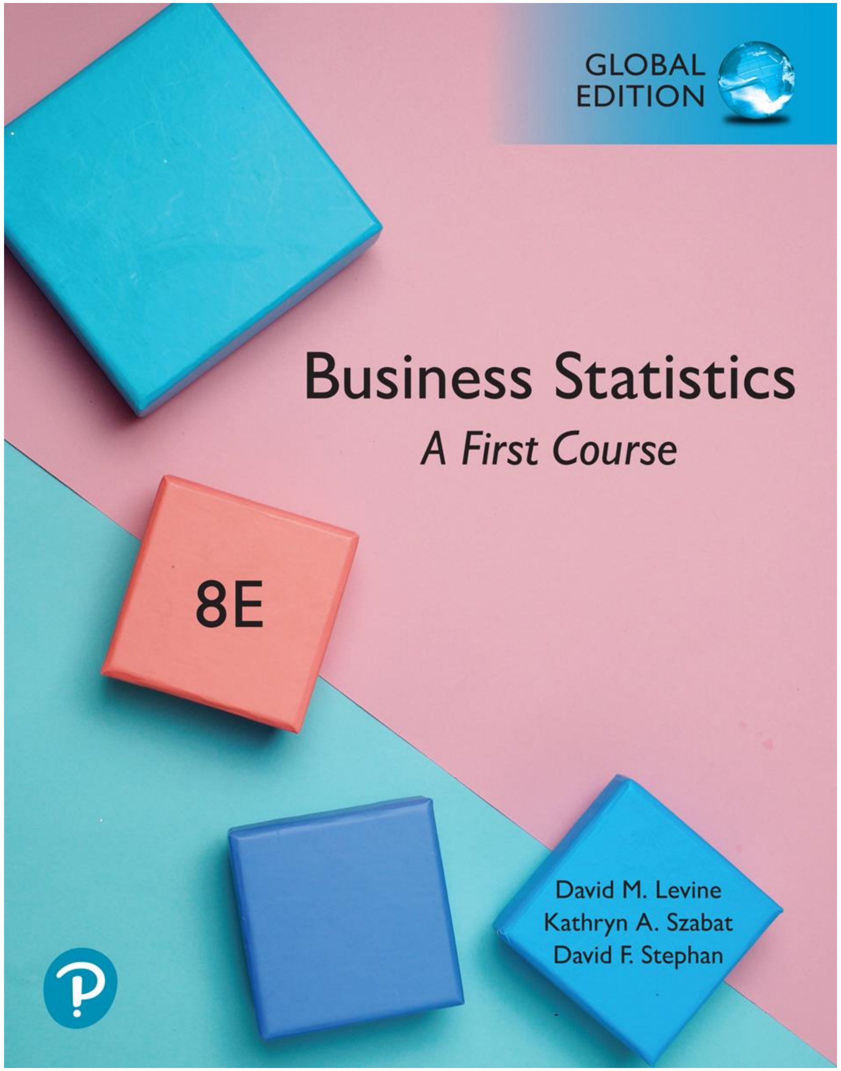


Chapter 1

Defining and Collecting Data

Objectives

In this chapter you learn:

- To understand issues that arise when defining variables.
- How to define variables.
- To understand the different measurement scales.
- How to collect data.
- To identify different ways to collect a sample.
- To understand the types of survey errors.

Classifying Variables By Type

DCOVA

- **Categorical** (***qualitative***) variables take categories as their values such as "yes", "no", or "blue", "brown", "green".

- **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***.

Examples of Types of Variables

DCOVA

Question	Responses	Variable Type
Do you have a Facebook profile?	Yes or No	Categorical
How many text messages have you sent in the past three days?	---------------	Numerical (discrete)
How long did the mobile app update take to download?	---------------	Numerical (continuous)

Measurement Scales

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

Measurement Scales (con't.)

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, DD, D
Student Grades	A, B, C, D, F

Measurement Scales (con't.)

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.

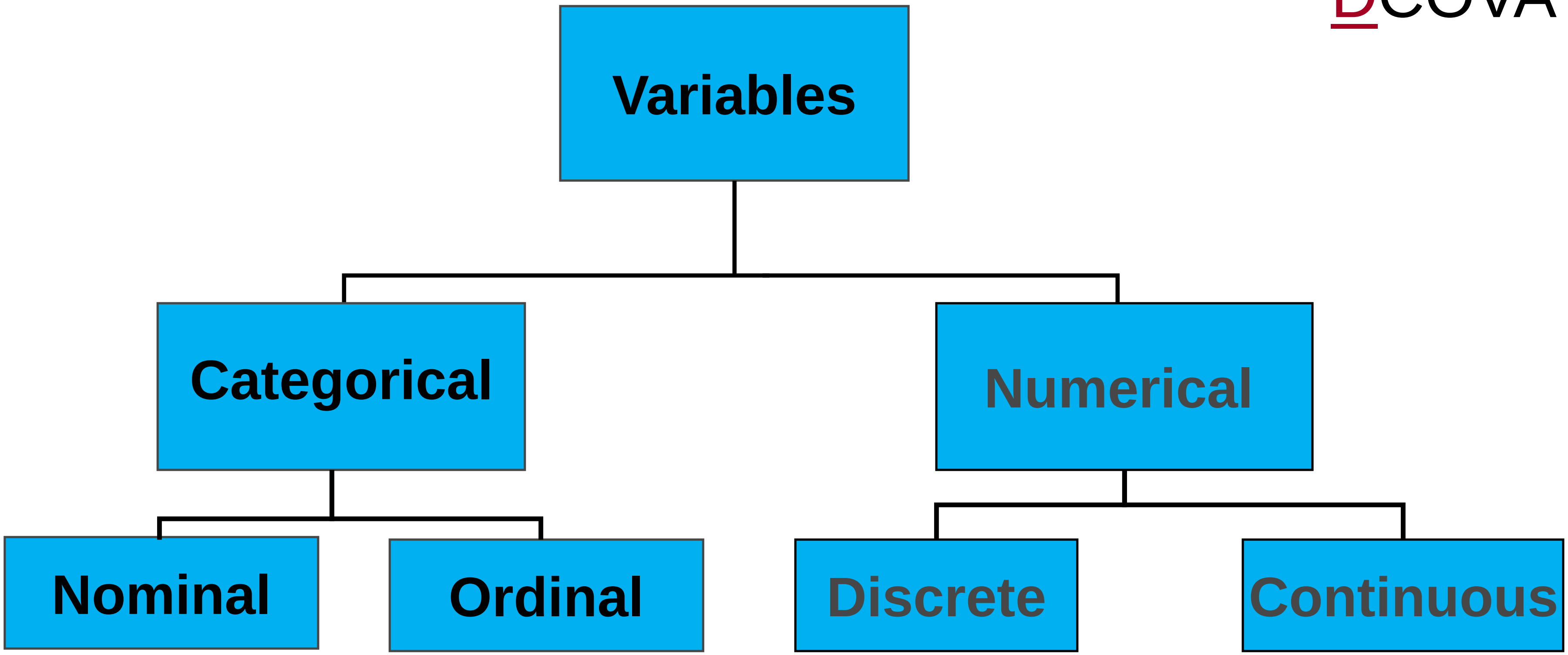
Interval and Ratio Scales

DCOVA

Numerical Variable	*Level of Measurement*
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

Types of Variables

DCOVA



Examples:

- Marital Status
- Political Party
- Eye Color

(Defined Categories)

Examples: Ratings

- Good, Better, Best
- Low, Med, High

(Ordered Categories)

Examples:

- Number of Children
- Defects per hour

(Counted items)

Examples:

- Weight
- Voltage

(Measured characteristics)

Data Is Collected From Either A Population or A Sample

DCOVA

POPULATION

A **population** contains all of the items or individuals of interest that you seek to study.

SAMPLE

A **sample** contains only a portion of a population of interest.

Population vs. Sample

DCOVA

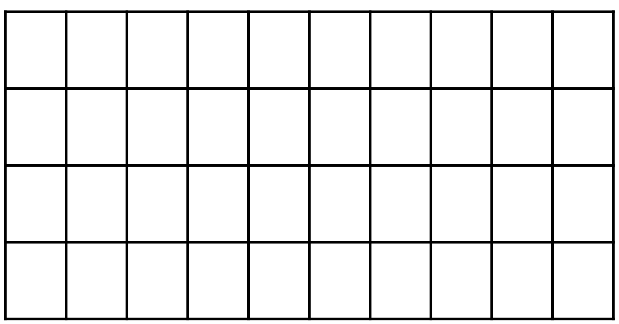
Population

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

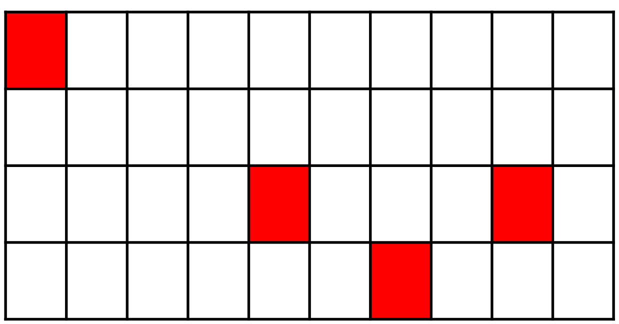
Sample

A portion of the population of items or individuals.

A Population of Size 40



A Sample of Size 4



Collecting Data Via Sampling Is Used When Doing So 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.

Parameter or Statistic?

DCOVA

- A **population parameter** summarizes the value of a specific variable for a population.

- A **sample statistic** summarizes the value of a specific variable for sample data.

Sources Of Data Arise From The Following Activities

DCOVA

- Capturing data generated by ongoing business activities.
- Distributing data compiled by an organization or individual.
- Compiling the responses from a survey.
- Conducting a designed experiment and recording the outcomes.
- Conducting an observational study and recording the results.

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.

Examples Of Data Distributed By An Organization or Individual

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 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.

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 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.

Observational Studies & Designed Experiments Have A Common Objective

DCOVA

- Both are attempting to quantify the effect that a process change (called a **treatment**) has on a variable of interest.
- In an observational study, there is no direct control over which items receive the treatment.
- In a designed experiment, there is direct control over which items receive the treatment.

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.

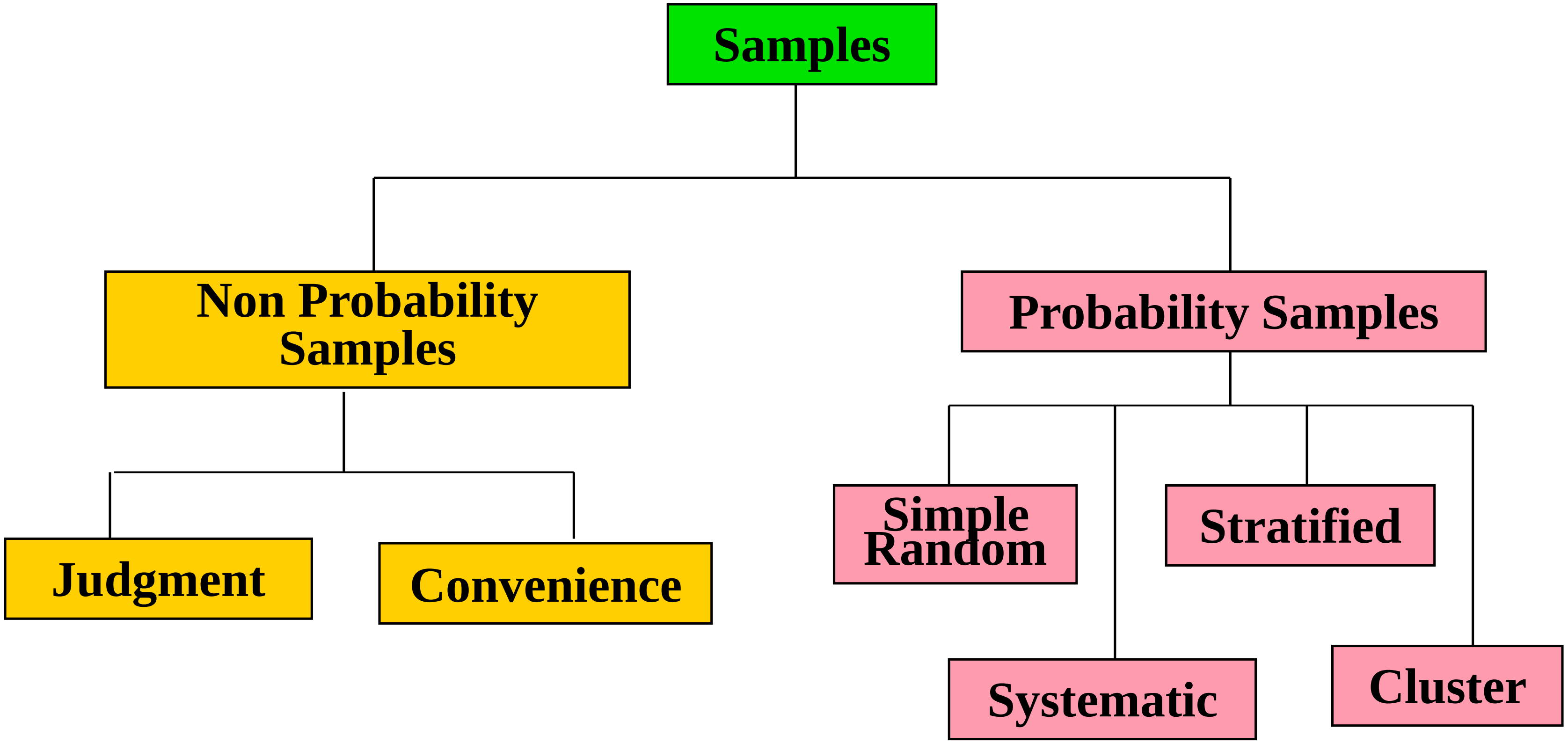
A Sampling Process Begins With A Sampling Frame

DCOVA

- The sampling frame is a listing of items that make up the population.
- Frames are data sources such as population lists, directories, or maps.
- Inaccurate or biased results can result if a frame excludes certain groups or portions of the population.
- Using different frames to generate data can lead to dissimilar conclusions.

Types of Samples

DCOVA



Types of Samples: Nonprobability Sample

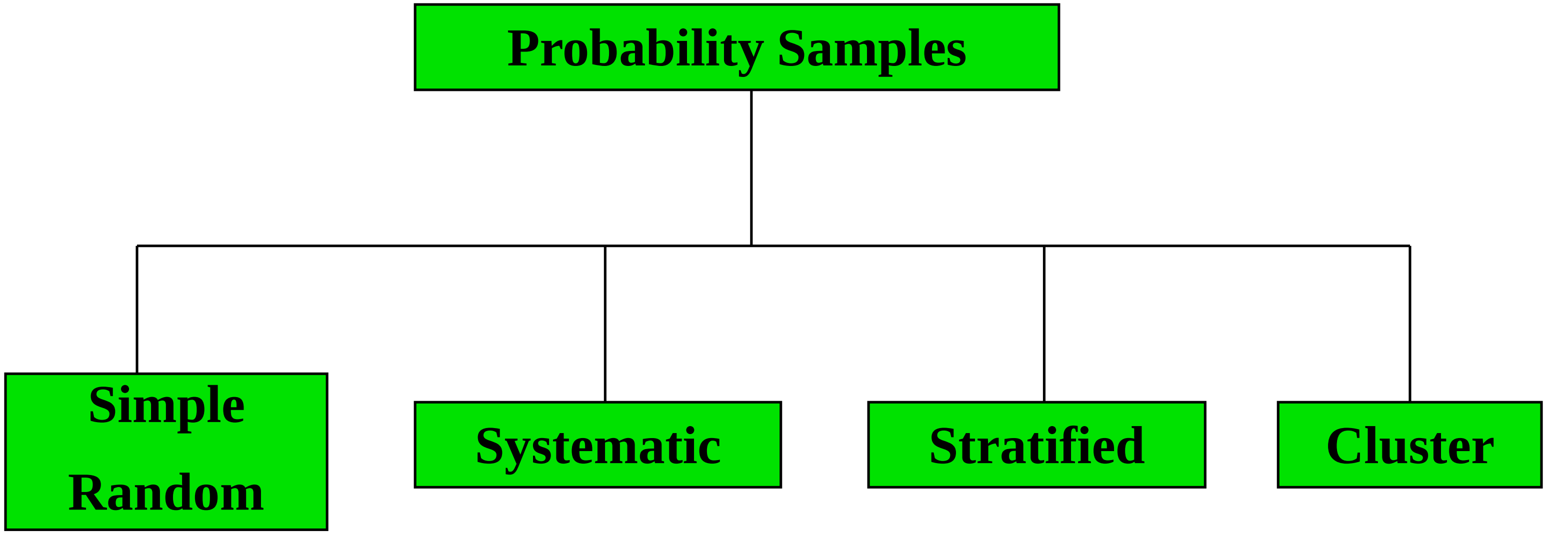
DCOVA

- In a nonprobability sample, items included are chosen without regard to their probability of occurrence.
  - In **convenience sampling**, items are selected based only on the fact that they are easy, inexpensive, or convenient to sample.
  - In a **judgment sample,** you get the opinions of pre-selected experts on the subject matter.

Types of Samples: Probability Sample

DCOVA

- In a **probability sample**, items in the sample are chosen on the basis of known probabilities.



Probability Sample: Simple Random Sample

DCOVA

- Every individual or item from the frame has an equal chance of being selected.
- Selection may be with replacement (selected individual is returned to frame for possible reselection) or without replacement (selected individual isn't returned to the frame).
- Samples obtained from table of random numbers or computer random number generators.

Selecting a Simple Random Sample Using A Random Number Table

DCOVA

Sampling Frame For Population With 850 Items

Item Name	Item #
Bev R.	001
Ulan X.	002
.	.
.	.
.	.
.	.
Joann P.	849
Paul F.	850

Portion Of A Random Number Table

49280	88924	35779	00283	81163	07275
11100	02340	12860	74697	96644	89439
09893	23997	20048	49420	88872	08401

The First 5 Items in a simple random sample

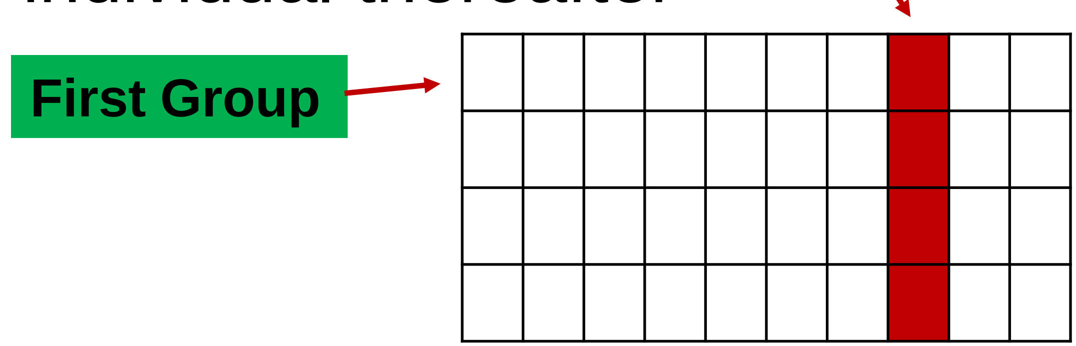
Item # 492
Item # 808
Item # 892 -- does not exist so ignore
Item # 435
Item # 779
Item # 002

Probability Sample: Systematic Sample

DCOVA

- Decide on sample size: n
- Divide frame of N individuals into groups of k individuals: $k=N/n$
- Randomly select one individual from the 1st group
- Select every k^{th} individual thereafter

$N = 40$
$n = 4$
$k = 10$



Probability Sample: Stratified Sample

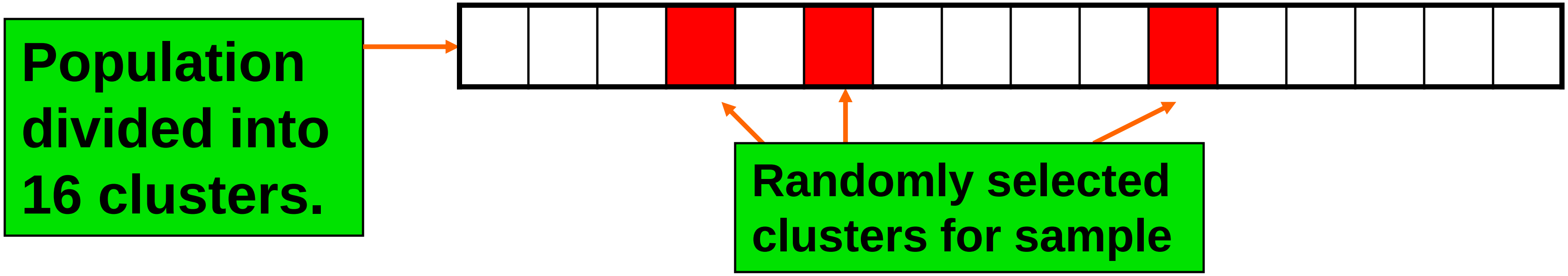
DCOVA

- Divide population into two or more subgroups (called *strata*) according to some common characteristic.
- A simple random sample is selected from each subgroup, with sample sizes proportional to strata sizes.
- Samples from subgroups are combined into one.
- This is a common technique when sampling population of voters, stratifying across racial or socio-economic lines.

Probability Sample
Cluster Sample

DCOVA

- Population is divided into several “clusters,” each representative of the population.
- A simple random sample of clusters is selected.
- All items in the selected clusters can be used, or items can be chosen from a cluster using another probability sampling technique.
- A common application of cluster sampling involves election exit polls, where certain election districts are selected and sampled.



Probability Sample: Comparing Sampling Methods

DCOVA

- Simple random sample and Systematic sample:
  - Simple to use.
  - May not be a good representation of the population’s underlying characteristics.
- Stratified sample:
  - Ensures representation of individuals across the entire population.
- Cluster sample:
  - More cost effective.
  - Less efficient (need larger sample to acquire the same level of precision).

Types of Survey Errors

DCOVA

- Coverage error or selection bias:
  - Exists if some groups are excluded from the frame and have no chance of being selected.
- Nonresponse error or bias:
  - People who do not respond may be different from those who do respond.
- Sampling error:
  - Variation from sample to sample will always exist.
- Measurement error:
  - Due to weaknesses in question design and / or respondent error.

Types of Survey Errors *(continued)*

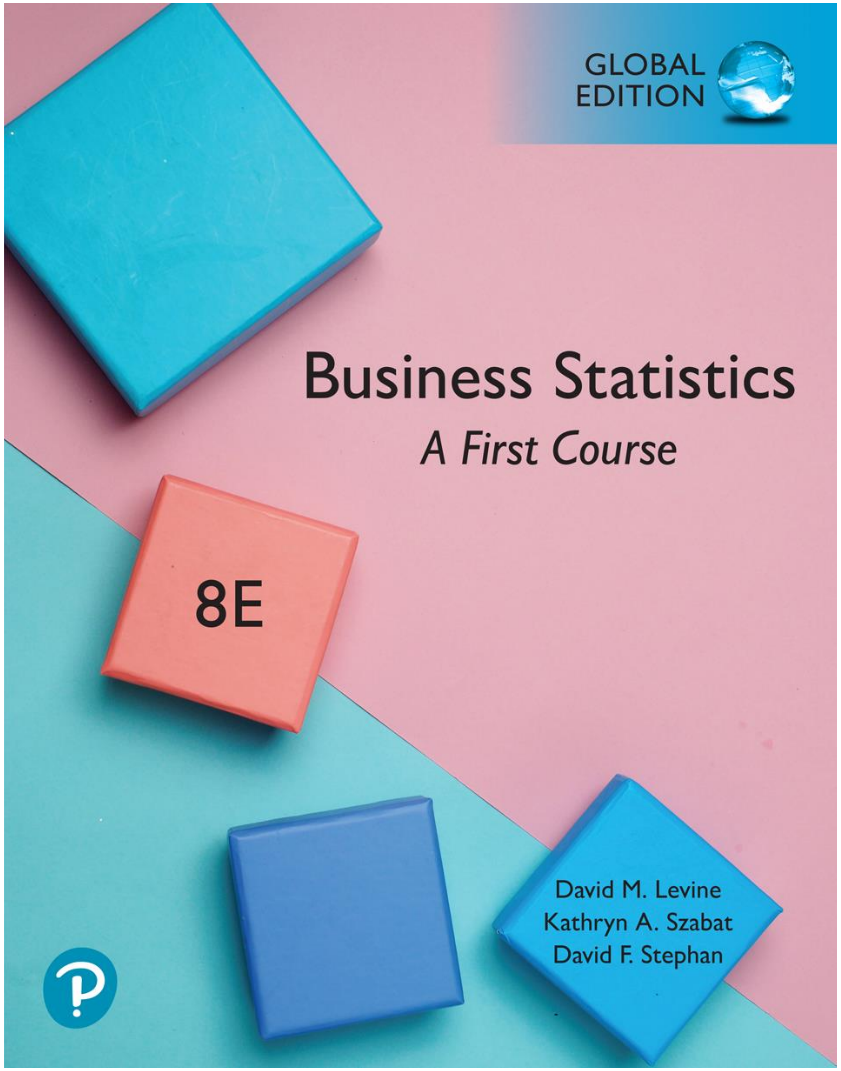
DCOVA

- 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**

Chapter Summary

In this chapter we have discussed:

- Understanding issues that arise when defining variables.
- How to define variables.
- Understanding the different measurement scales.
- How to collect data.
- Identifying different ways to collect a sample.
- Understanding the types of survey errors.



Chapter 2

Organizing and Visualizing Variables

Objectives

In this chapter you learn:

- How to organize and visualize categorical variables.
- How to organize and visualize numerical variables.
- How to visualizing Two Numerical Variables.

