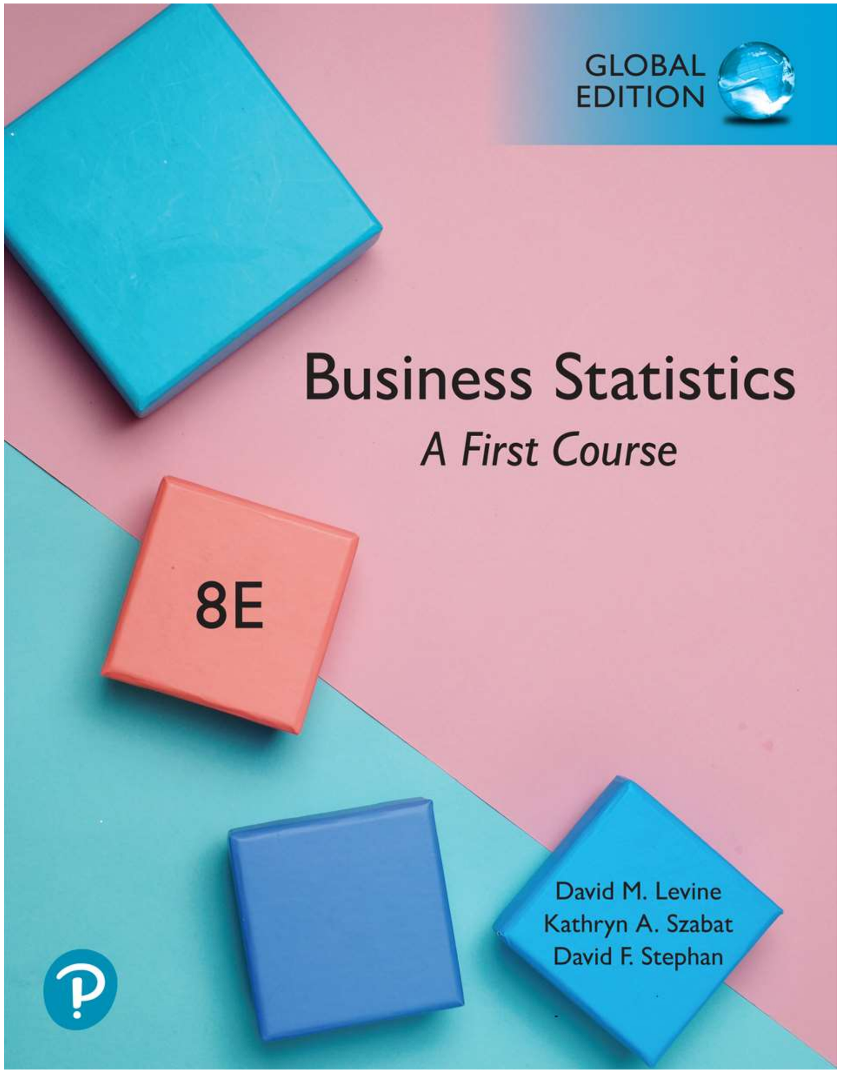


Chapter 1

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.)

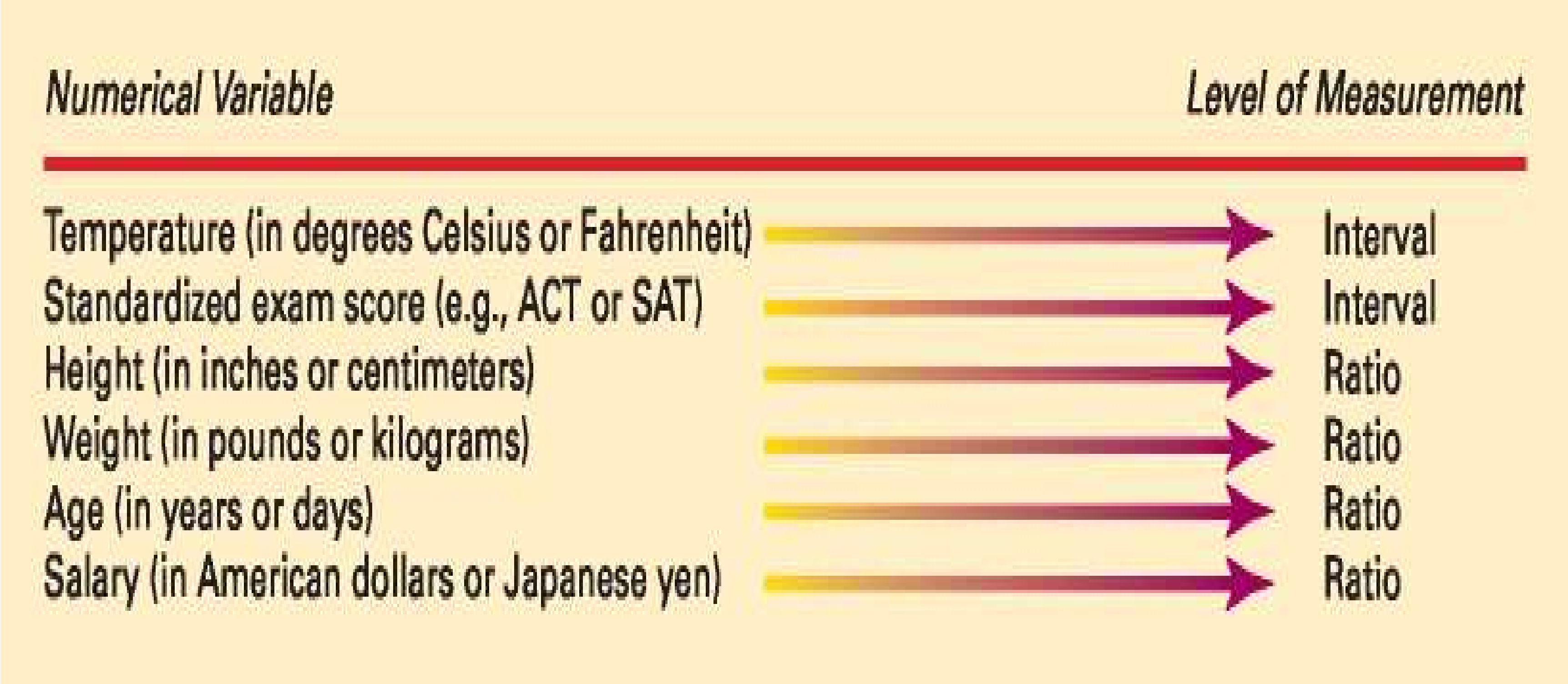
DCOVA

- 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

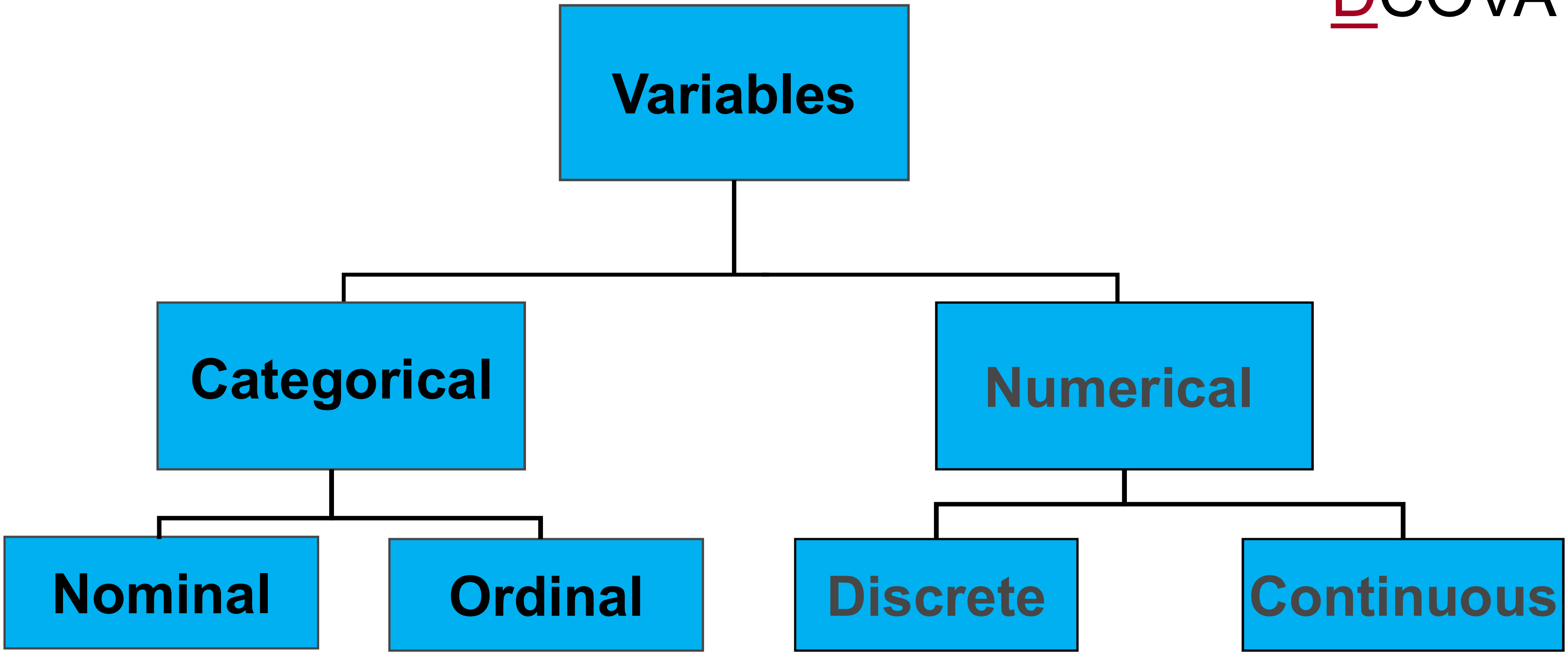
DCOVA



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



Variables

Categorical

Numerical

Nominal

Ordinal

Discrete

Continuous

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

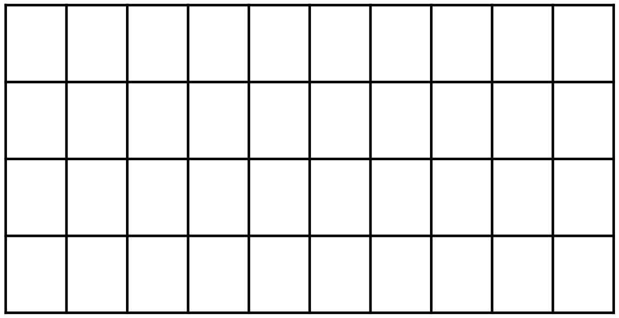
Population

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

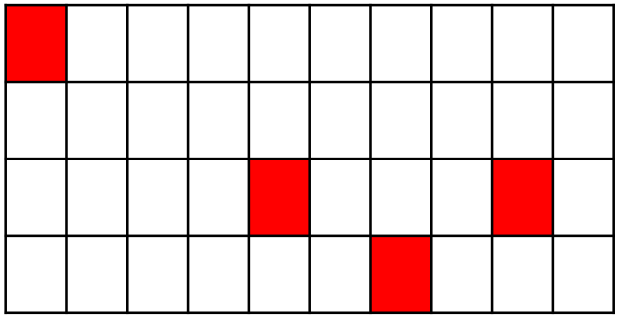
Sample

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.

