ADTA 5130/IPAC 4130

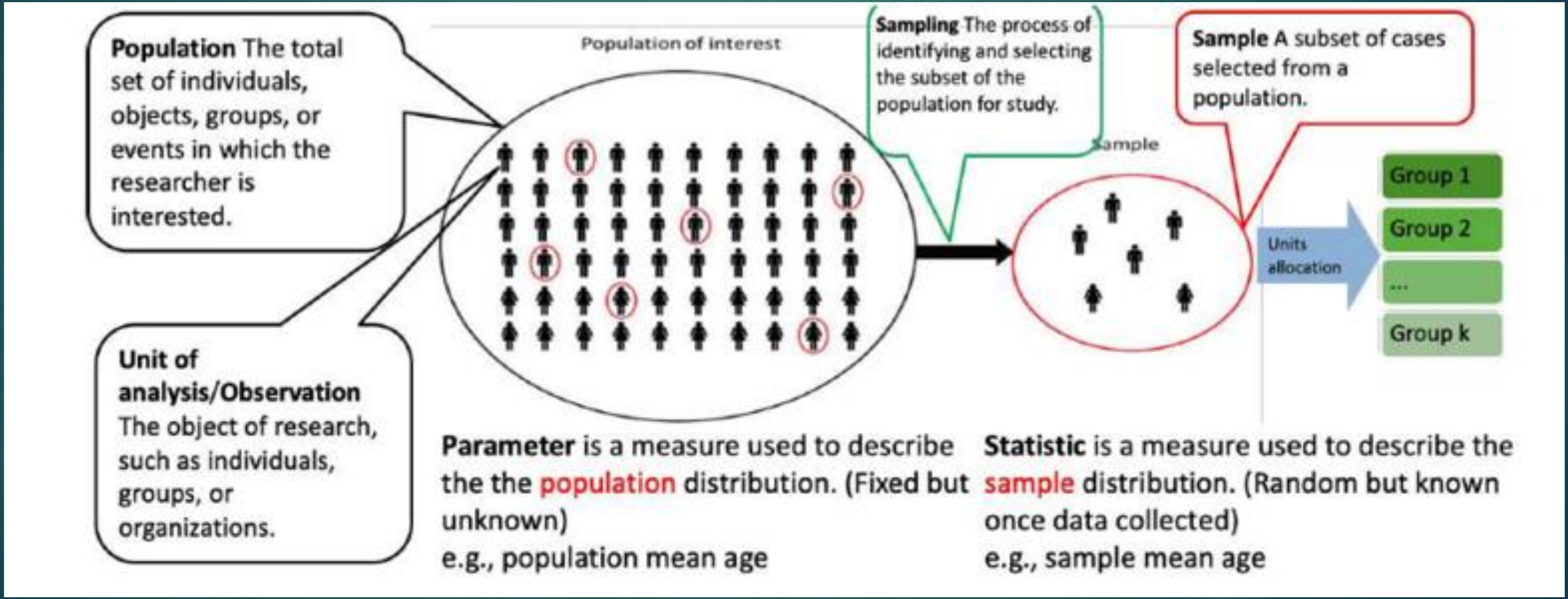
MODULE 1: CHAPTERS 1, 2, AND 3

Agenda

- Introduction
- Admin: Canvas, Syllabus, Expectations
- Chapters 1, 2, and 3
- Homework

Definitions

- Descriptive stats: characteristics of data; charts/tables/measures
- Inferential stats: drawing conclusions about data



Population The total set of individuals, objects, groups, or events in which the researcher is interested.

Unit of analysis/Observation The object of research, such as individuals, groups, or organizations.

Sampling The process of identifying and selecting the subset of the population for study.

Sample A subset of cases selected from a population.

Parameter is a measure used to describe the the population distribution. (Fixed but unknown)
e.g., population mean age

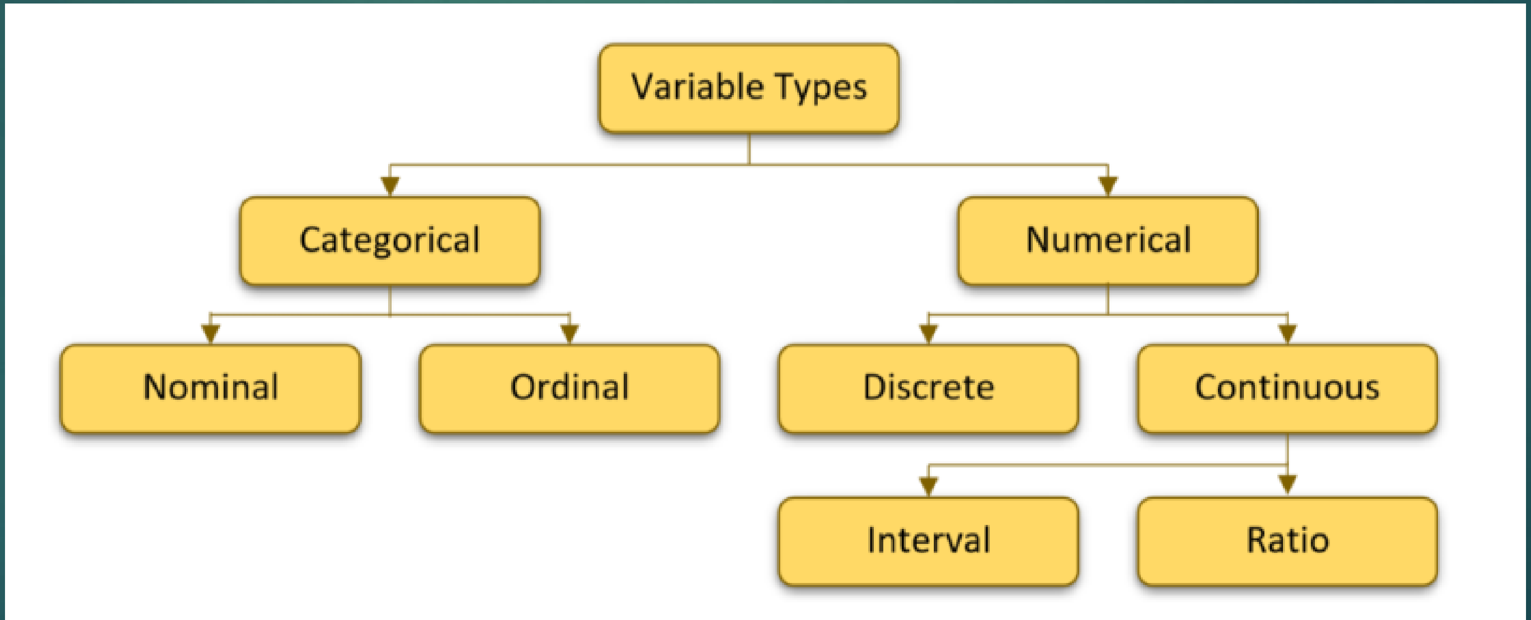
Statistic is a measure used to describe the sample distribution. (Random but known once data collected)
e.g., sample mean age

Definitions

- Structured Data: Row-column format (databases/spreadsheets)
- Unstructured Data: No predefined format (text, video, image)

Nominal – no order

Ordinal - order

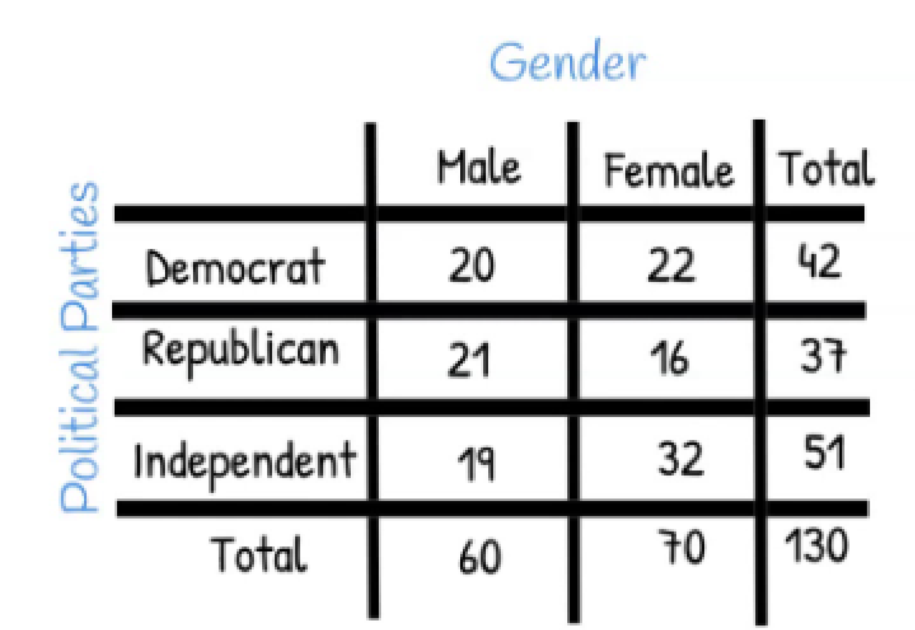


Discrete - counts

Continuous – can always find a measurement in between

Data Preparation

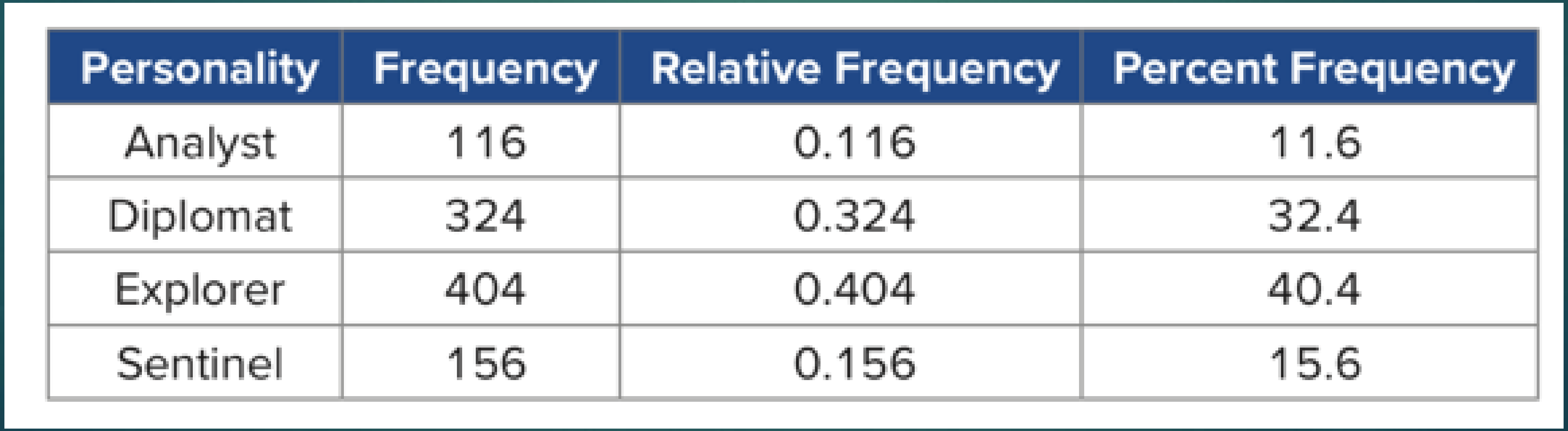
- Number of Observations
- Number of Variables
- Variable Types
- Missing Values
- Data Subsets
- Counts of Values
  - Categorical Variables: Contingency Table / Frequency Table
  - Numerical Variables: Histogram



		Gender	
Political Parties	Male	Female	Total
Democrat	20	22	42
Republican	21	16	37
Independent	19	32	51
Total	60	70	130

Frequency Distribution

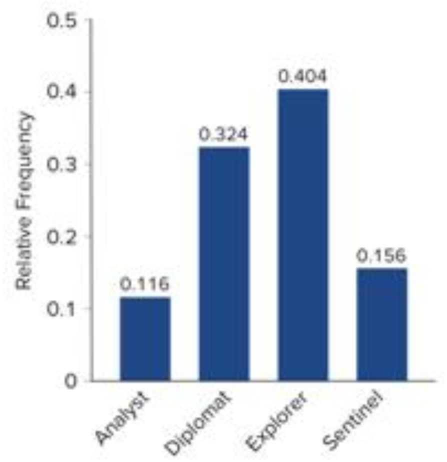
- Group the data into categories and record the number of observations that fall into each category.
- The relative frequency for each category is the proportion of observations in each category.
- The percent frequency is the relative frequency expressed as a percent (Multiply by 100).



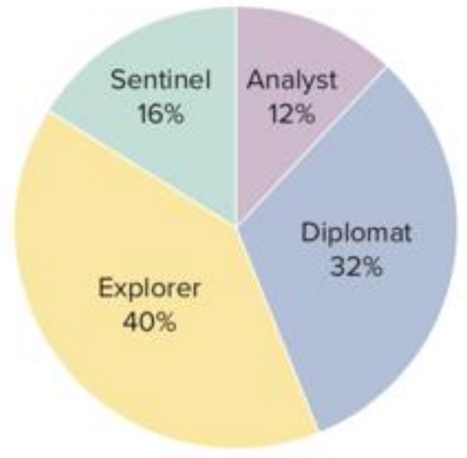
Personality	Frequency	Relative Frequency	Percent Frequency
Analyst	116	0.116	11.6
Diplomat	324	0.324	32.4
Explorer	404	0.404	40.4
Sentinel	156	0.156	15.6

