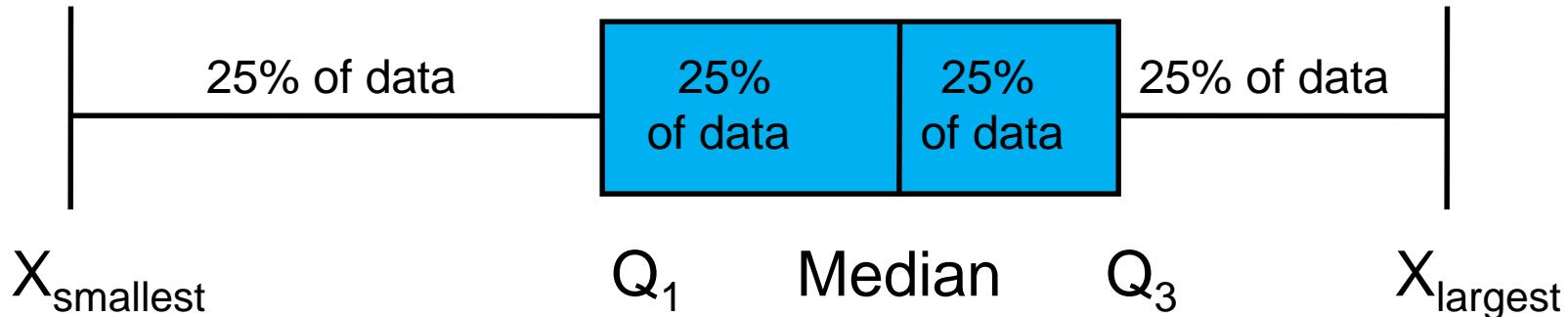
Five Number Summary and The Boxplot

DCOVA

- The Boxplot: A Graphical display of the data based on the five-number summary:

X_{smallest} -- Q_1 -- Median -- Q_3 -- X_{largest}

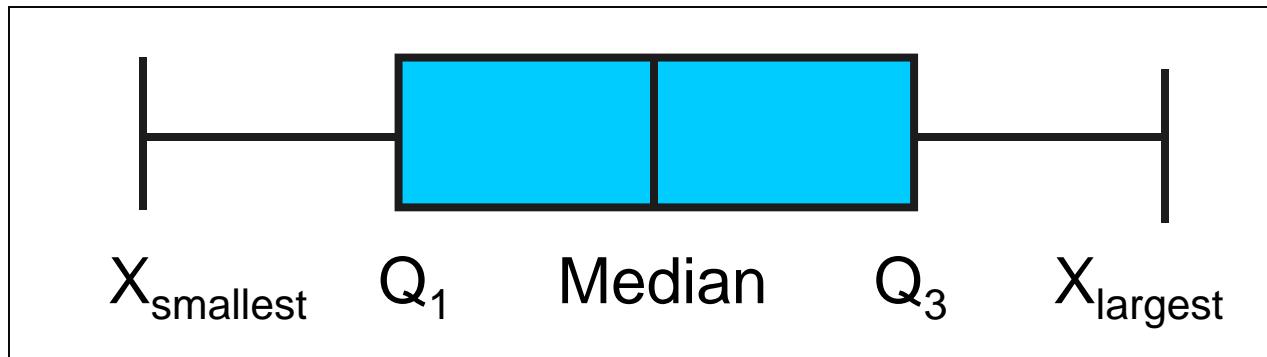
Example:



Five Number Summary: Shape of Boxplots

DCOVA

- If data are symmetric around the median then the box and central line are centered between the endpoints

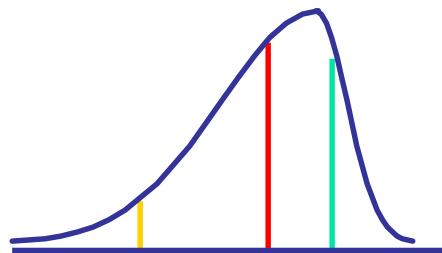


- A Boxplot can be shown in either a vertical or horizontal orientation

Distribution Shape and The Boxplot

DCOVA

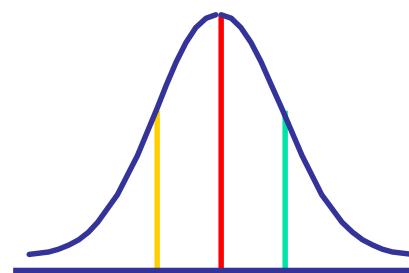
Left-Skewed



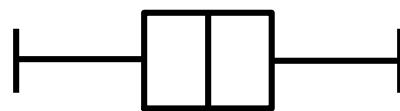
Q_1 Q_2 Q_3



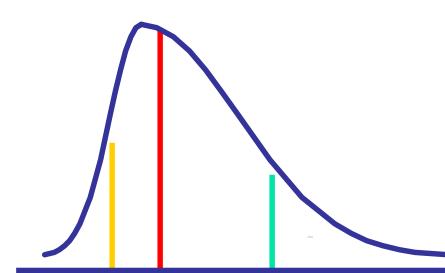
Symmetric



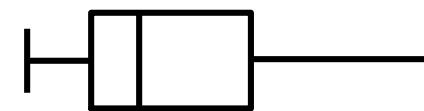
Q_1 Q_2 Q_3



Right-Skewed



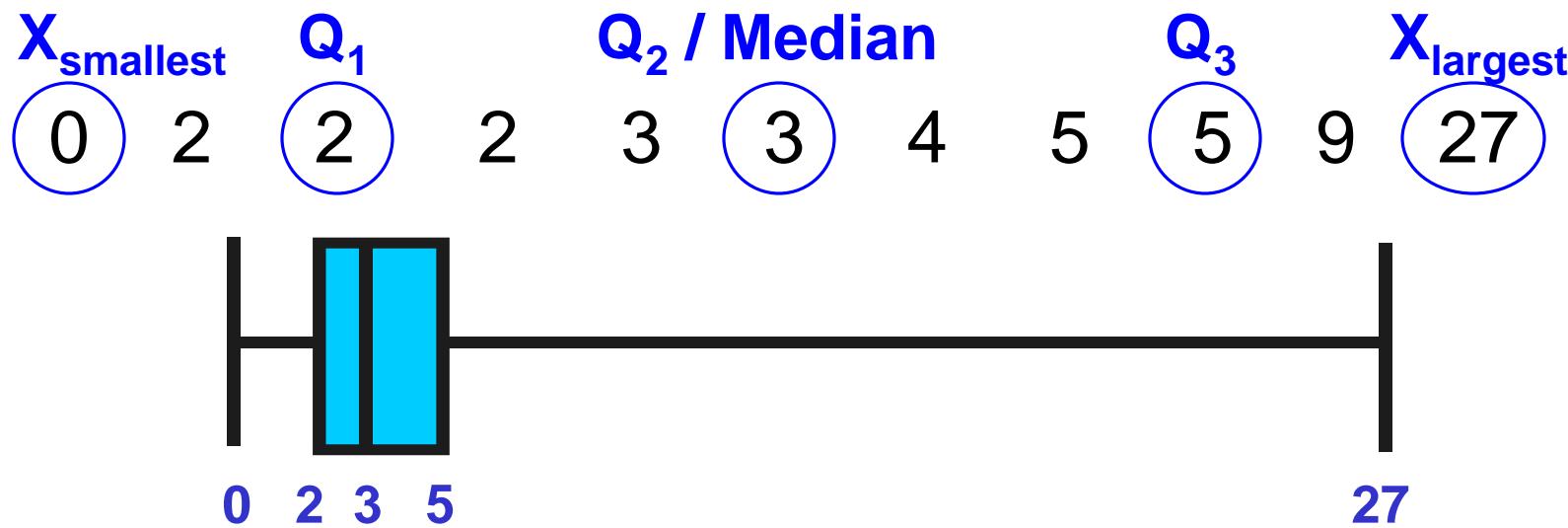
Q_1 Q_2 Q_3



Boxplot Example

DCOVA

- Below is a Boxplot for the following data:



- The data are right skewed, as the plot depicts

Numerical Descriptive Measures for a Population

DCOVA

- Descriptive statistics discussed previously described a *sample*, not the *population*.
- Summary measures describing a population, called **parameters**, are denoted with Greek letters.
- Important population parameters are the population mean, variance, and standard deviation.

Numerical Descriptive Measures for a Population: The mean μ

DCOVA

- The population mean is the sum of the values in the population divided by the population size, N

$$\mu = \frac{\sum_{i=1}^N X_i}{N} = \frac{X_1 + X_2 + \cdots + X_N}{N}$$

Where μ = population mean

N = population size

X_i = i^{th} value of the variable X

Numerical Descriptive Measures

For A Population: The Variance σ^2

DCOVA

- Average of squared deviations of values from the mean

- Population variance:

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N}$$

Where μ = population mean

N = population size

X_i = i^{th} value of the variable X

Numerical Descriptive Measures For A Population: The Standard Deviation σ

DCOV Δ

- Most commonly used measure of variation
- Shows variation about the mean
- Is the square root of the population variance
- Has the same units as the original data

- Population standard deviation:

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (X_i - \mu)^2}{N}}$$

Sample statistics versus population parameters

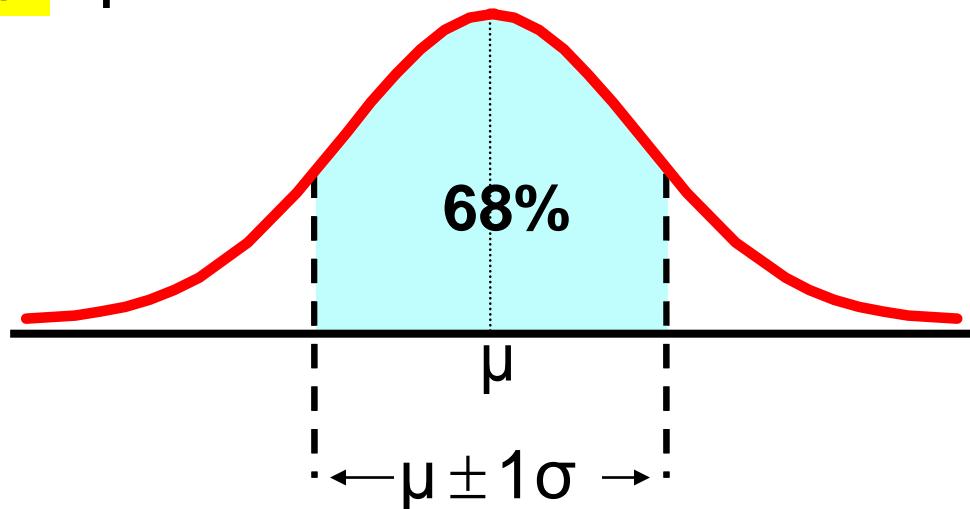
DCOVA

| Measure | Population Parameter | Sample Statistic |
|--------------------|----------------------|------------------|
| Mean | μ | \bar{X} |
| Variance | σ^2 | S^2 |
| Standard Deviation | σ | S |

The Empirical Rule

DCOV A

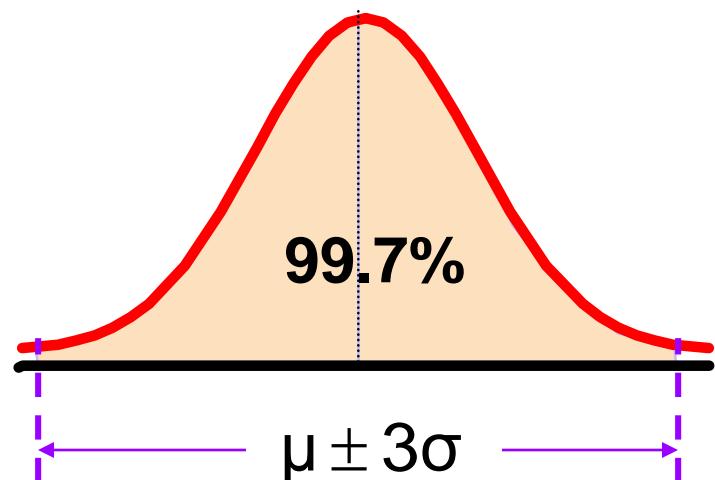
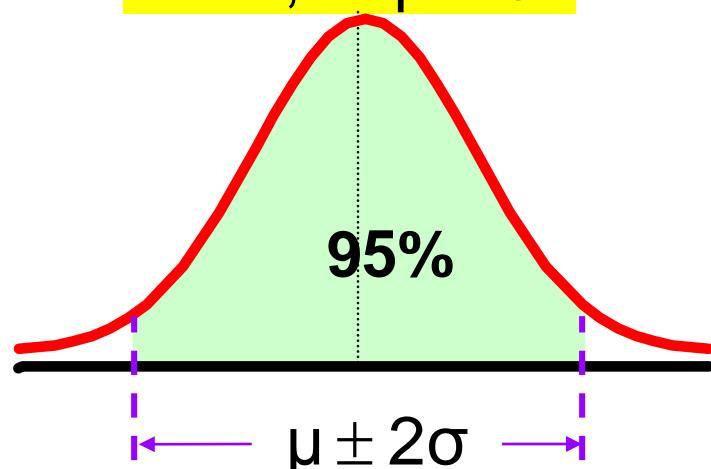
- The empirical rule approximates the variation of data in a bell-shaped distribution
- Approximately 68% of the data in a bell shaped distribution is within 1 standard deviation of the mean or $\mu \pm 1\sigma$



The Empirical Rule

DCOV A

- Approximately 95% of the data in a bell-shaped distribution lies within two standard deviations of the mean, or $\mu \pm 2\sigma$
- Approximately 99.7% of the data in a bell-shaped distribution lies within three standard deviations of the mean, or $\mu \pm 3\sigma$



The Empirical Rule Example

DCOVA

- Suppose that the variable Math SAT scores is bell-shaped with a mean of 500 and a standard deviation of 90. Then,
 - 68% of all test takers scored between 410 and 590 (500 ± 90).
 - 95% of all test takers scored between 320 and 680 (500 ± 180).
 - 99.7% of all test takers scored between 230 and 770 (500 ± 270).

Chebyshev Rule – used when you do not know the distribution

DCOV A

- Regardless of how the data are distributed, at least $(1 - 1/k^2) \times 100\%$ of the values will fall within k standard deviations of the mean (for $k > 1$)
 - Examples:

| At least | Within |
|--|-----------------------------|
| $(1 - 1/2^2) \times 100\% = 75\%$ | $k=2$ ($\mu \pm 2\sigma$) |
| $(1 - 1/3^2) \times 100\% = 88.89\%$ | $k=3$ ($\mu \pm 3\sigma$) |

We Discuss Two Measures Of The Relationship Between Two Numerical Variables

- Scatter plots allow you to visually examine the relationship between two numerical variables and now we will discuss two quantitative measures of such relationships.
- The Covariance
- The Coefficient of Correlation

The Covariance

DCOVA

- The covariance measures the strength of the linear relationship between **two numerical variables** (X & Y)
- The sample covariance:

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

- Only concerned with the strength of the relationship
- No causal effect is implied

Interpreting Covariance

DCOV A

- **Covariance** between two variables:

$\text{cov}(X, Y) > 0 \rightarrow X \text{ and } Y \text{ tend to move in the same direction}$

$\text{cov}(X, Y) < 0 \rightarrow X \text{ and } Y \text{ tend to move in opposite directions}$

$\text{cov}(X, Y) = 0 \rightarrow X \text{ and } Y \text{ are independent}$

- The covariance has a major flaw:

- It is not possible to determine the relative strength of the relationship from the size of the covariance

Coefficient of Correlation

DCOV A

- Measures the relative strength of the linear relationship between two numerical variables
- Sample coefficient of correlation:

$$r = \frac{\text{cov}(X, Y)}{S_x S_y}$$

where

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

$$S_x = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$$

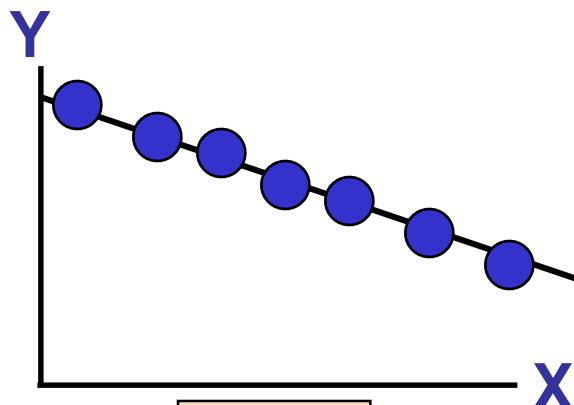
$$S_y = \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}}$$

Features of the Coefficient of Correlation

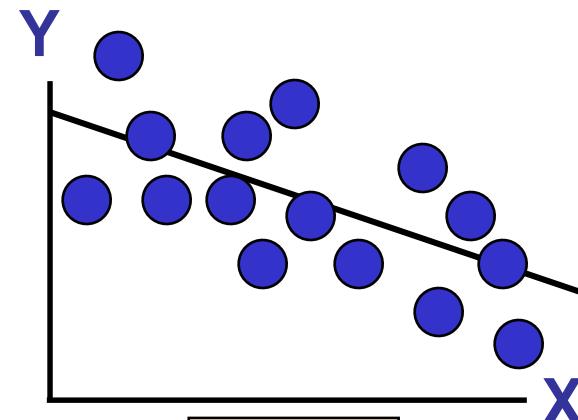
DCOV A

- The population coefficient of correlation is referred as ρ .
- The sample coefficient of correlation is referred to as r .
- Either ρ or r have the following features:
 - Unit free
 - Range between -1 and 1
 - The closer to -1 , the stronger the negative linear relationship
 - The closer to 1 , the stronger the positive linear relationship
 - The closer to 0 , the weaker the linear relationship
 - 0 means no correlation
 - -1 and 1 means perfect correlation

Scatter Plots of Sample Data with Various Coefficients of Correlation

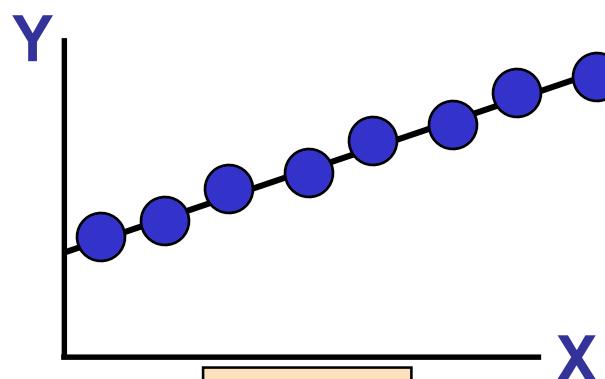


$$r = -1$$

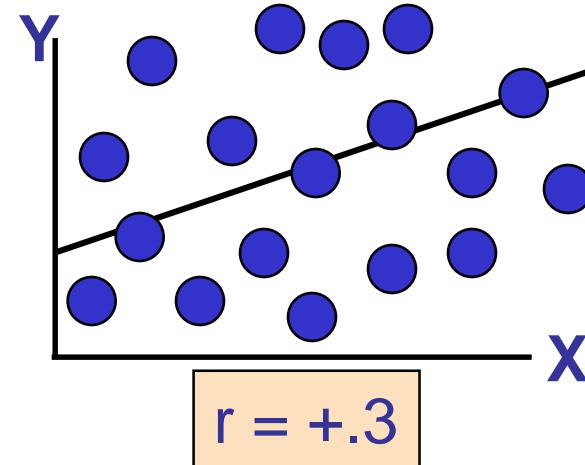


$$r = -.6$$

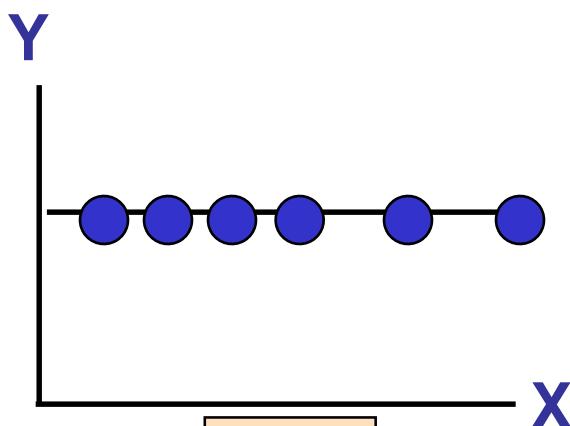
DCOVA



$$r = +1$$



$$r = +.3$$



$$r = 0$$

The Coefficient of Correlation Using Microsoft Excel Function

DCOVA

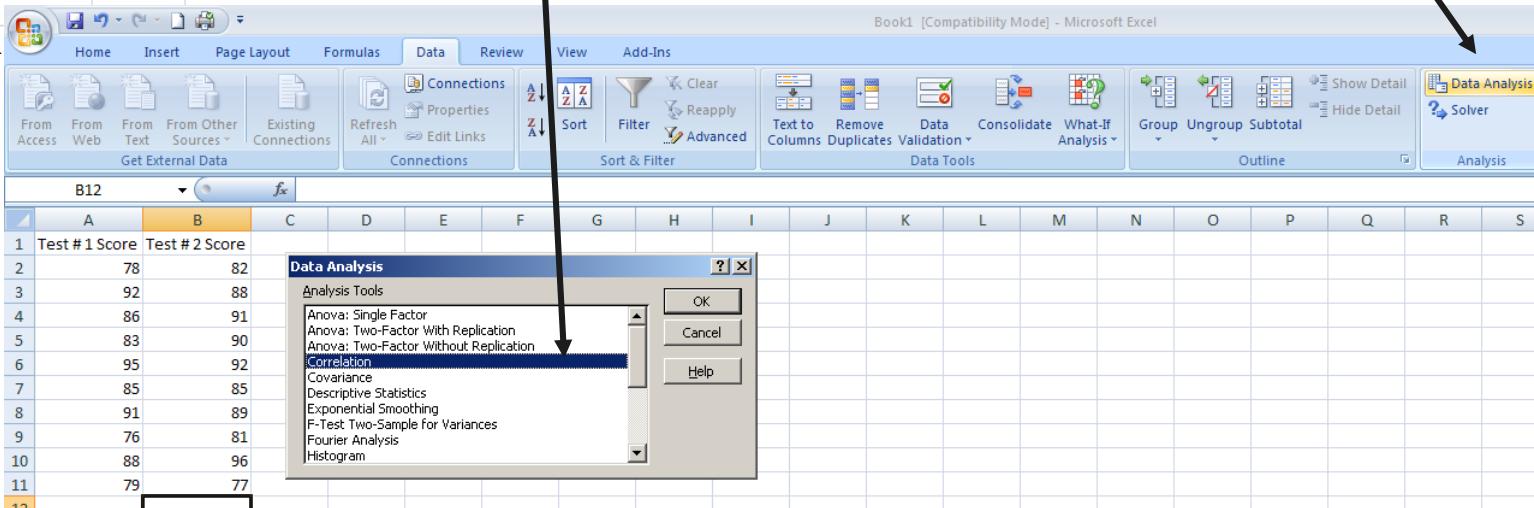
| Test #1 Score | Test #2 Score | | Correlation Coefficient |
|---------------|---------------|--|--------------------------------|
| 78 | 82 | | 0.7332 =CORREL(A2:A11,B2:B11) |
| 92 | 88 | | |
| 86 | 91 | | |
| 83 | 90 | | |
| 95 | 92 | | |
| 85 | 85 | | |
| 91 | 89 | | |
| 76 | 81 | | |
| 88 | 96 | | |
| 79 | 77 | | |

The Coefficient of Correlation Using Microsoft Excel Data Analysis Tool

| | A | B | C | D | E |
|----|---------------|---------------|---|---|---|
| 1 | Test #1 Score | Test #2 Score | | | |
| 2 | 78 | 82 | | | |
| 3 | 92 | 88 | | | |
| 4 | 86 | 91 | | | |
| 5 | 83 | 90 | | | |
| 6 | 95 | 92 | | | |
| 7 | 85 | 85 | | | |
| 8 | 91 | 89 | | | |
| 9 | 76 | 81 | | | |
| 10 | 88 | 96 | | | |
| 11 | 79 | 77 | | | |

1. Select Data
2. Choose Data Analysis
3. Choose Correlation & Click OK

DCOVA



The Coefficient of Correlation

Using Microsoft Excel

DCOV A

A screenshot of Microsoft Excel showing a data table in the background and the 'Correlation' dialog box in the foreground. The data table has columns 'Test # 1 Score' and 'Test # 2 Score' with values from 78 to 77. The dialog box shows the input range as \$A\$1:\$B\$11, grouped by columns, and labels in the first row checked. The output options are set to a new worksheet.

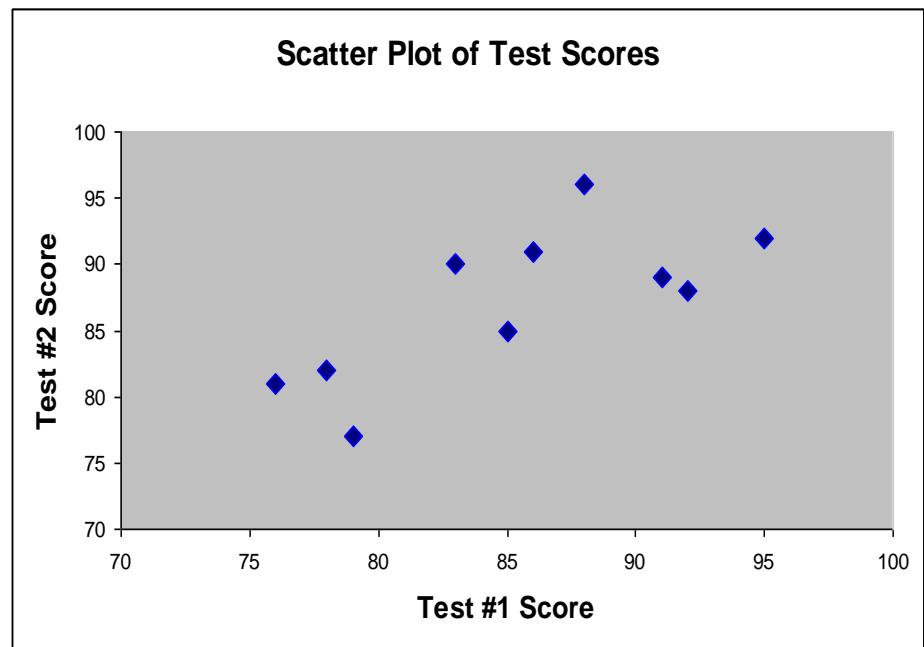
4. Input data range and select appropriate options
5. Click OK to get output

| | A | B | C |
|---|----------------|----------------|----------------|
| 1 | | Test # 1 Score | Test # 2 Score |
| 2 | Test # 1 Score | | 1 |
| 3 | Test # 2 Score | 0.733243705 | 1 |
| 4 | | | |

Interpreting the Coefficient of Correlation Using Microsoft Excel

DCOVA

- $r = .733$
- There is a relatively strong positive linear relationship between test score #1 and test score #2.
- Students who scored high on the first test tended to score high on second test.



Pitfalls in Numerical Descriptive Measures

DCOVA

- Data analysis is objective
 - Should report the summary measures that best describe and communicate the important aspects of the data set

- Data interpretation is subjective
 - Should be done in fair, neutral and clear manner

Ethical Considerations

DCOVA

Numerical descriptive measures:

- Should document both good and bad results
- Should be presented in a fair, objective and neutral manner
- Should not use inappropriate summary measures to distort facts



Chapter Summary

In this chapter we discussed

- Measures of central tendency
 - Mean, median, mode, geometric mean
- Measures of variation
 - Range, interquartile range, variance and standard deviation, coefficient of variation, Z-scores
- The shape of distributions
 - Skewness & Kurtosis
- Describing data using the 5-number summary
 - Boxplots

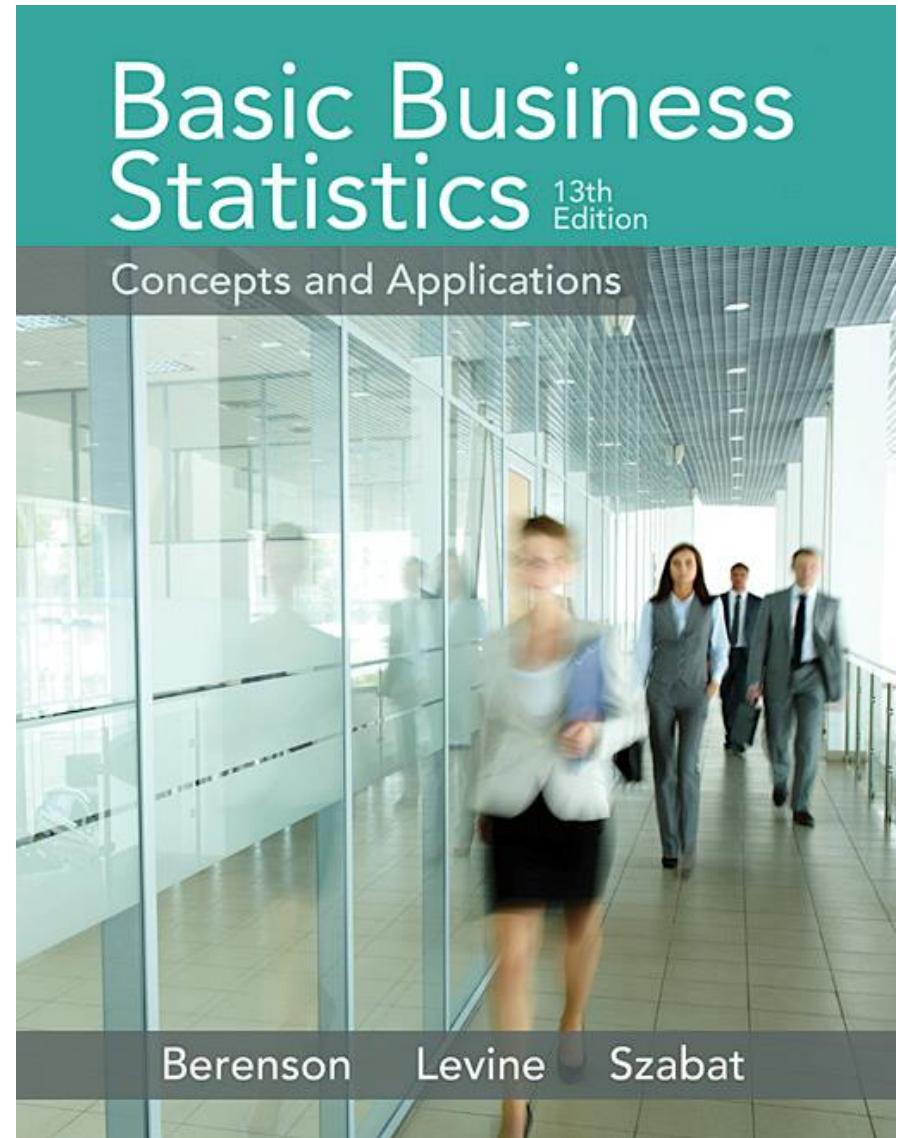
Chapter Summary

(continued)

- Covariance and correlation coefficient
- Pitfalls in numerical descriptive measures and ethical considerations

Chapter 4

Basic Probability



Learning Objectives

In this chapter, you learn:

- Basic probability concepts
- About conditional probability
- To use Bayes' Theorem to revise probabilities
- Various counting rules

Basic Probability Concepts

- **Probability** – the chance that an uncertain event will occur (always between 0 and 1)
- **Impossible Event** – an event that has no chance of occurring (probability = 0)
- **Certain Event** – an event that is sure to occur (probability = 1)

Assessing Probability

There are three approaches to assessing the probability of an uncertain event:

1. *a priori* -- based on prior knowledge of the process

probability of occurrence = $\frac{X}{T} = \frac{\text{number of ways in which the event occurs}}{\text{total number of possible outcomes}}$

Assuming
all
outcomes
are equally
likely

2. empirical probability

probability of occurrence = $\frac{\text{number of ways in which the event occurs}}{\text{total number of possible outcomes}}$

3. subjective probability

based on a combination of an individual's past experience, personal opinion, and analysis of a particular situation

Example of *a priori* probability

When randomly selecting a day from the year 2014 what is the probability the day is in January?

$$\text{Probability of Day In January} = \frac{X}{T} = \frac{\text{number of days in January}}{\text{total number of days in 2014}}$$

$$\frac{X}{T} = \frac{\text{31 days in January}}{\text{365 days in 2013}} = \frac{31}{365}$$

Example of empirical probability

Find the probability of selecting a male taking statistics from the population described in the following table:

| | Taking Stats | Not Taking Stats | Total |
|--------|--------------|------------------|-------|
| Male | 84 | 145 | 229 |
| Female | 76 | 134 | 210 |
| Total | 160 | 279 | 439 |

$$\text{Probability of male taking stats} = \frac{\text{number of males taking stats}}{\text{total number of people}} = \frac{84}{439} = 0.191$$

Subjective probability

- Subjective probability may differ from person to person
 - A media development team assigns a 60% probability of success to its new ad campaign.
 - The chief media officer of the company is less optimistic and assigns a 40% of success to the same campaign
- The assignment of a subjective probability is based on a person's experiences, opinions, and analysis of a particular situation
- Subjective probability is useful in situations when an empirical or a priori probability cannot be computed

Events

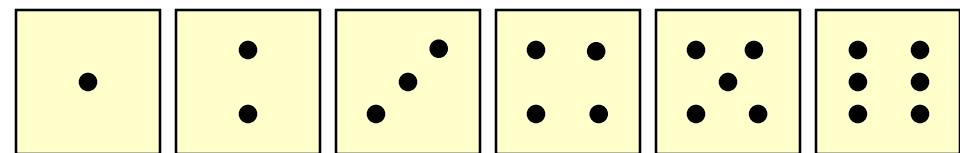
Each possible outcome of a variable is an **event**.

- **Simple event**
 - An event described by a single characteristic
 - e.g., A day in January from all days in 2014
- **Joint event**
 - An event described by two or more characteristics
 - e.g. A day in January that is also a Wednesday from all days in 2014
- **Complement of an event A (denoted A')**
 - All events that are not part of event A
 - e.g., All days from 2014 that are not in January

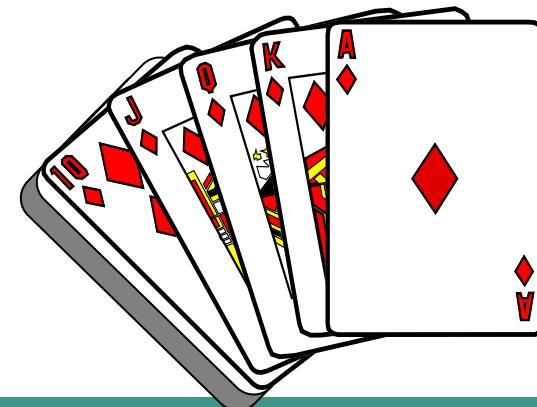
Sample Space

The **Sample Space** is the collection of all possible events

e.g. All 6 faces of a die:

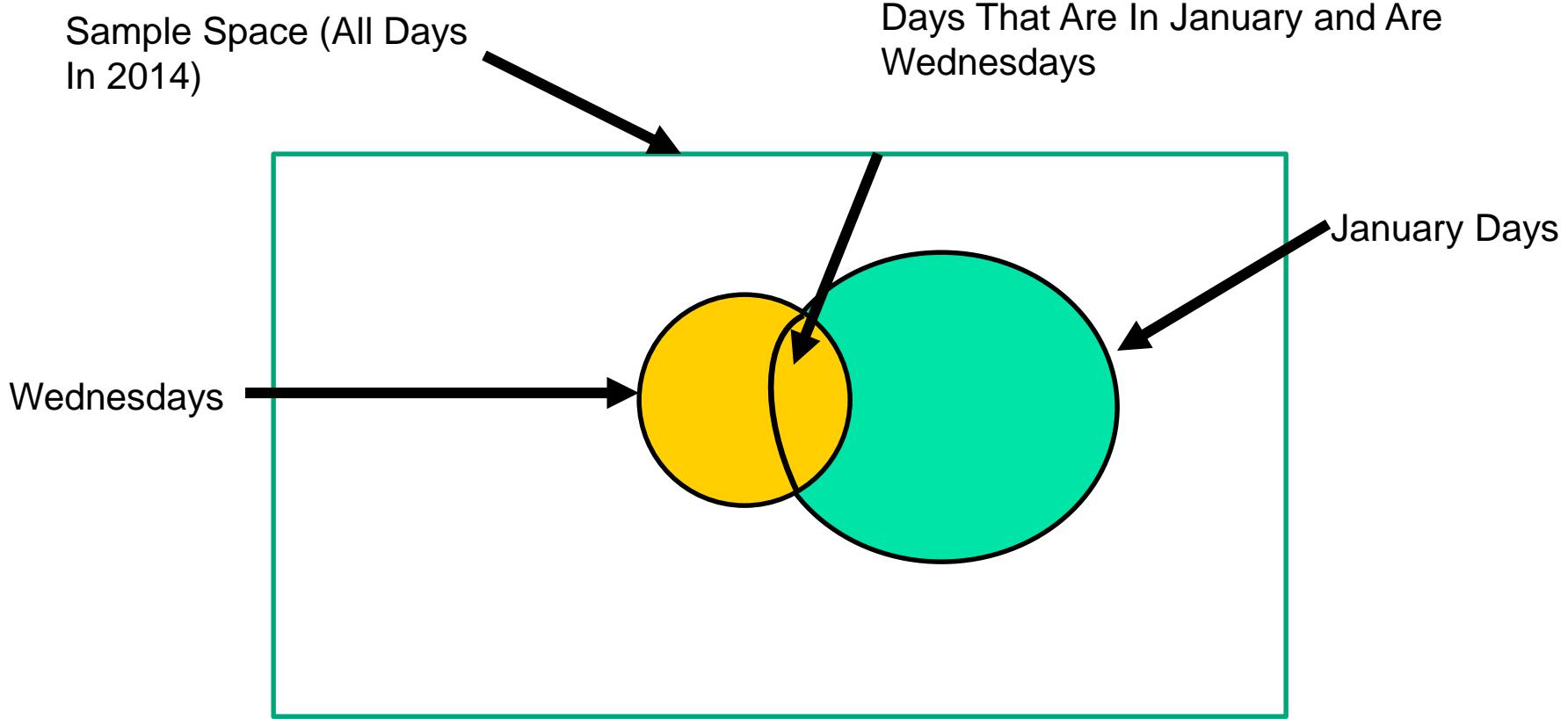


e.g. All 52 cards of a bridge deck:



Organizing & Visualizing Events

- Venn Diagram For All Days In 2014



Organizing & Visualizing Events

(continued)

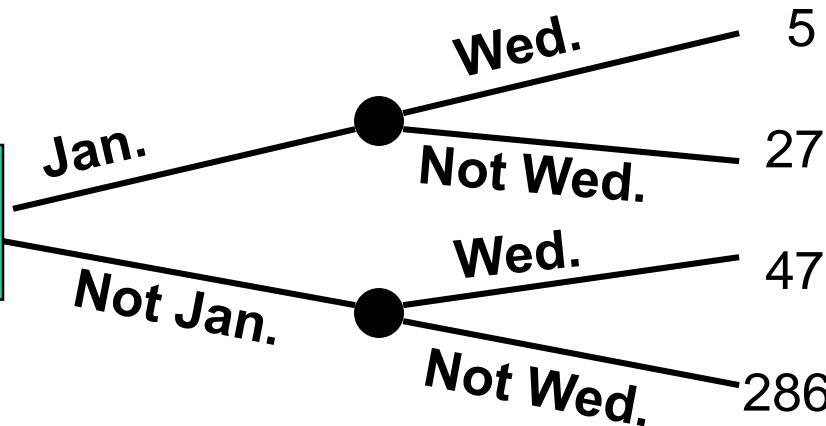
- Contingency Tables -- For All Days in 2014

| | Jan. | Not Jan. | Total |
|----------|------|----------|-------|
| Wed. | 5 | 47 | 52 |
| Not Wed. | 27 | 286 | 313 |
| Total | 32 | 333 | 365 |

- Decision Trees

Sample Space

All Days
In 2014



Total Number Of Sample Space Outcomes

Definition: Simple Probability

- Simple Probability refers to the probability of a simple event.
 - ex. $P(\text{Jan.})$
 - ex. $P(\text{Wed.})$

| | Jan. | Not Jan. | Total |
|----------|------|----------|-------|
| Wed. | 5 | 47 | 52 |
| Not Wed. | 27 | 286 | 313 |
| Total | 32 | 333 | 365 |

$$P(\text{Wed.}) = 52 / 365$$

$$P(\text{Jan.}) = 32 / 365$$

Definition: Joint Probability

- Joint Probability refers to the probability of an occurrence of two or more events (joint event).
 - ex. $P(\text{Jan. and Wed.})$
 - ex. $P(\text{Not Jan. and Not Wed.})$

| | Jan. | Not Jan. | Total |
|----------|------|----------|-------|
| Wed. | 5 | 47 | 52 |
| Not Wed. | 27 | 286 | 313 |
| Total | 32 | 333 | 365 |

$$P(\text{Not Jan. and Not Wed.}) = 286 / 365$$

$$P(\text{Jan. and Wed.}) = 5 / 365$$

Mutually Exclusive Events

- Mutually exclusive events
 - Events that cannot occur simultaneously

Example: Randomly choosing a day from 2014

A = day in January; B = day in February

- Events A and B are mutually exclusive

Collectively Exhaustive Events

- Collectively exhaustive events
 - One of the events must occur
 - The set of events covers the entire sample space

Example: Randomly choose a day from 2014

A = Weekday; B = Weekend;
C = January; D = Spring;

- Events A, B, C and D are collectively exhaustive (but not mutually exclusive – a weekday can be in January or in Spring)
- Events A and B are collectively exhaustive and also mutually exclusive

Computing Joint and Marginal Probabilities

- The probability of a joint event, A and B:

$$P(A \text{ and } B) = \frac{\text{number of outcomes satisfying } A \text{ and } B}{\text{total number of elementary outcomes}}$$

- Computing a marginal (or simple) probability:

$$P(A) = P(A \text{ and } B_1) + P(A \text{ and } B_2) + \cdots + P(A \text{ and } B_k)$$

- Where B_1, B_2, \dots, B_k are k mutually exclusive and collectively exhaustive events

Joint Probability Example

P(Jan. and Wed.)

$$= \frac{\text{number of days that are in Jan. and are Wed.}}{\text{total number of days in 2013}} = \frac{5}{365}$$

| | Jan. | Not Jan. | Total |
|----------|------|----------|-------|
| Wed. | 5 | 47 | 52 |
| Not Wed. | 27 | 286 | 313 |
| Total | 32 | 333 | 365 |

Marginal Probability Example

P(Wed.)

$$= P(\text{Jan. and Wed.}) + P(\text{Not Jan. and Wed.}) = \frac{5}{365} + \frac{48}{365} = \frac{53}{365}$$

| | Jan. | Not Jan. | Total |
|----------|------|----------|-------|
| Wed. | 5 | 48 | 52 |
| Not Wed. | 27 | 286 | 313 |
| Total | 31 | 334 | 365 |

Probability Summary So Far

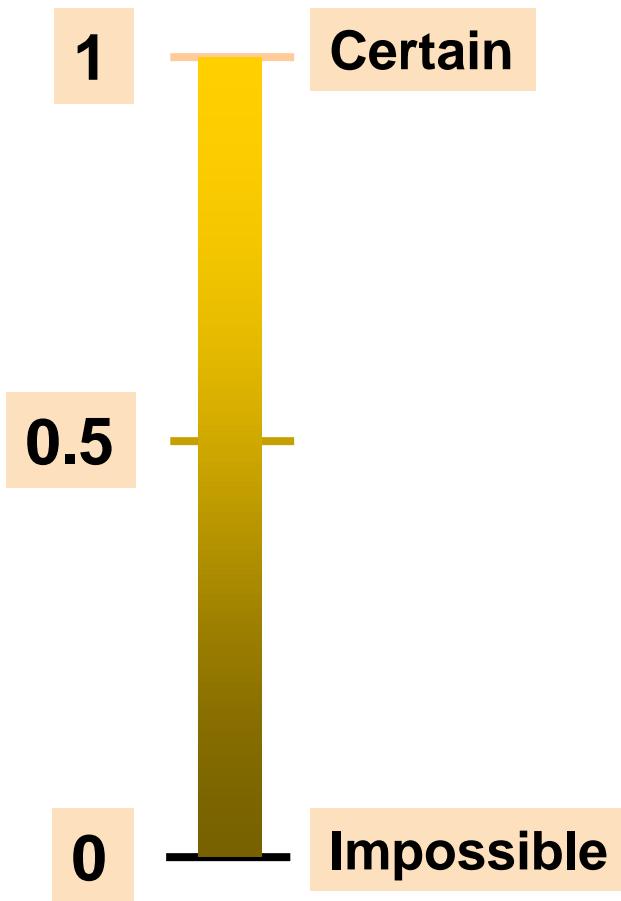
- Probability is the numerical measure of the likelihood that an event will occur
- The probability of any event must be between 0 and 1, inclusively

$$0 \leq P(A) \leq 1 \quad \text{For any event A}$$

- The sum of the probabilities of all mutually exclusive and collectively exhaustive events is 1

$$P(A) + P(B) + P(C) = 1$$

If A, B, and C are mutually exclusive and collectively exhaustive



General Addition Rule

General Addition Rule:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

If A and B are mutually exclusive, then

$P(A \text{ and } B) = 0$, so the rule can be simplified:

$$P(A \text{ or } B) = P(A) + P(B)$$

For mutually exclusive events A and B

General Addition Rule Example

$$P(\text{Jan. or Wed.}) = P(\text{Jan.}) + P(\text{Wed.}) - P(\text{Jan. and Wed.})$$

$$= 32/365 + 52/365 - 5/365 = 79/365$$

| | Jan. | Not Jan. | Total |
|----------|------|----------|-------|
| Wed. | 5 | 47 | 52 |
| Not Wed. | 27 | 286 | 313 |
| Total | 32 | 333 | 365 |

Don't count
the five
Wednesdays
in January
twice!

Computing Conditional Probabilities

- A conditional probability is the probability of one event, given that another event has occurred:

$$P(A | B) = \frac{P(A \text{ and } B)}{P(B)}$$



The conditional probability of A given that B has occurred

$$P(B | A) = \frac{P(A \text{ and } B)}{P(A)}$$



The conditional probability of B given that A has occurred

Where $P(A \text{ and } B)$ = joint probability of A and B

$P(A)$ = marginal or simple probability of A

$P(B)$ = marginal or simple probability of B

Conditional Probability Example

- Of the cars on a used car lot, 70% have air conditioning (AC) and 40% have a GPS. 20% of the cars have both.
- What is the probability that a car has a GPS, given that it has AC ?
i.e., we want to find $P(\text{GPS} | \text{AC})$

Conditional Probability Example

(continued)

- Of the cars on a used car lot, **70%** have air conditioning (AC) and **40%** have a GPS and **20%** of the cars have both.

| | GPS | No GPS | Total |
|-------|-----|--------|-------|
| AC | 0.2 | 0.5 | 0.7 |
| No AC | 0.2 | 0.1 | 0.3 |
| Total | 0.4 | 0.6 | 1.0 |

$$P(\text{GPS} | \text{AC}) = \frac{P(\text{GPS and AC})}{P(\text{AC})} = \frac{0.2}{0.7} = 0.2857$$

Conditional Probability Example

(continued)

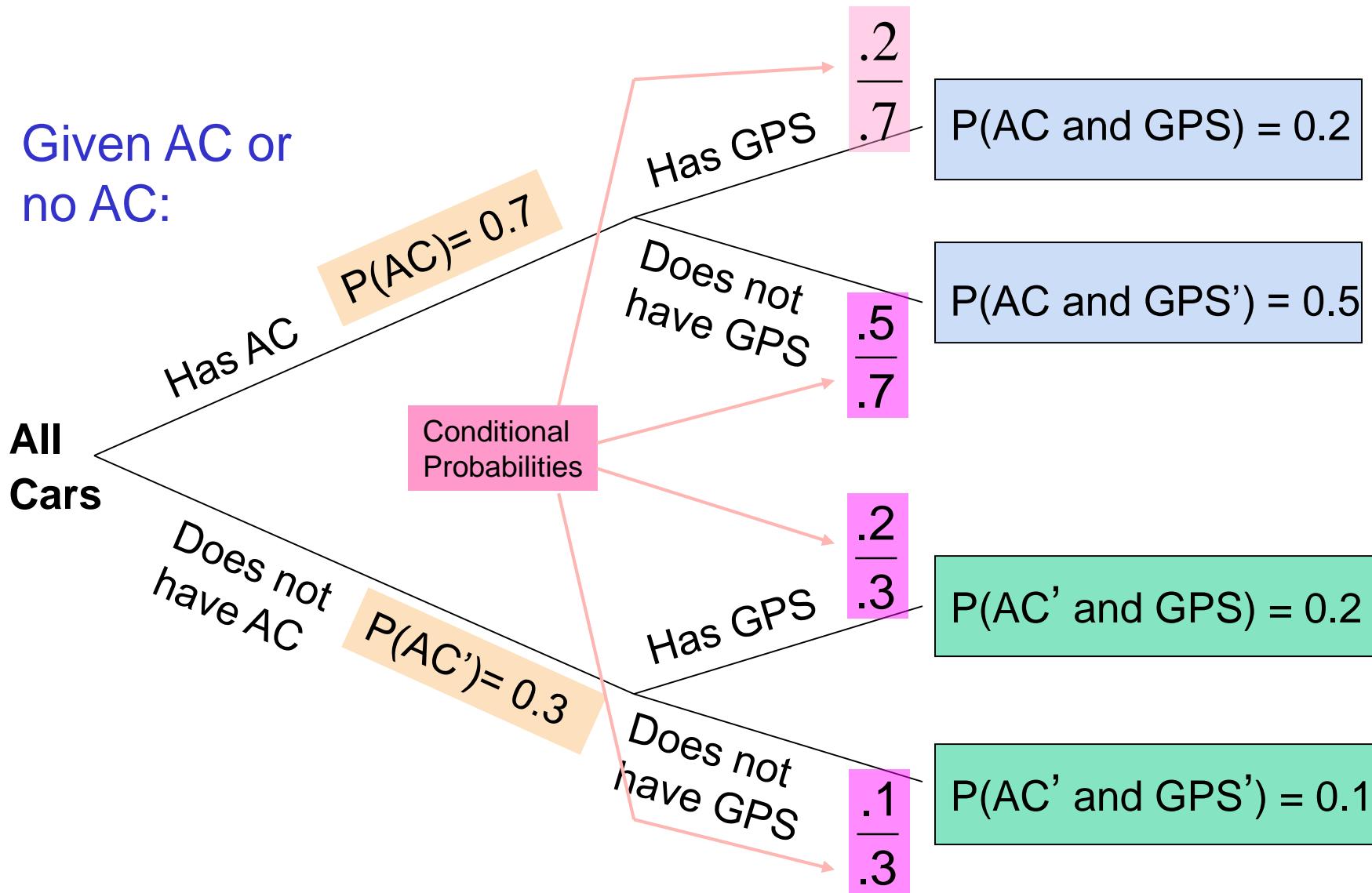
- Given AC, we only consider the top row (70% of the cars). Of these, 20% have a GPS. 20% of 70% is about 28.57%.

| | GPS | No GPS | Total |
|-------|-----|--------|-------|
| AC | 0.2 | 0.5 | 0.7 |
| No AC | 0.2 | 0.1 | 0.3 |
| Total | 0.4 | 0.6 | 1.0 |

$$P(\text{GPS} | \text{AC}) = \frac{P(\text{GPS and AC})}{P(\text{AC})} = \frac{0.2}{0.7} = 0.2857$$

Using Decision Trees

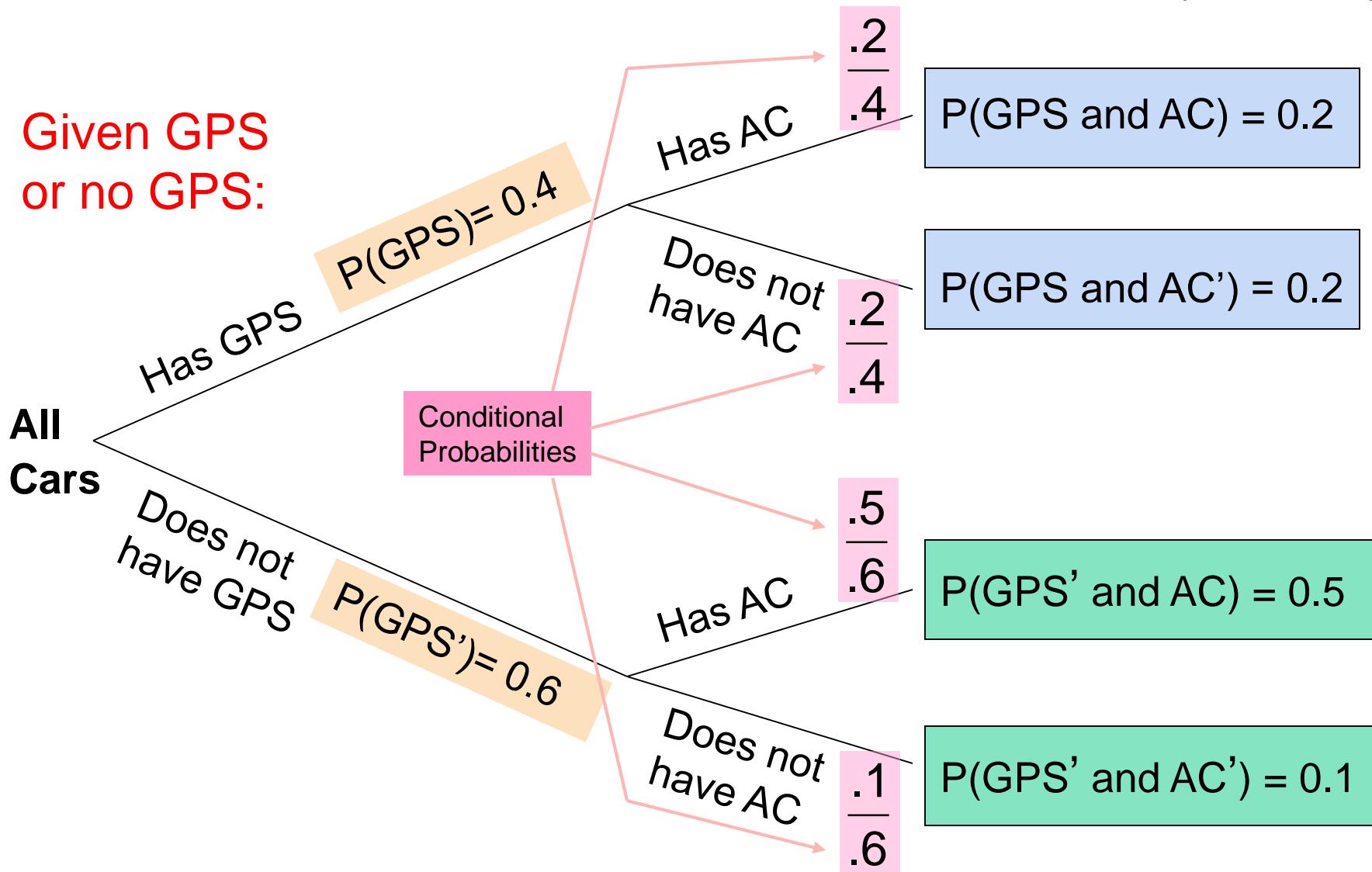
Given AC or
no AC:



Using Decision Trees

(continued)

Given GPS
or no GPS:



Independence

- Two events are **independent** if and only if:

$$P(A | B) = P(A)$$

- Events A and B are independent when the probability of one event is not affected by the fact that the other event has occurred

Multiplication Rules

- Multiplication rule for two events A and B:

$$P(A \text{ and } B) = P(A | B)P(B)$$

Note: If A and B are independent, then $P(A | B) = P(A)$ and the multiplication rule simplifies to

$$P(A \text{ and } B) = P(A)P(B)$$

Marginal Probability

- Marginal probability for event A:

$$P(A) = P(A | B_1)P(B_1) + P(A | B_2)P(B_2) + \cdots + P(A | B_k)P(B_k)$$

- Where B_1, B_2, \dots, B_k are k mutually exclusive and collectively exhaustive events

Bayes' Theorem

- Bayes' Theorem is used to revise previously calculated probabilities based on new information.
- Developed by Thomas Bayes in the 18th Century.
- It is an extension of conditional probability.