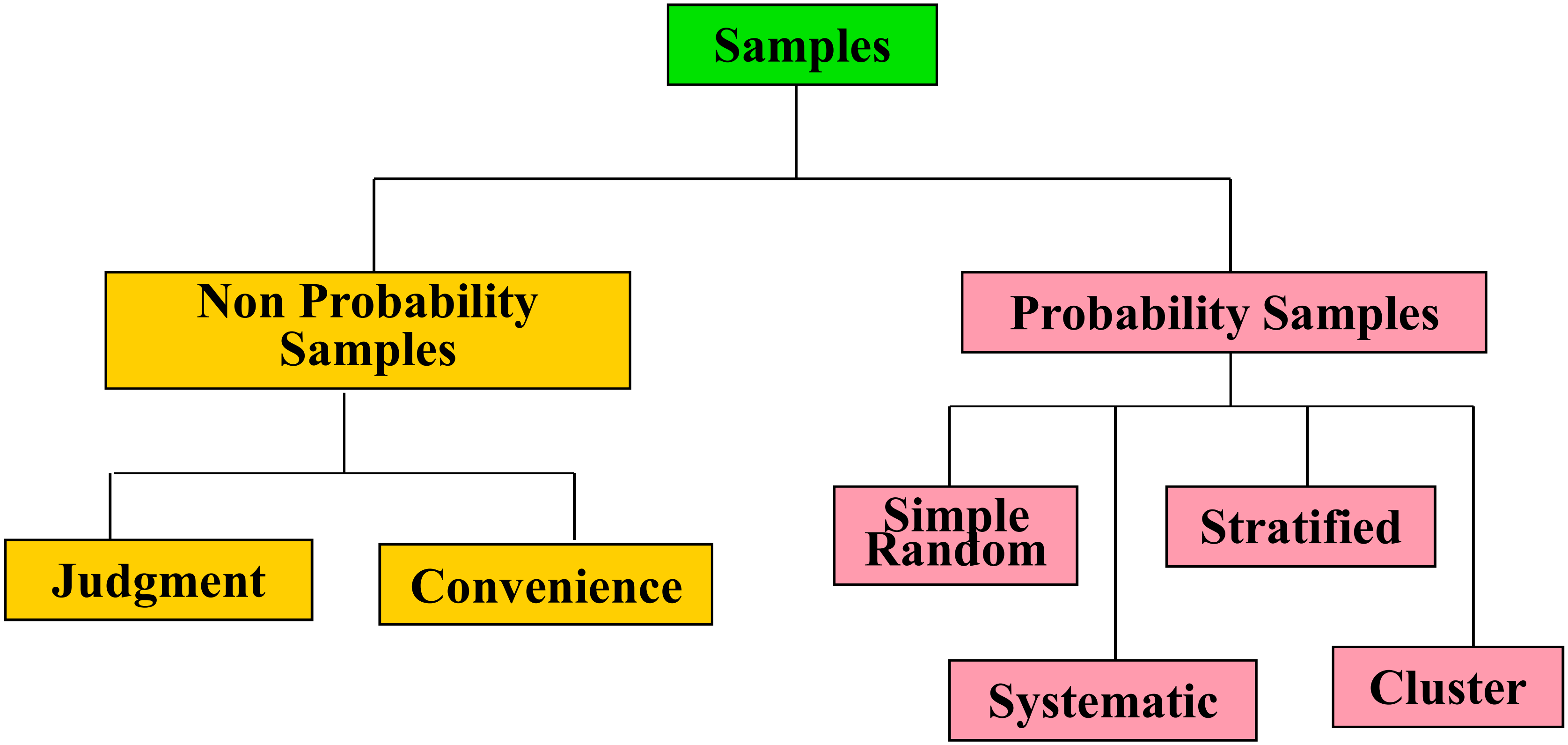
A Sampling Process Begins With A Sampling Frame

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

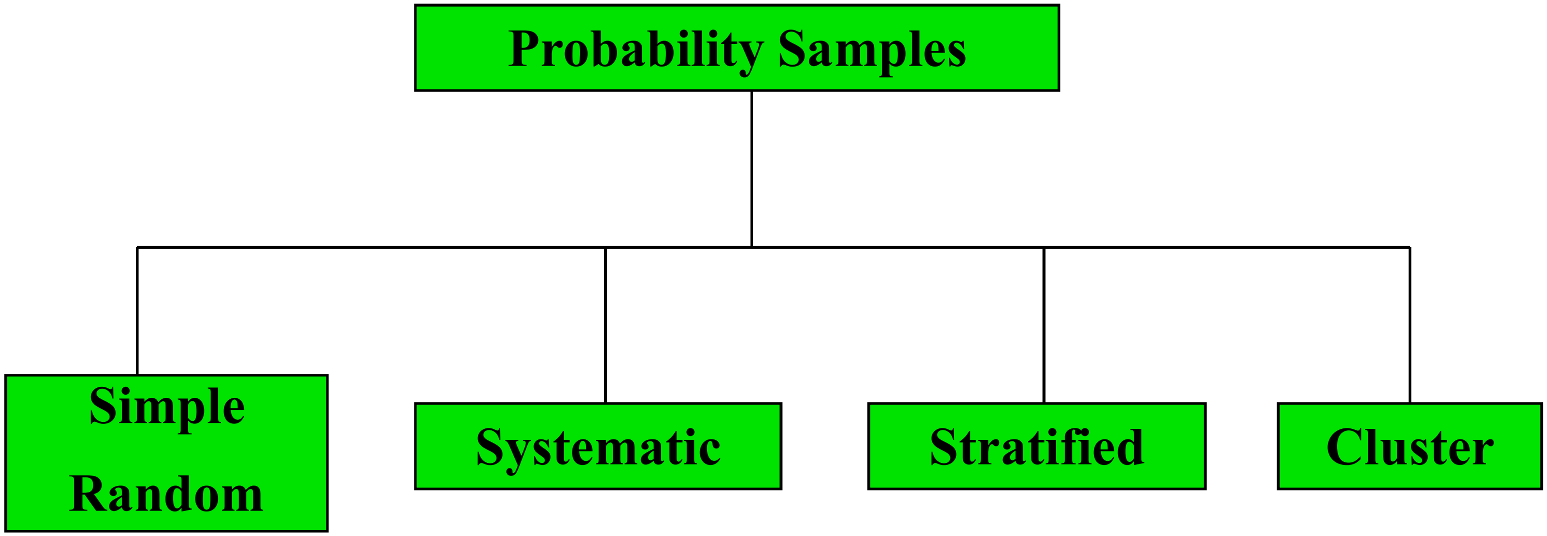
DCOVA

- 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

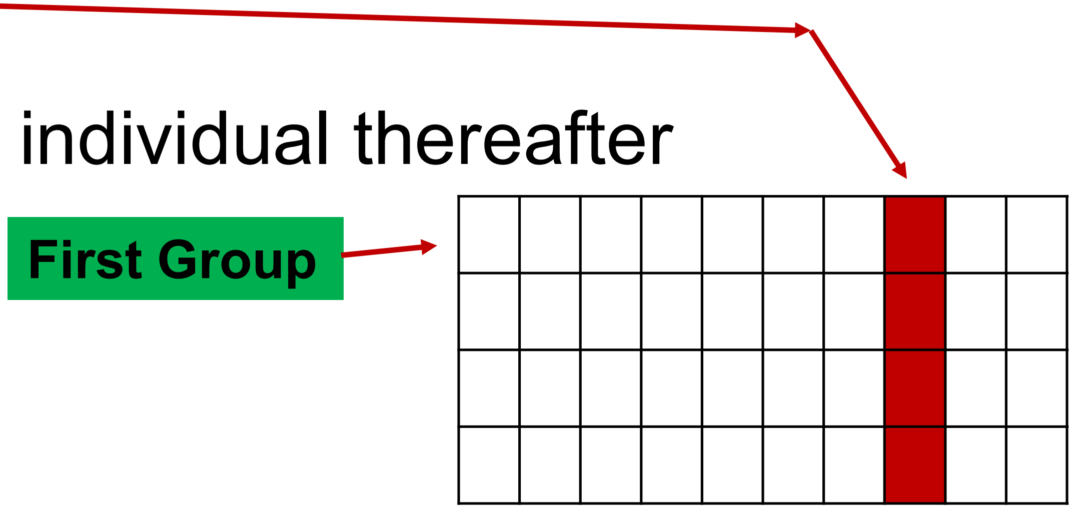
Item # 492
Item # 808
Item # 892 -- does not exist so ignore
Item # 435
Item # 779
Item # 002

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 kth individual thereafter

N = 40
n = 4
k = 10



Probability Sample: Stratified Sample

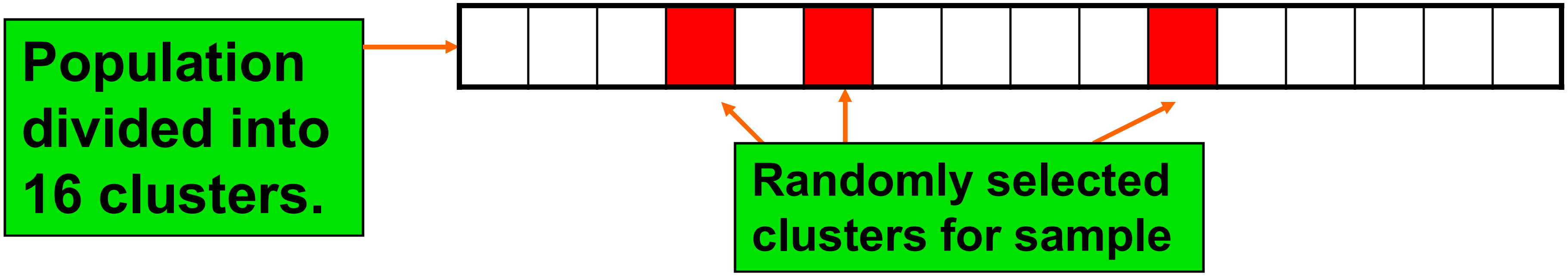
DCOVA

- 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.



Population divided into 16 clusters.