Visualizing Categorical Variables

One Variable

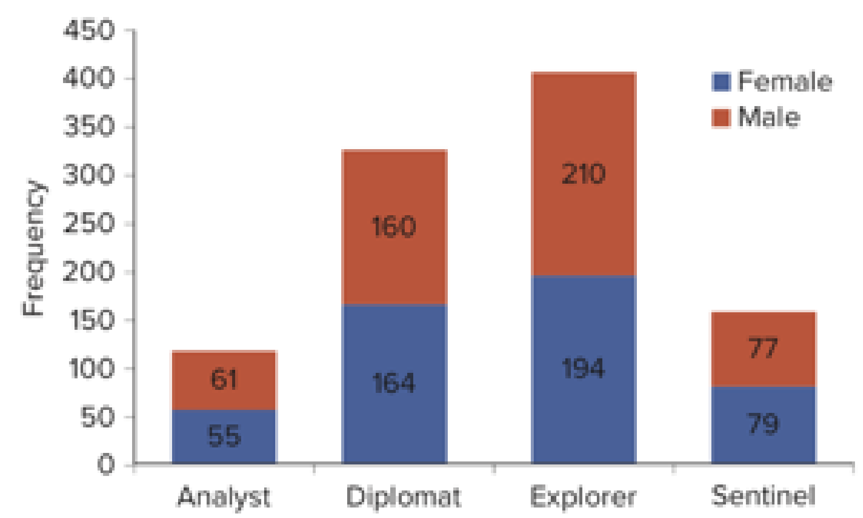


Bar Chart



Pie Chart

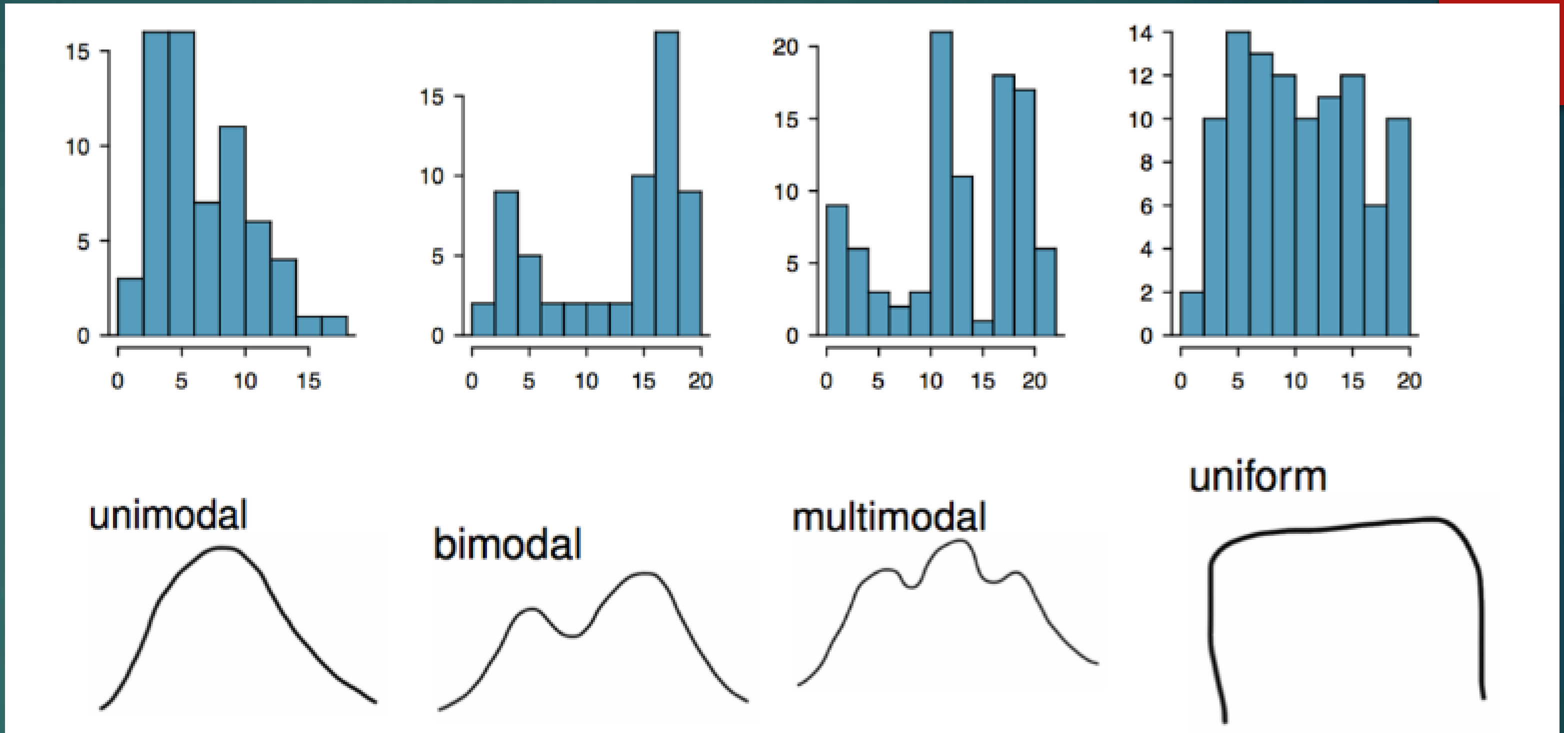
Two Variables



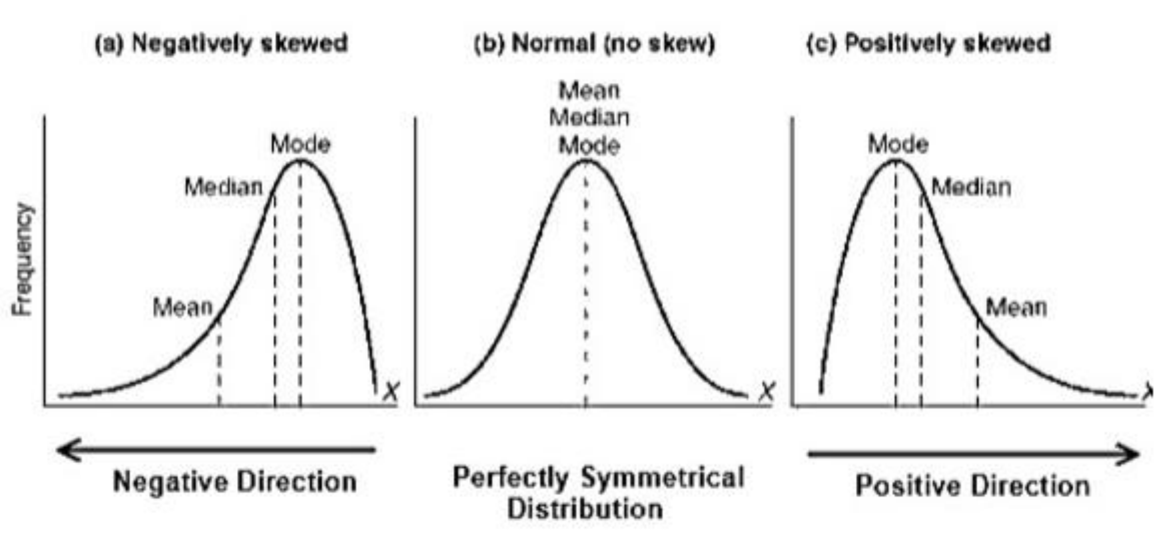
Stacked Bar Chart

Histogram

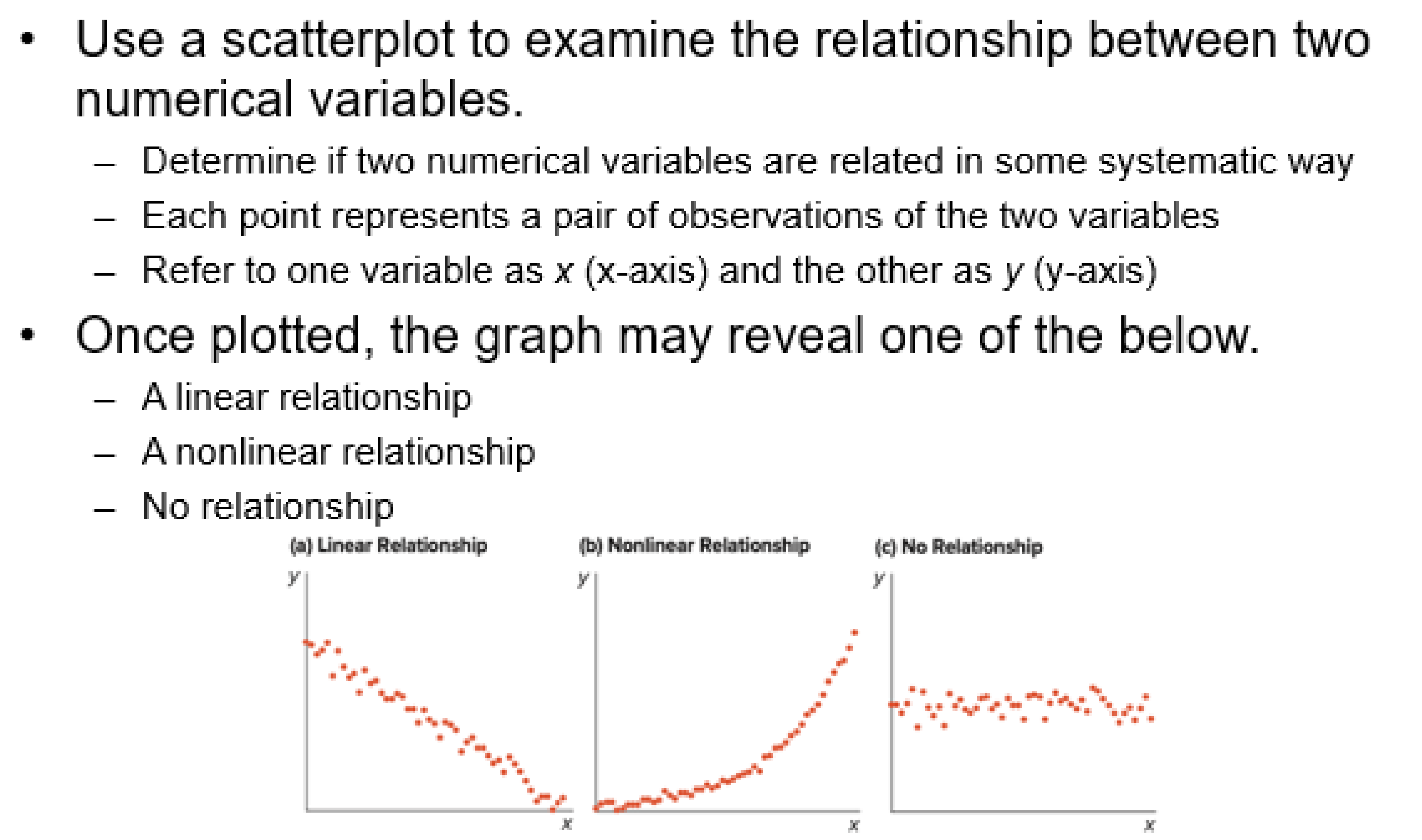
Modality



Skewness



Scatterplot



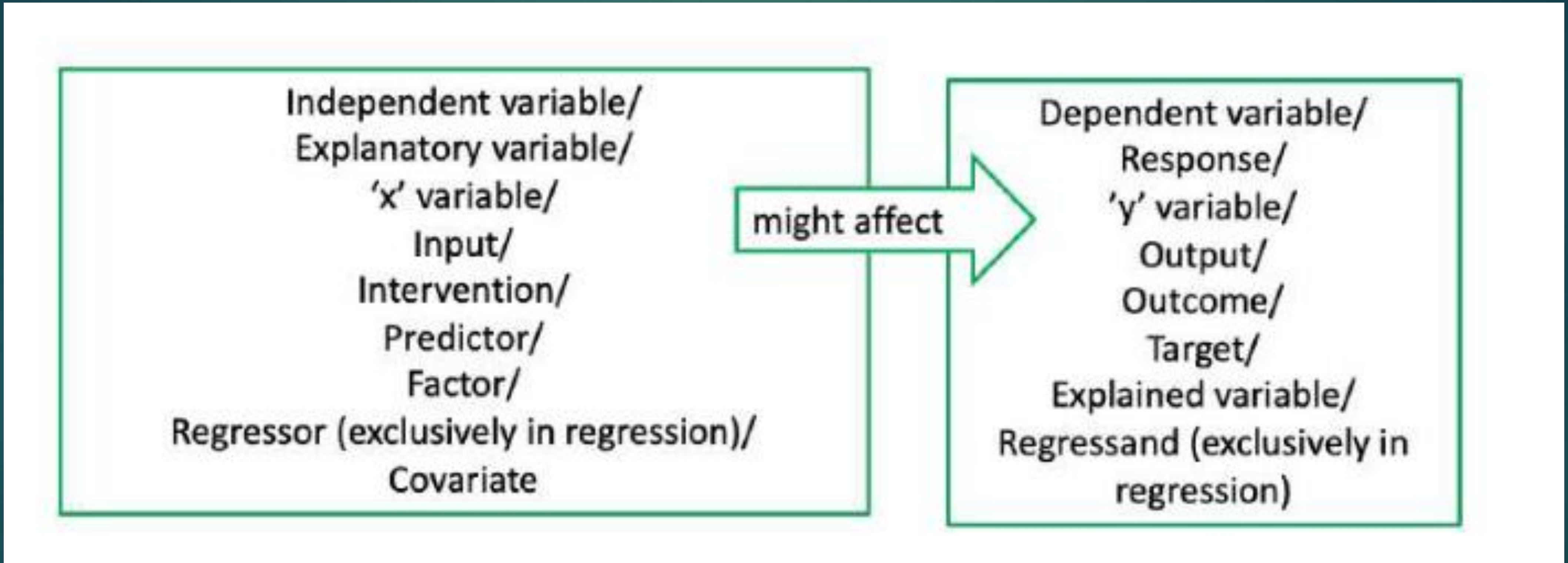
- Use a scatterplot to examine the relationship between two numerical variables.
  - Determine if two numerical variables are related in some systematic way
  - Each point represents a pair of observations of the two variables
  - Refer to one variable as *x* (x-axis) and the other as *y* (y-axis)
- Once plotted, the graph may reveal one of the below.
  - A linear relationship
  - A nonlinear relationship
  - No relationship

(a) Linear Relationship

(b) Nonlinear Relationship

(c) No Relationship

Variable Terminology



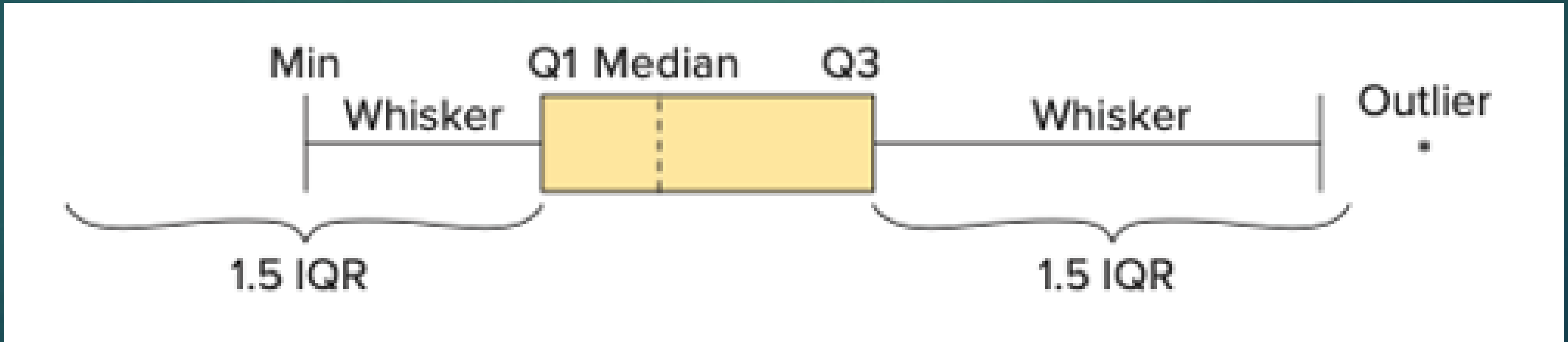
Independent variable/
Explanatory variable/
'x' variable/
Input/
Intervention/
Predictor/
Factor/
Regressor (exclusively in regression)/
Covariate

might affect

Dependent variable/
Response/
'y' variable/
Output/
Outcome/
Target/
Explained variable/
Regressand (exclusively in regression)

Measures of Central Tendency

- Mean (Arithmetic Mean or Average): Add up all the observations and divide by the number of observations
- Median: Middle value, or average of two middle values if no middle
- Mode: Observation that occurs most frequently
- Percentile: Divides data into two parts (e.g, 75% of data below 75th percentile)
- Boxplot: Box is Interquartile Range or IQR: 75th percentile [Q3] - 25th percentile [Q1]



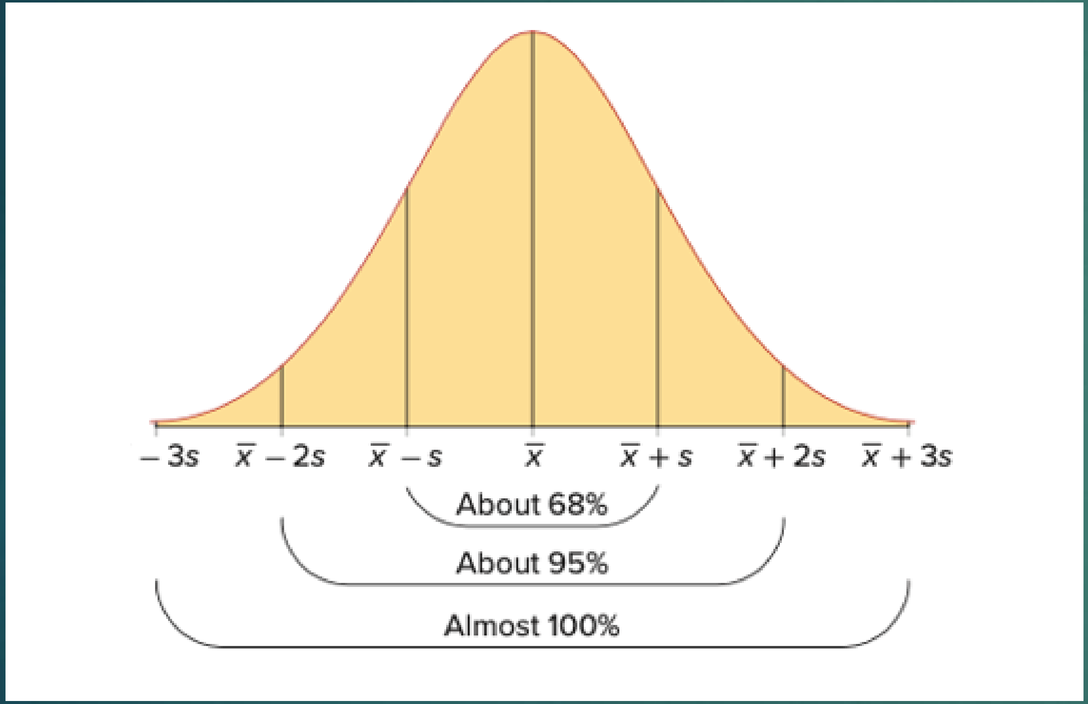
Measures of Dispersion or Variability

- Range: difference between the highest (maximum) and the lowest (minimum)
- Mean Absolute Difference (MAD): average of the absolute differences between the observations and the mean
- Variance: average of the squared differences between the observations and the mean
- Standard Deviation: square root of variance
- Coefficient of Variation: Used to compare variability of two or more data sets

Measures of Association

- Covariance: measures the direction of the linear relationship
- Correlation Coefficient: describes both the direction and strength of the linear relationship
  - -1 is perfect negative linear correlation
  - 0 is no correlation
  - 1 is perfect positive linear correlation

Relative Location



Empirical Rule

Outliers defined as outside -3s & 3s

Z score: distance of an observation (x) from the mean (x bar) in terms of standard deviations (s)

$$Z = \frac{x-\bar{x}}{s}$$

Resources

- R: https://www.rstudio.com/resources/cheatsheets/
- Python: https://www.pythoncheatsheet.org/
- Excel: https://www.investintech.com/resources/articles/excelcheatsheet/
- Statistics Formulas: http://web.mit.edu/~csvoss/Public/usabo/stats_handout.pdf

1

Data and Data Preparation

Business Statistics: Communicating with Numbers, 4e

By Sanjiv Jaggia and Alison Kelly

Chapter 1 Learning Objectives (LOs)

LO 1.1 Explain the various data types.

LO 1.2 Describe variables and types of measurement scales.

LO 1.3 Inspect and explore data.

LO 1.4 Apply data subsetting.

Introductory Case: Retail Customer Data [1]

Design a marketing campaign for Organic Food Superstore.

CustID	Sex	Race	BirthDate	…	Channel
1530016	Female	Black	12/16/1986	…	SM
1531136	Male	White	5/9/1993	…	TV
⋮	⋮	⋮	⋮	⋮	⋮
1579979	Male	White	7/5/1999	…	SM

Use the data set to:

- Identify Organic Food Superstore’s college-educated millennial customers.
- Compare the profiles of female and male college-educated millennial customers.

1.1: Types of Data 1

Data are compilations of facts, figures, or other contents, both numerical and non-numerical.

- All types/formats are generated from multiple sources.
- Customers/businesses use data from to help make decisions.
- Statistics is the language of data.

Statistics is the science that deals with the collection, preparation, analysis, interpretation, and presentation of data.

- First: find the right data and prepare it for the analysis.
- Second: use the appropriate statistical tool, which depends on the data.
- Third: clearly communicate information with actionable business insights.

1.1: Types of Data 2

There are two branches of statistics: descriptive and inferential statistics.

Descriptive statistics refers to the summary of important aspects of a data set.

- Includes collecting, organizing, and presenting the data in the form of charts and tables.
- Often calculate numerical measures (typical value, variability).

Inferential statistics refers to drawing conclusions about a larger set of data (population) based on a smaller set of data (sample).

- A population consists of all items/members of interest.
- A sample is a subset of the population.

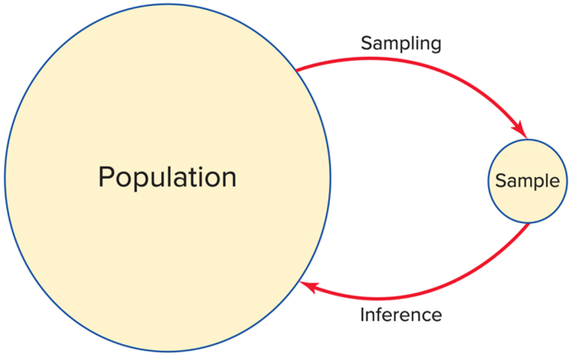
We rely on sample data to make inferences about various characteristics of the population.

1.1: Types of Data 3

We analyze sample data and calculate a sample statistic to make inferences about the unknown population parameter.

It is generally not feasible to obtain population data.

- Obtaining information on the entire population is expensive.
- It is impossible to examine every member of the population.



Sample data are generally collected in one of two ways.