Organizing Data Creates Both Tabular And Visual Summaries

DCOVA

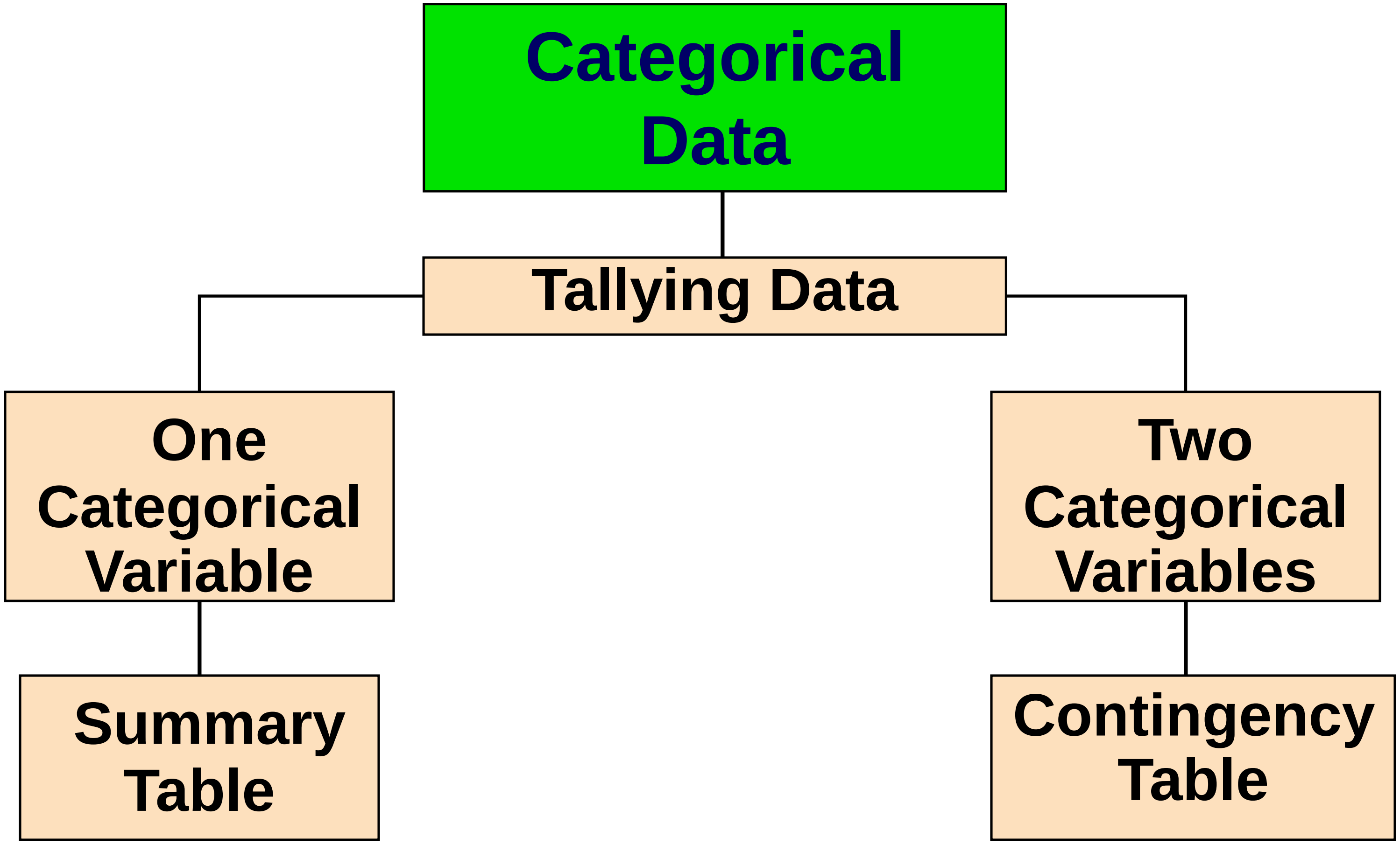
- Summaries both guide further exploration and sometimes facilitate decision making.

- Visual summaries enable rapid review of larger amounts of data & show possible significant patterns.

- Often, the **O**rganize and **V**isualize step in **DCOVA** occur concurrently.

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.

Devices Millennials Use to Watch Movies or Television Shows

Devices Used To Watch Movies or TV Shows	**Percent**
Television Set	49%
Tablet	9%
Smartphone	10%
Laptop / Desktop	32%

Source: Data extracted and adapted from A. Sharma, "Big Media Needs to Embrace Digital Shift Not Fight It," Wall Street Journal, June 22, 2016, p. 1-2.

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

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%

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

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

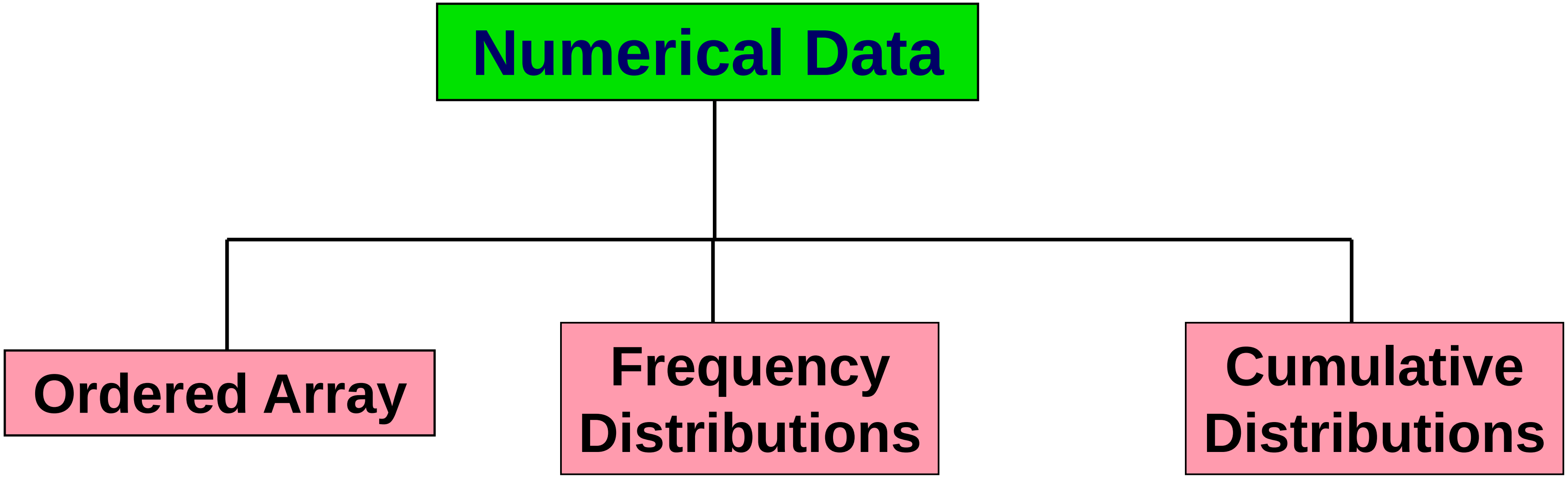
50.75% = 170 / 335
30.77% = 20 / 65

	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



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 in degrees Fahrenheit.

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 but less than 20.**
  - **Class 2: 20 but less than 30.**
  - **Class 3: 30 but less than 40.**
  - **Class 4: 40 but less than 50.**
  - **Class 5: 50 but 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

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

Relative Frequency = Frequency / Total, e.g. 0.10 = 2 / 20

Organizing Numerical Data: Cumulative Frequency Distribution Example

DCOVA

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%

Cumulative Percentage = Cumulative Frequency / Total * 100 e.g. 45% = 100*9/20

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).

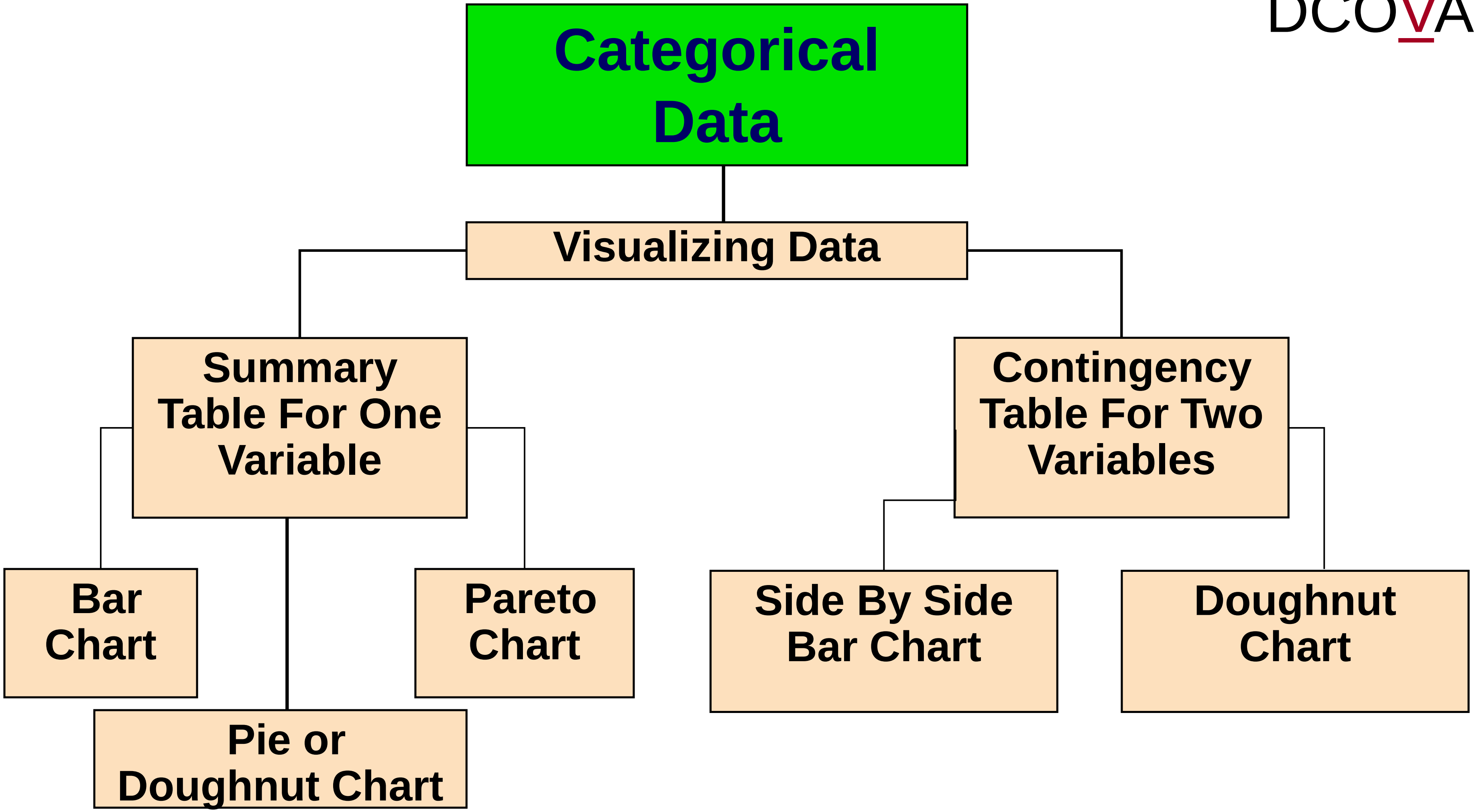
- Shifts in data concentration may show up when different class boundaries are chosen.

- 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

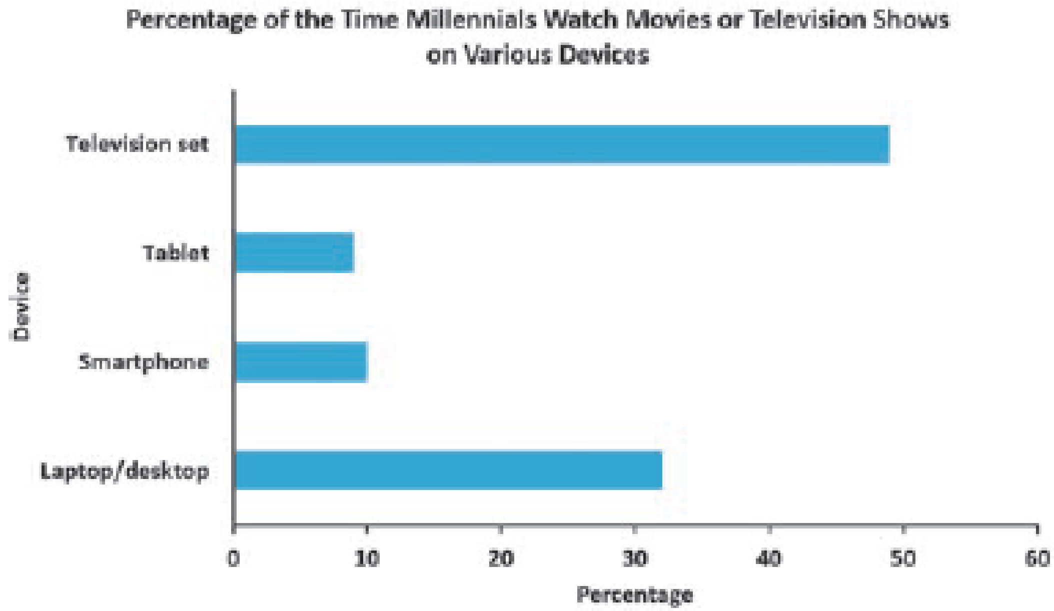


Visualizing Categorical Data: The Bar Chart

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.

Devices Used to Watch	Percent
Television Set	49%
Tablet	9%
Smartphone	10%
Laptop / Desktop	32%

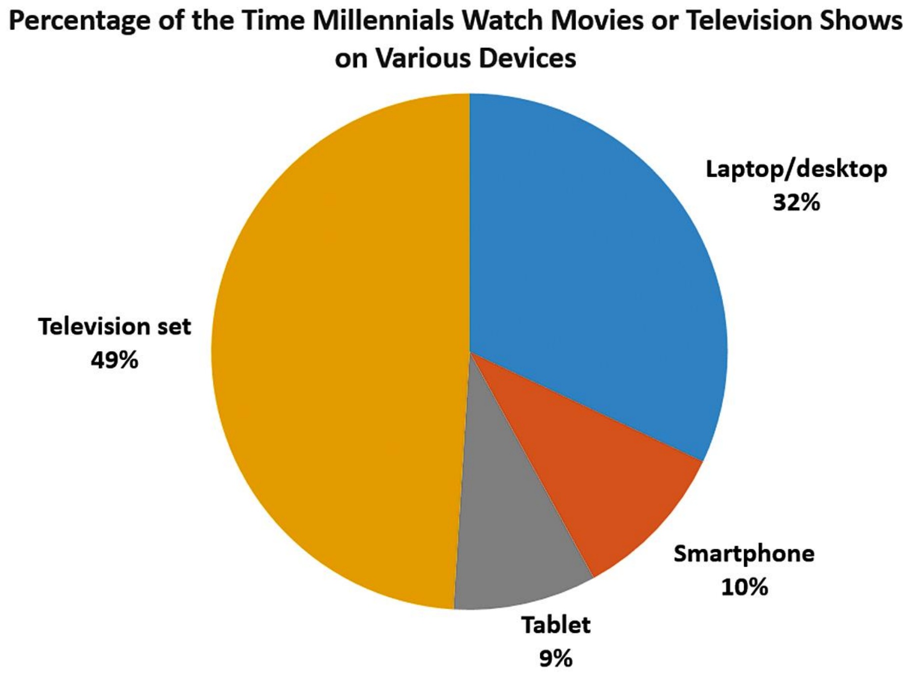


Visualizing Categorical Data: The Pie Chart

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.

Devices Used to Watch	Percent
Television Set	49%
Tablet	9%
Smartphone	10%
Laptop / Desktop	32%

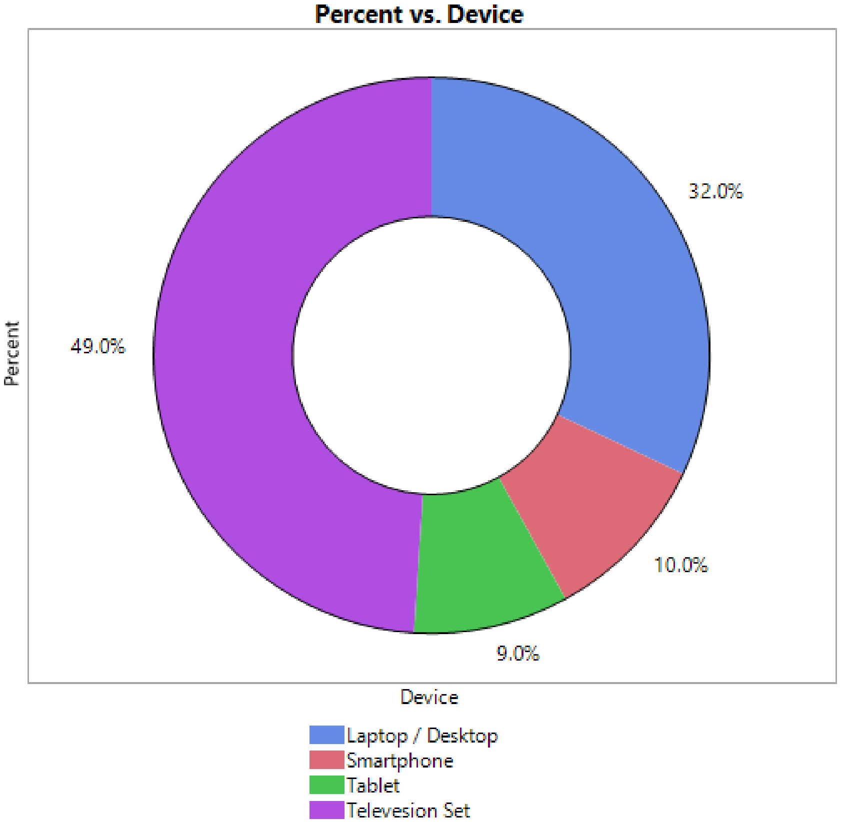


Visualizing Categorical Data: The Doughnut Chart

DCOVA

- The **doughnut chart** is the outer part of a circle broken up into pieces that represent categories. The size of each piece of the doughnut varies according to the percentage in each category.

Devices Used to Watch	Percent
Television Set	49%
Tablet	9%
Smartphone	10%
Laptop / Desktop	32%



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."

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

DCOVA

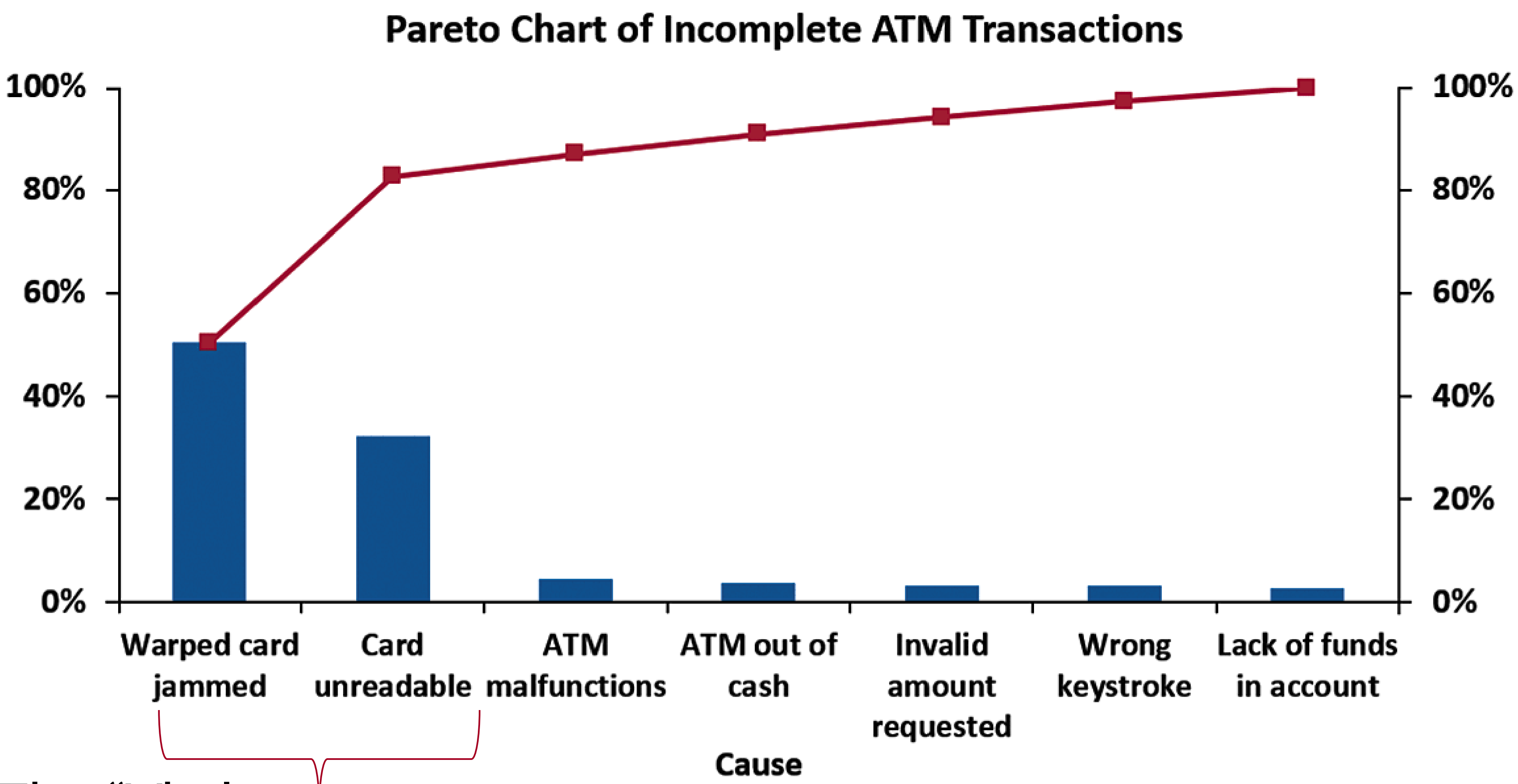
Ordered Summary Table For Causes Of Incomplete ATM Transactions

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



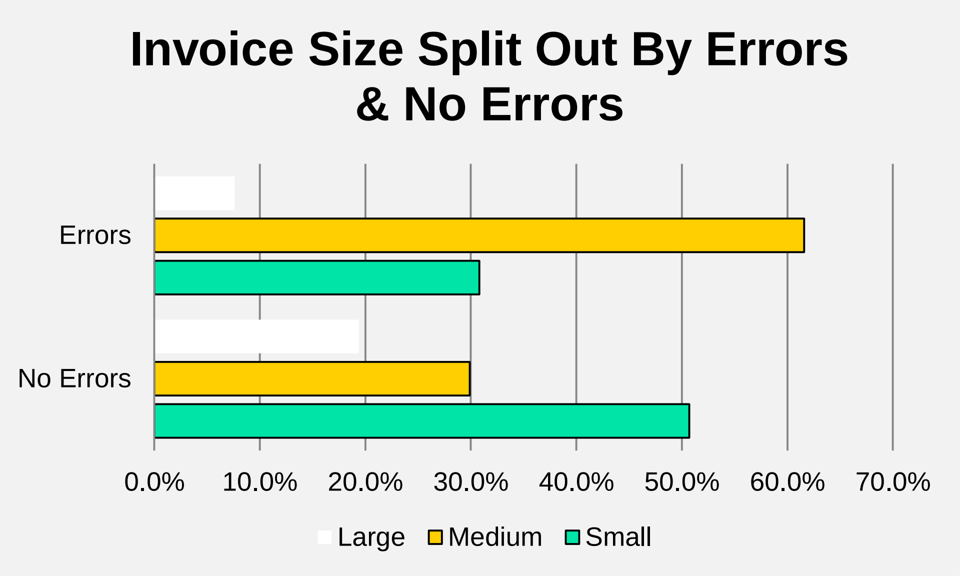
The "Vital Few"

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%



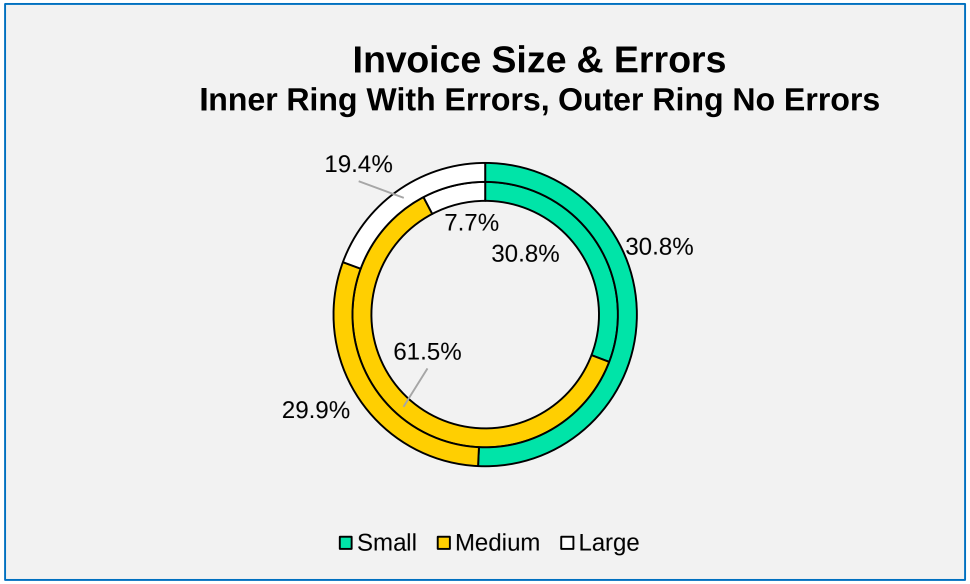
Invoices with errors are much more likely to be of medium size (61.5% vs 30.8% & 7.7%).

Visualizing Categorical Data: Doughnut Charts

DCOVA

- A **Doughnut Chart** can be used to represent the data from a contingency table.

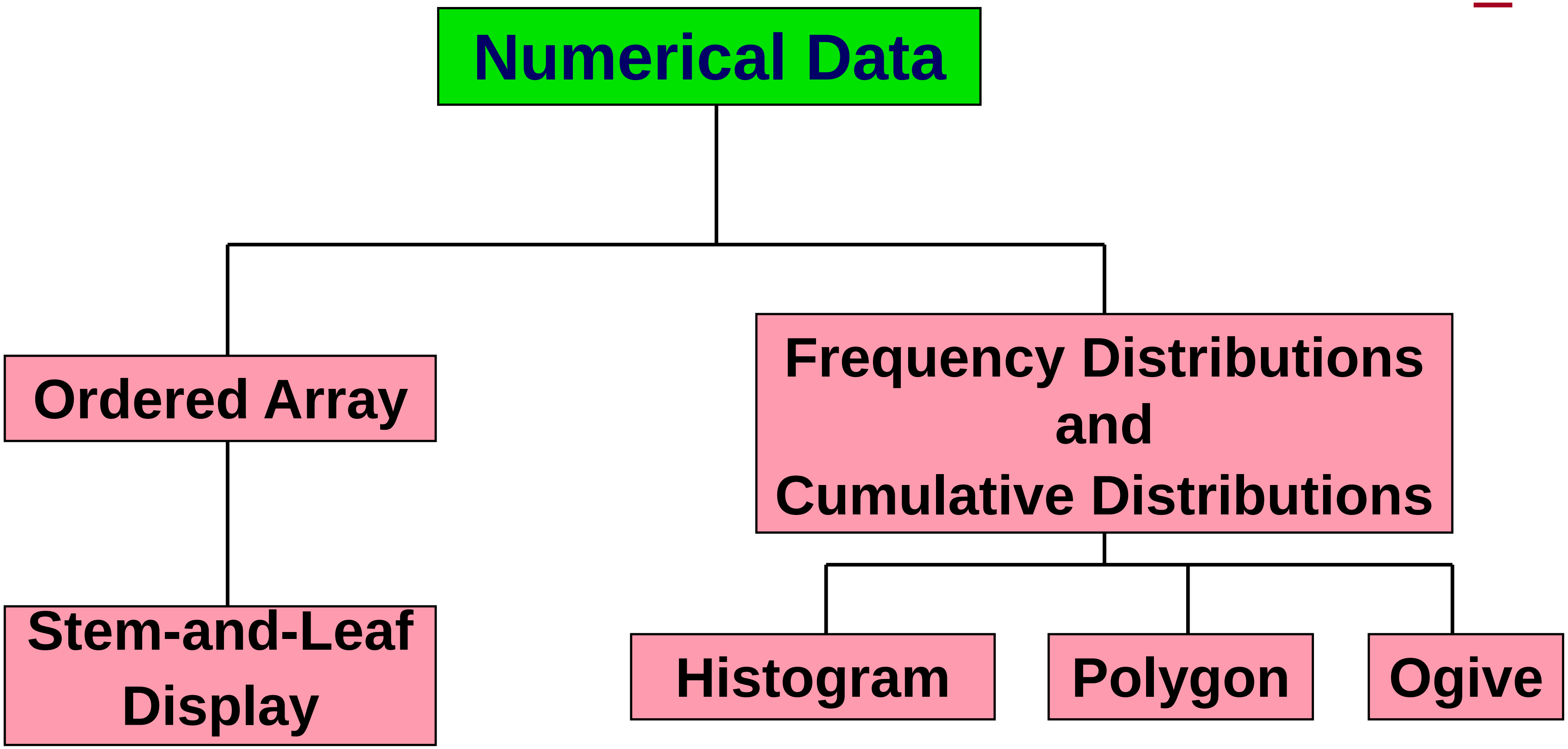
	No Errors	Errors	Total
Small Amount	50.75%	30.77%	47.50%
Medium Amount	29.85%	61.54%	35.00%
Large Amount	19.40%	7.69%	17.50%
Total	100.0%	100.0%	100.0%



Invoices with errors are much more likely to be of medium size (61.5% vs 30.8% & 7.7%).