Bayes' Theorem

$$P(B_i | A) = \frac{P(A | B_i)P(B_i)}{P(A | B_1)P(B_1) + P(A | B_2)P(B_2) + \cdots + P(A | B_k)P(B_k)}$$

- where:

B_i = i^{th} event of k mutually exclusive and collectively exhaustive events

A = new event that might impact $P(B_i)$

Bayes' Theorem Example

- A drilling company has estimated a 40% chance of striking oil for their new well.
- A detailed test has been scheduled for more information. Historically, 60% of successful wells have had detailed tests, and 20% of unsuccessful wells have had detailed tests.
- Given that this well has been scheduled for a detailed test, what is the probability that the well will be successful?



Bayes' Theorem Example

(continued)

- Let S = successful well
 U = unsuccessful well
- $P(S) = 0.4$, $P(U) = 0.6$ (prior probabilities)
- Define the detailed test event as D
- Conditional probabilities:
$$P(D|S) = 0.6 \qquad P(D|U) = 0.2$$
- Goal is to find $P(S|D)$



Bayes' Theorem Example

(continued)

Apply Bayes' Theorem:

$$\begin{aligned} P(S|D) &= \frac{P(D|S)P(S)}{P(D|S)P(S)+P(D|U)P(U)} \\ &= \frac{(0.6)(0.4)}{(0.6)(0.4)+(0.2)(0.6)} \\ &= \frac{0.24}{0.24+0.12} = 0.667 \end{aligned}$$



So the revised probability of success, given that this well has been scheduled for a detailed test, is 0.667

Bayes' Theorem Example

(continued)

- Given the detailed test, the revised probability of a successful well has risen to 0.667 from the original estimate of 0.4



| Event | Prior Prob. | Conditional Prob. | Joint Prob. | Revised Prob. |
|-------------------|-------------|-------------------|---------------------|---------------------|
| S (successful) | 0.4 | 0.6 | $(0.4)(0.6) = 0.24$ | $0.24/0.36 = 0.667$ |
| U (unsuccessful) | 0.6 | 0.2 | $(0.6)(0.2) = 0.12$ | $0.12/0.36 = 0.333$ |
| Sum = <u>0.36</u> | | | | |

Counting Rules Are Often Useful In Computing Probabilities

- In many cases, there are a large number of possible outcomes.
- Counting rules can be used in these cases to help compute probabilities.

Counting Rules

- Rules for counting the number of possible outcomes
- Counting Rule 1: Chance Rule
 - If any one of k different mutually exclusive and collectively exhaustive events can occur on each of n trials, the number of possible outcomes is equal to

$$k^n$$

- Example
 - If you roll a fair die 3 times then there are $6^3 = 216$ possible outcomes

Counting Rules

(continued)

- Counting Rule 2: **Choices Rule**

- If there are k_1 events on the first trial, k_2 events on the second trial, ... and k_n events on the n^{th} trial, the number of possible outcomes is

$$(k_1)(k_2) \cdots (k_n)$$

- Example:

- You want to go to a park, eat at a restaurant, and see a movie. There are 3 parks, 4 restaurants, and 6 movie choices. How many different possible combinations are there?
 - Answer: $(3)(4)(6) = 72$ different possibilities

Counting Rules

(continued)

■ Counting Rule 3: Order Rule

- The number of ways that n items can be arranged in order is

$$n! = (n)(n - 1) \cdots (1)$$

■ Example:

- You have five books to put on a bookshelf. How many different ways can these books be placed on the shelf?
- Answer: $5! = (5)(4)(3)(2)(1) = 120$ different possibilities

Counting Rules

(continued)

■ Counting Rule 4:

- **Permutations:** The number of ways of arranging X objects selected from n objects in order is

$${}_n P_x = \frac{n!}{(n-X)!}$$

■ Example:

- You have five books and are going to put three on a bookshelf. How many different ways can the books be ordered on the bookshelf?

- Answer:
$${}_n P_x = \frac{n!}{(n-X)!} = \frac{5!}{(5-3)!} = \frac{120}{2} = 60$$
 different possibilities

Counting Rules

(continued)

■ Counting Rule 5:

- **Combinations:** The number of ways of selecting X objects from n objects, irrespective of order, is

$${}^n C_x = \frac{n!}{X!(n-X)!}$$

■ Example:

- You have five books and are going to select three to read. How many different combinations are there, ignoring the order in which they are selected?

- Answer:
$${}^n C_x = \frac{n!}{X!(n-X)!} = \frac{5!}{3!(5-3)!} = \frac{120}{(6)(2)} = 10 \quad \text{different possibilities}$$

What's the Difference?

In English we use the word "combination" loosely, without thinking if the **order** of things is important. In other words:



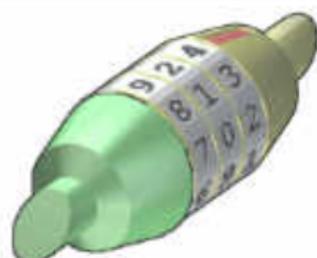
"My fruit salad is a combination of apples, grapes and bananas" We don't care what order the fruits are in, they could also be "bananas, grapes and apples" or "grapes, apples and bananas", it's the same fruit salad.



"The combination to the safe is 472". Now we **do** care about the order. "724" won't work, nor will "247". It has to be exactly **4-7-2**.

So, in Mathematics we use more *precise* language:

- When the order doesn't matter, it is a **Combination**.
- When the order **does** matter it is a **Permutation**.



So, we should really call this a "Permutation Lock"!

For example, let us say balls 1, 2 and 3 are chosen. These are the possibilities:

| Order does matter | Order doesn't matter |
|-------------------|----------------------|
| 1 2 3 | |
| 1 3 2 | |
| 2 1 3 | |
| 2 3 1 | |
| 3 1 2 | |
| 3 2 1 | 1 2 3 |

So, the permutations have 6 times as many possibilities.

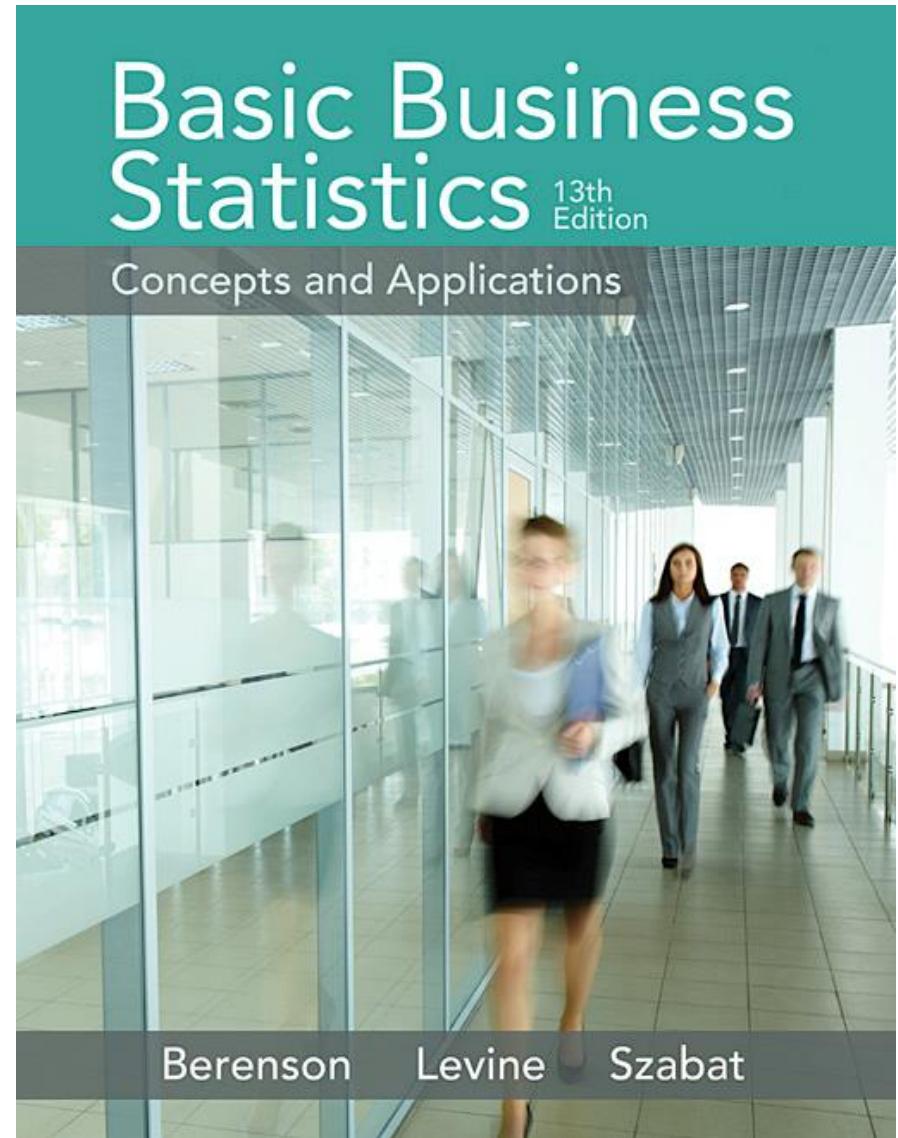
Chapter Summary

In this chapter we discussed:

- Basic probability concepts
 - Sample spaces and events, contingency tables, simple probability, and joint probability
- Basic probability rules
 - General addition rule, addition rule for mutually exclusive events, rule for collectively exhaustive events
- Conditional probability
 - Statistical independence, marginal probability, decision trees, and the multiplication rule
- Bayes' theorem
- Five useful counting rules

Chapter 6

The Normal Distribution and Other Continuous Distributions



Learning Objectives

In this chapter, you learn:

- To compute probabilities from the normal distribution
- How to use the normal distribution to solve business problems
- To use the normal probability plot to determine whether a set of data is approximately normally distributed
- To compute probabilities from the uniform distribution
- To compute probabilities from the exponential distribution

Continuous Probability Distributions

- A **continuous random variable** is a variable that can assume any value on a continuum (can assume an uncountable number of values)
 - thickness of an item
 - time required to complete a task
 - temperature of a solution
 - height, in inches
- These can potentially take on any value depending only on the ability to precisely and accurately measure

The Normal Distribution

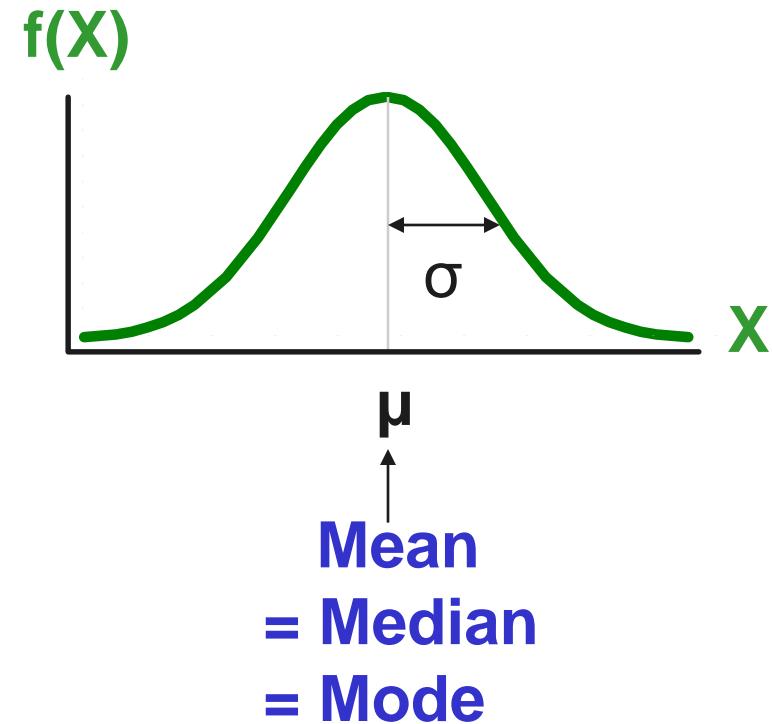
- ‘Bell Shaped’
- Symmetrical
- Mean, Median and Mode are Equal

Location is determined by the mean, μ

Spread is determined by the standard deviation, σ

The random variable has an infinite theoretical range:

$+\infty$ to $-\infty$



The Normal Distribution

Density Function

- The formula for the normal probability density function is

$$f(X) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{(X-\mu)}{\sigma}\right)^2}$$

Where e = the mathematical constant approximated by 2.71828

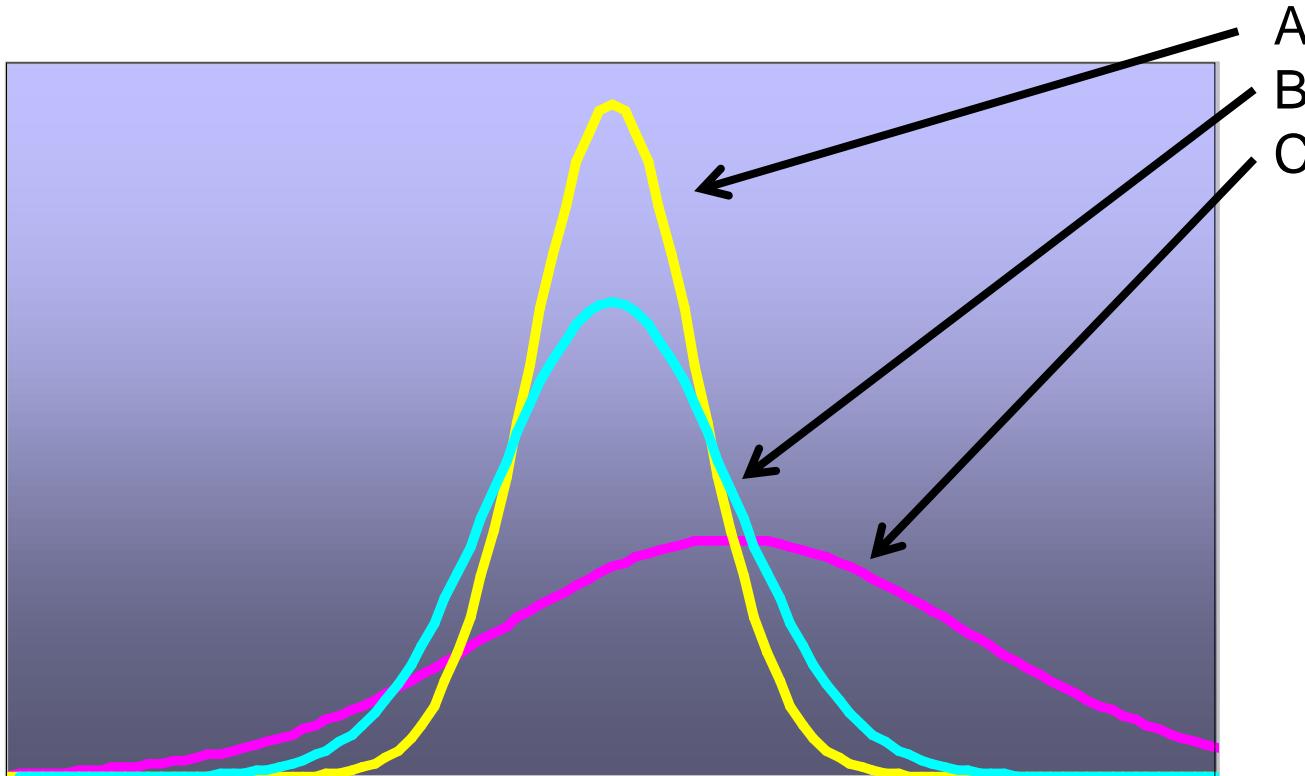
π = the mathematical constant approximated by 3.14159

μ = the population mean

σ = the population standard deviation

X = any value of the continuous variable

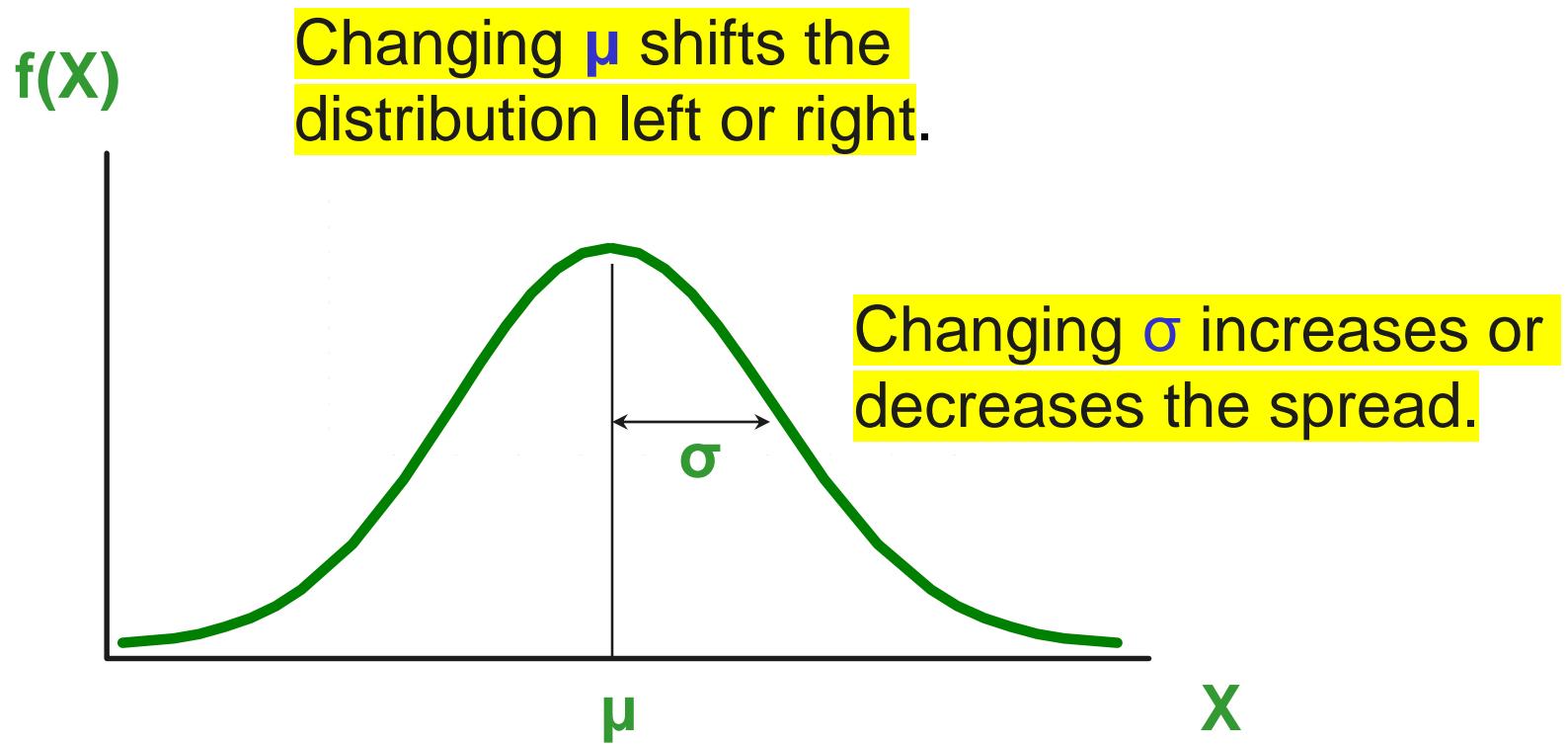
By varying the parameters μ and σ , we obtain different normal distributions



A and B have the same mean but different standard deviations.

B and C have different means and different standard deviations.

The Normal Distribution Shape



The Standardized Normal

- Any normal distribution (with any mean and standard deviation combination) can be transformed into the standardized normal distribution (Z)
- To compute normal probabilities we need to transform X units into Z units
- The standardized normal distribution (Z) has a mean of 0 and a standard deviation of 1

Translation to the Standardized Normal Distribution

- Translate from X to the standardized normal (the “ Z ” distribution) by subtracting the mean of X and dividing by its standard deviation:

$$Z = \frac{X - \mu}{\sigma}$$

The Z distribution always has mean = 0 and standard deviation = 1

The Standardized Normal Probability Density Function

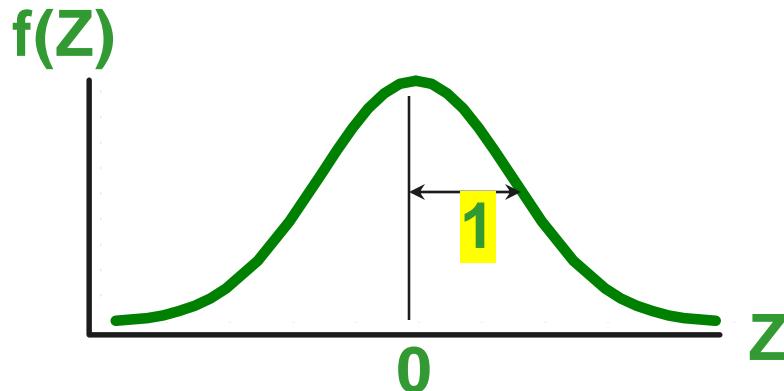
- The formula for the standardized normal probability density function is

$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-(1/2)Z^2}$$

Where e = the mathematical constant approximated by 2.71828
 π = the mathematical constant approximated by 3.14159
 Z = any value of the standardized normal distribution

The Standardized Normal Distribution

- Also known as the “Z” distribution
- Mean is 0
- Standard Deviation is 1



Values above the mean have positive Z-values.

Values below the mean have negative Z-values.

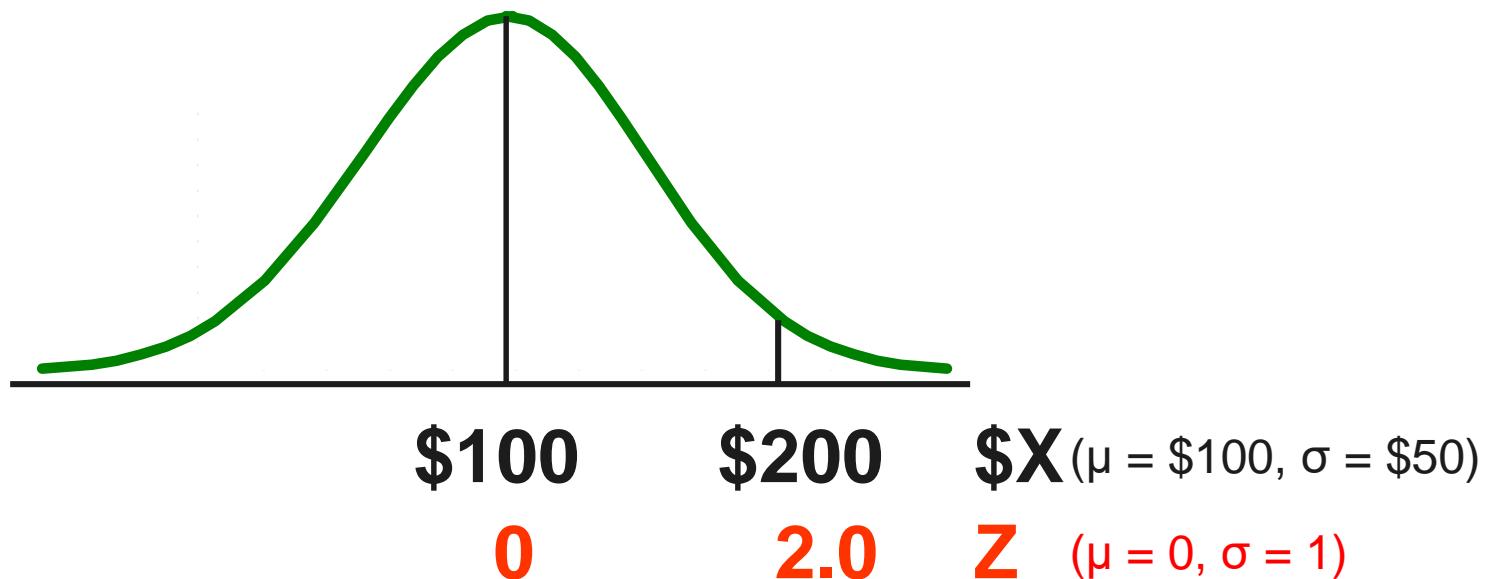
Example

- If X is distributed normally with mean of \$100 and standard deviation of \$50, the Z value for $X = \$200$ is

$$Z = \frac{X - \mu}{\sigma} = \frac{\$200 - \$100}{\$50} = 2.0$$

- This says that $X = \$200$ is two standard deviations (2 increments of \$50 units) above the mean of \$100.

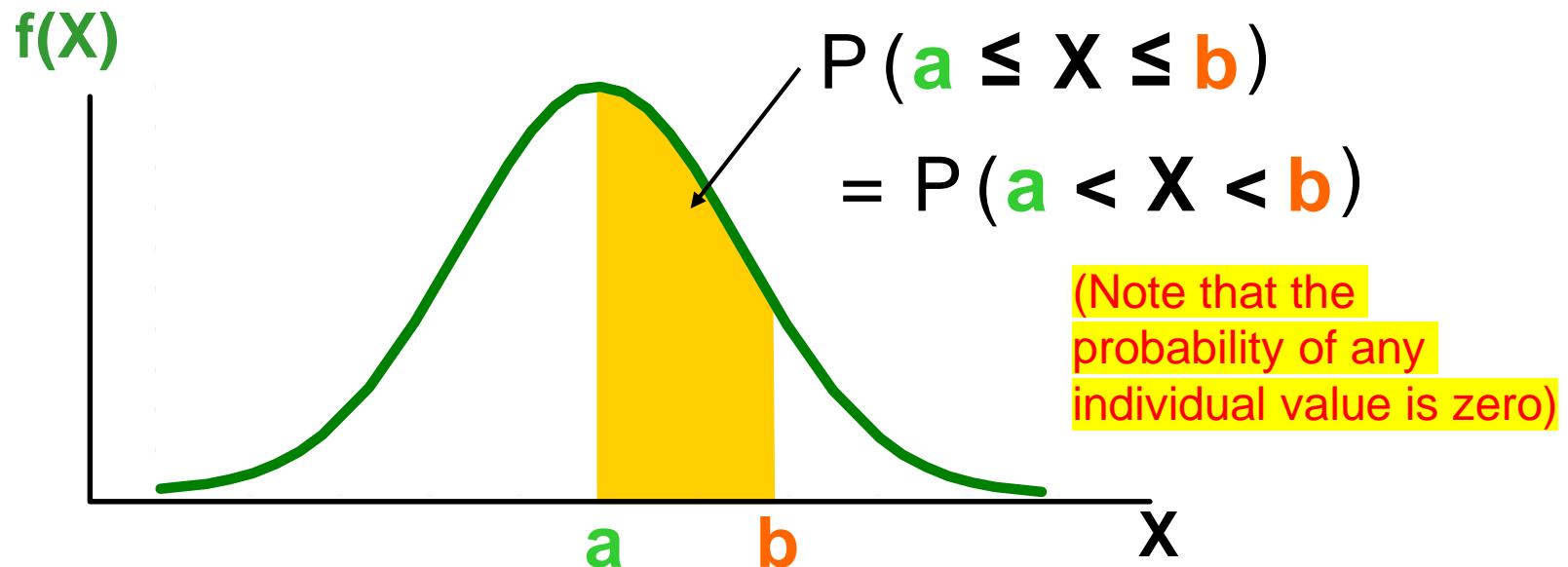
Comparing X and Z units



Note that the shape of the distribution is the same, only the scale has changed. We can express the problem in the original units (X in dollars) or in standardized units (Z)

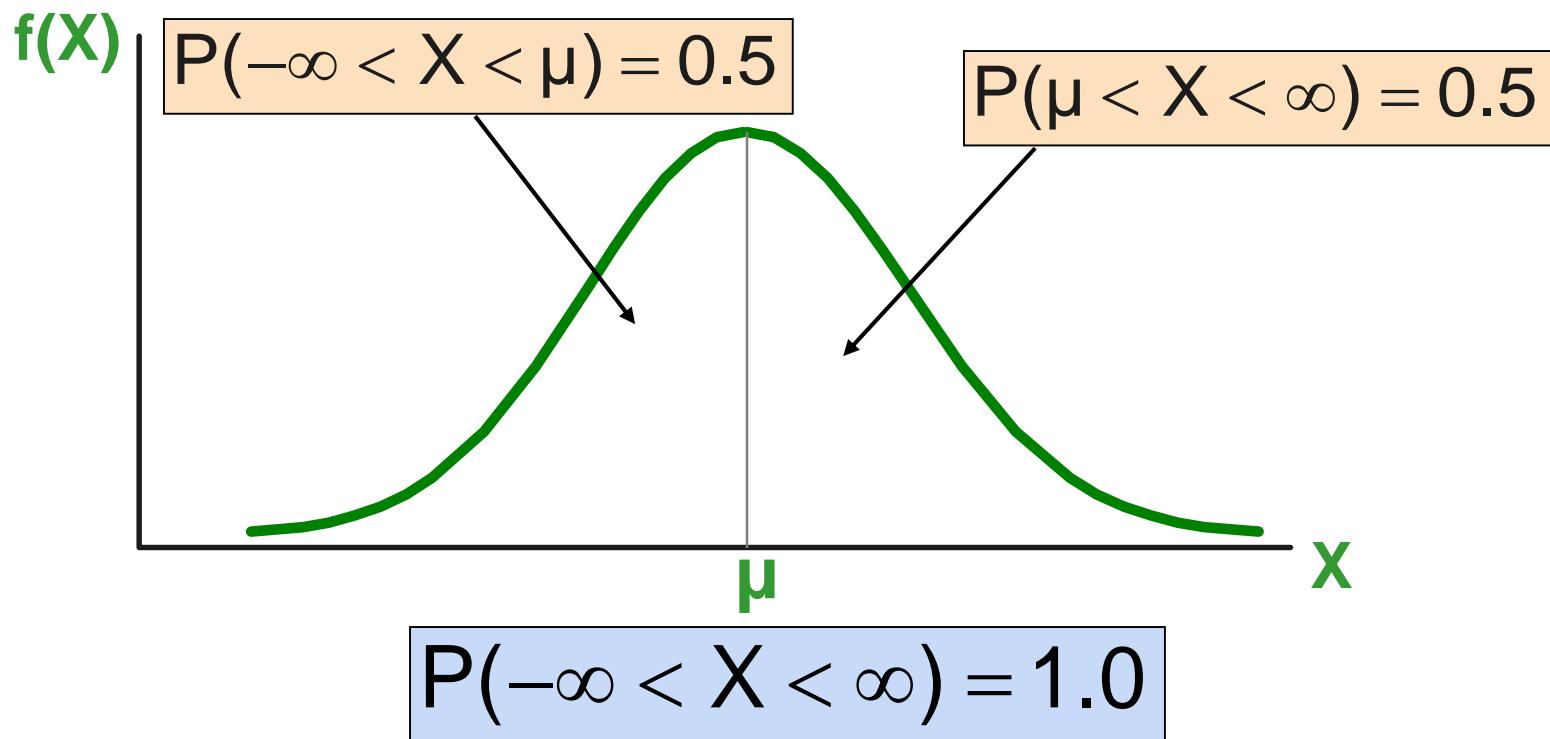
Finding Normal Probabilities

Probability is measured by the area
under the curve



Probability as Area Under the Curve

The total area under the curve is 1.0, and the curve is symmetric, so half is above the mean, half is below

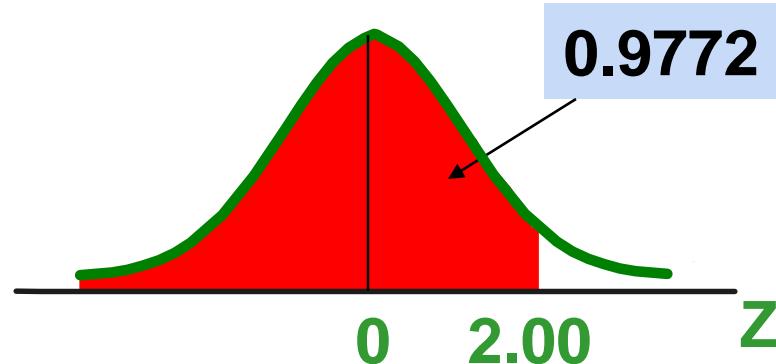


The Standardized Normal Table

- The Cumulative Standardized Normal table in the textbook (Appendix table E.2) gives the probability **less than** a desired value of Z (i.e., from negative infinity to Z)

Example:

$$P(Z < 2.00) = 0.9772$$



The Standardized Normal Table

(continued)

The column gives the value of Z to the second decimal point

| Z | 0.00 | 0.01 | 0.02 ... |
|-----|-------|------|----------|
| 0.0 | | | |
| 0.1 | | | |
| : | | | |
| : | | | |
| 2.0 | .9772 | | |

The row shows the value of Z to the first decimal point

The value within the table gives the probability from $Z = -\infty$ up to the desired Z value

$$P(Z < 2.00) = 0.9772$$

General Procedure for Finding Normal Probabilities

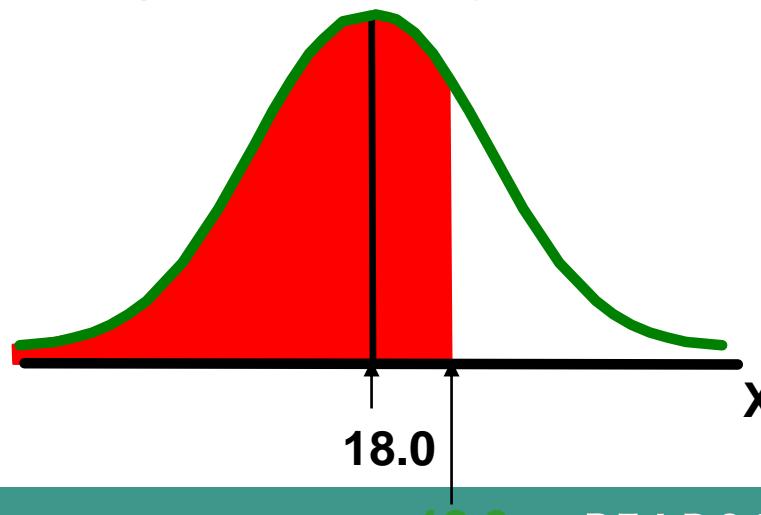
To find $P(a < X < b)$ when X is distributed normally:

- Draw the normal curve for the problem in terms of X
- Translate X -values to Z -values
- Use the Standardized Normal Table

Finding Normal Probabilities

Example

- Let X represent the time it takes (in seconds) to download an image file from the internet.
- Suppose X is normal with a mean of 18.0 seconds and a standard deviation of 5.0 seconds. Find $P(X < 18.6)$



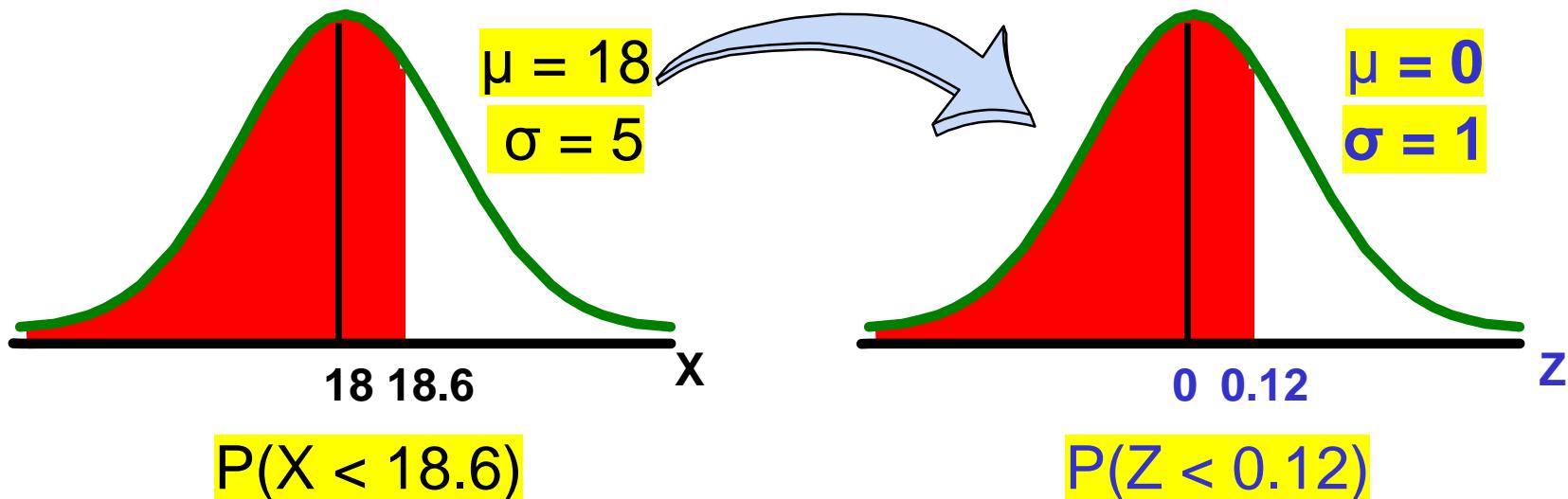
Finding Normal Probabilities

Example Cont.

(continued)

- Let X represent the time it takes, in seconds to download an image file from the internet.
- Suppose X is normal with a mean of 18.0 seconds and a standard deviation of 5.0 seconds. Find $P(X < 18.6)$

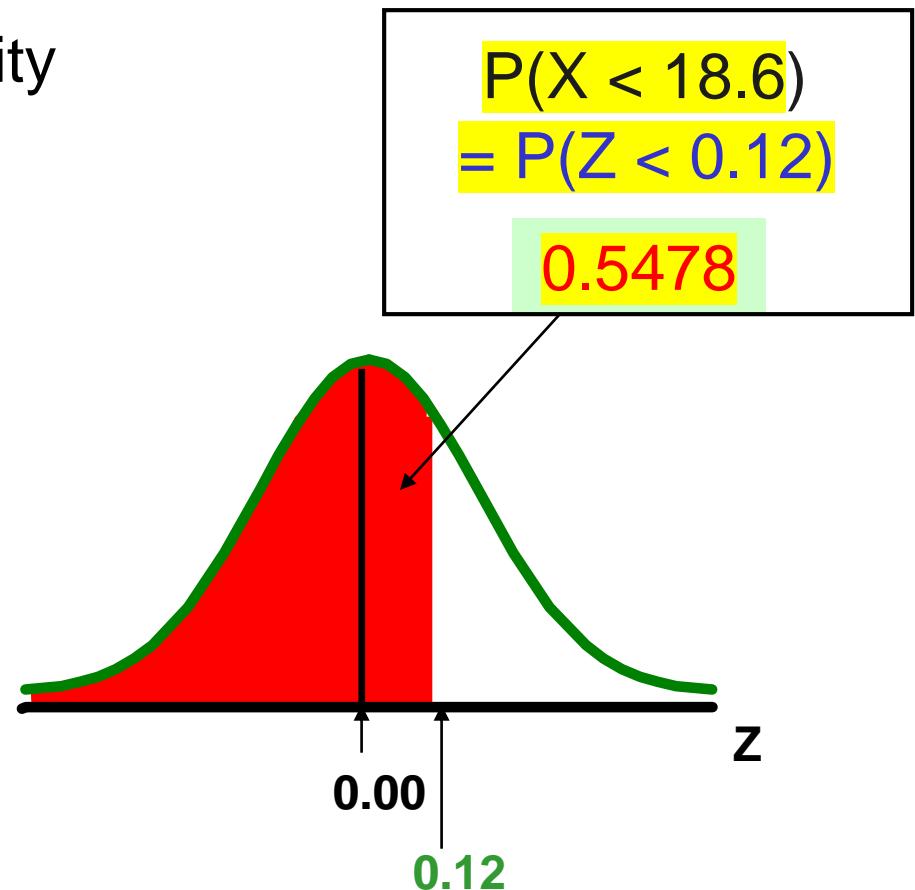
$$Z = \frac{X - \mu}{\sigma} = \frac{18.6 - 18.0}{5.0} = 0.12$$



Solution: Finding $P(Z < 0.12)$

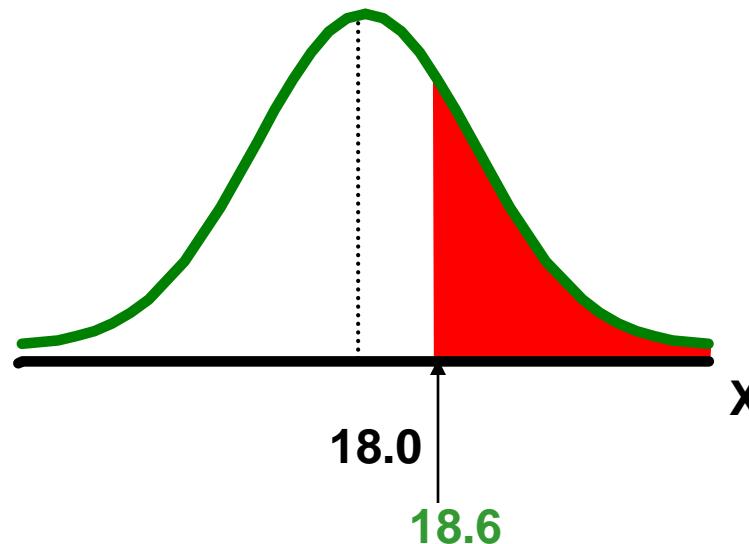
Standardized Normal Probability
Table (Portion)

| Z | .00 | .01 | .02 |
|-----|-------|-------|--------------|
| 0.0 | .5000 | .5040 | .5080 |
| 0.1 | .5398 | .5438 | .5478 |
| 0.2 | .5793 | .5832 | .5871 |
| 0.3 | .6179 | .6217 | .6255 |



Finding Normal Upper Tail Probabilities Example

- Suppose X is normal with mean 18.0 and standard deviation 5.0.
- Now Find $P(X > 18.6)$



Finding Normal Upper Tail Probabilities Example

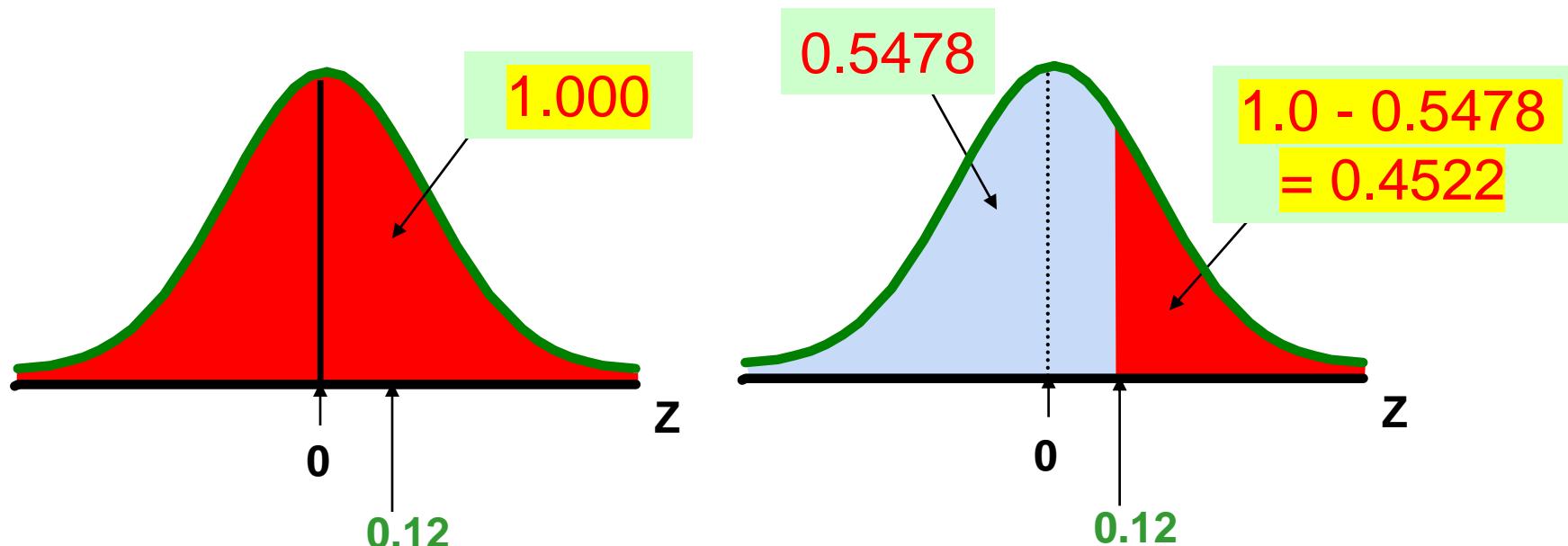
Cont.

(continued)

- Now Find $P(X > 18.6)$...

$$P(X > 18.6) = P(Z > 0.12) = 1.0 - P(Z \leq 0.12)$$

$$= 1.0 - 0.5478 = 0.4522$$

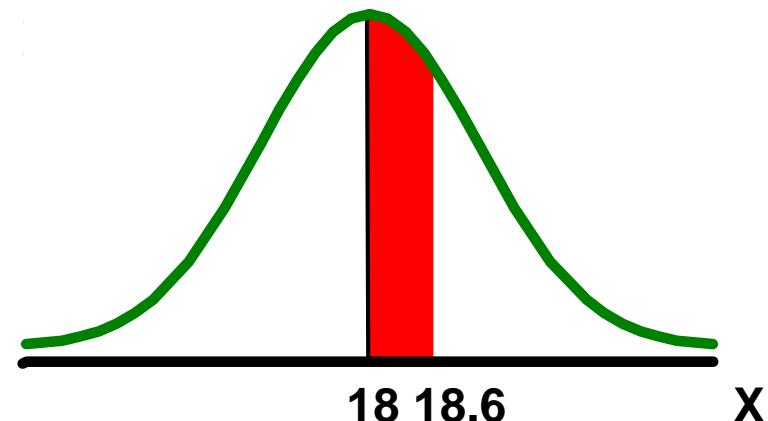


Finding a Normal Probability Between Two Values

- Suppose X is normal with mean 18.0 and standard deviation 5.0. Find $P(18 < X < 18.6)$

Calculate Z-values:

$$Z = \frac{X - \mu}{\sigma} = \frac{18 - 18}{5} = 0$$



$$Z = \frac{X - \mu}{\sigma} = \frac{18.6 - 18}{5} = 0.12$$

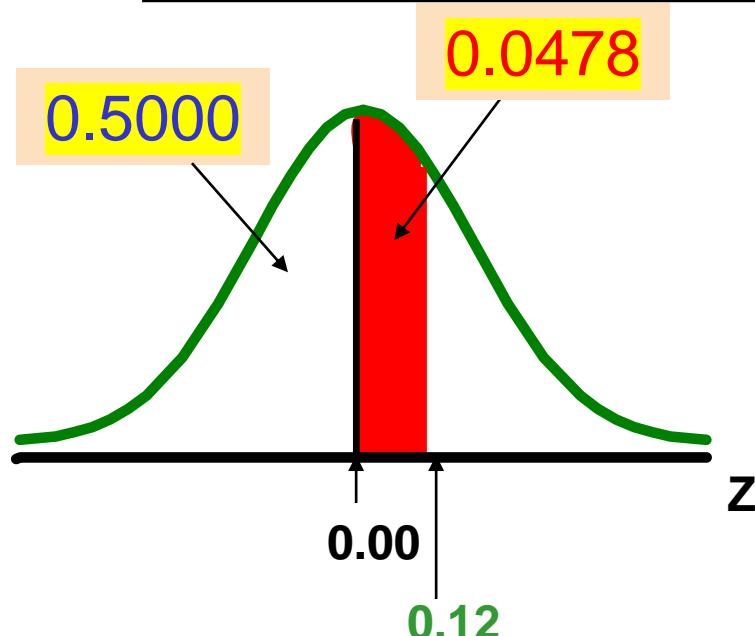
$$\begin{aligned}P(18 < X < 18.6) \\= P(0 < Z < 0.12)\end{aligned}$$

Solution: Finding $P(0 < Z < 0.12)$

Standardized Normal Probability Table (Portion)

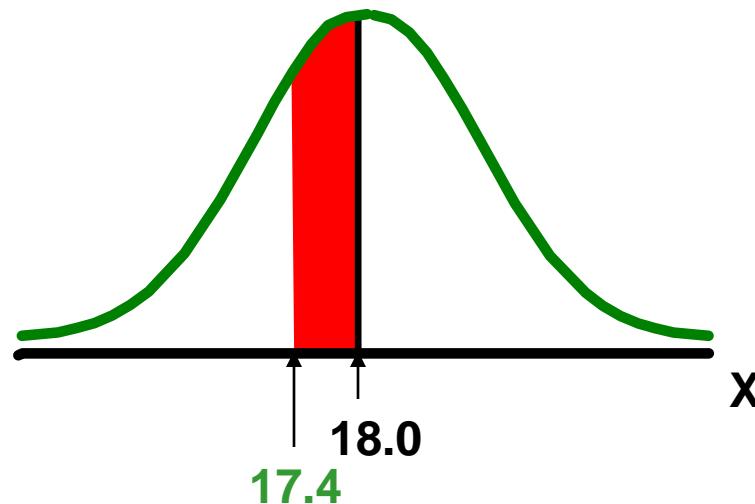
| Z | .00 | .01 | .02 |
|-----|-------|-------|-------|
| 0.0 | .5000 | .5040 | .5080 |
| 0.1 | .5398 | .5438 | .5478 |
| 0.2 | .5793 | .5832 | .5871 |
| 0.3 | .6179 | .6217 | .6255 |

$$\begin{aligned}P(18 < X < 18.6) \\&= P(0 < Z < 0.12) \\&= P(Z < 0.12) - P(Z \leq 0) \\&= 0.5478 - 0.5000 = 0.0478\end{aligned}$$



Probabilities in the Lower Tail

- Suppose X is normal with mean 18.0 and standard deviation 5.0.
- Now Find $P(17.4 < X < 18)$



Probabilities in the Lower Tail

(continued)

Now Find $P(17.4 < X < 18)$...