1.1: Types of Data 4

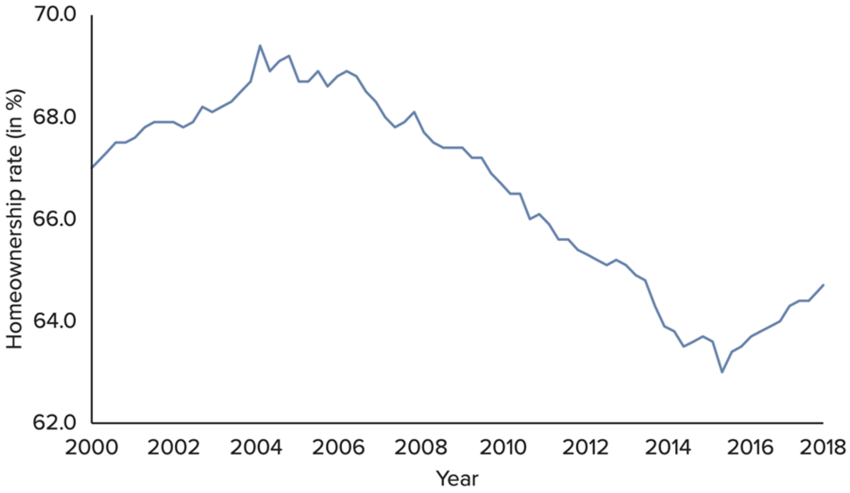
- Cross-sectional data refers to data collected by recording a characteristic of many subjects at the same point in time, or without regard to differences in time.
- Example: 2018-2019 NBA Eastern Conference standings.

Team name	Wins	Losses	Winning percentage
Milwaukee Bucks	60	22	0.732
Toronto Raptors*	58	24	0.707
Philadephia 76ers	51	31	0.622
Boston Celtics	49	33	0.598
Indiana Pacers	48	34	0.585
Brooklyn Nets	42	40	0.512
Orlando Magic	42	40	0.512
Detroit Pistons	41	41	0.500

*The Toronto Raptors won their first NBA title during the 2018-2019 season.

1.1: Types of Data 5

- Time series data refers to data collected over several time periods focusing on certain groups of people, specific events, or objects.
- Time series data can include hourly, daily, weekly, monthly, quarterly, or annual observations.
- Example: homeownership rates (%) in the US.

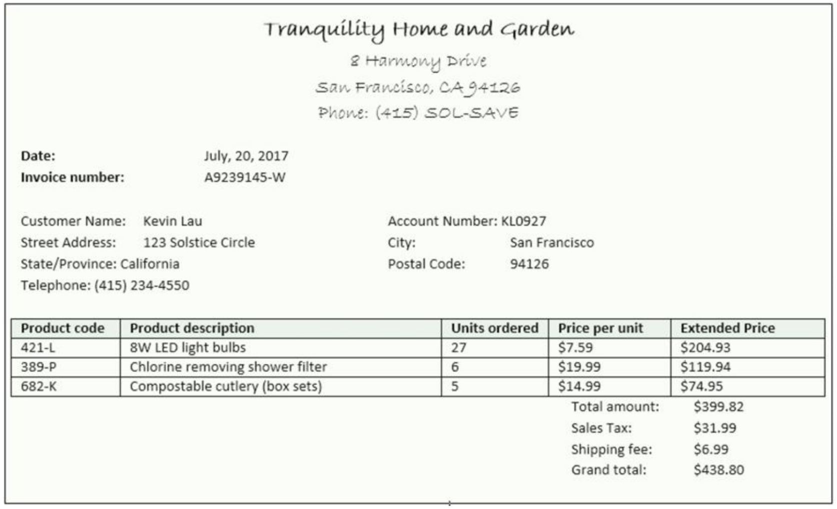


Access the text alternative for slide images.

1.1: Types of Data 6

Structured data,

- Reside in a pre-defined, row-column format.
- Spreadsheet or database applications.
- Enter, store, query, and analyze.
- Numerical information that is objective and not open to interpretation.



Tranquility Home and Garden

8 Harmony Drive

San Francisco, CA 94126

Phone: (415) SOL-SAVE

Date:	July, 20, 2017
Invoice number:	A9239145-W

Customer Name:	Kevin Lau	Account Number:	KL0927
Street Address:	123 Solstice Circle	City:	San Francisco
State/Province:	California	Postal Code:	94126
Telephone:	(415) 234-4550		

Product code	Product description	Units ordered	Price per unit	Extended Price
421-L	8W LED light bulbs	27	$7.59	$204.93
389-P	Chlorine removing shower filter	6	$19.99	$119.94
682-K	Compostable cutlery (box sets)	5	$14.99	$74.95

Total amount:	$399.82
Sales Tax:	$31.99
Shipping fee:	$6.99
Grand total:	$438.80

Access the text alternative for slide images.

1.1: Types of Data 7

Today, only about 20% of all data used in business decisions are structured.

Unstructured data.

- Do not conform to a pre-defined, row-column format.
- Textual and multimedia content.
- Do not conform to database structures.
- These data may have some implied structure.
  - Still considered unstructured.
- Do not conform to a row-column model required in most database systems.
- Example: social media data such as Twitter, YouTube, Facebook, and blogs.

1.1: Types of Data [8]

Businesses generate and gather more and more data at an increasing pace: Big Data.

- A massive volume of structured and unstructured data.
- Extremely difficult to manage, process, and analyze using traditional data processing tools.
- Presents great opportunities to gain knowledge and game-changing intelligence.

"[H]igh-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation" (www.gartner.com).

Does not imply complete (population) data.

Big data may not be used when available.

- Inconvenient and computationally burdensome.
- Benefits may not justify costs.

1.1: Types of Data [9]

There are three characteristics of big data.

- Volume: immense amount of data complied for a single or multiple sources.
- Velocity: generated at a rapid speed, management is a critical issue.
- Variety: all types, forms, granularity, structured or unstructured.

Additional characteristics.

- Veracity: credibility and quality of the data, reliability.
- Values: methodological plan for formulating questions, curating the right data and unlocking hidden potential.

Having a plethora of data does not guarantee that useful insights or measurable improvements will be generated.

1.1: Types of Data 10

There is an abundance of data on the Internet.

Many experts believe that 90% of the data in the world today was created in the last two years alone.

It is easy to access and find data by using a search engine like Google.

There are several sources of data.

- Bureau of Economic Analysis.
- Bureau of Labor Statistics.
- Federal Research Economic Data.
- U.S. Census Bureau.
- National Climate Data Center.
- Yahoo Finance, Google Finance.
- Zillow.
- ESPN.

1.2: Variables and Scales of Measurement 1

A variable is a characteristic of interest that differs in kind or degree among various observations (records).

There are two types of variables: categorical and numeric.

Categorical Data.

- Also called qualitative.
- Represent categories.
- Labels or names to identify distinguishing characteristics.
- Can be defined by two or more categories.
- Coded into numbers for data processing.
- Example: marital status, grade in a course.

1.2: Variables and Scales of Measurement 2

For a numerical variable, we use numbers to identify the distinguishing characteristic of each observation.

Numeric Data.

- Also called quantitative.
- Represent meaningful numbers.
- Either discrete or continuous.

A discrete variable assumes a countable number of values.

- The values need not be whole numbers.
- Example: number of children in a family.

1.2: Variables and Scales of Measurement 3

A continuous variable assumes an uncountable number of values within an interval.

- In practice, often measure in discrete values.
- Example: weight of a newborn baby.

In order to choose the appropriate techniques for summarizing and analyzing variables, we need to distinguish between the different measurement scales.

1.2: Variables and Scales of Measurement 4

There are four major scales: nominal, ordinal, interval, ratio.

Nominal and ordinal scales are used for categorical variables.

Nominal.

- Least sophisticated.
- Represent categories or groups.
- Values differ by label or name.
- Example: marital status.

Ordinal.

- Stronger level of measurement.
- Categorize and rank data with respect to some characteristic.
- Cannot interpret the difference between the ranked values, numbers are arbitrary.
- Example: reviews from 1 star (poor) to 5 starts (outstanding).

1.2: Variables and Scales of Measurement 5

Categorical variable are typically expressed in words but are coded into numbers for purposes of data processing.

- Typically count the number of observations that fall into each category (or find percentages).
- Unable to perform meaningful arithmetic operations.

1.2: Variables and Scales of Measurement 6

Interval and ratio scales are used for numerical variables.

Interval.

- Categorize and rank, differences are meaningful.
- Zero value is arbitrary and does not reflect absence of characteristic.
- Ratios are not meaningful.
- Example: temperature.

Ratio.

- Strongest level of measurement.
- A true zero point, reflects absence of characteristic.
- Ratios are meaningful.
- Example: profits.

Arithmetic operations are valid on interval- and ratio-scaled variable.

1.2: Variables and Scales of Measurement 7

- Example: The owner of a ski resort gathers data on tweens.

Tween	Music Streaming	Food Quality	Closing Time	Own Money Spent ($)
1	Apple Music	4	5:00 pm	20
2	Pandora	2	5:00 pm	10
⋮	⋮	⋮	⋮	⋮
20	Spotify	2	4:30 pm	10

- Music: nominal.
- Food quality: ordinal.
- Closing time: interval.
- Own money spent: ratio.

1.3: Data Preparation 1

We often spend a considerable amount of time inspecting and preparing the data for the subsequent analysis.

- Counting and sorting.
- Handling missing values.
- Subsetting.

Counting and Sorting.

- Among the very first tasks analysts perform.
- Gain a better understanding and insights into the data.
- Help to verify that the data set is complete or determine if there are missing values.
- Sorting allows us to review the range of values for each variable.
- Sort based on a single or multiple variables.

1.3: Data Preparation 2

There are two common strategies for dealing with missing values.

The omission strategy recommends that observations with missing values be excluded from subsequent analysis.

The imputation strategy recommends that the missing values be replaced with some reasonable imputed values.

- Numeric variables: replace with the average.
- Categorical variables: replace with the predominant category.

1.3: Data Preparation 3

Subsetting is the process of extracting a portion of the data set that is relevant for subsequent statistical analysis.

- The objective of the analysis is to compare two subsets of the data.
- Eliminate observations that contain missing values, low-quality data, or outliers.
- Excluding variables that contain redundant information, or variables with excessive amounts of missing values.

We can also subset data based on data ranges.

End of Main Content

2

Tabular and Graphical Methods

Business Statistics: Communicating with Numbers, 4e

By Sanjiv Jaggia and Alison Kelly

Chapter 2 Learning Objectives (LOs)

LO 2.1 Construct and interpret a frequency distribution for a categorical variable.

LO 2.2 Construct and interpret a bar chart and a pie chart.

LO 2.3 Construct and interpret a contingency table and a stacked bar chart for two categorical variables.

LO 2.4 Construct and interpret a frequency distribution for a numerical variable

LO 2.5 Construct and interpret a histogram, a polygon, and an ogive.

LO 2.6 Construct and interpret a scatterplot, a scatterplot with a categorical variable, and a line chart.

LO 2.7 Construct and interpret a stem-and-leaf diagram.

Introductory Case: House Prices in Punta Gorda 1

- A relocation specialist for a real estate firm gathers recent house sales data for a client.

Transaction	Price	Sqft	Beds	Baths	Built	Type
1	200	1684	3	2	2005	Single
2	435	2358	3	2.5	2017	Single
⋮	⋮	⋮	⋮	⋮	⋮	⋮
40	192	1154	2	2	2019	Condo

- Use the sample information to:

1. Make summary statements concerning the range of house prices.
2. Comment on where house prices tend to cluster.
3. Examine the relationship between a house price and its size.

2.1 Methods to Visualize a Categorical Variable 1

A categorical variable consists of observations that represent labels or names.

Summarize the data with a frequency distribution.

- Group the data into categories and record the number of observations that fall into each category.
- The relative frequency for each category is the proportion of observations in each category.
- Multiply the proportions by 100 to get percentages.

2.1 Methods to Visualize a Categorical Variable 2

- Example: Myers-Briggs assessment personality types for 1,000 employees at a technology firm.

Personality	Frequency	Relative Frequency	Percent Frequency
Analyst	116	0.116	11.6
Diplomat	324	0.324	32.4
Explorer	404	0.404	40.4
Sentinel	156	0.156	15.6

2.1 Methods to Visualize a Categorical Variable 3

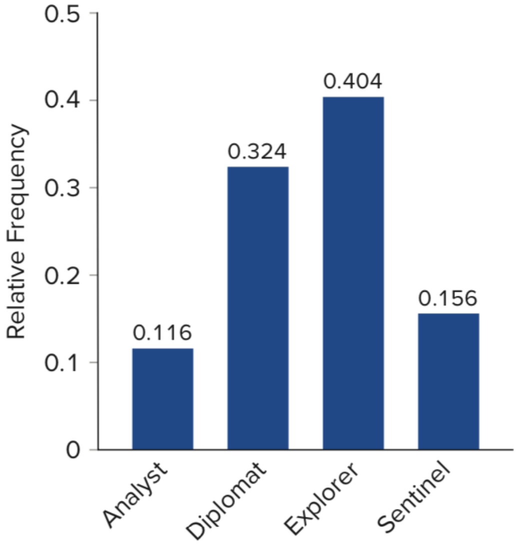
A bar chart depicts the frequency or relative frequency for each category of the categorial variable.

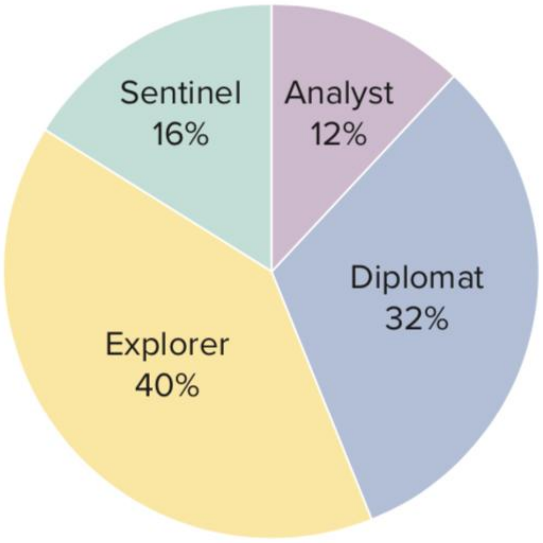
- Series of either horizontal or vertical bars.
- Bar lengths proportional to the values they are depicting.

A pie chart is a segmented circle whose segments portray the relative frequencies of the categories of a qualitative variable.

2.1 Methods to Visualize a Categorical Variable 4

- Example: Myers-Briggs assessment personality types for 1000 employees at a technology firm.

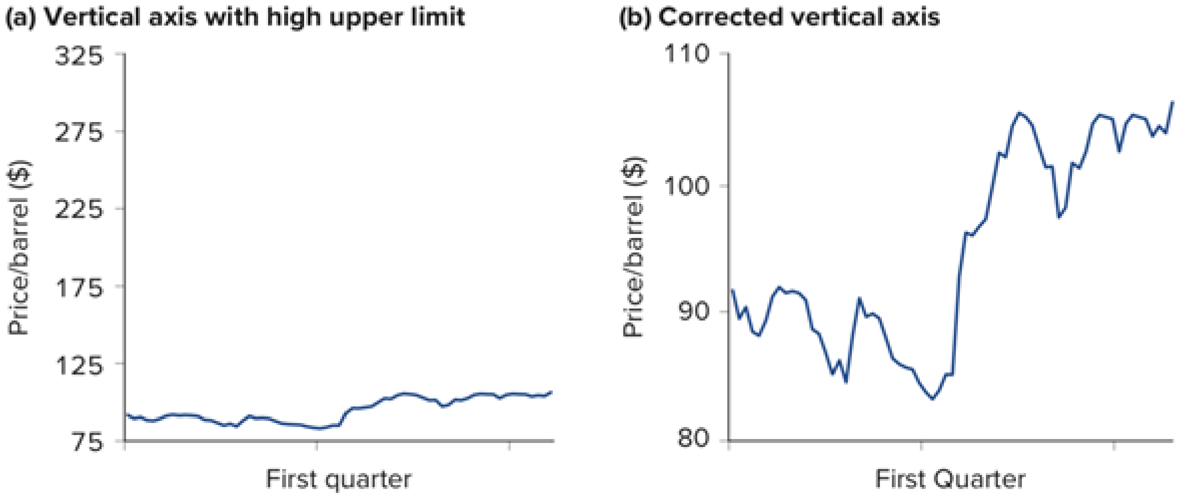




Access the text alternative for slide images.

2.1 Methods to Visualize a Categorical Variable 5

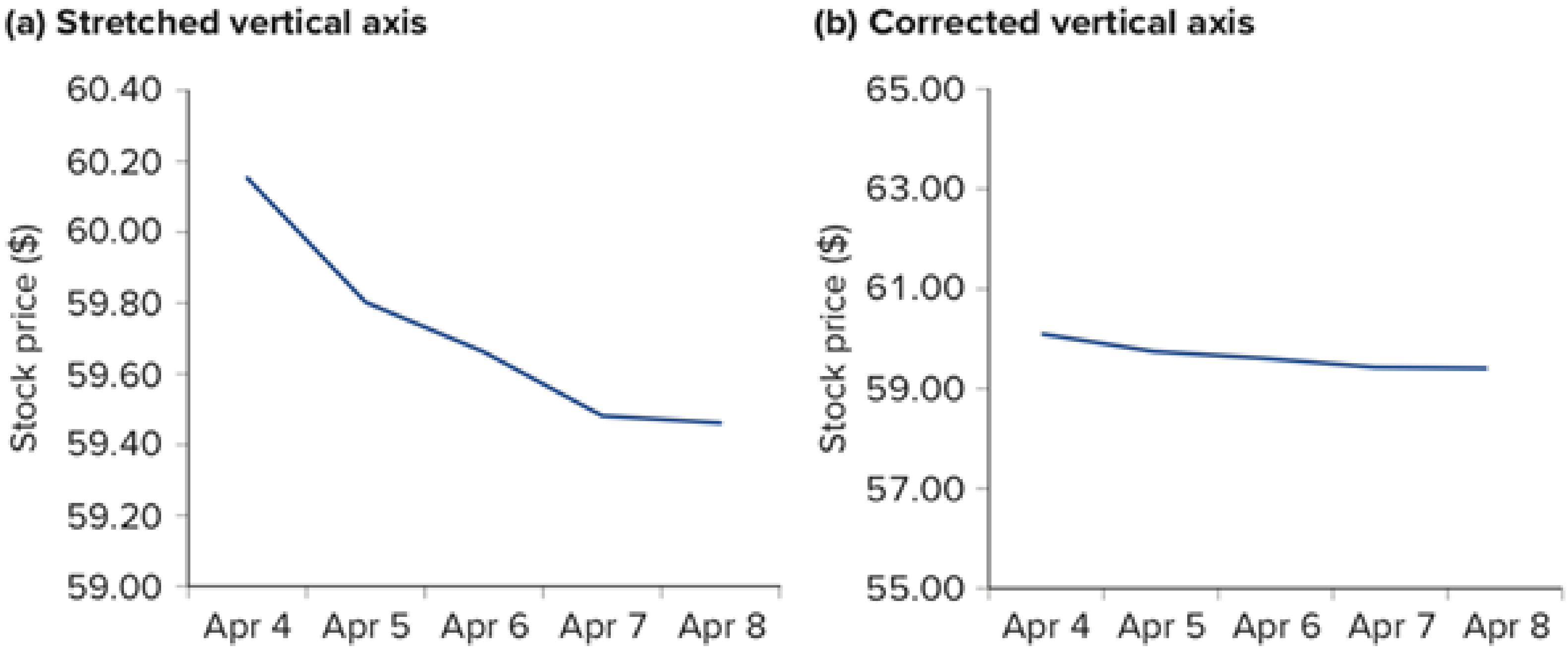
- The simplest graph should be used.
- Axes should be clearly marked with numbers and scales.
- Bars on bar charts should have the same width.
- Vertical axis should not have a very high values as an upper limit.