Randomly selected clusters for sample

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)*

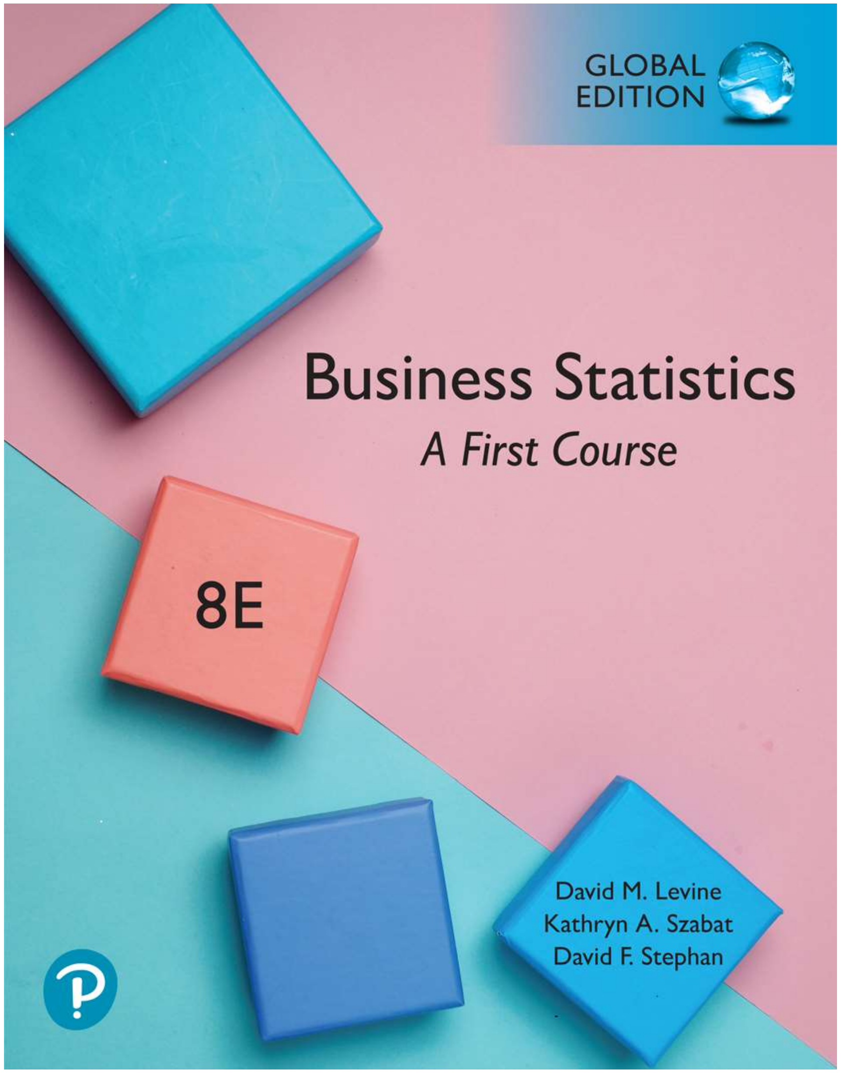
D<u>C</u>OVA

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

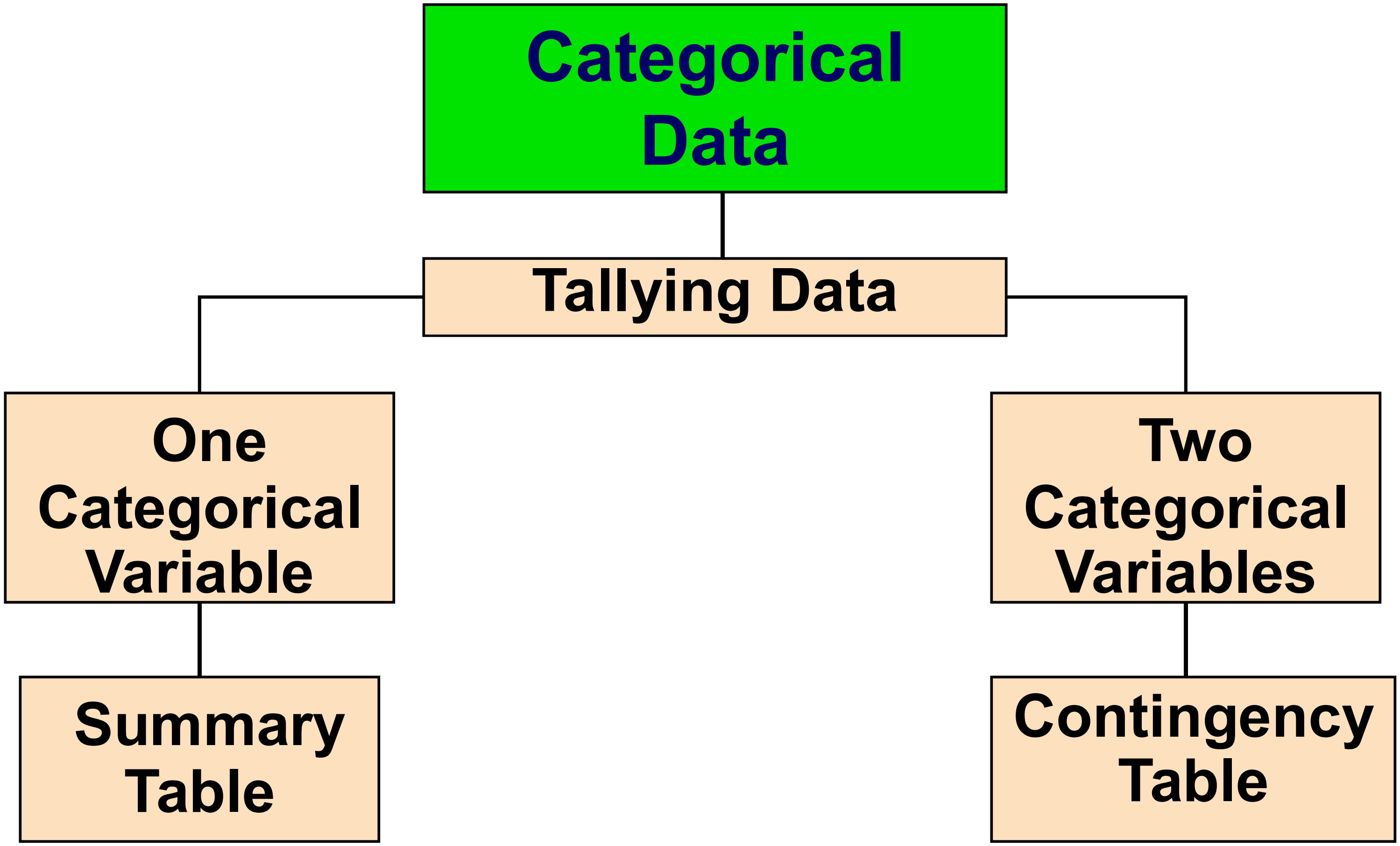
- Summaries both guide further exploration and sometimes facilitate decision making.

- 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



Categorical Data

Tallying Data

One Categorical Variable

Two Categorical Variables

Summary Table

Contingency Table

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

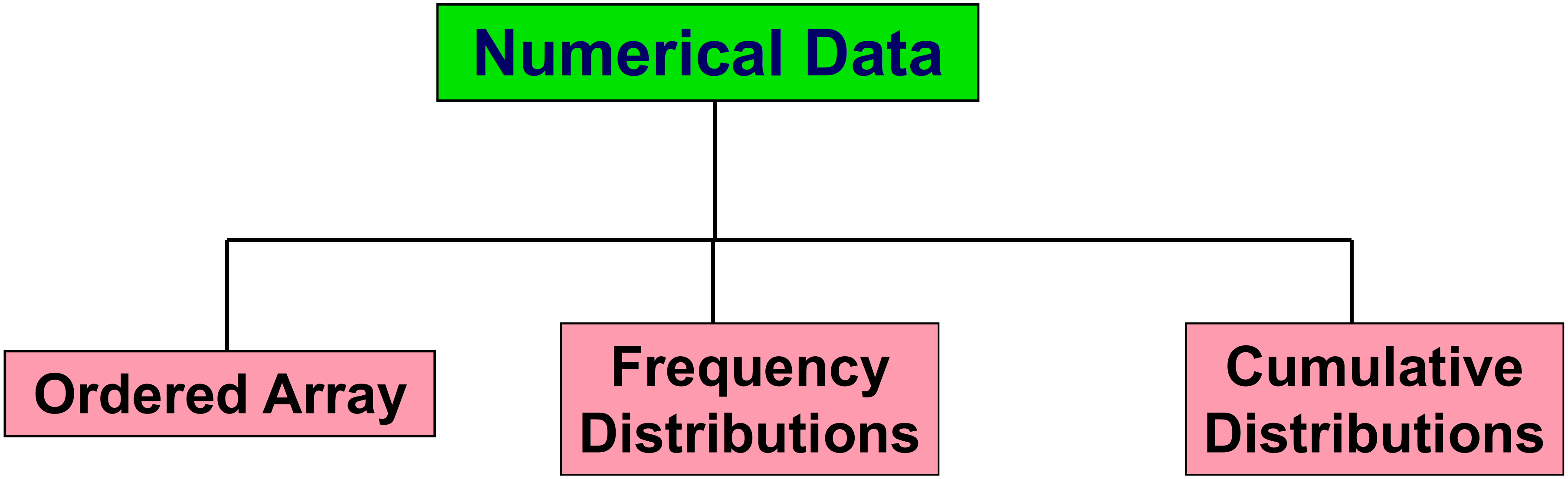
50.75% = 170 / 335
30.77% = 20 / 65

	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



Numerical Data

Ordered Array

Frequency Distributions

Cumulative Distributions

Organizing Numerical Data: Ordered Array

DCOVA

- An **ordered array** is a sequence of data, in rank order, from the smallest value to the largest value.
- Shows range (minimum value to maximum value).
- May help identify outliers (unusual observations).

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

Organizing Numerical Data: Frequency Distribution

DCOVA

- The **frequency distribution** is a summary table in which the data are arranged into numerically ordered classes.

- You must give attention to selecting the appropriate *number* of **class groupings** for the table, determining a suitable *width* of a class grouping, and establishing the *boundaries* of each class grouping to avoid overlapping.

- The number of classes depends on the number of values in the data. With a larger number of values, typically there are more classes. In general, a frequency distribution should have at least 5 but no more than 15 classes.

- To determine the **width of a class interval,** you divide the **range** (Highest value–Lowest value) of the data by the number of class groupings desired.