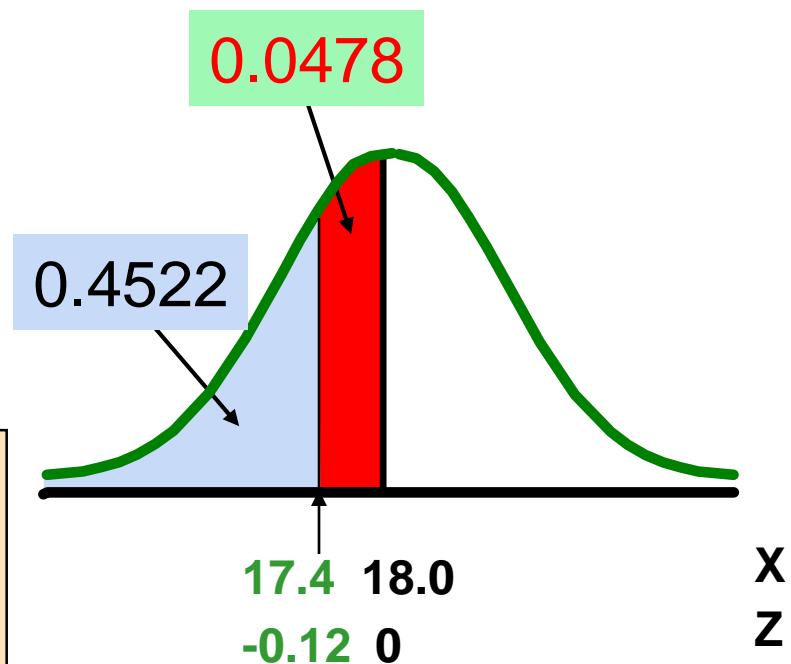
$$P(17.4 < X < 18)$$

$$= P(-0.12 < Z < 0)$$

$$= P(Z < 0) - P(Z \leq -0.12)$$

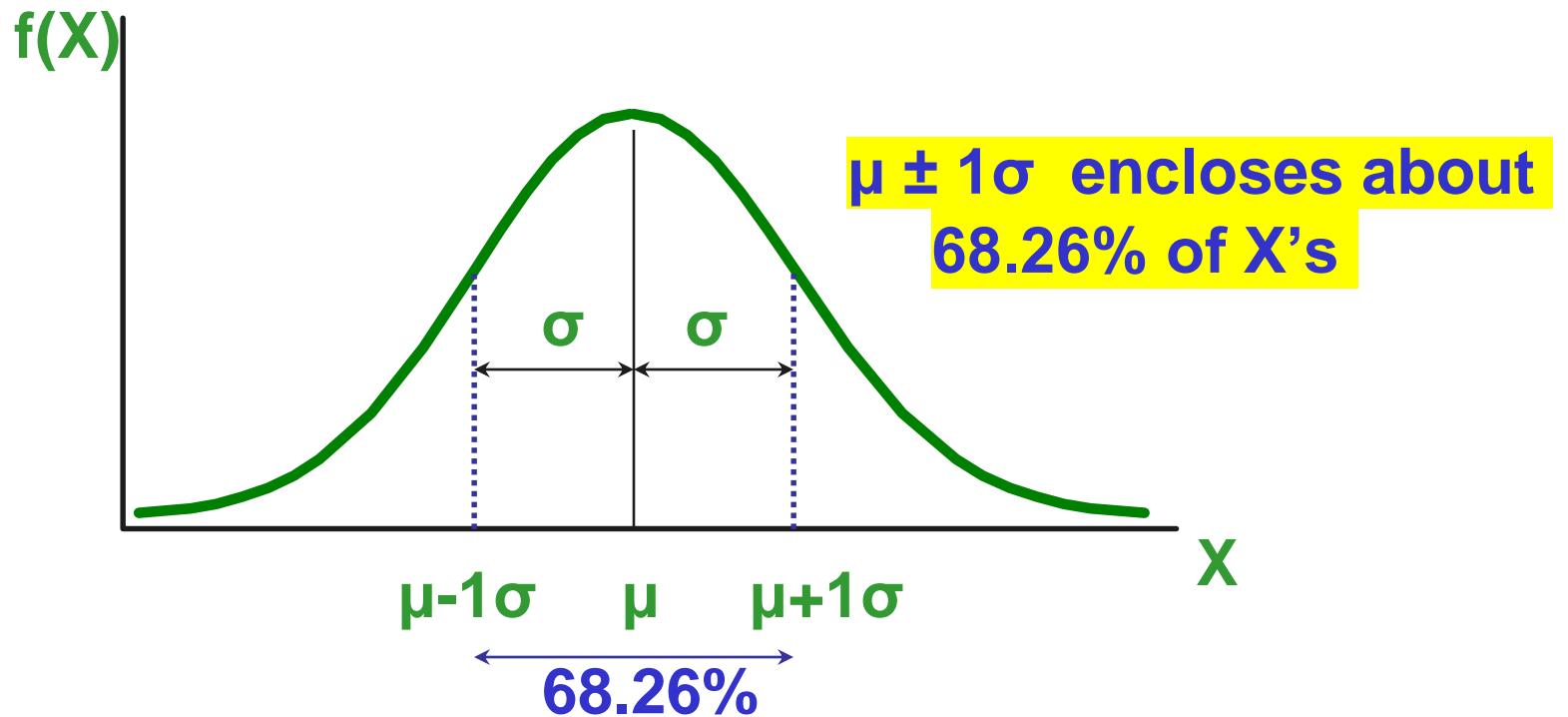
$$= 0.5000 - 0.4522 = \boxed{0.0478}$$

The Normal distribution is symmetric, so this probability is the same as $P(0 < Z < 0.12)$



Empirical Rule

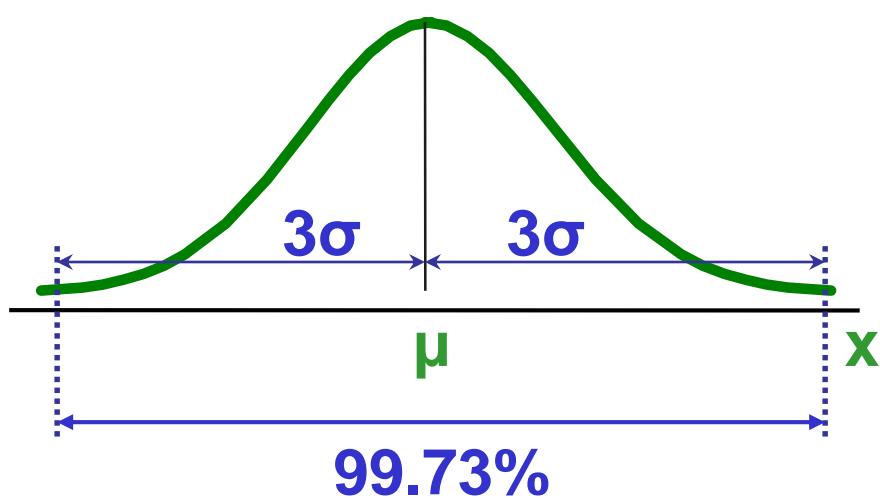
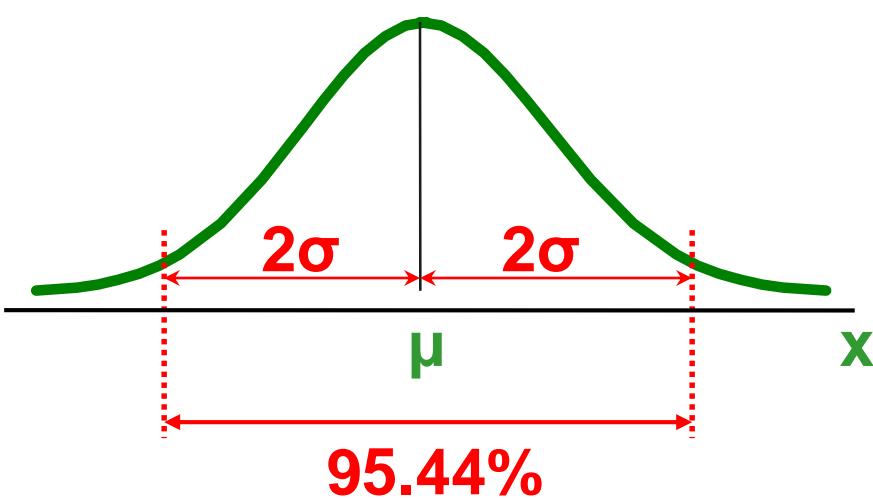
What can we say about the distribution of values around the mean? For any normal distribution:



The Empirical Rule

(continued)

- $\mu \pm 2\sigma$ covers about 95.44% of X's
- $\mu \pm 3\sigma$ covers about 99.73% of X's



Given a Normal Probability Find the X Value

- Steps to find the X value for a known probability:
 1. Find the Z value for the known probability
 2. Convert to X units using the formula:

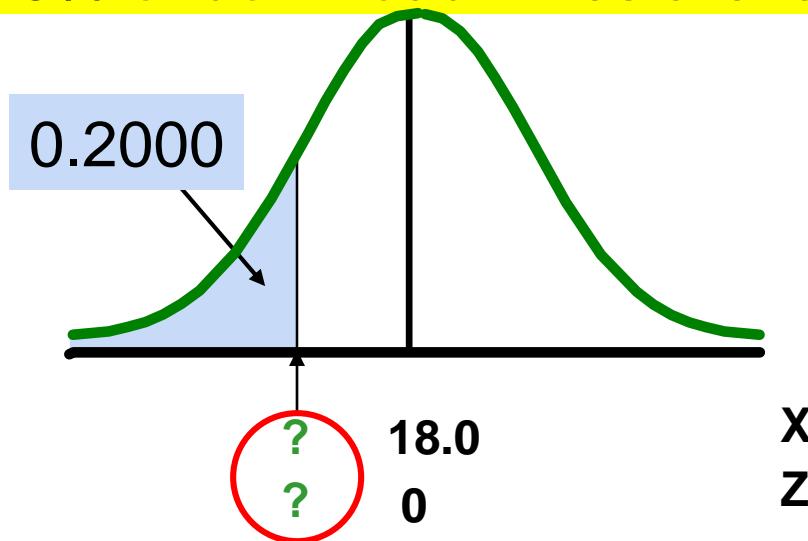
$$X = \mu + Z\sigma$$

Finding the X value for a Known Probability Example

(continued)

Example:

- Let X represent the time it takes (in seconds) to download an image file from the internet.
- Suppose X is normal with mean 18.0 and standard deviation 5.0
- Find X such that 20% of download times are less than X .



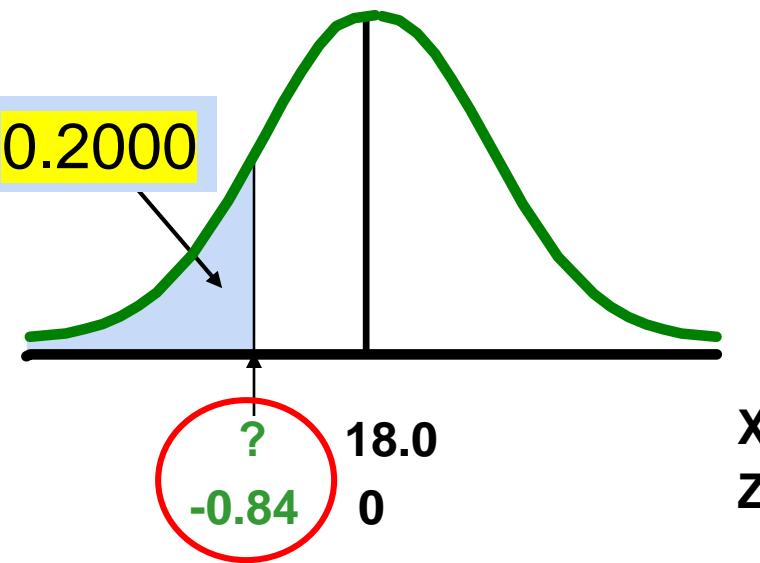
Find the Z value for 20% in the Lower Tail

1. Find the Z value for the known probability

Standardized Normal Probability Table (Portion)

| Z | ... | .03 | .04 | .05 |
|-------------|-----|-------|--------------|-------|
| -0.9 | ... | .1762 | .1736 | .1711 |
| -0.8 | ... | .2033 | .2005 | .1977 |
| -0.7 | ... | .2327 | .2296 | .2266 |

- 20% area in the lower tail is consistent with a Z value of **-0.84**



Finding the X value

2. Convert to X units using the formula:

$$\begin{aligned}X &= \mu + Z\sigma \\&= 18.0 + (-0.84)5.0 \\&= 13.8\end{aligned}$$

So 20% of the values from a distribution with mean 18.0 and standard deviation 5.0 are less than 13.80

Both Minitab & Excel Can Be Used To Find Normal Probabilities

Find $P(X < 9)$ where X is normal with a mean of 7 and a standard deviation of 2

Excel

| | A | B |
|----|--------------------------|--------|
| 1 | Normal Probabilities | |
| 2 | | |
| 3 | Common Data | |
| 4 | Mean | 7 |
| 5 | Standard Deviation | 2 |
| 6 | | |
| 7 | Probability for $X \leq$ | |
| 8 | X Value | 7 |
| 9 | Z Value | 0 |
| 10 | $P(X \leq 7)$ | 0.5000 |
| 11 | | |
| 12 | Probability for $X >$ | |
| 13 | X Value | 9 |
| 14 | Z Value | 1 |
| 15 | $P(X > 9)$ | 0.1587 |

Minitab

Cumulative Distribution Function

Normal with mean = 7 and standard deviation = 2

$$\begin{array}{ll} x & P(X \leq x) \\ 9 & 0.841345 \end{array}$$

=STANDARDIZE(B8, B4, B5)
=NORM.DIST(B8, B4, B5, TRUE)

=STANDARDIZE(B13, B4, B5)
=1 - NORM.DIST(B13, B4, B5, TRUE)

Evaluating Normality

- Not all continuous distributions are normal
- It is important to evaluate how well the data set is approximated by a normal distribution.
- Normally distributed data should approximate the theoretical normal distribution:
 - The normal distribution is bell shaped (symmetrical) where the mean is equal to the median.
 - The empirical rule applies to the normal distribution.
 - The interquartile range of a normal distribution is 1.33 standard deviations.

Evaluating Normality

(continued)

Comparing data characteristics to theoretical properties

■ Construct charts or graphs

- For small- or moderate-sized data sets, construct a stem-and-leaf display or a boxplot to check for symmetry
- For large data sets, does the histogram or polygon appear bell-shaped?

■ Compute descriptive summary measures

- Do the mean, median and mode have similar values?
- Is the interquartile range approximately 1.33σ ?
- Is the range approximately 6σ ?

Evaluating Normality

(continued)

Comparing data characteristics to theoretical properties

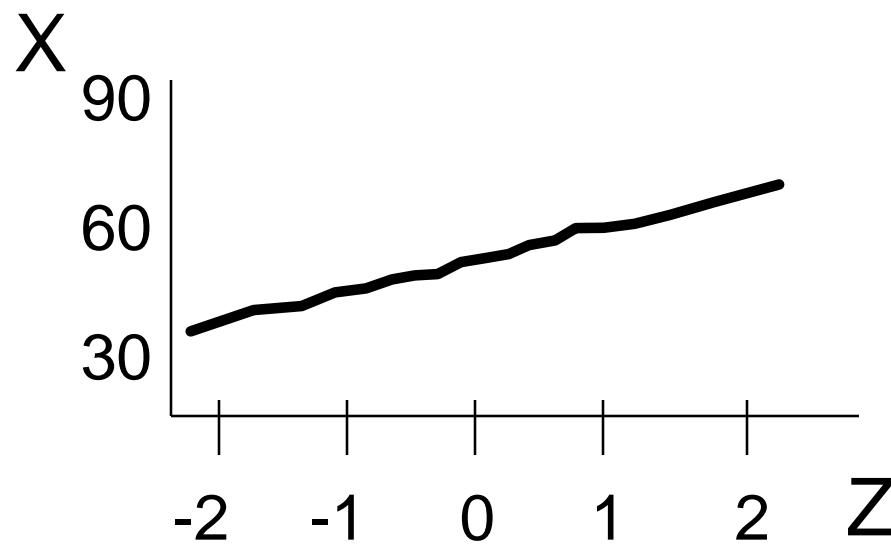
- Observe the distribution of the data set
 - Do approximately 2/3 of the observations lie within mean ± 1 standard deviation?
 - Do approximately 80% of the observations lie within mean ± 1.28 standard deviations?
 - Do approximately 95% of the observations lie within mean ± 2 standard deviations?
- Evaluate normal probability plot
 - Is the normal probability plot approximately linear (i.e. a straight line) with positive slope?

Constructing A Normal Probability Plot

- **Normal probability plot**
 - Arrange data into ordered array
 - Find corresponding standardized normal quantile values (Z)
 - Plot the pairs of points with observed data values (X) on the vertical axis and the standardized normal quantile values (Z) on the horizontal axis
 - Evaluate the plot for evidence of linearity

The Normal Probability Plot Interpretation

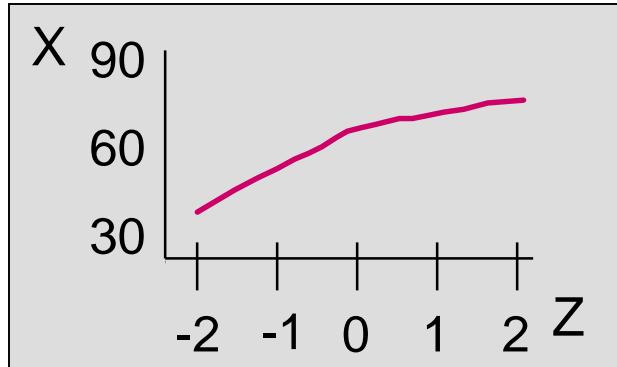
A normal probability plot for data from a normal distribution will be approximately linear:



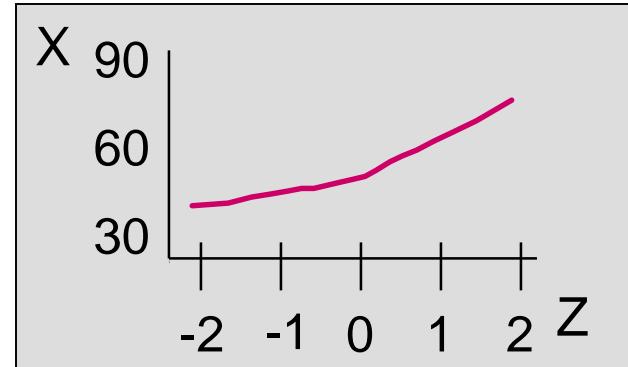
Normal Probability Plot Interpretation

(continued)

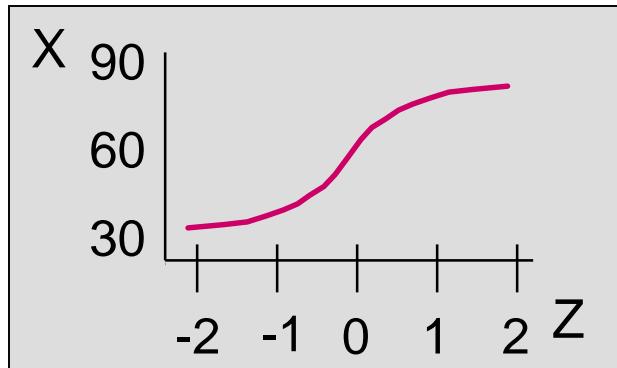
Left-Skewed



Right-Skewed



Rectangular



Nonlinear plots indicate a deviation from normality

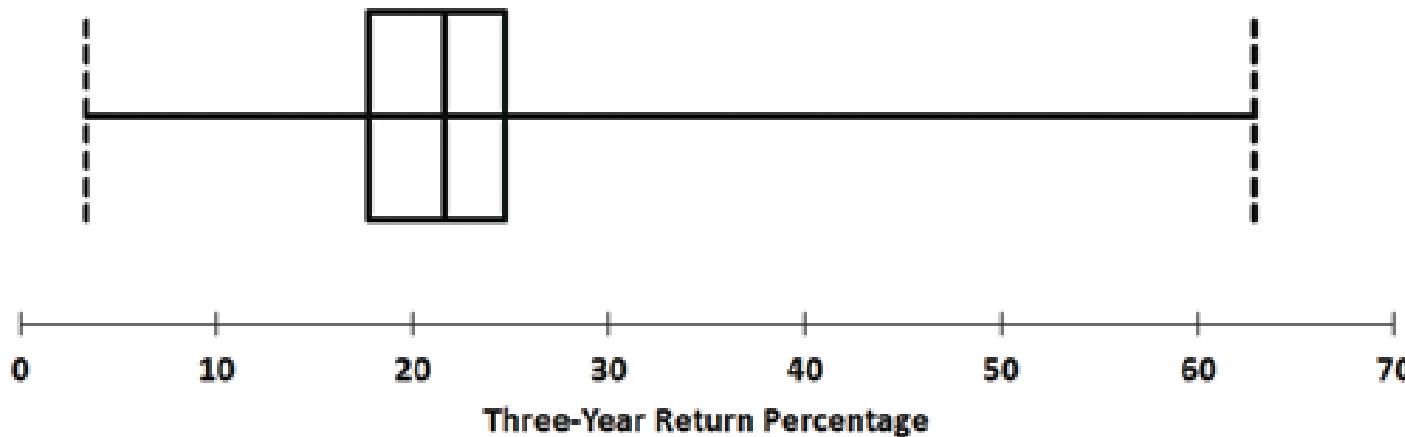
Evaluating Normality

An Example: Mutual Fund Returns

| Five-Number Summary | |
|---------------------|-------|
| Minimum | 3.39 |
| First quartile | 17.76 |
| Median | 21.65 |
| Third quartile | 24.74 |
| Maximum | 62.91 |

The boxplot is skewed to the right. (The normal distribution is symmetric.)

Boxplot for the Three-Year Return Percentage



Evaluating Normality

An Example: Mutual Fund Returns

(continued)

Descriptive Statistics

| | 3YrReturn% |
|---------------------|------------|
| Mean | 21.84 |
| Median | 21.65 |
| Mode | 21.74 |
| Minimum | 3.39 |
| Maximum | 62.91 |
| Range | 59.52 |
| Variance | 41.2968 |
| Standard Deviation | 6.4263 |
| Coeff. of Variation | 29.43% |
| Skewness | 1.6976 |
| Kurtosis | 8.4670 |
| Count | 318 |
| Standard Error | 0.3604 |

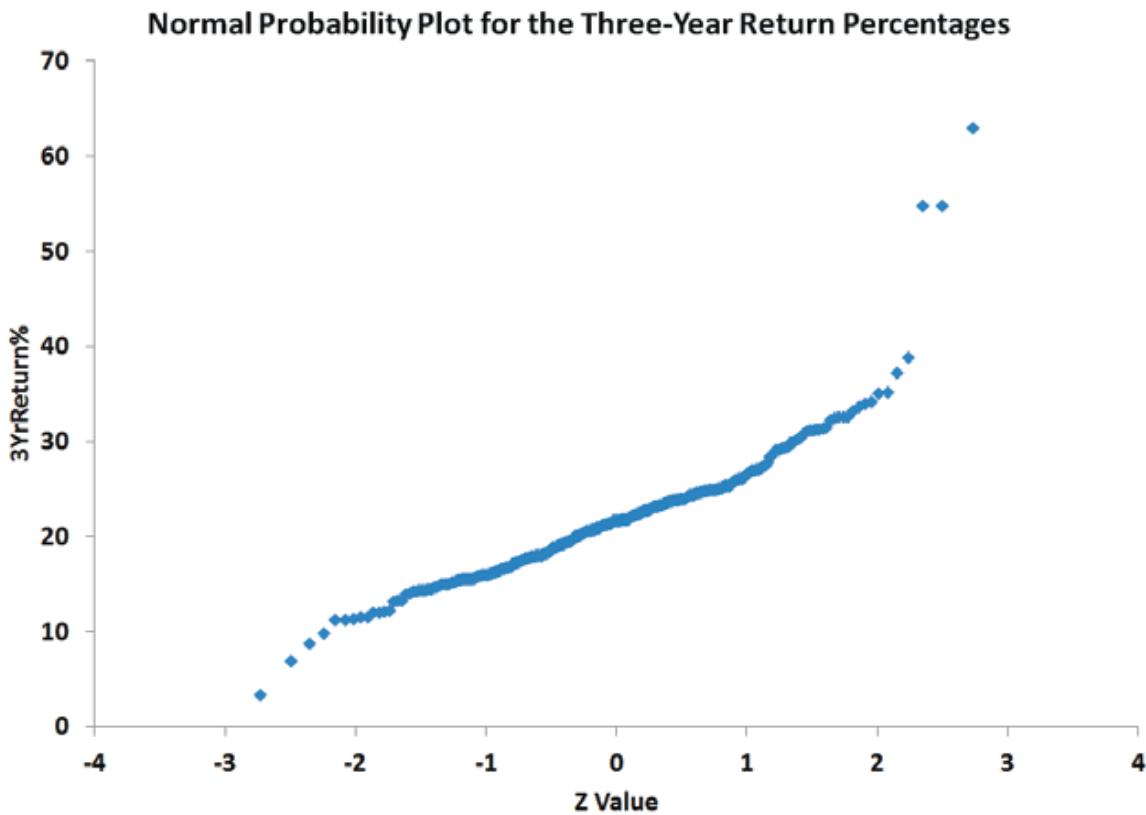
- The mean (21.84) is approximately the same as the median (21.65). (In a normal distribution the mean and median are equal.)
- The interquartile range of 6.98 is approximately 1.09 standard deviations. (In a normal distribution the interquartile range is 1.33 standard deviations.)
- The range of 59.52 is equal to 9.26 standard deviations. (In a normal distribution the range is 6 standard deviations.)
- 77.04% of the observations are within 1 standard deviation of the mean. (In a normal distribution this percentage is 68.26%).
- 86.79% of the observations are within 1.28 standard deviations of the mean. (In a normal distribution this percentage is 80%).
- 96.86% of the observations are within 2 standard deviations of the mean. (In a normal distribution this percentage is 95.44%).
- The skewness statistic is 1.698 and the kurtosis statistic is 8.467. (In a normal distribution, each of these statistics equals zero.)

Evaluating Normality Via Excel

An Example: Mutual Fund Returns

(continued)

Excel (quantile-quantile) normal probability plot



Plot is not a straight line and shows the distribution is skewed to the right. (But a normal distribution appears as a straight line.)

Evaluating Normality

An Example: Mutual Fund Returns

(continued)

Conclusions

- The returns are right-skewed
- The returns have more values concentrated around the mean than expected
- The range is larger than expected
- Normal probability plot is not a straight line
- Overall, this data set greatly differs from the theoretical properties of the normal distribution

THE END OF THIS CHAPTER

The Uniform Distribution

- The uniform distribution is a probability distribution that has equal probabilities for all possible outcomes of the random variable
- Also called a rectangular distribution

The Uniform Distribution

(continued)

The Continuous Uniform Distribution:

$$f(X) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq X \leq b \\ 0 & \text{otherwise} \end{cases}$$

where

$f(X)$ = value of the density function at any X value

a = minimum value of X

b = maximum value of X

Properties of the Uniform Distribution

- The mean of a uniform distribution is

$$\mu = \frac{a + b}{2}$$

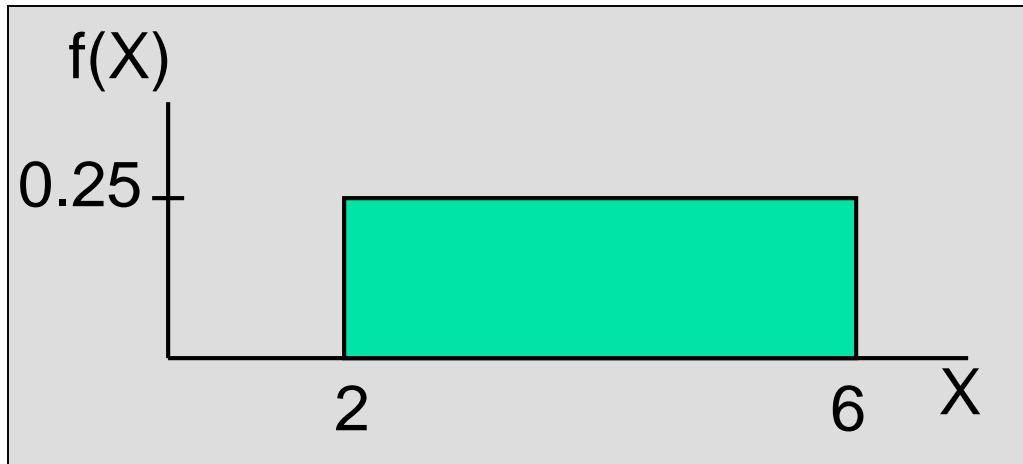
- The standard deviation is

$$\sigma = \sqrt{\frac{(b - a)^2}{12}}$$

Uniform Distribution Example

Example: Uniform probability distribution over the range $2 \leq X \leq 6$:

$$f(X) = \frac{1}{6 - 2} = 0.25 \quad \text{for } 2 \leq X \leq 6$$



$$\mu = \frac{a+b}{2} = \frac{2+6}{2} = 4$$

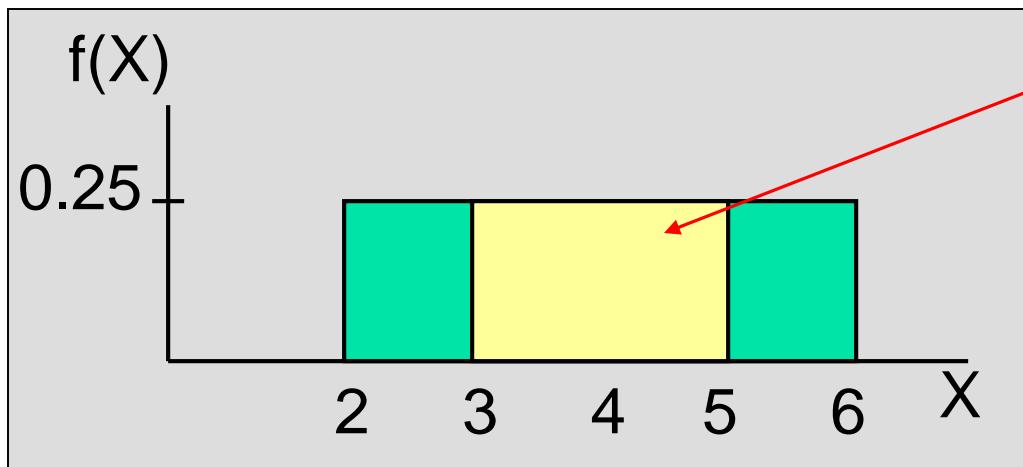
$$\sigma = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{(6-2)^2}{12}} = 1.1547$$

Uniform Distribution Example

(continued)

Example: Using the uniform probability distribution to find $P(3 \leq X \leq 5)$:

$$P(3 \leq X \leq 5) = (\text{Base})(\text{Height}) = (2)(0.25) = 0.5$$



The Exponential Distribution

- Often used to model the **length of time between two occurrences** of an event (the time between arrivals)
 - Examples:
 - Time between trucks arriving at an unloading dock
 - Time between transactions at an ATM Machine
 - Time between phone calls to the main operator

The Exponential Distribution

- Defined by a single parameter, its mean λ (lambda)
- Exponential Probability Density Function:

$$f(X) = \lambda e^{-\lambda X} \text{ for } X > 0$$

- The probability that an arrival time is less than some specified time X is

$$P(\text{arrival time} < X) = 1 - e^{-\lambda X}$$

where e = mathematical constant approximated by 2.71828

λ = the population mean number of arrivals per unit

X = any value of the continuous variable where $0 < X < \infty$

The Mean & Standard Deviation Of The Exponential Distribution

- The mean (μ) of the exponential distribution is given by:

$$\mu = 1/\lambda$$

- The standard deviation (σ) of the exponential distribution is given by:

$$\mu = 1/\lambda$$

∞

Exponential Distribution Example

Example: Customers arrive at the service counter at the rate of 15 per hour. What is the probability that the arrival time between consecutive customers is less than three minutes?

- The mean number of arrivals per hour is 15, so $\lambda = 15$
- Three minutes is 0.05 hours
- $P(\text{arrival time} < .05) = 1 - e^{-\lambda X} = 1 - e^{-(15)(0.05)} = \boxed{0.5276}$
- So there is a 52.76% chance that the arrival time between successive customers is less than three minutes

The Exponential Distribution In Excel

Calculating the probability that an exponential distribution with a mean of 20 is less than 0.1

| | A | B |
|---|-------------------------|----------------------------------|
| 1 | Exponential Probability | |
| 2 | | |
| 3 | Data | |
| 4 | Mean | 20 |
| 5 | X Value | 0.1 |
| 6 | | |
| 7 | Results | |
| 8 | P(<=X) | 0.8647 =EXPON.DIST(B5, B4, TRUE) |

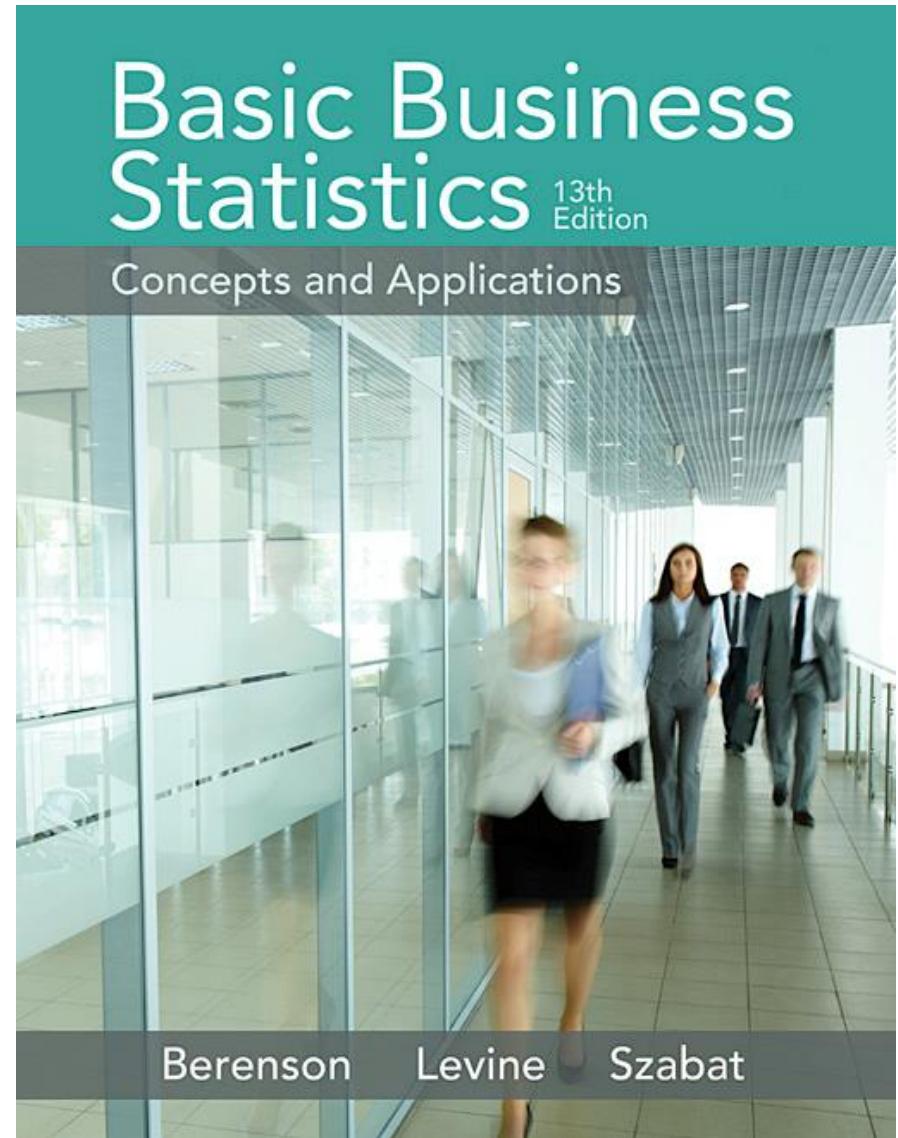
Chapter Summary

In this chapter we discussed

- Key continuous distributions
 - normal, uniform, exponential
- Finding probabilities using formulas and tables
- Recognizing when to apply different distributions
- Applying distributions to decision problems

Chapter 7

Sampling Distributions



Learning Objectives

In this chapter, you learn:

- The concept of the sampling distribution
- To compute probabilities related to the sample mean and the sample proportion
- The importance of the Central Limit Theorem

Sampling Distributions

DCOVA

- A sampling distribution is a distribution of all of the possible values of a sample statistic for a given sample size selected from a population.
- For example, suppose you sample 50 students from your college regarding their mean GPA. If you obtained many different samples of size 50, you will compute a different mean for each sample. We are interested in the distribution of all potential mean GPAs we might calculate for any sample of 50 students.

Developing a Sampling Distribution

DCOVA

- Assume there is a population ...
- Population size $N=4$
- Random variable, X ,
is age of individuals
- Values of X : 18, 20,
22, 24 (years)



Developing a Sampling Distribution

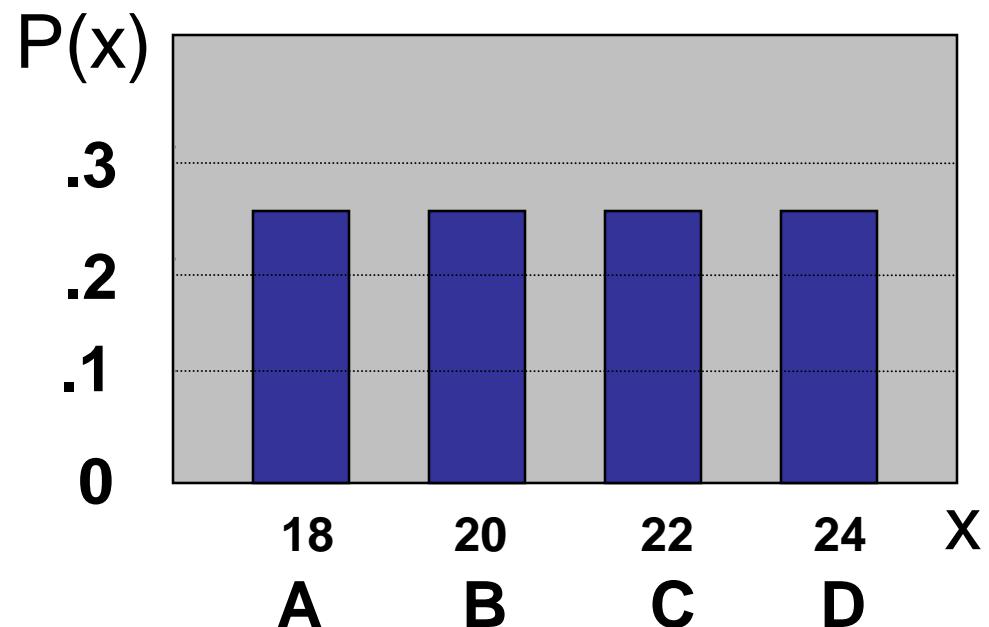
(continued)

DCOV A

Summary Measures for the Population Distribution:

$$\mu = \frac{\sum X_i}{N}$$
$$= \frac{18 + 20 + 22 + 24}{4} = 21$$

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}} = 2.236$$



Uniform Distribution

Developing a Sampling Distribution

(continued)

Now consider all possible samples of size $n=2$

DCOV A

| 1 st Obs | 2 nd Observation | | | |
|------------------------|-----------------------------|-------|-------|-------|
| | 18 | 20 | 22 | 24 |
| 18 | 18,18 | 18,20 | 18,22 | 18,24 |
| 20 | 20,18 | 20,20 | 20,22 | 20,24 |
| 22 | 22,18 | 22,20 | 22,22 | 22,24 |
| 24 | 24,18 | 24,20 | 24,22 | 24,24 |

16 possible samples
(sampling with replacement)

16 Sample Means



| 1st Obs | 2nd Observation | 18 | 20 | 22 | 24 |
|------------|-----------------|----|----|----|----|
| 18 | 18 | 18 | 19 | 20 | 21 |
| 20 | 19 | 19 | 20 | 21 | 22 |
| 22 | 20 | 20 | 21 | 22 | 23 |
| 24 | 21 | 21 | 22 | 23 | 24 |

Developing a Sampling Distribution

(continued)

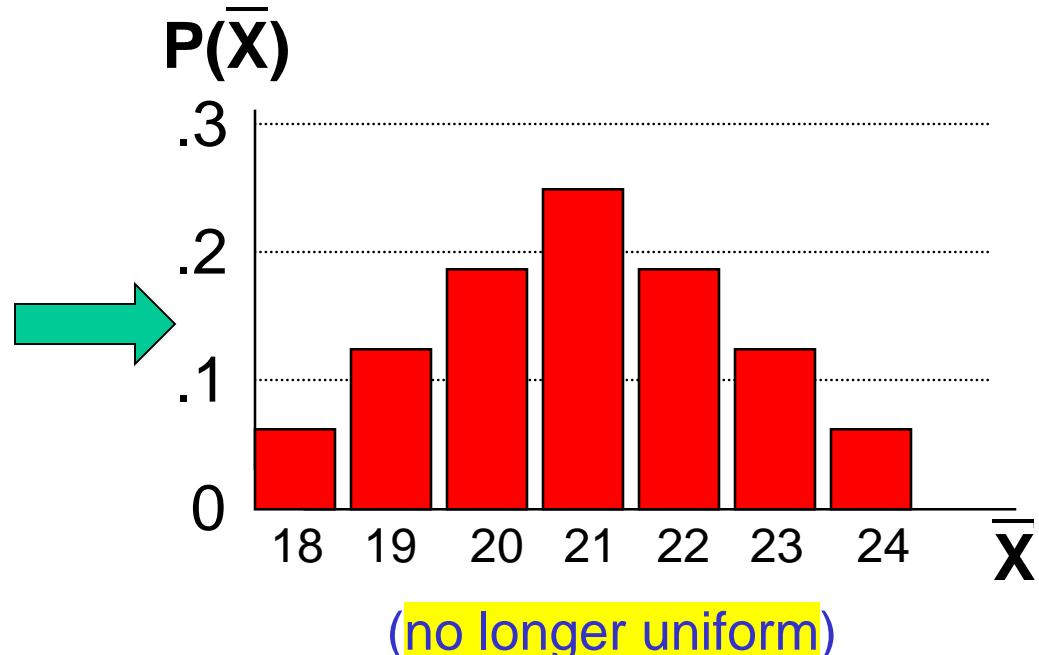
DCOVA

Sampling Distribution of All Sample Means

16 Sample Means

| 1st Obs | 2nd Observation | | | |
|------------|-----------------|----|----|----|
| | 18 | 20 | 22 | 24 |
| 18 | 18 | 19 | 20 | 21 |
| 20 | 19 | 20 | 21 | 22 |
| 22 | 20 | 21 | 22 | 23 |
| 24 | 21 | 22 | 23 | 24 |

Sample Means
Distribution



Developing A Sampling Distribution

(continued)

DCOVA

Summary Measures of this Sampling Distribution:

$$\mu_{\bar{x}} = \frac{18+19+19+\cdots+24}{16} = 21$$

$$\sigma_{\bar{x}} = \sqrt{\frac{(18-21)^2 + (19-21)^2 + \cdots + (24-21)^2}{16}} = 1.58$$

Note: Here we divide by 16 because there are 16
different samples of size 2.

Comparing the Population Distribution to the Sample Means Distribution

DCOV A

Population

$N = 4$

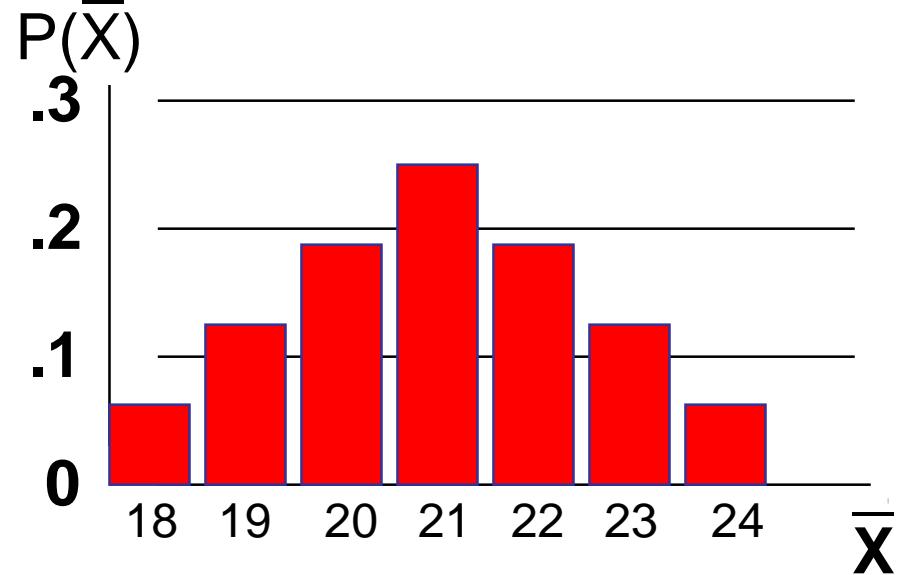
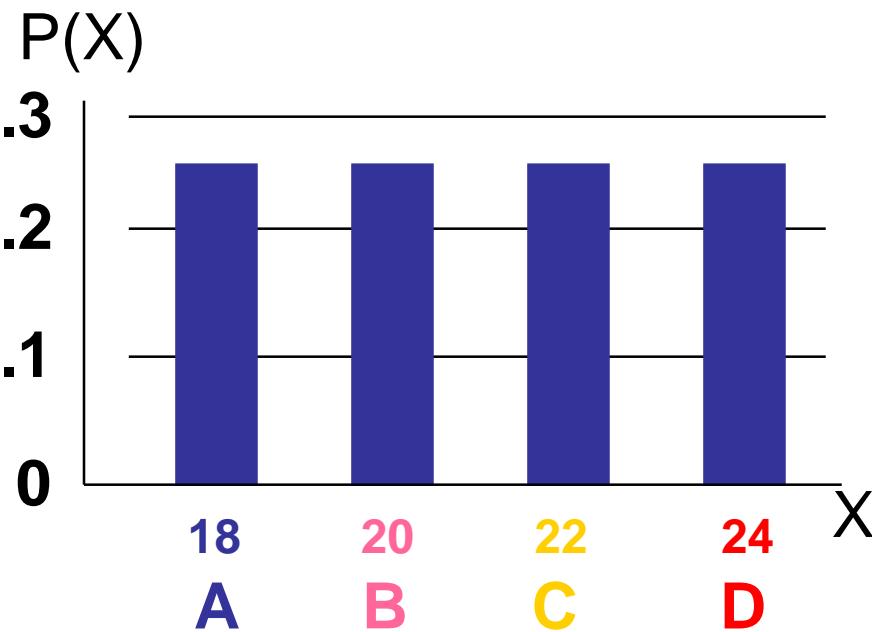
$\mu = 21$

$\sigma = 2.236$

Sample Means Distribution

$n = 2$

$\mu_{\bar{X}} = 21$ $\sigma_{\bar{X}} = 1.58$



Sample Mean Sampling Distribution: Standard Error of the Mean

DCOVA

- Different samples of the same size from the same population will yield different sample means
- A measure of the variability in the mean from sample to sample is given by the **Standard Error of the Mean**:
(This assumes that sampling is with replacement or sampling is without replacement from an infinite population)

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- Note that the standard error of the mean decreases as the sample size increases

Sample Mean Sampling Distribution: If the Population is Normal

DCOVA

- If a population is normal with mean μ and standard deviation σ , the sampling distribution of \bar{X} is also normally distributed with

$$\mu_{\bar{X}} = \mu$$

and

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Z-value for Sampling Distribution of the Mean

DCOVA

- Z-value for the sampling distribution of \bar{X} :

$$Z = \frac{(\bar{X} - \mu_{\bar{X}})}{\sigma_{\bar{X}}} = \frac{(\bar{X} - \mu)}{\frac{\sigma}{\sqrt{n}}}$$

where:

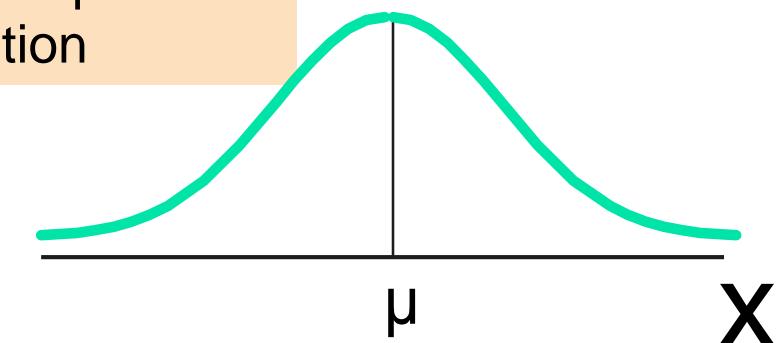
- \bar{X} = sample mean
- μ = population mean
- σ = population standard deviation
- n = sample size

Sampling Distribution Properties

DCOVA

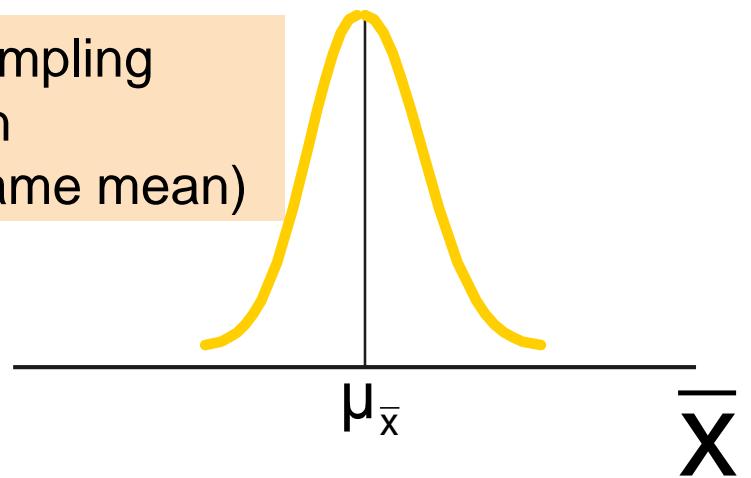
$$\mu_{\bar{x}} = \mu$$

Normal Population
Distribution



(i.e. \bar{X} is unbiased)

Normal Sampling
Distribution
(has the same mean)



Sampling Distribution Properties

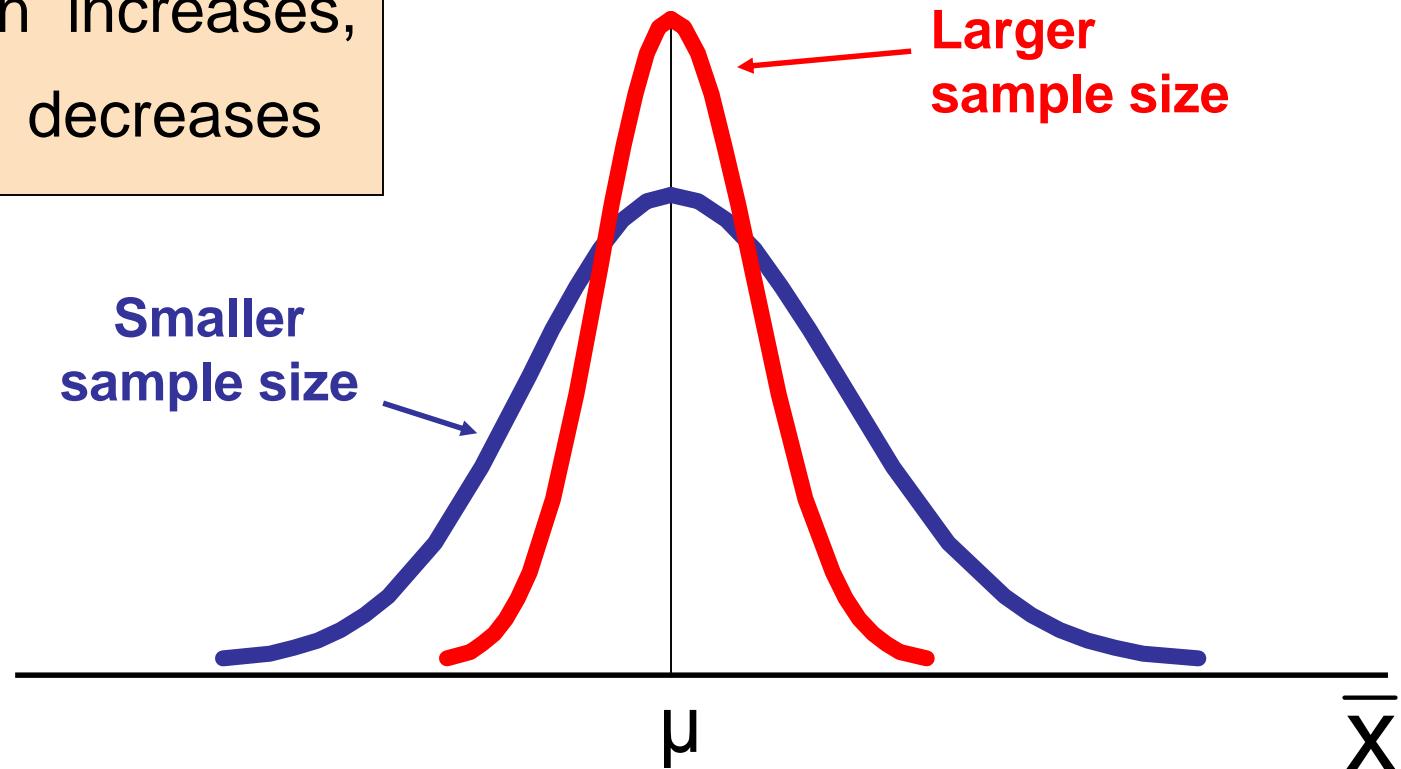
(continued)

DCOV A

As n increases,
 $\sigma_{\bar{X}}$ decreases

Smaller
sample size

Larger
sample size



Determining An Interval Including A Fixed Proportion of the Sample Means Example

DCOVA

Find a symmetrically distributed interval around μ that will include 95% of the sample means when $\mu = 368$, $\sigma = 15$, and $n = 25$.

- Since the interval contains 95% of the sample means 5% of the sample means will be outside the interval
- Since the interval is symmetric 2.5% will be above the upper limit and 2.5% will be below the lower limit.
- From the standardized normal table, the Z score with 2.5% (0.0250) below it is -1.96 and the Z score with 2.5% (0.0250) above it is 1.96.