Access the text alternative for slide images.

2.1 Methods to Visualize a Categorical Variable [6]

- Vertical axis should not be stretched.



Access the text alternative for slide images.

2.2 Methods to Visualize the Relationship Between Two Categorical Variables [1]

Use a contingency table to examine the relationship between two categorical variables.

- Frequencies for two categorical variables.
- Each cell represents a mutually exclusive combination of the pair of values.

Use a stacked column chart to visualize more than one categorical variable.

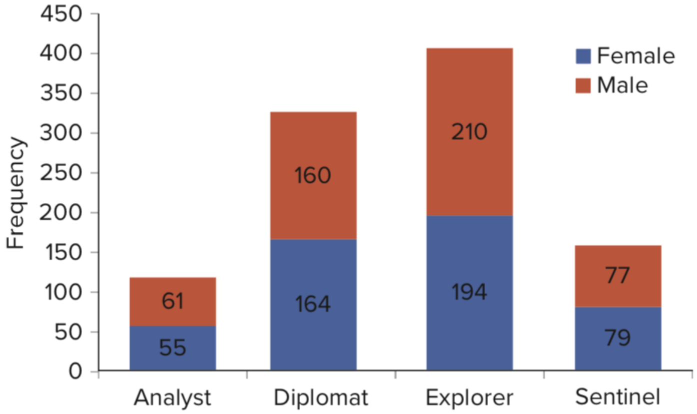
- Graphically shows the contingency table.
- Allows for the comparison compositive within each category.

2.2 Methods to Visualize the Relationship Between Two Categorical Variables 2

- Example: Myers-Brigg personality assessment and sex.

Personality

Sex	Analyst	Diplomat	Explorer	Sentinel
Female	55	164	194	79
Male	61	160	210	77



Access the text alternative for slide images.

2.3 Methods to Visualize a Numeric Variable [1]

- For a categorical, the raw data could be categorized in a well-defined way.
- With a numerical variable, each observation represents a meaningful amount or count.
- Use a frequency distribution to summarize a numerical variable.
- Instead of categories, we construct a series of intervals or classes.
- The data are more manageable using a frequency distribution, but some detail is lost.

2.3 Methods to Visualize a Numeric Variable 2

We have to make decisions about the number of intervals and the width of each interval.

- The intervals are mutually exclusive.
- The total number of intervals usually ranges from 5 to 20.
- The intervals are exhaustive.
- The intervals are easy to recognize and interpret.
- A starting point for approximating the width of each interval is given by

$$\frac{\text{Maximim} - \text{Minimum}}{\text{Number of Intervals}}.$$

2.3 Methods to Visualize a Numeric Variable [3]

- Example: house prices in Punta Gorda.
- Suppose we are going to have 6 intervals.
- The maximum is 649 and the minimum is 125.
- As a starting point, the width of each interval could be:

$$\frac{649-125}{6}=87.33$$

- This would not give limits that are easily recognizable, so we use 100.

Interval (in $1,000s)	Frequency
$100 < x \le 200$	9
$200 < x \le 300$	16
$300 < x \le 400$	8
$400 < x \le 500$	4
$500 < x \le 600$	2
$600 < x \le 700$	1

2.3 Methods to Visualize a Numeric Variable 4

In addition to a frequency distribution, there are three other items to compute.

- Relative frequency: proportion (or fraction) of observations that falls into each interval.
- Cumulative frequency: the number of observations that falls below the upper limit of a particular interval.
- Cumulative relative frequency: the proportion (or fraction) of observations that falls below the upper limit of a particular interval.

2.3 Methods to Visualize a Numeric Variable 6

- Example: house prices in Punta Gorda.
- Use the previous frequency distribution to determine the below.

Interval (in \$1,000s)	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
$100 < x \leq 200$	9	0.225	9	0.225
$200 < x \leq 300$	16	0.400	9 + 16 = 25	0.225 + 0.400 = 0.625
$300 < x \leq 400$	8	0.200	9 + 16 + 8 = 33	0.225 + 0.400 + 0.200 = 0.825
$400 < x \leq 500$	4	0.100	9 + 16 +....+ 4 = 37	0.225 + 0.400 +...+ 0.100 = 0.925
$500 < x \leq 600$	2	0.050	9 + 16 +....+ 2 = 39	0.225 + 0.400 +...+ 0.050 = 0.975
$600 < x \leq 700$	1	0.025	9 + 16 +....+ 1 = 40	0.225 + 0.400 +...+ 0.025 = 1.000

2.3 Methods to Visualize a Numeric Variable 7

A histogram is the counterpart to the vertical bar chart used for a categorical variable.

Graphically depict a frequency distribution for a numeric variable.

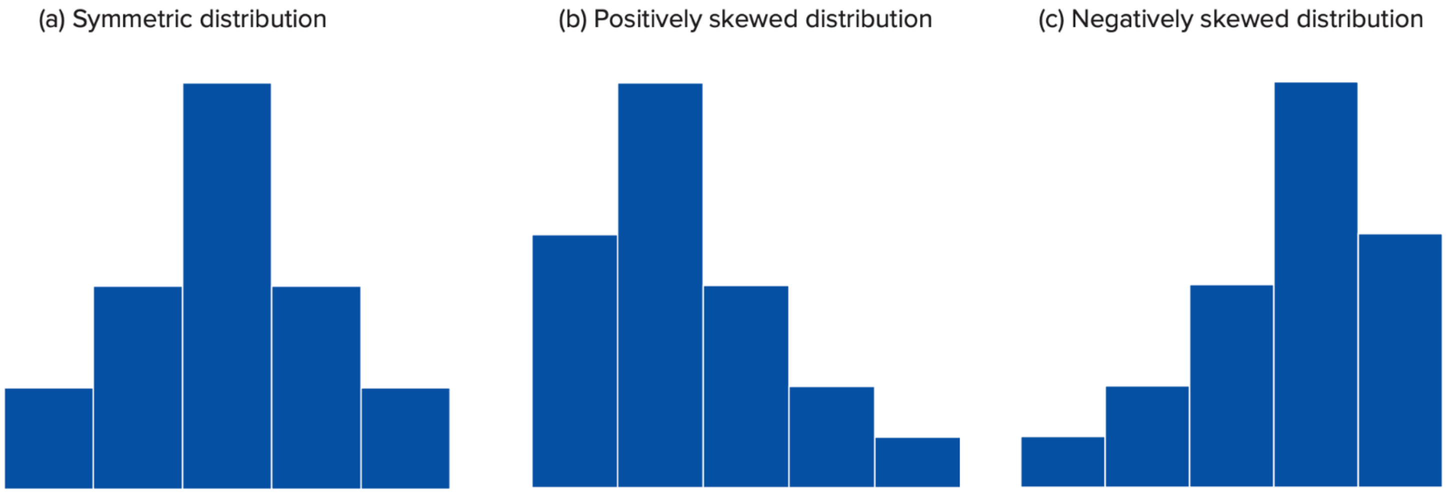
- A series of rectangles.
- Mark off the along the horizontal axis.
- The height of each bar represents the frequency or relative frequency for each interval.
- No gaps between bars/intervals.

2.3 Methods to Visualize a Numeric Variable 8

A histogram allows us to quickly see where most of the observations tend to cluster.

A histogram indicates the spread and shape of the variable.

- Symmetric: mirror image of itself on both sides of its center.
- Skewed: positive (elongated right tail) or negative (elongated left tail).

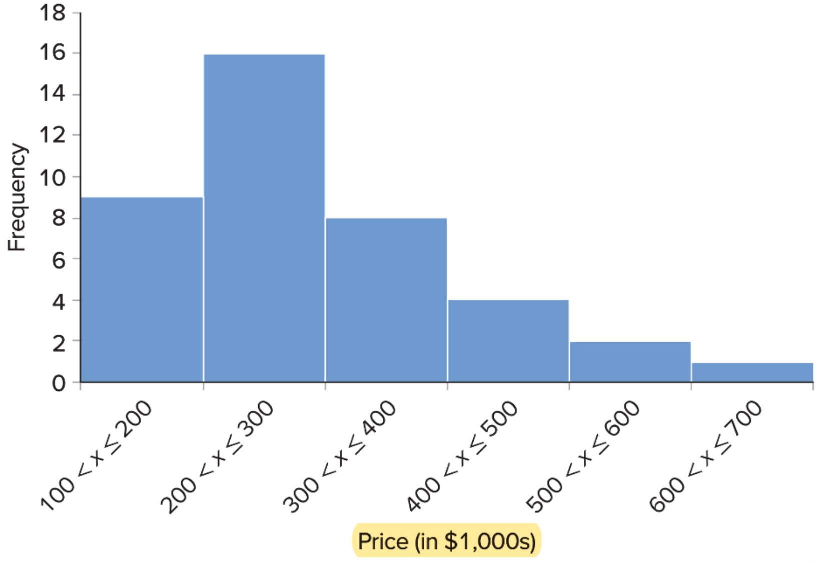


Access the text alternative for slide images.

2.3 Methods to Visualize a Numeric Variable [9]

- Example: house prices in Punta Gorda.

Interval (in $1,000s)	Frequency
100 < x ≤ 200	9
200 < x ≤ 300	16
300 < x ≤ 400	8
400 < x ≤ 500	4
500 < x ≤ 600	2
600 < x ≤ 700	1



Access the text alternative for slide images.

2.3 Methods to Visualize a Numeric Variable 10

A polygon provides another way of depicting a frequency distribution.

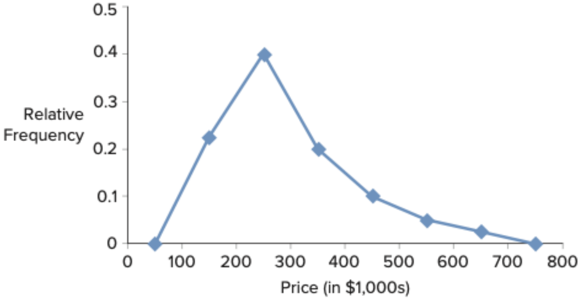
- Midpoint of each interval/class on the x-axis.
- Frequency or relative frequency on the y-axis.
- Connect neighboring points with a straight line.

A polygon gives a general idea about the shape of a distribution.

2.3 Methods to Visualize a Numeric Variable 11

- Example: house prices in Punta Gorda.

Interval	X-coordinate (midpoint)	Y-coordinate (relative frequency)
0 < x ≤ 100	50	0
100 < x ≤ 200	150	0.225
200 < x ≤ 300	250	0.400
300 < x ≤ 400	350	0.200
400 < x ≤ 500	450	0.100
500 < x ≤ 600	550	0.050
600 < x ≤ 700	650	0.25
700 < x ≤ 800	750	0



Access the text alternative for slide images.

2.3 Methods to Visualize a Numeric Variable 12

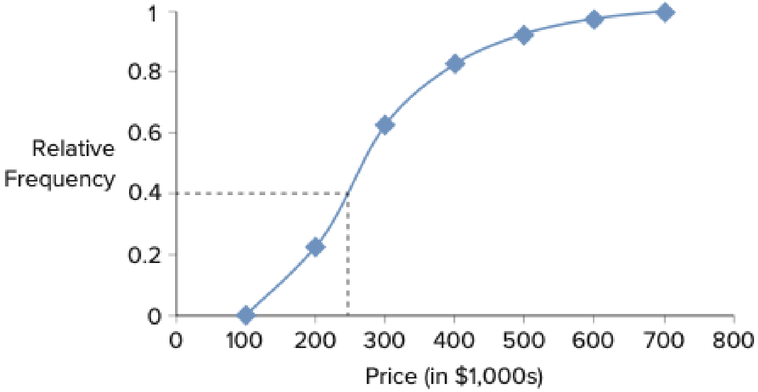
An ogive depicts a cumulative frequency or cumulative relative frequency.

- Upper limit of each interval/class on the x-axis.
- Cumulative frequency or cumulative relative frequency on the y-axis.
- Connect neighboring points with a straight line.
- Close the ogive at the lower end by intersecting the x-axis at the lower limit of the first interval.

2.3 Methods to Visualize a Numeric Variable

- Example: house prices in Punta Gorda.

Interval	X-coordinate (Upper limit)	Y-coordinate (cumulative relative frequency)
Lower limit of first class	100	0
100 < x ≤ 200	200	0.225
200 < x ≤ 300	300	0.625
300 < x ≤ 400	400	0.825
400 < x ≤ 500	500	0.925
500 < x ≤ 600	600	0.975
600 < x ≤ 700	700	1



Access the text alternative for slide images.

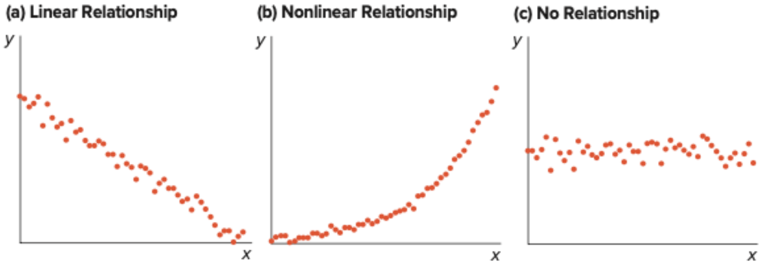
2.4 More Data Visualization Methods [1]

Use a scatterplot to examine the relationship between two numerical variables.

- Determine if two numerical variables are related in some systematic way.
- Each point represents a pair of observations of the two variables.
- Refer to one variable as *x* (x-axis) and the other as *y* (y-axis).

Once plotted, the graph may reveal one of the below.

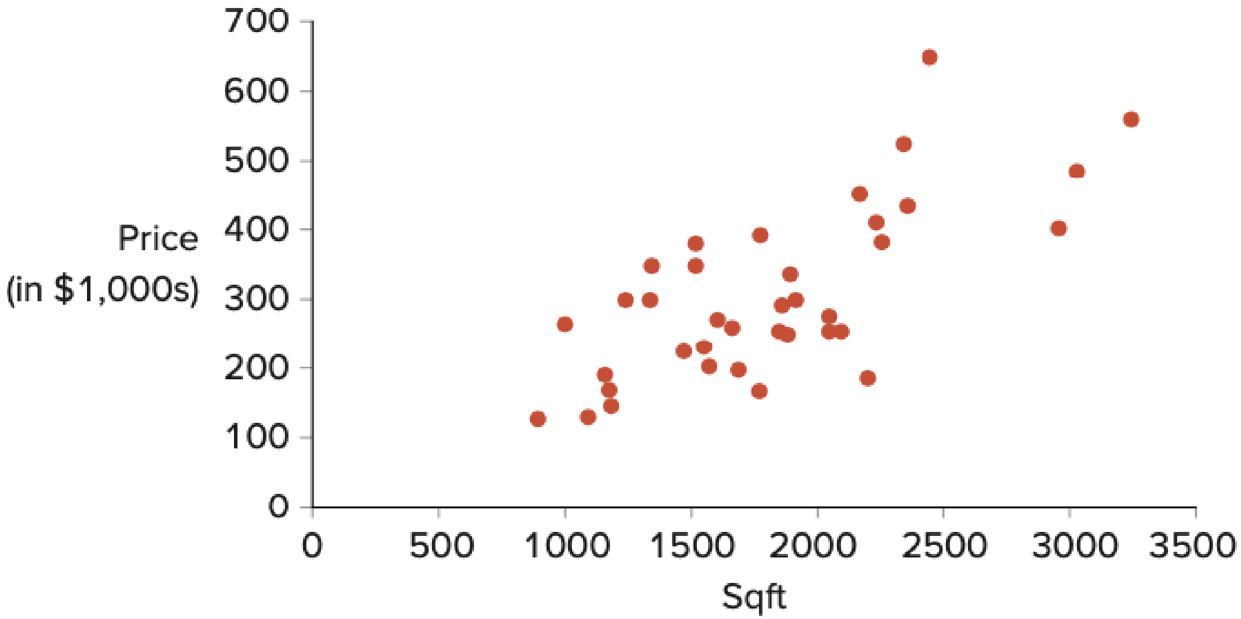
- A linear relationship.
- A nonlinear relationship.
- No relationship.



Access the text alternative for slide images.

2.4 More Data Visualization Methods 2

- Example: house prices and square footage in Punta Gorda.



Access the text alternative for slide images.