Organizing Numerical Data: Frequency Distribution Example

DCOVA

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

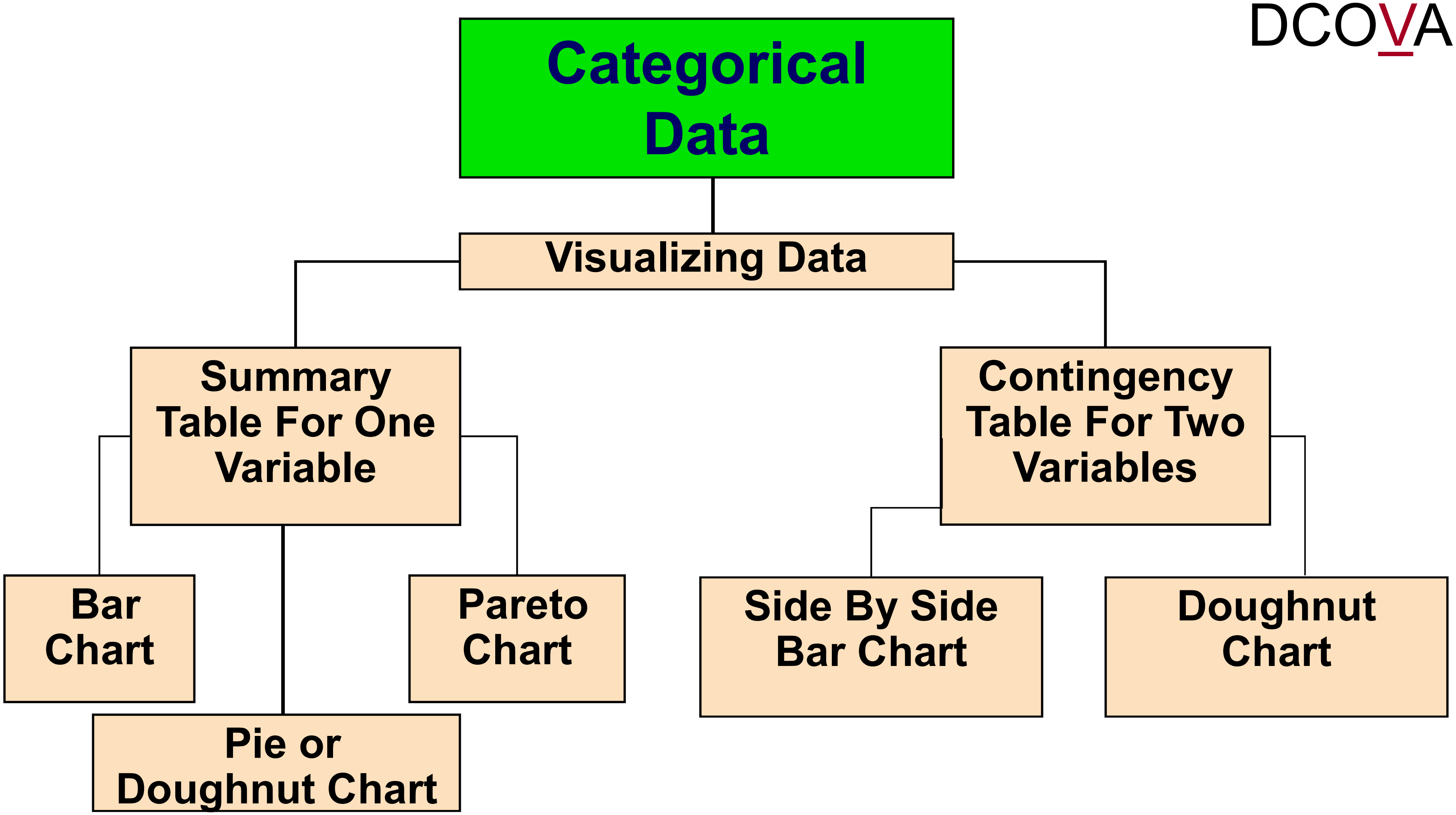
Frequency Distributions: Some Tips

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

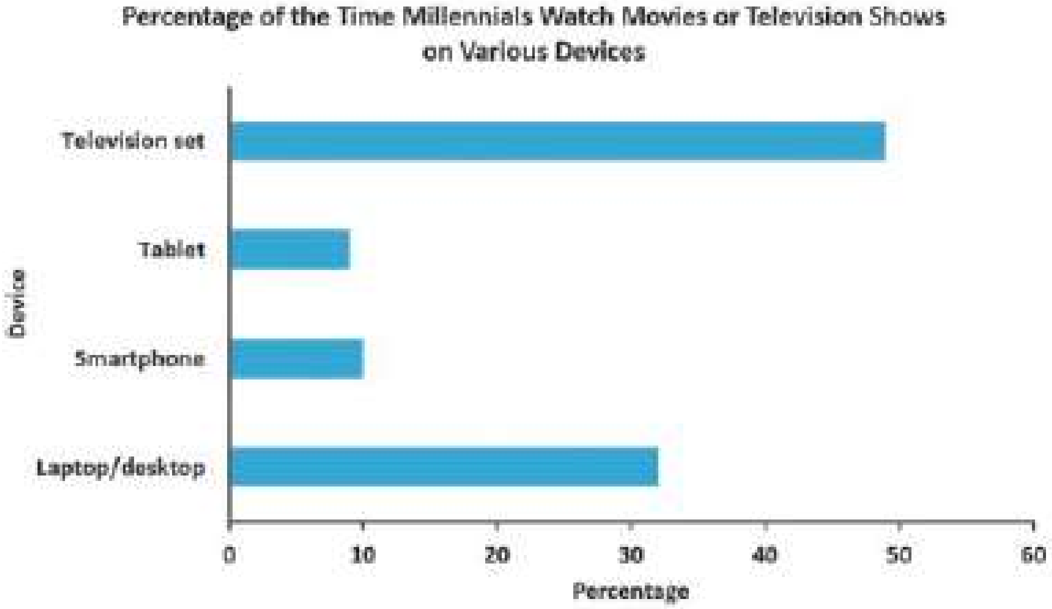


Visualizing Categorical Data: The Bar Chart

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%

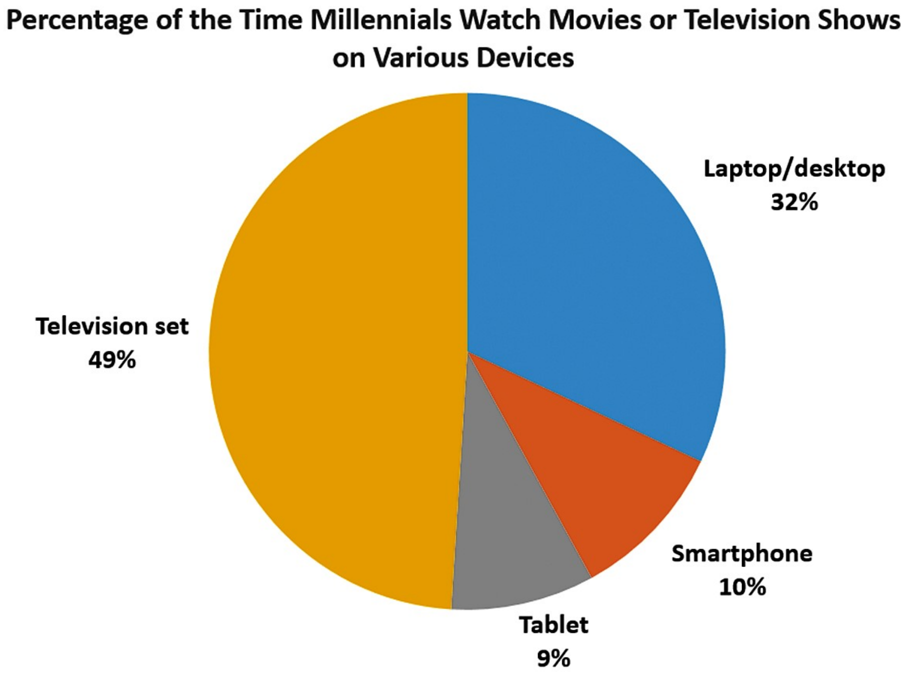


Visualizing Categorical Data: The Pie Chart

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%

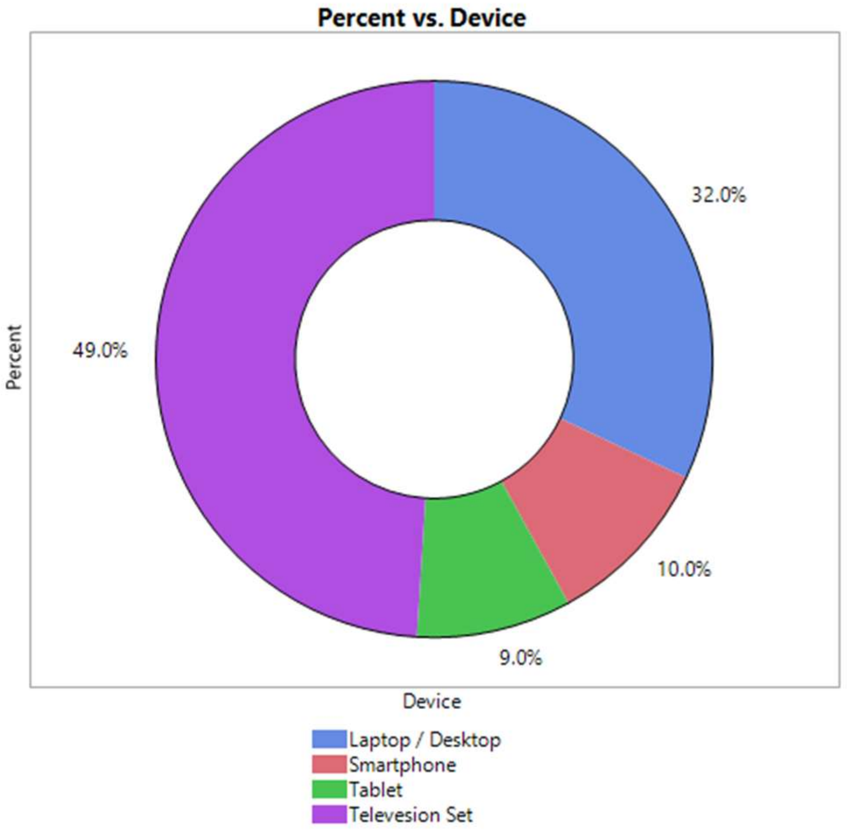


Visualizing Categorical Data: The Doughnut Chart

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

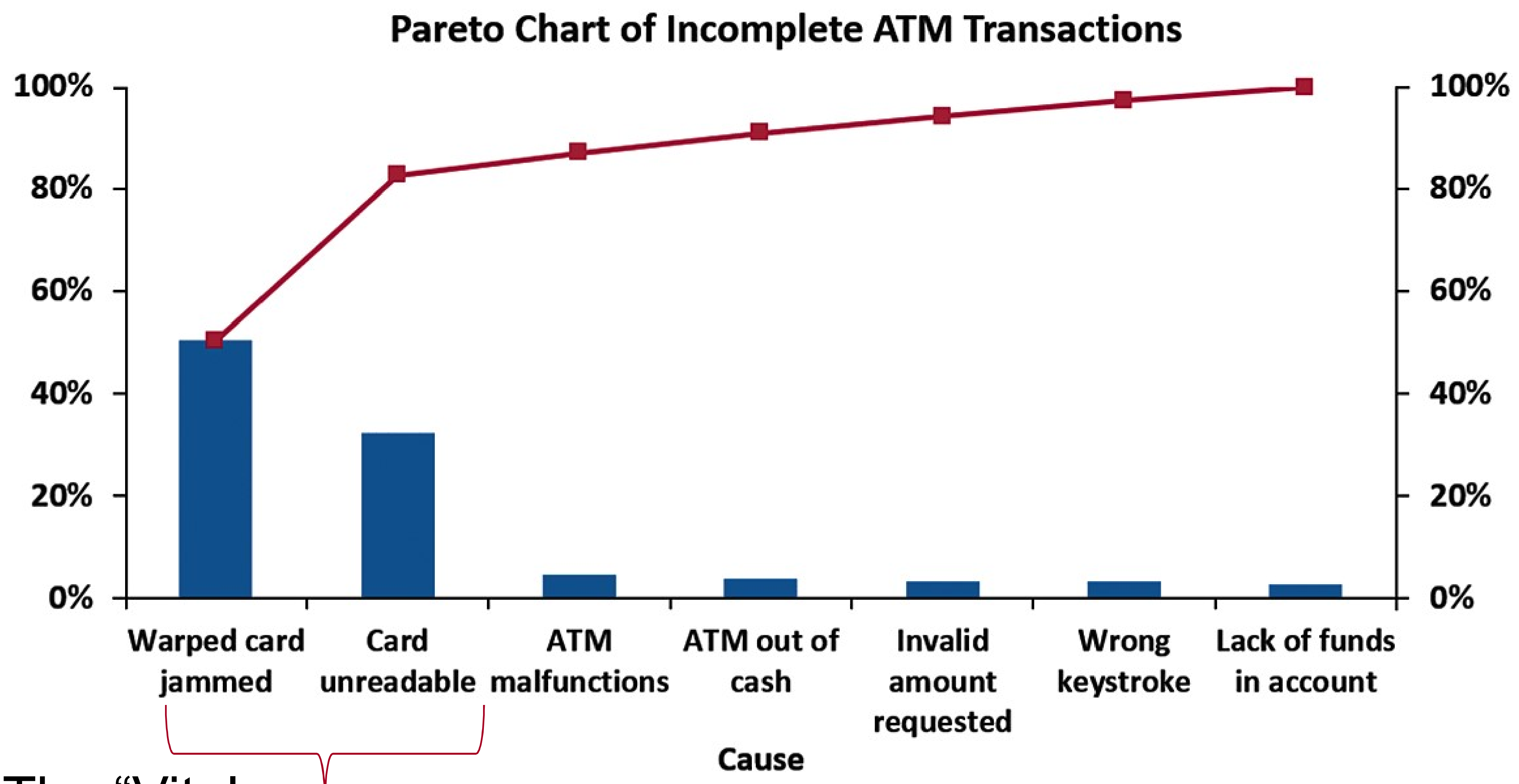
Ordered Summary Table For Causes Of Incomplete ATM Transactions

			Cumulative
Cause	**Frequency**	**Percent**	**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



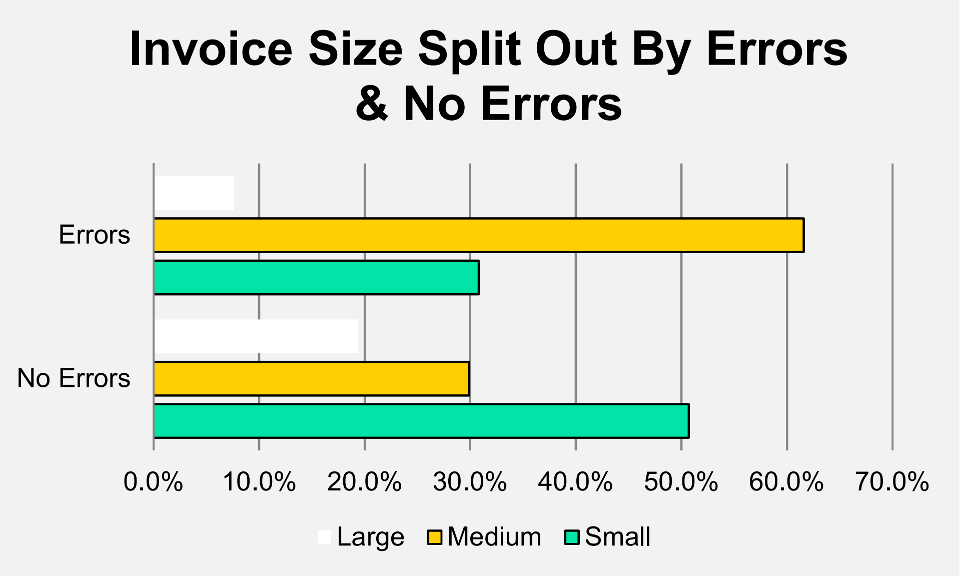
The "Vital Few"

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%



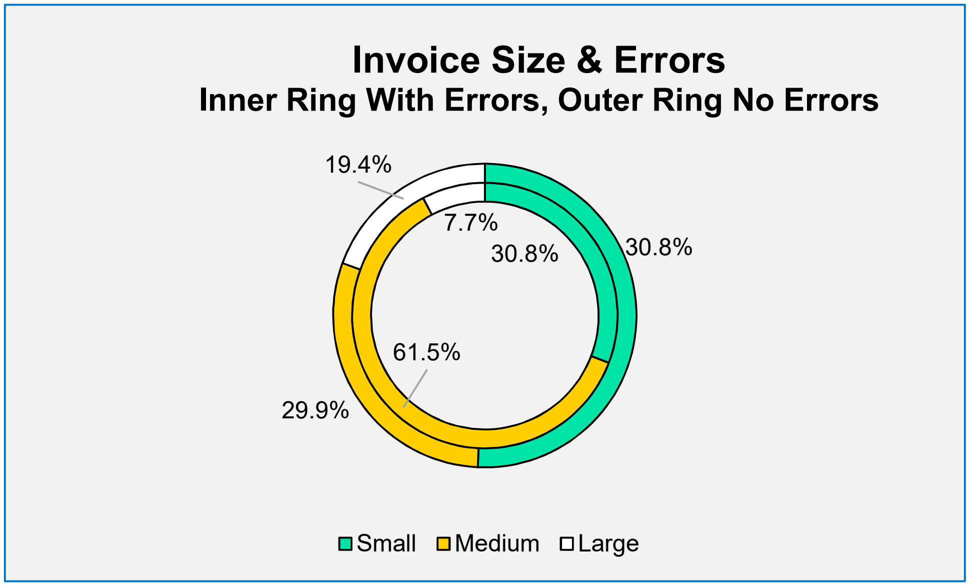
Invoices with errors are much more likely to be of medium size (61.5% vs 30.8% & 7.7%).