Determining An Interval Including A Fixed Proportion of the Sample Means

(continued)

DCOVA

- Calculating the lower limit of the interval

$$\bar{X}_L = \mu + Z \frac{\sigma}{\sqrt{n}} = 368 + (-1.96) \frac{15}{\sqrt{25}} = 362.12$$

- Calculating the upper limit of the interval

$$\bar{X}_U = \mu + Z \frac{\sigma}{\sqrt{n}} = 368 + (1.96) \frac{15}{\sqrt{25}} = 373.88$$

- 95% of all sample means of sample size 25 are between 362.12 and 373.88

Sample Mean Sampling Distribution: If the Population is **not** Normal

DCOVA

- We can apply the **Central Limit Theorem**:
 - Even if the population is **not normal**,
 - ...sample means from the population **will be approximately normal** as long as the sample size is large enough.
-

Properties of the sampling distribution:

$$\mu_{\bar{x}} = \mu$$

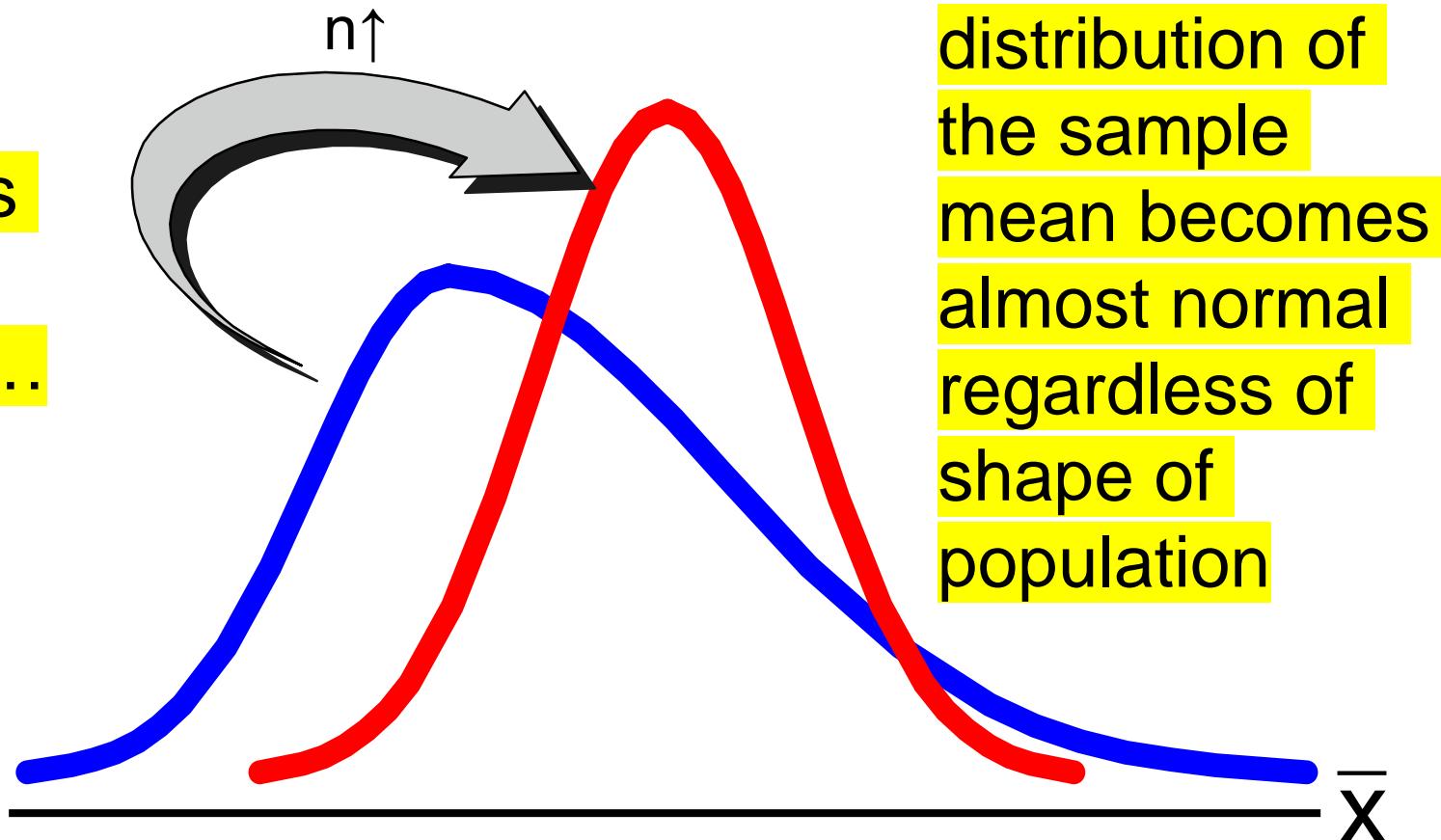
and

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Central Limit Theorem

DCOVA

As the sample size gets large enough...



the sampling distribution of the sample mean becomes almost normal regardless of shape of population

Sample Mean Sampling Distribution: If the Population is not Normal

(continued)

Sampling distribution properties:

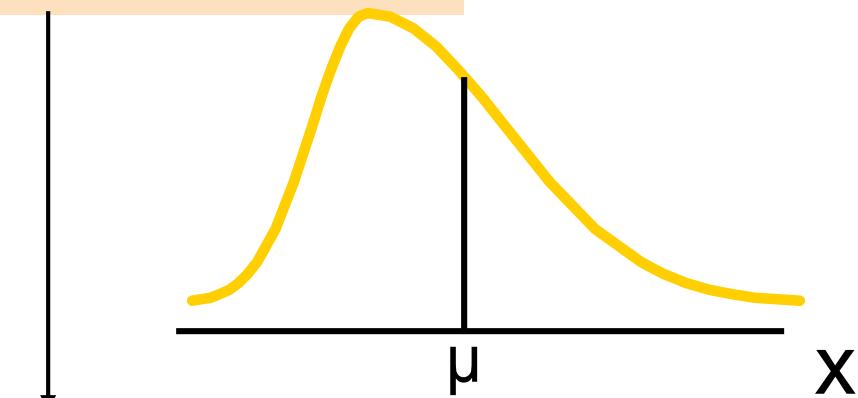
Central Tendency

$$\mu_{\bar{x}} = \mu$$

Variation

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Population Distribution

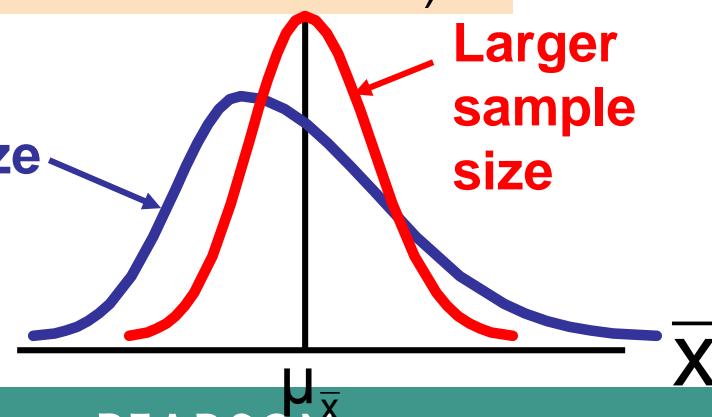


Sampling Distribution

(becomes normal as n increases)

Smaller sample size

Larger sample size



How Large is Large Enough?

DCOVA

- For most distributions, $n > 30$ will give a sampling distribution that is nearly normal
- For fairly symmetric distributions, $n > 15$
- For a normal population distribution, the sampling distribution of the mean is always normally distributed

Example

DCOVA

- Suppose a population has mean $\mu = 8$ and standard deviation $\sigma = 3$. Suppose a random sample of size $n = 36$ is selected.
- What is the probability that the sample mean is between 7.8 and 8.2?

Example

(continued)

DCOVA

Solution:

- Even if the population is not normally distributed, the central limit theorem can be used ($n > 30$)
- ... so the sampling distribution of \bar{X} is approximately normal
- ... with mean $\mu_{\bar{x}} = 8$
- ...and standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{36}} = 0.5$

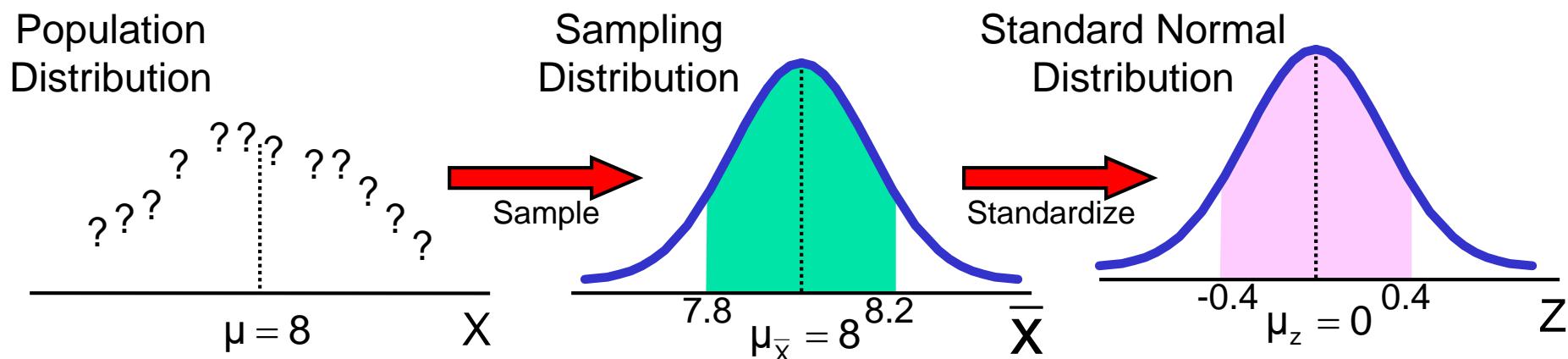
Example

(continued)

Solution (continued):

DCOVA

$$\begin{aligned} P(7.8 < \bar{X} < 8.2) &= P\left(\frac{7.8-8}{\sqrt{36}} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < \frac{8.2-8}{\sqrt{36}}\right) \\ &= P(-0.4 < Z < 0.4) = 0.6554 - 0.3446 = \boxed{0.3108} \end{aligned}$$



Population Proportions

END OF CHAPTER

DCOVA

π = the proportion of the population having some characteristic

- Sample proportion (p) provides an estimate of π :

$$p = \frac{X}{n} = \frac{\text{number of items in the sample having the characteristic of interest}}{\text{sample size}}$$

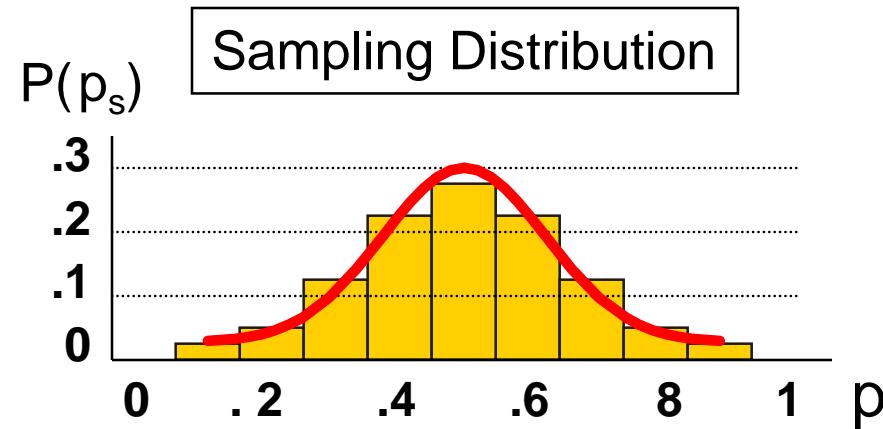
- $0 \leq p \leq 1$
- p is approximately distributed as a normal distribution when n is large
(assuming sampling with replacement from a finite population or without replacement from an infinite population)

Sampling Distribution of p

DCOV A

- Approximated by a normal distribution if:

- - $n\pi \geq 5$
 - and
 - $n(1 - \pi) \geq 5$



where

$$\mu_p = \pi$$

and

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

(where π = population proportion)

Z-Value for Proportions

DCOVA

Standardize p to a Z value with the formula:

$$Z = \frac{p - \pi}{\sigma_p} = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

Example

DCOVA

- If the true proportion of voters who support Proposition A is $\pi = 0.4$, what is the probability that a sample of size 200 yields a sample proportion between 0.40 and 0.45?

- i.e.: **if $\pi = 0.4$ and $n = 200$, what is $P(0.40 \leq p \leq 0.45)$?**

Example

(continued)

- if $\pi = 0.4$ and $n = 200$, what is $P(0.40 \leq p \leq 0.45)$?

DCOVA

Find σ_p :
$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.4(1-0.4)}{200}} = 0.03464$$

Convert to
standardized
normal:

$$\begin{aligned} P(0.40 \leq p \leq 0.45) &= P\left(\frac{0.40 - 0.40}{0.03464} \leq Z \leq \frac{0.45 - 0.40}{0.03464}\right) \\ &= P(0 \leq Z \leq 1.44) \end{aligned}$$

Example

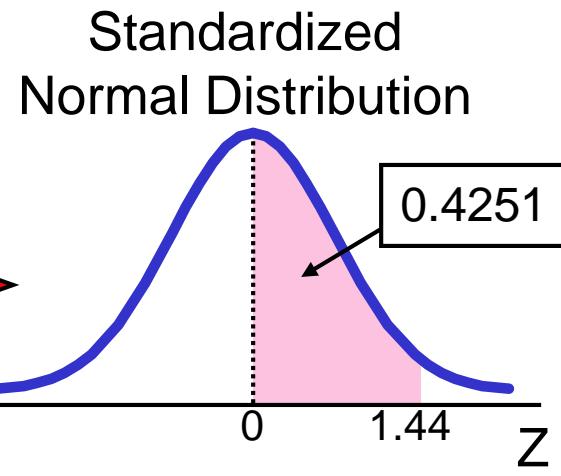
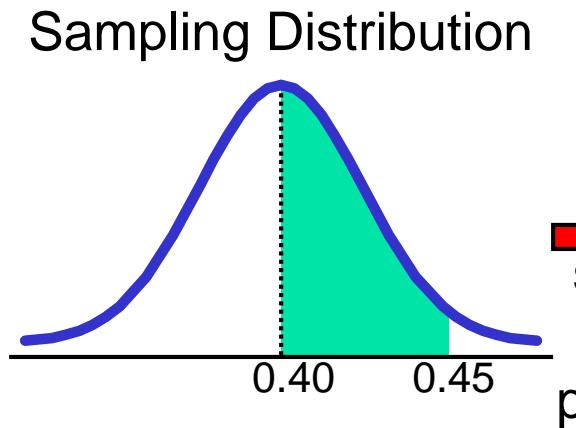
(continued)

- if $\pi = 0.4$ and $n = 200$, what is $P(0.40 \leq p \leq 0.45)$?

DCOVA

Utilize the cumulative normal table:

$$P(0 \leq Z \leq 1.44) = 0.9251 - 0.5000 = 0.4251$$



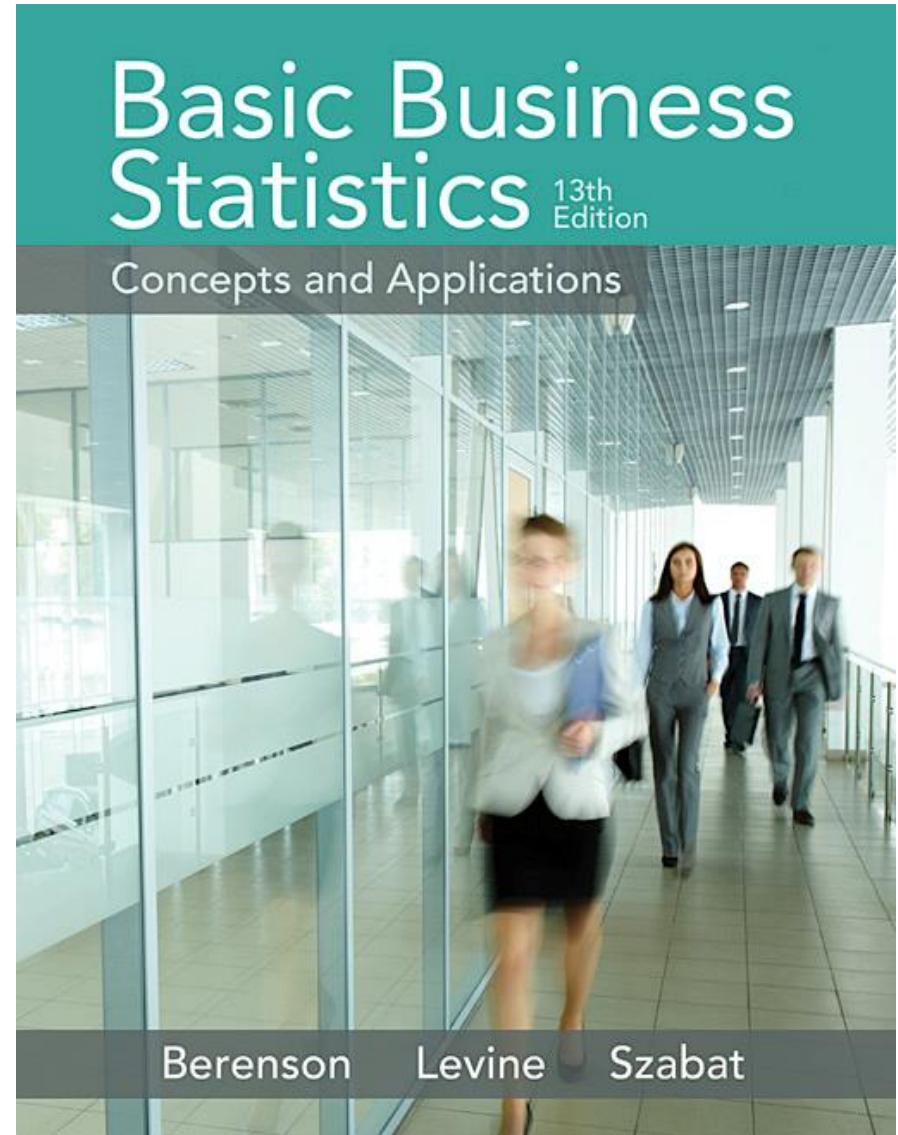
Chapter Summary

In this chapter we discussed

- Sampling distributions
- The sampling distribution of the mean
 - For normal populations
 - Using the Central Limit Theorem
- The sampling distribution of a proportion
- Calculating probabilities using sampling distributions

Chapter 8

Confidence Interval Estimation



Learning Objectives

In this chapter, you learn:

- To construct and interpret confidence interval estimates for the population mean and the population proportion
- To determine the sample size necessary to develop a confidence interval for the population mean or population proportion

Chapter Outline

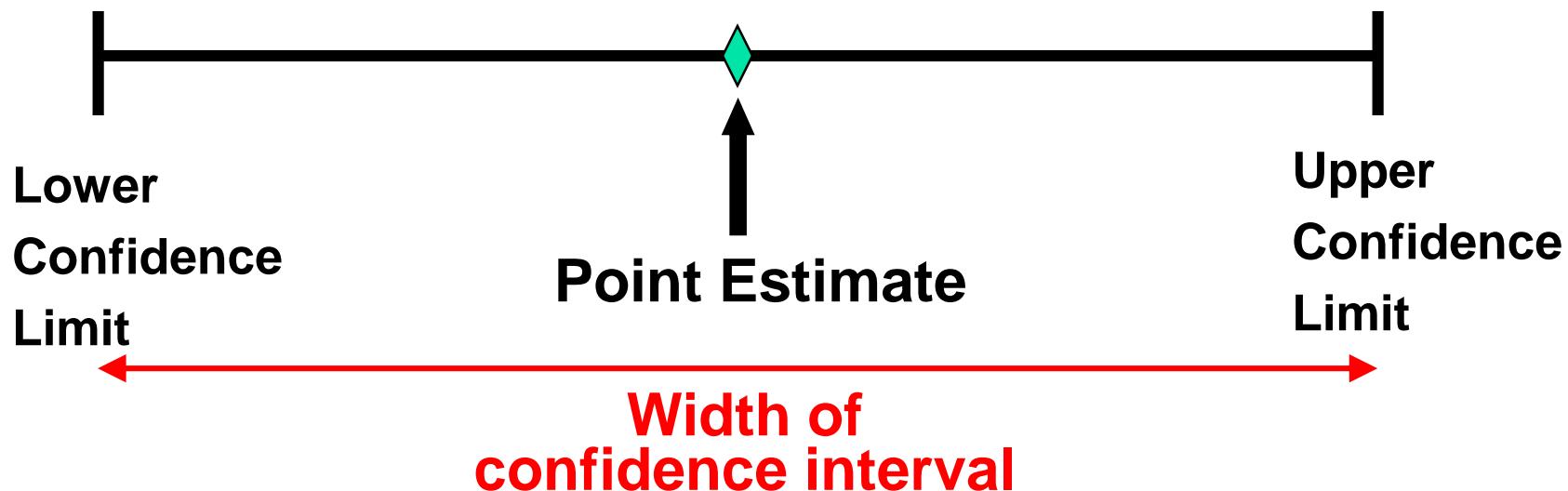
Content of this chapter

- Confidence Intervals for the Population Mean, μ
 - when Population Standard Deviation σ is Known
 - when Population Standard Deviation σ is Unknown
- Confidence Intervals for the Population Proportion, π
- Determining the Required Sample Size

Point and Interval Estimates

DCOVA

- A point estimate is a single number,
- a confidence interval provides additional information about the variability of the estimate



Point Estimates

DCOVA

| | | |
|--|-------|---|
| We can estimate a Population Parameter ... | | with a Sample Statistic (a Point Estimate) |
| Mean | μ | \bar{X} |
| Proportion | π | p |

Confidence Intervals

DCOVA

- How much uncertainty is associated with a point estimate of a population parameter?
- An **interval estimate** provides more information about a population characteristic than does a point estimate
- Such interval estimates are called **confidence intervals**

Confidence Interval Estimate

DCOVA

- An interval gives a range of values:
 - Takes into consideration variation in sample statistics from sample to sample
 - Based on observations from 1 sample
 - Gives information about closeness to unknown population parameters
 - Stated in terms of level of confidence
 - e.g. 95% confident, 99% confident
 - Can never be 100% confident

Confidence Interval Example

DCOVA

Cereal fill example

- Population has $\mu = 368$ and $\sigma = 15$.
- If you take a sample of size $n = 25$ you know
 - $368 \pm 1.96 * 15 / \sqrt{25} = (362.12, 373.88)$. 95% of the intervals formed in this manner will contain μ .
 - When you don't know μ , you use \bar{X} to estimate μ
 - If $\bar{X} = 362.3$ the interval is $362.3 \pm 1.96 * 15 / \sqrt{25} = (356.42, 368.18)$
 - Since $356.42 \leq \mu \leq 368.18$, the interval based on this sample makes a correct statement about μ .

But what about the intervals from other possible samples of size 25? Ans: μ will be in the interval 95% of the time

Confidence Interval Example

(continued)
DCOVA

| Sample # | \bar{X} | Lower Limit | Upper Limit | Contain μ ? |
|----------|-----------|-------------|-------------|-----------------|
| 1 | 362.30 | 356.42 | 368.18 | Yes |
| 2 | 369.50 | 363.62 | 375.38 | Yes |
| 3 | 360.00 | 354.12 | 365.88 | No |
| 4 | 362.12 | 356.24 | 368.00 | Yes |
| 5 | 373.88 | 368.00 | 379.76 | Yes |

Confidence Interval Example

(continued)

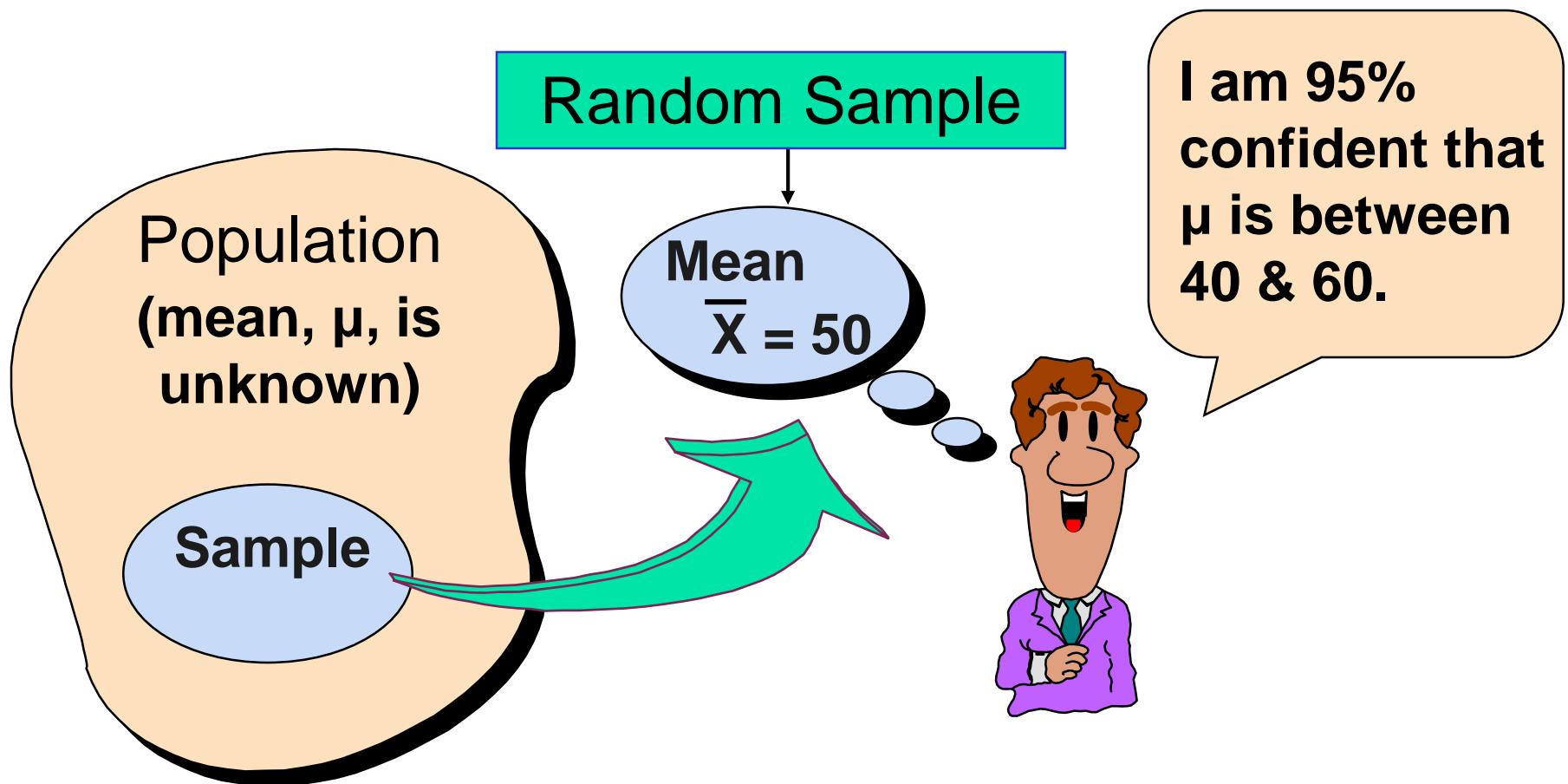
DCOVA

- In practice you only take one sample of size n
- In practice you do not know μ so you do not know if the interval actually contains μ
- However you do know that 95% of the intervals formed in this manner will contain μ
- Thus, based on the one sample, you actually selected you can be 95% confident your interval will contain μ (this is a 95% confidence interval)

Note: 95% confidence is based on the fact that we used $Z = 1.96$.

Estimation Process

DCOVA



General Formula

DCOVA

- The general formula for all confidence intervals is:

Point Estimate \pm (Critical Value)(Standard Error)

Where:

- Point Estimate is the sample statistic estimating the population parameter of interest
- Critical Value is a table value based on the sampling distribution of the point estimate and the desired confidence level
- Standard Error is the standard deviation of the point estimate

Confidence Level

DCOVA

- Confidence the interval will contain the unknown population parameter
- A percentage (less than 100%)

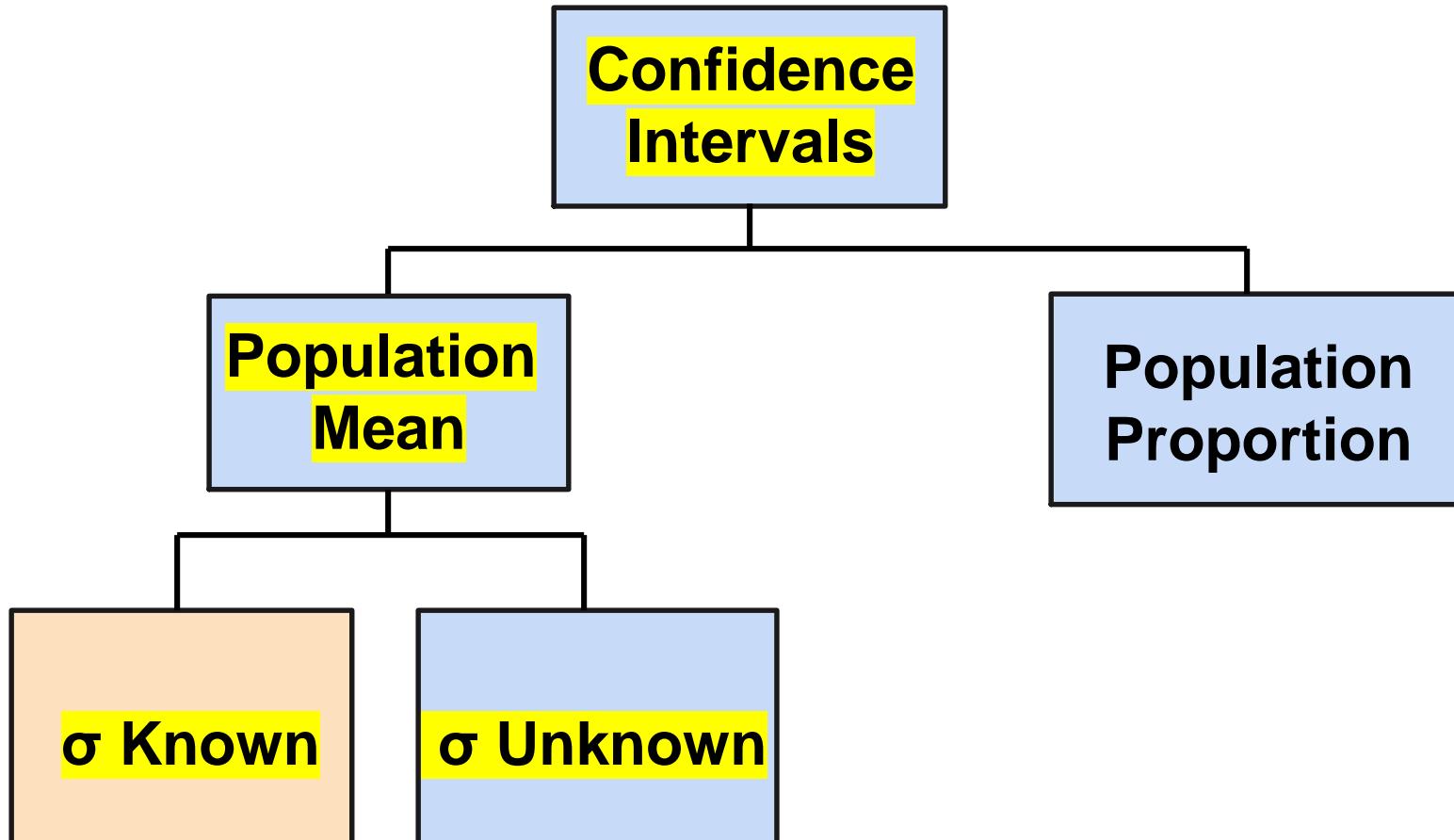
Confidence Level, $(1-\alpha)$ (continued)

DCOV A

- Suppose confidence level = 95%
- Also written $(1 - \alpha) = 0.95$, (so $\alpha = 0.05$)
- A relative frequency interpretation:
 - 95% of all the confidence intervals that can be constructed will contain the unknown true parameter
- A specific interval either will contain or will not contain the true parameter
 - No probability involved in a specific interval

Confidence Intervals

DCOVA



Confidence Interval for μ (σ Known)

DCOVA

- Assumptions

- Population standard deviation σ is known
- Population is normally distributed
- If population is not normal, use large sample ($n > 30$)

- Confidence interval estimate:

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

where \bar{X} is the point estimate

$Z_{\alpha/2}$ is the normal distribution critical value for a probability of $\alpha/2$ in each tail

σ/\sqrt{n} is the standard error

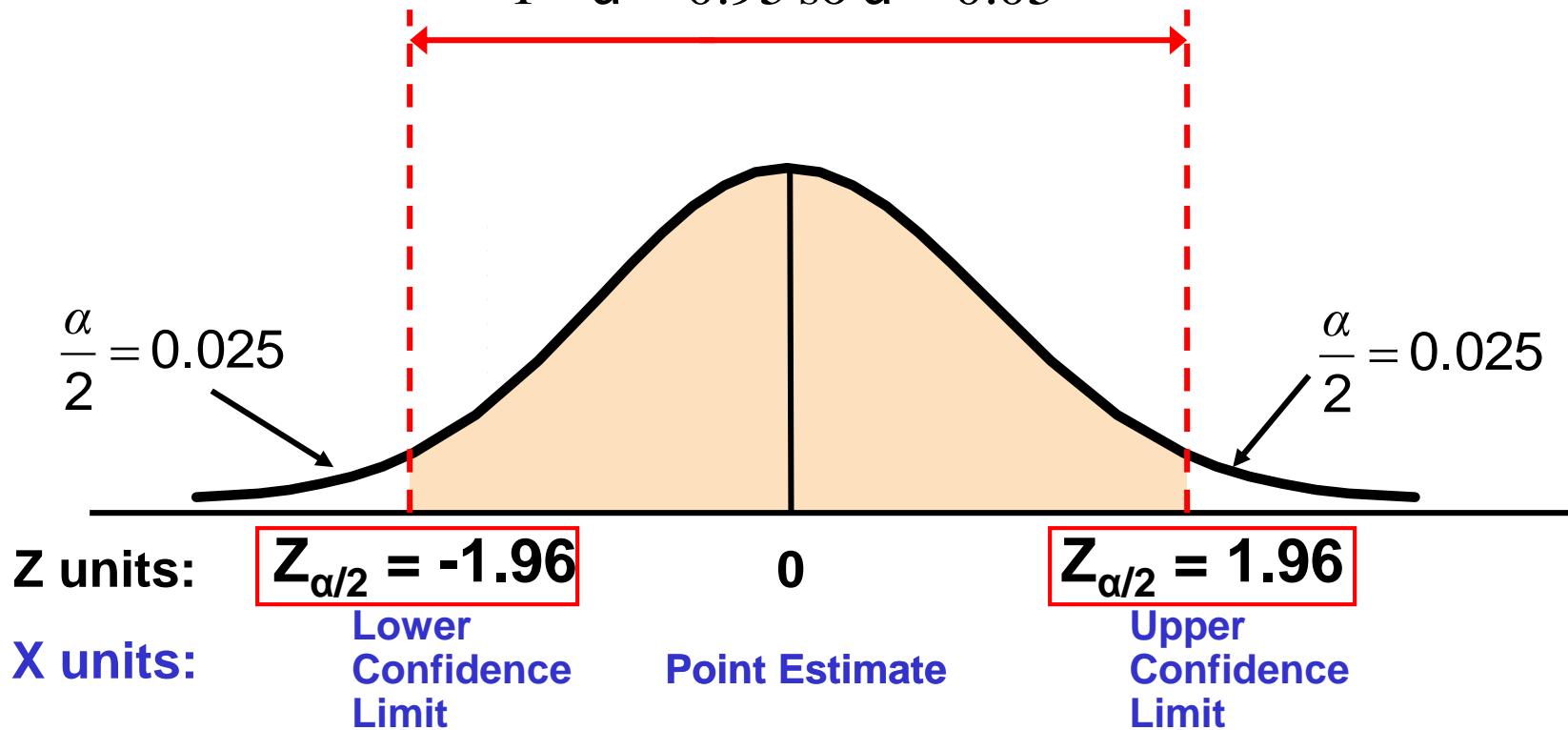
Finding the Critical Value, $Z_{\alpha/2}$

DCOV A

- Consider a 95% confidence interval:

$$Z_{\alpha/2} = \pm 1.96$$

$$1 - \alpha = 0.95 \text{ so } \alpha = 0.05$$



Common Levels of Confidence

[Important]

DCOVA

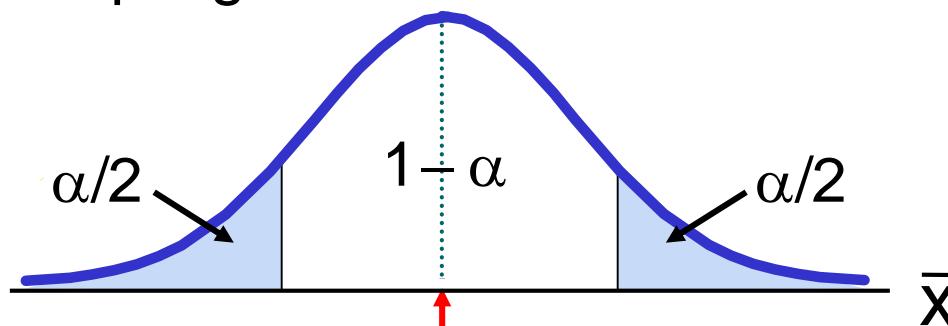
- Commonly used confidence levels are 90%, 95%, and 99%

| <i>Confidence Level</i> | <i>Confidence Coefficient, $1 - \alpha$</i> | $Z_{\alpha/2}$ value |
|--------------------------------|---|--|
| 80% | 0.80 | 1.28 |
| 90% | 0.90 | 1.645 |
| 95% | 0.95 | 1.96 |
| 98% | 0.98 | 2.33 |
| 99% | 0.99 | 2.58 |
| 99.8% | 0.998 | 3.08 |
| 99.9% | 0.999 | 3.27 |

Intervals and Level of Confidence

DCOVA

Sampling Distribution of the Mean

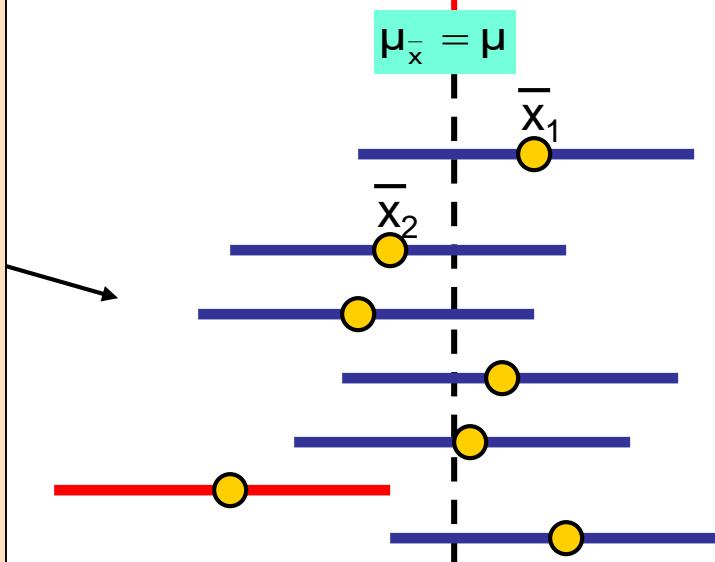


Intervals extend from

$$\bar{X} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

to

$$\bar{X} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$



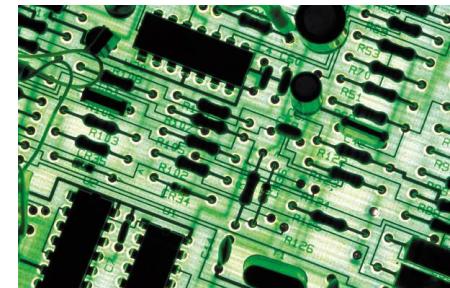
Confidence Intervals

(1-\alpha)100% of intervals constructed contain μ ; eg: 95%
(\alpha)100% do not. Eg: 5%

Example

DCOVA

- A sample of 11 circuits from a large normal population has a mean resistance of 2.20 ohms. We know from past testing that the population standard deviation is 0.35 ohms.
 - Determine a 95% confidence interval for the true mean resistance of the population.



Example

DCOVA

(continued)

- A sample of 11 circuits from a large normal population has a mean resistance of 2.20 ohms. We know from past testing that the population standard deviation is 0.35 ohms.

- **Solution:**

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$= 2.20 \pm 1.96(0.35/\sqrt{11})$$

$$= 2.20 \pm 0.2068$$

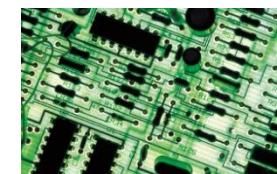
$$1.9932 \leq \mu \leq 2.4068$$



Interpretation

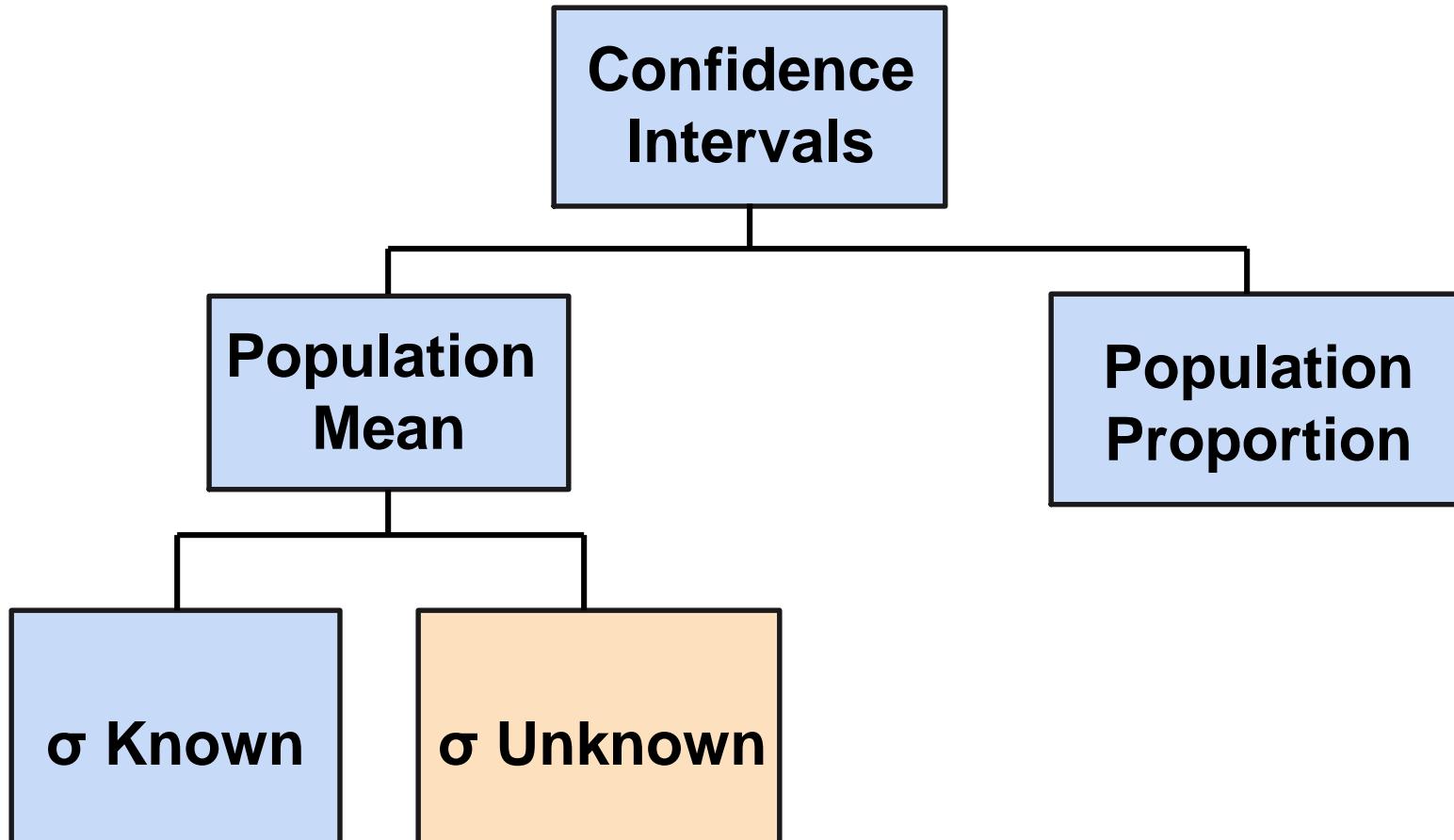
DCOVA

- We are 95% confident that the true mean resistance is between 1.9932 and 2.4068 ohms
- Although the true mean may or may not be in this interval, 95% of intervals formed in this manner will contain the true mean



Confidence Intervals

DCOVA



Do You Ever Truly Know σ ?

- Probably not!
- In virtually all real-world business situations, σ is not known.
- If there is a situation where σ is known, then μ is also known (since to calculate σ you need to know μ .)
- If you truly know μ there would be no need to gather a sample to estimate it.

Confidence Interval for μ (σ Unknown)

DCOVA

- If the population standard deviation σ is unknown, we can substitute the sample standard deviation, S
- This introduces extra uncertainty, since S is variable from sample to sample
- So we use the t distribution instead of the normal distribution

Confidence Interval for μ (σ Unknown)

(continued)

DCOV A

- Assumptions

- Population standard deviation is unknown
- Population is normally distributed
- If population is not normal, use large sample ($n > 30$)

- Use Student's t Distribution

- Confidence Interval Estimate:

$$\bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

(where $t_{\alpha/2}$ is the critical value of the t distribution with $n - 1$ degrees of freedom and an area of $\alpha/2$ in each tail)

Student's t Distribution

DCOVA

- The t is a family of distributions
- The $t_{\alpha/2}$ value depends on degrees of freedom (d.f.)
 - Number of observations that are free to vary after sample mean has been calculated

$$\text{d.f.} = n - 1$$

Degrees of Freedom (df)

DCOVA

Idea: Number of observations that are free to vary after sample mean has been calculated

Example: Suppose the mean of 3 numbers is 8.0

Let $X_1 = 7$

Let $X_2 = 8$

What is X_3 ?

If the mean of these three values is 8.0,
then X_3 must be 9
(i.e., X_3 is not free to vary)



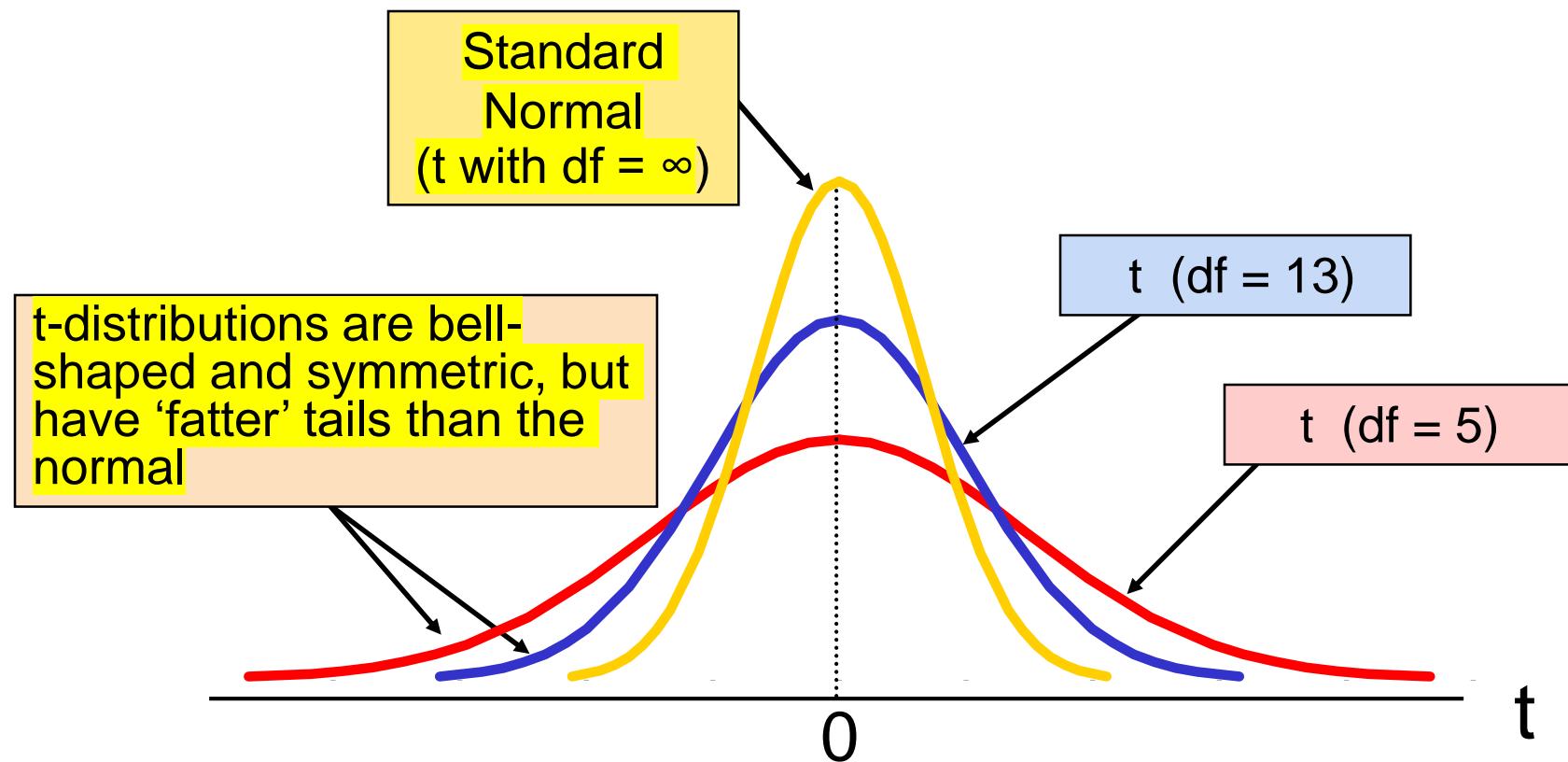
Here, $n = 3$, so degrees of freedom = $n - 1 = 3 - 1 = 2$

(2 values can be any numbers, but the third is not free to vary for a given mean)

Student's t Distribution

DCOV A

Note: $t \rightarrow Z$ as n increases



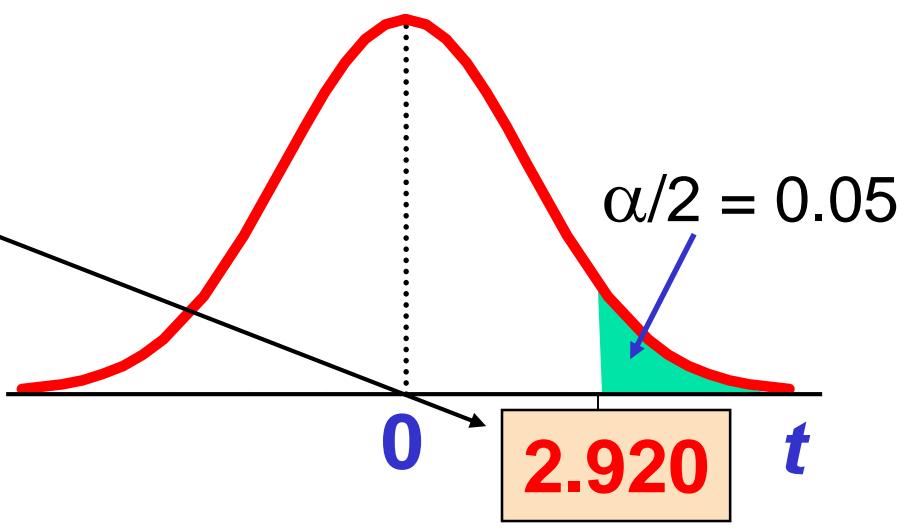
Student's t Table

DCOVA

| | | Upper Tail Area | | |
|----|-------|-----------------|-------|--------|
| | | .10 | .05 | .025 |
| df | 1 | 3.078 | 6.314 | 12.706 |
| 2 | 1.886 | 2.920 | | 4.303 |
| 3 | 1.638 | 2.353 | | 3.182 |

The body of the table
contains t values, not
probabilities

Let: $n = 3$
 $df = n - 1 = 2$
 $\alpha = 0.10$
 $\alpha/2 = 0.05$



Selected t distribution values

DCOVA

With comparison to the Z value

| Confidence Level | t (10 d.f.) | t (20 d.f.) | t (30 d.f.) | Z $(\infty \text{ d.f.})$ |
|-----------------------------|------------------------|------------------------|------------------------|---|
| 0.80 | 1.372 | 1.325 | 1.310 | 1.28 |
| 0.90 | 1.812 | 1.725 | 1.697 | 1.645 |
| 0.95 | 2.228 | 2.086 | 2.042 | 1.96 |
| 0.99 | 3.169 | 2.845 | 2.750 | 2.58 |

Note: $t \rightarrow Z$ as n increases

Example of t distribution confidence interval

DCOVA

A random sample of $n = 25$ has $\bar{X} = 50$ and $S = 8$. Form a 95% confidence interval for μ

- d.f. = $n - 1 = 24$, so $t_{\alpha/2} = t_{0.025} = 2.0639$

The confidence interval is

$$\bar{X} \pm t_{\alpha/2} \frac{S}{\sqrt{n}} = 50 \pm (2.0639) \frac{8}{\sqrt{25}}$$

$$46.698 \leq \mu \leq 53.302$$

Example of t distribution confidence interval

(continued)

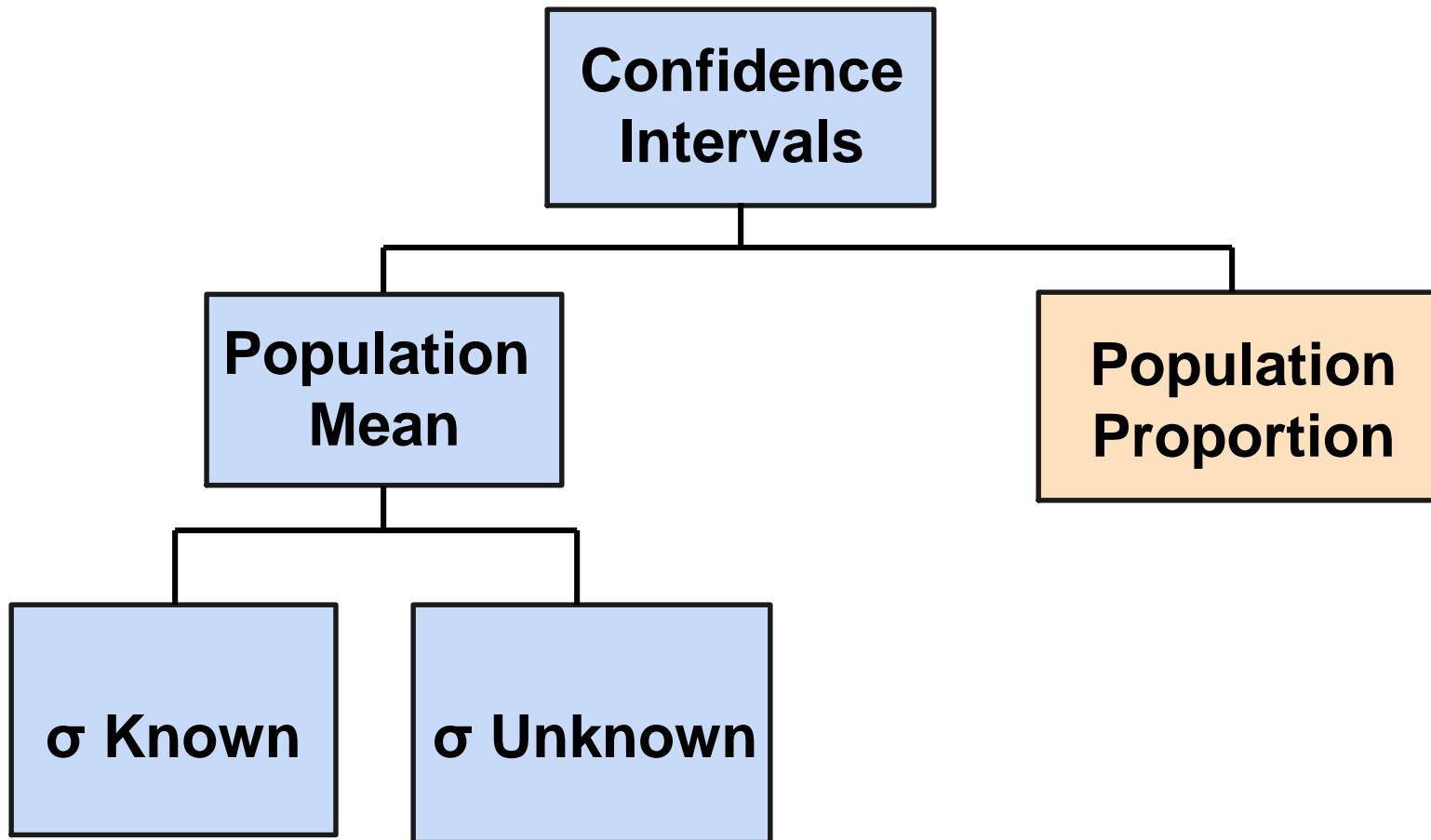
DCOVA

- Interpreting this interval requires the assumption that the population you are sampling from is approximately a normal distribution (especially since n is only 25).
- This condition can be checked by creating a:
 - Normal probability plot or
 - Boxplot

END OF CHAPTER

Confidence Intervals

DCOVA



Confidence Intervals for the Population Proportion, π

DCOVA

- An interval estimate for the population proportion (π) can be calculated by adding an allowance for uncertainty to the sample proportion (p)

Confidence Intervals for the Population Proportion, π

(continued)

- Recall that the distribution of the sample proportion is approximately normal if the sample size is large, with standard deviation DCOV A

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

- We will estimate this with sample data:

$$\sqrt{\frac{p(1-p)}{n}}$$

Confidence Interval Endpoints

DCOVA

- Upper and lower confidence limits for the population proportion are calculated with the formula

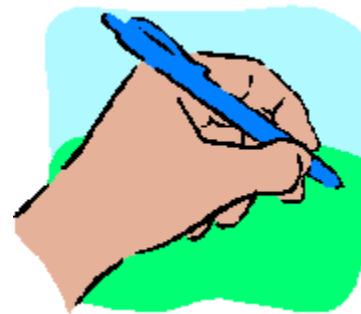
$$p \pm Z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

- where
 - $Z_{\alpha/2}$ is the standard normal value for the level of confidence desired
 - p is the sample proportion
 - n is the sample size
- Note: must have $np > 5$ and $n(1-p) > 5$

Example

DCOVA

- A random sample of 100 people shows that 25 are left-handed.
- Form a 95% confidence interval for the true proportion of left-handers



Example

DCOVA

(continued)

- A random sample of 100 people shows that 25 are left-handed. Form a 95% confidence interval for the true proportion of left-handers.

$$\begin{aligned} p \pm Z_{\alpha/2} \sqrt{p(1-p)/n} \\ = 25/100 \pm 1.96 \sqrt{0.25(0.75)/100} \\ = 0.25 \pm 1.96(0.0433) \\ = 0.1651 \leq \pi \leq 0.3349 \end{aligned}$$



Interpretation

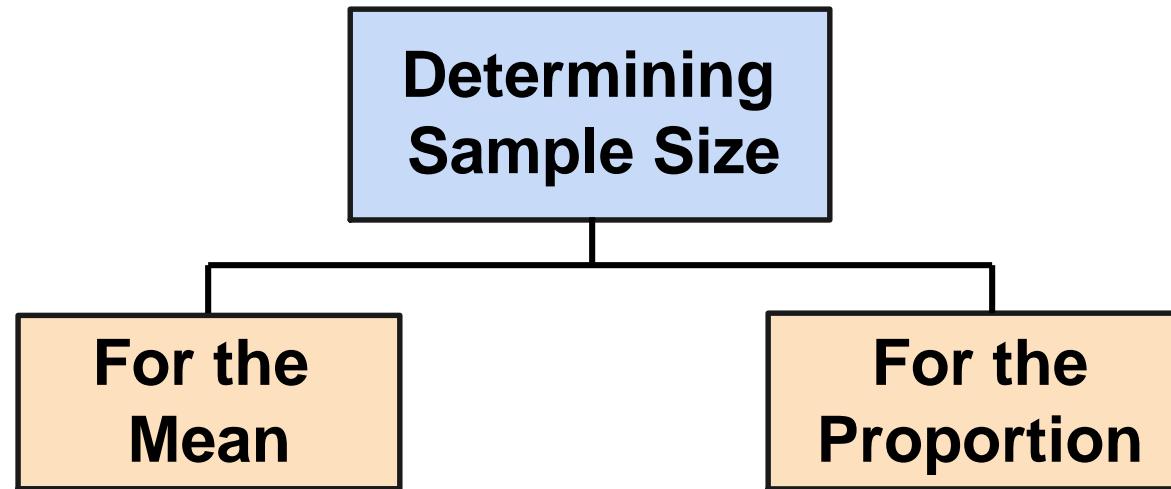
DCOVA

- We are 95% confident that the true percentage of left-handers in the population is between 16.51% and 33.49%.
- Although the interval from 0.1651 to 0.3349 may or may not contain the true proportion, 95% of intervals formed from samples of size 100 in this manner will contain the true proportion.