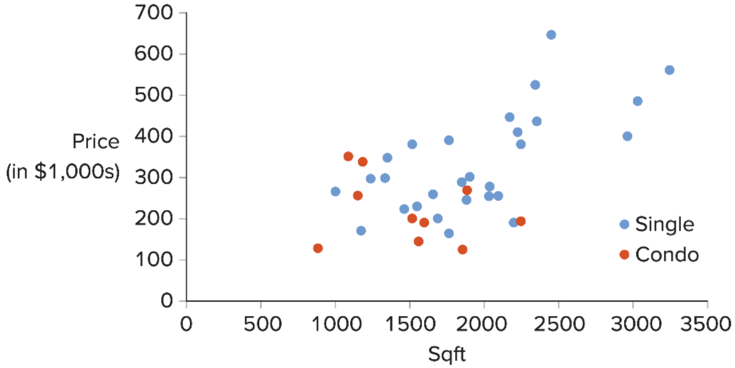
2.4 More Data Visualization Methods 3

A scatterplot with a categorical variable modifies a basic scatterplot.

- Incorporate a categorical variable in addition to the two numeric variables.
- Encode the categorical variable with color.
- Giving each point a distinct hue makes it easy to show its category.

This allows you to determine if the relationship between *x* and *y* differs across the values of the categorical variable.

Example: house prices and square footage by type in Punta Gorda.



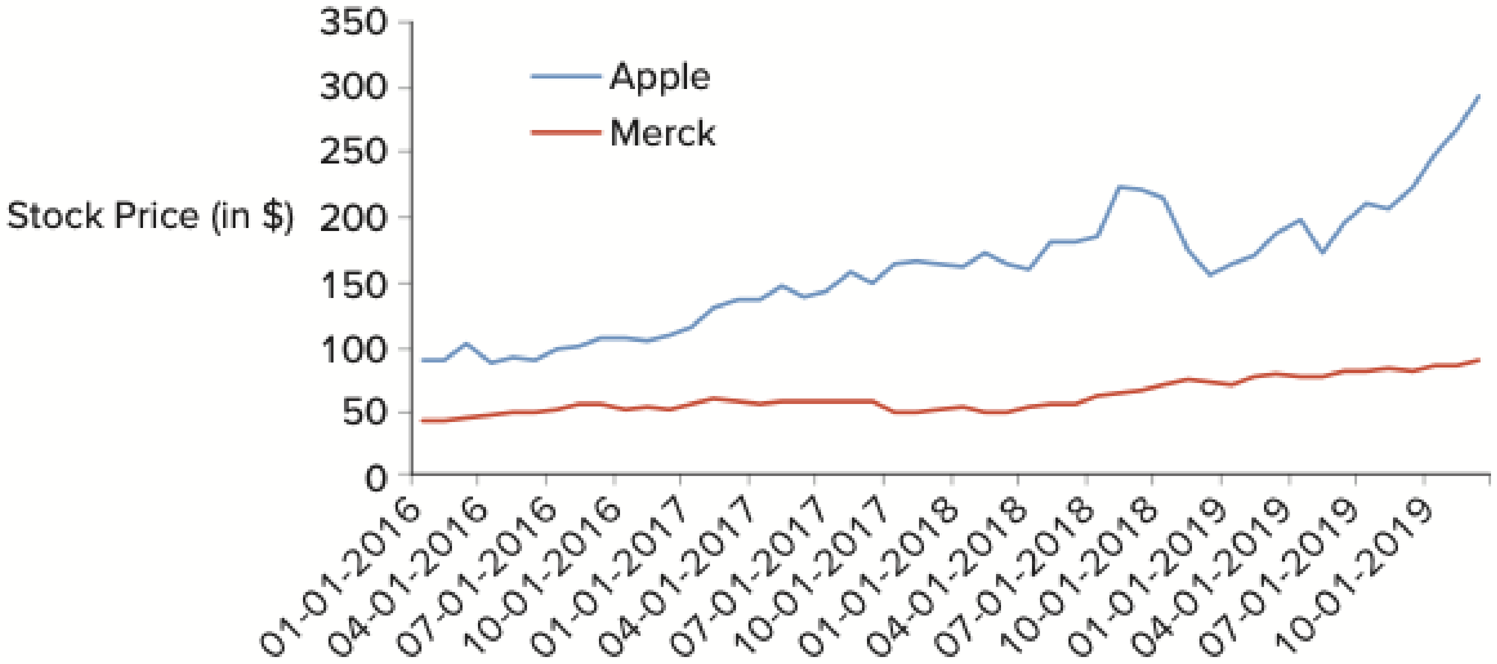
Access the text alternative for slide images.

2.4 More Data Visualization Methods 4

- A line chart displays a numerical variable as a series of consecutive observations connected by a line.
- A line chart is especially useful for tracking changes or trends over time.
- It is also easy for us to identify any major changes that happened in the past on a line chart.
- When multiple lines are plotted in the same chart, we can compare these observations on one or more dimensions.

2.4 More Data Visualization Methods 5

- Example: monthly stock prices for Apple and Merck.



Access the text alternative for slide images.

2.5 A Stem-And-Leaf Diagram 1

A stem-and-leaf diagram provides another visual method for displaying a numerical variable.

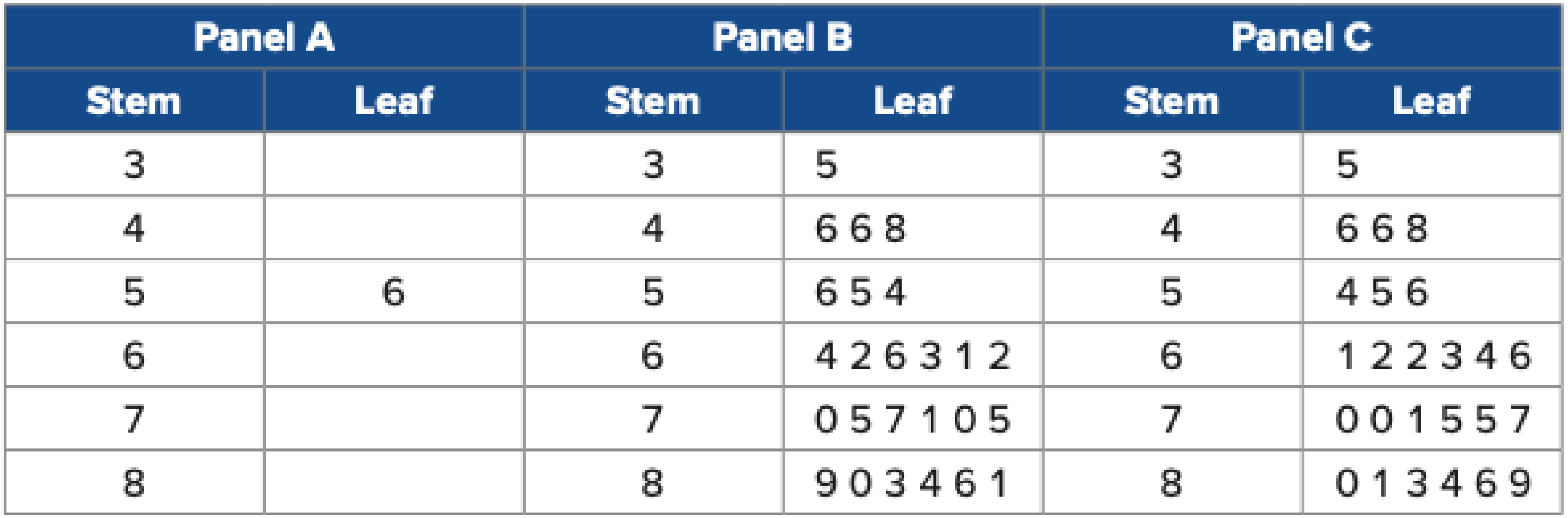
It gives an overall picture of where the observations are centered and how they are dispersed from the center.

Separate each observation of a variable into two parts.

- Stem: left-most digits.
- Leaf: the last digit.

2.5 A Stem-And-Leaf Diagram $_2$

- Example: age of the wealthiest people in the world.



Panel A		Panel B		Panel C	
Stem	Leaf	Stem	Leaf	Stem	Leaf
3		3	5	3	5
4		4	6 6 8	4	6 6 8
5	6	5	6 5 4	5	4 5 6
6		6	4 2 6 3 1 2	6	1 2 2 3 4 6
7		7	0 5 7 1 0 5	7	0 0 1 5 5 7
8		8	9 0 3 4 6 1	8	0 1 3 4 6 9

Access the text alternative for slide images.

End of Main Content

3

Numerical Descriptive Measures

Business Statistics:
Communicating with Numbers, 4e

By Sanjiv Jaggia and Alison Kelly

Chapter 3 Learning Objectives (LOs)

LO 3.1 Calculate and interpret measures of central location.

LO 3.2 Interpret a percentile and a boxplot.

LO 3.3 Calculate and interpret a geometric mean return and an average growth rate.

LO 3.4 Calculate and interpret measures of dispersion.

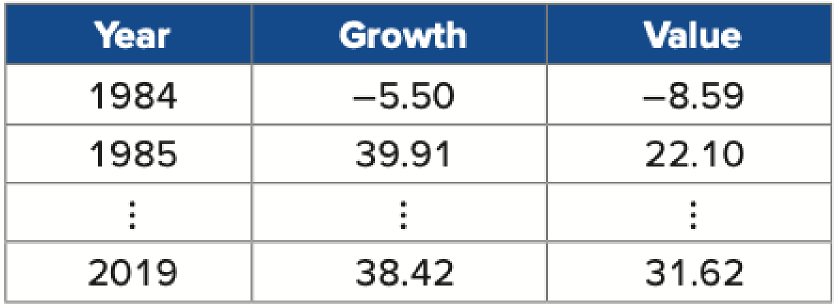
LO 3.5 Explain mean-variance analysis and the Sharpe ratio.

LO 3.6 Apply Chebyshev’s theorem, the empirical rule, and *z*-scores.

LO 3.7 Calculate and interpret measures of association.

Introductory Case: Investment Decision

- Dorothy works as a financial advisor at a large investment firm.
- She meets with an inexperienced investor who has some questions regarding two approaches to mutual fund investing.
- The investor shows Dorothy the annual return data for Fidelity's Growth Index mutual fund (Growth) and Fidelity's Value Index mutual fund (Value).



Year	Growth	Value
1984	−5.50	−8.59
1985	39.91	22.10
⋮	⋮	⋮
2019	38.42	31.62

- Dorothy will use the sample information for the following tasks.
  1. Calculate and interpret the typical return for these two mutual funds.
  2. Calculate and interpret the investment risk for these two mutual funds.
  3. Determine which mutual fund provides the greater return relative to risk.

3.1 Measures of Central Location (1)

- The term central location refers to how numerical data tend to cluster around some middle or central value.
- Measures of central location attempt to find a typical or central value that describes the data.
- The arithmetic mean is the primary measure of central location.
  - Referred to as the mean or the average.
  - Add up all the observations and divide by the number of observations.

3.1 Measures of Central Location (2)

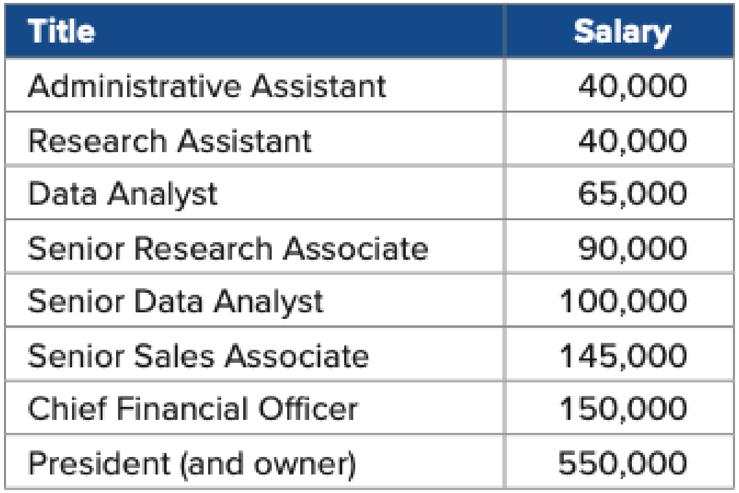
- The only thing that differs between a population mean and a sample mean is the notation.
- The population mean is denoted as μ.
  - N observations in the population: $x_1, x_2, \ldots, x_N$
  - $\mu = \frac{\sum x_i}{N}$
  - μ is a parameter (describes a population)
- The sample mean is dented as $\bar{x}$.
  - n observations in the sample: $x_1, x_2, \ldots, x_n$
  - $\bar{x} \ = \frac{\sum x_i}{n}$
  - $\bar{x}$ is a statistic (describes a sample)
  - Can be misleading in the presence of extremely small or large observations, or outliers

3.1 Measures of Central Location (3)

- Example: the mean return for Growth and the mean return for Value
- Growth: $\frac{(-5.50)+39.91+\cdots+38.42}{36} = 15.775$
- Value: $\frac{(-8.59)+22.10+\cdots+31.62}{36} = 12.005$
- Over the 36-year period, the mean return for Growth was greater than the mean return for Value.
- As we will see, we would be ill-advised to invest in a mutual fund solely on the basis of its average return.

3.1 Measures of Central Location (4)

- Example: salaries of employees at Acetech



Title	Salary
Administrative Assistant	40,000
Research Assistant	40,000
Data Analyst	65,000
Senior Research Associate	90,000
Senior Data Analyst	100,000
Senior Sales Associate	145,000
Chief Financial Officer	150,000
President (and owner)	550,000

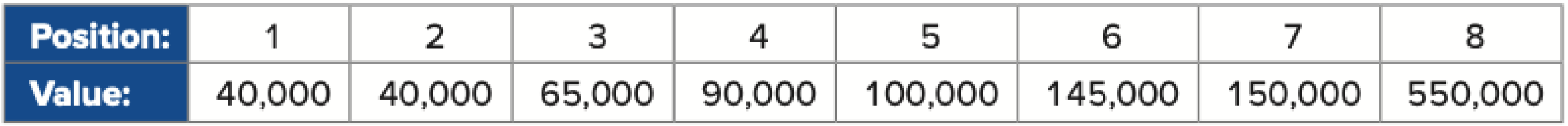
- Since these are all employees, calculate the population mean.
- This value does not reflect the typical salary; 6 of the 8 employees earn less than the mean.

3.1 Measures of Central Location (5)

- Since the mean can be affected by outliers, we often calculate the median.
- The median is the middle value of a variable.
  - It divides the data in half
  - An equal number of observations lie above and below the median
  - The middle value if *n* (or *N*) is odd
  - The average of the two middle values if *n* (or *N*) is even
- The median is especially useful when outliers are present.
- The mean and median are both typically published.
- If they differ, then the variable likely contains outliers.

3.1 Measures of Central Location (6)

- Example: the median salary of Acetech employees
- Arrange the data in ascending order



Position:	1	2	3	4	5	6	7	8
Value:	40,000	40,000	65,000	90,000	100,000	145,000	150,000	550,000

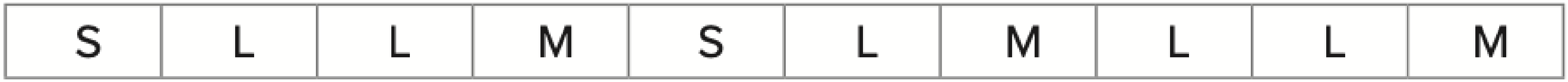
- Since there are 8 salaries, the median is the average of the observations in the 4th and 5th positions.
- The median is $\frac{90{,}000+100{,}000}{2} = \$95{,}000.$
- Four salaries are less than $95,000 and four salaries are greater than $95,000.
- Compared to the mean of $147,500, the median better reflects the typical salary.

3.1 Measures of Central Location (7)

- The mode of a variable is the observation that occurs most frequently.
- There can be one or more modes, or even no mode.
  - One mode: unimodal
  - Two modes: bimodal
- The mode is a less useful measure of centrality when there are more than three modes.
- Example: modal salary for employees at Acetech
  - $40,000 is earned by two employees, every other salary occurs just once
  - $40,000 is the mode
  - Most employees earn considerably more than this
- Just because an observation occurs with the most frequency does not guarantee that it best reflects the center of the variable.

3.1 Measures of Central Location (8)

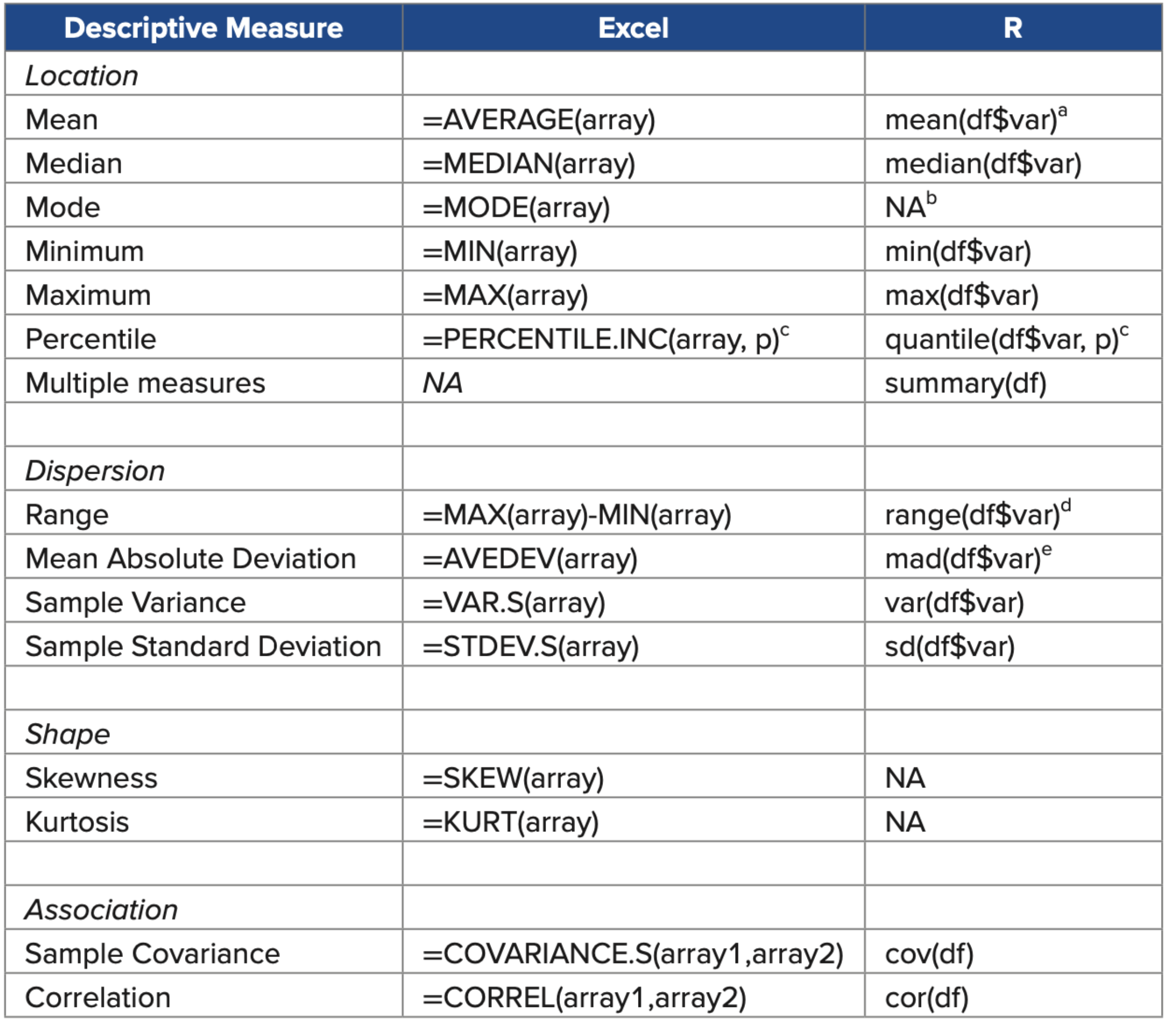
- To summarize a categorical variable, the mode is the only meaningful measure of central location.
- Example: sizes of women’s sweatshirts



S	L	L	M	S	L	M	L	L	M

  - The sizes are S, M, or L.
  - S and M appear 2 times each, L appears 5 times.
  - The modal size is L.