Visualizing Categorical Data: Doughnut Charts

DCOVA

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

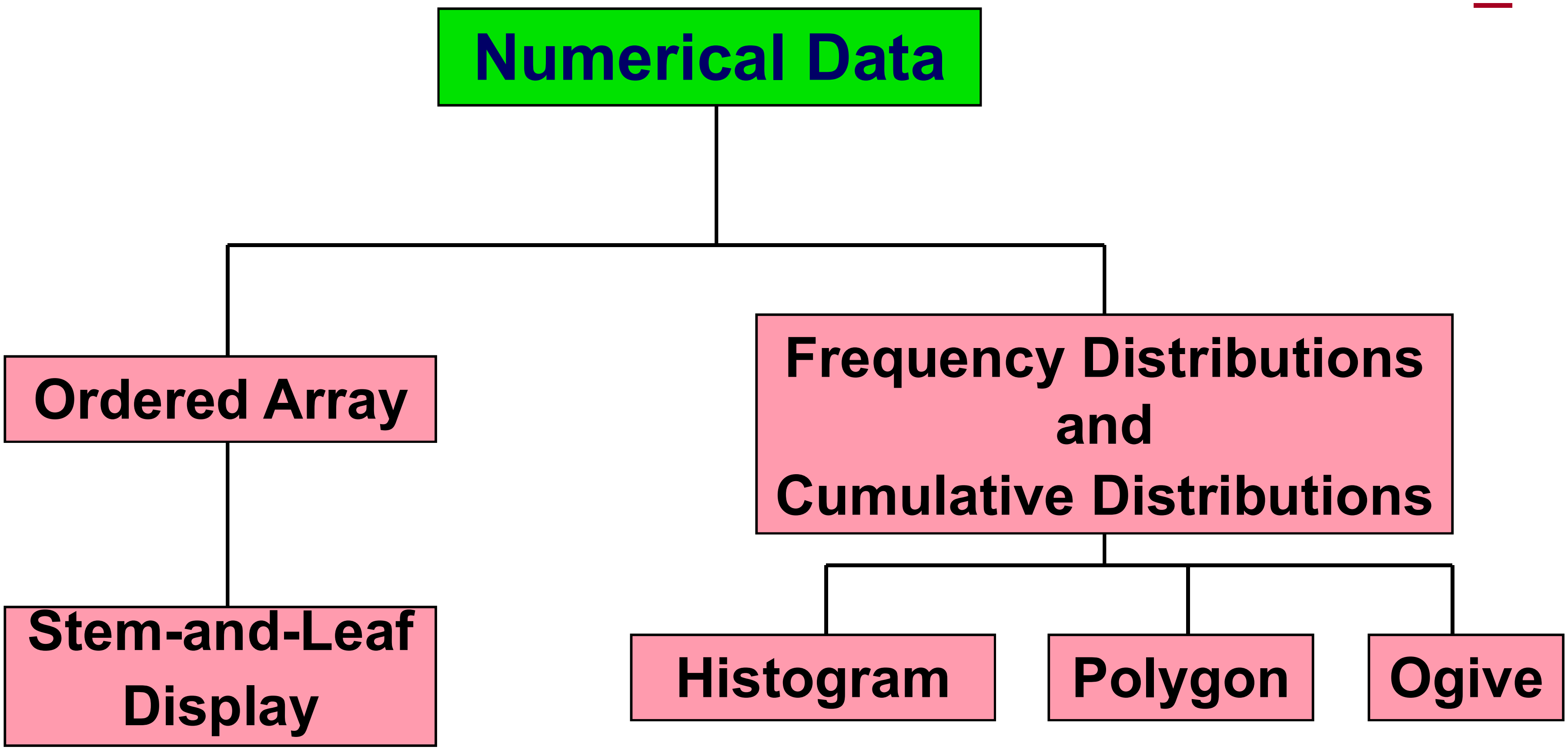
	No Errors	Errors	Total
Small Amount	50.75%	30.77%	47.50%
Medium Amount	29.85%	61.54%	35.00%
Large Amount	19.40%	7.69%	17.50%
Total	100.0%	100.0%	100.0%



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**.

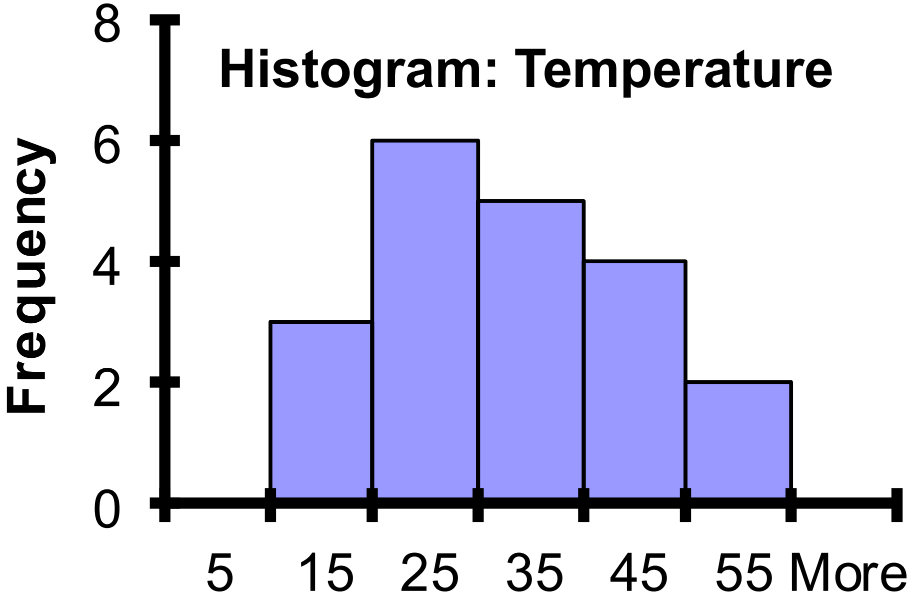
- The height of the bars represent the frequency, relative frequency, or percentage.

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.

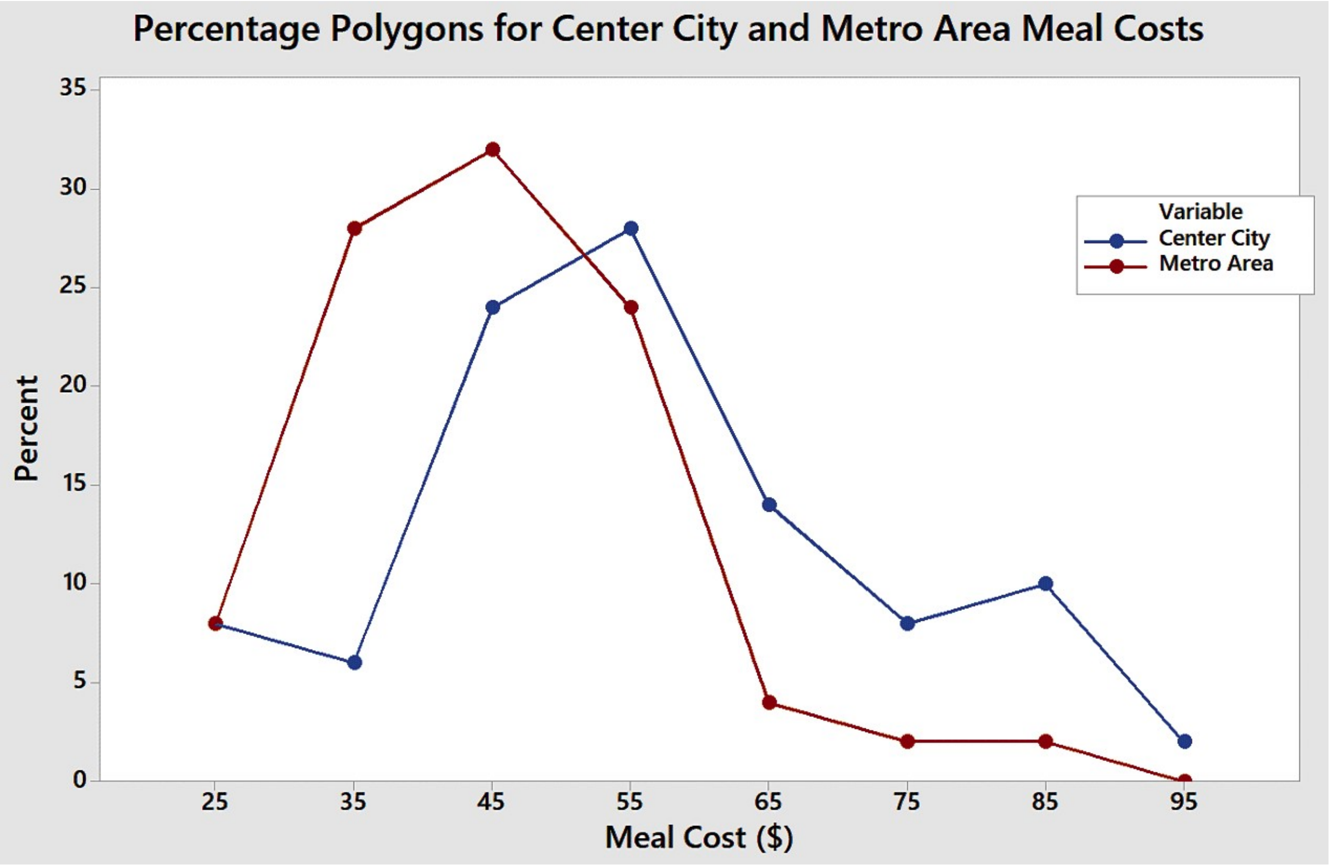
- The **cumulative percentage polygon, or ogive,** displays the variable of interest along the *X* axis, and the cumulative percentages along the *Y* axis.