Determining Sample Size

DCOVA



Sampling Error

DCOVA

- The required sample size can be found to reach a desired margin of error (e) with a specified level of confidence ($1 - \alpha$)

- The margin of error is also called sampling error
 - the amount of imprecision in the estimate of the population parameter
 - the amount added and subtracted to the point estimate to form the confidence interval

Determining Sample Size

DCOVA

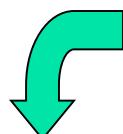
Determining Sample Size

For the Mean

Sampling error
(margin of error)

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$e = Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$



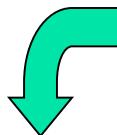
Determining Sample Size

(continued)

DCOVA

Determining Sample Size

For the Mean



$$e = Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Now solve
for n to get

$$n = \frac{Z_{\alpha/2}^2 \sigma^2}{e^2}$$

Determining Sample Size

(continued)

DCOVA

- To determine the required sample size for the mean, you must know:
 - The desired level of confidence ($1 - \alpha$), which determines the critical value, $Z_{\alpha/2}$
 - The acceptable sampling error, e
 - The standard deviation, σ

Required Sample Size Example

DCOVA

If $\sigma = 45$, what sample size is needed to estimate the mean within ± 5 with 90% confidence?

$$n = \frac{Z^2 \sigma^2}{e^2} = \frac{(1.645)^2(45)^2}{5^2} = 219.19$$

So the required sample size is **n = 220**

(Always round up)

If σ is unknown

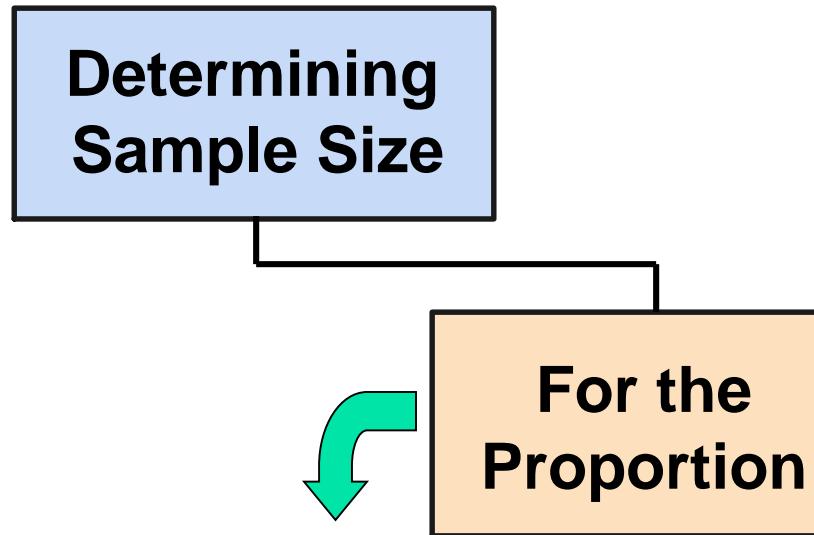
DCOVA

- If unknown, σ can be estimated when using the required sample size formula
 - Use a value for σ that is expected to be at least as large as the true σ
 - Select a pilot sample and estimate σ with the sample standard deviation, S

Determining Sample Size

(continued)

DCOVA



$$e = Z \sqrt{\frac{\pi(1-\pi)}{n}}$$

Now solve for n to get

$$n = \frac{Z_{\alpha/2}^2 \pi (1-\pi)}{e^2}$$

Determining Sample Size

(continued)

DCOVA

- To determine the required sample size for the proportion, you must know:
 - The desired level of confidence ($1 - \alpha$), which determines the critical value, $Z_{\alpha/2}$
 - The acceptable sampling error, e
 - The true proportion of events of interest, π
 - π can be estimated with a pilot sample if necessary (or conservatively use 0.5 as an estimate of π)

Required Sample Size Example

DCOVA

How large a sample would be necessary to estimate the true proportion defective in a large population **within $\pm 3\%$, with 95% confidence?**

(Assume a pilot sample yields $p = 0.12$)

Required Sample Size Example

(continued)

Solution:

DCOVA

For 95% confidence, use $Z_{\alpha/2} = 1.96$

$e = 0.03$

$p = 0.12$, so use this to estimate π

$$n = \frac{Z_{\alpha/2}^2 \pi (1 - \pi)}{e^2} = \frac{(1.96)^2 (0.12)(1 - 0.12)}{(0.03)^2} = 450.74$$

So use $n = 451$

Ethical Issues

- A confidence interval estimate (reflecting sampling error) should always be included when reporting a point estimate
- The level of confidence should always be reported
- The sample size should be reported
- An interpretation of the confidence interval estimate should also be provided

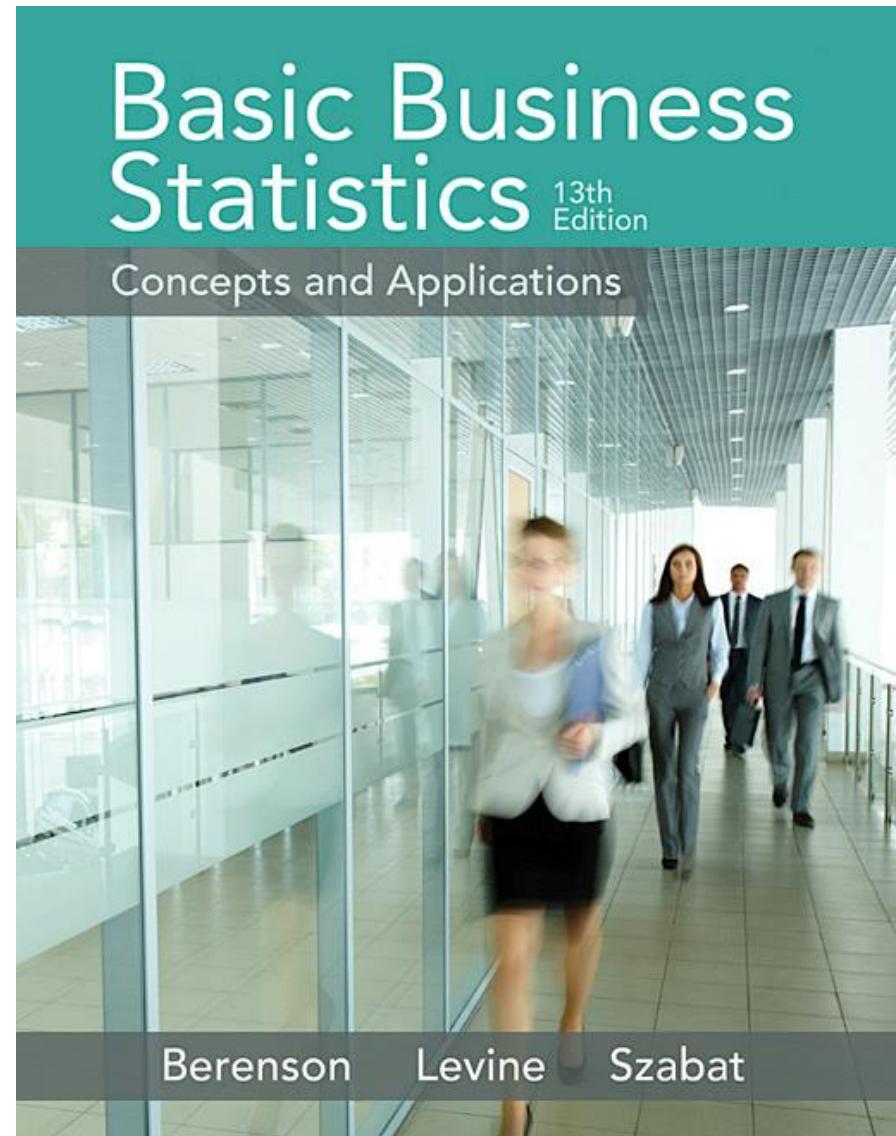
Chapter Summary

In this chapter we discussed

- The concept of confidence intervals
- Point estimates & confidence interval estimates
- Finding confidence interval estimates for the mean (σ known)
- Finding confidence interval estimates for the mean (σ unknown)
- Finding confidence interval estimates for the proportion
- Determining required sample size for mean and proportion
- Ethical issues in confidence interval estimation

Chapter 9

Fundamentals of Hypothesis Testing: One-Sample Tests



Learning Objectives

In this chapter, you learn:

- The basic principles of hypothesis testing
- How to use hypothesis testing to test a mean or proportion
- The assumptions of each hypothesis-testing procedure, how to evaluate them, and the consequences if they are seriously violated
- Pitfalls & ethical issues involved in hypothesis testing
- How to avoid the pitfalls involved in hypothesis testing

What is a Hypothesis?

DCOVA

- A hypothesis is a claim (assertion) about a population parameter:

- population mean

Example: The mean monthly cell phone bill in this city is $\mu = \$42$

- population proportion

Example: The proportion of adults in this city with cell phones is $\pi = 0.68$



The Null Hypothesis, H_0

DCOV A

- States the claim or assertion to be tested

Example: The mean diameter of a manufactured bolt is 30mm ($H_0 : \mu = 30$)

$$H_0 : \mu = 30$$

$$H_0 : \bar{X} = 30$$



The Null Hypothesis, H_0

DCOV**A**
(continued)

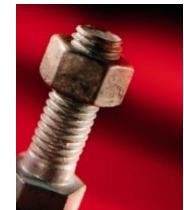
- Begin with the assumption that the null hypothesis is true
 - Similar to the notion of innocent until proven guilty
- Refers to the status quo or historical value
- Always contains “=”, or “≤”, or “≥” sign
- May or may not be rejected



The Alternative Hypothesis, H_1

DCOV A

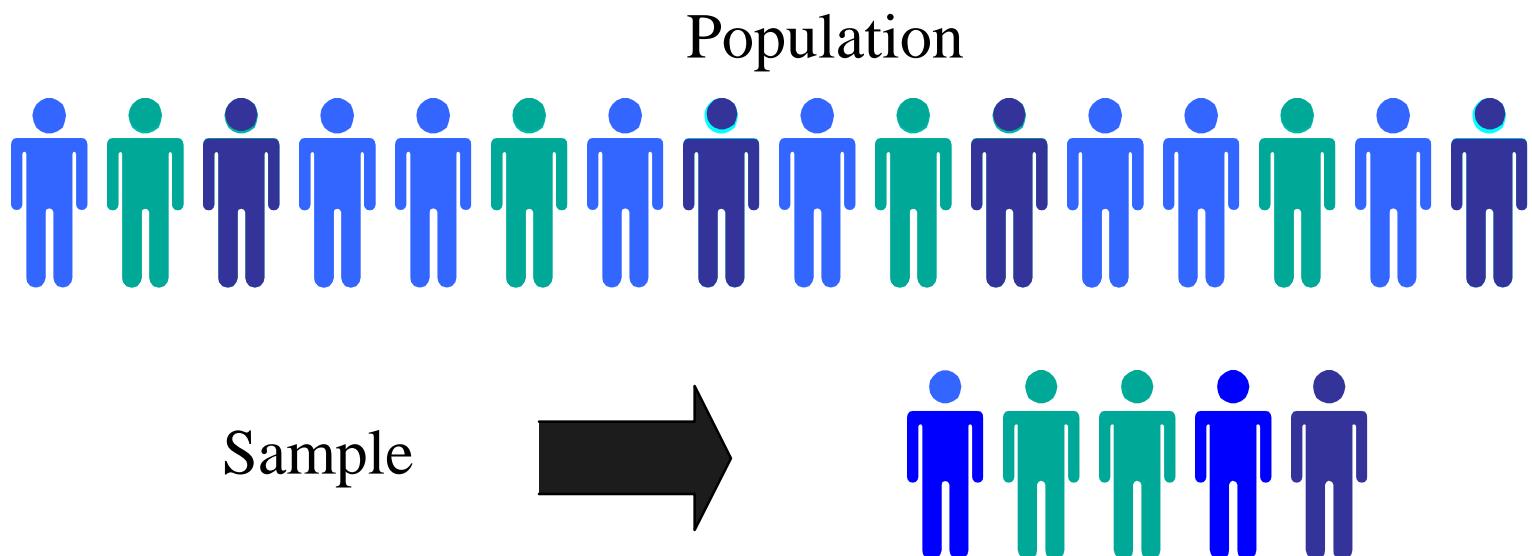
- Is the opposite of the null hypothesis
 - e.g., The average diameter of a manufactured bolt is not equal to 30mm ($H_1: \mu \neq 30$)
- Challenges the status quo
- Never contains the “=”, or “≤”, or “≥” sign
- May or may not be proven
- Is generally the hypothesis that the researcher is trying to prove



The Hypothesis Testing Process

DCOVA

- Claim: The population mean age is 50.
 - $H_0: \mu = 50$, $H_1: \mu \neq 50$
- Sample the population and find the sample mean.



The Hypothesis Testing

Process(continued)

DCOV A

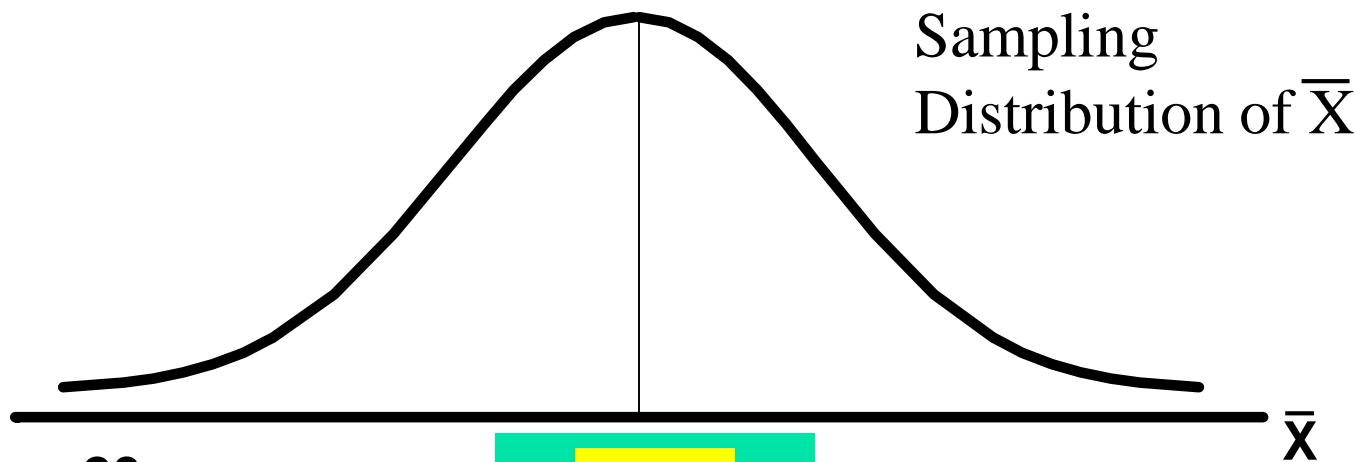
- Suppose the sample mean age was $\bar{X} = 20$.
- This is significantly lower than the claimed mean population age of 50.
- If the null hypothesis were true, the probability of getting such a different sample mean would be very small, **so you reject the null hypothesis.**
- In other words, getting a sample mean of 20 is so unlikely if the population mean was 50, you conclude that the population mean must not be 50.

The Hypothesis Testing

(continued)

Process

DCOVA



If it is unlikely that you would get a sample mean of this value ...

$\mu = 50$
If H_0 is true

... When in fact this were the population mean...

... then you reject the null hypothesis that $\mu = 50$.

The Test Statistic and Critical Values

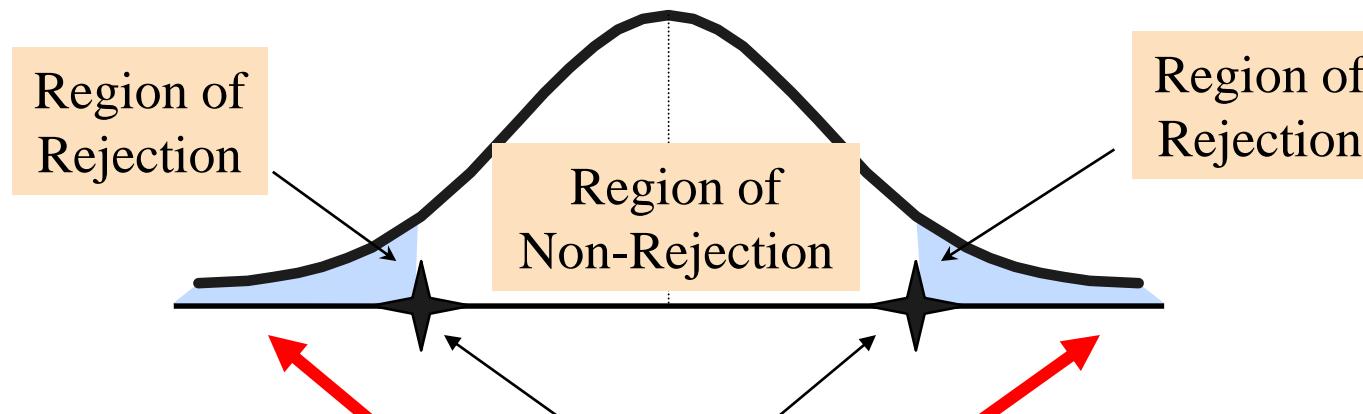
DCOV A

- If the sample mean is close to the stated population mean, the null hypothesis is not rejected.
- If the sample mean is far from the stated population mean, the null hypothesis is rejected.
- How far is “far enough” to reject H_0 ?
- The critical value of a test statistic creates a “line in the sand” for decision making -- it answers the question of how far is far enough.

The Test Statistic and Critical Values

DCOVA

Sampling Distribution of the test statistic



“Too Far Away” From Mean of Sampling Distribution

Possible Errors in Hypothesis Test Decision Making

DCOVA

- **Type I Error**

- Reject a true null hypothesis
- Considered a serious type of error
- The probability of a Type I Error is α

False Alarm

Alpha

- Called level of significance of the test
- Set by researcher in advance

- **Type II Error**

- Failure to reject a false null hypothesis
- The probability of a Type II Error is β

Missed Opportunity

Possible Errors in Hypothesis Test

Decision Making

DCOVA
(continued)

Possible Hypothesis Test Outcomes

| | | Actual Situation |
|---------------------|--------------------------------------|--------------------------------------|
| Decision | H_0 True | H_0 False |
| Do Not Reject H_0 | No Error Probability $1 - \alpha$ | Type II Error Probability β |
| Reject H_0 | Type I Error Probability α | No Error Power $1 - \beta$ |

Possible Errors in Hypothesis Test

Decision Making

DCOV
(continued)

- The confidence coefficient ($1-\alpha$) is the probability of not rejecting H_0 when it is true.
- The confidence level of a hypothesis test is $(1-\alpha)*100\%$.
- The power of a statistical test ($1-\beta$) is the probability of rejecting H_0 when it is false.

Type I & II Error Relationship

DCOV A

- Type I and Type II errors cannot happen at the same time
 - A Type I error can only occur if H_0 is true
 - A Type II error can only occur if H_0 is false

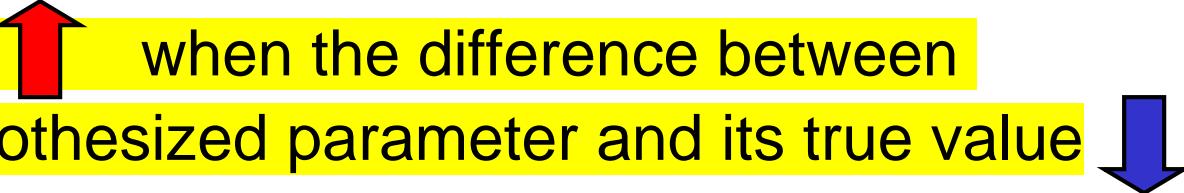
If Type I error probability (α)  , then

Type II error probability (β) 

Factors Affecting Type II Error

DCOV A

- All else equal,

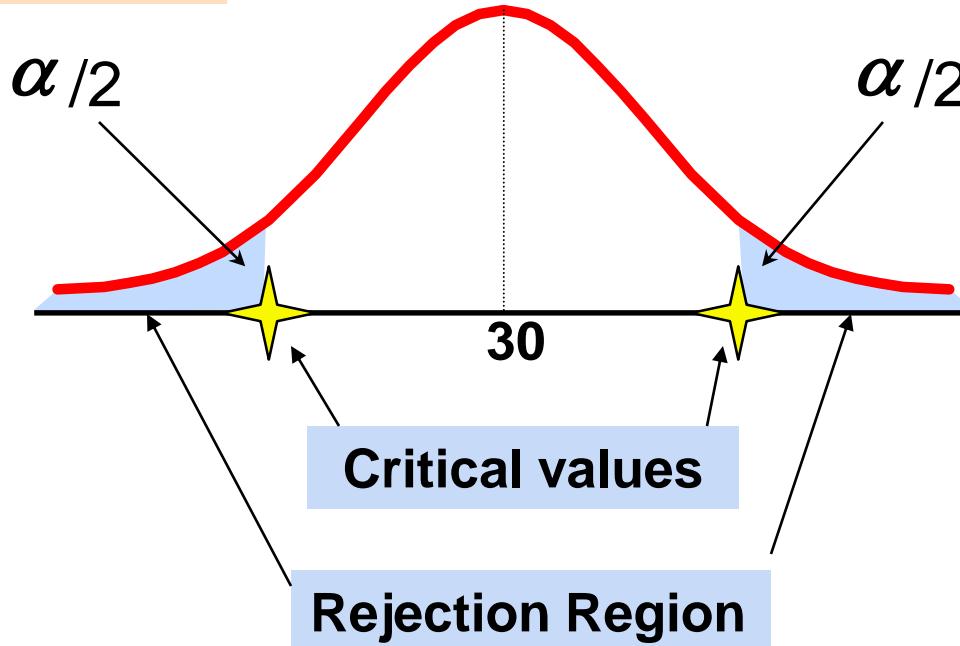
- $\beta \uparrow$ when the difference between hypothesized parameter and its true value 
- $\beta \uparrow$ when $\alpha \downarrow$ 
- $\beta \uparrow$ when $\sigma \uparrow$ 
- $\beta \uparrow$ when $n \downarrow$ 

Level of Significance and the Rejection Region

DCOV A

$$\begin{aligned} H_0: \mu &= 30 \\ H_1: \mu &\neq 30 \end{aligned}$$

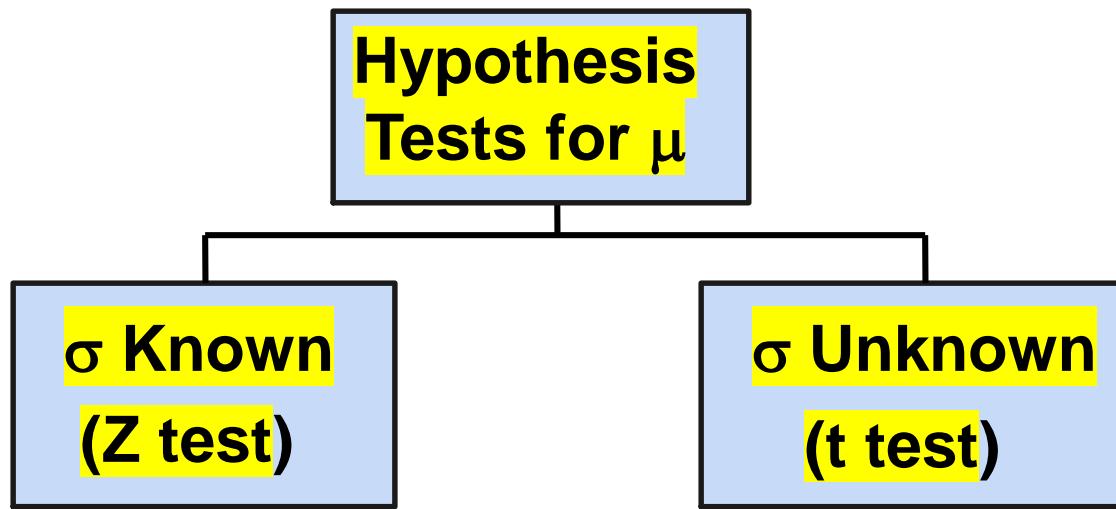
Level of significance = α



This is a **two-tail test** because there is a rejection region in both tails

Hypothesis Tests for the Mean

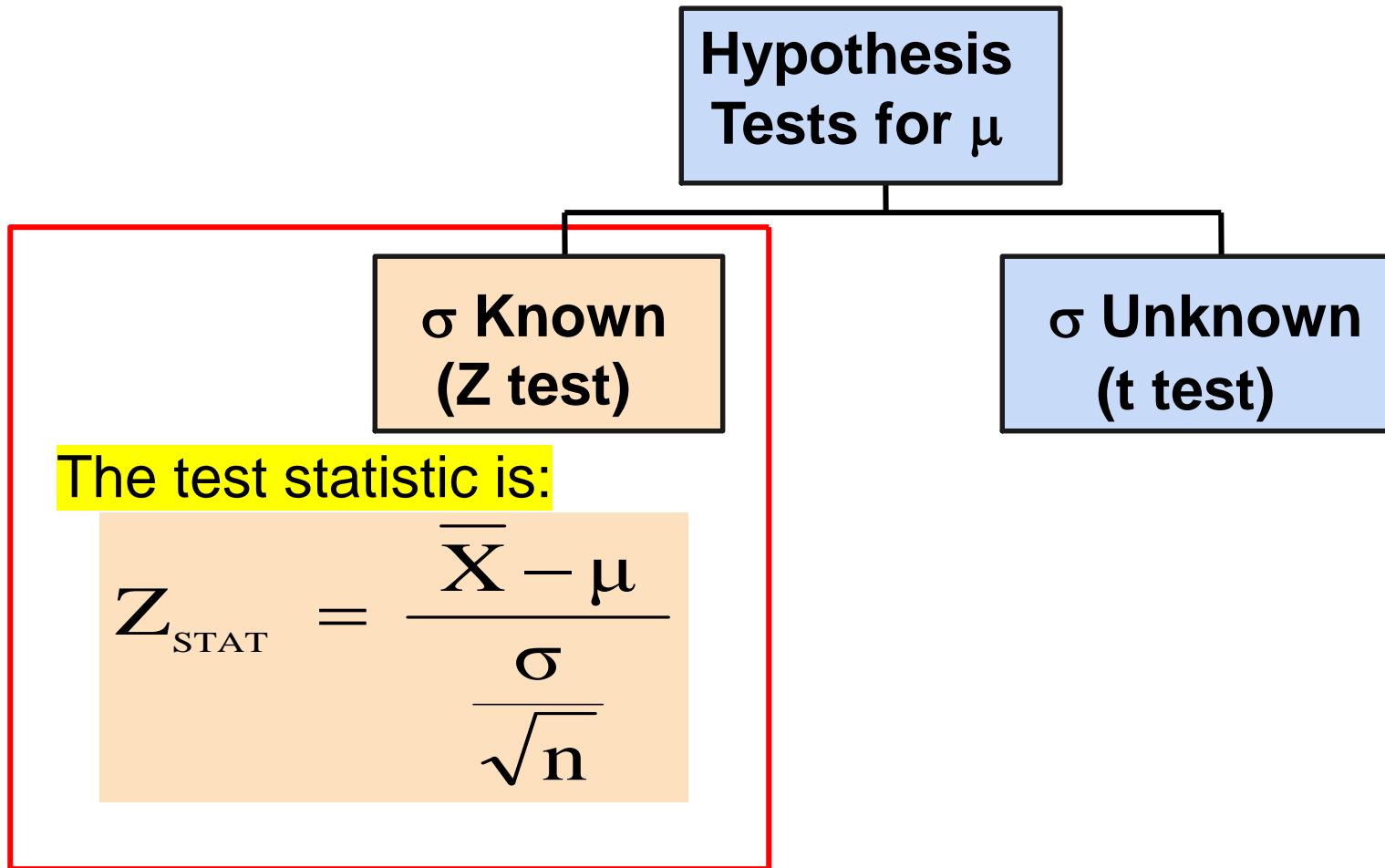
DCOVA



Z Test of Hypothesis for the Mean (σ Known)

DCOV A

- Convert sample statistic (\bar{X}) to a Z_{STAT} test statistic



Critical Value Approach to Testing

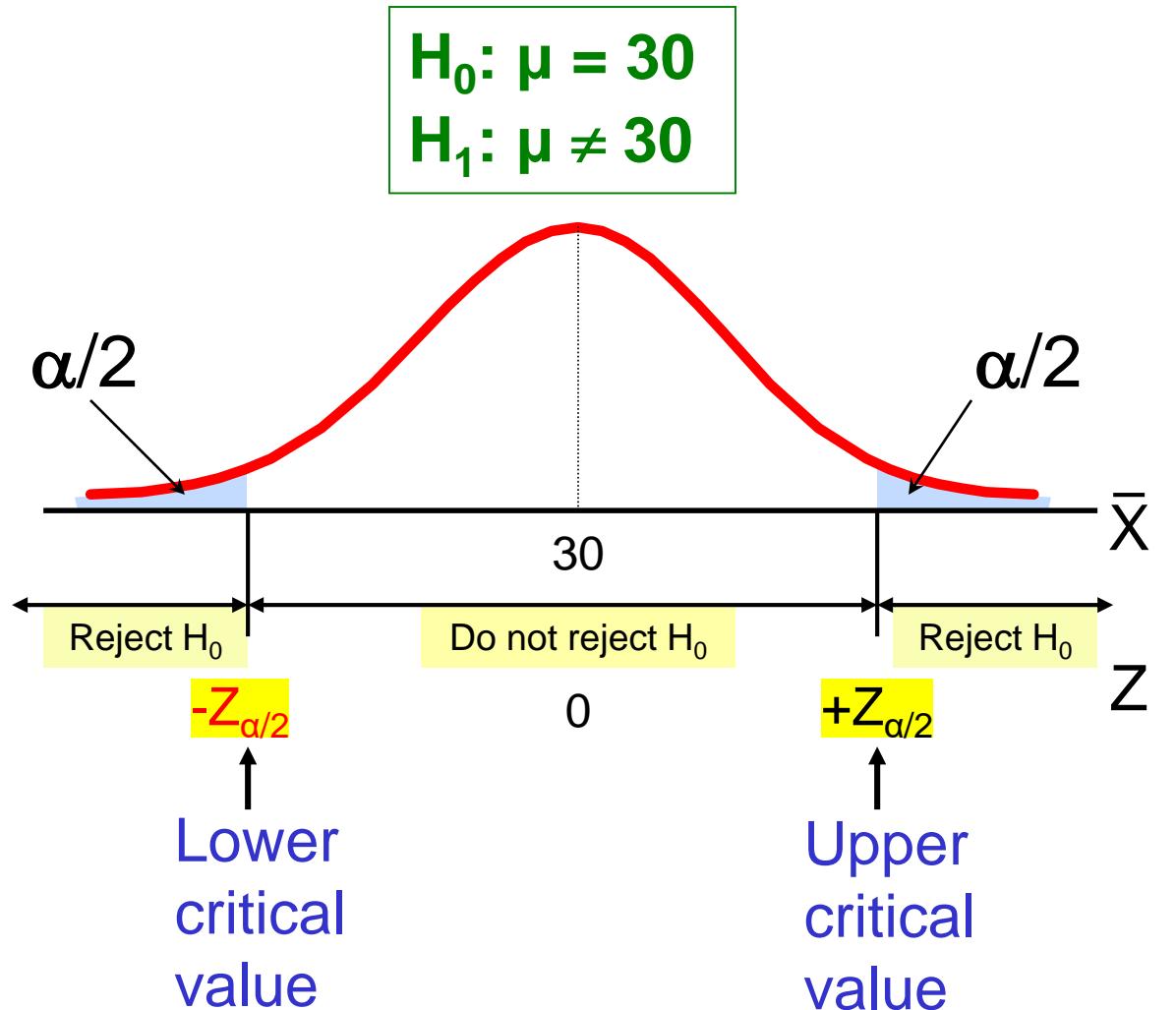
DCOV A

- For a two-tail test for the mean, σ known:
- Convert sample statistic (\bar{X}) to test statistic (Z_{STAT})
- Determine the critical Z values for a specified level of significance α from a table or computer
- **Decision Rule:** If the test statistic falls in the rejection region, reject H_0 ; otherwise do not reject H_0

Two-Tail Tests

DCOV A

- There are two cutoff values (critical values), defining the regions of rejection



6 Steps in Hypothesis Testing

DCOV A

1. State the null hypothesis, H_0 and the alternative hypothesis, H_1
2. Choose the level of significance, α , and the sample size, n
3. Determine the appropriate test statistic and sampling distribution
4. Determine the critical values that divide the rejection and nonrejection regions

6 Steps in Hypothesis Testing

DCOV**A**

(continued)

5. Collect data and compute the value of the test statistic
6. Make the statistical decision and state the managerial conclusion. If the test statistic falls into the nonrejection region, do not reject the null hypothesis H_0 . If the test statistic falls into the rejection region, reject the null hypothesis. Express the managerial conclusion in the context of the problem

Hypothesis Testing Example

DCOVA

**Test the claim that the true mean diameter
of a manufactured bolt is 30mm.
(Assume $\sigma = 0.8$)**

1. State the appropriate null and alternative hypotheses
 - $H_0: \mu = 30$ $H_1: \mu \neq 30$ (This is a two-tail test)
2. Specify the desired level of significance and the sample size
 - Suppose that $\alpha = 0.05$ and $n = 100$ are chosen for this test



Hypothesis Testing Example

(continued)

DCOVA

3. Determine the appropriate technique
 - σ is assumed known so this is a Z test.
4. Determine the critical values
 - For $\alpha = 0.05$ the critical Z values are ± 1.96
5. Collect the data and compute the test statistic
 - Suppose the sample results are
 $n = 100, \bar{X} = 29.84$ ($\sigma = 0.8$ is assumed known)

So the test statistic is:

$$Z_{\text{STAT}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{29.84 - 30}{\frac{0.8}{\sqrt{100}}} = \frac{-0.16}{0.08} = -2.0$$

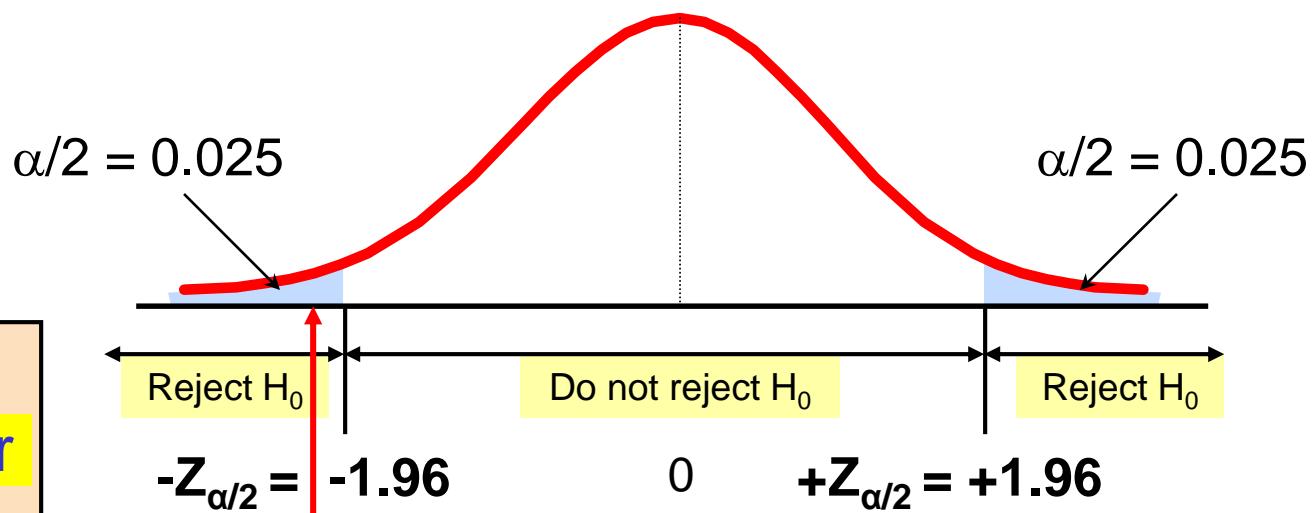


Hypothesis Testing Example

(continued)

DCOV A

- 6. Is the test statistic in the rejection region?

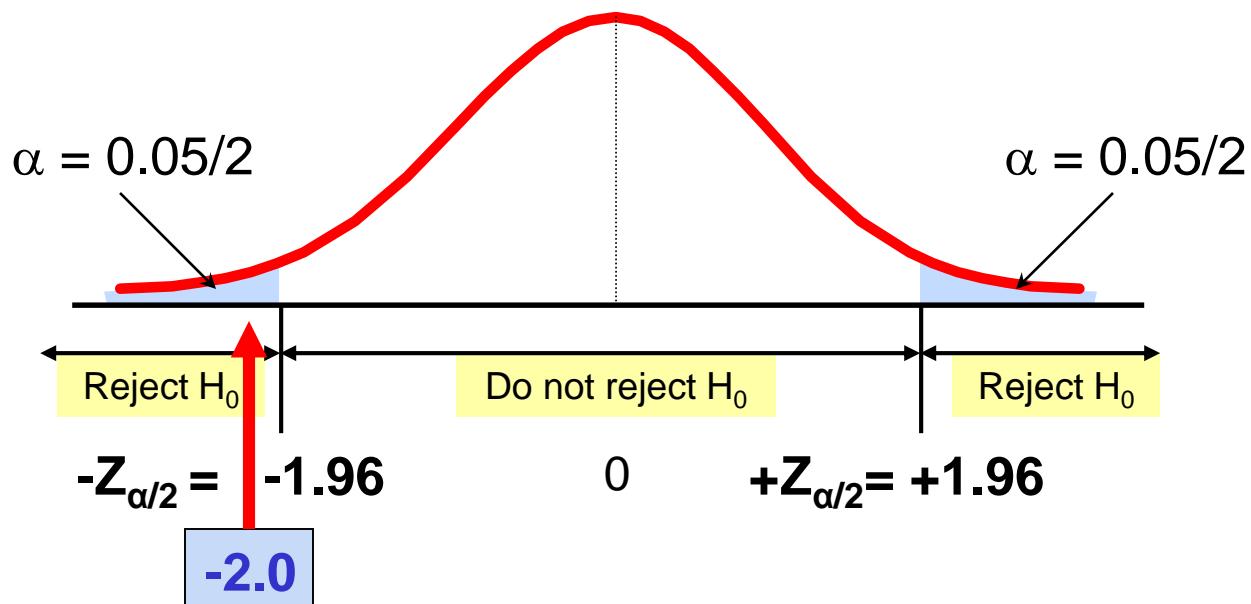


Hypothesis Testing Example

(continued)

DCOV A

6 (continued). Reach a decision and interpret the result



Since $Z_{STAT} = -2.0 < -1.96$, reject the null hypothesis and conclude there is sufficient evidence that the mean diameter of a manufactured bolt is not equal to 30



p-Value Approach to Testing

DCOV A

- p-value: Probability of obtaining a test statistic equal to or more extreme than the observed sample value given H_0 is true
 - The p-value is also called the observed level of significance
 - It is the smallest value of α for which H_0 can be rejected

p-Value Approach to Testing: Interpreting the p-value

DCOV A

- Compare the p-value with α

- If p-value $< \alpha$, reject H_0
- If p-value $\geq \alpha$, do not reject H_0

- Remember

- If the p-value is low then H_0 must go

The 5 Step p-value approach to Hypothesis Testing

DCOVA

1. State the null hypothesis, H_0 and the alternative hypothesis, H_1
2. Choose the level of significance, α , and the sample size, n
3. Determine the appropriate test statistic and sampling distribution
4. Collect data and compute the value of the test statistic and the p-value
5. Make the statistical decision and state the managerial conclusion. If the p-value is $< \alpha$ then reject H_0 , otherwise do not reject H_0 . State the managerial conclusion in the context of the problem

p-value Hypothesis Testing

Example

DCOVA

**Test the claim that the true mean diameter of a manufactured bolt is 30mm.
(Assume $\sigma = 0.8$)**

1. State the appropriate null and alternative hypotheses
 - $H_0: \mu = 30$ $H_1: \mu \neq 30$ (This is a two-tail test)
2. Specify the desired level of significance and the sample size
 - Suppose that $\alpha = 0.05$ and $n = 100$ are chosen for this test



p-value Hypothesis Testing Example

(continued)

DCOVA

3. Determine the appropriate technique
 - σ is assumed known so this is a Z test.
4. Collect the data, compute the test statistic and the p-value
 - Suppose the sample results are
 $n = 100, \bar{X} = 29.84$ ($\sigma = 0.8$ is assumed known)

So the test statistic is:

$$Z_{\text{STAT}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{29.84 - 30}{\frac{0.8}{\sqrt{100}}} = \frac{-0.16}{0.08} = -2.0$$



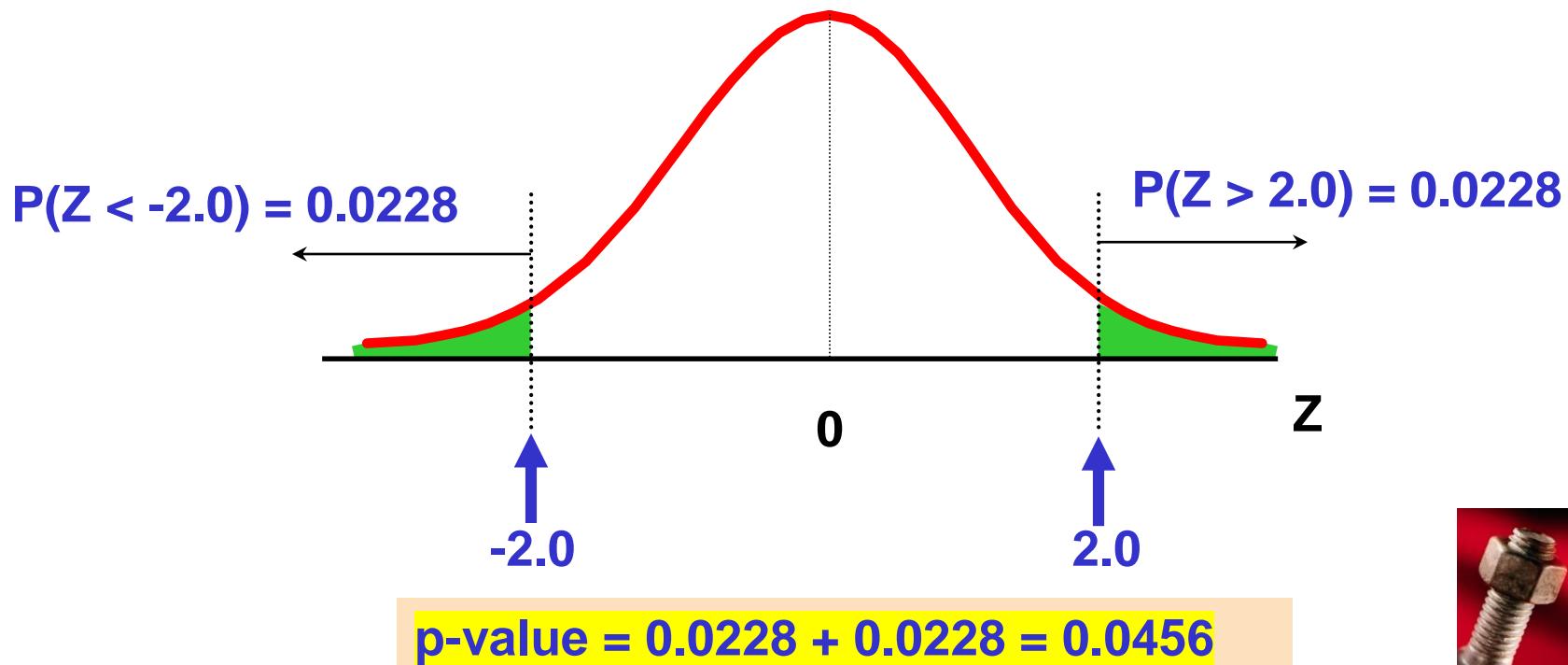
p-Value Hypothesis Testing Example: Calculating the p-value

(continued)

DCOVA

4. (continued) Calculate the p-value.

- How likely is it to get a Z_{STAT} of -2 (or something further from the mean (0), in either direction) if H_0 is true?

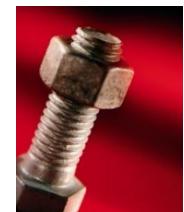


p-value Hypothesis Testing Example

(continued)

DCOVA

- 5. Is the p-value $< \alpha$?
 - Since $p\text{-value} = 0.0456 < \alpha = 0.05$, Reject H_0
- 5. (continued) State the managerial conclusion in the context of the situation.
 - There is sufficient evidence to conclude the average diameter of a manufactured bolt is not equal to 30mm.



Connection Between Two Tail Tests and Confidence Intervals

DCOV A

- For $\bar{X} = 29.84$, $\sigma = 0.8$ and $n = 100$, the 95% confidence interval is:

$$29.84 - (1.96) \frac{0.8}{\sqrt{100}} \text{ to } 29.84 + (1.96) \frac{0.8}{\sqrt{100}}$$

$$29.6832 \leq \mu \leq 29.9968$$

- Since this interval does not contain the hypothesized mean (30), we reject the null hypothesis at $\alpha = 0.05$



Do You Ever Truly Know σ ?

DCOVA

- Probably not!
- In virtually all real world business situations, σ is not known.
- If there is a situation where σ is known then μ is also known (since to calculate σ you need to know μ .)
- If you truly know μ there would be no need to gather a sample to estimate it.