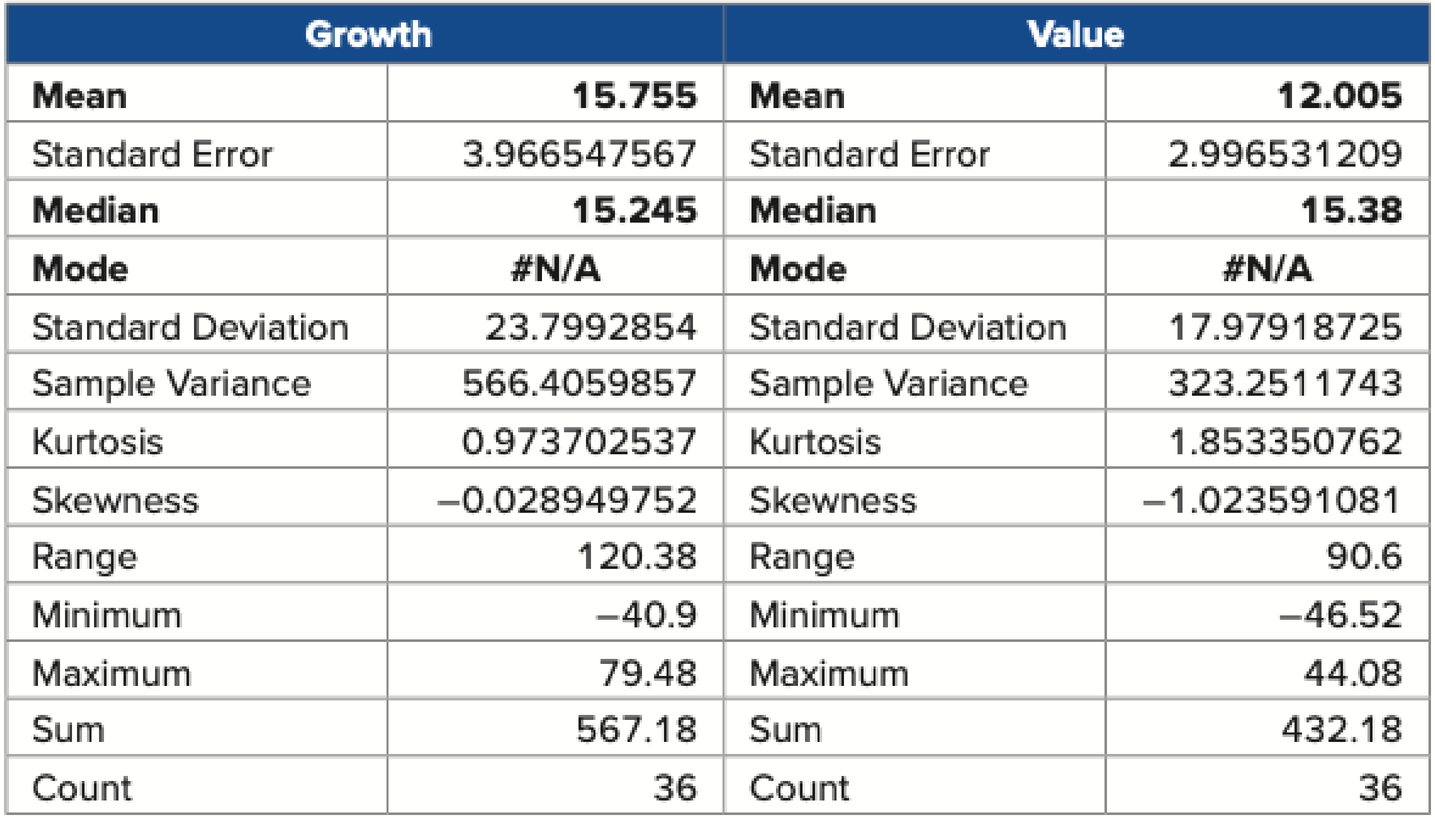
3.1 Measures of Central Location (9)



Descriptive Measure	Excel	R
Location		
Mean	=AVERAGE(array)	mean(df$var)[a]
Median	=MEDIAN(array)	median(df$var)
Mode	=MODE(array)	NA[b]
Minimum	=MIN(array)	min(df$var)
Maximum	=MAX(array)	max(df$var)
Percentile	=PERCENTILE.INC(array, p)[c]	quantile(df$var, p)[c]
Multiple measures	*NA*	summary(df)
Dispersion		
Range	=MAX(array)-MIN(array)	range(df$var)[d]
Mean Absolute Deviation	=AVEDEV(array)	mad(df$var)[e]
Sample Variance	=VAR.S(array)	var(df$var)
Sample Standard Deviation	=STDEV.S(array)	sd(df$var)
Shape		
Skewness	=SKEW(array)	NA
Kurtosis	=KURT(array)	NA
Association		
Sample Covariance	=COVARIANCE.S(array1,array2)	cov(df)
Correlation	=CORREL(array1,array2)	cor(df)

3.1 Measures of Central Location (10)

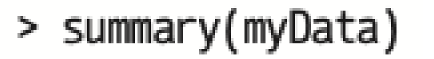
- Example: measures of centrality for Growth and Value
- With Excel



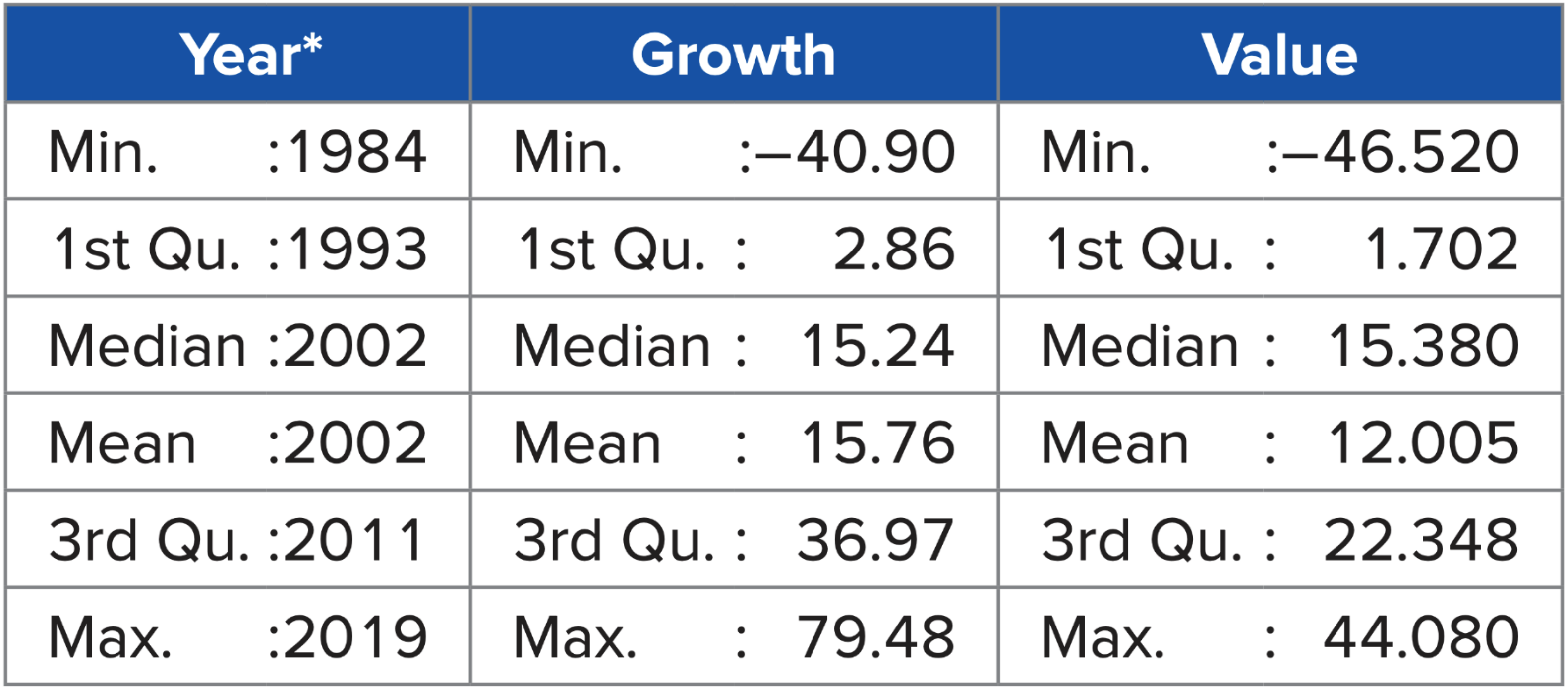
Growth		Value	
Mean	**15.755**	**Mean**	**12.005**
Standard Error	3.966547567	Standard Error	2.996531209
Median	**15.245**	**Median**	**15.38**
Mode	**#N/A**	**Mode**	**#N/A**
Standard Deviation	23.7992854	Standard Deviation	17.97918725
Sample Variance	566.4059857	Sample Variance	323.2511743
Kurtosis	0.973702537	Kurtosis	1.853350762
Skewness	–0.028949752	Skewness	–1.023591081
Range	120.38	Range	90.6
Minimum	–40.9	Minimum	–46.52
Maximum	79.48	Maximum	44.08
Sum	567.18	Sum	432.18
Count	36	Count	36

3.1 Measures of Central Location (11)

- Example: measures of centrality for Growth and Value
- With R



```
> summary(myData)
```

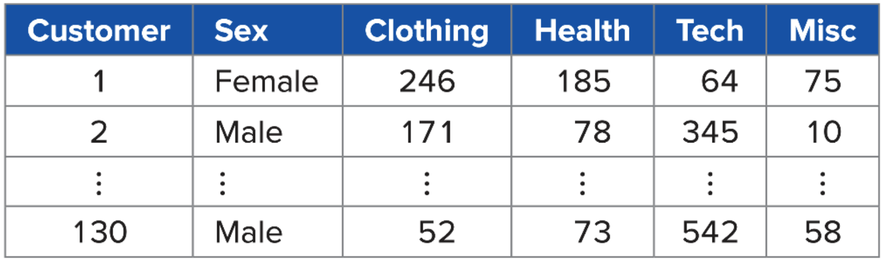


Year*	Growth	Value
Min. :1984	Min. :−40.90	Min. :−46.520
1st Qu. :1993	1st Qu. : 2.86	1st Qu. : 1.702
Median :2002	Median : 15.24	Median : 15.380
Mean :2002	Mean : 15.76	Mean : 12.005
3rd Qu. :2011	3rd Qu. : 36.97	3rd Qu. : 22.348
Max. :2019	Max. : 79.48	Max. : 44.080

*Note that in this example, the summary statistics for the variable Year are not useful.

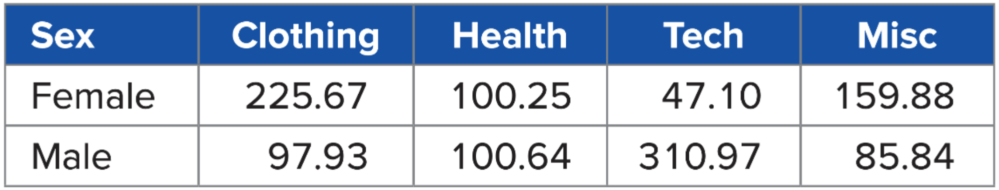
3.1 Measures of Central Location (12)

- It is sometimes useful to subset the observations, and compute means for each subset.
- Example: mean spending by sex



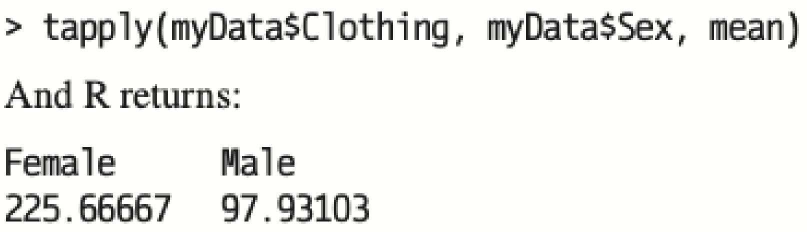
Customer	Sex	Clothing	Health	Tech	Misc
1	Female	246	185	64	75
2	Male	171	78	345	10
⋮	⋮	⋮	⋮	⋮	⋮
130	Male	52	73	542	58

  - With Excel



Sex	Clothing	Health	Tech	Misc
Female	225.67	100.25	47.10	159.88
Male	97.93	100.64	310.97	85.84

  - With R



```
> tapply(myData$Clothing, myData$Sex, mean)
```

And R returns:

```
Female     Male
225.66667  97.93103
```

3.1 Measures of Central Location (13)

- So far each observation of a variable contributed equally to the mean.
- The weighted mean is relevant when some observations contribute more than others.
- Let $w_1, w_2, \cdots, w_n$ denote the weight of sample observations $x_1, x_2, \cdots, x_n$ such that $w_1 + w_2 + \cdots + w_n = 1$.
- The weighted mean is computed as $\bar{x} = \sum w_i x_i$.
- For a frequency distribution, we substitute the relative frequencies for w_i and the midpoint of each interval for x_i.
- The weighted mean for the population is computed similarly.

3.1 Measures of Central Location (14)

- Example: a student scores 60 on Exam 1, 70 on Exam 2, and 80 on Exam 3. What is the student’s average score for the course if Exams 1, 2, and 3 are worth 25%, 25%, and 50% of the grade, respectively?
- Let $w_1 = 0.25$, $w_2 = 0.25$, $w_3 = 0.50$
- The average score is the weighted mean.

$$\bar{x} = 0.25(60) + 0.25(70) + 0.50(80) = 75.50$$

- Note the unweighted mean is only 70.

3.1 Measures of Central Location (15)

- A distribution is symmetric if one side of the histogram is a mirror image of the other.
- For a symmetric symmetric and unimodal distribution, the mean, median, and mode are equal.
- Positively skewed: mean is usually greater than the median
- Negatively skewed: the mean is usually less than the median
- The skewness coefficient is a measure of skewness.
  - Zero: observations are evenly distributed on both sides of the mean (symmetric)
  - Positive: extreme observations in the right tail, pulling the mean up relative to the median
  - Negative: extreme observations in the left tail, pulling the mean down relative the median

3.2 Percentiles and Boxplots (1)

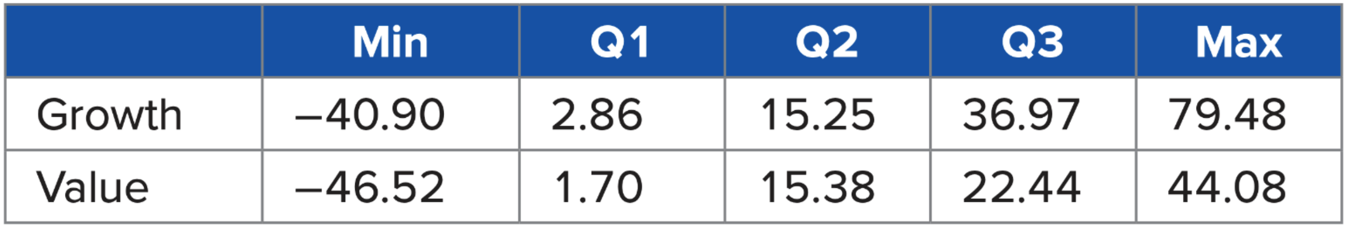
- A percentile divides the data into two parts.
  - Approximately *p* percent of the observations are less than the p^{th} percentile
  - Approximately *1-p* percent of the observations are greater than the p^{th} percentile
- A percentile is a measure of location.
- It is also used as a measure of relative position because it is easy to interpret.