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

Visualizing Numerical Data: The Frequency Polygon

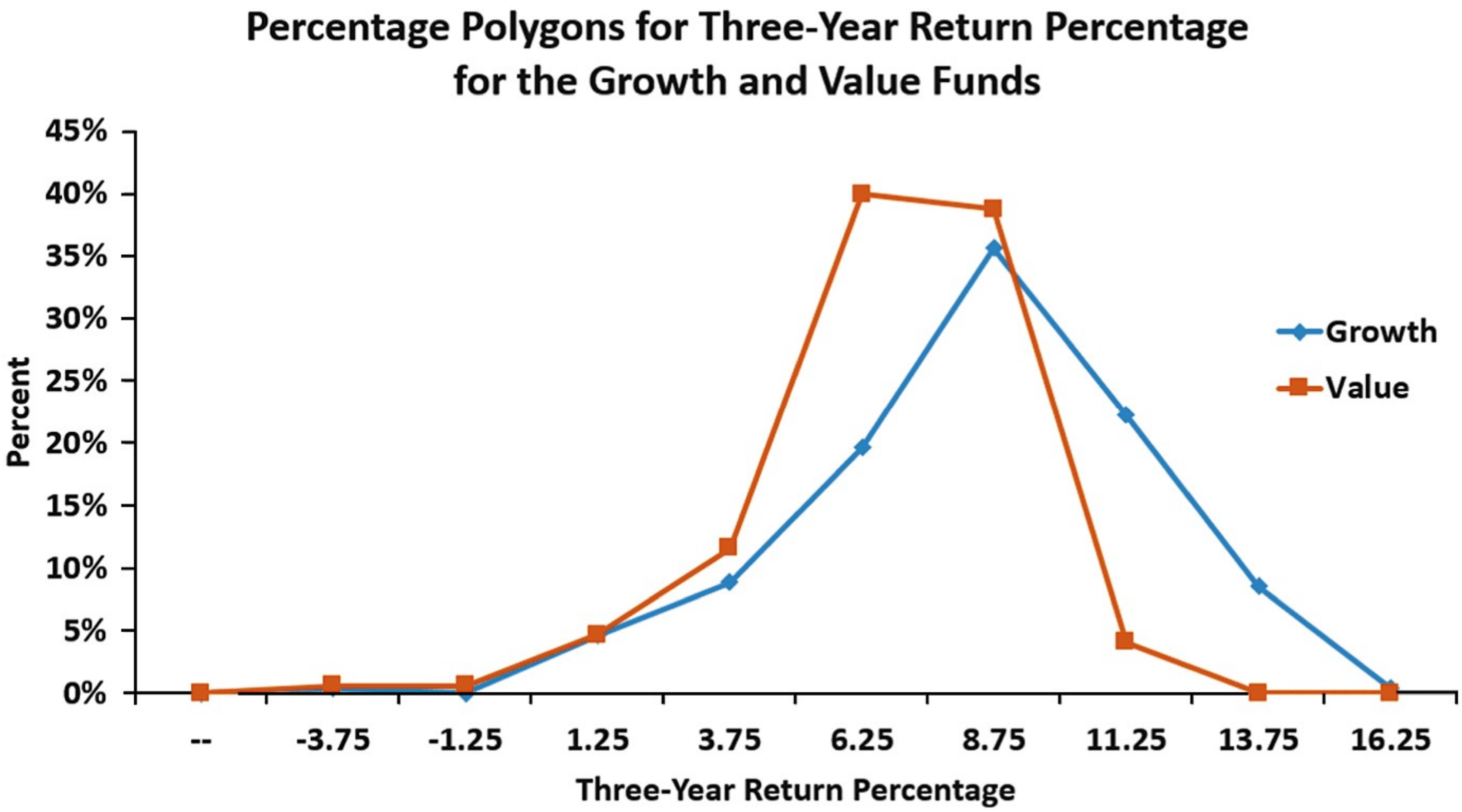
DCOVA

Useful When Comparing Two or More Groups



Visualizing Numerical Data: The Percentage Polygon

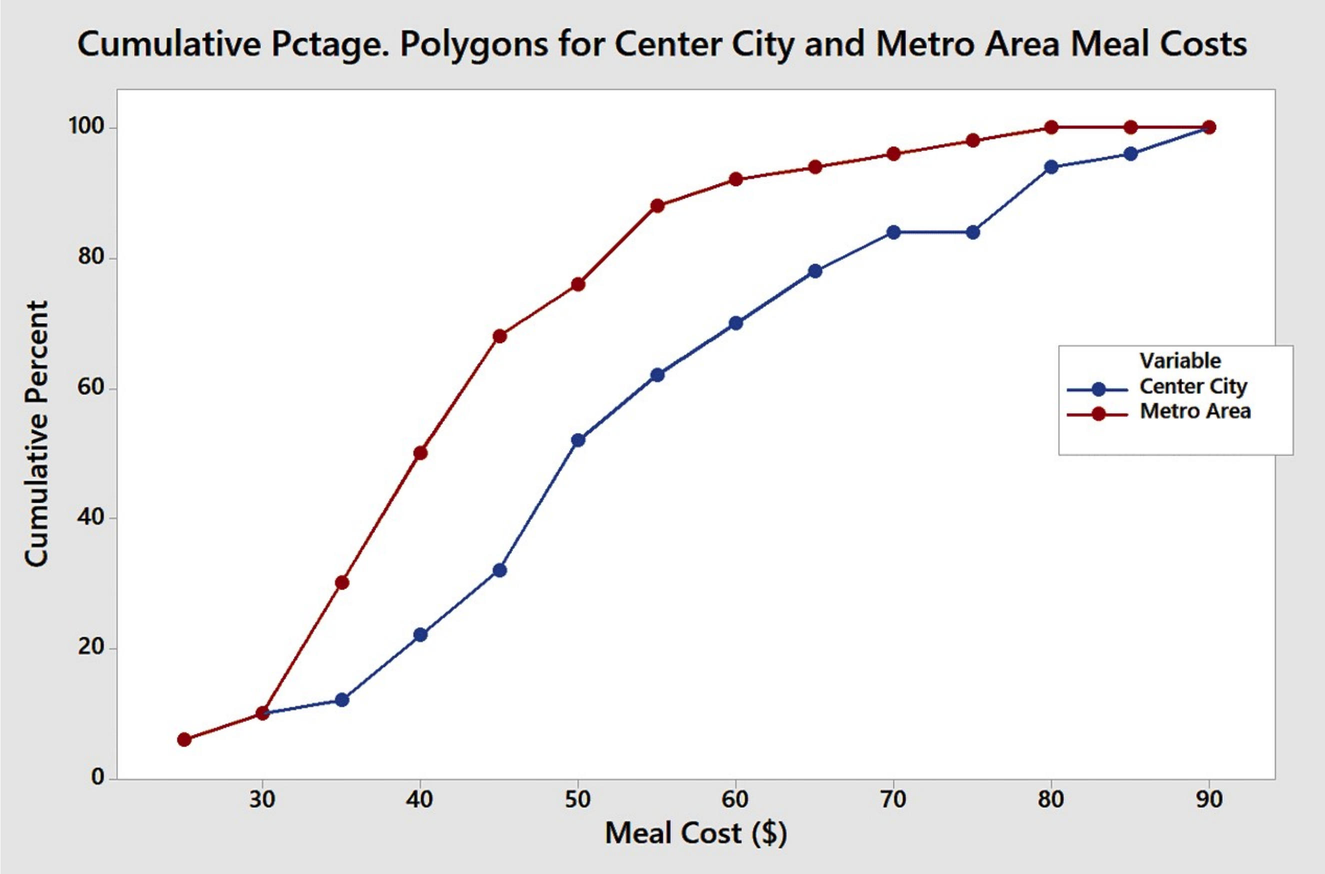
DCOVA



Visualizing Numerical Data: The Cumulative Percentage Polygon (Ogive)

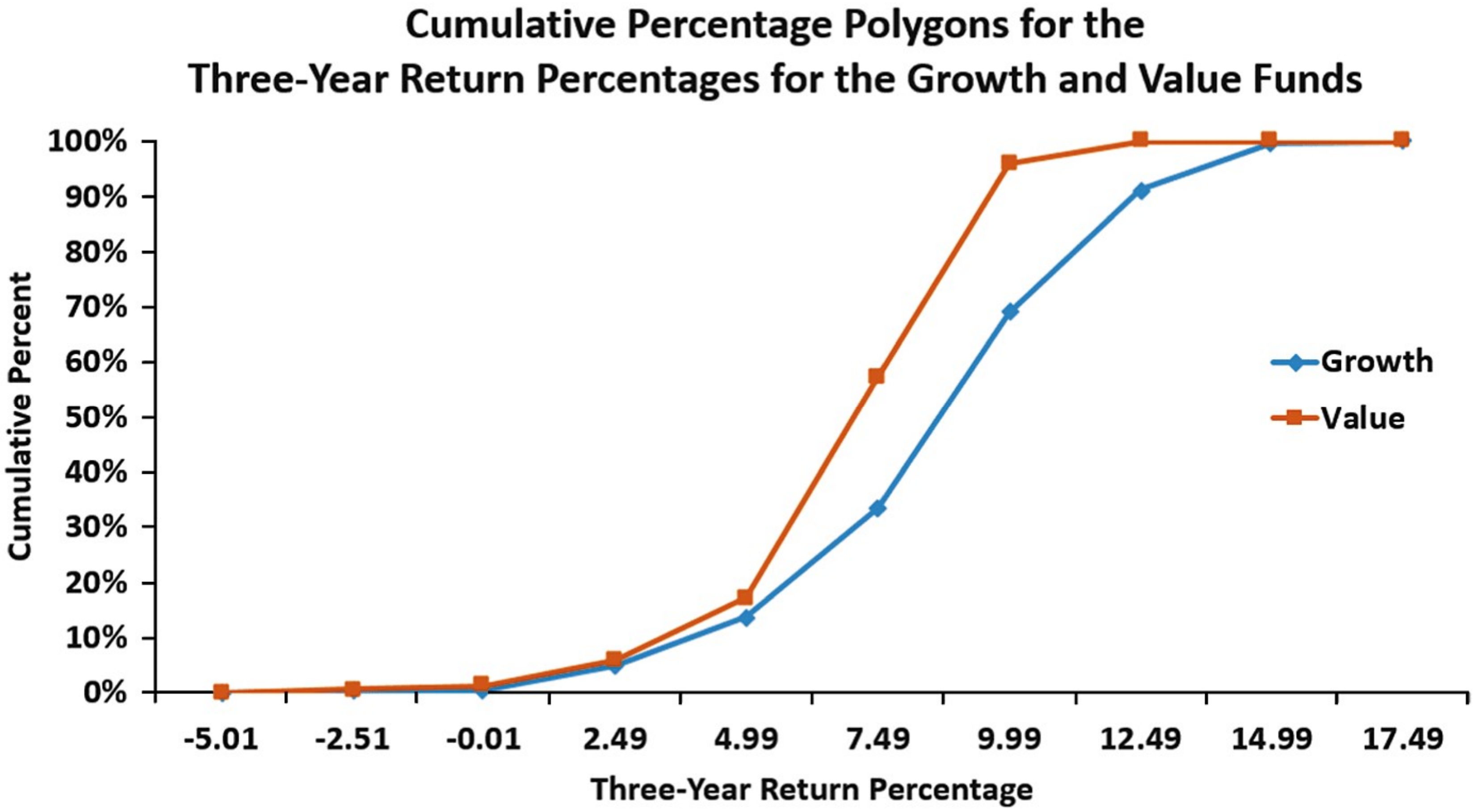
DCOVA

Useful When Comparing Two or More Groups



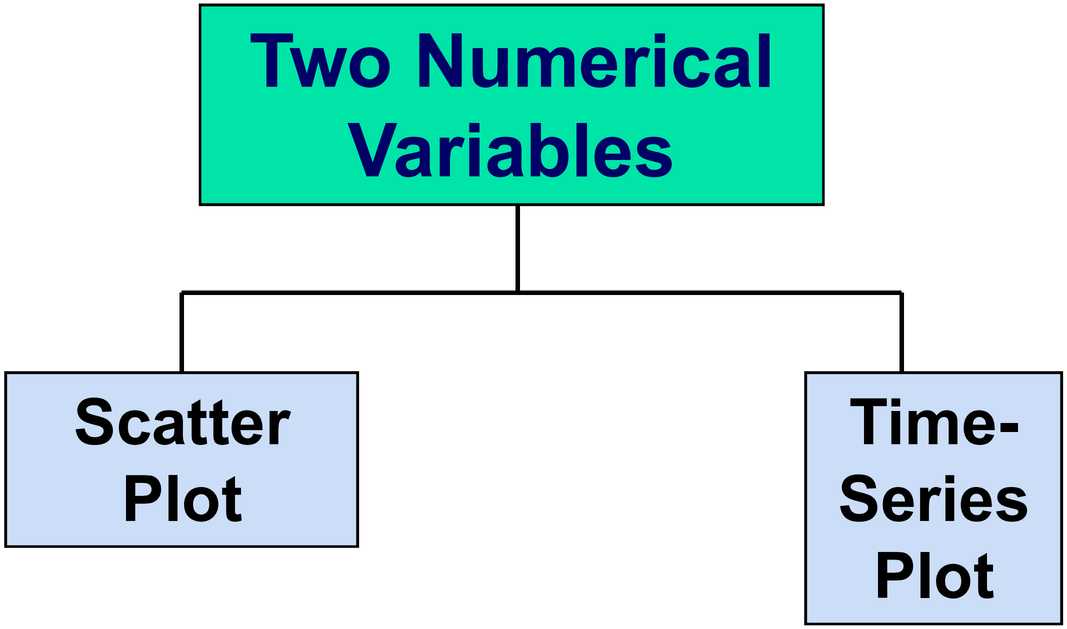
Visualizing Numerical Data: The Cumulative Percentage Polygon (Ogive)