Hypothesis Testing: σ Unknown

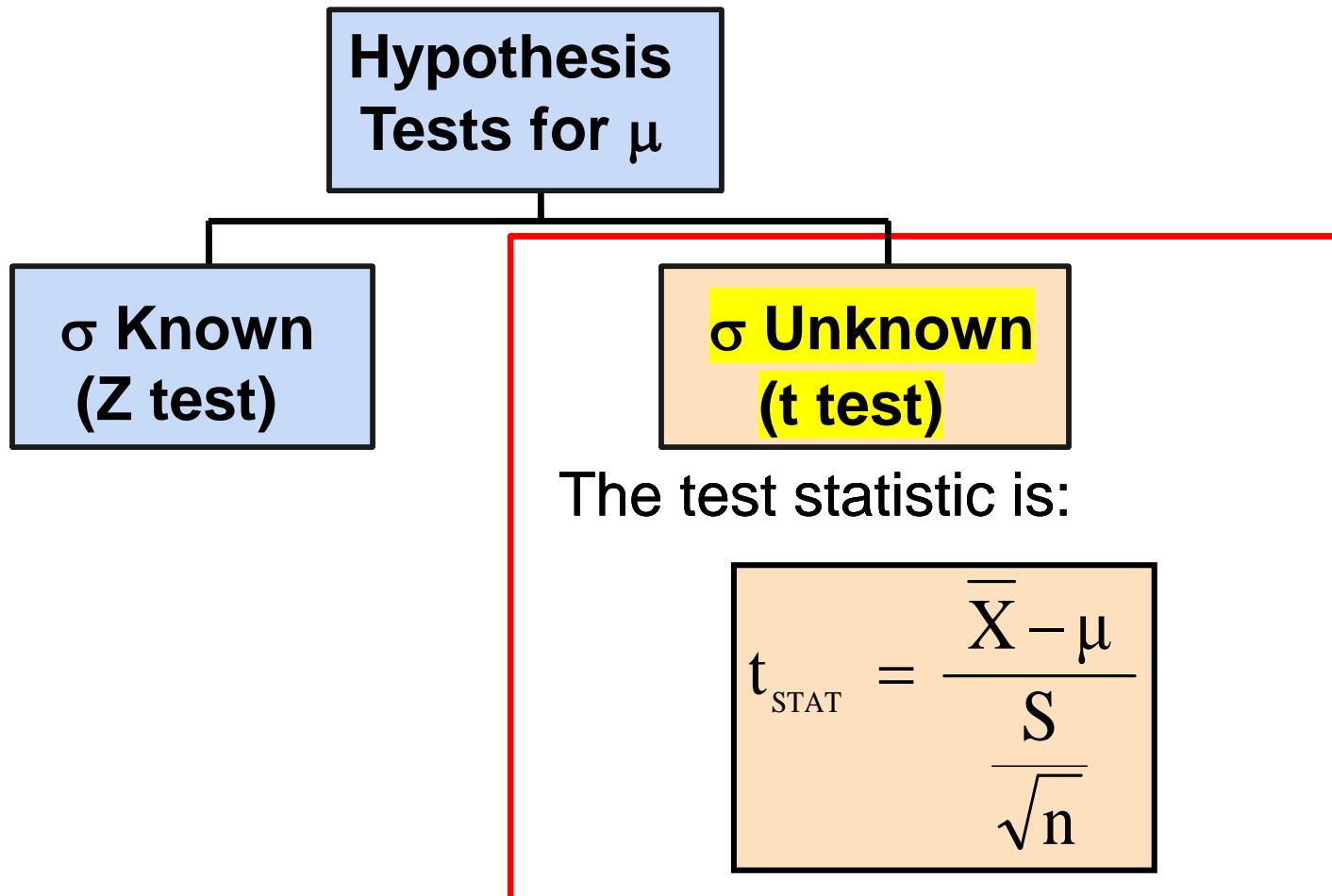
DCOVA

- If the population standard deviation is unknown, you instead use the sample standard deviation S.
- Because of this change, you use the t distribution instead of the Z distribution to test the null hypothesis about the mean.
- When using the t distribution you must assume the population you are sampling from follows a normal distribution.
- All other steps, concepts, and conclusions are the same.

t Test of Hypothesis for the Mean (σ Unknown)

DCOVA

- Convert sample statistic (\bar{X}) to a t_{STAT} test statistic

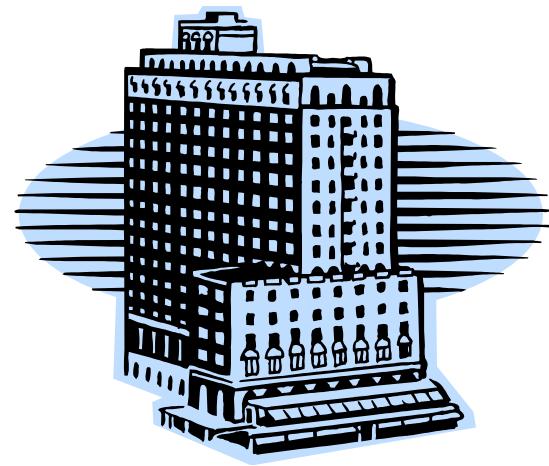


Example: Two-Tail Test (σ Unknown)

DCOVA

The average cost of a hotel room in New York is said to be \$168 per night. To determine if this is true, a random sample of 25 hotels is taken and resulted in an \bar{X} of \$172.50 and an S of \$15.40. Test the appropriate hypotheses at $\alpha = 0.05$.

(Assume the population distribution is normal)



$$\begin{aligned}H_0: \mu &= 168 \\H_1: \mu &\neq 168\end{aligned}$$

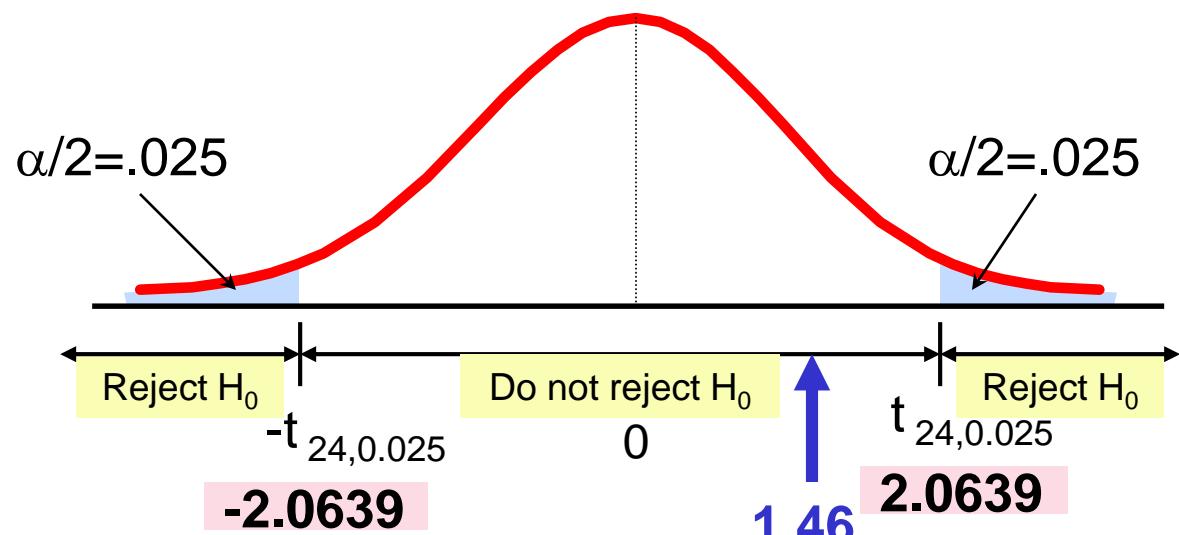
Example Solution: Two-Tail t Test

DCOV A

$$\begin{aligned} H_0: \mu &= 168 \\ H_1: \mu &\neq 168 \end{aligned}$$

- $\alpha = 0.05$
- $n = 25, df = 25-1=24$
- σ is unknown, so use a **t statistic**
- Critical Value:

$$\pm t_{24,0.025} = \pm 2.0639$$



$$t_{\text{STAT}} = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{172.50 - 168}{\frac{15.40}{\sqrt{25}}} = 1.46$$

Do not reject H_0 : insufficient evidence that true mean cost is different from \$168

To Use the t-test Must Assume

the Population Is Normal

DCOV A

- As long as the sample size is not very small and the population is not very skewed, the t-test can be used.
- To evaluate the normality assumption you can use:
 - How closely sample statistics match the normal distribution's theoretical properties.
 - A histogram or stem-and-leaf display or boxplot or a normal probability plot.
 - Section 6.3 has more details on evaluating normality.

Example Two-Tail t Test Using A p-value from Excel

DCOV A

- Since this is a t-test we cannot calculate the p-value without some calculation aid.
- The Excel output below does this:

t Test for the Hypothesis of the Mean

| Data | | |
|---------------------------|---------|-----------|
| Null Hypothesis | $\mu =$ | \$ 168.00 |
| Level of Significance | | 0.05 |
| Sample Size | | 25 |
| Sample Mean | | \$ 172.50 |
| Sample Standard Deviation | | \$ 15.40 |

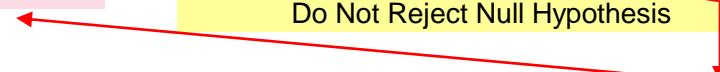
Intermediate Calculations

| | | | |
|----------------------------|----|------|--------------|
| Standard Error of the Mean | \$ | 3.08 | =B8/SQRT(B6) |
| Degrees of Freedom | | 24 | =B6-1 |
| t test statistic | | 1.46 | =(B7-B4)/B11 |

Two-Tail Test

| | | |
|-------------------------------|---------|---|
| Lower Critical Value | -2.0639 | =-TINV(B5,B12) |
| Upper Critical Value | 2.0639 | =TINV(B5,B12) |
| p-value | 0.157 | =TDIST(ABS(B13),B12,2) =IF(B18<B5, "Reject null hypothesis", "Do not reject null hypothesis") |
| Do Not Reject Null Hypothesis | | |

p-value > α
So do not reject H_0



Connection of Two Tail Tests to Confidence Intervals

DCOV A

- For $\bar{X} = 172.5$, $S = 15.40$ and $n = 25$, the 95% confidence interval for μ is:

$$172.5 - (2.0639) \frac{15.4}{\sqrt{25}}$$

to

$$172.5 + (2.0639) \frac{15.4}{\sqrt{25}}$$

$$166.14 \leq \mu \leq 178.86$$

- Since this interval contains the Hypothesized mean (168), we do not reject the null hypothesis at $\alpha = 0.05$

One-Tail Tests

DCOVA

- In many cases, the alternative hypothesis focuses on a particular direction

$$H_0: \mu \geq 3$$

$$H_1: \mu < 3$$



This is a **lower**-tail test since the alternative hypothesis is focused on the lower tail below the mean of 3

$$H_0: \mu \leq 3$$

$$H_1: \mu > 3$$

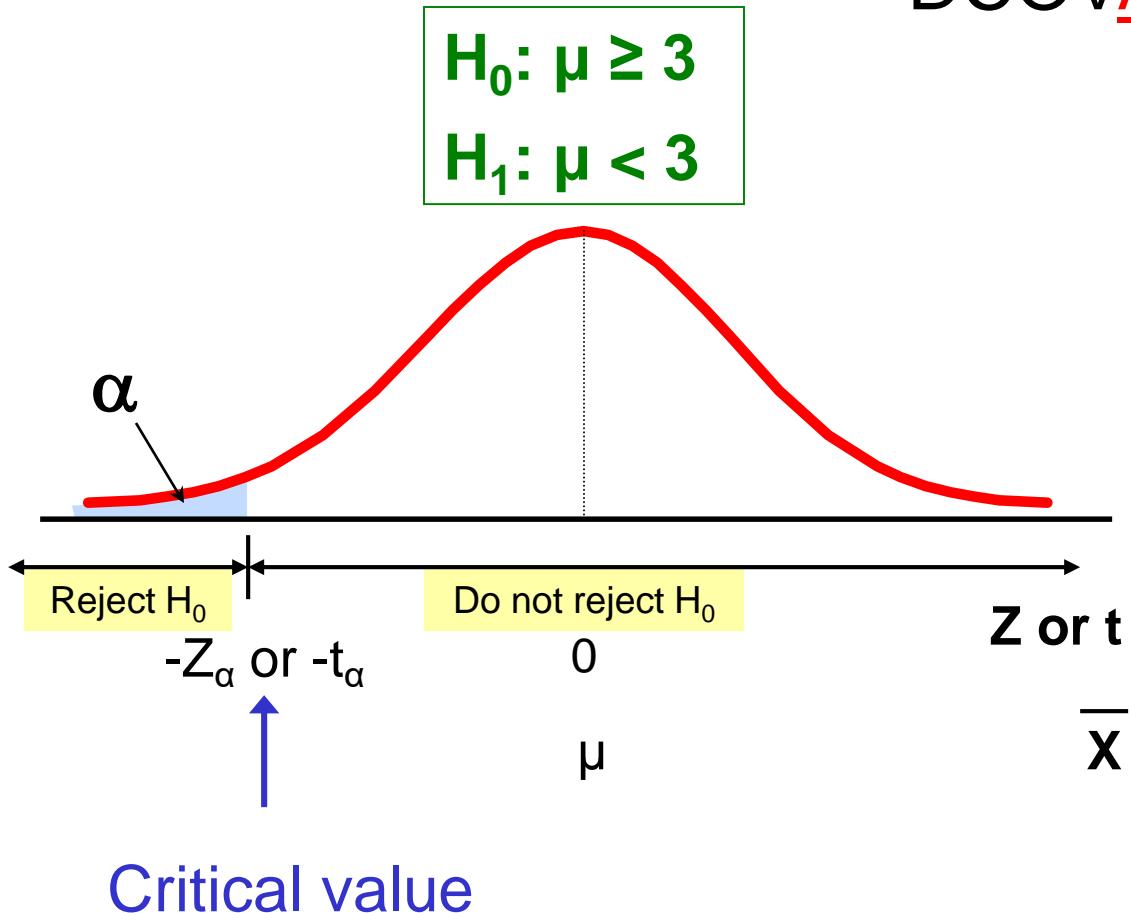


This is an **upper**-tail test since the alternative hypothesis is focused on the upper tail above the mean of 3

Lower-Tail Tests

DCOV A

- There is only one critical value, since the rejection area is in only one tail

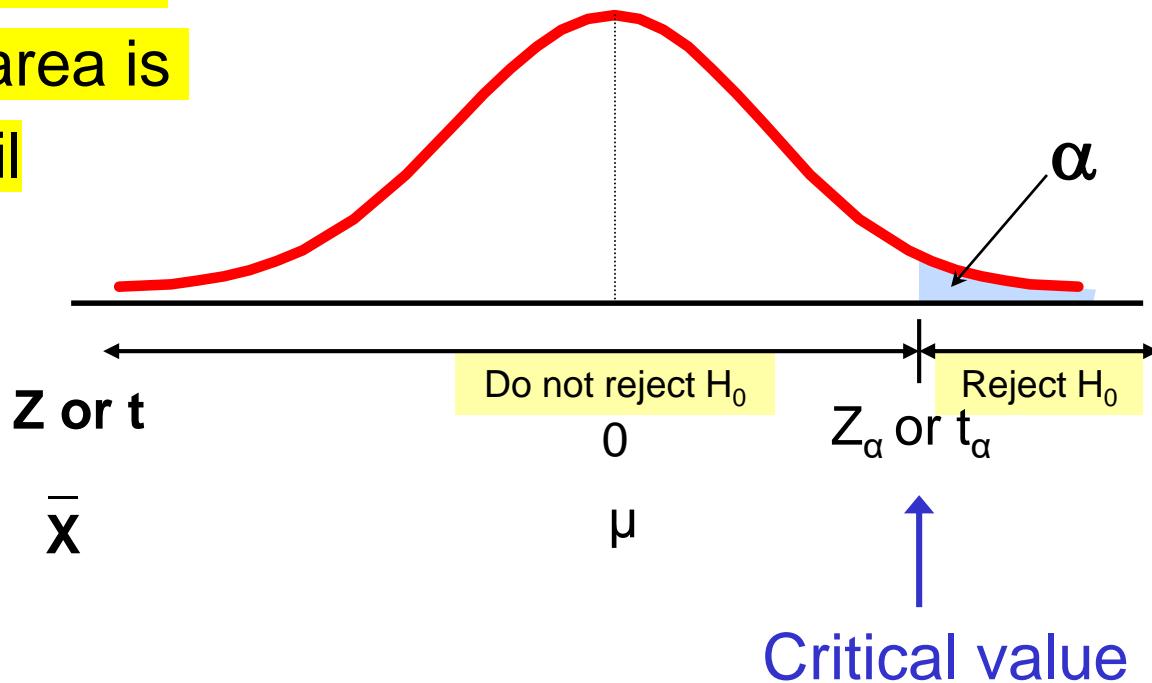


Upper-Tail Tests

DCOV A

- There is only one critical value, since the rejection area is in only one tail

$$\begin{aligned}H_0: \mu &\leq 3 \\H_1: \mu &> 3\end{aligned}$$



Example: Upper-Tail t Test for Mean (σ unknown)

DCOVA

A phone industry manager thinks that customer monthly cell phone bills have increased, and now average over \$52 per month. The company wishes to test this claim. (Assume a normal population)

Form hypothesis test:

$H_0: \mu \leq 52$ the average is not over \$52 per month

$H_1: \mu > 52$ the average **is** greater than \$52 per month
(i.e., sufficient evidence exists to support the manager's claim)



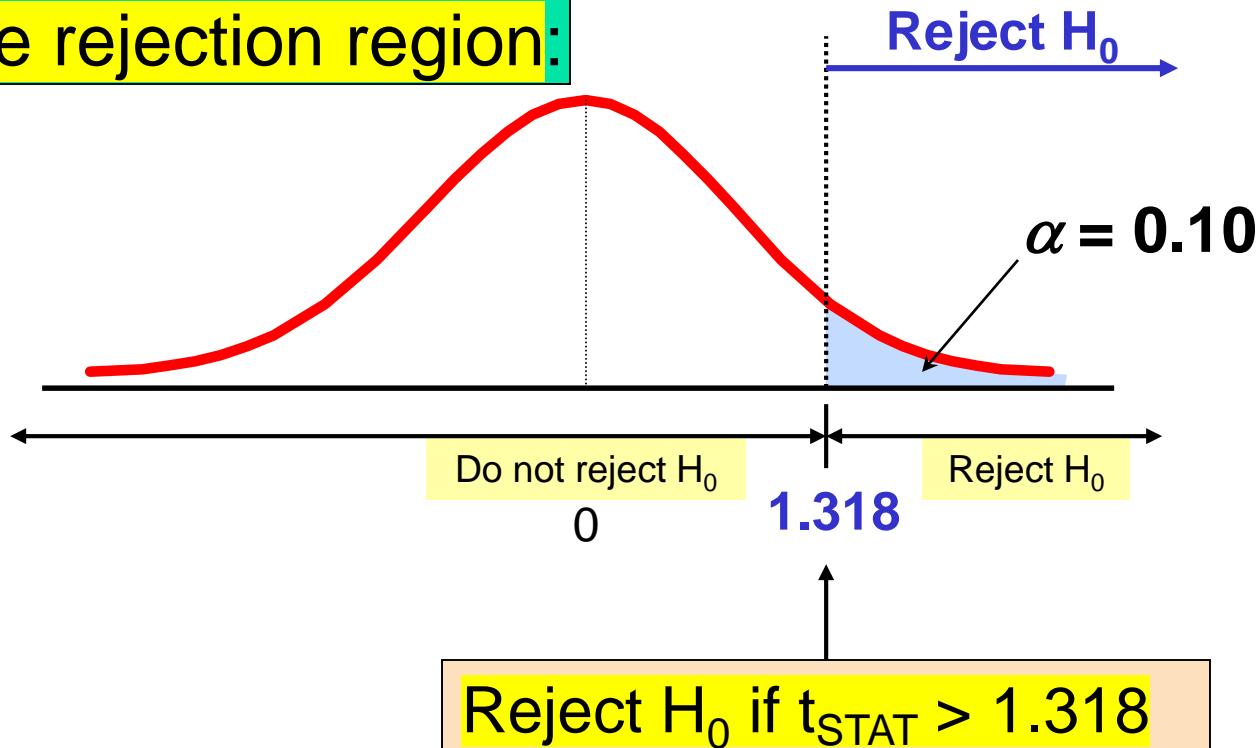
Example: Find Rejection Region

(continued)

DCOVA

- Suppose that $\alpha = 0.10$ is chosen for this test and $n = 25$.

Find the rejection region:



Example: Test Statistic

(continued)

DCOVA

Obtain sample and compute the test statistic

Suppose a sample n is taken with the following results: $n = 25$, $\bar{X} = 53.1$, and $S = 10$

- Then the test statistic is:

$$t_{\text{STAT}} = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{53.1 - 52}{\frac{10}{\sqrt{25}}} = 0.55$$

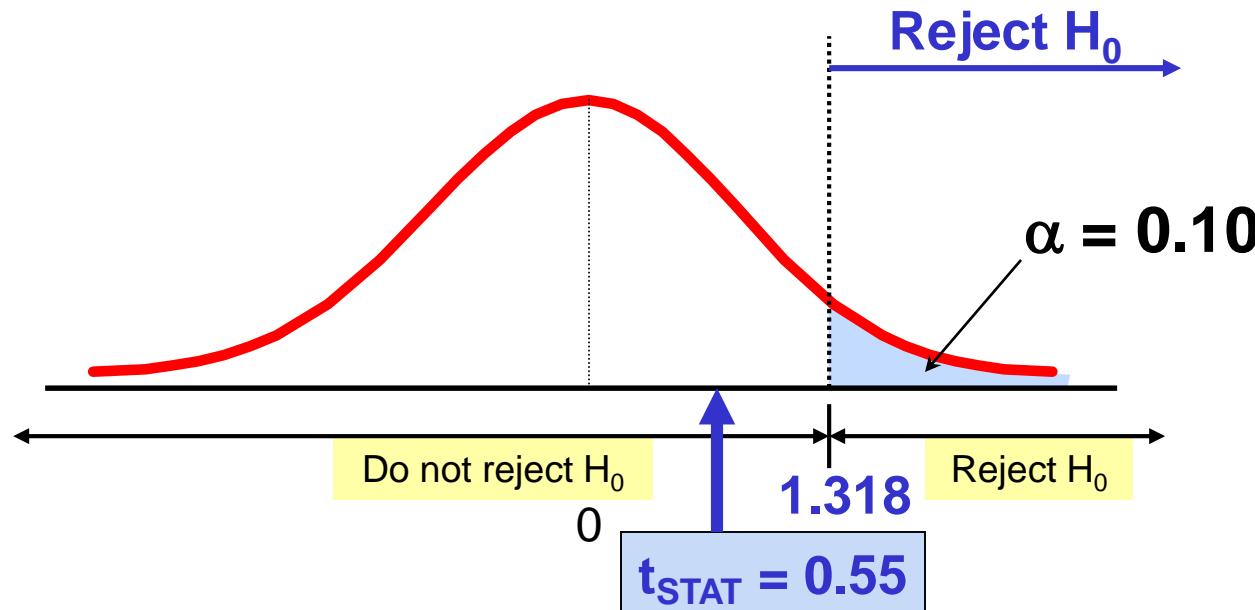


Example: Decision

(continued)

DCOVA

Reach a decision and interpret the result:



Do not reject H_0 since $t_{STAT} = 0.55 \leq 1.318$

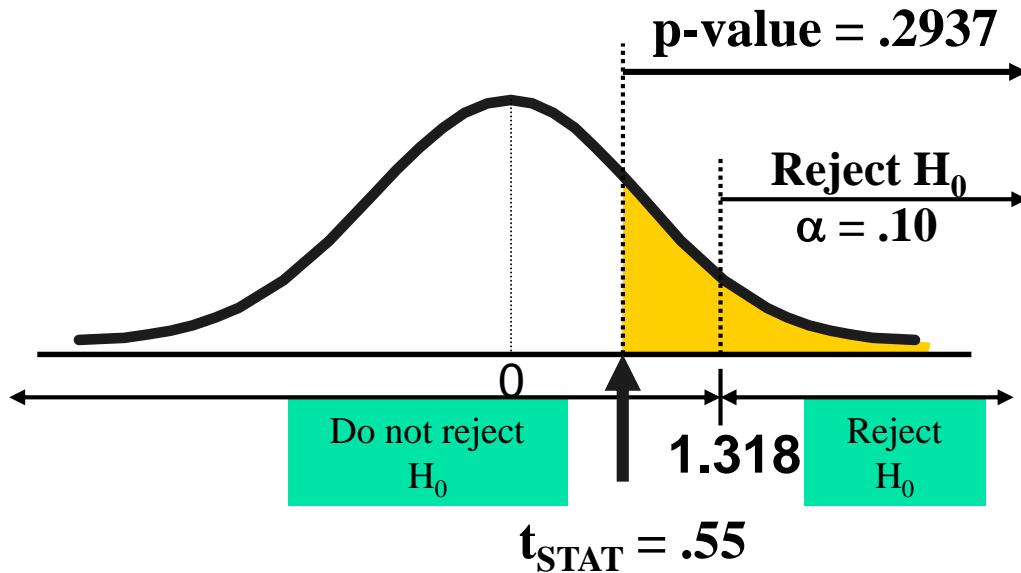
there is not sufficient evidence that the mean bill is over \$52



Example: Utilizing The p-value for The Test

DCOV A

- Calculate the p-value and compare to α (p-value below calculated using excel spreadsheet on next page)



Do not reject H_0 since $p\text{-value} = .2937 > \alpha = .10$

Excel Spreadsheet Calculating The p-value for The Upper Tail t Test

End of Chapter

DCOVA

t Test for the Hypothesis of the Mean

| Data | | |
|---------------------------|---------|-------|
| Null Hypothesis | $\mu =$ | 52.00 |
| Level of Significance | | 0.1 |
| Sample Size | | 25 |
| Sample Mean | | 53.10 |
| Sample Standard Deviation | | 10.00 |

| Intermediate Calculations | |
|-------------------------------|--------|
| Standard Error of the Mean | 2.00 |
| Degrees of Freedom | 24 |
| t test statistic | 0.55 |
| Upper Tail Test | |
| Upper Critical Value | 1.318 |
| p-value | 0.2937 |
| Do Not Reject Null Hypothesis | |

=B8/SQRT(B6)

=B6-1

=(B7-B4)/B11

=TINV(2*B5,B12)

=TDIST(ABS(B13),B12,1)

=IF(B18<B5, "Reject null hypothesis",
"Do not reject null hypothesis")

Hypothesis Tests for Proportions

DCOVA

- Involves categorical variables
- Two possible outcomes
 - Possesses characteristic of interest
 - Does not possess characteristic of interest
- Fraction or proportion of the population in the category of interest is denoted by π

Proportions

(continued)

DCOVA

- Sample proportion in the category of interest is denoted by p

- $$p = \frac{X}{n} = \frac{\text{number in category of interest in sample}}{\text{sample size}}$$

- When both $n\pi$ and $n(1-\pi)$ are at least 5, p can be approximated by a normal distribution with mean and standard deviation

-

$$\mu_p = \pi$$

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

Hypothesis Tests for Proportions

DCOVA

- The sampling distribution of p is approximately normal, so the test statistic is a Z_{STAT} value:

$$Z_{\text{STAT}} = \frac{p - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}}$$

Hypothesis Tests for p

$n\pi \geq 5$
and
 $n(1-\pi) \geq 5$

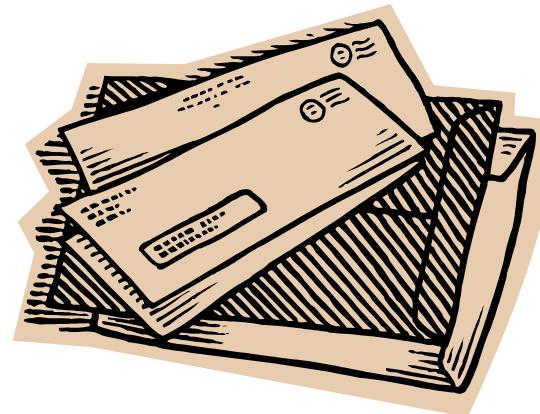
$n\pi < 5$
or
 $n(1-\pi) < 5$

Not discussed
in this chapter

Example: Z Test for Proportion

DCOVA

A marketing company claims that it receives 8% responses from its mailing. To test this claim, a random sample of 500 were surveyed with 25 responses. Test at the $\alpha = 0.05$ significance level.



Check:

$$n\pi = (500)(.08) = 40$$

$$n(1-\pi) = (500)(.92) = 460$$



Z Test for Proportion: Solution

DCOVA

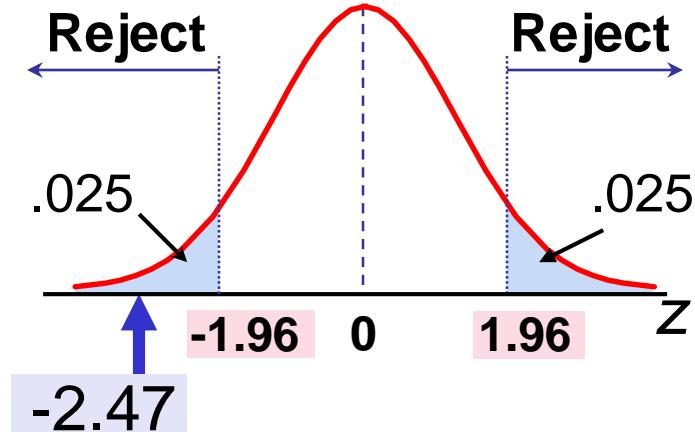
$$H_0: \pi = 0.08$$

$$H_1: \pi \neq 0.08$$

$$\alpha = 0.05$$

$$n = 500, p = 0.05$$

Critical Values: ± 1.96



Test Statistic:

$$Z_{\text{STAT}} = \frac{p - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}} = \frac{.05 - .08}{\sqrt{\frac{.08(1 - .08)}{500}}} = -2.47$$

Decision:

Reject H_0 at $\alpha = 0.05$

Conclusion:

There is sufficient evidence to reject the company's claim of 8% response rate.

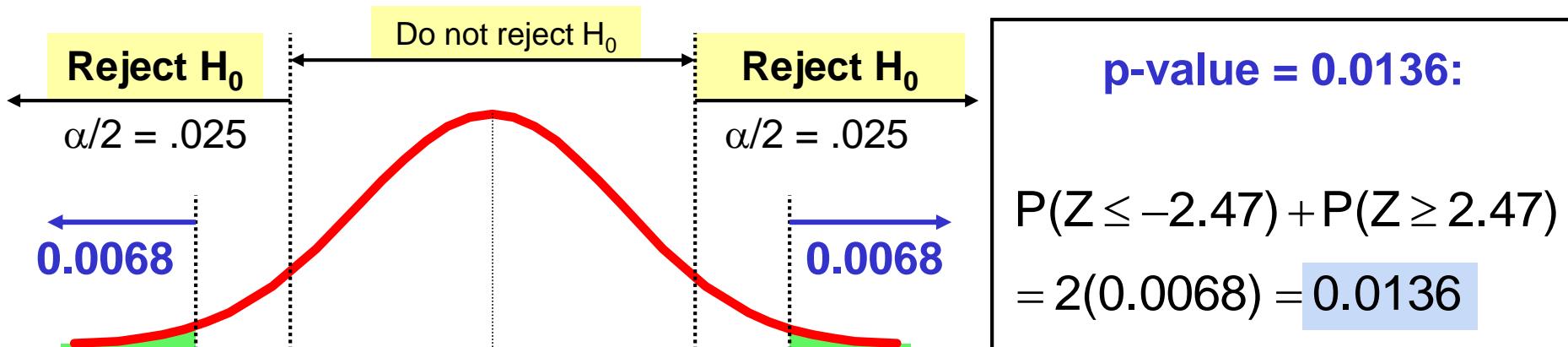
p-Value Solution

(continued)

DCOV A

Calculate the p-value and compare to α

(For a two-tail test the p-value is always two-tail)



$Z = -2.47$

$Z = 2.47$

Reject H_0 since p-value = 0.0136 < $\alpha = 0.05$

Questions To Address In The Planning Stage

- What is the goal of the survey, study, or experiment?
- How can you translate this goal into a null and an alternative hypothesis?
- Is the hypothesis test one or two tailed?
- Can a random sample be selected?
- What types of data will be collected? Numerical? Categorical?
- What level of significance should be used?
- Is the intended sample size large enough to achieve the desired power?
- What statistical test procedure should be used?
- What conclusions & interpretations can you reach from the results of the planned hypothesis test?

Failing to consider these questions can lead to bias or incomplete results

Statistical Significance vs Practical Significance

- Statistically significant results (rejecting the null hypothesis) are not always of practical significance
 - This is more likely to happen when the sample size gets very large
- Practically significant results might be found to be statistically insignificant (failing to reject the null hypothesis)
 - This is more likely to happen when the sample size is relatively small

Chapter Summary

In this chapter we discussed

- Hypothesis testing methodology
- Performing a Z Test for the mean (σ known)
- Critical value and p-value approaches to hypothesis testing
- Performing one-tail and two-tail tests

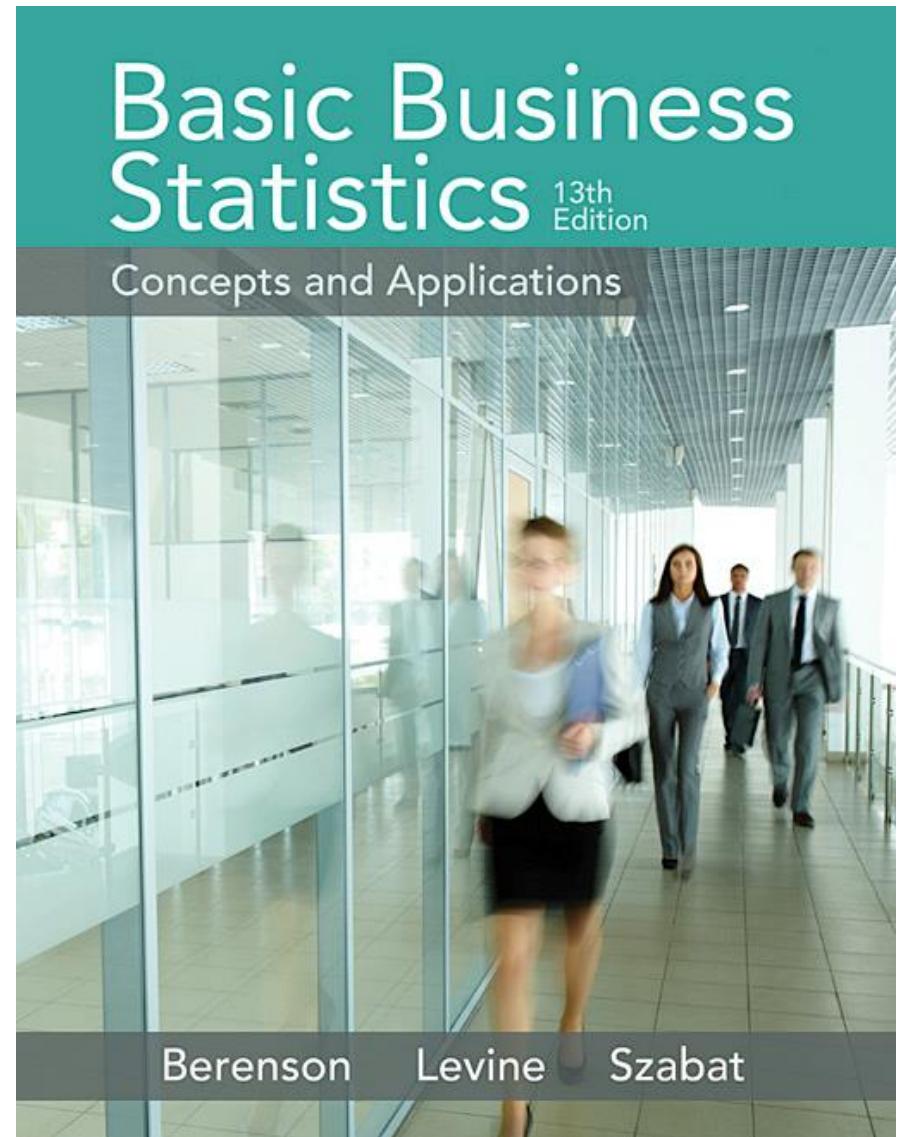
Chapter Summary

(continued)

- Performing a t test for the mean (σ unknown)
- Performing a Z test for the proportion
- Statistical and practical significance
- Pitfalls and ethical issues

Chapter 13

Simple Linear Regression



Learning Objectives

In this chapter, you learn:

- How to use regression analysis to predict the value of a dependent variable based on a value of an independent variable
- The meaning of the regression coefficients b_0 and b_1
- How to evaluate the assumptions of regression analysis and know what to do if the assumptions are violated
- To make inferences about the slope and correlation coefficient
- To estimate mean values and predict individual values

Correlation vs. Regression

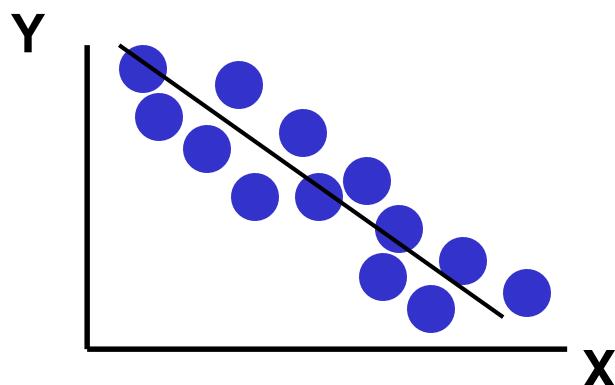
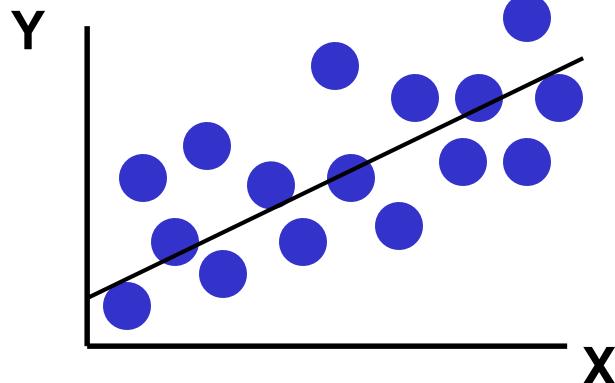
DCOV A

- A scatter plot can be used to show the relationship between two variables
- Correlation analysis is used to measure the strength of the association (linear relationship) between two variables
 - Correlation is only concerned with strength of the relationship
 - No causal effect is implied with correlation
 - Scatter plots were first presented in Ch. 2
 - Correlation was first presented in Ch. 3

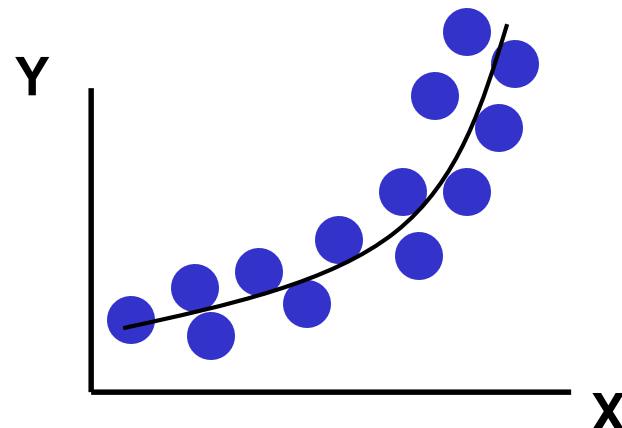
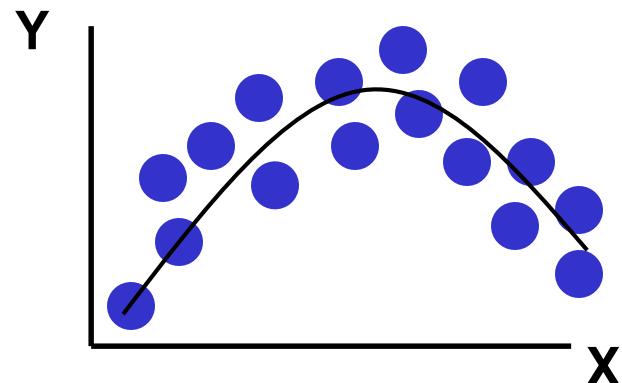
Types of Relationships

DCOVA

Linear relationships



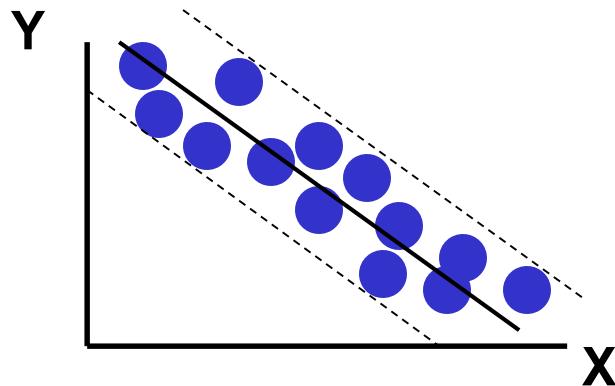
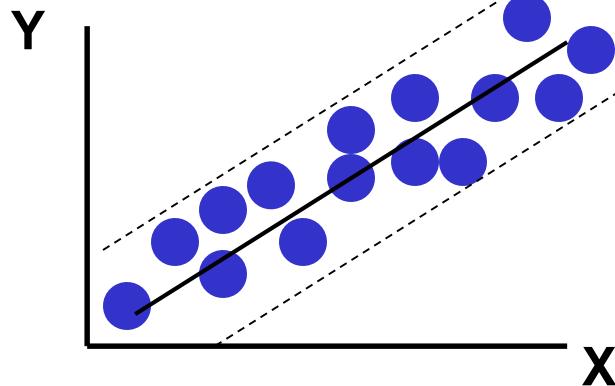
Curvilinear relationships



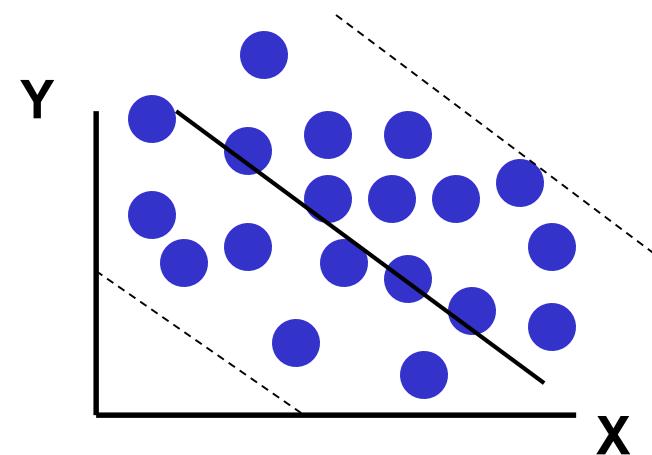
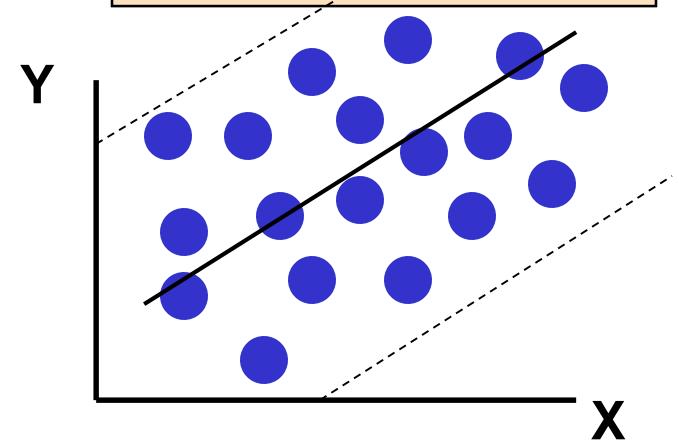
Types of Relationships

DCOVA
(continued)

Strong relationships



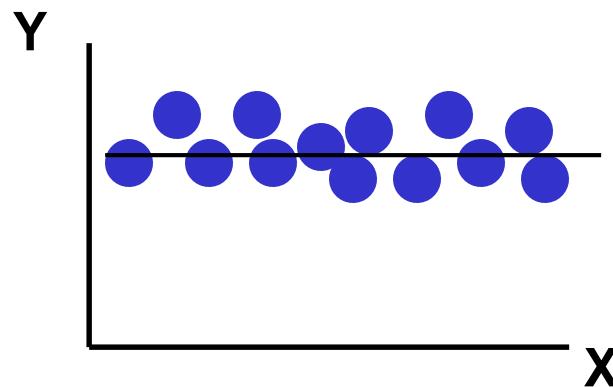
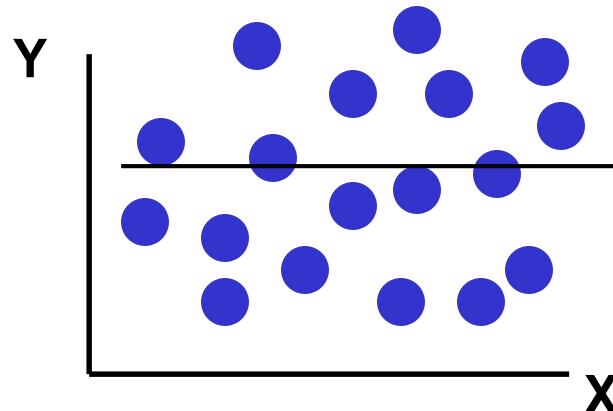
Weak relationships



Types of Relationships

DCOVA
(continued)

No relationship



Introduction to Regression Analysis

DCOVA

- **Regression analysis is used to:**

- Predict the value of a dependent variable based on the value of at least one independent variable
- Explain the impact of changes in an independent variable on the dependent variable

Dependent variable: the variable we wish to predict or explain

Independent variable: the variable used to predict or explain the dependent variable

Simple Linear Regression Model

DCOVA

- Only **one** independent variable, X
- Relationship between X and Y is described by a linear function
- Changes in Y are assumed to be related to changes in X

Simple Linear Regression Model

DCOV_A

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Dependent Variable

Population Y intercept

Population Slope Coefficient

Independent Variable

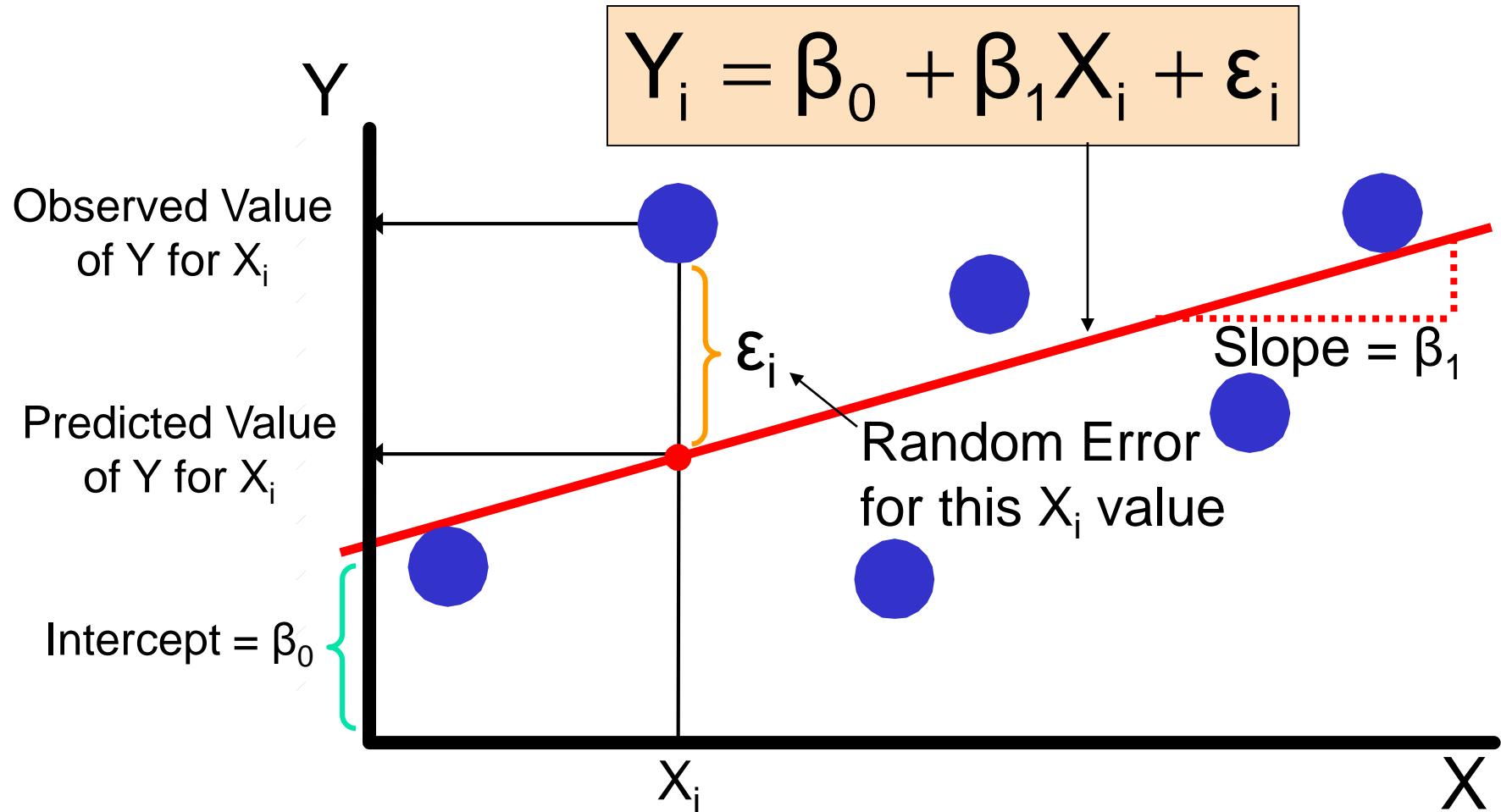
Random Error term

Linear component

Random Error component

Simple Linear Regression Model

DCOV A
(continued)



Simple Linear Regression Equation (Prediction Line)

DCOVA

The simple linear regression equation provides an estimate of the population regression line

Estimated
(or predicted)
Y value for
observation i

Estimate of
the regression
intercept

Estimate of the
regression slope

Value of X for
observation i

$$\hat{Y}_i = b_0 + b_1 X_i$$

The Least Squares Method

DCOVA

b_0 and b_1 are obtained by finding the values of
that minimize the sum of the squared
differences between Y (actual value) and \hat{Y}
(predicted value):

$$\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - (b_0 + b_1 X_i))^2$$

Finding the Least Squares Equation

DCOVA

- The coefficients b_0 and b_1 , and other regression results in this chapter, will be found using Excel or Minitab

Formulas are shown in the text for those
who are interested

Interpretation of the Slope and the Intercept

DCOVA

- b_0 (intercept) is the estimated average value of Y (dependent variable) when the value of X (independent variable) is zero
- b_1 (slope) is the estimated change in the average value of Y as a result of a one-unit increase in X

Simple Linear Regression Example

DCOVA

- A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)
- A random sample of 10 houses is selected
 - Dependent variable (Y) = house price in \$1000s
 - Independent variable (X) = square feet



Simple Linear Regression Example: Data

DCOVA

| House Price in \$1000s (Y) | Square Feet (X) |
|----------------------------|-----------------|
| 245 | 1400 |
| 312 | 1600 |
| 279 | 1700 |
| 308 | 1875 |
| 199 | 1100 |
| 219 | 1550 |
| 405 | 2350 |
| 324 | 2450 |
| 319 | 1425 |
| 255 | 1700 |

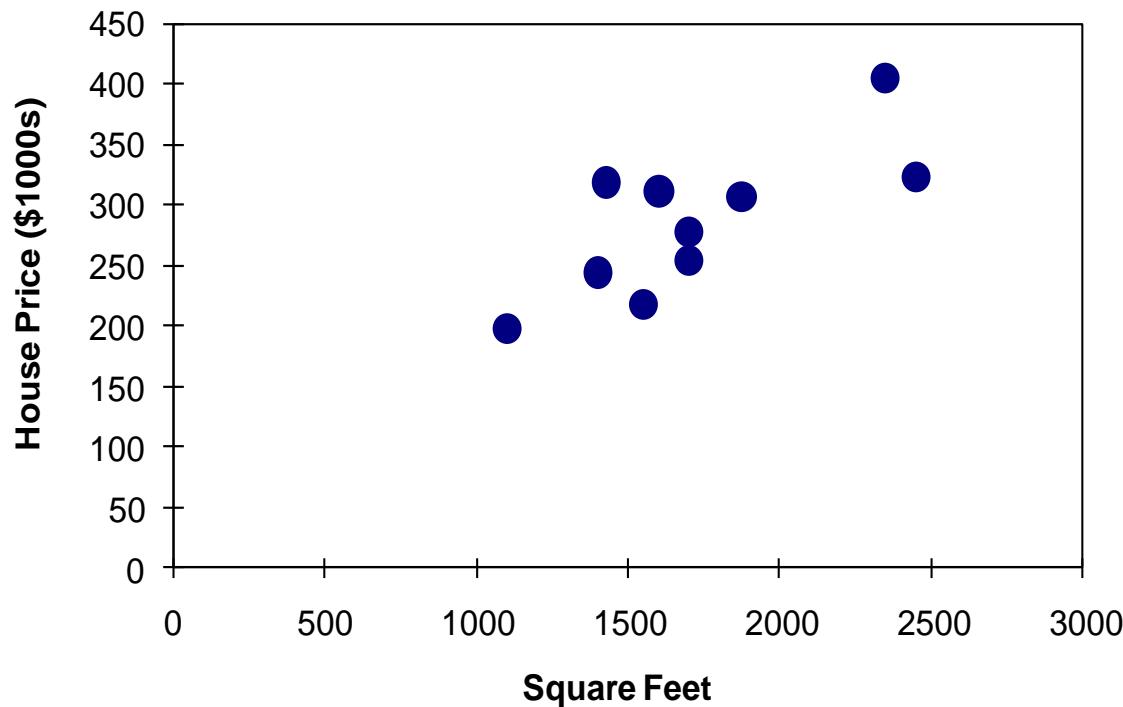


Simple Linear Regression

Example: Scatter Plot

DCOVA

House price model: Scatter Plot



Simple Linear Regression Example: Using Excel Data Analysis Function

DCOV A

1. Choose Data

2. Choose Data Analysis

3. Choose Regression

The screenshot shows a Microsoft Excel spreadsheet titled "Chapter 13 examples.xlsx". The data consists of two columns: "House Price" and "Square Feet", with 11 data points from row 1 to row 11. The "Data" tab is selected in the ribbon. A red arrow points from the "Data" tab to the "Data Analysis" button in the "Analysis" group. Another red arrow points from the "Data Analysis" button to the "Data Analysis" dialog box. Inside the dialog box, a third red arrow points to the "Regression" option under the "Analysis Tools" list.