Visualizing Numerical Data By Using Graphical Displays

DCOVA



Stem-and-Leaf Display

DCOVA

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

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

Organizing Numerical Data: Stem and Leaf Display

DCOVA

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

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

Age of College Students

Day Students

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

Night Students

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

Visualizing Numerical Data: The Histogram

DCOVA

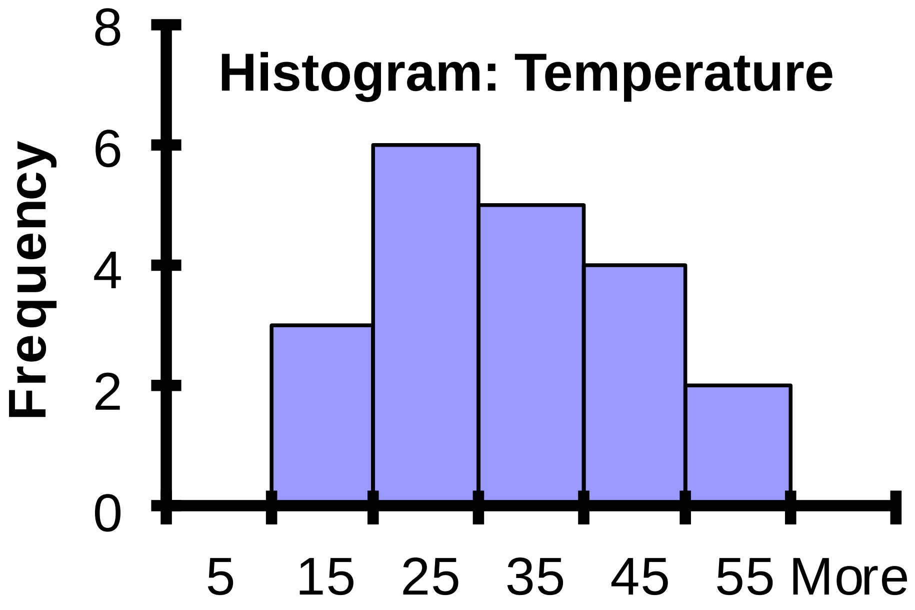
- A vertical bar chart of the data in a frequency distribution is called a **histogram.**
- In a histogram there are no gaps between adjacent bars.
- The **class boundaries** (or **class midpoints**) are shown on the horizontal axis.
- The vertical axis is either **frequency, relative frequency,** or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

Visualizing Numerical Data: The Histogram

DCOVA

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

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



Visualizing Numerical Data: The Percentage Polygon

DCOVA

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

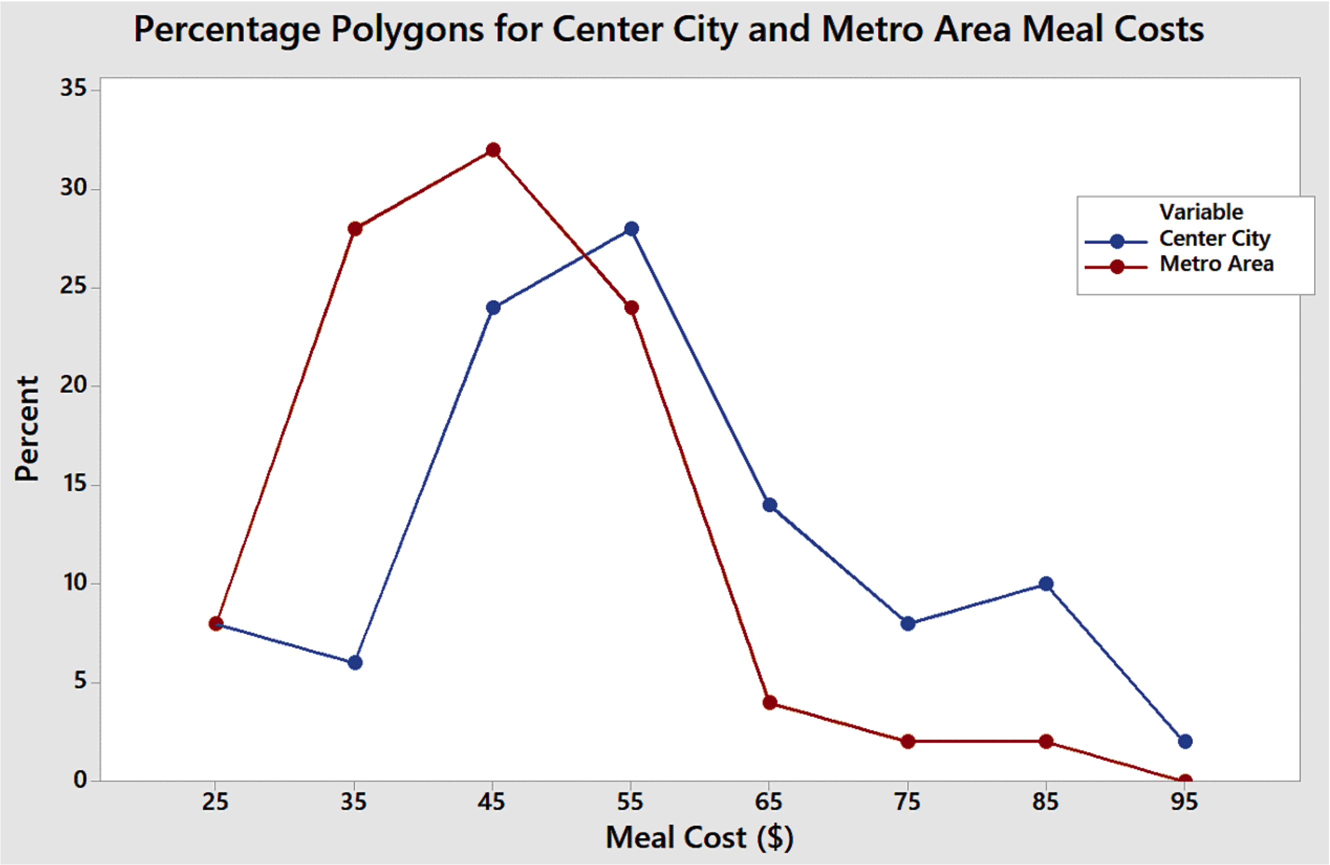
- The **cumulative percentage polygon,** or **ogive,** displays the variable of interest along the X axis, and the cumulative percentages along the Y axis.

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

Visualizing Numerical Data: The Frequency Polygon

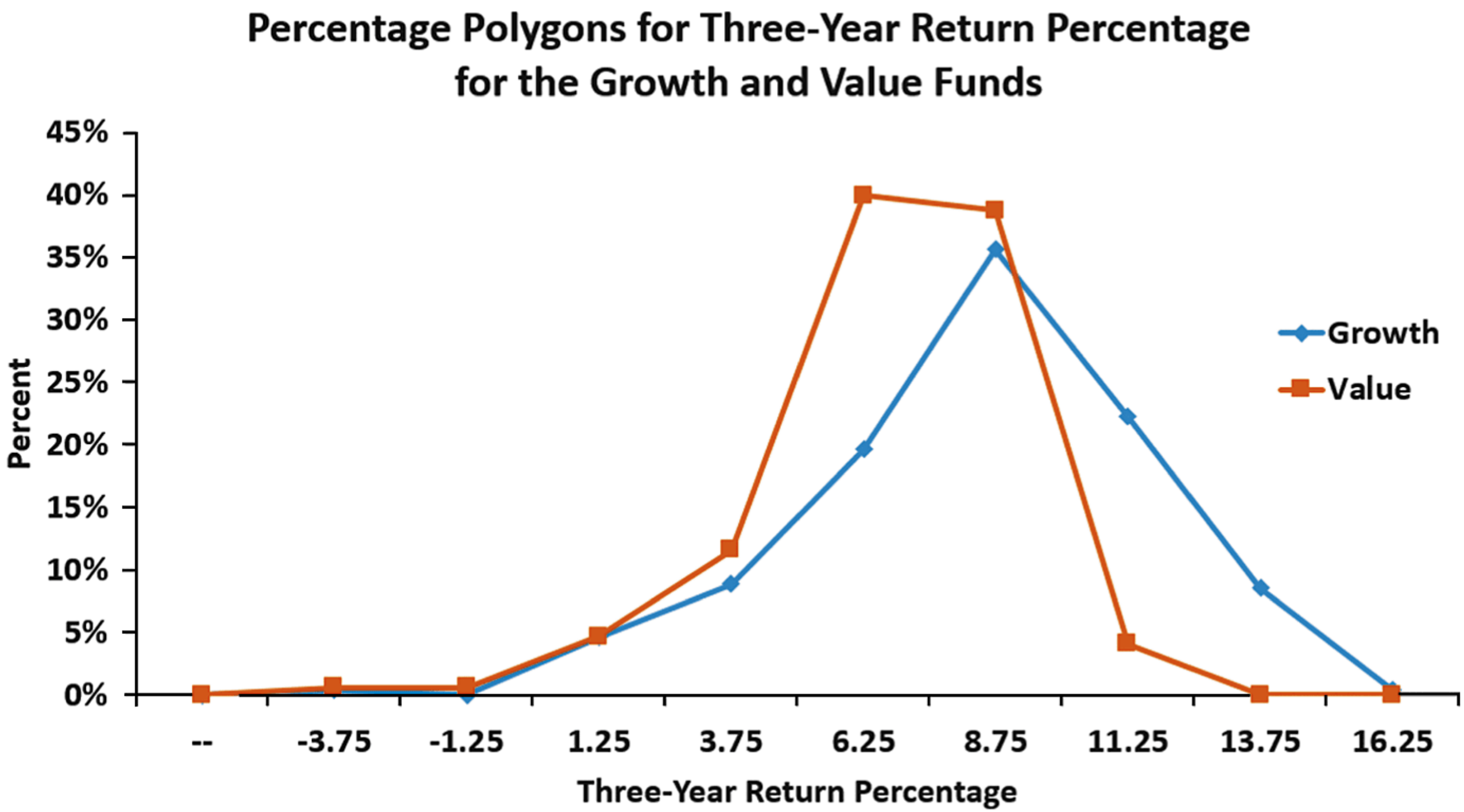
DCOVA

Useful When Comparing Two or More Groups



Visualizing Numerical Data: The Percentage Polygon

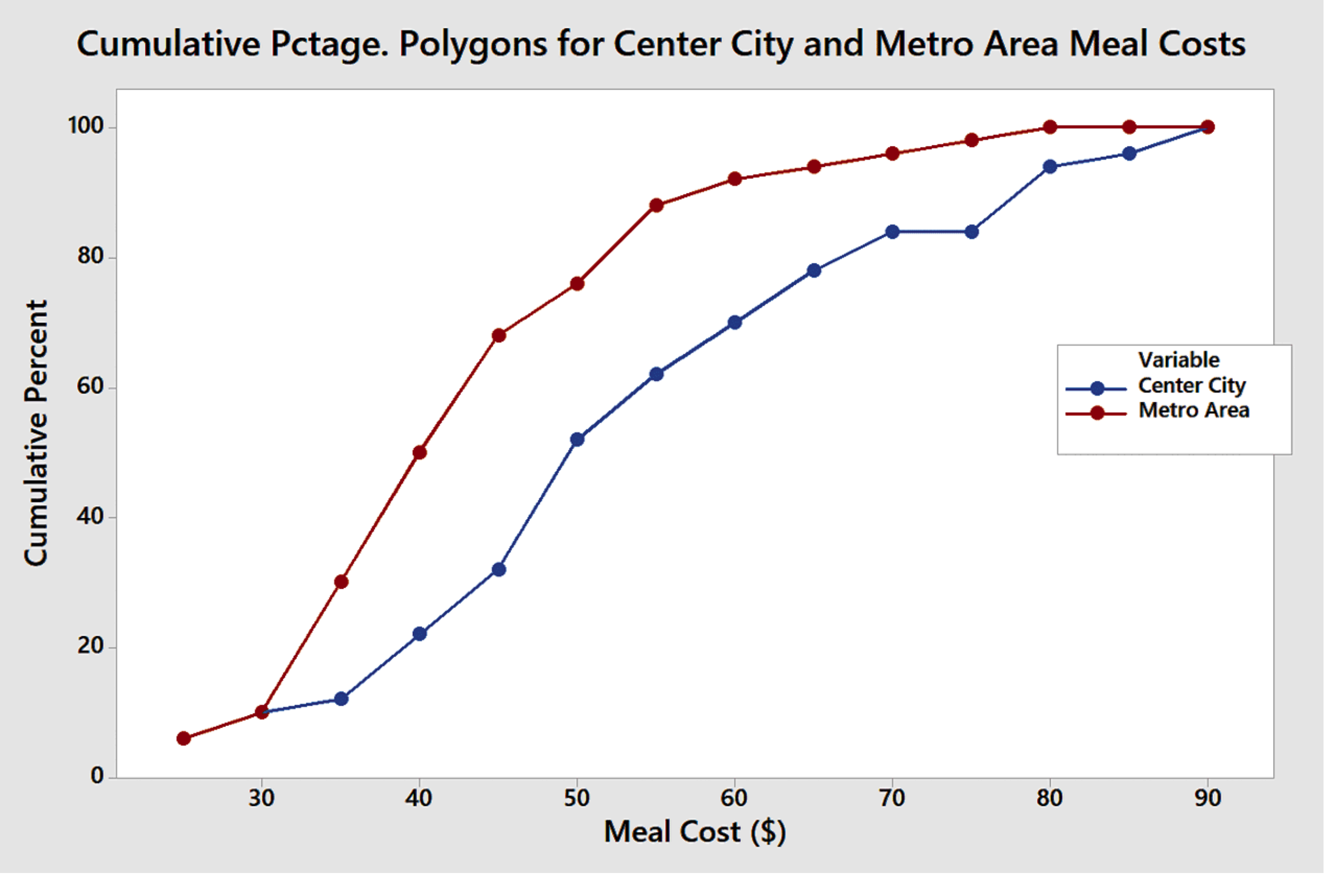
DCOVA



Visualizing Numerical Data: The Cumulative Percentage Polygon (Ogive)

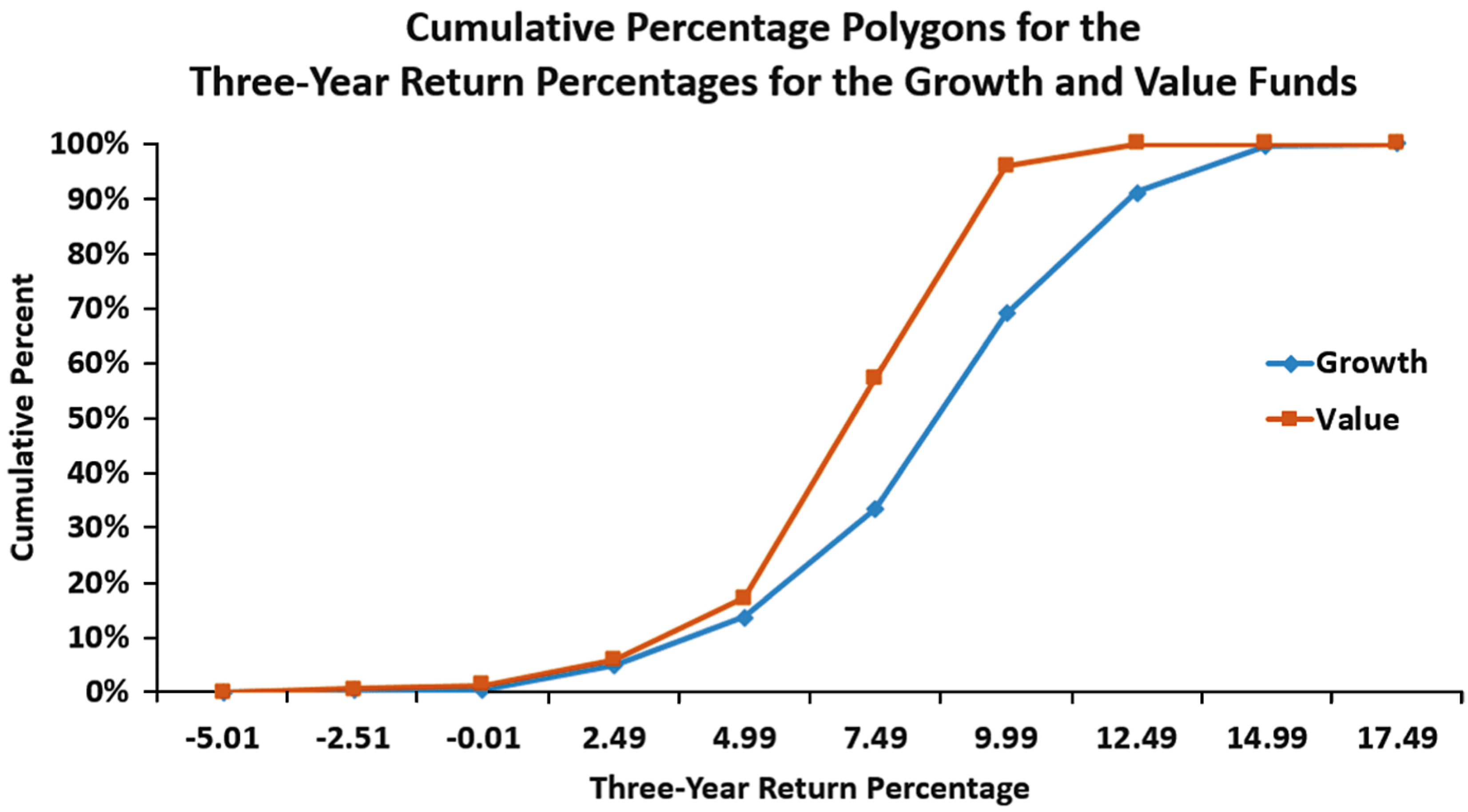
DCOVA

Useful When Comparing Two or More Groups



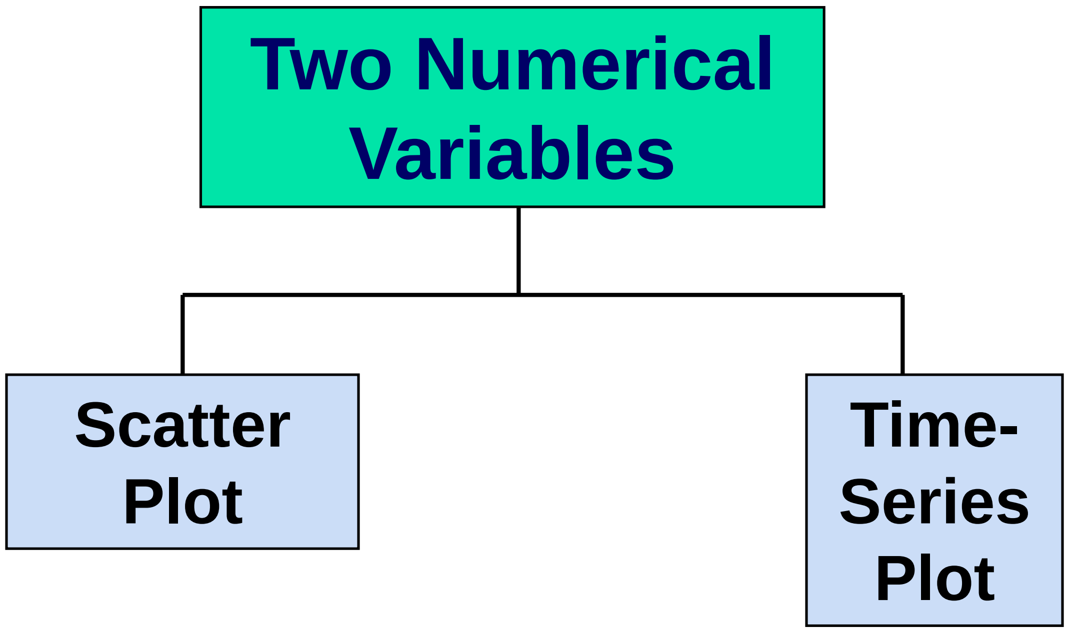
Visualizing Numerical Data: The Cumulative Percentage Polygon (Ogive)

DCOVA



Visualizing Two Numerical Variables By Using Graphical Displays

DCOVA



Visualizing Two Numerical Variables: The Scatter Plot

DCOVA

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

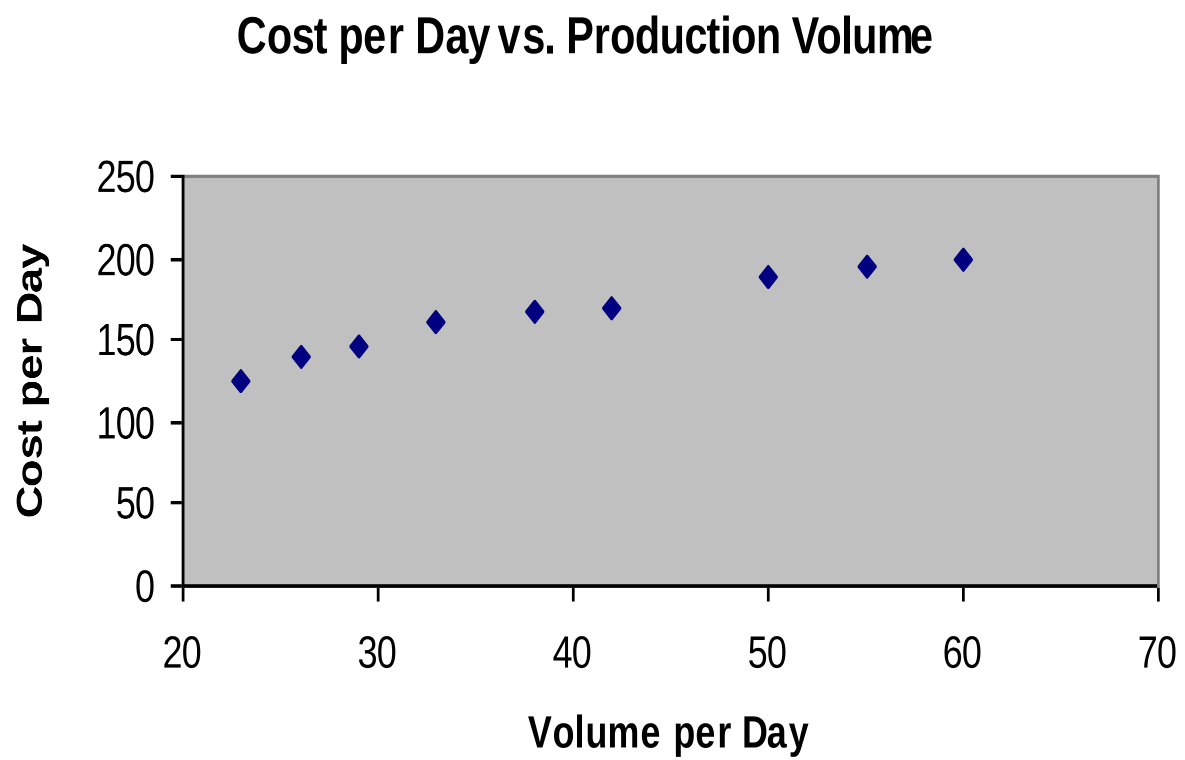
- One variable's values are displayed on the horizontal or X axis and the other variable's values are displayed on the vertical or Y axis.

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

Scatter Plot Example

DCOVA

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



Visualizing Two Numerical Variables: The Time Series Plot

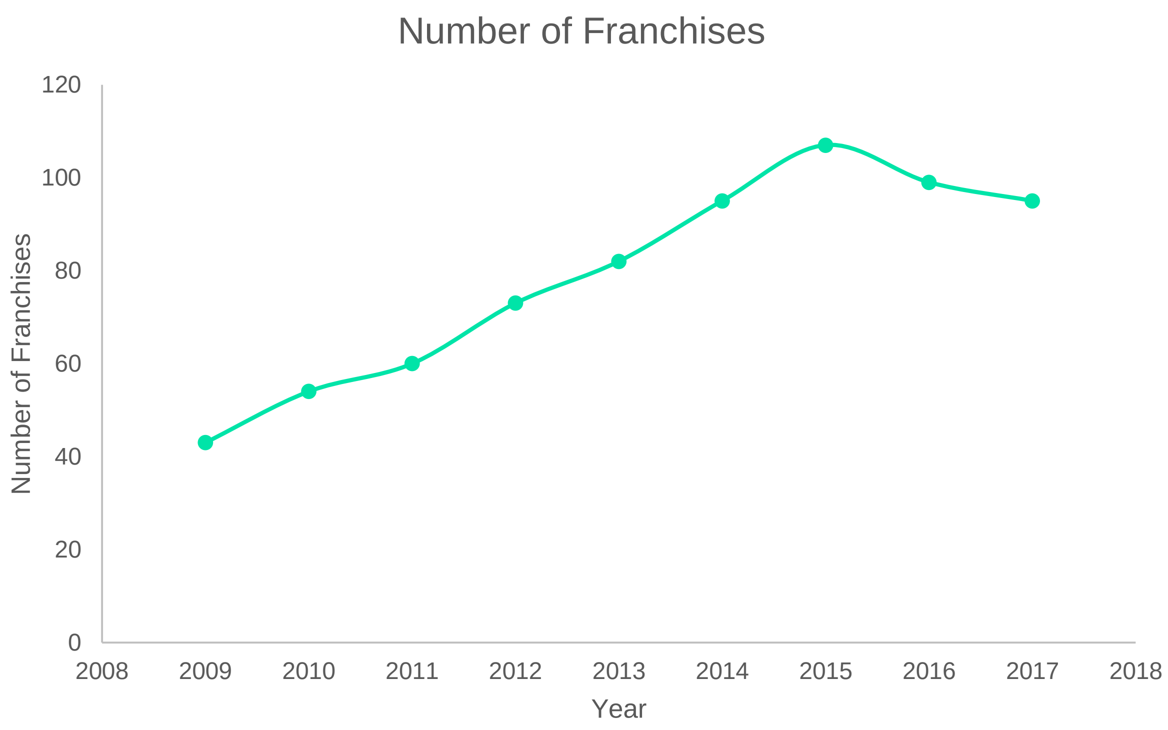
DCOVA

- A Time-Series Plot is used to study patterns in the values of a numeric variable over time.
- The Time-Series Plot:
  - Numeric variable's values are on the vertical axis and the time period is on the horizontal axis.