DCOVA



Visualizing Two Numerical Variables By Using Graphical Displays

DCOVA



Visualizing Two Numerical Variables: The Scatter Plot

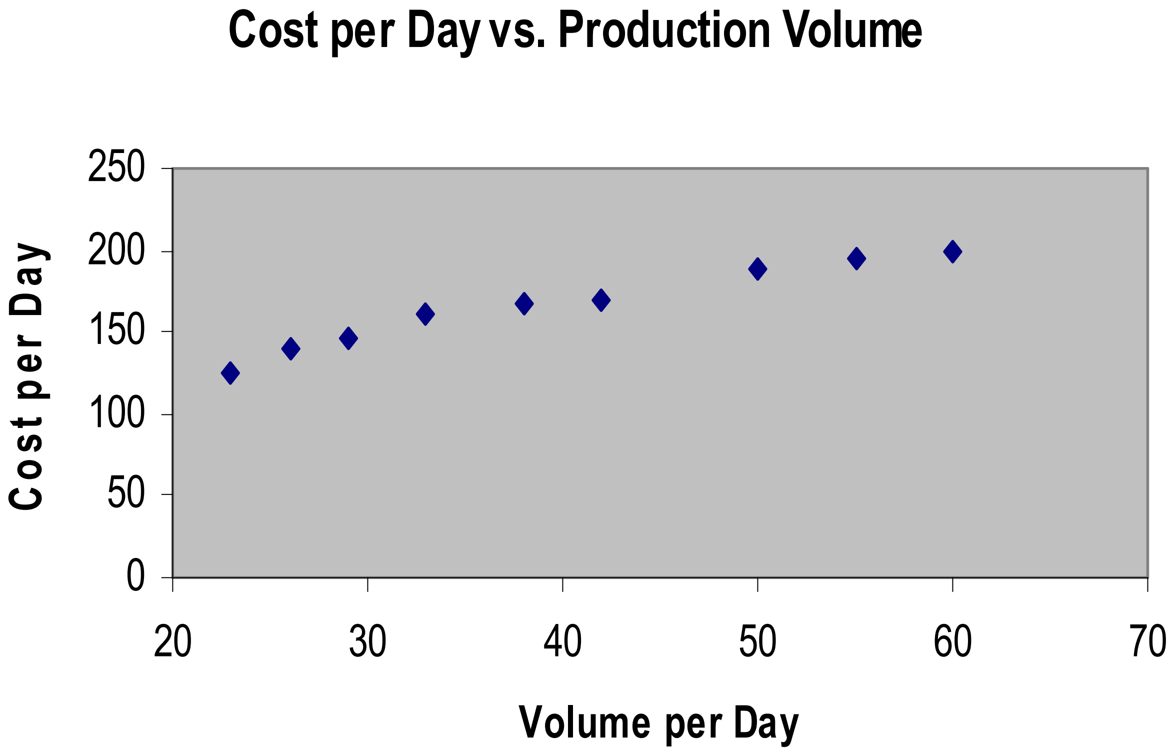
DCOVA

- **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

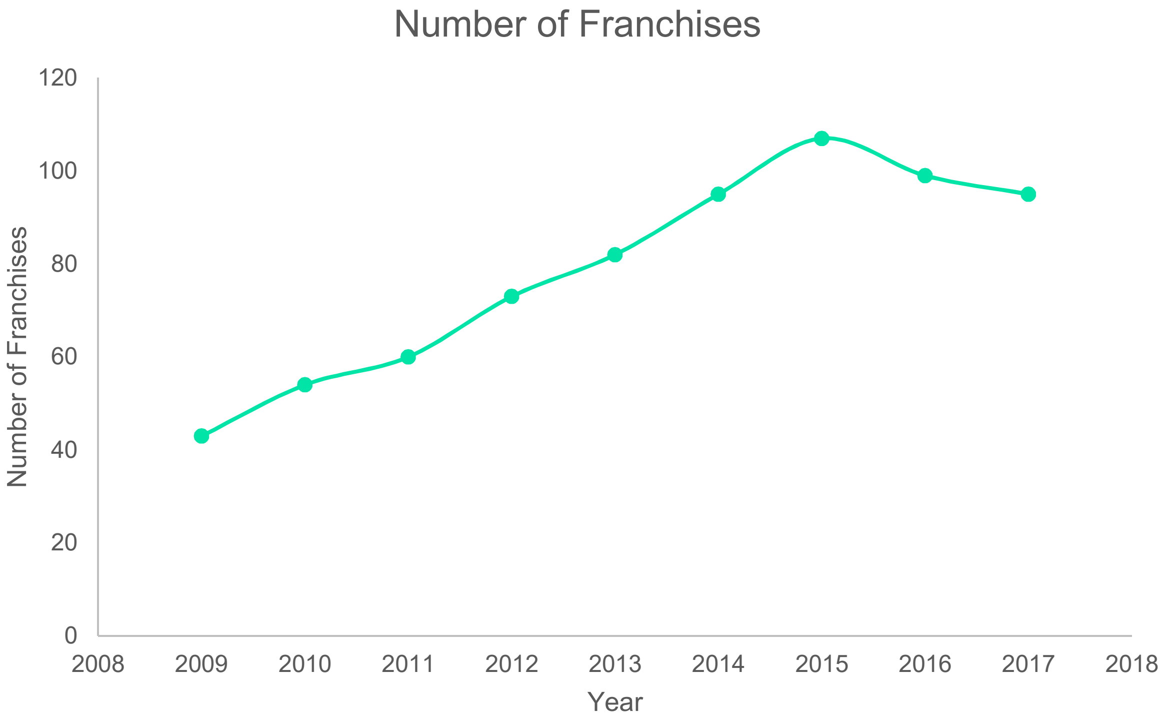
- A Time-Series Plot is used to study patterns in the values of a numeric variable over time.

- 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

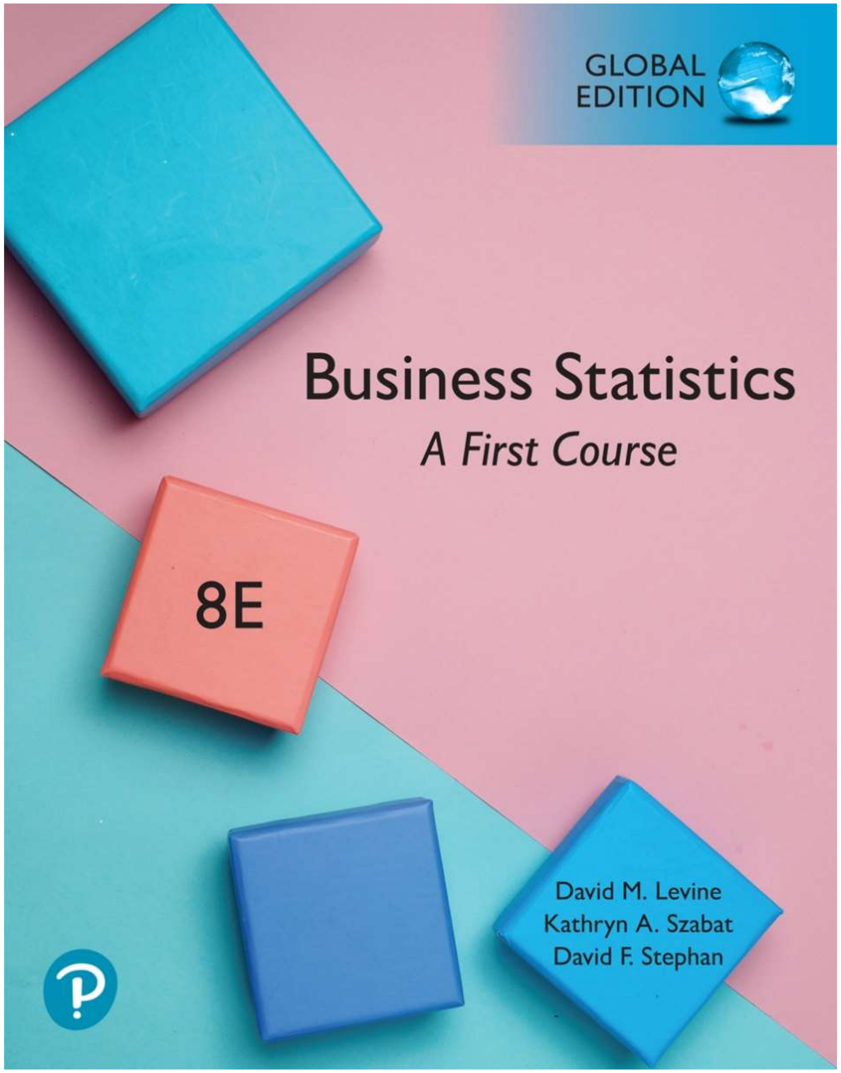
Year	Number of Franchises
2009	43
2010	54
2011	60
2012	73
2013	82
2014	95
2015	107
2016	99
2017	95



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:

Pronounced X-bar

The i^{th} value

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

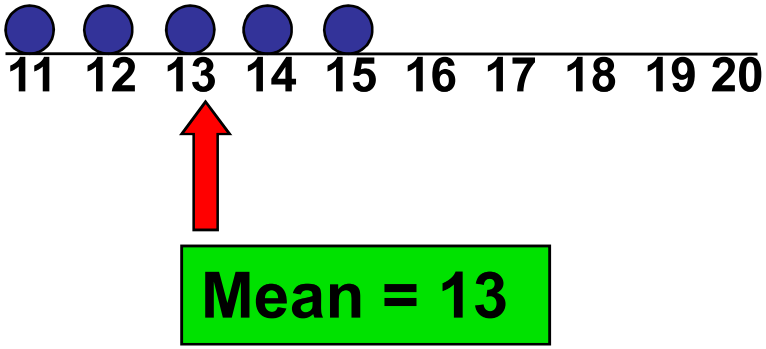
Sample size

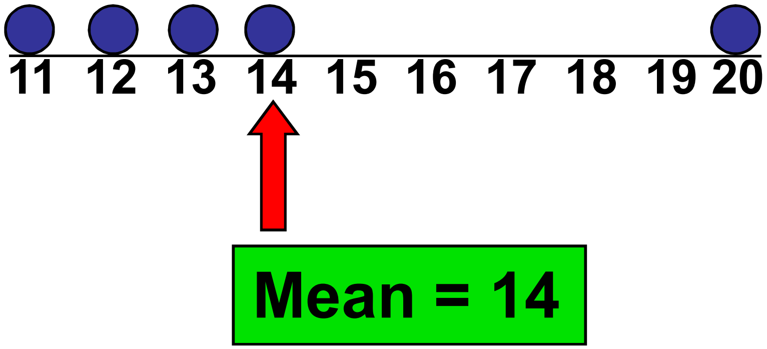
Observed values

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).





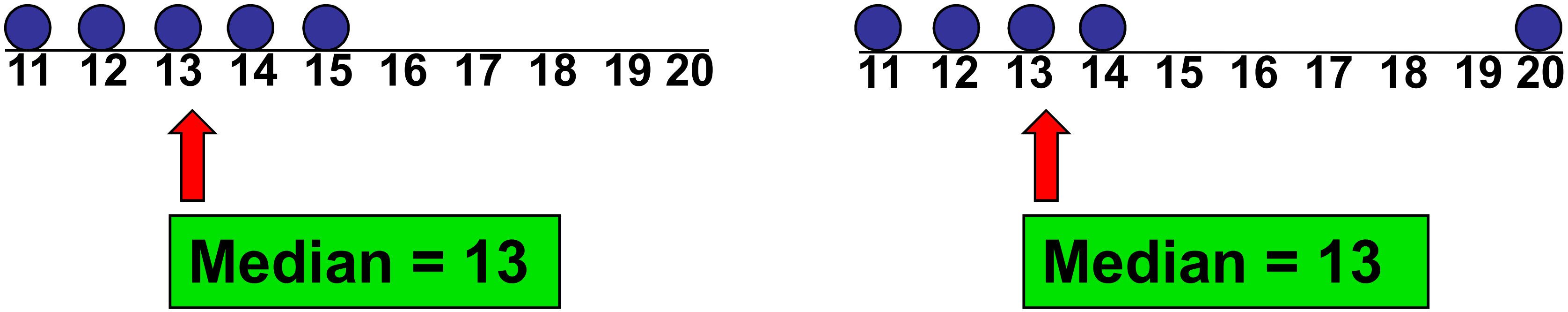
$$\frac{11+12+13+14+15}{5} = \frac{65}{5} = 13$$

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

Measures of Central Tendency: The Median

DCOVA

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



Median = 13

Median = 13

- 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.

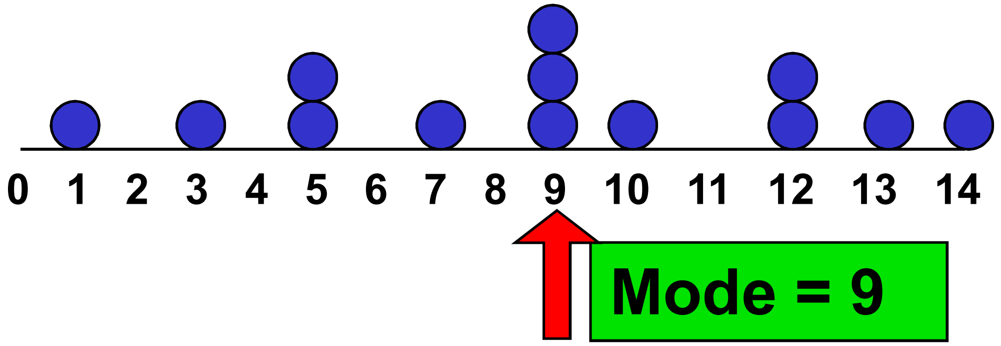
- If the number of values is even, the median is the average of the two middle numbers.

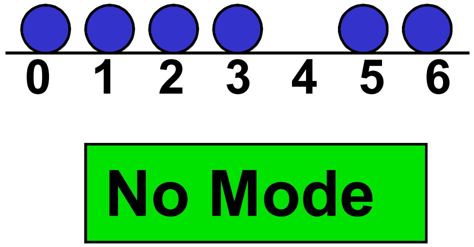
Note that $\frac{n+1}{2}$ is not the *value* of the median, only the *position* of the median in the ranked data.

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**

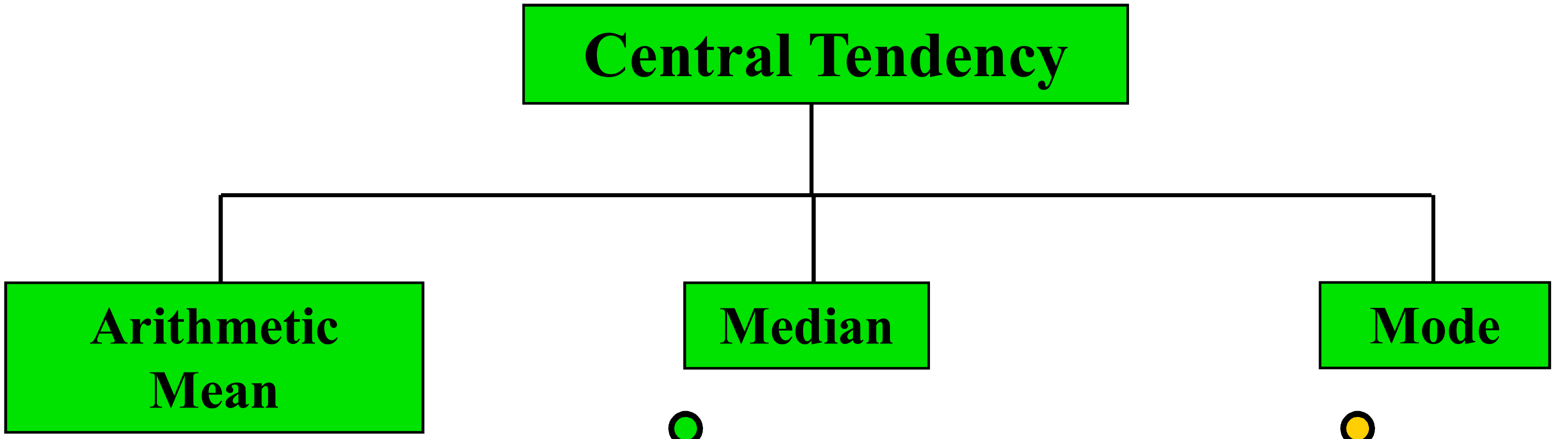
Measures of Central Tendency: Which Measure to Choose?