3.2 Percentiles and Boxplots (2)

- When we calculate the 25th, 50th, and 75th percentiles for a variable, we have divided it into four equal parts or quarter.
  - 25th percentile: Q1
  - 50th percentile: Q2
  - 75th percentile: Q3
- It only makes sense to calculate percentile for larger data sets.
- Rely on Excel and R to compute these.
  - Different algorithms so the values might be slightly different
  - With larger sample size, the differences tend to be negligible
- A five-number summary is commonly reported: the minimum, quartiles, and the maximum.
- Note IQR = Q3 – Q1: range of the middle 50% of values

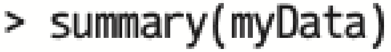
3.2 Percentiles and Boxplots (3)

- Example: five-number summary for the Growth and Value variables.
- With Excel

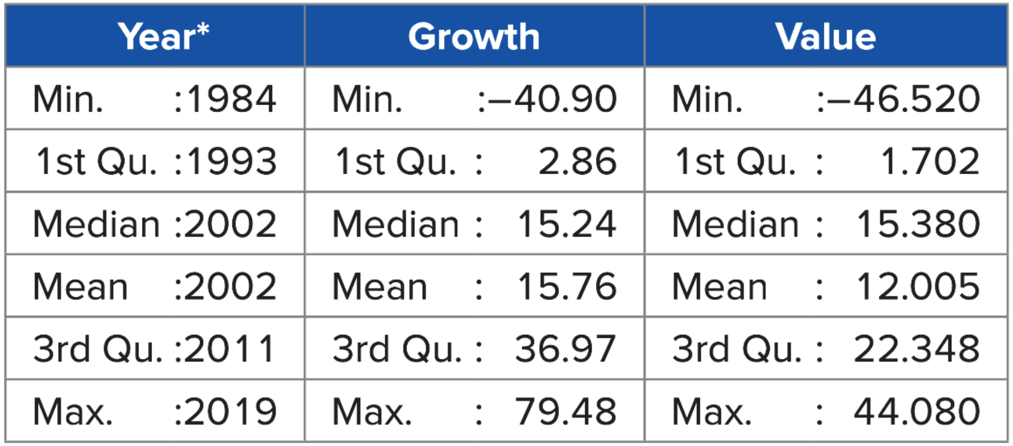


	Min	Q1	Q2	Q3	Max
Growth	–40.90	2.86	15.25	36.97	79.48
Value	–46.52	1.70	15.38	22.44	44.08

- With R



```
> summary(myData)
```



Year*	Growth	Value
Min. :1984	Min. :–40.90	Min. :–46.520
1st Qu. :1993	1st Qu. : 2.86	1st Qu. : 1.702
Median :2002	Median : 15.24	Median : 15.380
Mean :2002	Mean : 15.76	Mean : 12.005
3rd Qu. :2011	3rd Qu. : 36.97	3rd Qu. : 22.348
Max. :2019	Max. : 79.48	Max. : 44.080

*Note that in this example, the summary statistics for the variable Year are not useful.

3.2 Percentiles and Boxplots (4)

- A boxplot, also referred to as a box-and-whisker plot, is a way to graphically display a five-number summary.

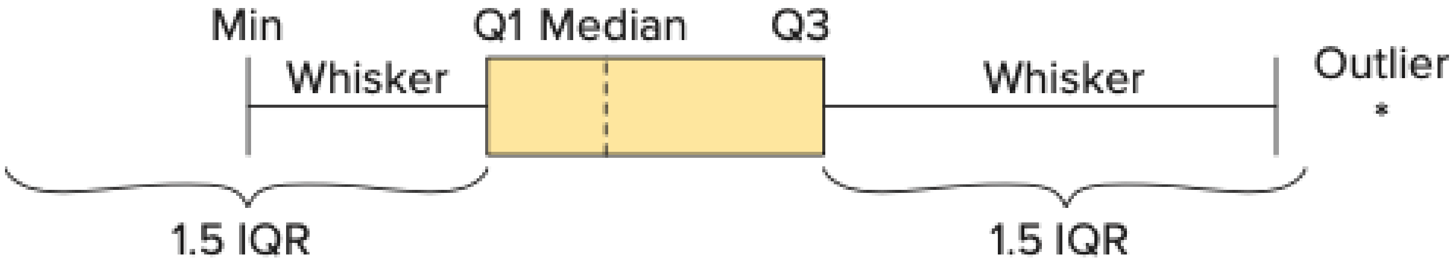
a. Plot the five-number summary values in ascending order on the horizontal axis.

b. Draw a box encompassing the first and third quartiles.

c. Draw a dashed vertical line in the box at the median.

d. Draw a line (“whisker”) that extends from Q1 to the minimum value that is not further from 1.5*IQR from Q1 (similarly for the other side).

e. Use an asterisk (or another symbol) to indicate observations that are farther than 1.5*IQR from the box (outliers).

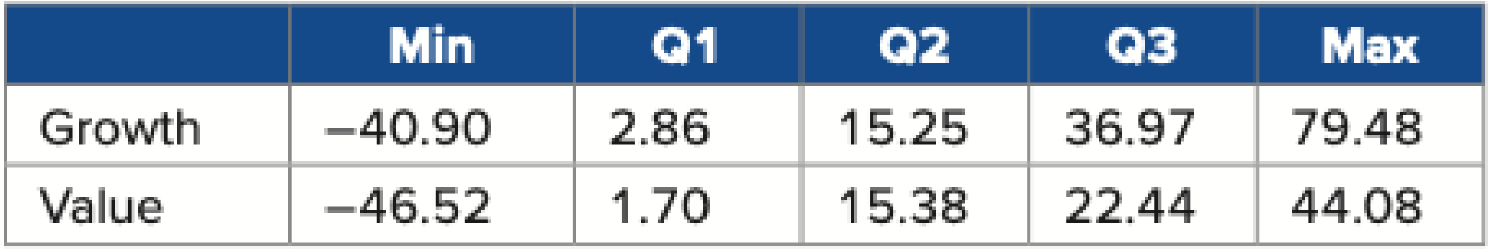


3.2 Percentiles and Boxplots (5)

- A boxplot is used to informally gauge the shape of the distribution.
  - Symmetry: median is in the center of the box, and the left and right whiskers are equally distant from their respective quartiles
  - Positively skewed: median is left of center and the right whisker is longer than the left whisker
  - Negatively skewed: median is right of center and the left whisker is longer than the right whisker
- If outliers exist, we need to include them when comparing the length of the whiskers.
- Excel does not provide a simple and straightforward way to construct a boxplot.
- R and other statistical packages offer this option.

3.2 Percentiles and Boxplots (6)

- Example: using the five-number summaries for Growth and Value



	Min	Q1	Q2	Q3	Max
Growth	−40.90	2.86	15.25	36.97	79.48
Value	−46.52	1.70	15.38	22.44	44.08

- Find the IQR for Growth.
  - Determine whether any outliers exist
  - Repeat for Value
- Is the distribution for Growth symmetric?
  - If not, comment on its skewness
  - Repeat for value
- Use R to construct boxplots for Growth and Value.
- Are the results consistent with the previous parts?

3.2 Percentiles and Boxplots (7)

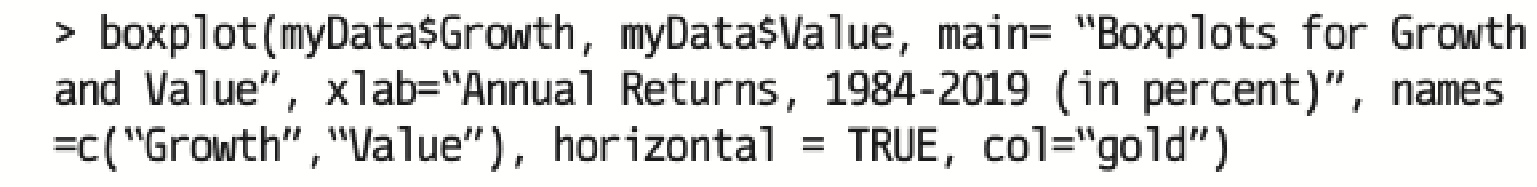
- Example continued
- Growth
  - IQR = Q3 − Q1 = 36.94 − 2.86 = 34.11
  - Limit = 1.5*IQR = 1.5*34.11 = 51.17
  - Q1− Min = 2.86 − (−40.90) = 43.76 (left whisker)
  - Max − Q3 = 79.48 − 36.94 = 42.51 (right whisker)
  - Both are less than 51.17, no outliers
- Value
  - IQR = Q3 − Q1 = 22.44 − 1.70 = 20.74
  - Limit = 1.5*IQR = 1.5*20.74 = 31.11
  - Q1- Min = 1.70 − (−46.52) = 48.22 (left whisker)
  - Max − Q3 = 44.08 − 22.44 = 21.64 (right whisker)
  - Left whisker limit exceeds the limit, there is an outlier(s) on the left side

3.2 Percentiles and Boxplots (8)

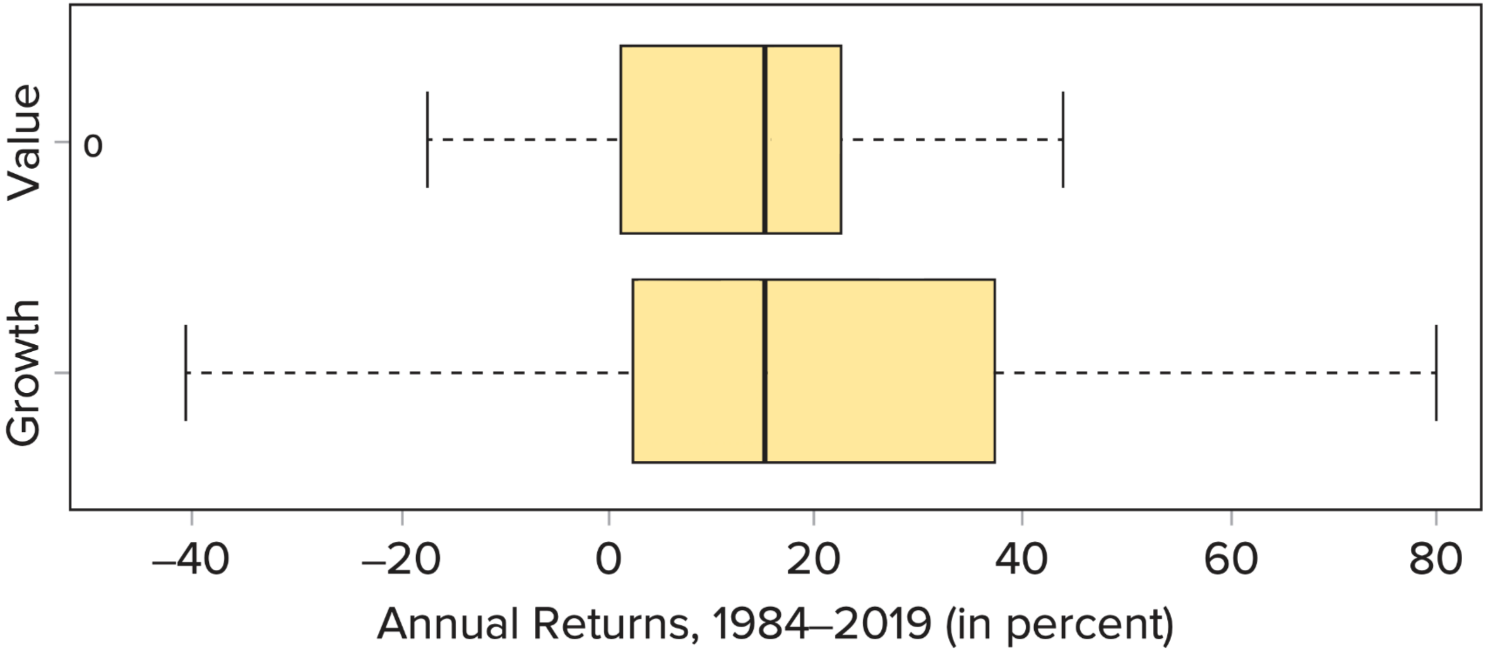
- Example continued
- Growth
  - Median − Q1 = 15.25 − 2.86 = 12.39
  - Q3 − Median = 36.97 − 15.25 = 21.72
  - Since 12.39 < 21.72, the median is left of center
  - Left whisker is longer than the right whisker
  - Skewness coefficient is negative
  - The distribution is negatively skewed
- Value (similar calculations)
  - Median falls right of center
  - Left whisker is longer than the right whisker
  - Skewness coefficient is negative
  - The distribution is negatively skewed

3.2 Percentiles and Boxplots (9)

- Example continued
- Boxplot with R



```
> boxplot(myData$Growth, myData$Value, main= “Boxplots for Growth
and Value”, xlab=“Annual Returns, 1984-2019 (in percent)”, names
=c(“Growth”,“Value”), horizontal = TRUE, col=“gold”)
```



3.3 The Geometric Mean (1)

- The arithmetic mean is additive.
  - Ignores the effects of compounding
  - Suitable for analyzing a one-year investment
- The geometric mean is a multiplicative average.
  - Smaller than the arithmetic mean
  - Less sensitive to outliers
- The geometric mean is
  - Relevant measure when evaluating investment returns over several years
  - Calculating the average growth rates
- For *n* multiperiod returns $R_1, R_2, \cdots, R_3$, the geometric mean return $G_R$is given by the below.

$$G_R = \sqrt[n]{(1 + R_1)(1 + R_2)\cdots(1 + R_n)} - 1$$

3.3 The Geometric Mean (2)

- Example: given an initial investment of $1,000 over two years

Year	Return %	Value at the End of Year
1	10	1,000 + 1,000 × 0.10 = 1,100
2	−10	1,100 + 1,100 × (−0.10) = 990

- The arithmetic mean return is $\bar{x} = \frac{10+(-10)}{2} = 0\%$
- The geometric mean is

$$G_R = \sqrt[2]{(1+0.10)(1+(-0.10))} - 1 = -0.005 \text{ or } -0.5\%$$

- We can interpret the geometric return as the annualized return from the two year-period.

Year	Annualized Return	Value at the End of Year
1	−0.5	1,000 + 1,000 × (−0.005) = 995
2	−0.5	995 + 995 × (−0.005) = 990

3.3 The Geometric Mean (3)

- Also use the geometric mean when we calculate an average growth rate.
- For *n* growth rates $g_1, g_2, \cdots, g_n$, the average growth rate is computed as $G_g = \sqrt[n]{(1+g_1)(1+g_2)\cdots(1+g_n)} - 1$.
- There is a simpler way to compute the average growth rate from the values rather than growth rates.
- For *n* observations $x_1, x_2, \cdots, x_n$, the average growth rate is computed using n-1 distinct growth rates.

$$G_g = \sqrt[n-1]{\frac{x_n}{x_{n-1}} \frac{x_{n-1}}{x_{n-2}} \frac{x_{n-2}}{x_{n-3}} \cdots \frac{x_2}{x_1}} - 1 = \sqrt[n-1]{\frac{x_n}{x_1}} - 1$$

3.3 The Geometric Mean (4)

- Example: sales for multinational corporation

Year	1	2	3	4	5
Sales	13,322	14,883	14,203	14,534	16,915

- The growth rate after five years is 6.15%.

$$G_g = \sqrt[5-1]{\frac{16{,}915}{13{,}322}} - 1 = 0.0615 \; or \; 6.15\%$$

3.4 Measures of Dispersion (1)

- Measures of central location do not describe the underlying dispersion.
- Measures of dispersion gauge the variability of a variable.
  - 0 indicates all the observations are identical
  - Increases as the observations become more diverse
- The range is the simplest measure.
  - Difference between the maximum and minimum
  - Range = Max – Min
  - Not good because it focuses solely on extreme observations
- The interquartile range (IQR) is the difference between the third quartile and the first quartile.
  - $IQR = Q_3 - Q_1$
  - The range of the middle 50% of the variable
  - Does not depend on the extreme observations
- Neither the range or IQR incorporate all the observations.

3.4 Measures of Dispersion (2)

- A good measure of dispersion should consider differences of all observations from the mean.
- If we average all the differences from the mean, the positives and negatives will cancel out.
- The mean absolute difference (MAD) is the average of the absolute differences between the observations and the mean.
- For sample observations $x_1, x_2, \cdots, x_n$, the sample MAD is computed as sample MAD = $\frac{\sum |x_i - \bar{x}|}{n}$.
- For population observations $x_1, x_2, \cdots, x_N$, the population MAD is computed as population MAD = $\frac{\sum |x_i - \mu|}{N}$.

3.4 Measures of Dispersion (3)

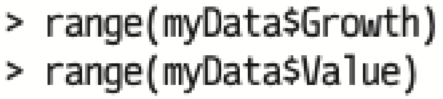
- The variance and the standard deviation are the two most widely used measures of dispersion.
  - Compute the average of the squared differences
  - The squaring of the differences emphasizes larger differences
- The variance is defined as the average of the squared differences between the observations and the mean.
- Whatever the units of original variable, the variance has squared units.
- To return to the original units, we take the positive square root of the variance which gives the standard deviation.

3.4 Measures of Dispersion (4)

- For sample observations $x_1, x_2, \cdots, x_n$, the sample variance s^2 and standard deviation s are:
  - $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$
  - $s = \sqrt{s^2}$
- For population observations $x_1, x_2, \cdots, x_N$, the population variance σ^2 and standard deviation σ are:
  - $\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$
  - $\sigma = \sqrt{\sigma^2}$
- The sample variance uses *n-1* rather than *n* in the denominator to ensure the sample variance is an unbiased estimator (Appendix 7.2).

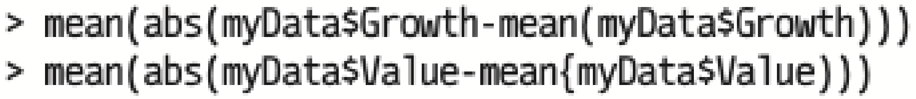
3.4 Measures of Dispersion (5)

- Summary of Excel and R Commands
- Range
  - Excel: MAX – MIN
  - R:



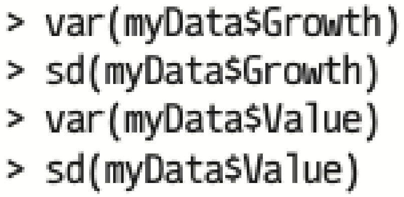
```
> range(myData$Growth)
> range(myData$Value)
```

- MAD
  - Excel: AVEDEF
  - R:



```
> mean(abs(myData$Growth-mean(myData$Growth)))
> mean(abs(myData$Value-mean{myData$Value)))
```

- Standard deviation and variance
  - Excel: VAR.S and STDEV.S
  - R:



```
> var(myData$Growth)
> sd(myData$Growth)
> var(myData$Value)
> sd(myData$Value)
```

3.4 Measures of Dispersion (6)

- Example: measures of dispersion for Growth and Value
- Range
  - Growth: 120.38
  - Value: 90.6
- MAD
  - Growth: 17.49
  - Value: 13.67
- Variance and standard deviation
  - Growth: 566.41, 23.80
  - Value: 323.25, 17.98

3.4 Measures of Dispersion (7)

- We want to compare the variability of two or more data sets that have different means or units.
- The coefficient of variation is the relevant measure.
  - Denoted CV
  - Adjusts for differences in the magnitudes of the means
  - Unitless, allowing easy comparisons of mean-adjusted dispersion across different data sets
- Sample CV: $\frac{s}{\bar{x}}$
- Population CV: $\frac{\sigma}{\mu}$

3.4 Measures of Dispersion (8)

- Example: CV for Growth and Value
- Growth CV: $\frac{s}{\bar{x}} = \frac{23.80}{15.755}$ =1.51
- Value CV: $\frac{s}{\bar{x}} = \frac{17.98}{12.005}$ =1.50
- The coefficient of variation indicates that the relative dispersion of the two variables is about the same.