Time Series Plot Example

DCOVA

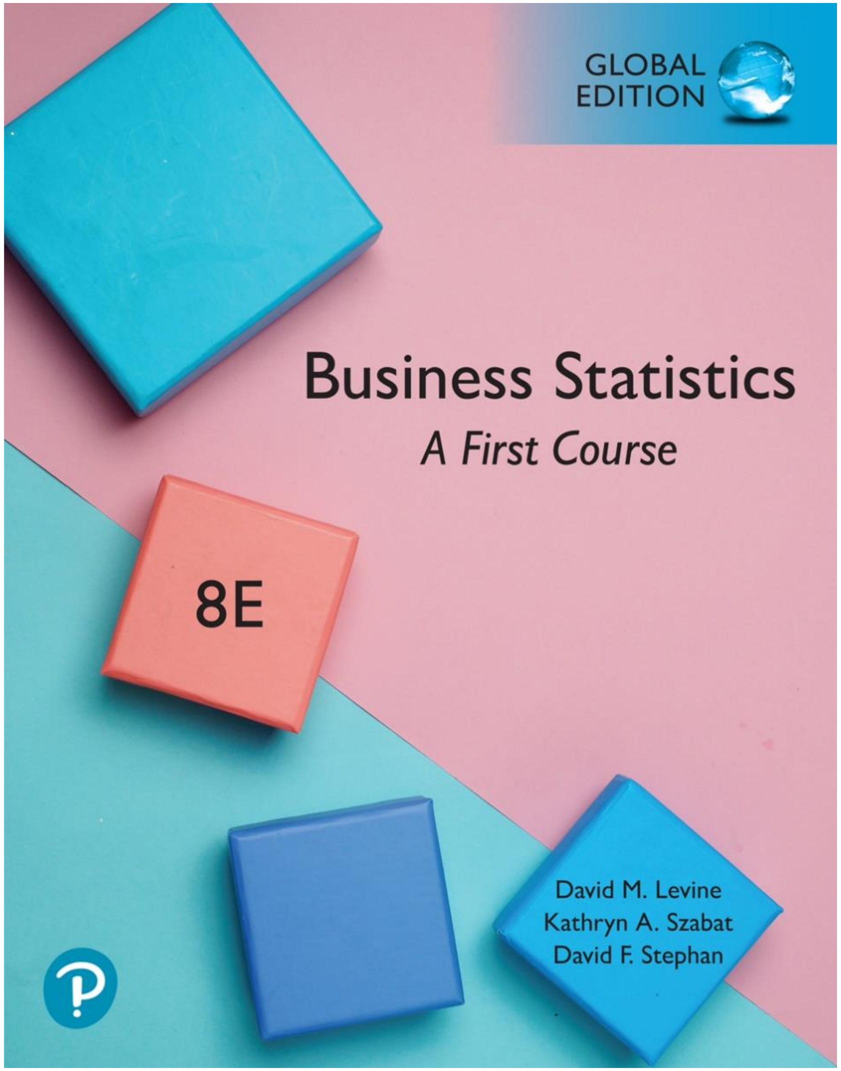
Year	Number of Franchises
2009	43
2010	54
2011	60
2012	73
2013	82
2014	95
2015	107
2016	99
2017	95



Chapter Summary

In this chapter we covered:

- Organizing and visualizing categorical variables.
- Organizing and visualizing numerical variables.
- How to visualizing Two Numerical Variables.



Chapter 3

Numerical Descriptive Measures

Objectives

In this chapter, you learn to:

- Describe the properties of central tendency, variation, and shape in numerical variables.
- Construct and interpret a boxplot.
- Compute descriptive summary measures for a population.
- Calculate the covariance and the coefficient of correlation.

Summary Definitions

DCOVA

- The **central tendency** is the extent to which the values of a numerical variable group around a typical or central value.
- The **variation** is the amount of dispersion or scattering away from a central value that the values of a numerical variable show.
- 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:

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

Pronounced X-bar

The i^{th} value

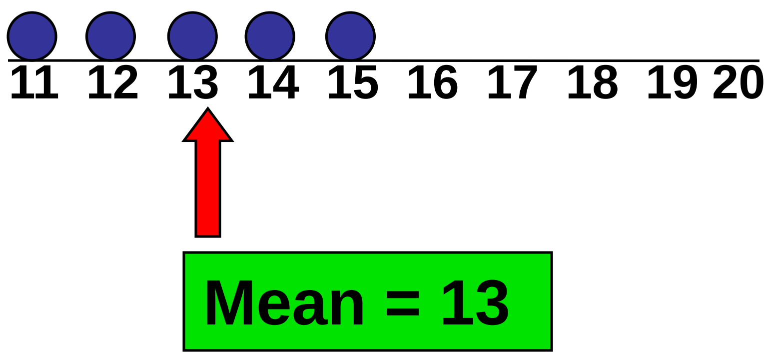
Sample size

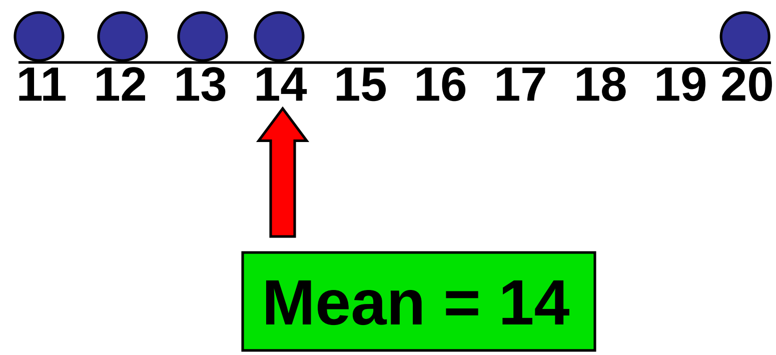
Observed values

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

DCOVA

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





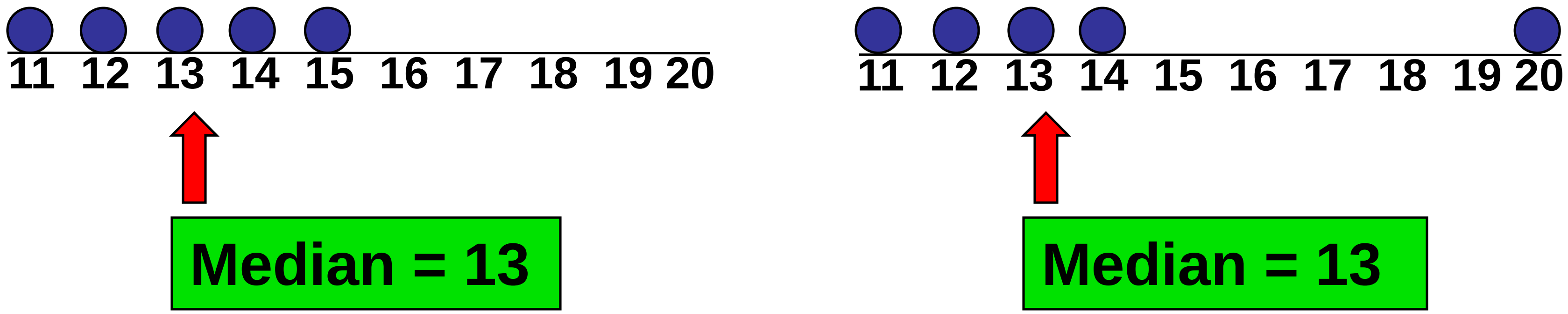
$$\frac{11+12+13+14+15}{5} = \frac{65}{5} = 13$$

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

Measures of Central Tendency: The Median

DCOVA

- 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

DCOVA

- 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.

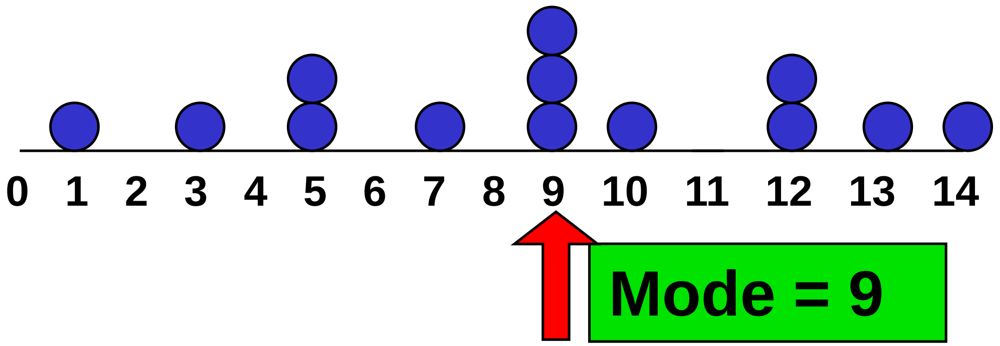
- If the number of values is even, the median is the average of the two middle numbers.

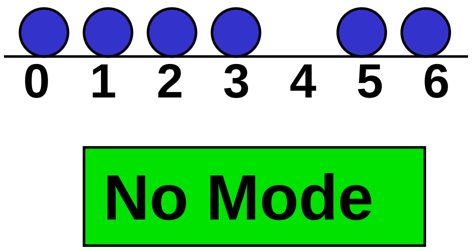
Note that $\frac{n+1}{2}$ is not the *value* of the median, only the *position* of the median in the ranked data.

Measures of Central Tendency: The Mode

DCOVA

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





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.
- In many situations it makes sense to report both the **mean** and the **median**.

Measures of Central Tendency: Summary

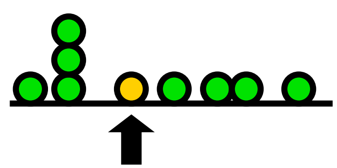
DCOVA

Central Tendency

Arithmetic Mean

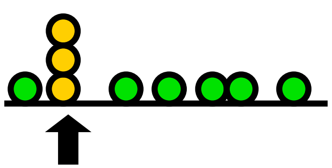
$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Median



Middle value in the ordered array

Mode



Most frequently observed value

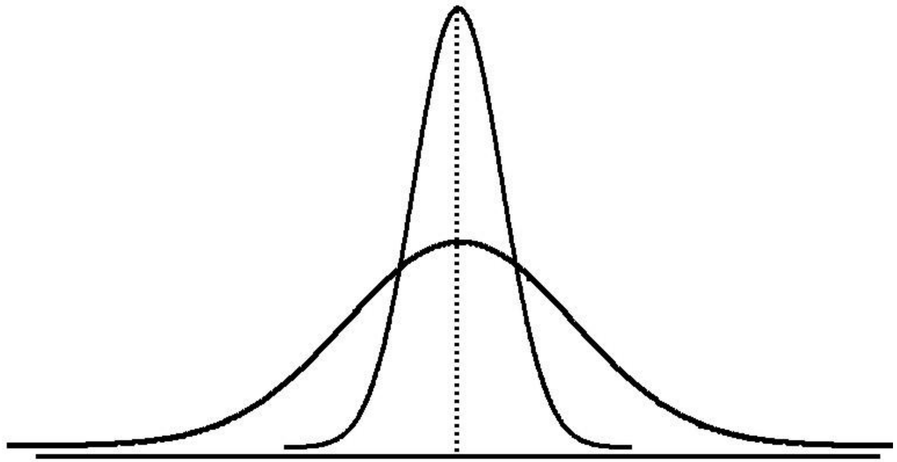
Measures of Variation

DCOVA

Variation

- **Range**
- **Variance**
- **Standard Deviation**
- **Coefficient of Variation**

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



Same center, different variation

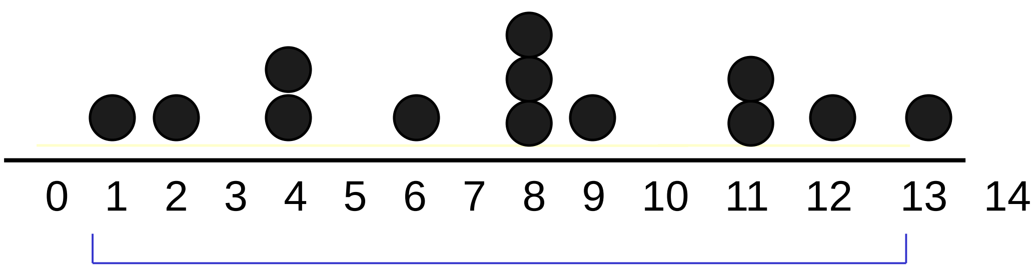
Measures of Variation: The Range

DCOVA

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

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

Example:

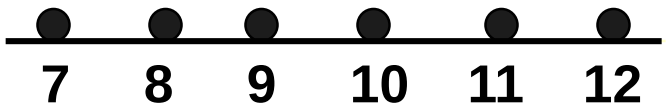


Range = 13 - 1 = 12

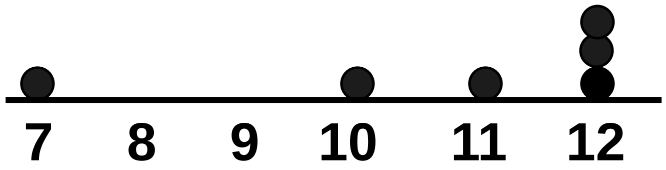
Measures of Variation: Why The Range Can Be Misleading

DCOVA

- Does not account for how the data are distributed.



Range = 12 - 7 = 5



Range = 12 - 7 = 5

- Sensitive to outliers

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

Range = 5 - 1 = 4

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

Range = 120 - 1 = 119

Measures of Variation: The Sample Variance

DCOVA

- Average (approximately) of squared deviations of values from the mean.
  - 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

DCOVA

- 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.

  - Sample standard deviation:

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

Measures of Variation: The Sample 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   Mean = $\bar{X}$ = 16

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

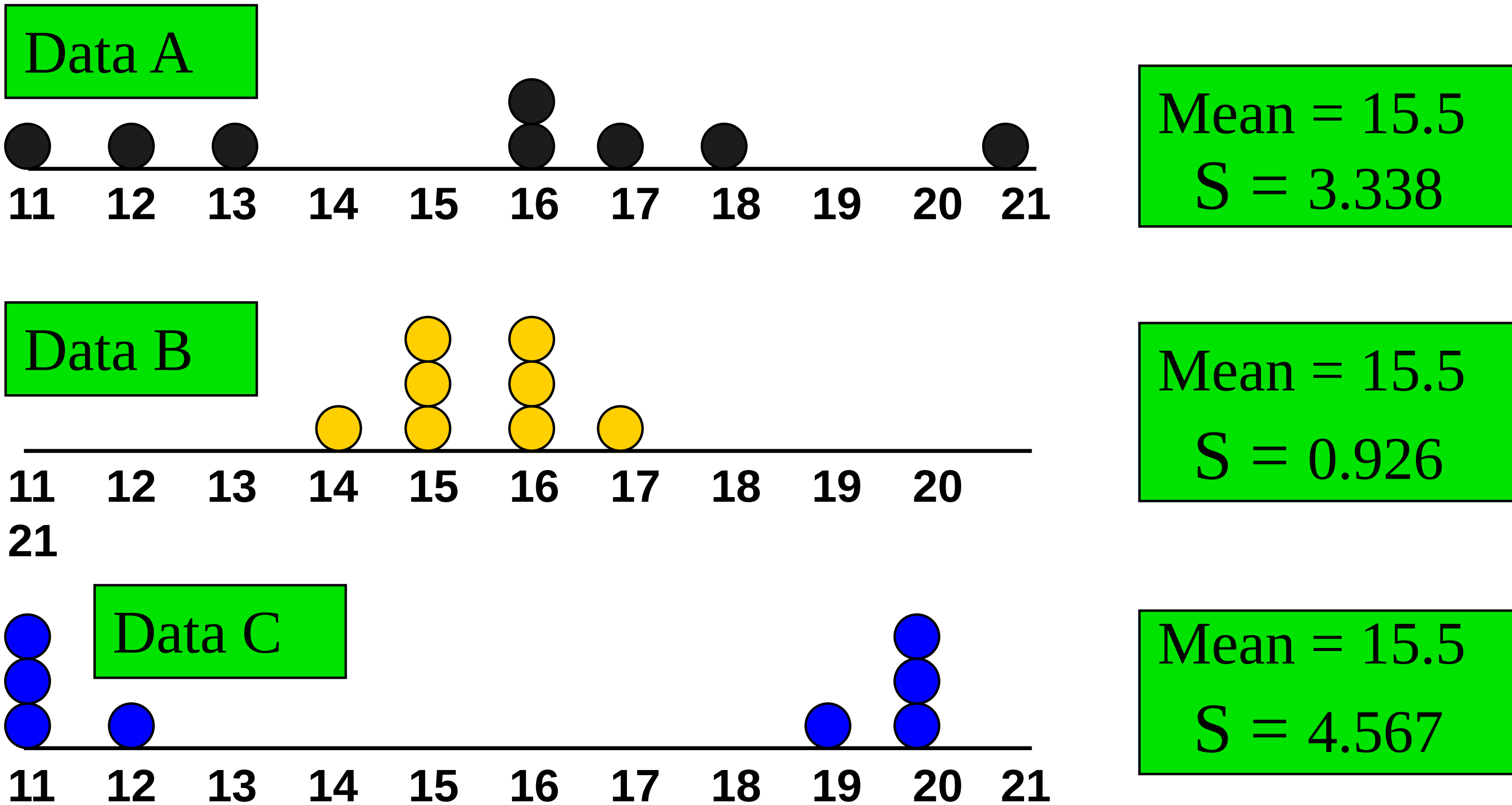
$$= \sqrt{\frac{(10-16)^2 + (12-16)^2 + (14-16)^2 + \cdots + (24-16)^2}{8-1}}$$

$$= \sqrt{\frac{130}{7}} = 4.3095$$

A measure of the "average" scatter around the mean.

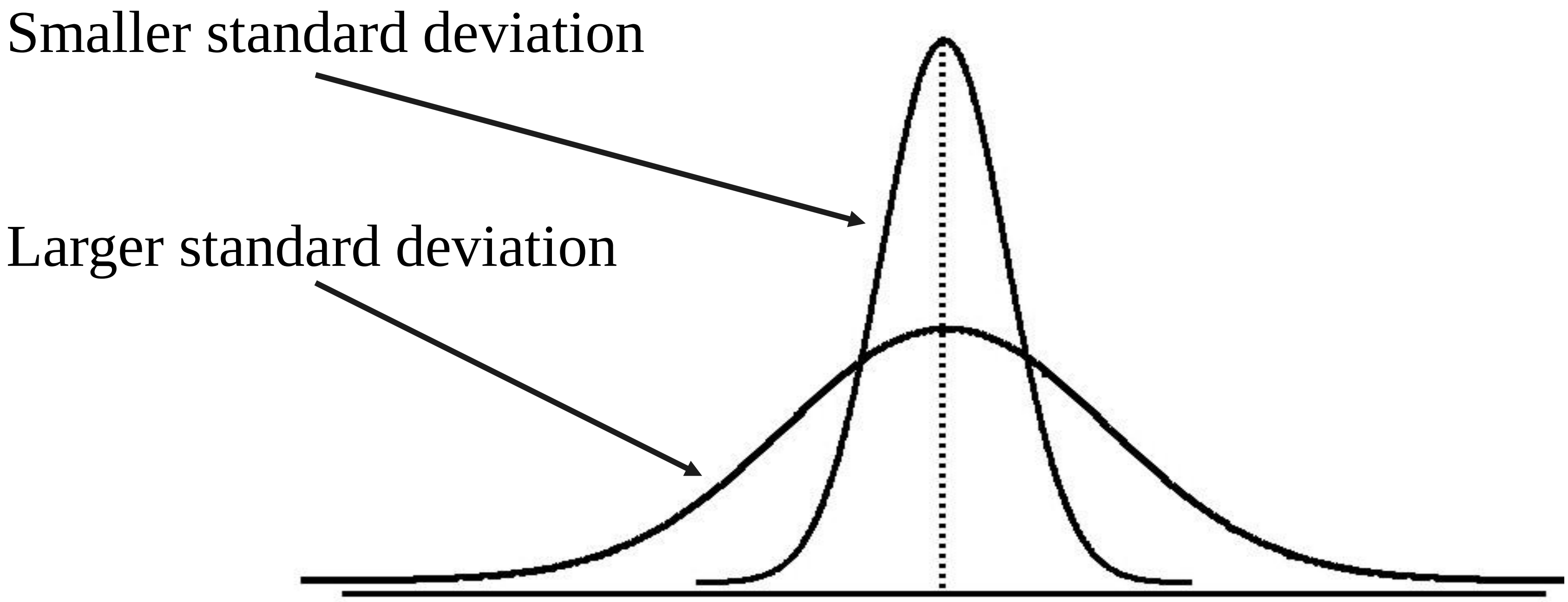
Measures of Variation: Comparing Standard Deviations

DCOVA



Measures of Variation: Comparing Standard Deviations

DCOVA



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

DCOVA

- Stock A:
  - Mean 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:
  - Mean 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 mean price.

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

DCOVA

- Stock A:
  - Mean price last year = $50.
  - Standard deviation = $5.

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

- Stock C:
  - Mean 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

DCOVA

- 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

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

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 and would not be considered an outlier.

Shape of a Distribution

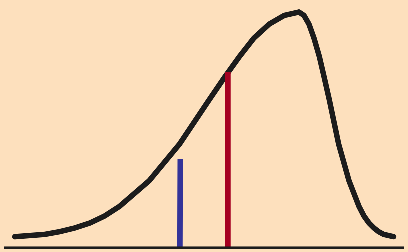
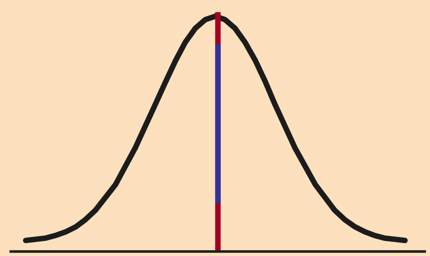
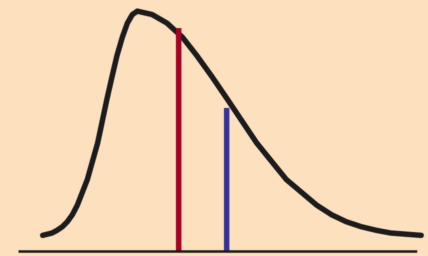
DCOVA

- Describes how data are distributed.
- Two useful shape related statistics are:
  - Skewness:
    - Measures the extent to which data values are not symmetrical.
  - Kurtosis:
    - Kurtosis measures 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 (Skewness)

DCOVA

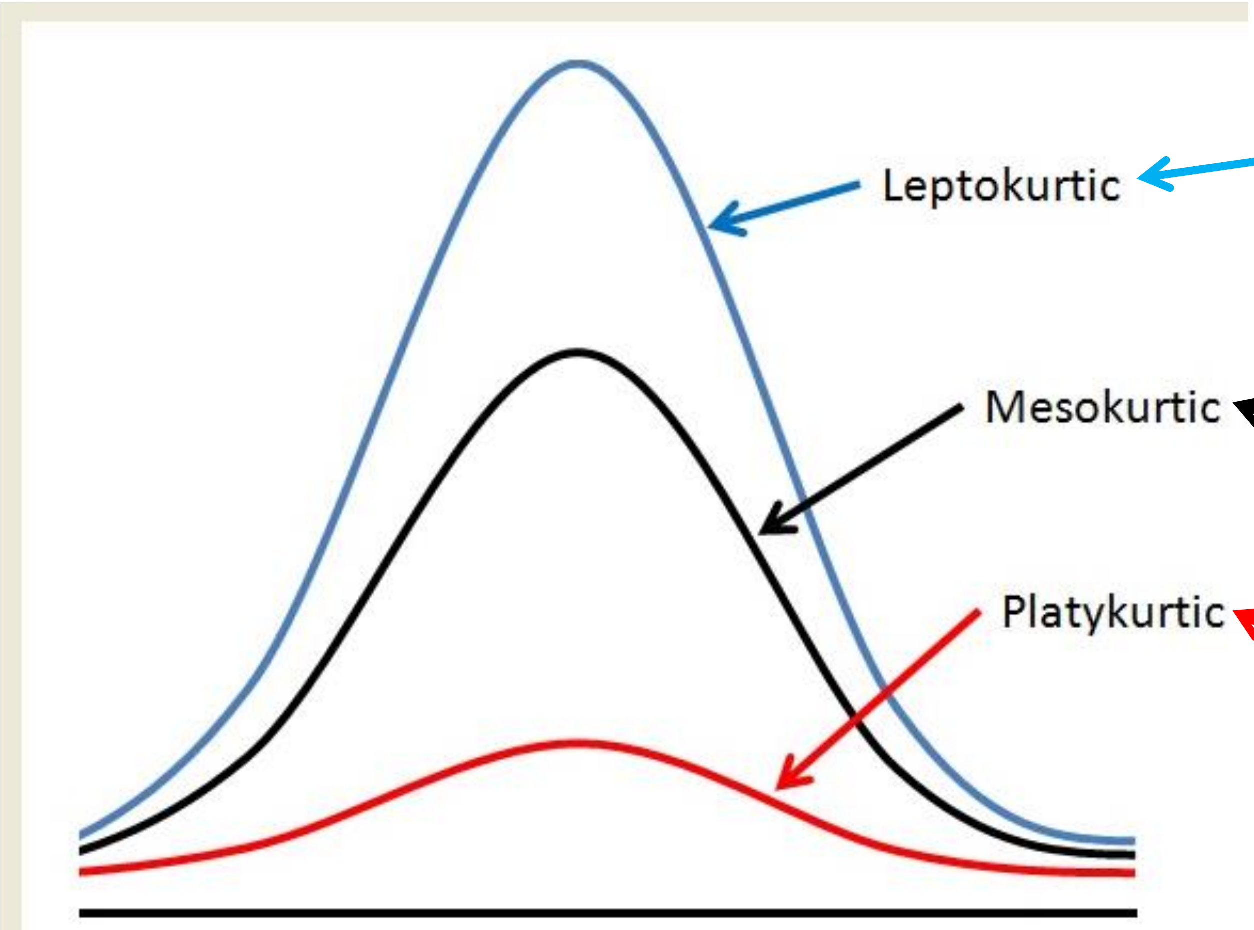
- Measures the extent to which data is not symmetrical.

Left-Skewed	Symmetric	Right-Skewed
Mean < Median	Mean = Median	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



Sharper Peak Than Bell-Shaped (Kurtosis > 0)

Bell-Shaped (Kurtosis = 0)

Flatter Than Bell-Shaped (Kurtosis < 0)

Exploring Numerical Data Using Quartiles

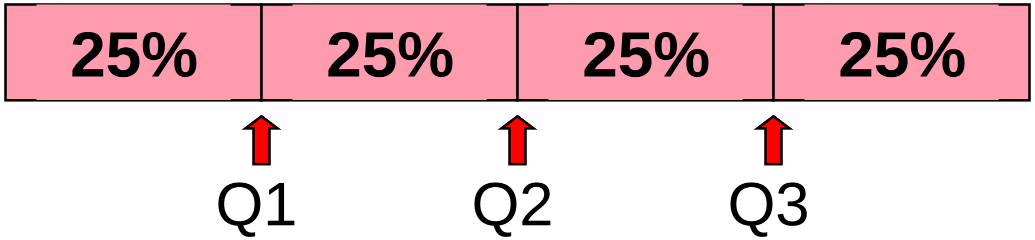
DCOVA

- Can visualize the distribution of the values for a numerical variable by computing:
  - The quartiles.
  - The five-number summary.
  - Constructing a boxplot.