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



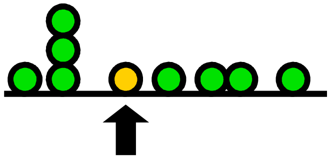
Central Tendency

Arithmetic Mean

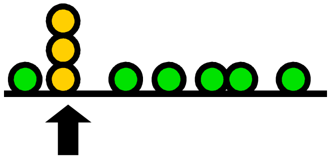
Median

Mode

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



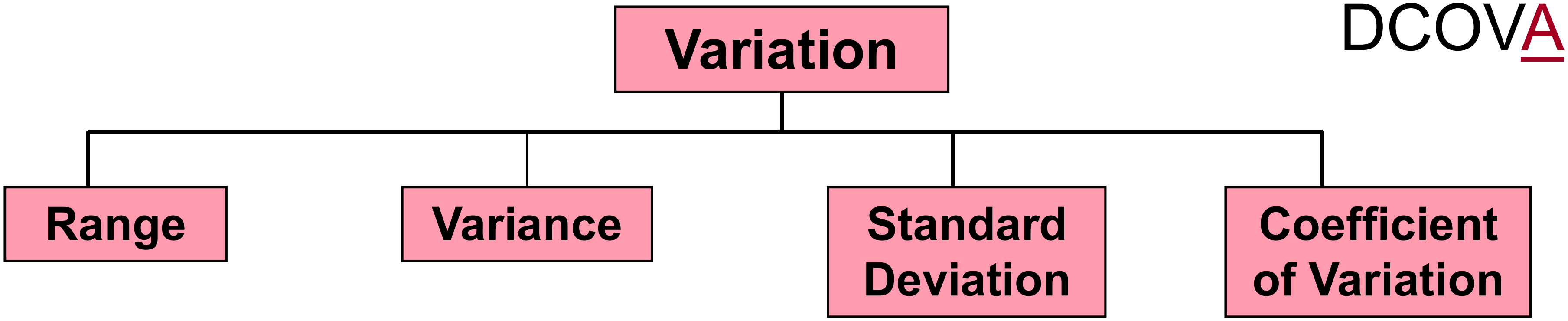
Middle value in the ordered array



Most frequently observed value

Measures of Variation

DCOVA



Variation

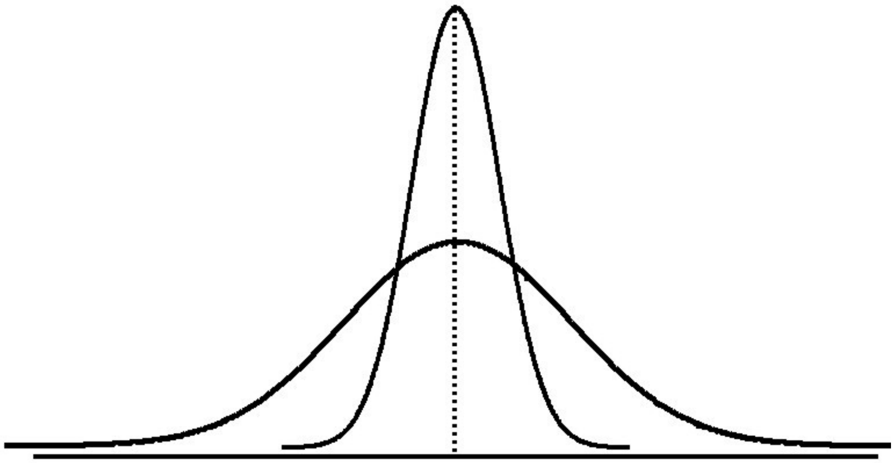
Range

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

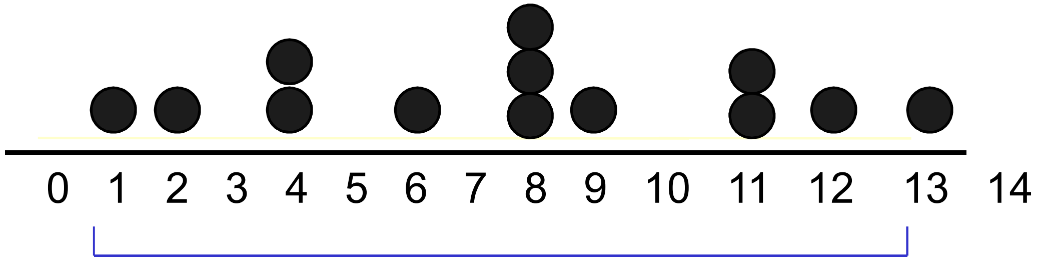
Measures of Variation: The Range

DCOVA

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

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

Example:

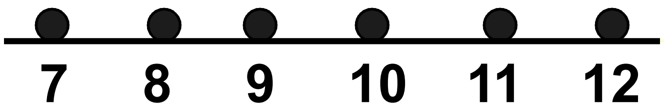


Range = 13 - 1 = 12

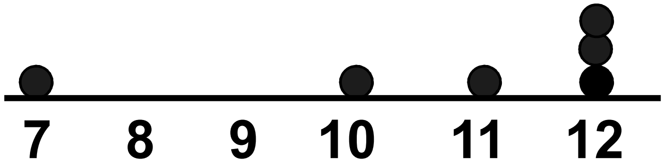
Measures of Variation:
Why The Range Can Be Misleading

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:

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

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:

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

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

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

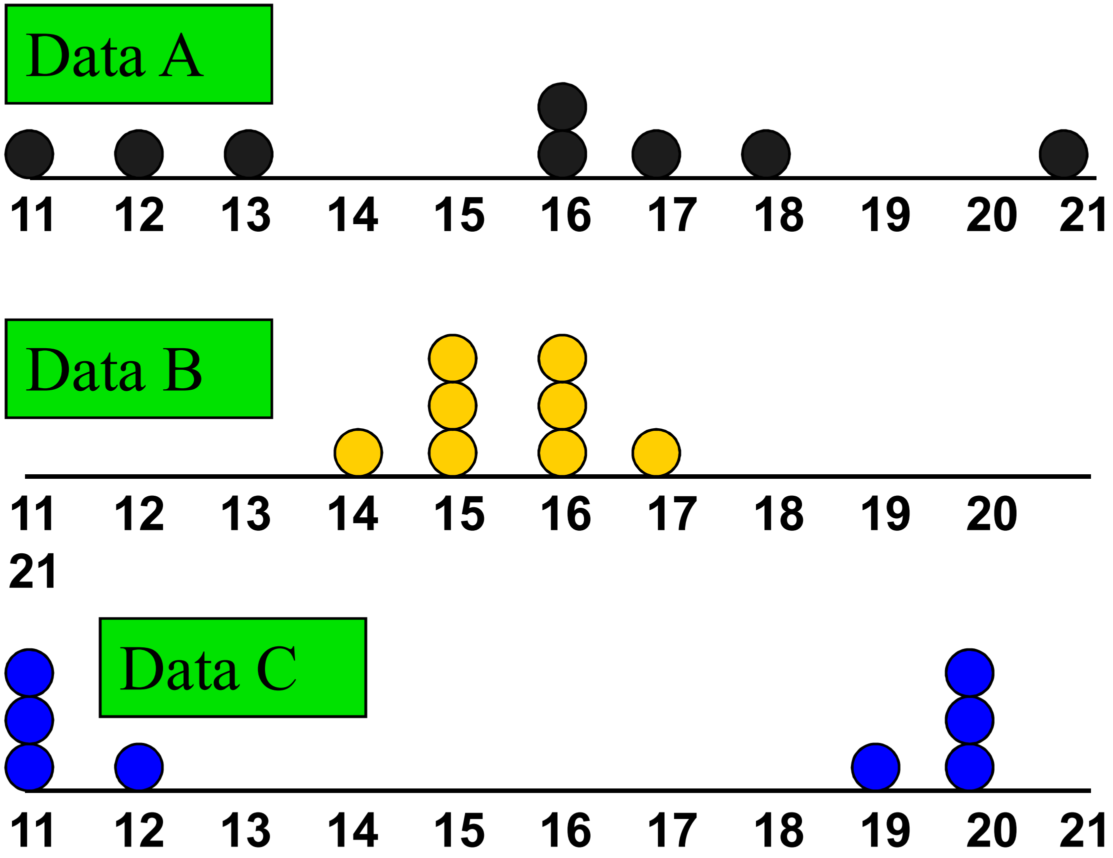
$$= \sqrt{\frac{(10-16)^2 + (12-16)^2 + (14-16)^2 + \cdots + (24-16)^2}{8-1}}$$

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

→ A measure of the “average” scatter around the mean.

Measures of Variation: Comparing Standard Deviations

DCOVA



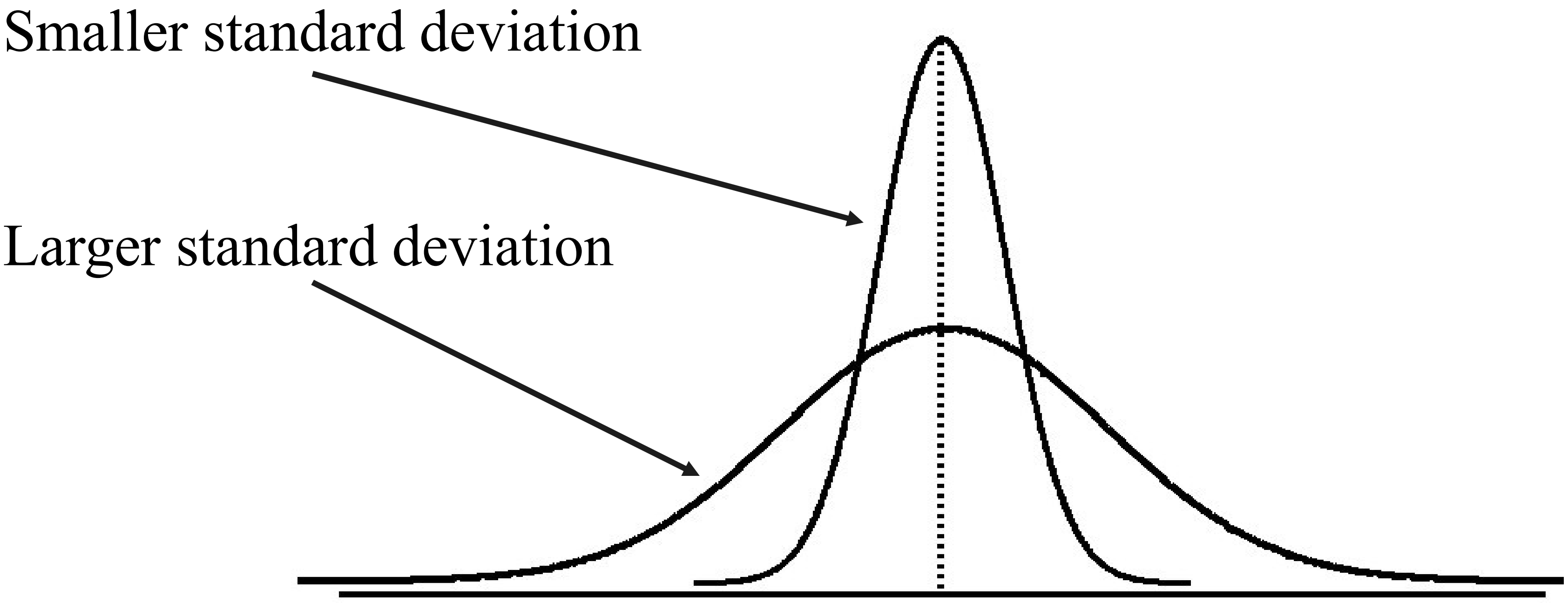
Mean = 15.5
S = 3.338

Mean = 15.5
S = 0.926

Mean = 15.5
S = 4.567

Measures of Variation: Comparing Standard Deviations

DCOVA



Smaller standard deviation

Larger standard deviation

Measures of Variation: Summary Characteristics

DCOVA

- The more the data are spread out, the greater the range, variance, and standard deviation.

- The more the data are concentrated, the smaller the range, variance, and standard deviation.

- If the values are all the same (no variation), all these measures will be zero.

- None of these measures are ever negative.

Measures of Variation: The Coefficient of Variation

DCOVA

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

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

Measures of Variation: Comparing Coefficients of Variation

DCOVA

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

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

- Stock B:
  - Mean price last year = $100.
  - Standard deviation = $5.

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

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.

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

- Stock C:
  - Mean price last year = $8.
  - Standard deviation = $2.

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

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

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

where X represents the data value

$\overline{X}$ is the sample mean

S is the sample standard deviation

Locating Extreme Outliers: Z-Score

DCOV<u>A</u>

- 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.