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R |
|----|-------------|-------------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | House Price | Square Feet | | | | | | | | | | | | | | | | |
| 2 | 245 | 1400 | | | | | | | | | | | | | | | | |
| 3 | 312 | 1600 | | | | | | | | | | | | | | | | |
| 4 | 279 | 1700 | | | | | | | | | | | | | | | | |
| 5 | 308 | 1875 | | | | | | | | | | | | | | | | |
| 6 | 199 | 1100 | | | | | | | | | | | | | | | | |
| 7 | 219 | 1550 | | | | | | | | | | | | | | | | |
| 8 | 405 | 2350 | | | | | | | | | | | | | | | | |
| 9 | 324 | 2450 | | | | | | | | | | | | | | | | |
| 10 | 319 | 1425 | | | | | | | | | | | | | | | | |
| 11 | 255 | 1700 | | | | | | | | | | | | | | | | |
| 12 | | | | | | | | | | | | | | | | | | |

Data Analysis
Analysis Tools
Histogram
Moving Average
Random Number Generation
Rank and Percentile
Regression
Sampling
t-Test: Paired Two Sample for Means
t-Test: Two-Sample Assuming Equal Variances
t-Test: Two-Sample Assuming Unequal Variances
z-Test: Two Sample for Means



Simple Linear Regression Example: Using Excel Data Analysis Function

(continued)

Enter Y range and X range and desired options

DCOV A

| | A | B | C | D | E | F | G | H | I |
|----|-------------|-------------|---|---|---|---|---|---|---|
| 1 | House Price | Square Feet | | | | | | | |
| 2 | 245 | 1400 | | | | | | | |
| 3 | 312 | 1600 | | | | | | | |
| 4 | 279 | 1700 | | | | | | | |
| 5 | 308 | 1875 | | | | | | | |
| 6 | 199 | 1100 | | | | | | | |
| 7 | 219 | 1550 | | | | | | | |
| 8 | 405 | 2350 | | | | | | | |
| 9 | 324 | 2450 | | | | | | | |
| 10 | 319 | 1425 | | | | | | | |
| 11 | 255 | 1700 | | | | | | | |
| 12 | | | | | | | | | |
| 13 | | | | | | | | | |
| 14 | | | | | | | | | |
| 15 | | | | | | | | | |
| 16 | | | | | | | | | |
| 17 | | | | | | | | | |
| 18 | | | | | | | | | |
| 19 | | | | | | | | | |
| 20 | | | | | | | | | |

The screenshot shows an Excel spreadsheet with data in columns A and B. Column A is labeled "House Price" and column B is labeled "Square Feet". Row 13 is highlighted in orange, indicating the starting point for the regression analysis.

A "Regression" dialog box is open over the spreadsheet. The "Input" section has "Input Y Range" set to \$A\$2:\$A\$11 and "Input X Range" set to \$B\$2:\$B\$11. Under "Output options", the radio button for "Output Range" is selected, with the output range set to \$D\$1. Other options like "Residuals" and "Normal Probability Plots" are available but not selected.



(SKIP) Simple Linear Regression Example: Using PHStat

Add-Ins: PHStat: Regression: Simple Linear Regression

The screenshot shows the Microsoft Excel ribbon with the 'Add-Ins' tab selected. The 'PHStat' add-in is installed and its ribbon is visible. The 'Regression' section is highlighted, and the 'Simple Linear Regression...' option is selected, indicated by a yellow background. A red arrow points from this option to a 'Simple Linear Regression' dialog box on the right. The dialog box contains the following settings:

- Data:**
 - Y Variable Cell Range: Sheet1!\$A\$2:\$A\$11
 - X Variable Cell Range: Sheet1!\$B\$2:\$B\$11
 - First cells in both ranges contain label
 - Confidence level for regression coefficients: 95 %
- Regression Tool Output Options:**
 - Regression Statistics Table
 - ANOVA and Coefficients Table
 - Residuals Table
 - Residual Plot
- Output Options:**
 - Title: [empty text field]
 - Scatter Plot
 - Durbin-Watson Statistic
 - Confidence and Prediction Interval for X = [empty text field]
Confidence level for interval estimates: [empty text field] %

The Excel worksheet on the left displays a table of house data with columns for House Price and Square Feet.

| | A | B | C | D | E | F | G | H |
|----|-------------|-------------|---|---|---|---|---|---|
| 1 | House Price | Square Feet | | | | | | |
| 2 | 245 | 1400 | | | | | | |
| 3 | 312 | 1600 | | | | | | |
| 4 | 279 | 1700 | | | | | | |
| 5 | 308 | 1875 | | | | | | |
| 6 | 199 | 1100 | | | | | | |
| 7 | 219 | 1550 | | | | | | |
| 8 | 405 | 2350 | | | | | | |
| 9 | 324 | 2450 | | | | | | |
| 10 | 319 | 1425 | | | | | | |
| 11 | 255 | 1700 | | | | | | |
| 12 | | | | | | | | |
| 13 | | | | | | | | |
| 14 | | | | | | | | |
| 15 | | | | | | | | |
| 16 | | | | | | | | |
| 17 | | | | | | | | |
| 18 | | | | | | | | |
| 19 | | | | | | | | |
| 20 | | | | | | | | |
| 21 | | | | | | | | |



Simple Linear Regression Example: Excel Output

DCOV_A

| Regression Statistics | |
|-----------------------|----------|
| Multiple R | 0.76211 |
| R Square | 0.58082 |
| Adjusted R Square | 0.52842 |
| Standard Error | 41.33032 |
| Observations | 10 |

The regression equation is:

$$\text{house price} = \widehat{98.24833 + 0.10977 (\text{square feet})}$$

| ANOVA | | | | | |
|------------|----|------------|------------|---------|----------------|
| | df | SS | MS | F | Significance F |
| Regression | 1 | 18934.9348 | 18934.9348 | 11.0848 | 0.01039 |
| Residual | 8 | 13665.5652 | 1708.1957 | | |
| Total | 9 | 32600.5000 | | | |

| | Coefficients | Standard Error | t Stat | P-value | Lower 95% | Upper 95% |
|-------------|--------------|----------------|---------|---------|-----------|-----------|
| Intercept | 98.24833 | 58.03348 | 1.69296 | 0.12892 | -35.57720 | 232.07386 |
| Square Feet | 0.10977 | 0.03297 | 3.32938 | 0.01039 | 0.03374 | 0.18580 |



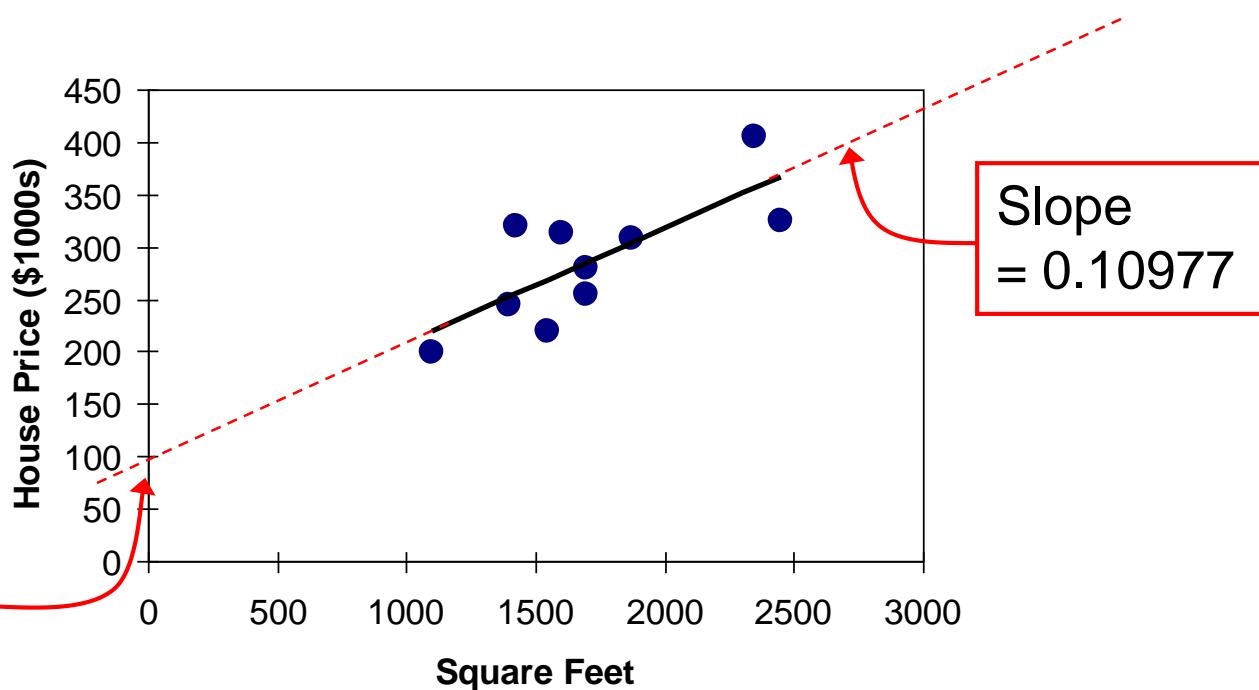
Simple Linear Regression Example: Graphical Representation

DCOV A

House price model: Scatter Plot and Prediction Line



Intercept
= 98.248



$$\widehat{\text{house price}} = 98.24833 + 0.10977 \text{ (square feet)}$$

Simple Linear Regression

Example: Interpretation of b_0

DCOV A

$$\widehat{\text{house price}} = 98.24833 + 0.10977 \text{ (square feet)}$$

- b_0 is the estimated average value of Y when the value of X is zero (if $X = 0$ is in the range of observed X values)
- Because a house cannot have a square footage of 0, b_0 has no practical application



Simple Linear Regression

Example: Interpreting b_1

DCOV_A

$$\widehat{\text{house price}} = 98.24833 + 0.10977 \text{ (square feet)}$$

- b_1 estimates the change in the average value of Y as a result of a one-unit increase in X
 - Here, $b_1 = 0.10977$ tells us that the mean value of a house increases by $.10977(\$1000) = \109.77 , on average, for each additional one square foot of size



Simple Linear Regression Example: Making Predictions

DCOV A

Predict the price for a house
with 2000 square feet:

$$\begin{aligned}\widehat{\text{house price}} &= 98.25 + 0.1098 (\text{sq.ft.}) \\ &= 98.25 + 0.1098(2000) \\ &= 317.85\end{aligned}$$

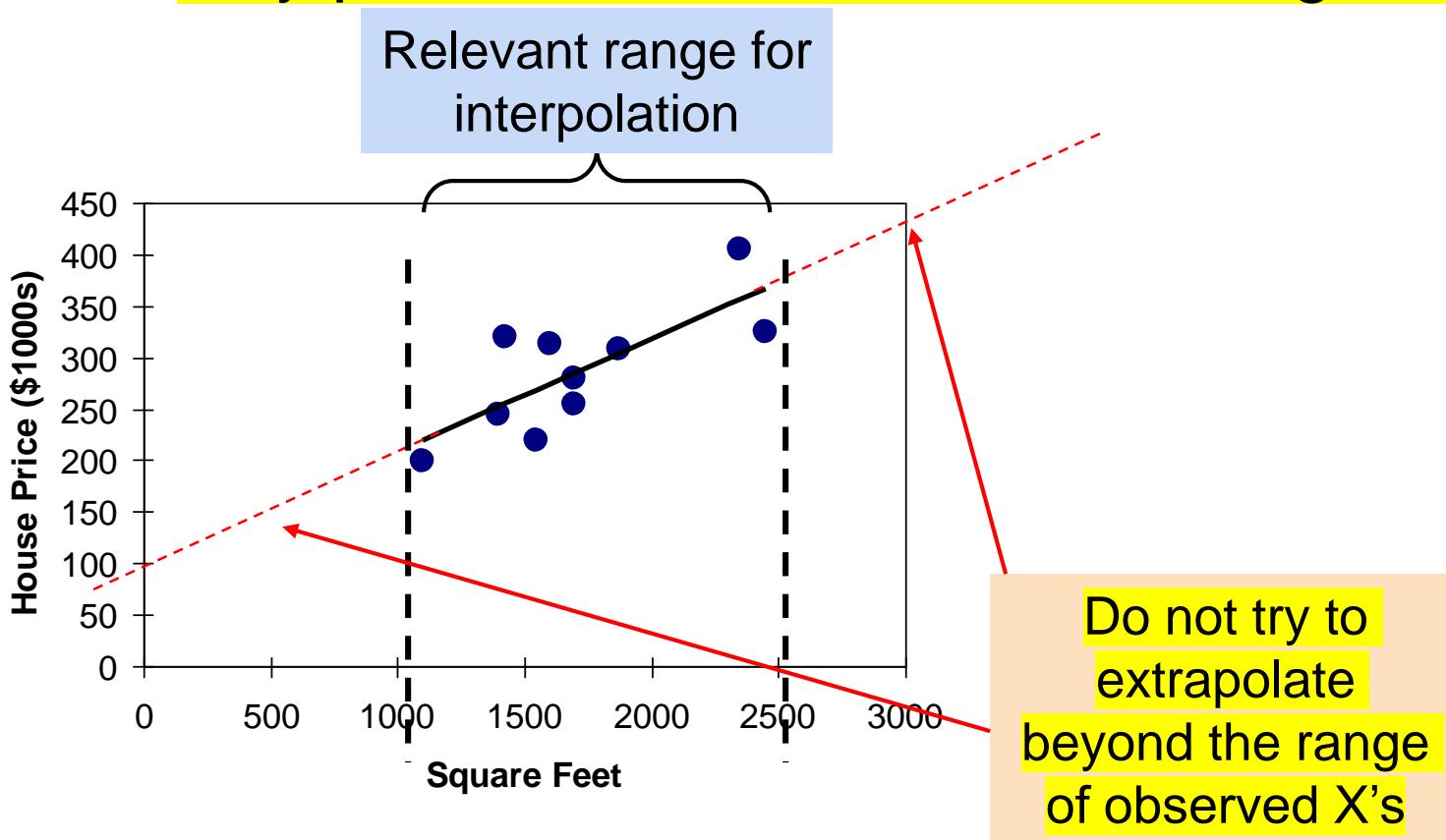
The predicted price for a house with 2000 square feet is 317.85(\$1,000s) = \$317,850



Simple Linear Regression Example: Making Predictions

DCOVA

- When using a regression model for prediction, only predict within the relevant range of data



Measures of Variation

DCOV A

- Total variation is made up of two parts:

$$\text{SST} = \text{SSR} + \text{SSE}$$

Total Sum of
Squares

Regression Sum
of Squares

Error Sum of
Squares

$$\text{SST} = \sum (Y_i - \bar{Y})^2$$

$$\text{SSR} = \sum (\hat{Y}_i - \bar{Y})^2$$

$$\text{SSE} = \sum (Y_i - \hat{Y}_i)^2$$

where:

\bar{Y} = Mean value of the dependent variable

Y_i = Observed value of the dependent variable

\hat{Y}_i = Predicted value of Y for the given X_i value

Measures of Variation

(continued)

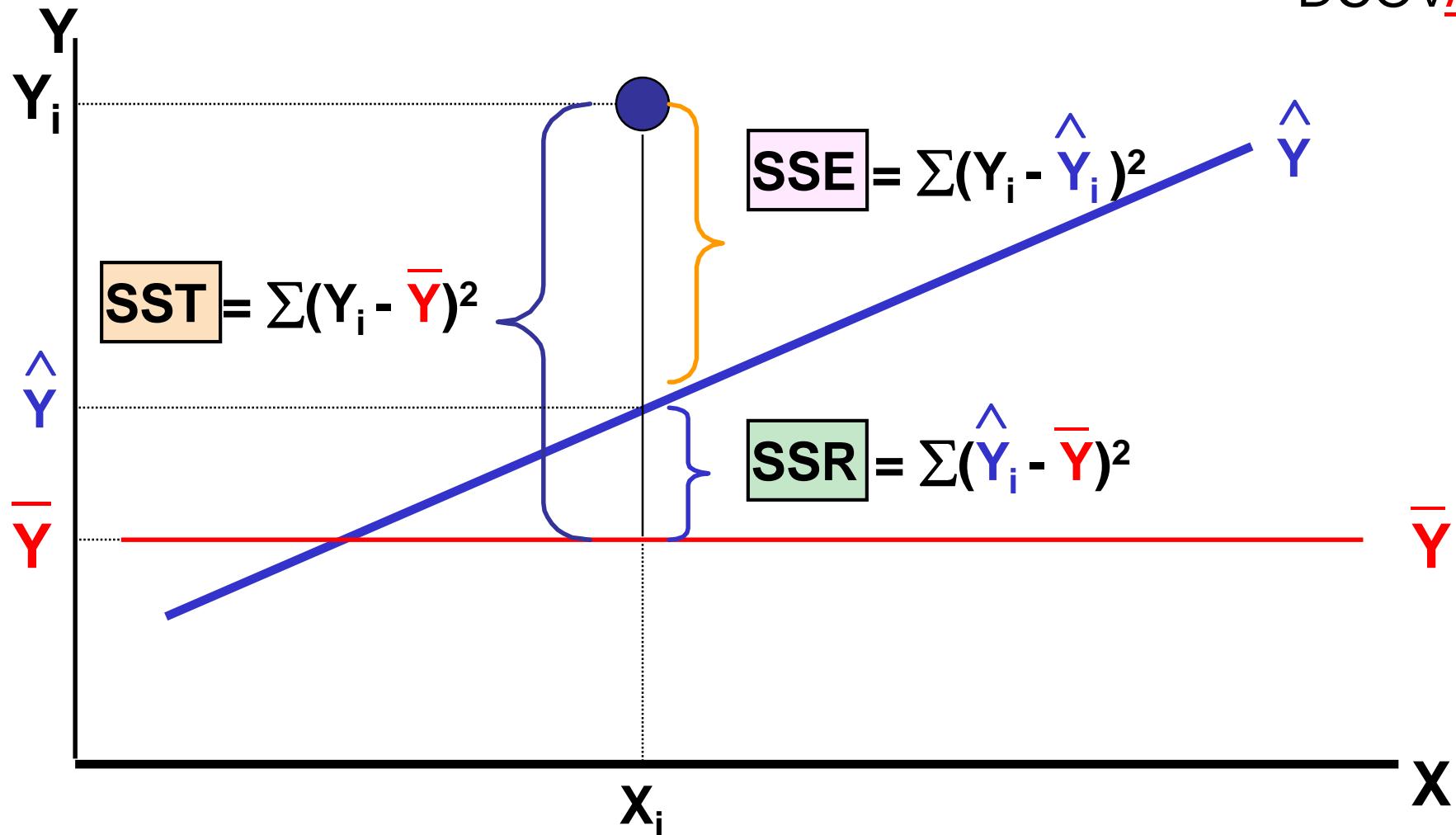
DCOV A

- SST = total sum of squares (Total Variation)
 - Measures the variation of the Y_i values around their mean \bar{Y}
- SSR = regression sum of squares (Explained Variation)
 - Variation attributable to the relationship between X and Y
- SSE = error sum of squares (Unexplained Variation)
 - Variation in Y attributable to factors other than X

Measures of Variation

(continued)

DCOV A



Coefficient of Determination, r^2

DCOV A

- The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable
- The coefficient of determination is also called r-squared and is denoted as r^2

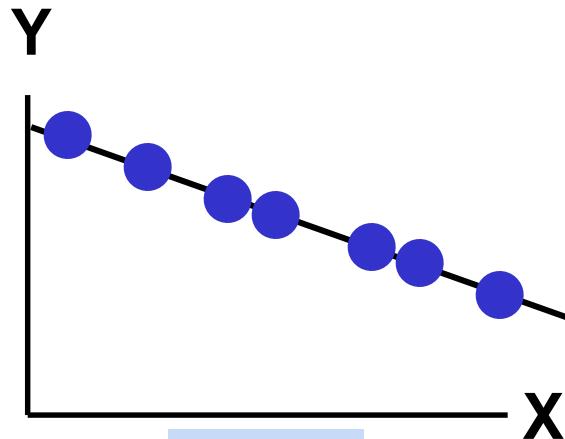
$$r^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

note:

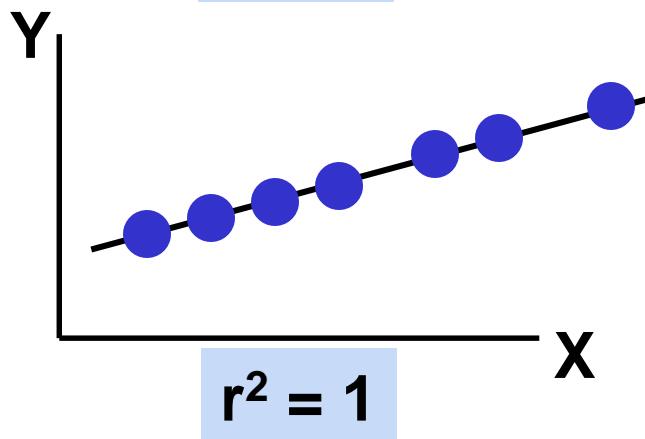
$$0 \leq r^2 \leq 1$$

Examples of Approximate r^2 Values

DCOV A



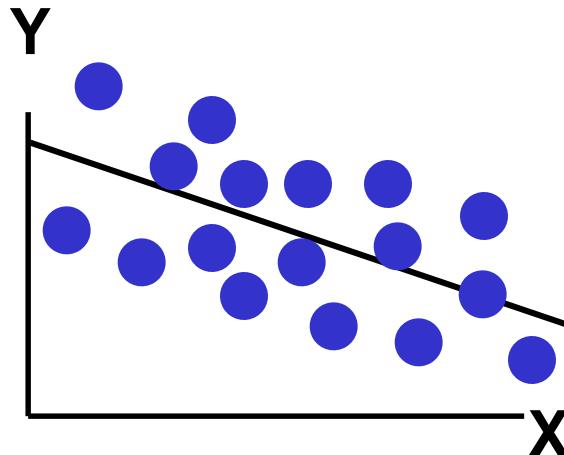
**Perfect linear relationship
between X and Y:**



**100% of the variation in Y is
explained by variation in X**

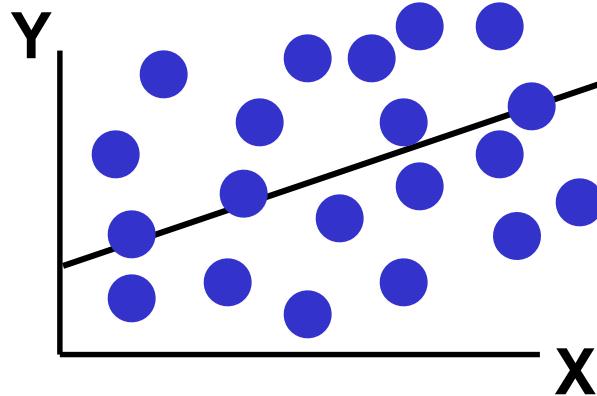
Examples of Approximate r^2 Values

DCOV A



$$0 < r^2 < 1$$

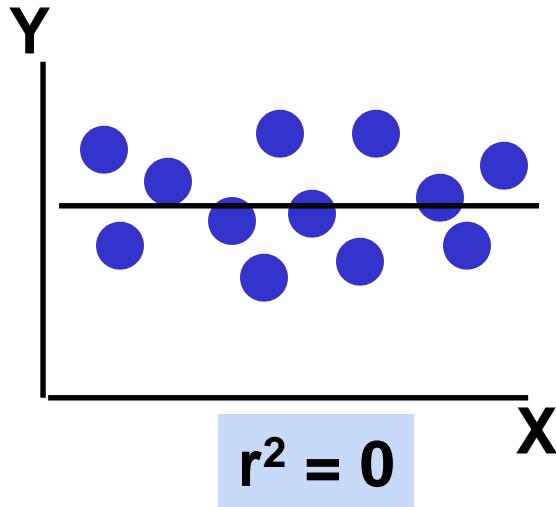
Weaker linear relationships
between X and Y:



Some but not all of the
variation in Y is explained
by variation in X

Examples of Approximate r^2 Values

DCOV A



$$r^2 = 0$$

No linear relationship
between X and Y:

The value of Y does not
depend on X. (None of the
variation in Y is explained
by variation in X)

Simple Linear Regression Example: Coefficient of Determination, r^2 in Excel

DCOV A

| Regression Statistics | |
|-----------------------|----------------|
| Multiple R | 0.76211 |
| R Square | 0.58082 |
| Adjusted R Square | 0.52842 |
| Standard Error | 41.33032 |
| Observations | 10 |

$$r^2 = \frac{SSR}{SST} = \frac{18934.9348}{32600.5000} = 0.58082$$

58.08% of the variation in house prices is explained by variation in square feet

| ANOVA | | df | SS | MS | F | Significance F |
|------------|--|----|------------|------------|---------|----------------|
| Regression | | 1 | 18934.9348 | 18934.9348 | 11.0848 | 0.01039 |
| Residual | | 8 | 13665.5652 | 1708.1957 | | |
| Total | | 9 | 32600.5000 | | | |

| | Coefficients | Standard Error | t Stat | P-value | Lower 95% | Upper 95% |
|-------------|--------------|----------------|---------|---------|-----------|-----------|
| Intercept | 98.24833 | 58.03348 | 1.69296 | 0.12892 | -35.57720 | 232.07386 |
| Square Feet | 0.10977 | 0.03297 | 3.32938 | 0.01039 | 0.03374 | 0.18580 |



Standard Error of Estimate

DCOV A

- The standard deviation of the variation of observations around the regression line is estimated by

$$S_{YX} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}}$$

Where

SSE = error sum of squares

n = sample size

Simple Linear Regression Example: Standard Error of Estimate in Excel

Regression Statistics

DCOV A

| | |
|-------------------|----------|
| Multiple R | 0.76211 |
| R Square | 0.58082 |
| Adjusted R Square | 0.52842 |
| Standard Error | 41.33032 |
| Observations | 10 |

$$S_{YX} = 41.33032$$

ANOVA

| | df | SS | MS | F | Significance F |
|------------|----|------------|------------|---------|----------------|
| Regression | 1 | 18934.9348 | 18934.9348 | 11.0848 | 0.01039 |
| Residual | 8 | 13665.5652 | 1708.1957 | | |
| Total | 9 | 32600.5000 | | | |

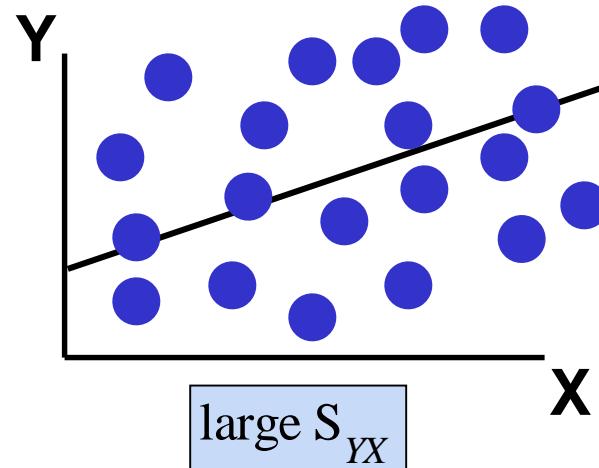
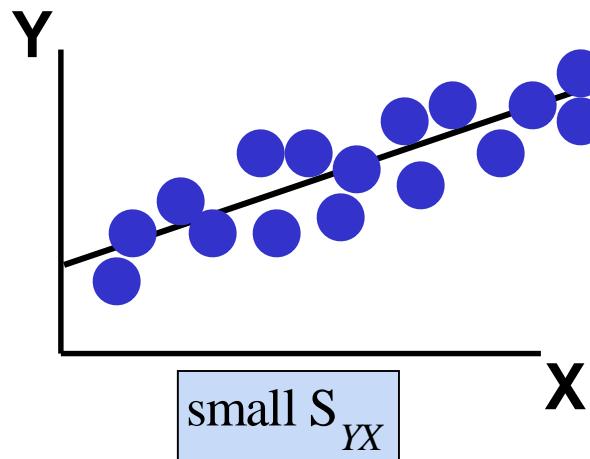
| | Coefficients | Standard Error | t Stat | P-value | Lower 95% | Upper 95% |
|-------------|--------------|----------------|---------|---------|-----------|-----------|
| Intercept | 98.24833 | 58.03348 | 1.69296 | 0.12892 | -35.57720 | 232.07386 |
| Square Feet | 0.10977 | 0.03297 | 3.32938 | 0.01039 | 0.03374 | 0.18580 |



Comparing Standard Errors

DCOVA

S_{YX} is a measure of the variation of observed Y values from the regression line



The magnitude of S_{YX} should always be judged relative to the size of the Y values in the sample data

i.e., $S_{YX} = \$41.33K$ is moderately small relative to house prices in the $\$200K - \$400K$ range

Assumptions of Regression

L.I.N.E

DCOVA

- Linearity
 - The relationship between X and Y is linear
- Independence of Errors
 - Error values are statistically independent
- Normality of Error
 - Error values are normally distributed for any given value of X
- Equal Variance (also called homoscedasticity)
 - The probability distribution of the errors has constant variance

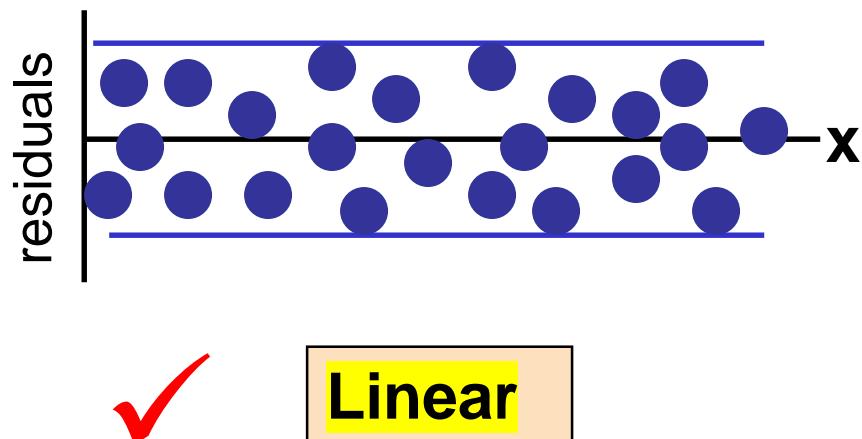
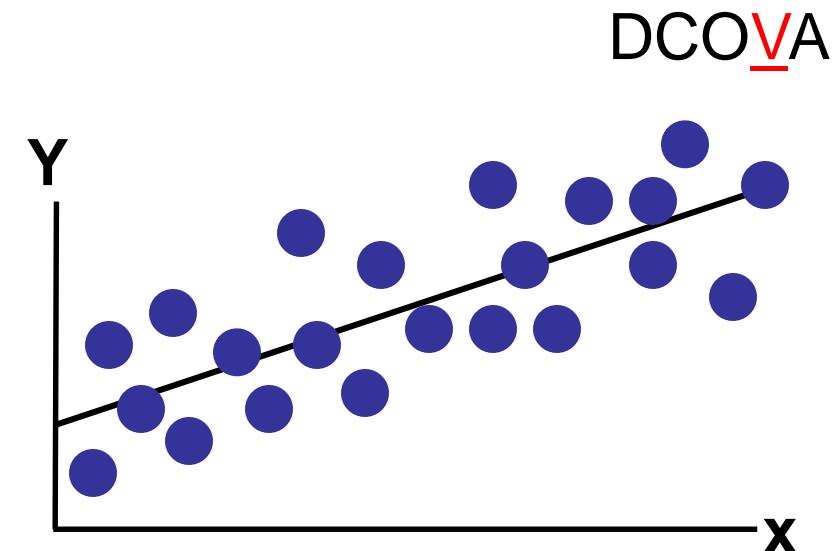
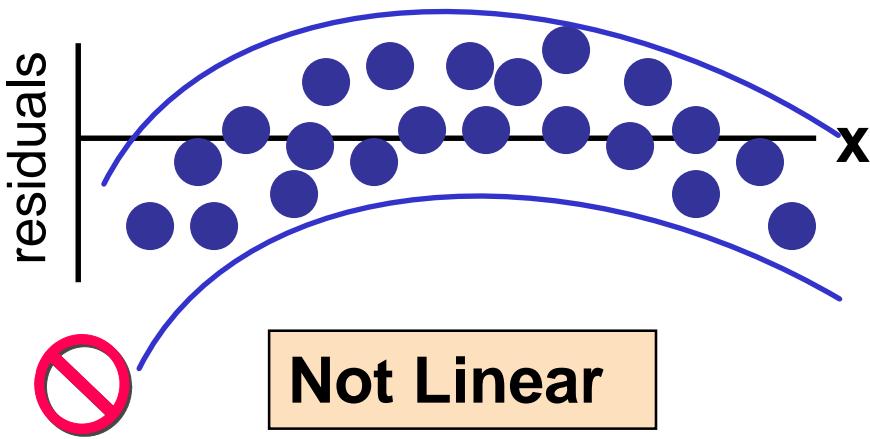
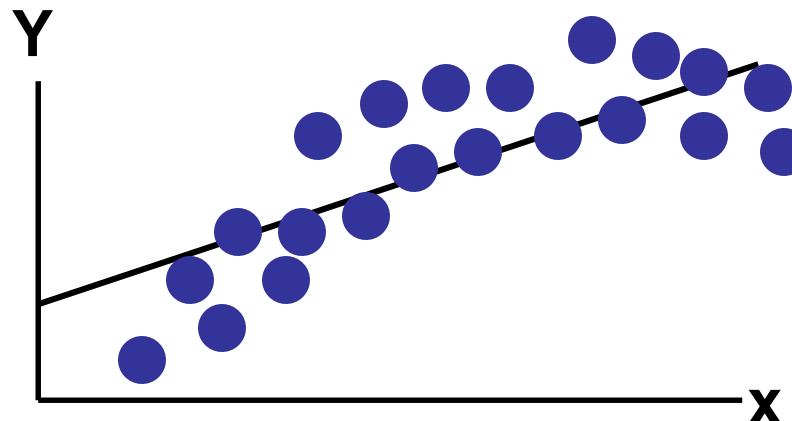
Residual Analysis

DCOVA

$$e_i = Y_i - \hat{Y}_i$$

- The residual for observation i , e_i , is the difference between its observed and predicted value
- Check the assumptions of regression by examining the residuals
 - Examine for linearity assumption
 - Evaluate independence assumption
 - Evaluate normal distribution assumption
 - Examine for constant variance for all levels of X (homoscedasticity)
- Graphical Analysis of Residuals
 - Can plot residuals vs. X

Residual Analysis for Linearity



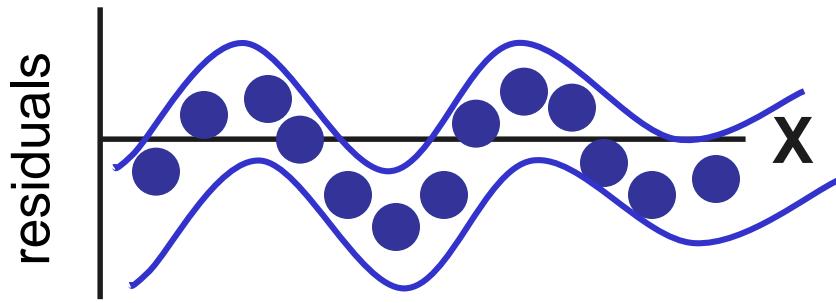
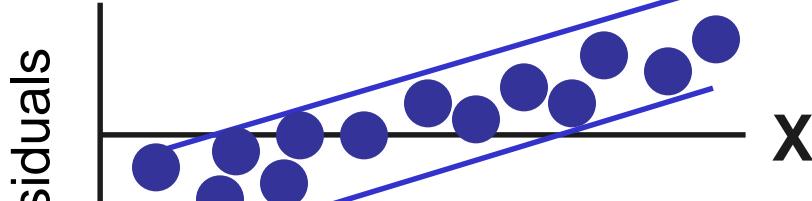
DCOVA

Residual Analysis for Independence

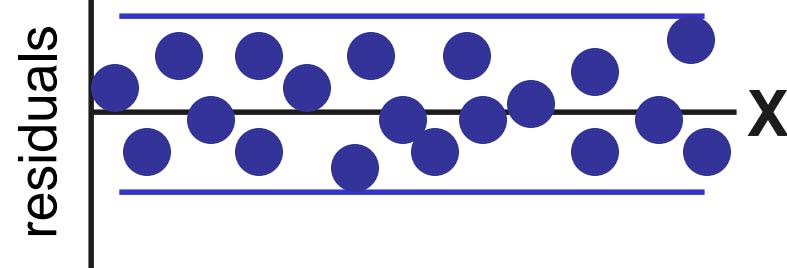
DCOVA



Not Independent



Independent



Checking for Normality

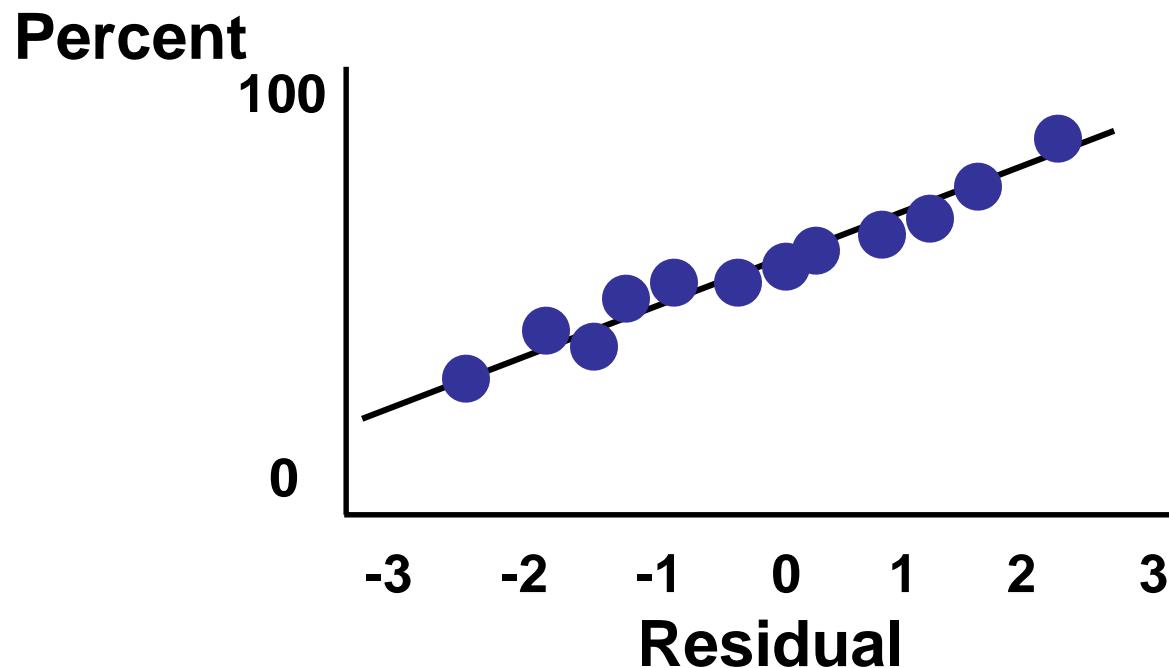
DCOVA

- Examine the Stem-and-Leaf Display of the Residuals
- Examine the Boxplot of the Residuals
- Examine the Histogram of the Residuals
- Construct a Normal Probability Plot of the Residuals

Residual Analysis for Normality

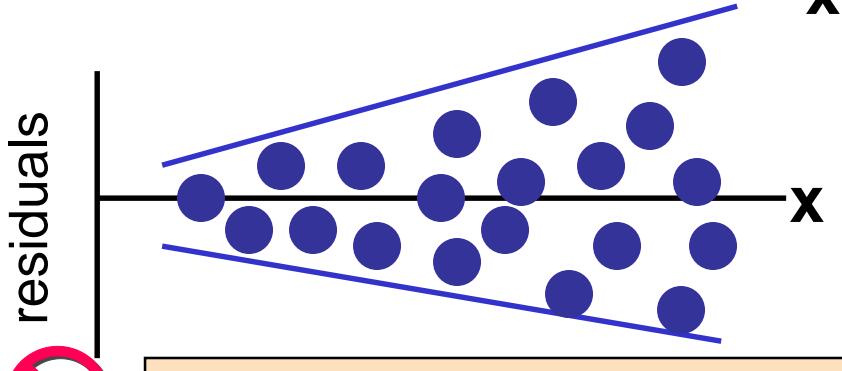
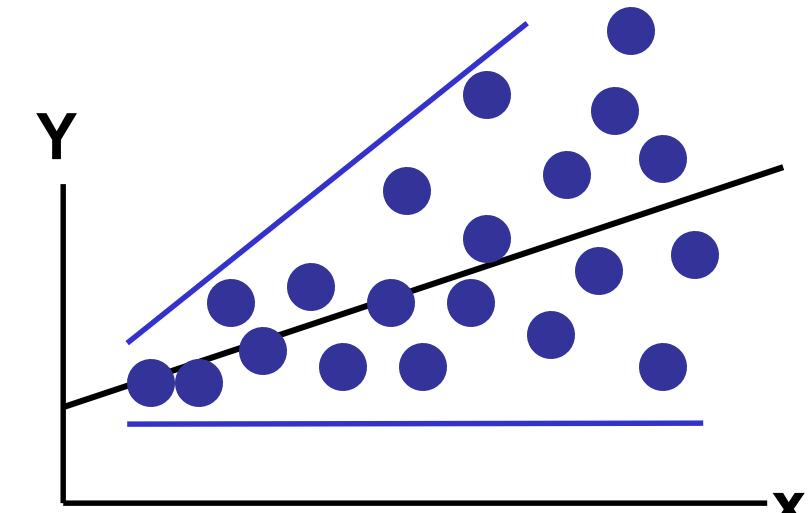
DCOVA

When using a normal probability plot, normal errors will approximately display in a straight line

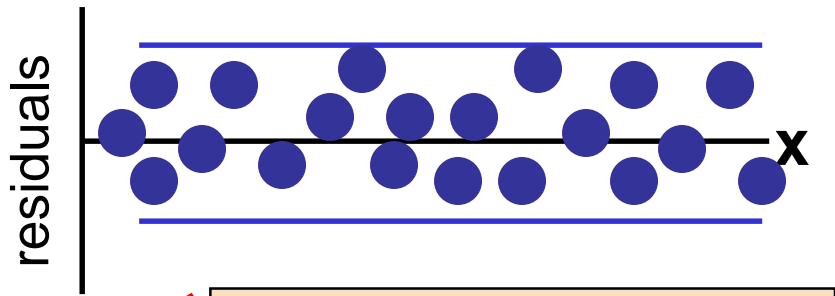
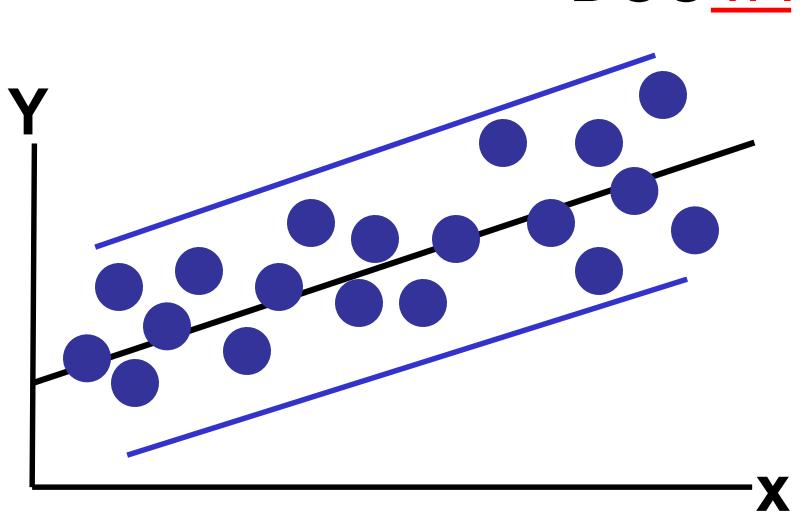


Residual Analysis for Equal Variance

DCOVA



Non-constant variance



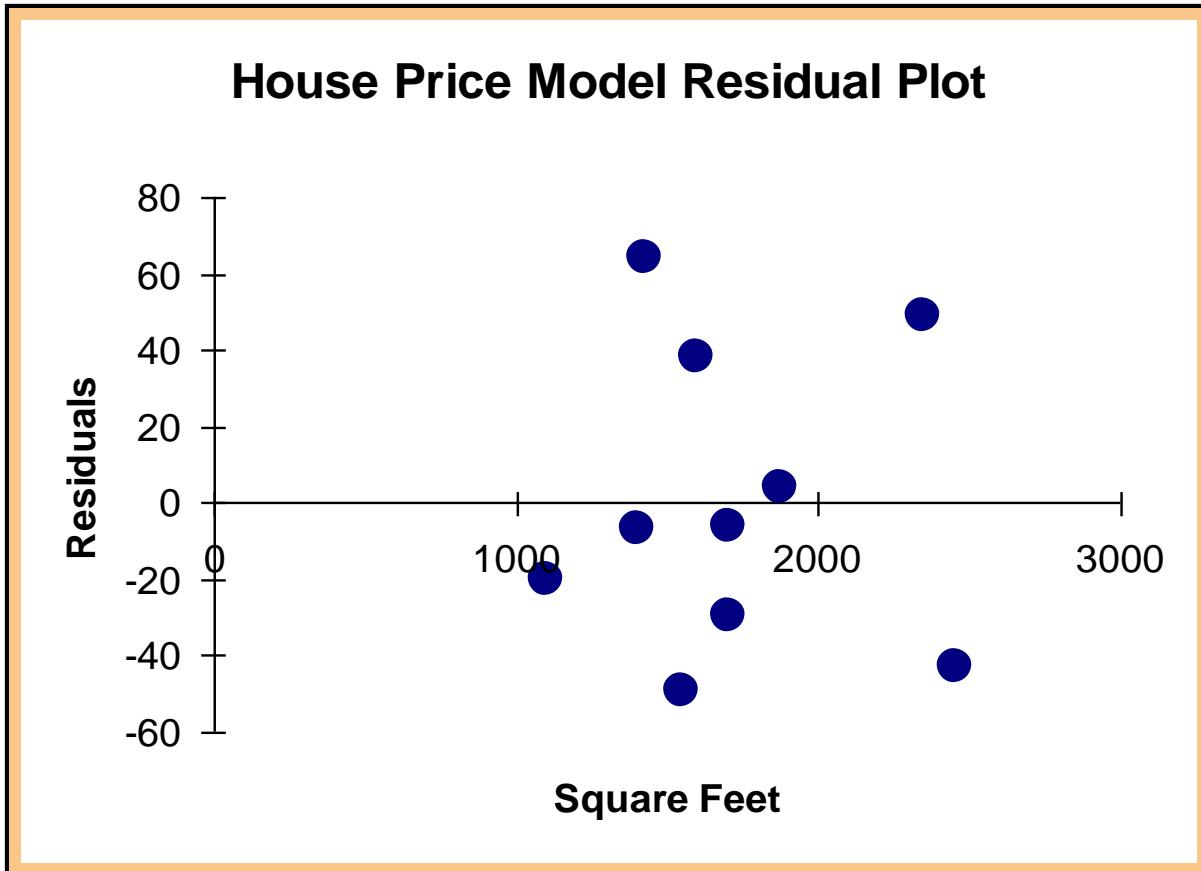
Constant variance

Simple Linear Regression

Example: Excel Residual Output

DCOVA

| RESIDUAL OUTPUT | | |
|-----------------|------------------------------|------------------|
| | <i>Predicted House Price</i> | <i>Residuals</i> |
| 1 | 251.92316 | -6.923162 |
| 2 | 273.87671 | 38.12329 |
| 3 | 284.85348 | -5.853484 |
| 4 | 304.06284 | 3.937162 |
| 5 | 218.99284 | -19.99284 |
| 6 | 268.38832 | -49.38832 |
| 7 | 356.20251 | 48.79749 |
| 8 | 367.17929 | -43.17929 |
| 9 | 254.6674 | 64.33264 |
| 10 | 284.85348 | -29.85348 |

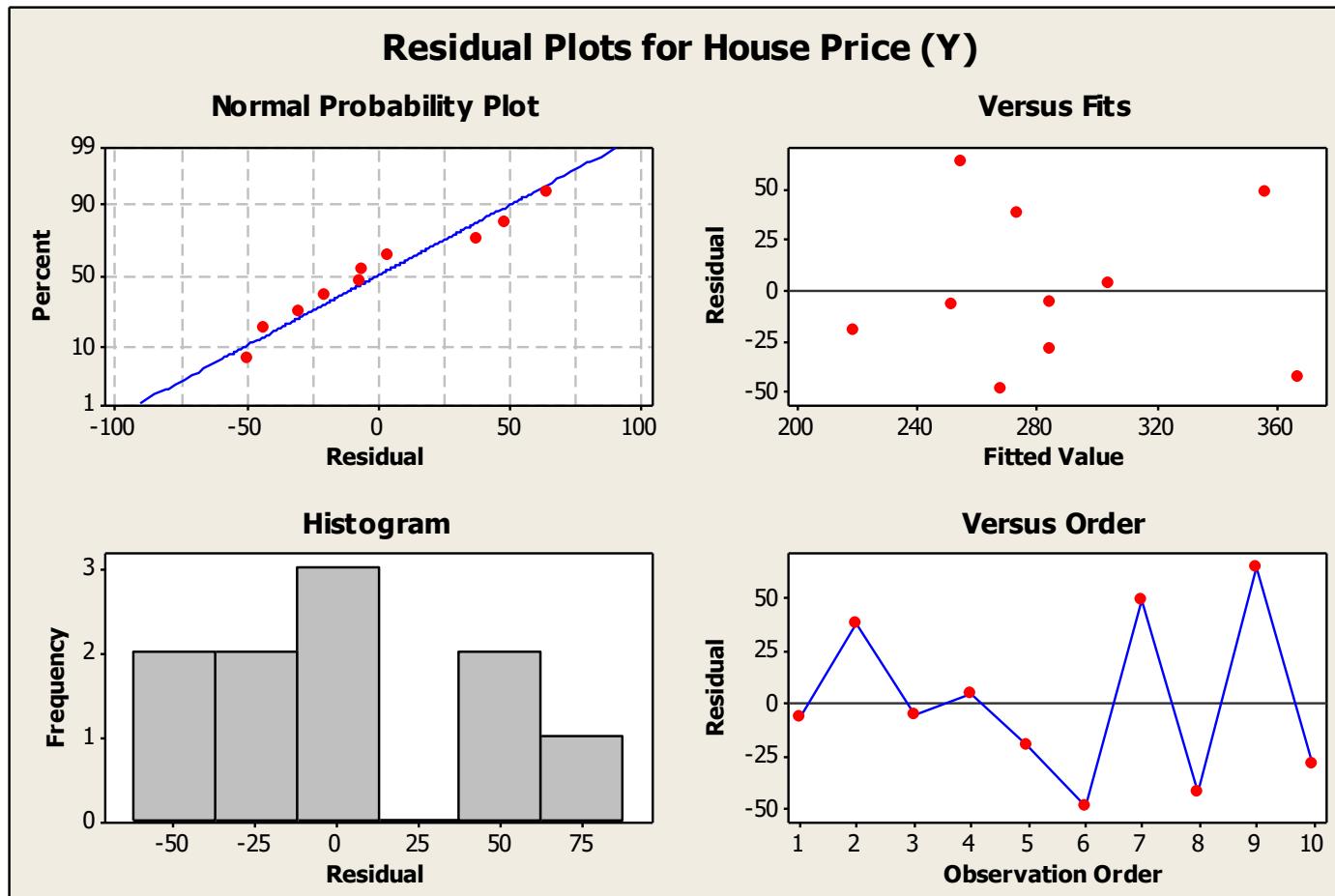


Does not appear to violate
any regression assumptions

Simple Linear Regression

Example: Minitab Residual Output

DCOVA



Does not appear to violate any regression assumptions

Measuring Autocorrelation: The Durbin-Watson Statistic

DCOV A

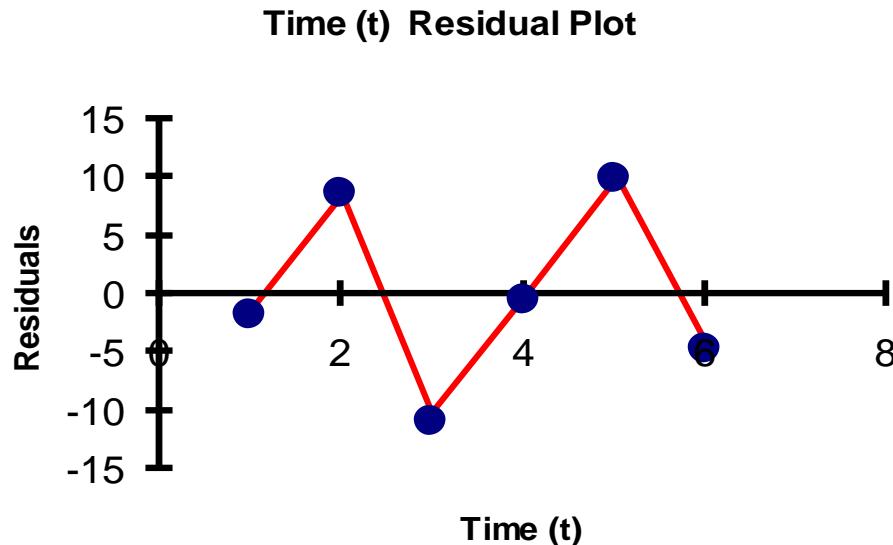
- Used when data are collected over time to detect if autocorrelation is present
- Autocorrelation exists if residuals in one time period are related to residuals in another period

Autocorrelation

DCOVA

- Autocorrelation is correlation of the errors (residuals) over time

- Here, residuals show a cyclic pattern, not random. Cyclical patterns are a sign of positive autocorrelation



- Violates the regression assumption that residuals are random and independent

The Durbin-Watson Statistic

DCOV A

- The Durbin-Watson statistic is used to test for autocorrelation

H_0 : residuals are not correlated

H_1 : positive autocorrelation is present

$$D = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2}$$

- The possible range is $0 \leq D \leq 4$
- D should be close to 2 if H_0 is true
- D less than 2 may signal positive autocorrelation, D greater than 2 may signal negative autocorrelation

Testing for Positive Autocorrelation

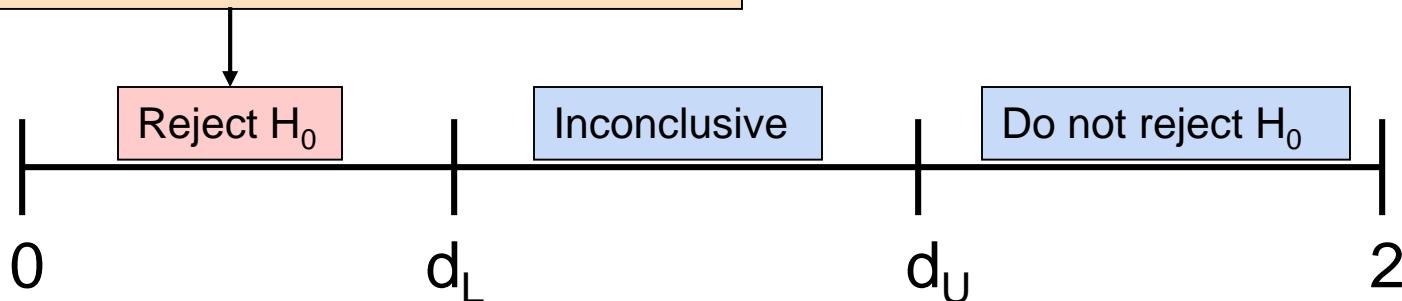
DCOV A

H_0 : positive autocorrelation does not exist

H_1 : positive autocorrelation is present

- Calculate the Durbin-Watson test statistic = D
(The Durbin-Watson Statistic can be found using Excel or Minitab)
- Find the values d_L and d_U from the Durbin-Watson table
(for sample size n and number of independent variables k)

Decision rule: reject H_0 if $D < d_L$

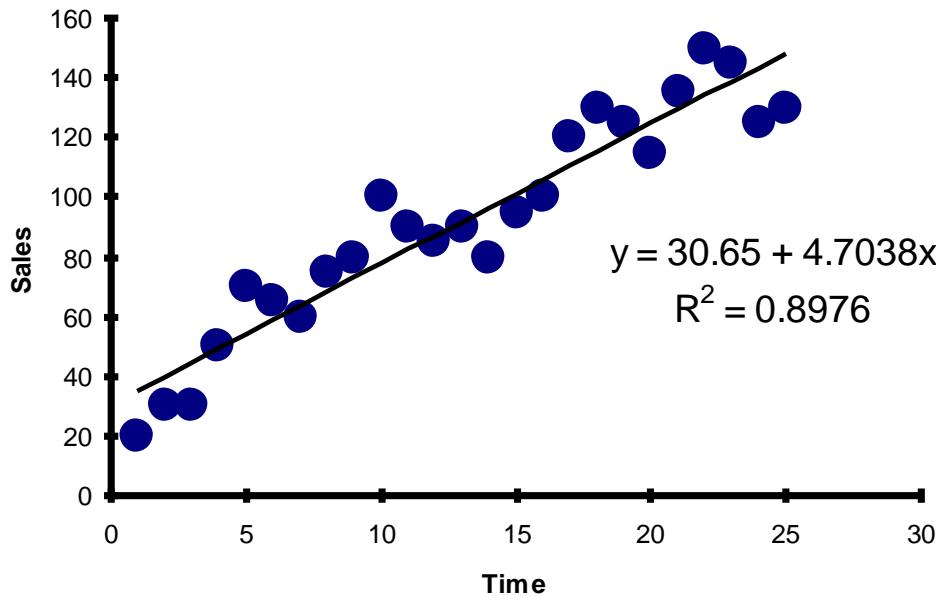


Testing for Positive Autocorrelation

(continued)

DCOVA

- Suppose we have the following time series data:



- Is there autocorrelation?