Quartile Measures

DCOVA

- 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 values are smaller and 75% are larger.
- Q_2 is the same as the median (50% of the values are smaller and 50% are larger).
- Only 25% of the values 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
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,

so **Q_1 = (12+13)/2 = 12.5.**

Q_2 is in the (9+1)/2 = 5th position of the ranked data,

so **Q_2 = median = 16.**

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

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)

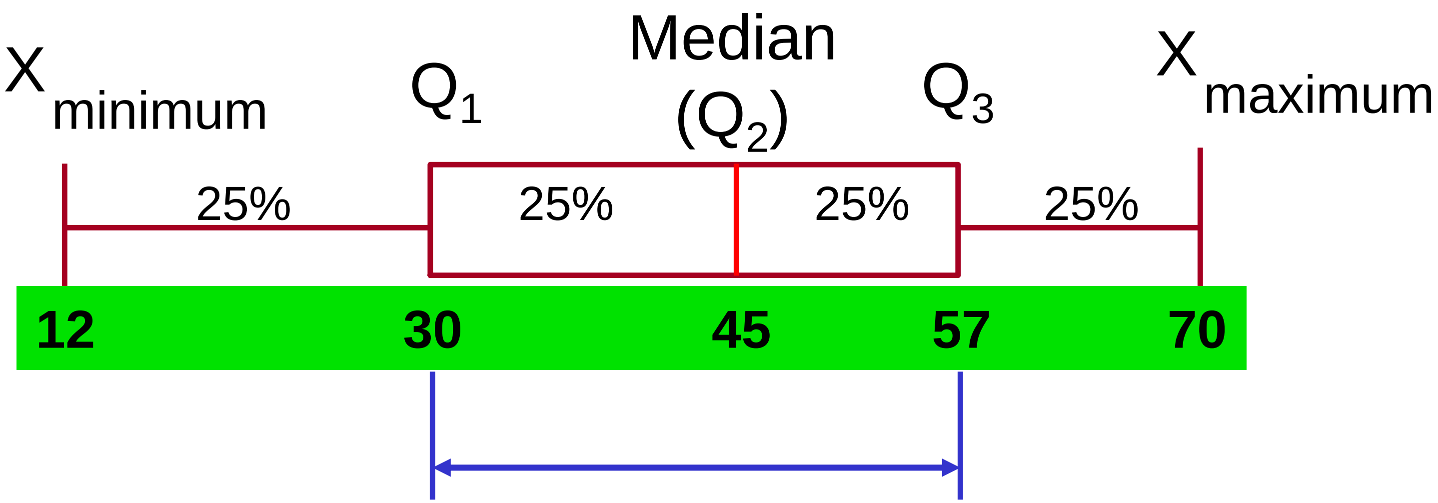
DCOVA

- 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:



Interquartile range
= 57 – 30 = 27

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).
- $X_{largest.}$

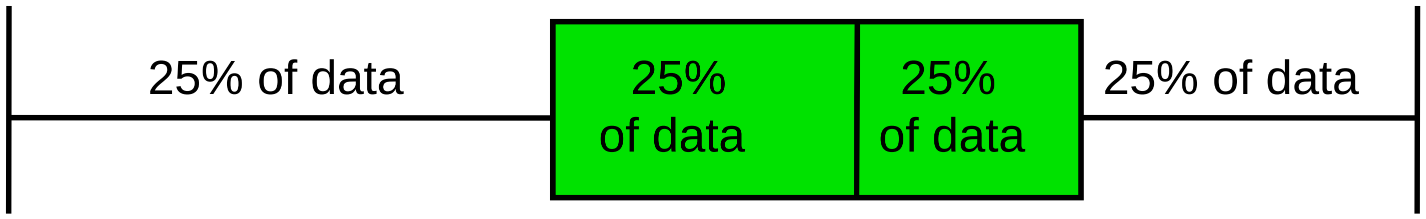
Five Number Summary and The Boxplot

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:

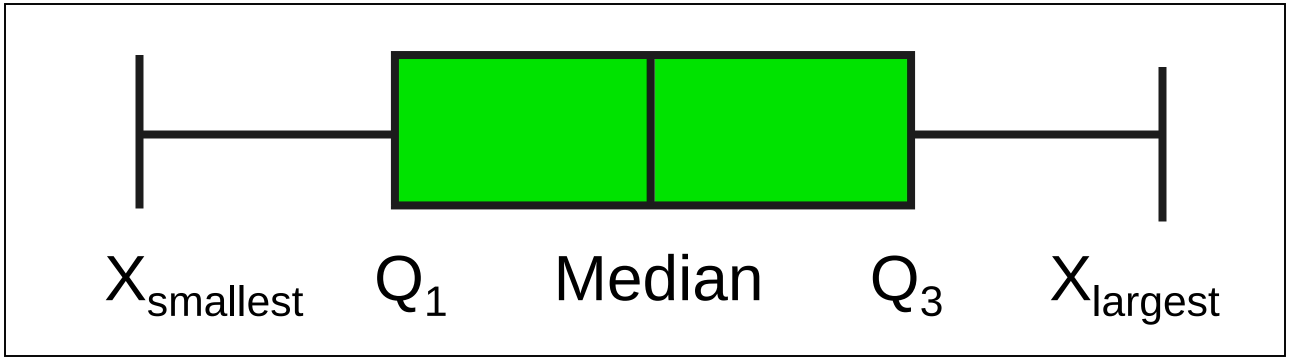


$X_{smallest}$ Q_1 Median Q_3 $X_{largest}$

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.

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

DCOVA

- 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

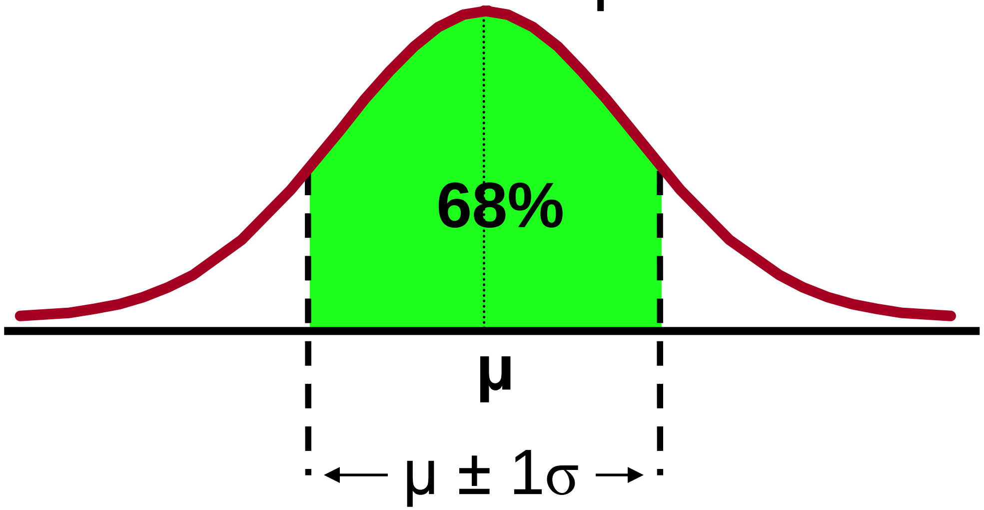
DCOVA

Measure	Population Parameter	Sample Statistic
Mean	μ	$\overline{X}$
Variance	σ^2	S^2
Standard Deviation	σ	S

The Empirical Rule

DCOVA

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

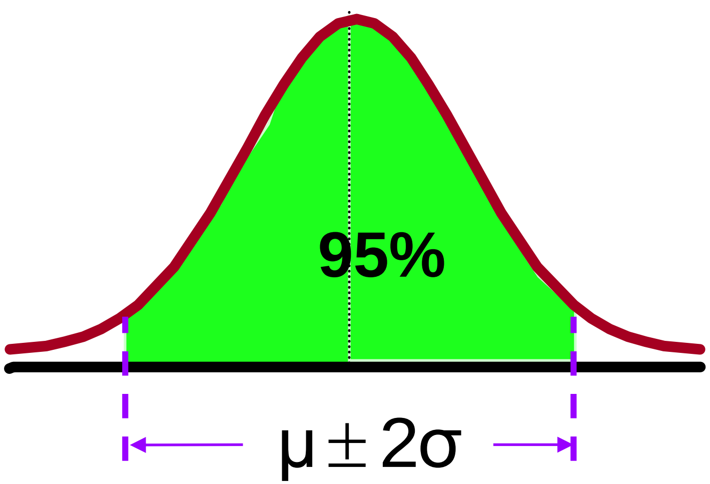


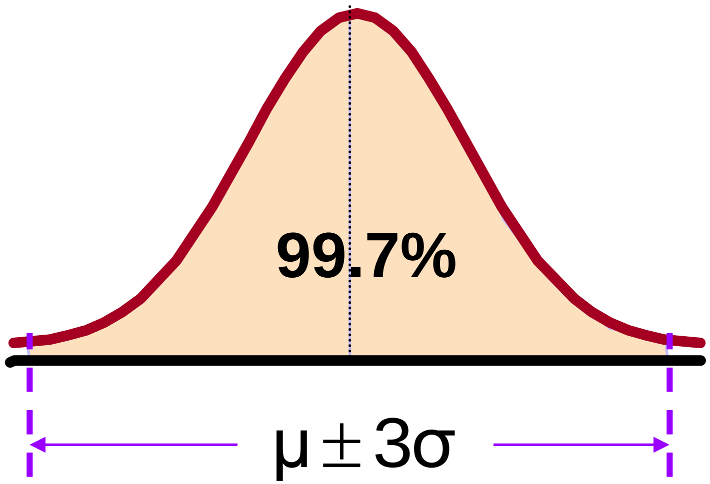
The Empirical Rule

DCOVA

- Approximately 95% of the data in a symmetric mound-shaped distribution lies within two standard deviations of the mean, or $\mu \pm 2\sigma$.

- Approximately 99.7% of the data in a symmetric mound-shaped distribution lies within three standard deviations of the mean, or $\mu \pm 3\sigma$.





Using the Empirical Rule

DCOVA

- Suppose that the variable Math SAT scores is bell-shaped with a mean of 500 and a standard deviation of 90. Then:

  - Approximately 68% of all test takers scored between 410 and 590, (500 ± 90).

  - Approximately 95% of all test takers scored between 320 and 680, (500 ± 180).

  - Approximately 99.7% of all test takers scored between 230 and 770, (500 ± 270).

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

DCOVA

- **Covariance** between two variables:

cov(X,Y) > 0 ⟶ X and Y tend to move in the same direction.

cov(X,Y) < 0 ⟶ X and Y tend to move in opposite directions.

cov(X,Y) = 0 ⟶ X and Y 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

DCOVA

- 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,

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

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

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

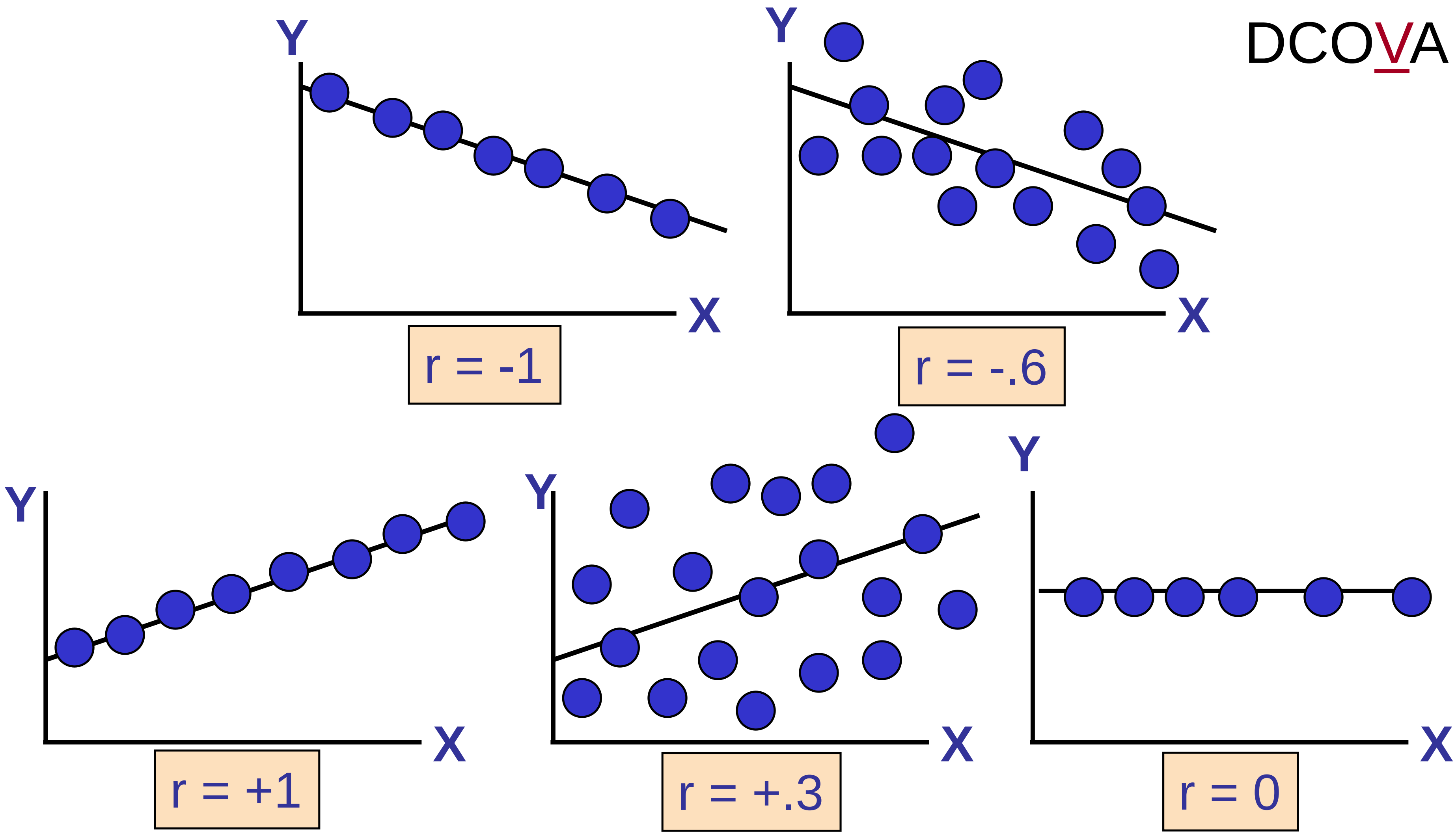
Features of the Coefficient of Correlation

DCOVA

- 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.

Scatter Plots of Sample Data with Various Coefficients of Correlation

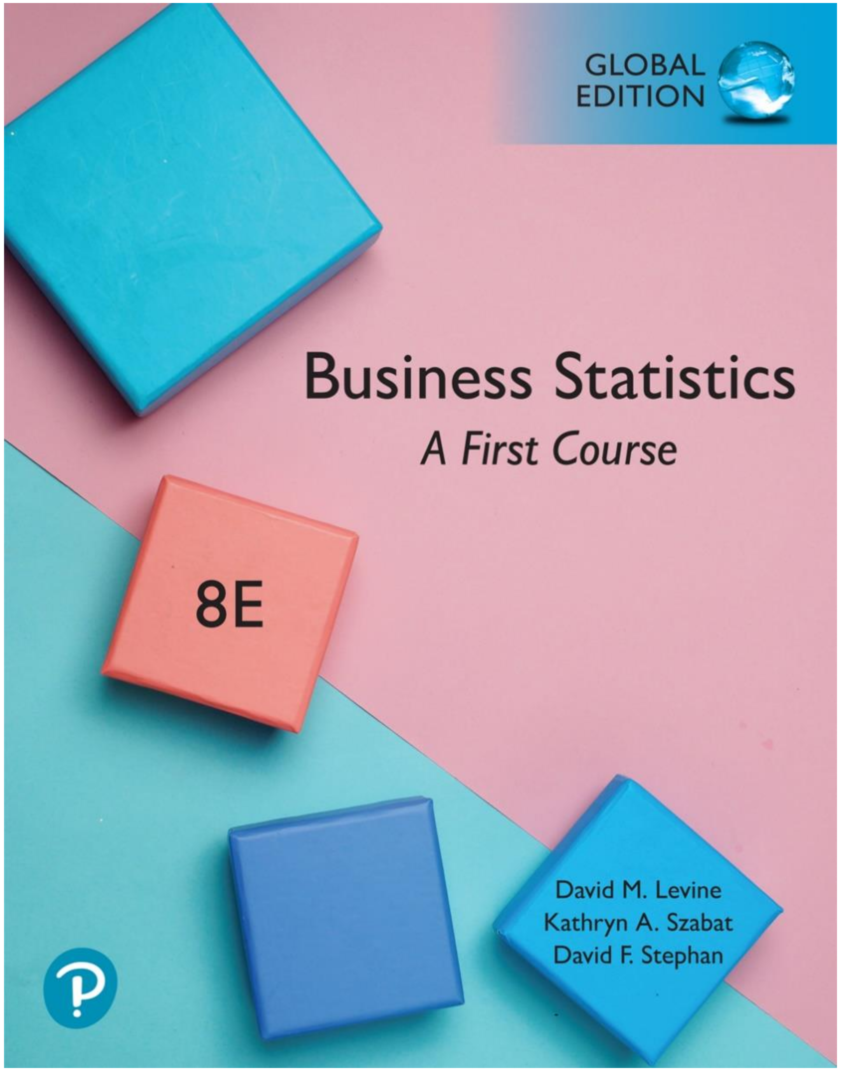
DCOVA



Chapter Summary

In this chapter we have discussed:

- Describing the properties of central tendency, variation, and shape in numerical variables.
- Constructing and interpreting a boxplot.
- Computing descriptive summary measures for a population.
- Calculating the covariance and the coefficient of correlation.



Chapter 4

Basic Probability

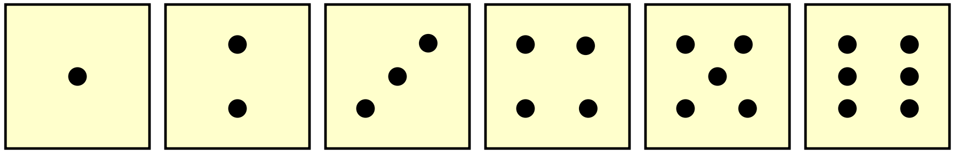
Objectives

The objectives for this chapter are:

- To understand basic probability concepts.
- To understand conditional probability.
- Use Bayes’ theorem to revise probabilities.
- Apply counting rules.

The Sample Space Is The Collection Of All Possible Outcomes Of A Variable

e.g. All 6 faces of a die:



e.g. All 52 cards of a bridge deck

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 2019.
- 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 2019.
- Complement of an event A (denoted A'):
  - All events that are not part of event A.
  - e.g., All days from 2019 that are not in January.

Basic Probability Concepts

- Probability – the numerical value representing the chance, likelihood, or possibility that a certain 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).

Mutually Exclusive Events

- Mutually exclusive events:
  - Events that cannot occur simultaneously.

Example: Randomly choosing a day from 2019

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 2019.

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.

Three Approaches To Assessing Probability Of An Event

Assuming all outcomes are equally likely

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

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

2. empirical probability -- *based on observed data*

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

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 2019 what is the probability the day is in January?

$$\textbf{Probability of Day In January} = \frac{X}{T} = \frac{\text{number of days in January}}{\text{total number of days in 2019}}$$

$$\frac{X}{T} = \frac{\text{31 days in January}}{\text{365 days in 2018}} = \frac{31}{365}$$

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 Differs From Person To Person

- What is the probability a new ad campaign is successful?
  - 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.

Summarizing Sample Spaces

Contingency Table -- M&R Survey Results.

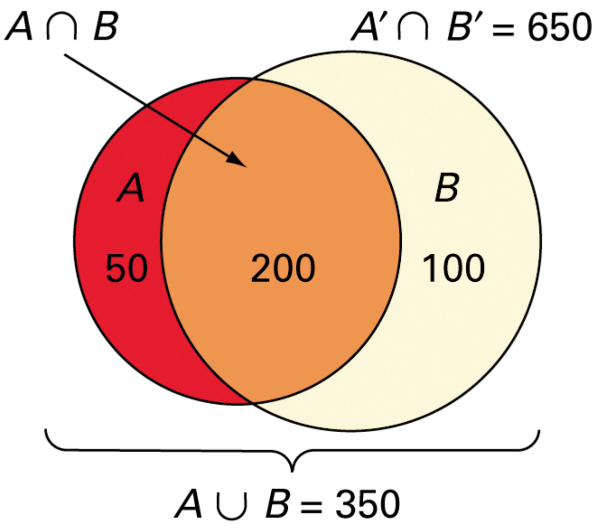
	Actually Purchased TV		
Planned To Purchase TV	Yes	No	Total
Yes	200	50	250
No	100	650	750
Total	300	700	1,000

Total Number
Of Sample
Space Outcomes.

Summarizing Sample Spaces

Venn Diagram -- M&R Survey Results.

A = Planned to Purchase
A' = Did not Plan To Purchase
B = Actually Purchased
B' = Did not Purchase



Planned To Purchase TV	Actually Purchased TV: Yes	No	Total
Yes	200	50	250
No	100	650	750
Total	300	700	1,000

Simple Probability: Definition & Computing

- Simple Probability refers to the probability of a simple event.
  - P(Planned to purchase)
  - P(Actually purchased)

$$P(A) = \frac{\textit{number of outcomes satisfying } A}{\textit{total number of outcomes}}$$

P(Purchase) = 250 / 1000

Planned To Purchase TV	Actually Purchased TV: Yes	No	Total
Yes	200	50	250
No	100	650	750
Total	300	700	1,000

P(Actually Purchased) = 300 / 1000

Joint Probability: Definition & Computing

- Joint Probability refers to the probability of an occurrence of two or more events (joint event).
  - ex. P(Plan to Purchase and Purchase).
  - ex. P(No Plan and Purchase).

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

P(Plan to Purchase and Purchase) = 200 / 1000

Planned To Purchase TV	Actually Purchased TV: Yes	No	Total
Yes	200	50	250
No	100	650	750
Total	300	700	1,000

P(No Plan and Purchase) = 50 / 1000

Computing A Marginal Probability Via Joint Probabilities

- 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, …, B_k are k mutually exclusive and collectively exhaustive events.

P(Planned) = P(Yes and Yes) + P(Yes and No) = 200 / 1000 + 50 / 1000 = 250 / 1000

Planned To Purchase TV	Actually Purchased TV: Yes	No	Total
Yes	200	50	250
No	100	650	750
Total	300	700	1,000

Marginal & Joint Probabilities In A Contingency Table

Event	Event: B_1	Event: B_2	Total
A_1	$P(A_1 \text{ and } B_1)$	$P(A_1 \text{ and } B_2)$	$P(A_1)$
A_2	$P(A_2 \text{ and } B_1)$	$P(A_2 \text{ and } B_2)$	$P(A_2)$
Total	$P(B_1)$	$P(B_2)$	1

Joint Probabilities

Marginal (Simple) Probabilities

Probability Summary So Far

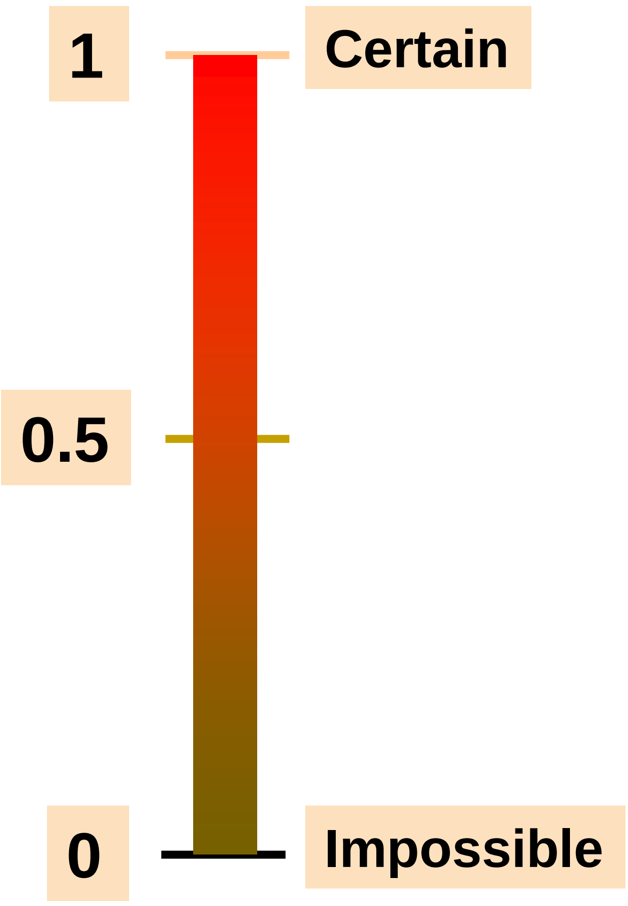
- Probability is the numerical measure of the likelihood that an event will occur.
- The probability of any event must be between 0 and 1, inclusively.

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

- 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 or B) = P(A) + P(B) - P(A and B)

If A and B are mutually exclusive, then

P(A and B) = 0, so the rule can be simplified:

P(A or B) = P(A) + P(B)

For mutually exclusive events A and B

General Addition Rule Example

P(Planned or Purchased) =

P(Planned) + P(Purchased) – P(Planned and Purchased) =

250 / 1,000 + 300 / 1,000 – 200 / 1,000 = 350 / 1,000

Planned To Purchase TV	Actually Purchased TV: Yes	No	Total
Yes	200	50	250
No	100	650	750
Total	300	700	1,000

Computing Conditional Probabilities

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

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

→ The conditional probability of A given that B has occurred.

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

→ The conditional probability of B given that A has occurred.

Where P(A 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

P(Purchased | Planned) = P(Purchased and Planned) / P(Planned) =

(200 / 1000) / (250 / 1000) = 200 / 250

Planned To Purchase TV	Actually Purchased TV Yes	No	Total
Yes	200	50	250
No	100	650	750
Total	300	700	1,000

Since Planned is given we only need to consider the top row of the table.

Independent Events

- 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.

Are The Events Planned and Purchased Independent?

Does P(Purchased | Planned) = P(Purchased)?

P(Purchased | Planned) = 200 / 250 = 0.8.

P(Purchased) = 700 / 1000 = 0.7.

Since these two probabilities are not equal, these two events are dependent.

	Actually Purchased TV		
Planned To Purchase TV	Yes	No	Total
Yes	200	50	250
No	100	650	750
Total	300	700	1,000

Multiplication Rules For Two Events

The General Multiplication Rule

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

Solving for P(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 Using The General Multiplication Rule

- Marginal probability for event A:

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

  - Where B_1, B_2, …, B_k are k mutually exclusive and collectively exhaustive events.