3.5 Mean-Variance Analysis and the Sharpe Ratio (1)

- Investments include financial assets such as stocks, bonds, and mutual funds.
- The average return represents an investor’s reward, whereas variance, or equivalently standard deviation, corresponds to risk.
- Mean-variance analysis postulates the performance of an asset is measured by its rate of return, evaluated in terms of reward (mean) and risk (variance).
- Investments with higher returns also carry higher risk.

3.5 Mean-Variance Analysis and the Sharpe Ratio (2)

- The Sharpe ratio is the “reward-to-variability” ratio.
  - Characterizes how well the return compensates for the risk
  - Measures the extra reward per unit of risk
- Calculated as $\frac{\bar{x}_I - R_i}{s_I}$, where R_i is the mean return for a risk-free asset such as a Treasure bill (T-bill)
- The numerator measures the extra reward for the added risk, the difference is excess return
- The higher Sharpe ratio, the better the investment compensates its investors for risk.

3.5 Mean-Variance Analysis and the Sharpe Ratio (3)

- Example: compute the Sharpe ratios for the Growth and Value fund assuming $\bar{R}_f = 2\%$.
- Growth: $\frac{15.755-2}{23.799} = 0.58$
- Value: $\frac{12.005-2}{17.797} = 0.56$
- The Growth funds have a higher rate of return (good) along with a higher variance (bad).
- Growth funds offer more reward per unit risk than the Value funds.

3.6 Analysis of Relative Location (1)

- The mean and standard deviation alone don’t tell us about relative location.
  - Low standard deviation: observations are close to the mean
  - High standard deviation: observations are spread out
- Chebyshev’s Theorem
  - The proportion of observations that lie within k standard deviations from the mean is at least $1 - 1/k^2$, where k is any number greater than 1
  - At least 75% of the observations fall within 2 standard deviations from the mean
  - At least 89% of the observations fall within 3 standard deviations from the mean
  - Applies to all variables, regardless of the shape of the distribution
  - Conservative bounds for the percentage of observations falling into interval

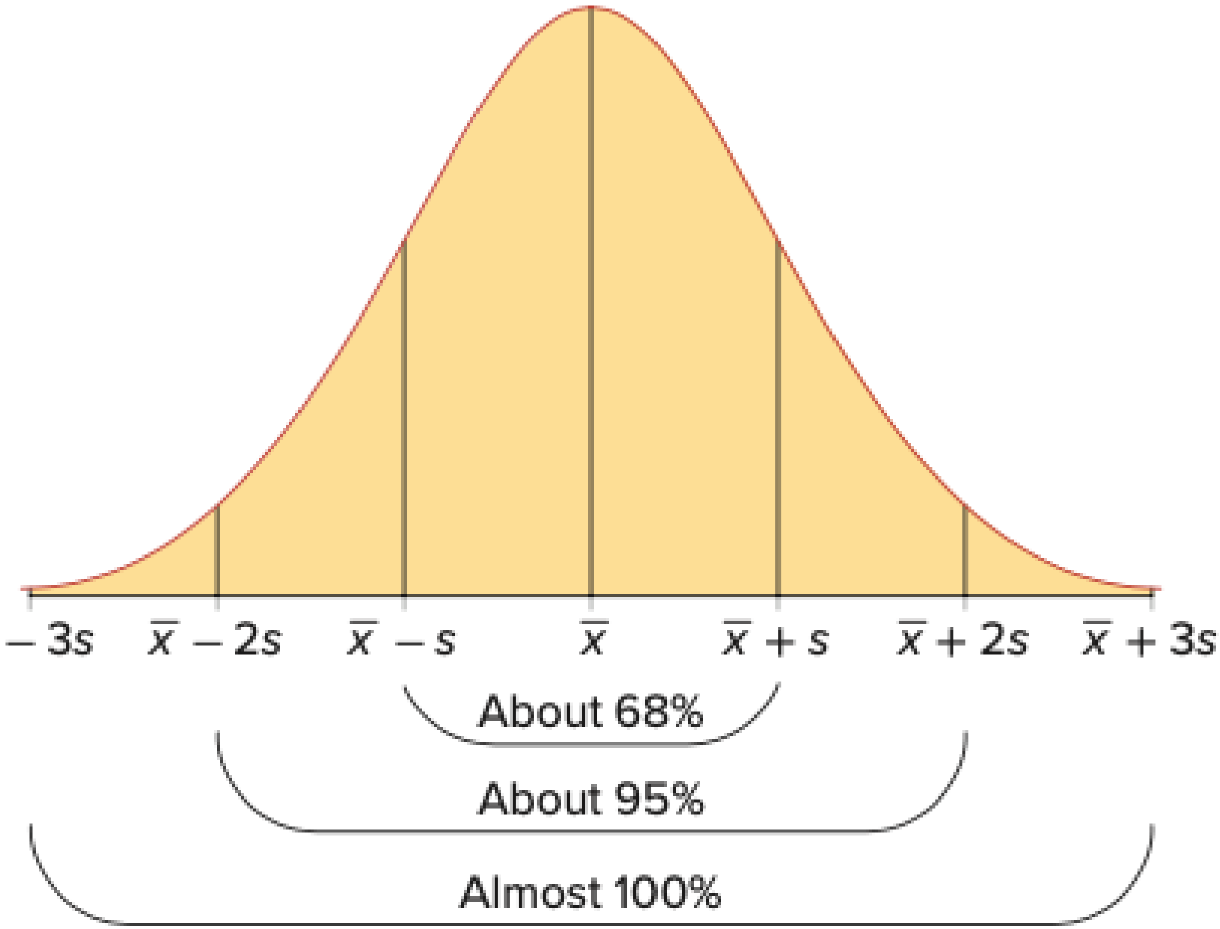
3.6 Analysis of Relative Location (2)

- Example: a large lecture class has 280 students.
  - The mean score on an exam is 74 with a standard deviation of 8.
  - At least how many students scored within 58 and 90?
- 58 and 90 are two standard deviations below and above the mean
  - $58 = 74 - 2 * 8$
  - $90 = 74 + 2 * 8$
- $k = 2$ and $1 - 1/2^2 = 0.75$
- At least 75% of the scores will fall within 58 and 90
- At least 75% of 280 students, or 210 students, scored within 58 and 90

3.6 Analysis of Relative Location (3)

- Chebyshev’s Theorem gives a conservative bound, the empirical rule makes more precise statements.
- Assume the observations are drawn from a relatively symmetric and bell-shaped distribution.
  - Inspection of a histogram
  - Other numerical measures (e.g. skewness)
- The empirical rule provides the approximate percentage of observations that fall within a specified number of standard deviations from the mean.
  - Approximately 68% of all observations fall in the interval $\bar{x} \pm s$
  - Approximately 95% of all observations fall in the interval $\bar{x} \pm 2s$
  - Almost all all observations fall in the interval $\bar{x} \pm 3s$
- If the distribution is symmetric and bell-shaped, the empirical rule is preferable to Chebyshev’s Theorem.

3.6 Analysis of Relative Location (4)



3.6 Analysis of Relative Location (5)

- Example: a large lecture class has 280 students.
  - The mean score on an exam is 74 with a standard deviation of 8.
  - Approximately how many students scored within 58 and 90?
  - Approximately how many students scored more than 90?
- 58 and 90 are two standard deviations below and above the mean.
- So about 95% of 280 students, or 266, scored within 58 and 90.
- 90 is two standard deviations above the mean.
  - 95% of the observations fall within two standard deviations of the mean, 5% fall outside the interval
  - Given the symmetry of the distribution, 2.5% scored above 90
  - 2.5% of the 280 students, or about 7, scored above 90 on the exam

3.6 Analysis of Relative Location (6)

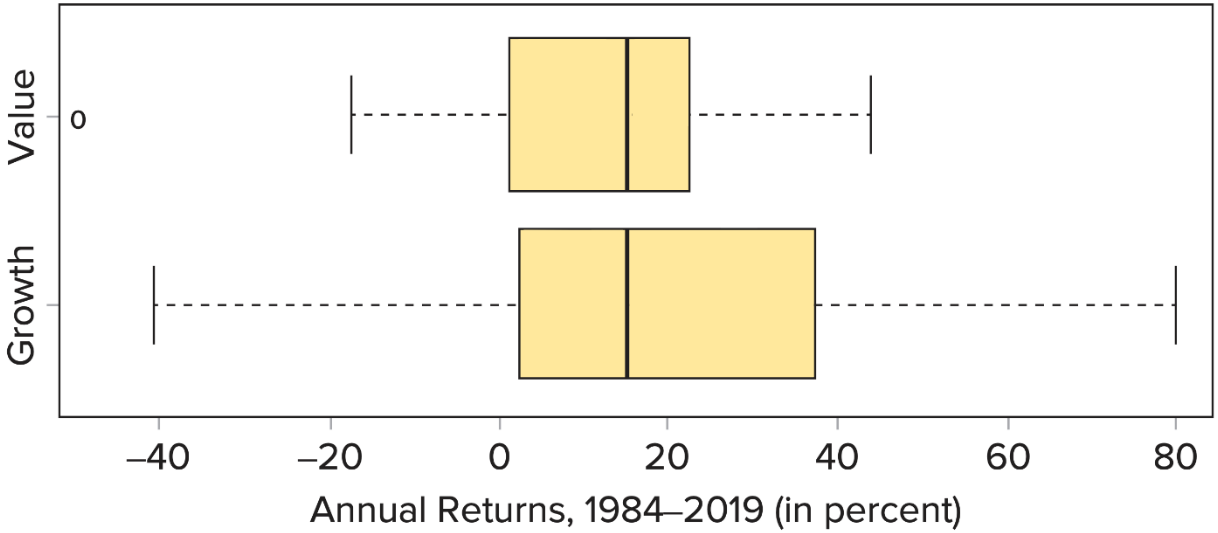
- Use the mean and the standard deviation to find the relative location of an observation within a distribution.
- The *z*-score gives the relative position of an observation within a distribution by $z = \frac{x - \bar{x}}{s}$.
  - Unitless measure
  - Measures the distance of an observation from the mean in terms of standard deviations
  - Also called standardizing

3.6 Analysis of Relative Location (7)

- If the distribution is relatively symmetric and bell-shaped, use *z*-scores to detect outliers.
- Since almost all observations are within three standard deviations of the mean, an observation is an outlier if its *z*-score is more than 3 or less than −3.
- Such observations must be reviewed to determine if they should remain in the data set.

3.6 Analysis of Relative Location (8)

- Example: compute the z-scores of the smallest and largest observations in Growth, and determine if they are outliers.
- Smallest: $z = \frac{-40.90-15.775}{23.799} = -2.38$
- Largest: $z = \frac{79.48-15.775}{23.799} = 2.68$
- Neither are outliers, which is consistent with the Growth's boxplot.



3.7 Measure of Association (1)

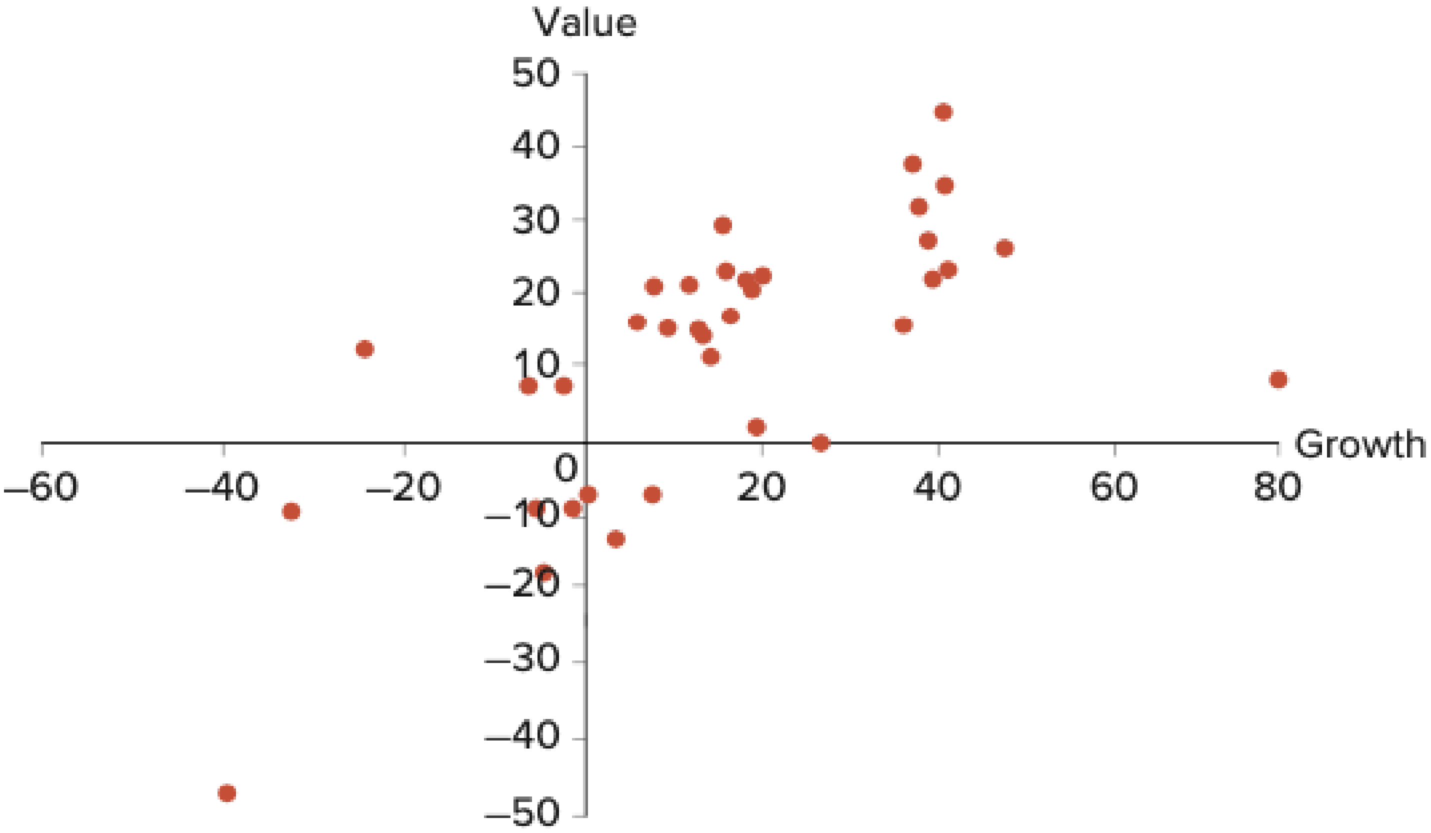
- In Chapter 2, we used a scatterplot to visually assess whether two variables had some type of linear relationship.
- If the relationship is linear, there are two numerical measures of association that quantify the direction and strength of the linear relationship.
- Covariance measures the direction of the linear relationship.
  - Population: $\sigma_{xy} = \frac{\sum(x_i - \mu_x)(y_i - \mu_y)}{N}$
  - Sample: $s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$
  - Negative: negative linear relationship
  - Positive: positive linear relationship
  - Zero: no linear relationship
- Covariance is hard to interpret because it is sensitive to the units of measurement. We cannot comment on the strength of the linear relationship.

3.7 Measure of Association (2)

- The correlation coefficient describes both the direction and strength of the linear relationship between *x* and *y*.
  - Population: $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$
  - Sample: $r_{xy} = \frac{s_{xy}}{s_x s_y}$
  - Negative: negative linear relationship
  - Positive: positive linear relationship
  - Zero: no linear relationship
- The correlation is unit-free.
- The correlation is between -1 and 1.
  - Correlation is -1: perfect negative linear relationship
  - Correlation is 0: not linearly related
  - Correlation is 1: perfect positive linear relationship

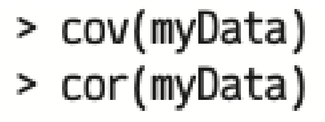
3.7 Measure of Association (3)

- Example: using Excel or R, find the covariance and the correlation coefficient for Growth and Value.

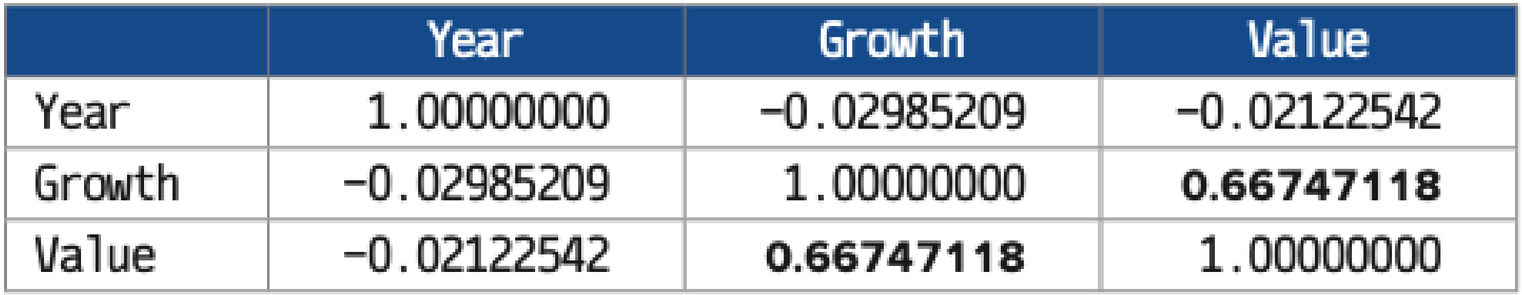


3.7 Measure of Association (4)

- Example: continued
- With Excel
  - COVARIANCE.S
  - CORREL
- With R



```
> cov(myData)
> cor(myData)
```



	Year	Growth	Value
Year	1.00000000	-0.02985209	-0.02122542
Growth	-0.02985209	1.00000000	**0.66747118**
Value	-0.02122542	**0.66747118**	1.00000000