Testing for Positive Autocorrelation

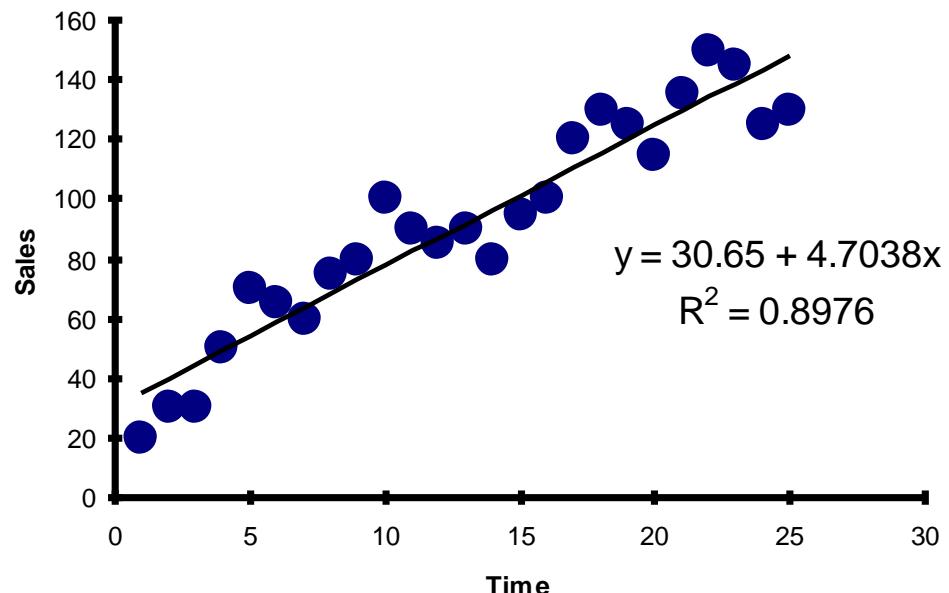
(continued)

DCOV A

- Example with $n = 25$:

Excel/PHStat output:

| Durbin-Watson Calculations | |
|--|---------|
| Sum of Squared Difference of Residuals | 3296.18 |
| Sum of Squared Residuals | 3279.98 |
| Durbin-Watson Statistic | 1.00494 |



$$D = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2} = \frac{3296.18}{3279.98} = 1.00494$$

Testing for Positive Autocorrelation

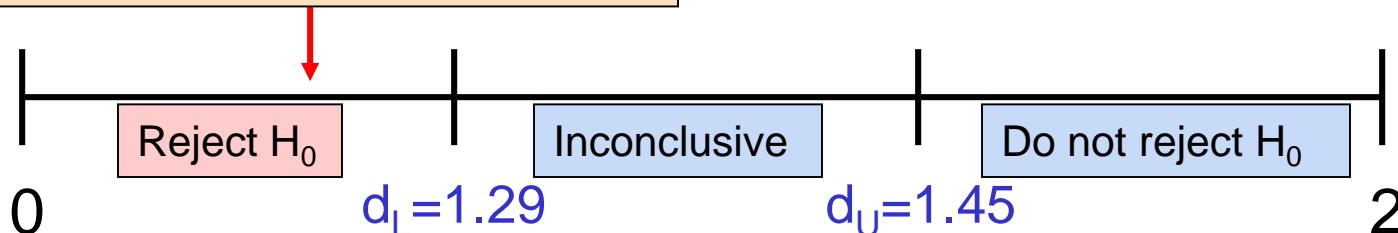
(continued)

DCOVA

- Here, $n = 25$ and there is $k = 1$ one independent variable
- Using the Durbin-Watson table, $d_L = 1.29$ and $d_U = 1.45$
- $D = 1.00494 < d_L = 1.29$, so reject H_0 and conclude that significant positive autocorrelation exists

Decision: **reject H_0** since

$$D = 1.00494 < d_L$$



Inferences About the Slope

DCOV A

- The standard error of the regression slope coefficient (b_1) is estimated by

$$S_{b_1} = \frac{S_{YX}}{\sqrt{SSX}} = \frac{S_{YX}}{\sqrt{\sum(X_i - \bar{X})^2}}$$

where:

S_{b_1} = Estimate of the standard error of the slope

$S_{YX} = \sqrt{\frac{SSE}{n-2}}$ = Standard error of the estimate

Inferences About the Slope: t Test

DCOVA

- t test for a population slope
 - Is there a linear relationship between X and Y?
- Null and alternative hypotheses
 - $H_0: \beta_1 = 0$ (no linear relationship)
 - $H_1: \beta_1 \neq 0$ (linear relationship does exist)
- Test statistic

$$t_{\text{STAT}} = \frac{b_1 - \beta_1}{S_{b_1}}$$

$$\text{d.f.} = n - 2$$

where:

b_1 = regression slope coefficient

β_1 = hypothesized slope

S_{b_1} = standard error of the slope

Inferences About the Slope: t Test Example

DCOVA

| House Price in \$1000s (y) | Square Feet (x) |
|----------------------------------|--------------------|
| 245 | 1400 |
| 312 | 1600 |
| 279 | 1700 |
| 308 | 1875 |
| 199 | 1100 |
| 219 | 1550 |
| 405 | 2350 |
| 324 | 2450 |
| 319 | 1425 |
| 255 | 1700 |

Estimated Regression Equation:

$$\text{house price} = 98.25 + 0.1098 (\text{sq.ft.})$$

The slope of this model is 0.1098

Is there a relationship between the square footage of the house and its sales price?

Inferences About the Slope: t Test Example

DCOVA

From Excel output:

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

| | Coefficients | Standard Error | t Stat | P-value |
|-------------|--------------|----------------|---------|---------|
| Intercept | 98.24833 | 58.03348 | 1.69296 | 0.12892 |
| Square Feet | 0.10977 | 0.03297 | 3.32938 | 0.01039 |

From Minitab output:

$$\begin{matrix} b_1 \\ S_{b_1} \end{matrix}$$

| Predictor | Coef | SE Coef | T | P |
|-------------|---------|---------|------|-------|
| Constant | 98.25 | 58.03 | 1.69 | 0.129 |
| Square Feet | 0.10977 | 0.03297 | 3.33 | 0.010 |

$$b_1$$

$$S_{b_1}$$

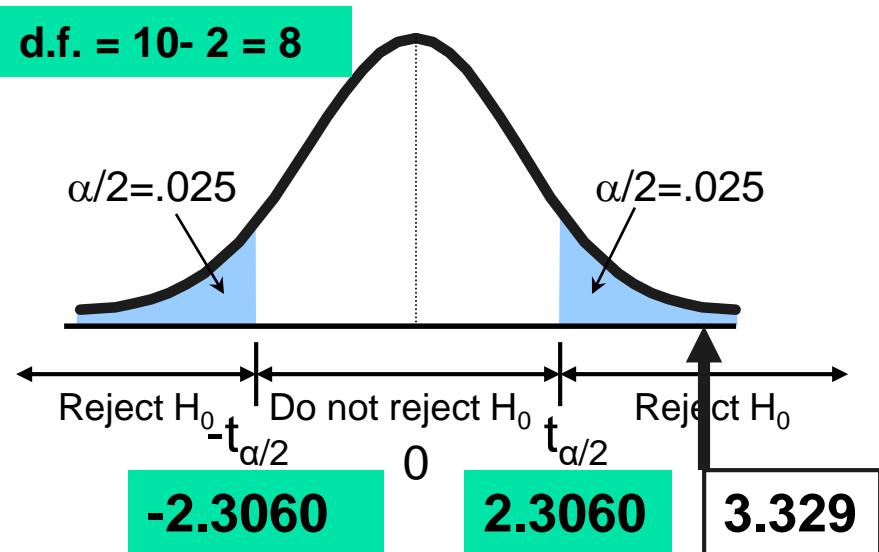
$$t_{\text{STAT}} = \frac{b_1 - \beta_1}{S_{b_1}} = \frac{0.10977 - 0}{0.03297} = 3.32938$$

Inferences About the Slope: t Test Example

DCOV
A

Test Statistic: $t_{\text{STAT}} = 3.329$

$$\begin{aligned}H_0: \beta_1 &= 0 \\H_1: \beta_1 &\neq 0\end{aligned}$$



Decision: Reject H_0

There is sufficient evidence
that square footage affects
house price

Inferences About the Slope: t Test Example

DCOVA

$$H_0: \beta_1 = 0$$

From Excel output: $H_1: \beta_1 \neq 0$

| | Coefficients | Standard Error | t Stat | P-value |
|-------------|--------------|----------------|---------|---------|
| Intercept | 98.24833 | 58.03348 | 1.69296 | 0.12892 |
| Square Feet | 0.10977 | 0.03297 | 3.32938 | 0.01039 |

From Minitab output:

| Predictor | Coef | SE Coef | T | P |
|-------------|---------|---------|------|-------|
| Constant | 98.25 | 58.03 | 1.69 | 0.129 |
| Square Feet | 0.10977 | 0.03297 | 3.33 | 0.010 |

p-value

Decision: Reject H_0 , since p-value < α

There is sufficient evidence that
square footage affects house price.

Jump to slide 63 instead

F Test for Significance

DCOVA

- F Test statistic:

$$F_{STAT} = \frac{MSR}{MSE}$$

where

$$MSR = \frac{SSR}{k}$$

$$MSE = \frac{SSE}{n - k - 1}$$

where F_{STAT} follows an F distribution with k numerator and $(n - k - 1)$ denominator degrees of freedom

$(k = \text{the number of independent variables in the regression model})$

F-Test for Significance

Excel Output

DCOV A

| Regression Statistics | |
|-----------------------|----------|
| Multiple R | 0.76211 |
| R Square | 0.58082 |
| Adjusted R Square | 0.52842 |
| Standard Error | 41.33032 |
| Observations | 10 |

| ANOVA | | | | | |
|------------|----|------------|------------|---------|----------------|
| | df | SS | MS | F | Significance F |
| Regression | 1 | 18934.9348 | 18934.9348 | 11.0848 | 0.01039 |
| Residual | 8 | 13665.5652 | 1708.1957 | | |
| Total | 9 | 32600.5000 | | | |

$$F_{\text{STAT}} = \frac{\text{MSR}}{\text{MSE}} = \frac{18934.9348}{1708.1957} = 11.0848$$

With 1 and 8 degrees of freedom

p-value for the F-Test

F Test for Significance

(continued)

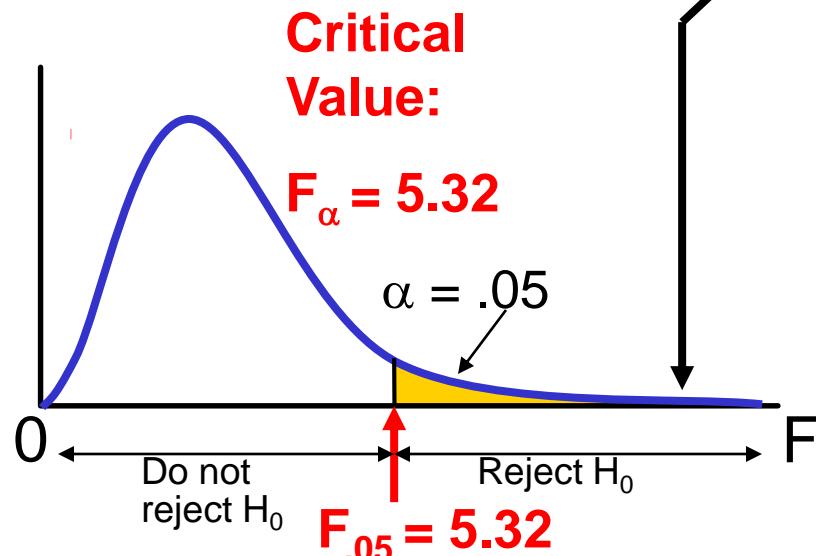
DCOV A

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

$$\alpha = .05$$

$$df_1 = 1 \quad df_2 = 8$$



Test Statistic:

$$F_{\text{STAT}} = \frac{MSR}{MSE} = 11.08$$

Decision:

Reject H_0 at $\alpha = 0.05$

Conclusion:

There is sufficient evidence that house size affects selling price

Pitfalls of Regression Analysis

- Lacking an awareness of the assumptions underlying least-squares regression
- Not knowing how to evaluate the assumptions
- Not knowing the alternatives to least-squares regression if a particular assumption is violated
- Using a regression model without knowledge of the subject matter
- Extrapolating outside the relevant range

Strategies for Avoiding the Pitfalls of Regression

- Start with a scatter plot of X vs. Y to observe possible relationship
- Perform residual analysis to check the assumptions
 - Plot the residuals vs. X to check for violations of assumptions such as homoscedasticity
 - Use a histogram, stem-and-leaf display, boxplot, or normal probability plot of the residuals to uncover possible non-normality

Strategies for Avoiding the Pitfalls of Regression

(continued)

- If there is violation of any assumption, use alternative methods or models
- If there is no evidence of assumption violation, then test for the significance of the regression coefficients and construct confidence intervals and prediction intervals
- Avoid making predictions or forecasts outside the relevant range

Chapter Summary

In this chapter we discussed

- Types of regression models
- The assumptions of regression and correlation
- Determining the simple linear regression equation
- Measures of variation
- Residual analysis
- Measuring autocorrelation

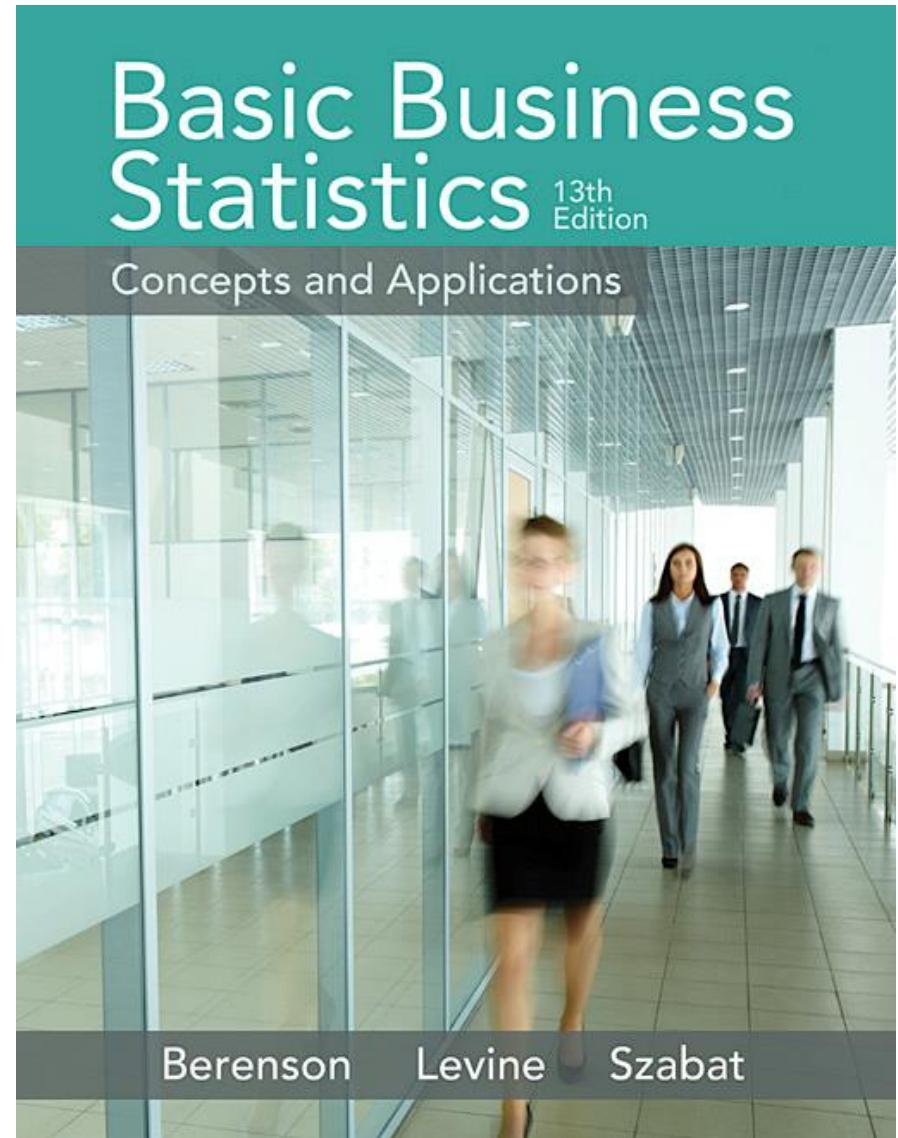
Chapter Summary

(continued)

- Making inferences about the slope
- Correlation -- measuring the strength of the association
- The estimation of mean values and prediction of individual values
- Possible pitfalls in regression and recommended strategies to avoid them

Chapter 14

Introduction to Multiple Regression



Learning Objectives

In this chapter, you learn:

- How to develop a multiple regression model
- How to interpret the regression coefficients
- How to determine which independent variables to include in the regression model
- How to determine which independent variables are most important in predicting a dependent variable
- How to use categorical independent variables in a regression model
- How to predict a categorical dependent variable using logistic regression
- How to identify individual observations that may be unduly influencing the multiple regression model

The Multiple Regression Model

DCOVA

Idea: Examine the linear relationship between
1 dependent (Y) & 2 or more independent variables (X_i)

Multiple Regression Model with k Independent Variables:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

Y-intercept Population slopes Random Error

```
graph TD; A[Y-intercept] --> B[Y_i = β₀ + β₁X₁ᵢ + β₂X₂ᵢ + ... + βₖXₖᵢ + εᵢ]; C[Population slopes] --> B; D[Random Error] --> B;
```

Multiple Regression Equation

DCOVA

The coefficients of the multiple regression model are estimated using sample data

Multiple regression equation with k independent variables:

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_k X_{ki}$$

Estimated (or predicted) value of Y

Estimated intercept

Estimated slope coefficients

```
graph TD; A[Estimated (or predicted) value of Y] --> Y_hat_i; B[Estimated intercept] --> X1i; C[Estimated slope coefficients] --> X2i; C --> Xki;
```

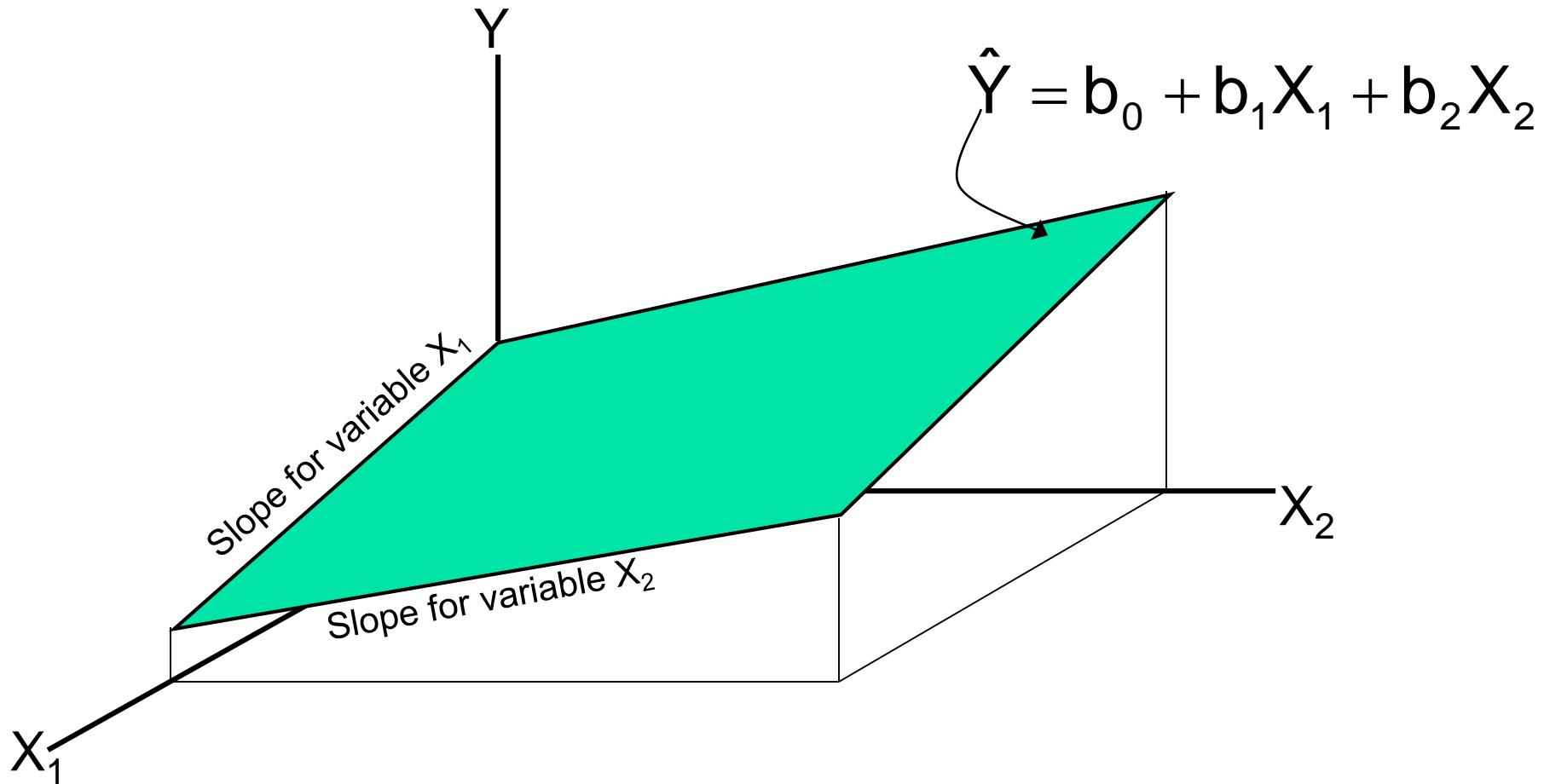
In this chapter we will use Excel to obtain the regression slope coefficients and other regression summary measures.

Multiple Regression Equation

(continued)

Two variable model

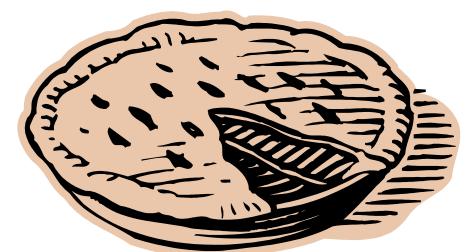
DCOVA



Example: 2 Independent Variables

DCOVA

- A distributor of frozen dessert pies wants to evaluate factors thought to influence demand
 - Dependent variable: Pie sales (units per week)
 - Independent variables: { Price (in \$)
 Advertising (\$100's)
- Data are collected for 15 weeks



Pie Sales Example

| Week | Pie Sales | Price (\$) | Advertising (\$100s) |
|------|-----------|------------|----------------------|
| 1 | 350 | 5.50 | 3.3 |
| 2 | 460 | 7.50 | 3.3 |
| 3 | 350 | 8.00 | 3.0 |
| 4 | 430 | 8.00 | 4.5 |
| 5 | 350 | 6.80 | 3.0 |
| 6 | 380 | 7.50 | 4.0 |
| 7 | 430 | 4.50 | 3.0 |
| 8 | 470 | 6.40 | 3.7 |
| 9 | 450 | 7.00 | 3.5 |
| 10 | 490 | 5.00 | 4.0 |
| 11 | 340 | 7.20 | 3.5 |
| 12 | 300 | 7.90 | 3.2 |
| 13 | 440 | 5.90 | 4.0 |
| 14 | 450 | 5.00 | 3.5 |
| 15 | 300 | 7.00 | 2.7 |

DCOVA
Multiple regression equation:

$$\widehat{\text{Sales}} = b_0 + b_1 * (\text{Price}) + b_2 * (\text{Advertising})$$



Excel Multiple Regression Output

DCOVA

| Regression Statistics | | | | | | |
|--|--------------|----------------|-----------|---------|----------------|-----------|
| Multiple R | 0.72213 | | | | | |
| R Square | 0.52148 | | | | | |
| Adjusted R Square | 0.44172 | | | | | |
| Standard Error | 47.46341 | | | | | |
| Observations | 15 | | | | | |
| $\widehat{\text{Sales}} = 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising})$ | | | | | | |
| ANOVA | | | | | | |
| | df | SS | MS | F | Significance F | |
| Regression | 2 | 29460.027 | 14730.013 | 6.53861 | 0.01201 | |
| Residual | 12 | 27033.306 | 2252.776 | | | |
| Total | 14 | 56493.333 | | | | |
| Coefficients | | | | | | |
| | Coefficients | Standard Error | t Stat | P-value | Lower 95% | Upper 95% |
| Intercept | 306.52619 | 114.25389 | 2.68285 | 0.01993 | 57.58835 | 555.46404 |
| Price | -24.97509 | 10.83213 | -2.30565 | 0.03979 | -48.57626 | -1.37392 |
| Advertising | 74.13096 | 25.96732 | 2.85478 | 0.01449 | 17.55303 | 130.70888 |



The Multiple Regression Equation

DCOVA

$$\widehat{\text{Sales}} = 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising})$$

where

Sales is in number of pies per week

Price is in \$

Advertising is in \$100's.

$b_1 = -24.975$: sales will decrease, on average, by 24.975 pies per week for each \$1 increase in selling price, net of the effects of changes due to advertising

$b_2 = 74.131$: sales will increase, on average, by 74.131 pies per week for each \$100 increase in advertising, net of the effects of changes due to price



Using The Equation to Make Predictions

DCOVA

Predict sales for a week in which the selling price is \$5.50 and advertising is \$350:

$$\begin{aligned}\widehat{\text{Sales}} &= 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising}) \\ &= 306.526 - 24.975(5.50) + 74.131(3.5) \\ &= 428.62\end{aligned}$$

Predicted sales
is 428.62 pies

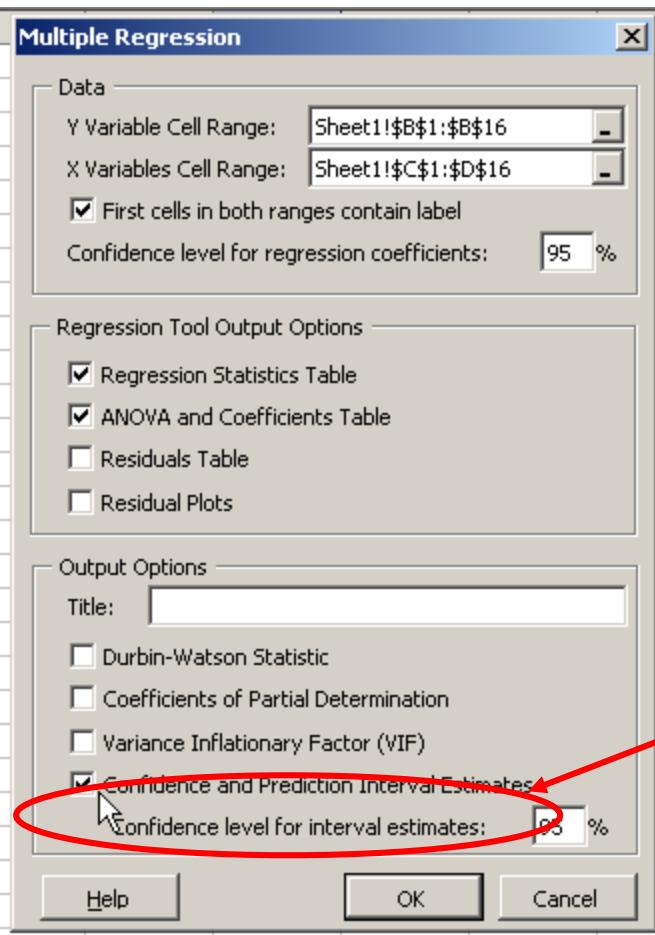
Note that Advertising is
in \$100s, so \$350 means
that $X_2 = 3.5$

Predictions in Excel using PHStat

DCOVA

- PHStat | regression | multiple regression ...

| | A | B | C | D |
|----|------|-----------|-------|-------------|
| 1 | Week | Pie Sales | Price | Advertising |
| 2 | 1 | 350 | 5.5 | 3.3 |
| 3 | 2 | 460 | 7.5 | 3.3 |
| 4 | 3 | 350 | 8 | 3 |
| 5 | 4 | 430 | 8 | 4.5 |
| 6 | 5 | 350 | 6.8 | 3 |
| 7 | 6 | 380 | 7.5 | 4 |
| 8 | 7 | 430 | 4.5 | 3 |
| 9 | 8 | 470 | 6.4 | 3.7 |
| 10 | 9 | 450 | 7 | 3.5 |
| 11 | 10 | 490 | 5 | 4 |
| 12 | 11 | 340 | 7.2 | 3.5 |
| 13 | 12 | 300 | 7.9 | 3.2 |
| 14 | 13 | 440 | 5.9 | 4 |
| 15 | 14 | 450 | 5 | 3.5 |
| 16 | 15 | 300 | 7 | 2.7 |
| 17 | | | | |
| 18 | | | | |
| 19 | | | | |
| 20 | | | | |
| 21 | | | | |
| 22 | | | | |
| 23 | | | | |
| 24 | | | | |
| 25 | | | | |
| 26 | | | | |
| 27 | | | | |

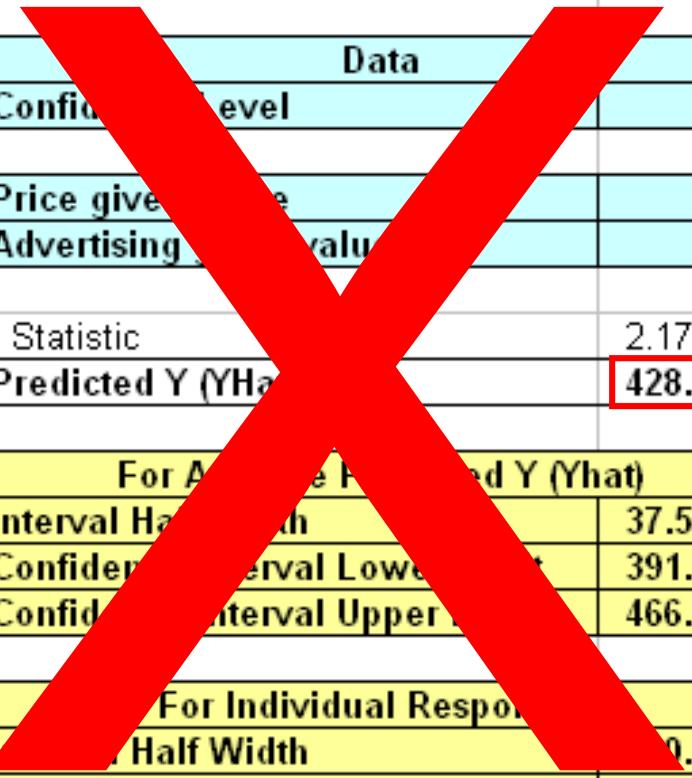


Check the
“confidence and
prediction interval
estimates” box

Predictions in PHStat

(continued)

DCOVA



| A | B |
|--|----------|
| 1 Confidence and Prediction Estimate Intervals | |
| 2 | |
| 3 Data | |
| 4 Confidence Level | 95% |
| 5 | |
| 6 Price given | 5.5 |
| 7 Advertising | 3.5 |
| 8 | |
| 20 t Statistic | 2.178813 |
| 21 Predicted Y (\hat{Y}) | 428.6216 |
| 22 | |
| 23 For A Single Predicted Y (\hat{Y}) | |
| 24 Interval Half Width | 37.50306 |
| 25 Confidence Interval Lower Limit | 391.1185 |
| 26 Confidence Interval Upper Limit | 466.1246 |
| 27 | |
| 28 For Individual Response | |
| 29 Interval Half Width | 0.0041 |
| 30 Prediction Interval Lower Limit | 318.6174 |
| 31 Prediction Interval Upper Limit | 538.6257 |

Can't do
CI without
PH Stat

Input values

Predicted \hat{Y} value

Confidence interval for the
mean value of Y, given
these X values

Prediction interval for an
individual Y value, given
these X values

The Coefficient of Multiple Determination, r^2

DCOVA

- Reports the proportion of total variation in Y explained by all X variables taken together

$$r^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

Multiple Coefficient of Determination In Excel

DCOVA

| Regression Statistics | | | | | | |
|--|--------------|----------------|-----------|---------|----------------|-----------|
| Multiple R | 0.72213 | | | | | |
| R Square | 0.52148 | | | | | |
| Adjusted R Square | 0.44172 | | | | | |
| Standard Error | 47.46341 | | | | | |
| Observations | 15 | | | | | |
| $r^2 = \frac{SSR}{SST} = \frac{29460.0}{56493.3} = .52148$ | | | | | | |
| 52.1% of the variation in pie sales is explained by the variation in price and advertising | | | | | | |
| ANOVA | | | | | | |
| | df | SS | MS | F | Significance F | |
| Regression | 2 | 29460.027 | 14730.013 | 6.53861 | 0.01201 | |
| Residual | 12 | 27033.306 | 2252.776 | | | |
| Total | 14 | 56493.333 | | | | |
| | Coefficients | Standard Error | t Stat | P-value | Lower 95% | Upper 95% |
| Intercept | 306.52619 | 114.25389 | 2.68285 | 0.01993 | 57.58835 | 555.46404 |
| Price | -24.97509 | 10.83213 | -2.30565 | 0.03979 | -48.57626 | -1.37392 |
| Advertising | 74.13096 | 25.96732 | 2.85478 | 0.01449 | 17.55303 | 130.70888 |

Adjusted r^2

DCOV A

- r^2 never decreases when a new X variable is added to the model
 - This can be a disadvantage when comparing models
- What is the net effect of adding a new variable?
 - We lose a degree of freedom when a new X variable is added
 - Did the new X variable add enough explanatory power to offset the loss of one degree of freedom?

(continued)

Adjusted r^2

DCOVA

- Shows the proportion of variation in Y explained by all X variables adjusted for the number of X variables used

$$r_{adj}^2 = 1 - \left[(1 - r^2) \left(\frac{n - 1}{n - k - 1} \right) \right]$$

(where n = sample size, k = number of independent variables)

- Penalizes excessive use of unimportant independent variables
- Smaller than r^2
- Useful in comparing among models

Adjusted r^2 in Excel

DCOVA

| Regression Statistics | |
|-----------------------|----------|
| Multiple R | 0.72213 |
| R Square | 0.52148 |
| Adjusted R Square | 0.44172 |
| Standard Error | 47.46341 |
| Observations | 15 |

| | | $r^2_{adj} = .44172$ | |
|---|--|----------------------|--|
| 44.2% of the variation in pie sales is explained by the variation in price and advertising, taking into account the sample size and number of independent variables | | | |

| ANOVA | | df | SS | MS | F | Significance F |
|------------|--|----|-----------|-----------|---------|----------------|
| Regression | | 2 | 29460.027 | 14730.013 | 6.53861 | 0.01201 |
| Residual | | 12 | 27033.306 | 2252.776 | | |
| Total | | 14 | 56493.333 | | | |

| | Coefficients | Standard Error | t Stat | P-value | Lower 95% | Upper 95% |
|-------------|--------------|----------------|----------|---------|-----------|-----------|
| Intercept | 306.52619 | 114.25389 | 2.68285 | 0.01993 | 57.58835 | 555.46404 |
| Price | -24.97509 | 10.83213 | -2.30565 | 0.03979 | -48.57626 | -1.37392 |
| Advertising | 74.13096 | 25.96732 | 2.85478 | 0.01449 | 17.55303 | 130.70888 |



Is the Model Significant?

DCOVA

- F Test for Overall Significance of the Model
- Shows if there is a linear relationship between all of the X variables considered together and Y
- Use F-test statistic
- Hypotheses:

$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ (no linear relationship)

$H_1: \text{at least one } \beta_i \neq 0$ (at least one independent variable affects Y)

F Test for Overall Significance

DCOVA

- Test statistic:

$$F_{STAT} = \frac{MSR}{MSE} = \frac{\frac{SSR}{k}}{\frac{SSE}{n - k - 1}}$$

where F_{STAT} has numerator d.f. = k and
denominator d.f. = $(n - k - 1)$

F Test for Overall Significance In Excel

DCOVA
(continued)

| Regression Statistics | | | | | | |
|--|-----------|-----------------------|---------------|----------------|-----------------------|------------------|
| Multiple R | 0.72213 | | | | | |
| R Square | 0.52148 | | | | | |
| Adjusted R Square | 0.44172 | | | | | |
| Standard Error | 47.46341 | | | | | |
| Observations | 15 | | | | | |
| $F_{STAT} = \frac{MSR}{MSE} = \frac{14730.0}{2252.8} = 6.5386$ | | | | | | |
| With 2 and 12 degrees of freedom | | | | | | |
| ANOVA | | | | | | |
| | <i>df</i> | SS | MS | <i>F</i> | <i>Significance F</i> | |
| Regression | 2 | 29460.027 | 14730.013 | 6.53861 | 0.01201 | |
| Residual | 12 | 27033.306 | 2252.776 | | | |
| Total | 14 | 56493.333 | | | | |
| Coefficients | | | | | | |
| | | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> | <i>Lower 95%</i> | <i>Upper 95%</i> |
| Intercept | 306.52619 | 114.25389 | 2.68285 | 0.01993 | 57.58835 | 555.46404 |
| Price | -24.97509 | 10.83213 | -2.30565 | 0.03979 | -48.57626 | -1.37392 |
| Advertising | 74.13096 | 25.96732 | 2.85478 | 0.01449 | 17.55303 | 130.70888 |

F Test for Overall Significance

(continued)

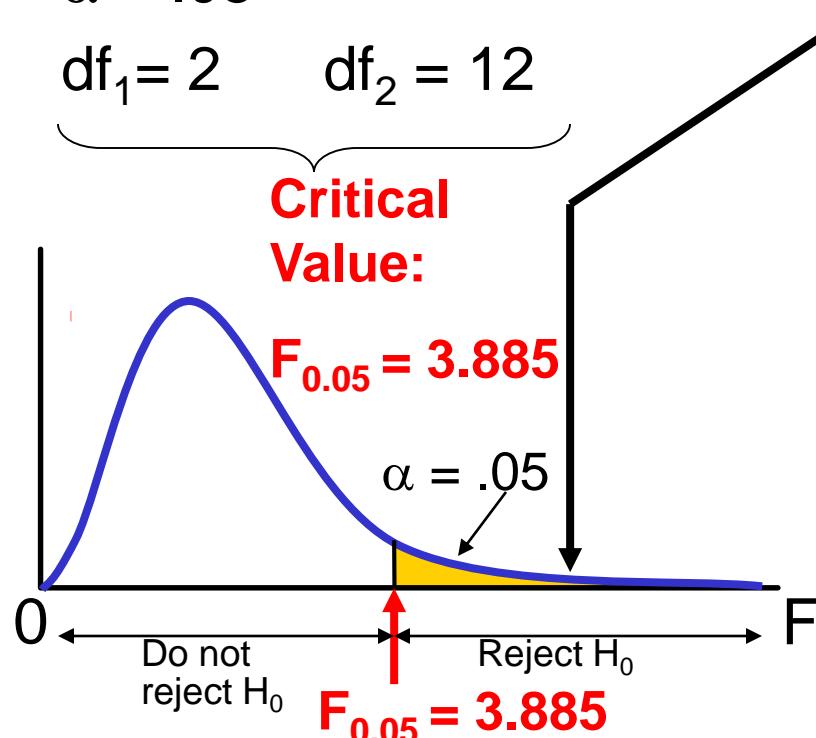
DCOVA

$$H_0: \beta_1 = \beta_2 = 0$$

$$H_1: \beta_1 \text{ and } \beta_2 \text{ not both zero}$$

$$\alpha = .05$$

$$df_1 = 2 \quad df_2 = 12$$



Test Statistic:

$$F_{\text{STAT}} = \frac{MSR}{MSE} = 6.5386$$

Decision:

Since F_{STAT} test statistic is in the rejection region ($p\text{-value} < .05$), reject H_0

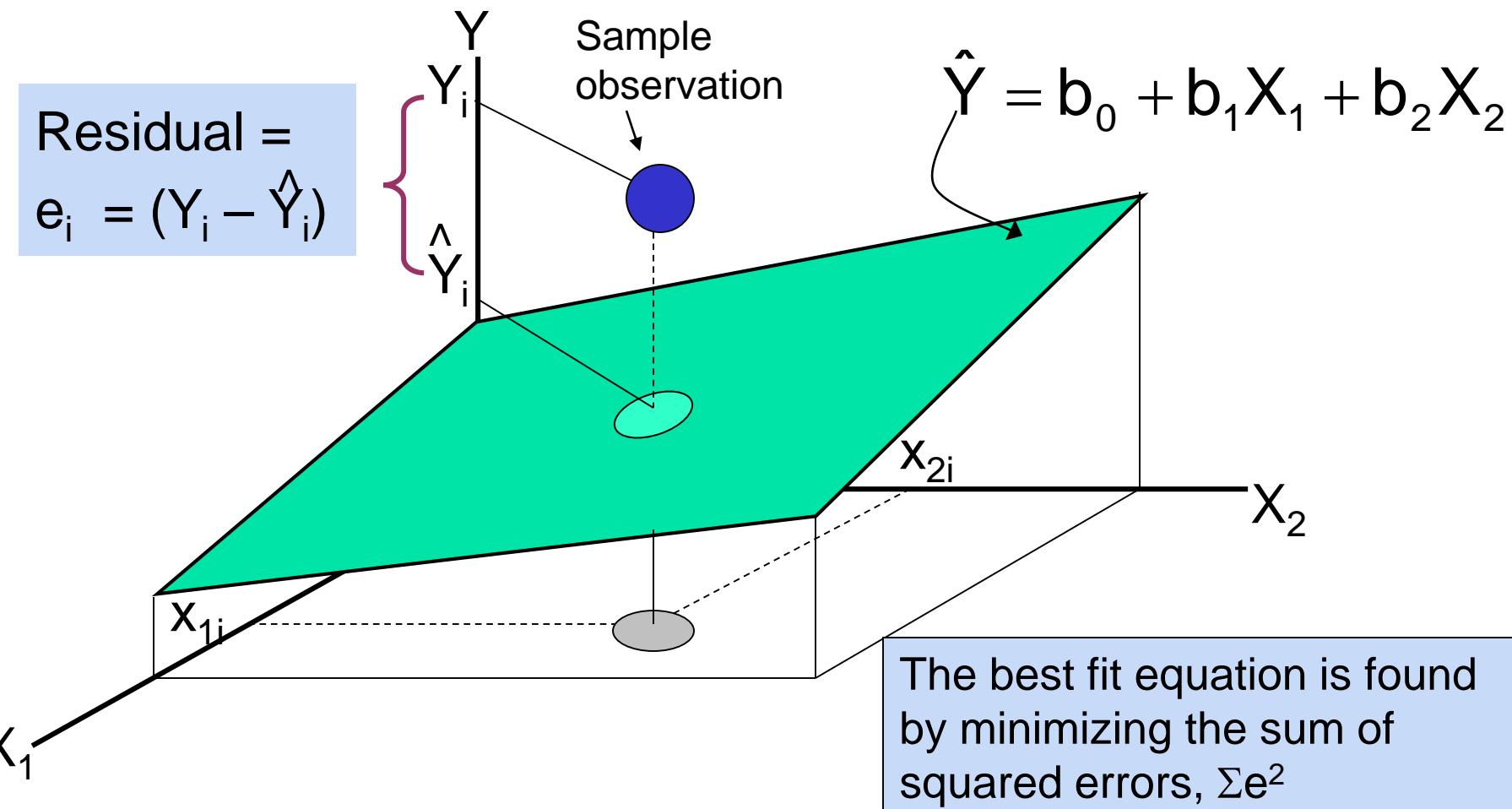
Conclusion:

There is evidence that at least one independent variable affects Y

Residuals in Multiple Regression

DCOVA

Two variable model



Multiple Regression Assumptions

DCOVA

Errors (residuals) from the regression model:

$$e_i = (Y_i - \hat{Y}_i)$$

Assumptions:

- The errors are normally distributed
- Errors have a constant variance
- The model errors are independent

Residual Plots Used in Multiple Regression

DCOVA

- These residual plots are used in multiple regression:
 - Residuals vs. \hat{Y}_i
 - Residuals vs. X_{1i}
 - Residuals vs. X_{2i}
 - Residuals vs. time (if time series data)

Use the residual plots to check for violations of regression assumptions

Are Individual Variables Significant?

DCOVA

- Use t tests of individual variable slopes
- Shows if there is a linear relationship between the variable X_j and Y holding constant the effects of other X variables
- Hypotheses:

- $H_0: \beta_j = 0$ (no linear relationship)
- $H_1: \beta_j \neq 0$ (linear relationship does exist between X_j and Y)

Are Individual Variables Significant?

(continued)

DCOV**A**

$H_0: \beta_j = 0$ (no linear relationship between X_j and Y)

$H_1: \beta_j \neq 0$ (linear relationship does exist between X_j and Y)

Test Statistic:

$$t_{STAT} = \frac{b_j - 0}{S_{b_j}}$$

(df = n - k - 1)

Are Individual Variables Significant? Excel Output

(continued)
DCOVA

| Regression Statistics | | | | | | |
|-----------------------|--------------|----------------|-----------|---------|----------------|-----------|
| Multiple R | 0.72213 | | | | | |
| R Square | 0.52148 | | | | | |
| Adjusted R Square | 0.44172 | | | | | |
| Standard Error | 47.46341 | | | | | |
| Observations | 15 | | | | | |
| ANOVA | | | | | | |
| | df | SS | MS | F | Significance F | |
| Regression | 2 | 29460.027 | 14730.013 | 6.53861 | 0.01201 | |
| Residual | 12 | 27033.306 | 2252.776 | | | |
| Total | 14 | 56493.333 | | | | |
| Coefficients | | | | | | |
| | Coefficients | Standard Error | t Stat | P-value | Lower 95% | Upper 95% |
| Intercept | 306.52619 | 114.25389 | 2.68285 | 0.01993 | 57.58835 | 555.46404 |
| Price | -24.97509 | 10.83213 | -2.30565 | 0.03979 | -48.57626 | -1.37392 |
| Advertising | 74.13096 | 25.96732 | 2.85478 | 0.01449 | 17.55303 | 130.70888 |

Inferences about the Slope: t Test Example

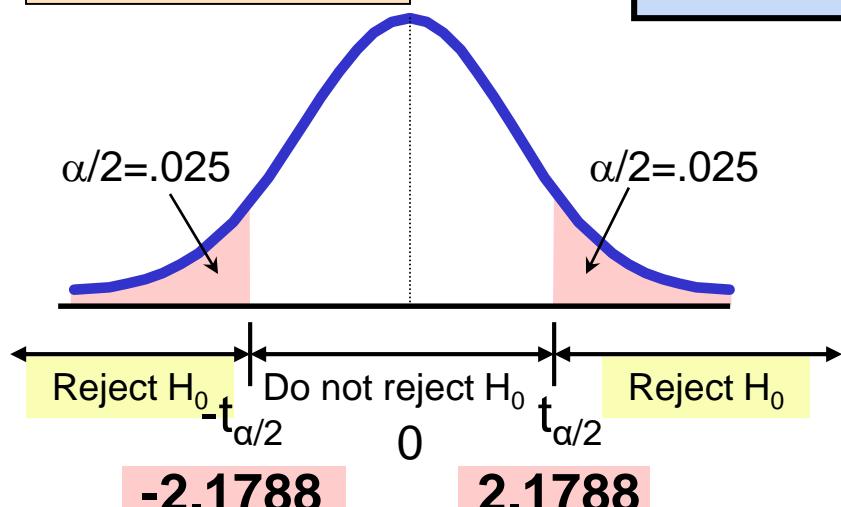
DCOVA

$$H_0: \beta_j = 0$$
$$H_1: \beta_j \neq 0$$

$$d.f. = 15-2-1 = 12$$

$$\alpha = .05$$

$$t_{\alpha/2} = 2.1788$$



From the Excel output:

For Price $t_{STAT} = -2.306$, with p-value .0398

For Advertising $t_{STAT} = 2.855$, with p-value .0145

The test statistic for each variable falls in the rejection region (p-values < .05)

Decision:

Reject H_0 for each variable

Conclusion:

There is evidence that both Price and Advertising affect pie sales at $\alpha = .05$

Confidence Interval Estimate for the Slope

DCOVA

Confidence interval for the population slope β_j

$$b_j \pm t_{\alpha/2} S_{b_j}$$

where t has
(n - k - 1) d.f.

| | Coefficients | Standard Error |
|-------------|--------------|----------------|
| Intercept | 306.52619 | 114.25389 |
| Price | -24.97509 | 10.83213 |
| Advertising | 74.13096 | 25.96732 |

Here, t has
(15 - 2 - 1) = 12 d.f.

Example: Form a 95% confidence interval for the effect of changes in price (X_1) on pie sales:

$$-24.975 \pm (2.1788)(10.832)$$

So the interval is (-48.576 , -1.374)

(This interval does not contain zero, so price has a significant effect on sales)

Confidence Interval Estimate for the Slope

DCOVA
(continued)

Confidence interval for the population slope β_j

| | Coefficients | Standard Error | ... | Lower 95% | Upper 95% |
|-------------|--------------|----------------|-----|-----------|-----------|
| Intercept | 306.52619 | 114.25389 | ... | 57.58835 | 555.46404 |
| Price | -24.97509 | 10.83213 | ... | -48.57626 | -1.37392 |
| Advertising | 74.13096 | 25.96732 | ... | 17.55303 | 130.70888 |



Example: Excel output also reports these interval endpoints:

Weekly sales are estimated to be reduced by between 1.37 to 48.58 pies for each increase of \$1 in the selling price, holding the effect of advertising constant

Testing Portions of the Multiple Regression Model

DCOVA

- Contribution of a Single Independent Variable X_j

$SSR(X_j | \text{all variables except } X_j)$

$$= SSR(\text{all variables}) - SSR(\text{all variables except } X_j)$$

- Measures the contribution of X_j in explaining the total variation in Y (SST)

Testing Portions of the Multiple Regression Model

(continued)

DCOV_A

Contribution of a Single Independent Variable X_j ,
assuming all other variables are already included
(consider here a 2-variable model):

$$\begin{aligned} \text{SSR}(X_1 | X_2) \\ = \text{SSR} (\text{all variables}) - \text{SSR}(X_2) \end{aligned}$$

From ANOVA section of regression for

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

From ANOVA section of regression for

$$\hat{Y} = b_0 + b_2 X_2$$

Measures the contribution of X_1 in explaining SST

Chapter Summary

In this chapter we discussed

- The multiple regression model
- Testing the significance of the multiple regression model
- Adjusted r^2
- Using residual plots to check model assumptions
- Testing individual regression coefficients
- Testing portions of the regression model