Let A = Planned, B_1 = Purchase, & B_2 = No Purchase

$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) =$

(200/300)(300/1000) + (50/700)(700/1000) = 0.25

Planned To Purchase TV		Actually Purchased TV: Yes	No	Total
	Yes	200	50	250
	No	100	650	750
	Total	300	700	1,000

Bayes' Theorem

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

Bayes' Theorem

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

- 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 P(D|U) = 0.2

- Goal is to find P(S|D)

Bayes' Theorem Example

(continued)

Apply Bayes' Theorem:

$$P(S|D) = \frac{P(D|S)P(S)}{P(D|S)P(S) + P(D|U)P(U)}$$

$$= \frac{(0.6)(0.4)}{(0.6)(0.4) + (0.2)(0.6)}$$

$$= \frac{0.24}{0.24 + 0.12} = 0.667$$

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 = 0.36

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:
  - 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:
  - 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:
  - 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

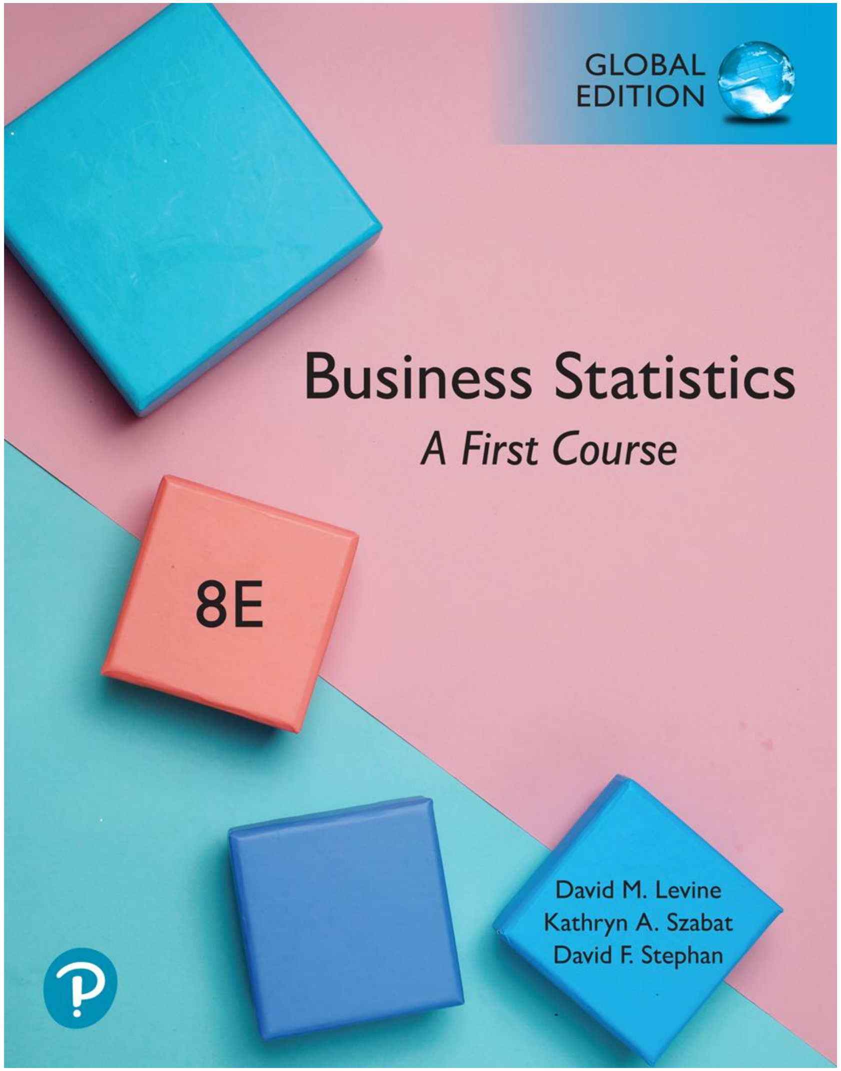
$$_{n}C_{x} = \frac{n!}{X!(n-X)!}$$

  - Example:
    - You have five books and are going to select three are to read. How many different combinations are there, ignoring the order in which they are selected?
    - Answer: $_{n}C_{x} = \frac{n!}{X!(n-X)!} = \frac{5!}{3!(5-3)!} = \frac{120}{(6)(2)} = 10$ different possibilities

Chapter Summary

In this chapter we covered:

- Using basic probability concepts.
- Using conditional probability.
- Using Bayes' theorem to revise probabilities.
- Using counting rules.



Chapter 5

Discrete Probability Distributions

Objectives

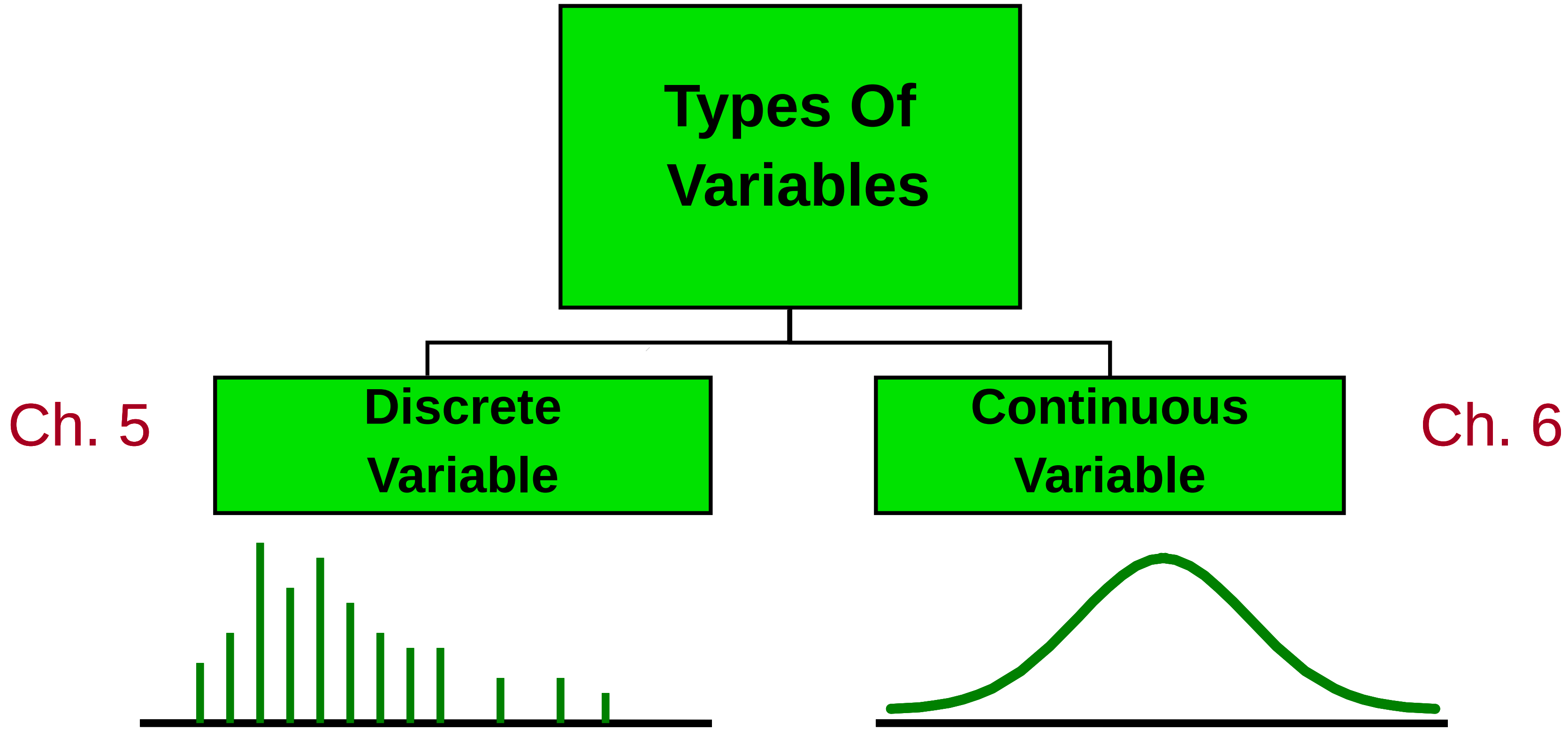
In this chapter, you learn:

- The properties of a probability distribution.
- How to calculate the expected value and variance of a probability distribution.
- How to calculate probabilities from binomial and Poisson distributions.
- How to use the binomial and Poisson distributions to solve business problems.

Definitions

- **Discrete** variables produce outcomes that come from a counting process (e.g. number of classes you are taking).
- **Continuous** variables produce outcomes that come from a measurement (e.g. your annual salary, or your weight).

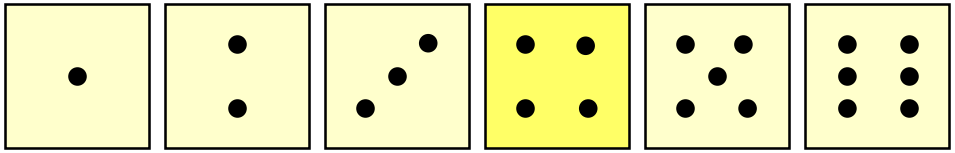
Types Of Variables



Discrete Variables

- Can only assume a countable number of values.

Examples:



- **Roll a die twice**
  Let X be the number of times 4 occurs
  (then X could be 0, 1, or 2 times).

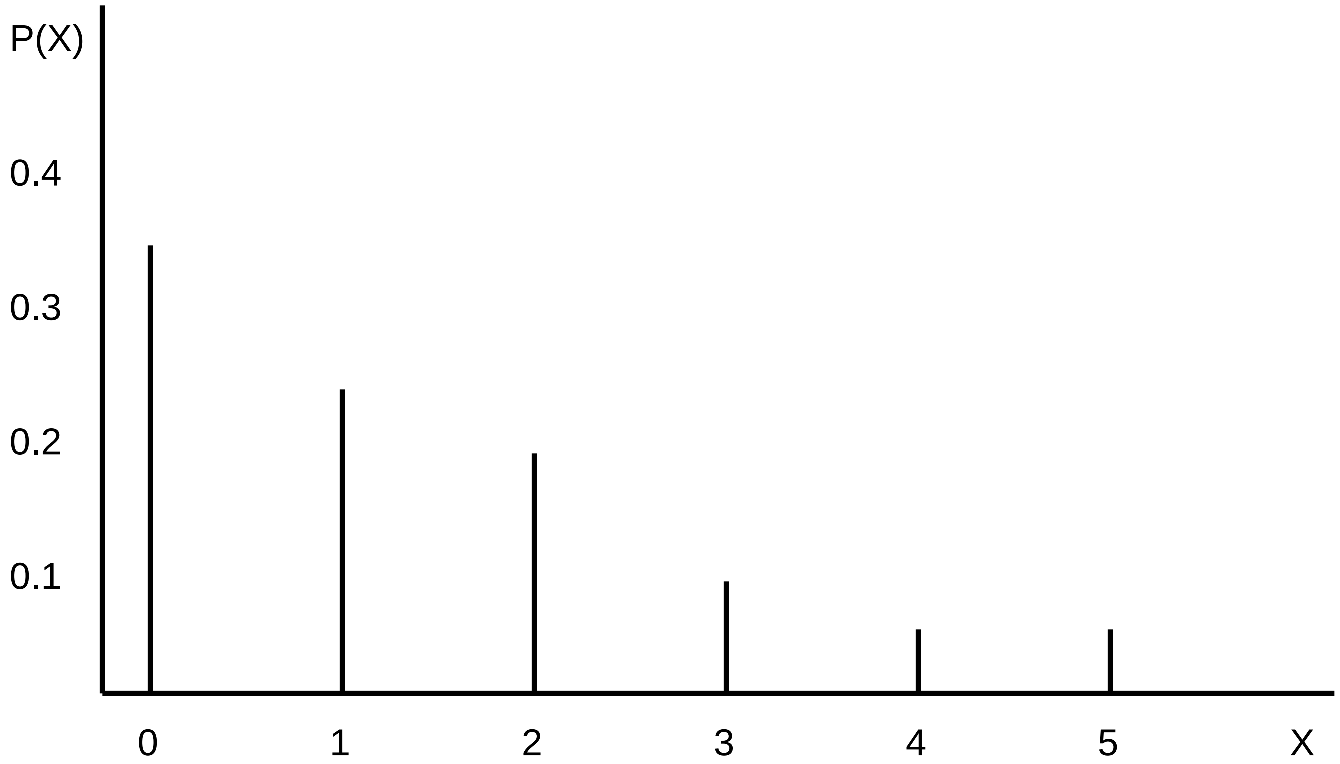
- **Toss a coin 5 times.**
  Let X be the number of heads
  (then X = 0, 1, 2, 3, 4, or 5).

Probability Distribution For A Discrete Variable

- A **probability distribution for a discrete variable** is a mutually exclusive list of all possible numerical outcomes for that variable and a probability of occurrence associated with each outcome.

Interruptions Per Day In Computer Network	Probability
0	0.35
1	0.25
2	0.20
3	0.10
4	0.05
5	0.05

Probability Distributions Are Often Represented Graphically



Expected Value Of Discrete Variables, Measuring Center

- Expected Value (or mean) of a discrete variable (Weighted Average):

$$\mu = \mathrm{E}(\mathrm{X}) = \sum_{i=1}^{N} x_i P(X = x_i)$$

Interruptions Per Day In Computer Network (x_i)	Probability $P(X = x_i)$	$x_i P(X = x_i)$
0	0.35	(0)(0.35) = 0.00
1	0.25	(1)(0.25) = 0.25
2	0.20	(2)(0.20) = 0.40
3	0.10	(3)(0.10) = 0.30
4	0.05	(4)(0.05) = 0.20
5	0.05	(5)(0.05) = 0.25
	1.00	μ = E(X) = 1.40

Discrete Variables: Measuring Dispersion

- Variance of a discrete variable.

$$\sigma^2 = \sum_{i=1}^{N} [x_i - E(X)]^2 P(X = x_i)$$

- Standard Deviation of a discrete variable.

$$\sigma = \sqrt{\sigma^2} = \sqrt{\sum_{i=1}^{N} [x_i - E(X)]^2 P(X = x_i)}$$

where:

E(X) = Expected value of the discrete variable X

x_i = the i^{th} outcome of X

$P(X=x_i)$ = Probability of the i^{th} occurrence of X

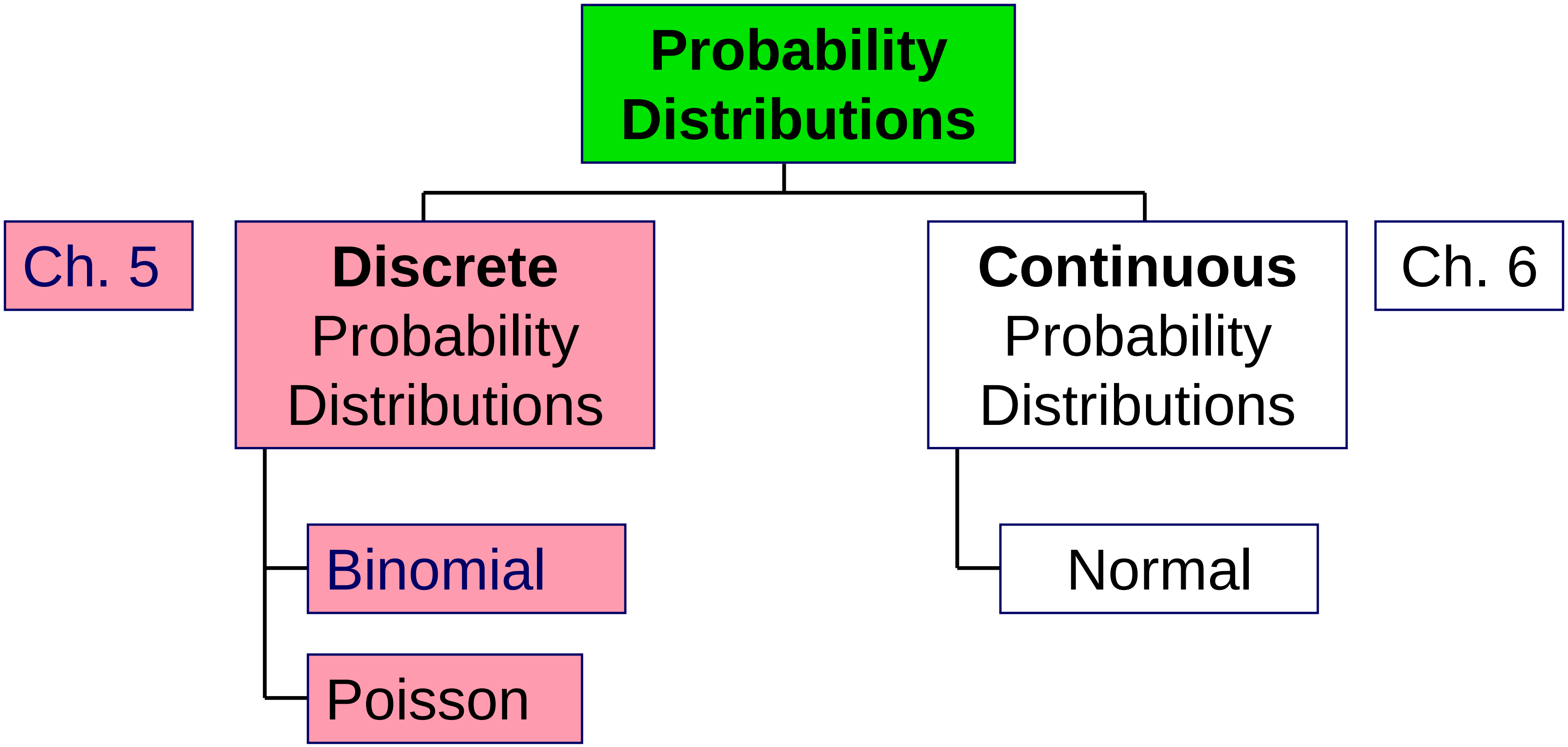
Discrete Variables: Measuring Dispersion

(continued)

$$\sigma = \sqrt{\sum_{i=1}^{N} [x_i - E(X)]^2 P(X = x_i)}$$

Interruptions Per Day In Computer Network (x_i)	Probability $P(X = x_i)$	$[x_i - E(X)]^2$	$[x_i - E(X)]^2 P(X = x_i)$
0	0.35	$(0 - 1.4)^2 = 1.96$	(1.96)(0.35) = 0.686
1	0.25	$(1 - 1.4)^2 = 0.16$	(0.16)(0.25) = 0.040
2	0.20	$(2 - 1.4)^2 = 0.36$	(0.36)(0.20) = 0.072
3	0.10	$(3 - 1.4)^2 = 2.56$	(2.56)(0.10) = 0.256
4	0.05	$(4 - 1.4)^2 = 6.76$	(6.76)(0.05) = 0.338
5	0.05	$(5 - 1.4)^2 = 12.96$	(12.96)(0.05) = 0.648
			$\sigma^2 = 2.04$, $\sigma = 1.4283$

Probability Distributions



Binomial Probability Distribution

- A fixed number of observations, n.
  - e.g., 15 tosses of a coin; ten light bulbs taken from a warehouse.
- Each observation is classified into one of two mutually exclusive & collectively exhaustive categories.
  - e.g., head or tail in each toss of a coin; defective or not defective light bulb.
- The probability of being classified as the event of interest, π, is constant from observation to observation.
  - Probability of getting a tail is the same each time we toss the coin.
  - Since the two categories are mutually exclusive and collectively exhaustive, when the probability of the event of interest is π, the probability of the event of interest not occurring is $1 - \pi$.
- The value of any observation is independent of the value of any other observation.

Possible Applications for the Binomial Distribution

- A manufacturing plant labels items as either defective or acceptable.
- A firm bidding for contracts will either get a contract or not.
- A marketing research firm receives survey responses of “yes I will buy” or “no I will not.”
- New job applicants either accept the offer or reject it.

The Binomial Distribution Counting Techniques

- Suppose the event of interest is obtaining heads on the toss of a fair coin. You are to toss the coin three times. In how many ways can you get two heads?
- Possible ways: HHT, HTH, THH, so there are three ways you can getting two heads.
- This situation is fairly simple. We need to be able to count the number of ways for more complicated situations.

Counting Techniques
Rule of Combinations

- The number of combinations of selecting x objects out of n objects is:

$$ {}_nC_x = \frac{n!}{x!(n-x)!} $$

where:

$n! = (n)(n-1)(n-2)\cdots(2)(1)$

$x! = (X)(X-1)(X-2)\cdots(2)(1)$

$0! = 1$ (by definition)

Counting Techniques
Rule of Combinations

- How many possible 3 scoop combinations could you create at an ice cream parlor if you have 31 flavors to select from and no flavor can be used more than once in the 3 scoops?
- The total choices is n = 31, and we select X = 3.

$$_{31}C_3 = \frac{31!}{3!(31-3)!} = \frac{31!}{3!28!} = \frac{31 \bullet 30 \bullet 29 \bullet 28!}{3 \bullet 2 \bullet 1 \bullet 28!} = 31 \bullet 5 \bullet 29 = 4{,}495$$

Binomial Distribution Formula

$$P(X=x\,|n,\pi) = \frac{n!}{x!\,(n-x)!}\,\pi^{x}\,(1-\pi)^{n-x}$$

$P(X=x|n,\pi)$ = probability that X = **x** events of interest, given **n** and **π**

x = number of "events of interest" in sample, (x = 0, 1, 2, ..., n)

n = sample size (number of trials or observations)

π = probability of "event of interest"

1 – π = probability of not having an event of interest

Example: Flip a coin four times, let x = # heads:

n = 4

π = 0.5

1 - π = (1 - 0.5) = 0.5

X = 0, 1, 2, 3, 4

Example: Calculating a Binomial Probability

What is the probability of one success in five observations if the probability of an event of interest is 0.1?

x = 1, n = 5, and π = 0.1

$$P(X=1|5,0.1) = \frac{n!}{x!(n-x)!}\pi^x(1-\pi)^{n-x}$$

$$= \frac{5!}{1!(5-1)!}(0.1)^1(1-0.1)^{5-1}$$

$$= (5)(0.1)(0.9)^4$$

$$= 0.32805$$

The Binomial Distribution Example

Suppose the probability of an invoice payment being late is 0.10. What is the probability of 1 late invoice payment in a group of 4 invoices?

$x = 1$, $n = 4$, and $\pi = 0.10$

$$
\begin{aligned}
P(X=1|4, 0.10) &= \frac{n!}{x!(n-x)!}\pi^{x}(1-\pi)^{n-x} \\
&= \frac{4!}{1!(4-1)!}(0.10)^{1}(1-0.10)^{4-1} \\
&= (4)(0.10)(0.729) \\
&= 0.2916
\end{aligned}
$$

Excel, JMP, & Minitab Can Be Used To Calculate Binomial Probabilities

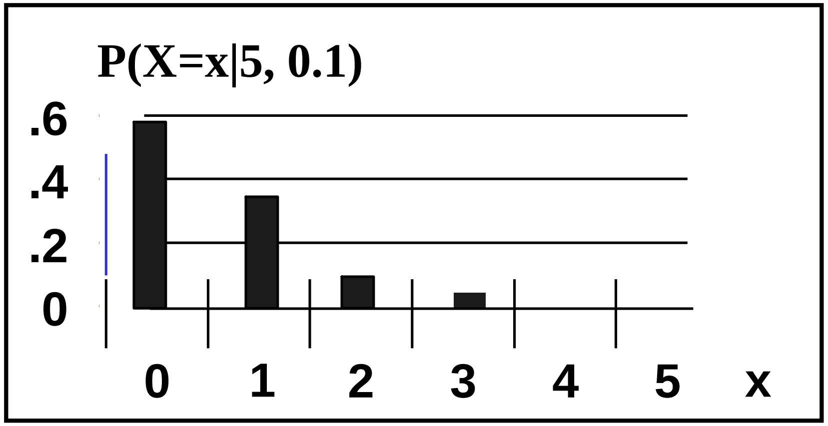
	A	B
1	Binomial Probabilities	
2		
3	Data	
4	Sample size	4
5	Probability of an event of interest	0.1
6		
7	Parameters	
8	Mean	0.4
9	Variance	0.36
10	Standard deviation	0.6
11		
12	Binomial Probabilities Table	
13	*X*	*P(X)*
14	0	0.6561
15	1	0.2916
16	2	0.0486
17	3	0.0036
18	4	0.0001

2/1 Cols / 5/0 Rows	X	P(X)
1	0	0.6561
2	1	0.2916
3	2	0.0486
4	3	0.0036
5	4	0.0001

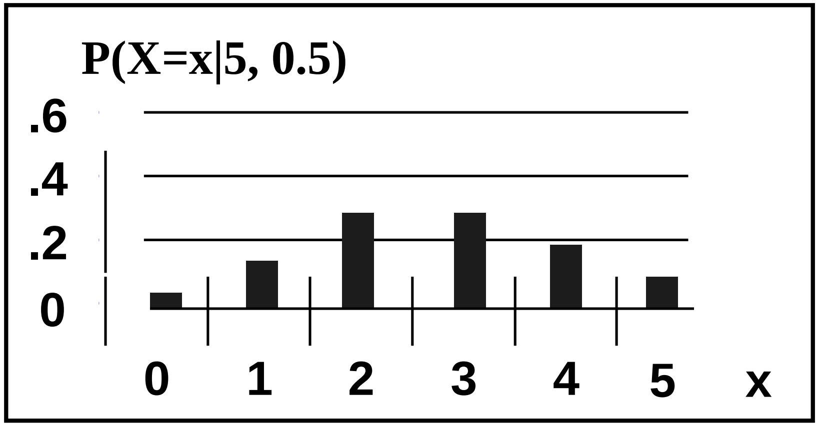
↓	C1	C2
	X	P(X)
1	0	0.6561
2	1	0.2916
3	2	0.0486
4	3	0.0036
5	4	0.0001

The Binomial Distribution Shape

- The shape of the binomial distribution depends on the values of π and n.



- Here, n = 5 and π = 0.1.



- Here, n = 5 and π = 0.5.

Binomial Distribution Characteristics

- Mean:

$$\mu = E(X) = n\pi$$

- Variance and Standard Deviation:

$$\sigma^2 = n\pi(1 - \pi)$$

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

Where n = sample size

π = probability of the event of interest for any trial

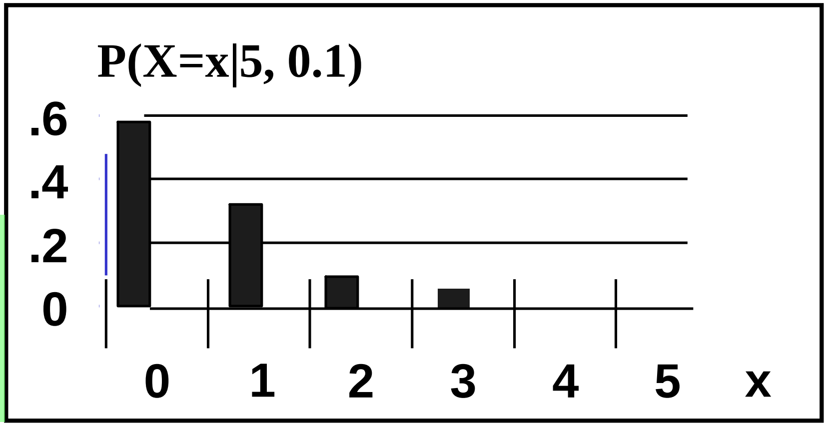
$(1 – \pi)$ = probability of no event of interest for any trial

The Binomial Distribution Characteristics

Examples

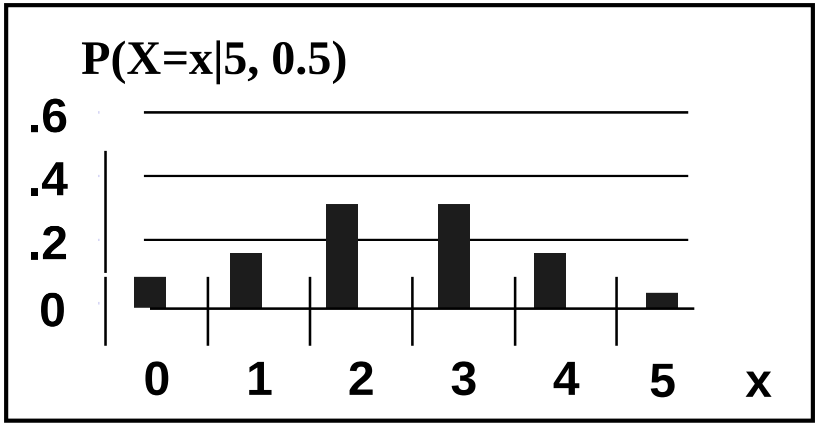
$$\mu = n\pi = (5)(.1) = 0.5$$

$$\sigma = \sqrt{n\pi(1-\pi)} = \sqrt{(5)(.1)(1-.1)} = 0.6708$$



$$\mu = n\pi = (5)(.5) = 2.5$$

$$\sigma = \sqrt{n\pi(1-\pi)} = \sqrt{(5)(.5)(1-.5)} = 1.118$$



The Poisson Distribution Definitions

- You use the **Poisson distribution** when you are interested in the number of times an event occurs in a given **area of opportunity**.
- An **area of opportunity** is a continuous unit or interval of time, volume, or such area in which more than one occurrence of an event can occur.
  - The number of scratches in a car's paint.
  - The number of mosquito bites on a person.
  - The number of computer crashes in a day

The Poisson Distribution

- Apply the Poisson Distribution when:
  - You are interested in counting the number of times a particular event occurs in a given area of opportunity. An area of opportunity is defined by time, length, surface area, and so forth.
  - The probability that an event occurs in a given area of opportunity is the same for all the areas of opportunity.
  - The number of events that occur in one area of opportunity is independent of the number of events that occur in any other area of opportunity.
  - The probability that two or more events will occur in an area of opportunity approaches zero as the area of opportunity becomes smaller.
- The average number of events per unit is λ (lambda).

Poisson Distribution Formula

$$P(X = x \mid \lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

where:

x = number of events in an area of opportunity

λ = expected number of events

e = base of the natural logarithm system (2.71828...)

Poisson Distribution Characteristics

- Mean:

$$\mu = \lambda$$

- Variance and Standard Deviation:

$$\sigma^2 = \lambda$$

$$\sigma = \sqrt{\lambda}$$

where λ = expected number of events.

The Poisson Distribution Example

The mean number of customers who arrive per minute at a bank during the noon-to-1pm hour is 3.0. What is the probability that 2 customers arrive in a given minute?

$$x = 2, \lambda = 3.0$$

$$P(X = 2|3.0) = \frac{e^{-\lambda}\lambda^{x}}{x!} = \frac{e^{-3.0}3.0^{2}}{2!}$$

$$= \frac{9}{2.71828^{3}(2)}$$

$$= 0.2240$$

Excel & Minitab & JMP Can Automate Poisson Probability Calculations

The mean number of customers who arrive per minute at a bank during the noon-to-1pm hour is 3.0.

$$\lambda = 3.0$$

	A	B	C	D	E
1	Poisson Probabilities				
2					
3	Data				
4	Mean/Expected number of events of interest:				3
5					
6	Poisson Probabilities Table				
7	*X*	*P(X)*			
8	0	0.0498	=POISSON.DIST(A8, E4, FALSE)		
9	1	0.1494	=POISSON.DIST(A9, E4, FALSE)		
10	2	0.2240	=POISSON.DIST(A10, E4, FALSE)		
11	3	0.2240	=POISSON.DIST(A11, E4, FALSE)		
12	4	0.1680	=POISSON.DIST(A12, E4, FALSE)		
13	5	0.1008	=POISSON.DIST(A13, E4, FALSE)		
14	6	0.0504	=POISSON.DIST(A14, E4, FALSE)		
15	7	0.0216	=POISSON.DIST(A15, E4, FALSE)		
16	8	0.0081	=POISSON.DIST(A16, E4, FALSE)		
17	9	0.0027	=POISSON.DIST(A17, E4, FALSE)		
18	10	0.0008	=POISSON.DIST(A18, E4, FALSE)		
19	11	0.0002	=POISSON.DIST(A19, E4, FALSE)		
20	12	0.0001	=POISSON.DIST(A20, E4, FALSE)		
21	13	0.0000	=POISSON.DIST(A21, E4, FALSE)		
22	14	0.0000	=POISSON.DIST(A22, E4, FALSE)		
23	15	0.0000	=POISSON.DIST(A23, E4, FALSE)		

Probability Density Function

Poisson with mean = 3

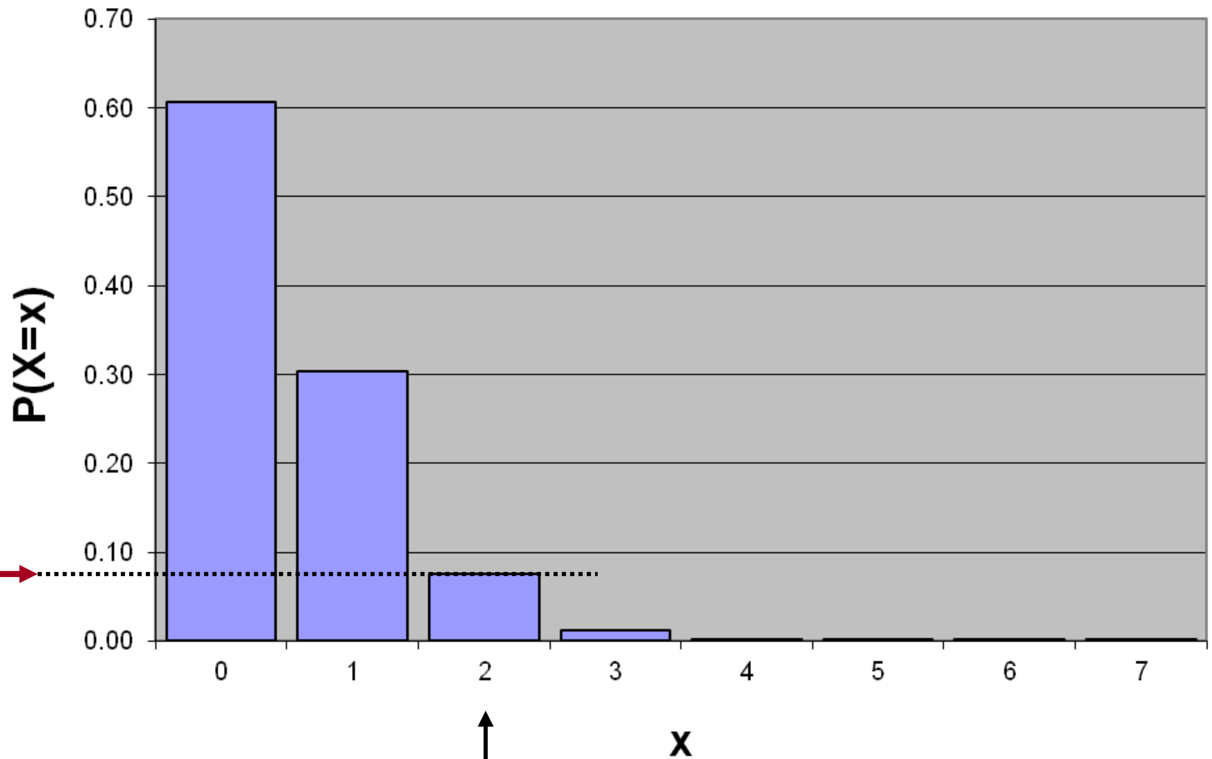
x	P(X = x)
0	0.049787
1	0.149361
2	0.224042
3	0.224042
4	0.168031
5	0.100819
6	0.050409
7	0.021604
8	0.008102
9	0.002701
10	0.000810
11	0.000221
12	0.000055
13	0.000013
14	0.000003
15	0.000001

Graph of Poisson Probabilities

Graphically:

λ = 0.50

X	**λ = 0.50**
0	0.6065
1	0.3033
2	0.0758
3	0.0126
4	0.0016
5	0.0002
6	0.0000
7	0.0000

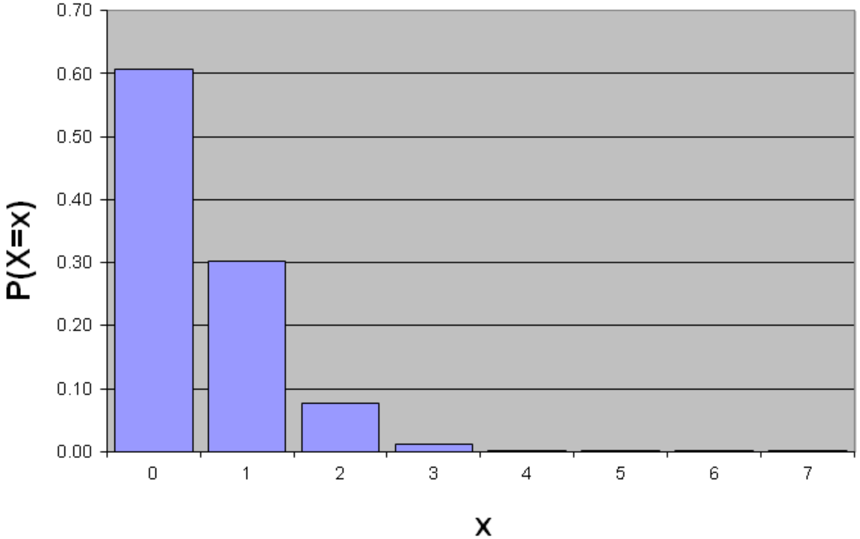


$P(X = 2 \mid \lambda=0.50) = 0.0758$

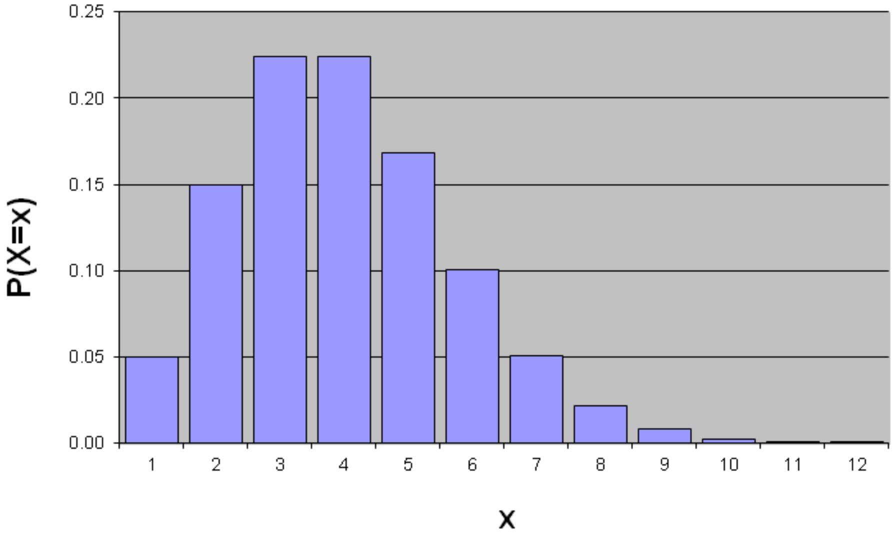
Poisson Distribution Shape

- The shape of the Poisson Distribution depends on the parameter λ:

$\lambda = 0.50$



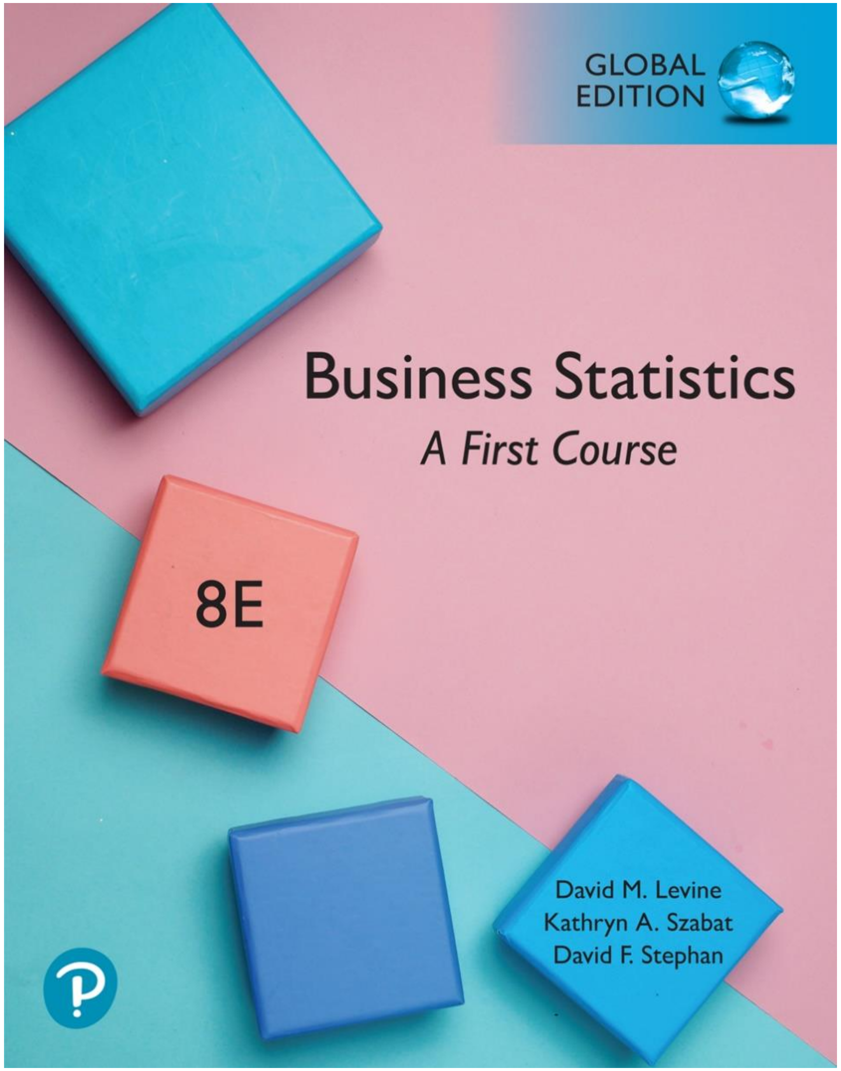
$\lambda = 3.00$



Chapter Summary

In this chapter we covered:

- The properties of a probability distribution.
- Computing the expected value and variance of a probability distribution.
- Computing probabilities from binomial and Poisson distributions.
- Using the binomial and Poisson distributions to solve business problems.



Chapter 6

The Normal Distribution

Objectives

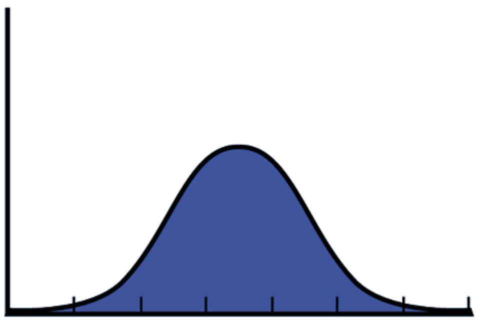
In this chapter, you learn:

- 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.

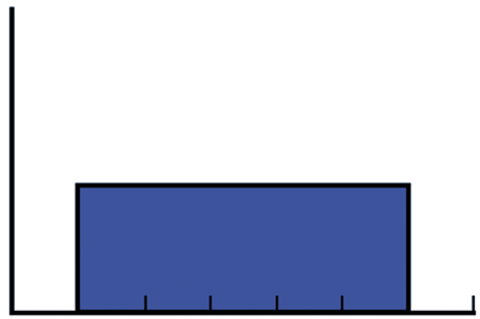
Continuous Probability Distributions

- A continuous 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.

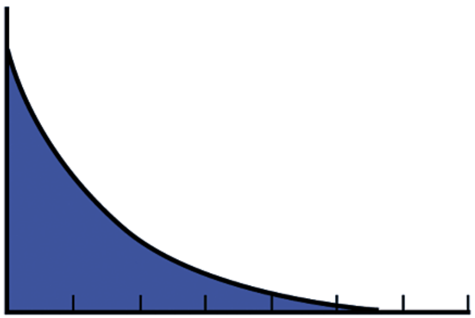
Continuous Probability Distributions Vary By Shape



Normal Distribution



Uniform Distribution



Exponential Distribution

- Symmetrical
- Bell-shaped
- Ranges from negative to positive infinity

- Symmetrical
- Also known as Rectangular Distribution
- Every value between the smallest & largest is equally likely

- Right skewed
- Mean > Median
- Ranges from zero to positive infinity

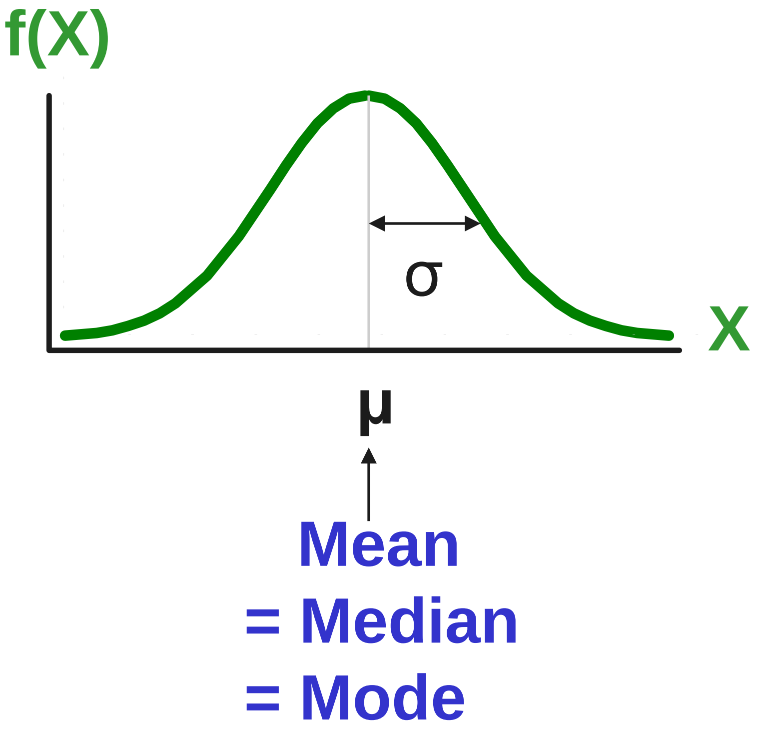
The Normal Distribution

- **Bell Shaped.**
- **Symmetrical.**
- **Mean, Median and Mode are equal.**

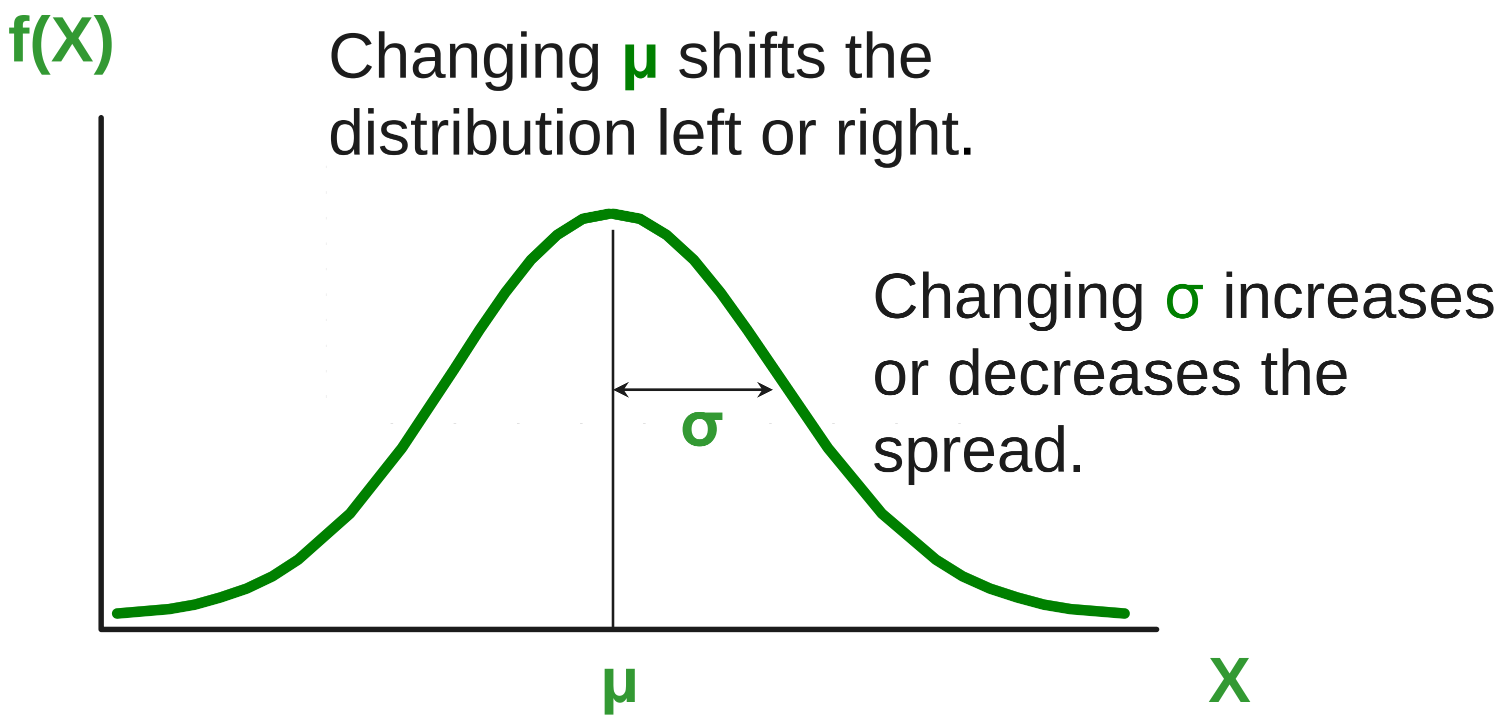
Location is determined by the mean, μ.

Spread is determined by the standard deviation, σ.

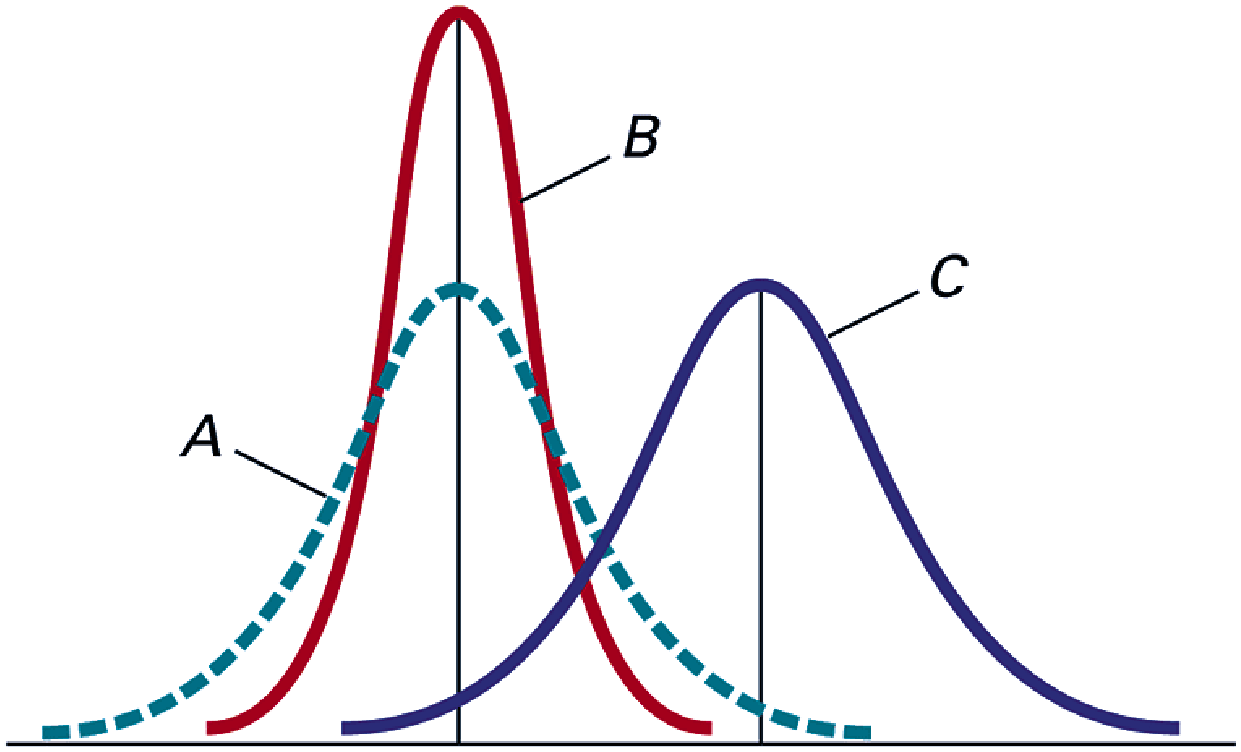
The random variable has an infinite theoretical range: -∞ to +∞.



The Normal Distribution Shape



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 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

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 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

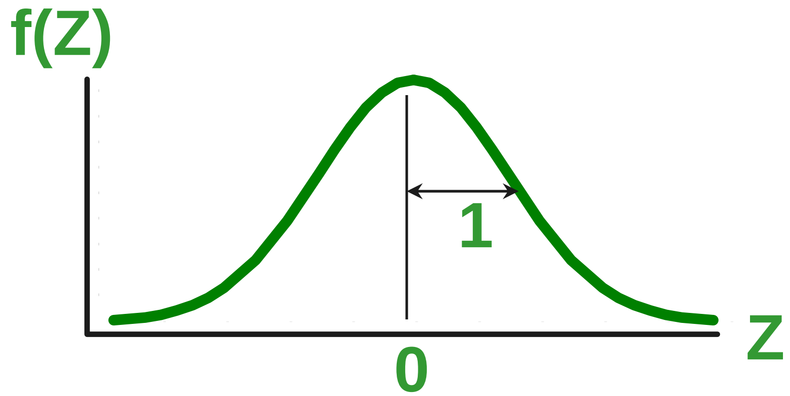
- Translate from X to the standardized normal (the "Z" distribution) by subtracting the mean of X and dividing by its standard deviation:

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

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

The Standardized Normal 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.

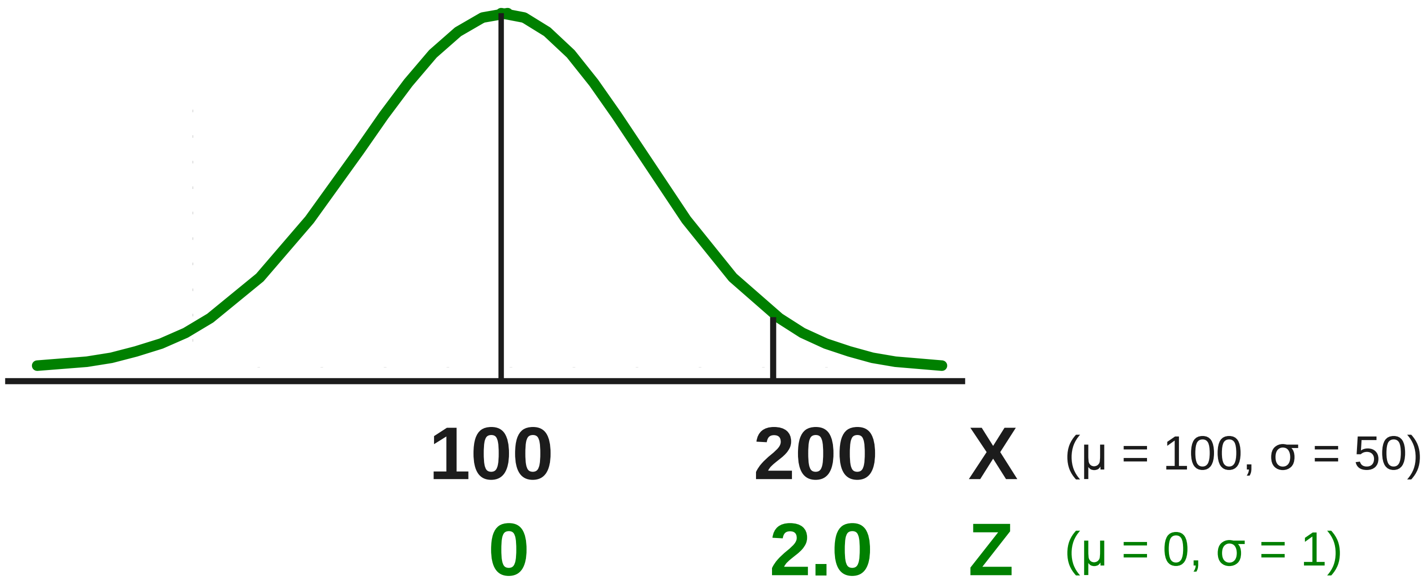
Example

- 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.

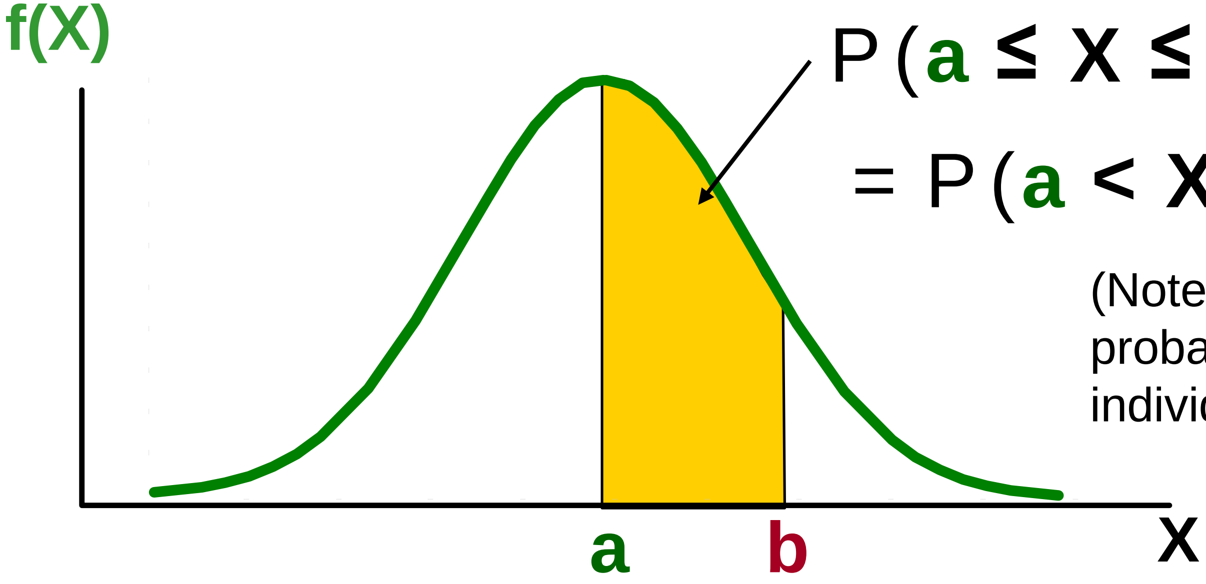
Comparing X and Z units



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

Finding Normal Probabilities

Probability is measured by the area under the curve.



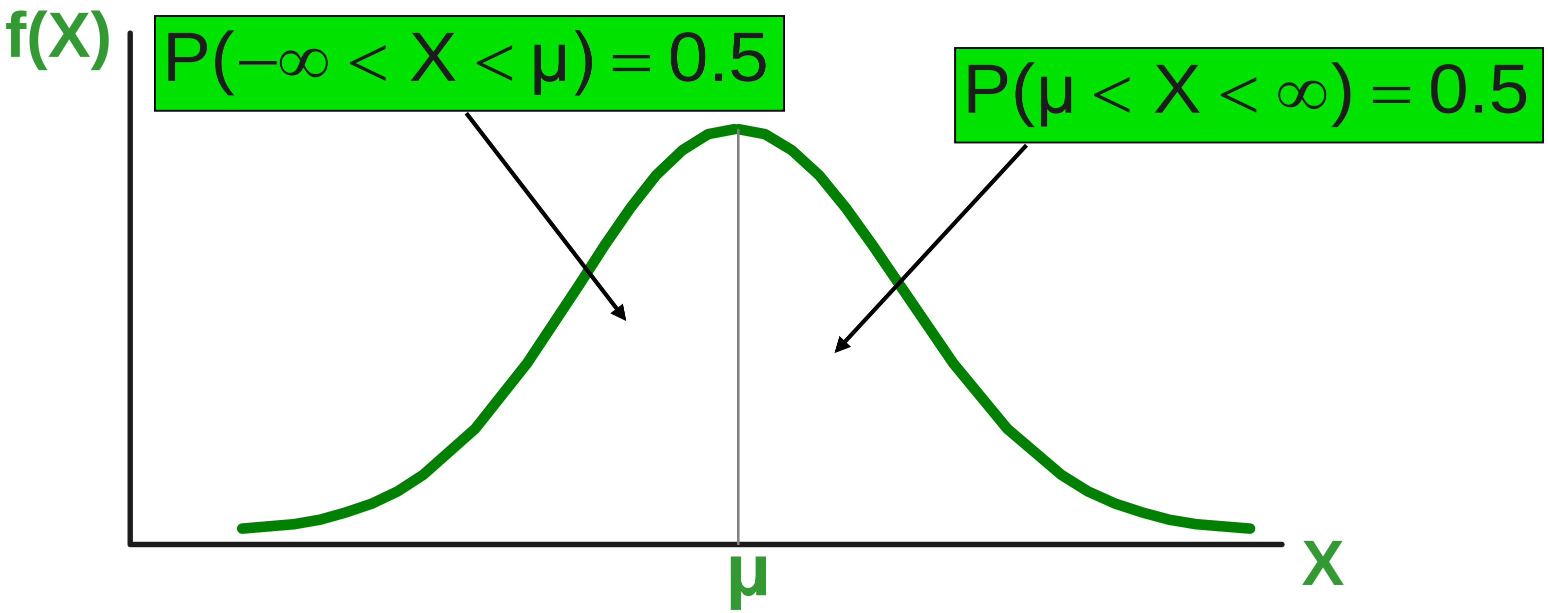
$$P(a \le X \le b)$$

$$= P(a < X < b)$$

(Note that the probability of any individual value is zero.)

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.



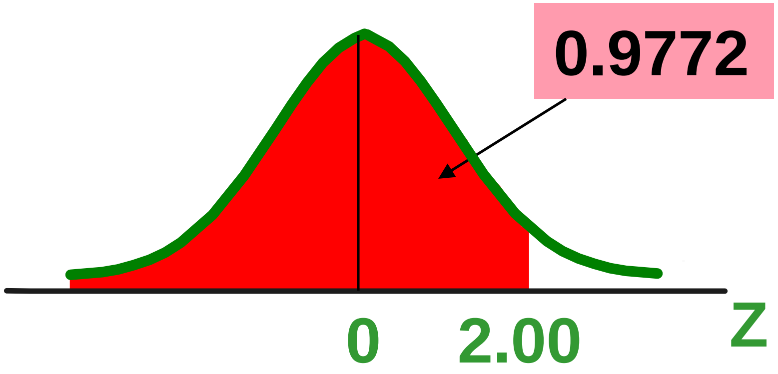
$$P(-\infty < X < \infty) = 1.0$$

The Cumulative 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

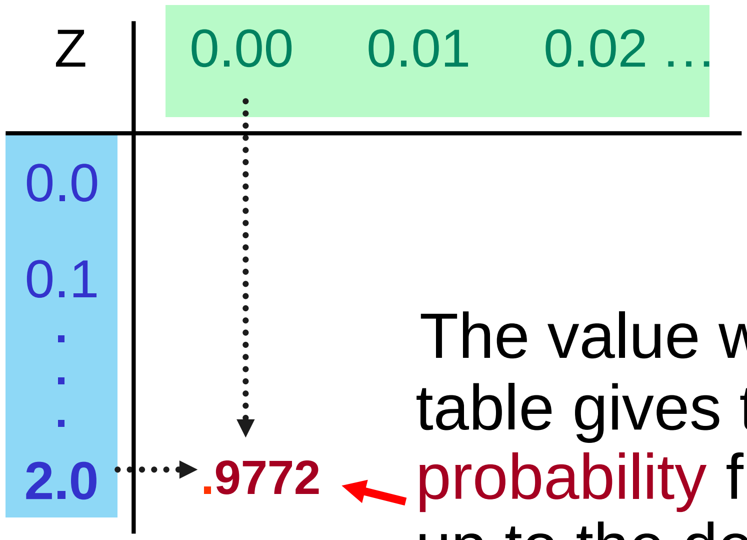


The Cumulative Standardized Normal Table

(continued)

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

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

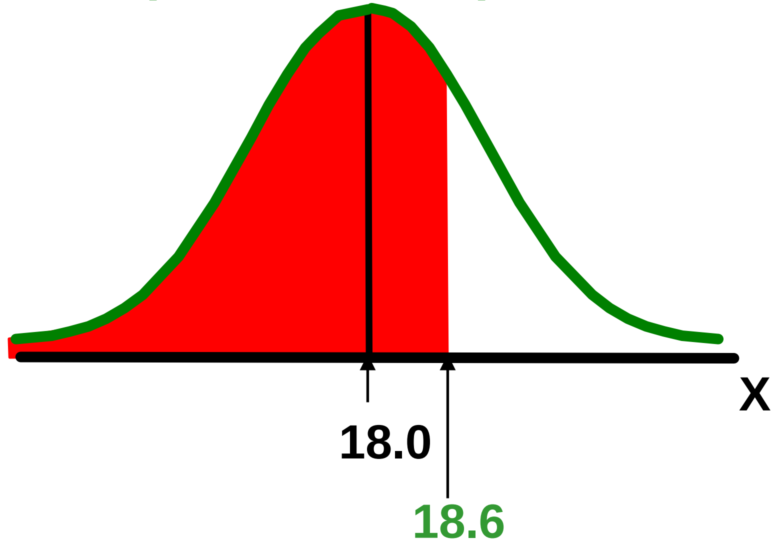
General Procedure for Finding Normal Probabilities

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 Cumulative Standardized Normal Table.

Finding Normal Probabilities

- 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 $P(X < 18.6)$.

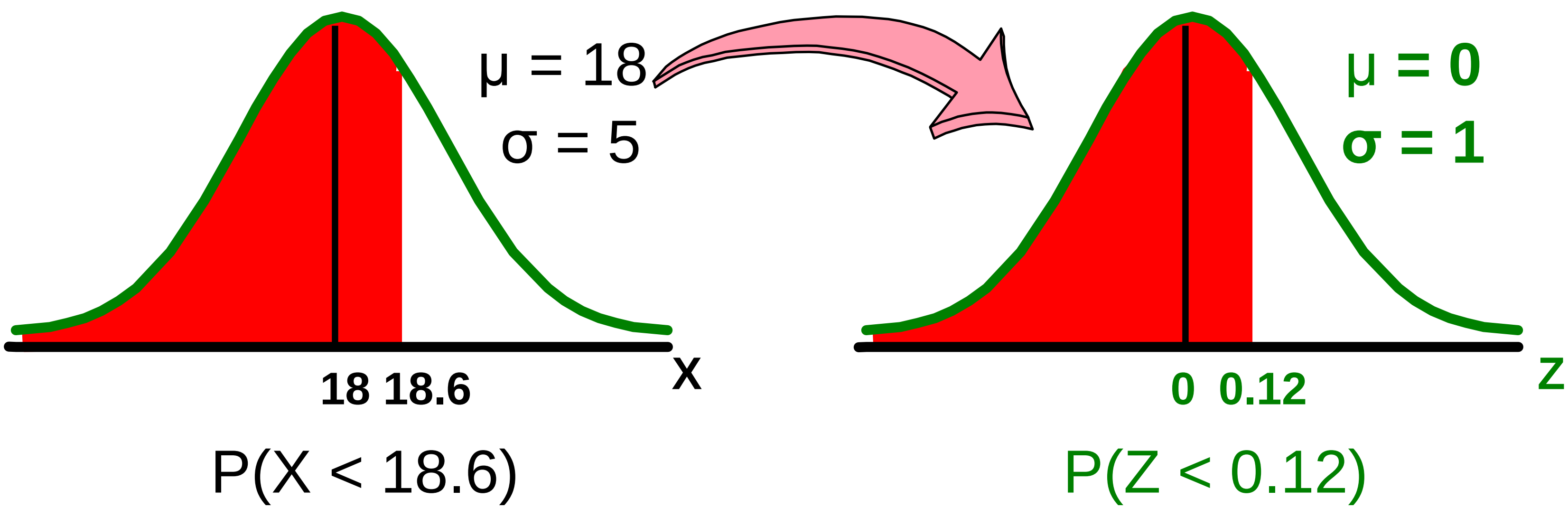


Finding Normal Probabilities

(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):**

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



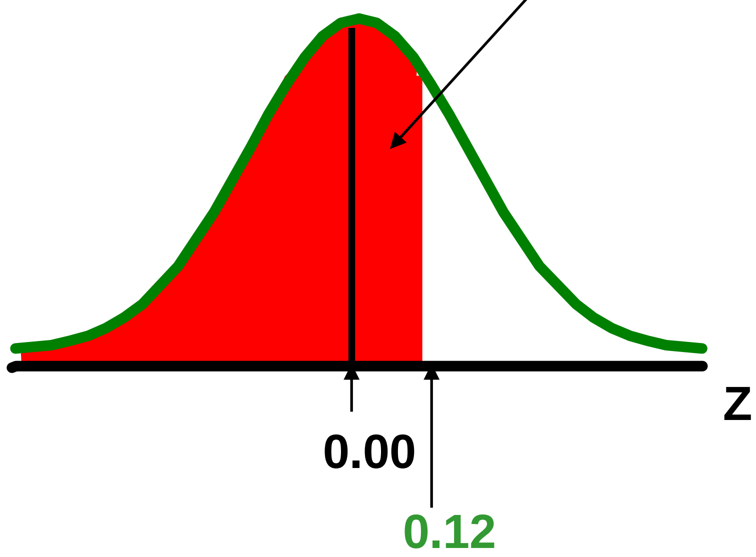
Solution: Finding P(Z < 0.12)

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

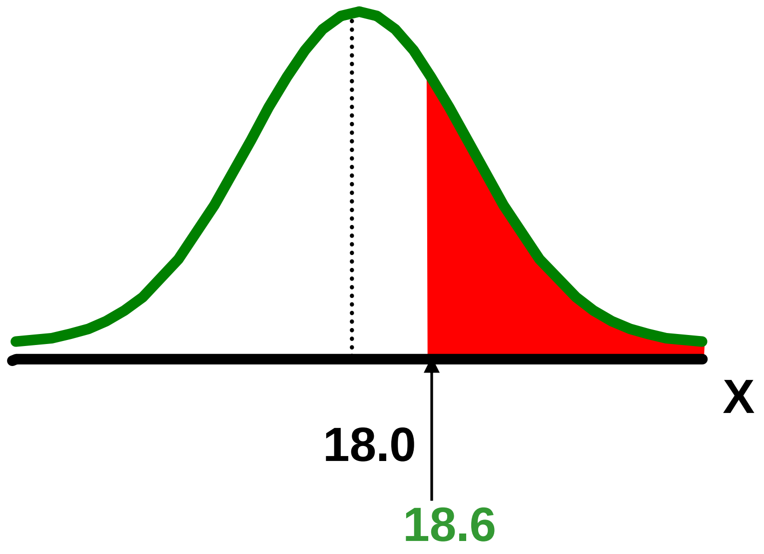
P(X < 18.6)
= P(Z < 0.12)

0.5478



Finding Normal Upper Tail Probabilities

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



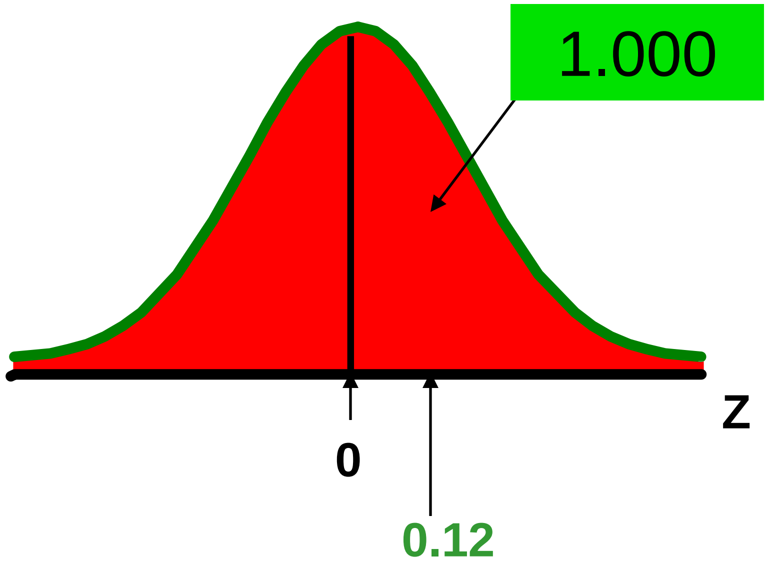
Finding Normal Upper Tail Probabilities

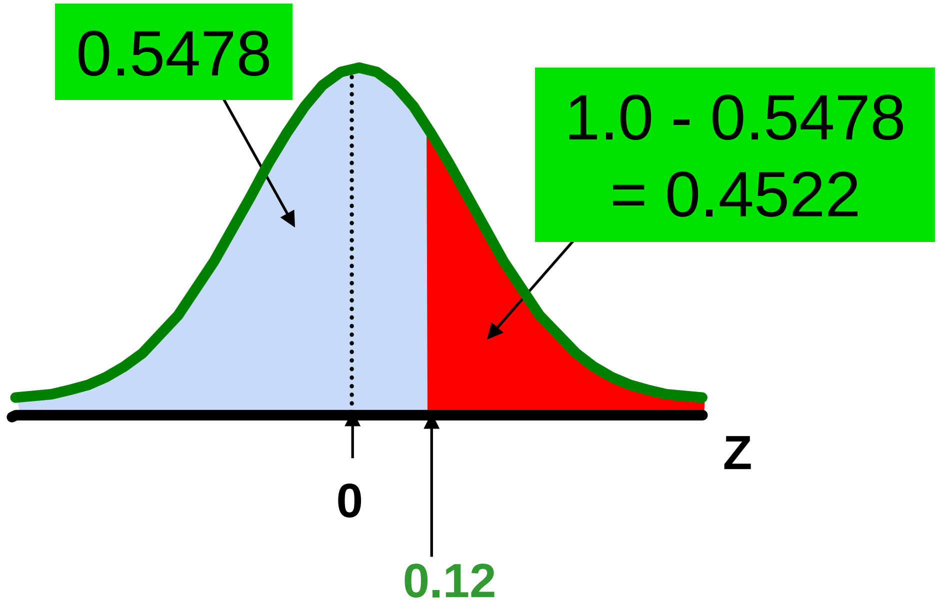
(continued)

- Now Find P(X > 18.6).

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

$$= 1.0 - 0.5478 = \boxed{0.4522}$$





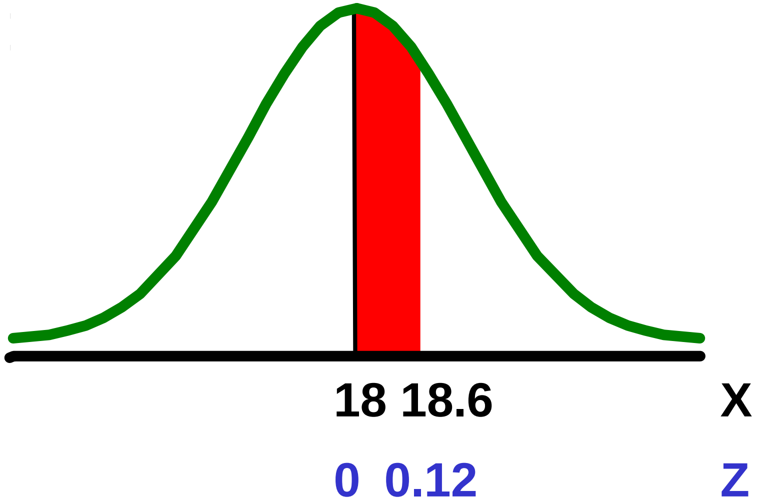
Finding a Normal Probability Between Two Values

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

Calculate Z-values:

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

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



P(18 < X < 18.6)

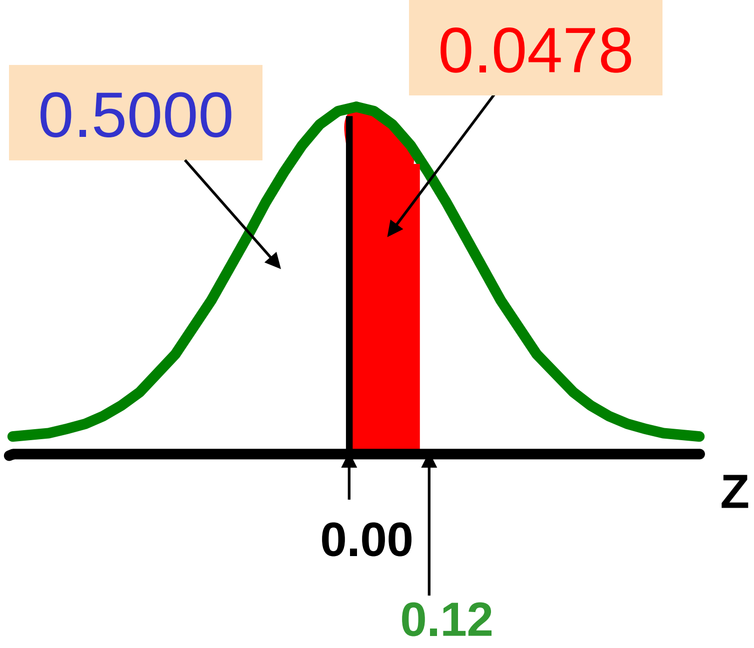
= P(0 < Z < 0.12)

Solution: Finding P(0 < Z < 0.12)

Standardized Normal Probability Table (Portion)

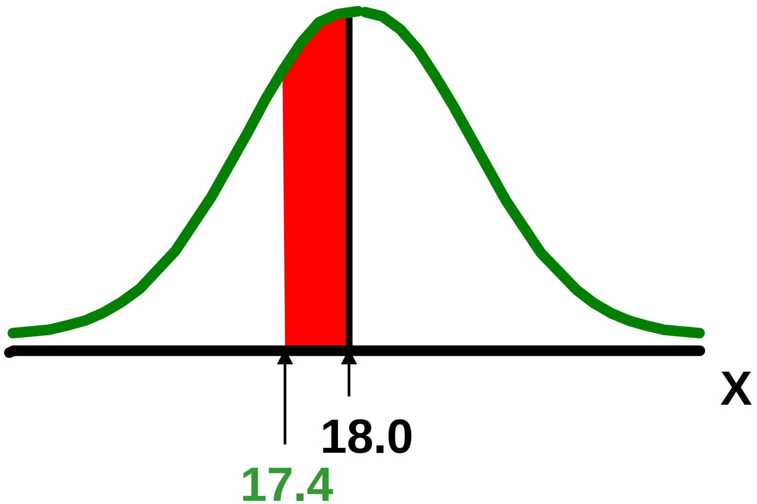
Z	.00	.01	.02
0.0	.5000	.5040	.5080
0.1	.5398	.5438	.5478
0.2	.5793	.5832	.5871
0.3	.6179	.6217	.6255

$P(18 < X < 18.6)$
$= P(0 < Z < 0.12)$
$= P(Z < 0.12) – P(Z \leq 0)$
$= 0.5478 - 0.5000 = 0.0478$



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):

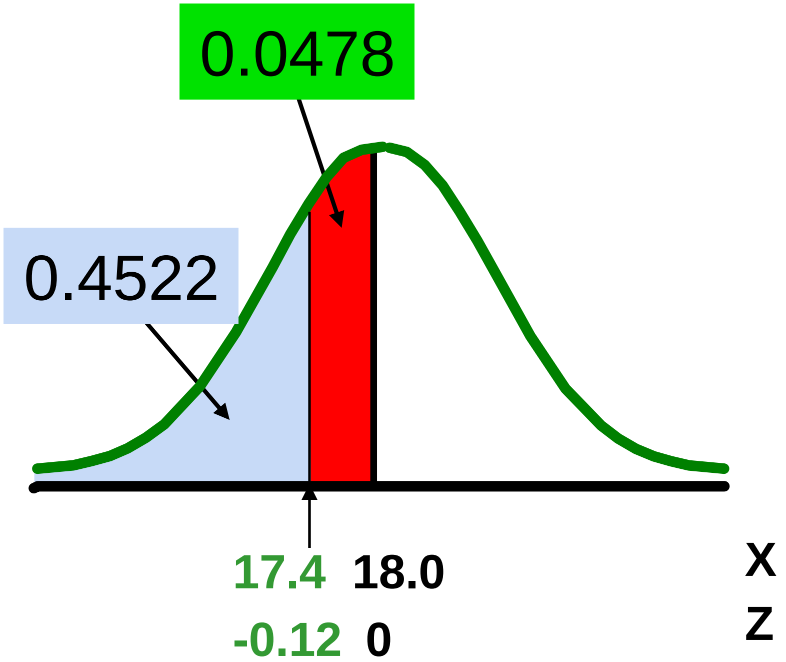
$P(17.4 < X < 18)$

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

$= P(Z < 0) – P(Z \le -0.12)$

$= 0.5000 - 0.4522 =$ **0.0478**

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



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:

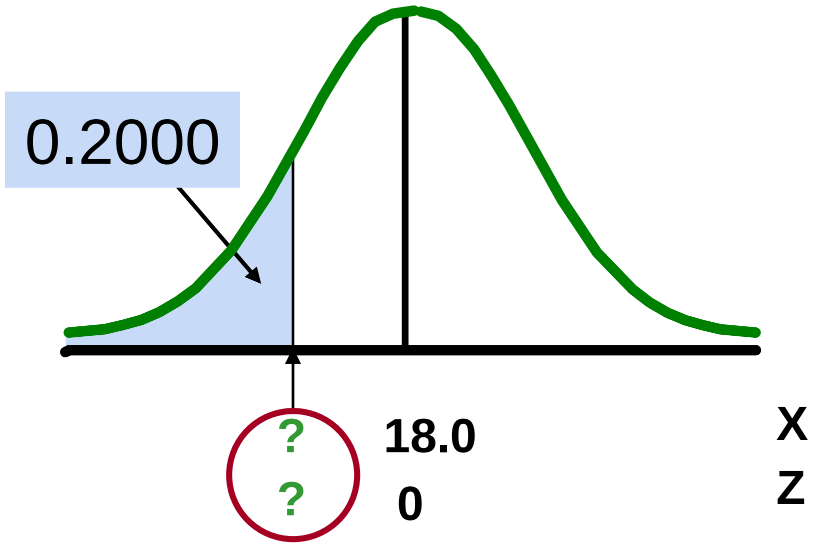
$$X = \mu + Z\sigma$$

Finding the X value for a Known Probability

(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.



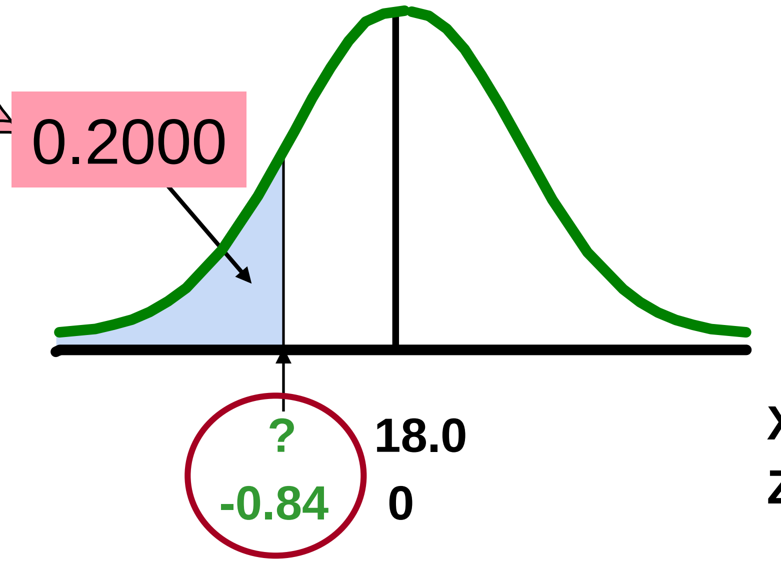
Find the Z value for 20% in the Lower Tail

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:

$$X = \mu + Z\sigma$$
$$= 18.0 + (-0.84)5.0$$
$$= 13.8$$

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

Chapter Summary

In this chapter we discussed:

- Computing probabilities from the normal distribution.
- Using the normal distribution to solve business problems.
- Using the normal probability plot to determine whether a set of data is approximately normally distributed.