RWTH BUSINESS SCHOOL

Mathematics & Statistics
M.Sc. Data Analytics and Decision Science

Prof. Dr. Thomas S. Lontzek

AACSB ACCREDITED

BUSINESS SCHOOL | RWTH AACHEN UNIVERSITY

Outline

- Data Sources
- Descriptive Statistics
- Statistical Inference
- Analytics
- Big Data and Data Mining
- Descriptive Statistics: Tabular and Graphical Displays
- Measures of Location
- Measures of Variability
- Measures of Distribution Shape, Relative Location, and Detecting Outliers.
- Measures of Association Between Two Variables

What is Statistics?

- The term “statistics” can refer to numerical facts such as averages, medians, percentages, and maximums that help us understand a variety of business and economic situations.
- Statistics can also refer to the art and science of collecting, analyzing, presenting, and interpreting data.

Application to Business and Economics

- Accounting
  - Public accounting firms use statistical sampling procedures when conducting audits for their clients.
- Economics
  - Economists use statistical information in making forecasts about the future of the economy or some aspect of it.
- Finance
  - Financial advisors use price-earnings ratios and dividend yields to guide their investment advice.

Application to Business and Economics

- Marketing
  - Electronic point-of-sale scanners at retail checkout counters are used to collect data for a variety of marketing research applications.
- Production
  - A variety of statistical quality control charts are used to monitor the output of a production process.
- Information Systems
  - A variety of statistical information helps administrators assess the performance of computer networks.

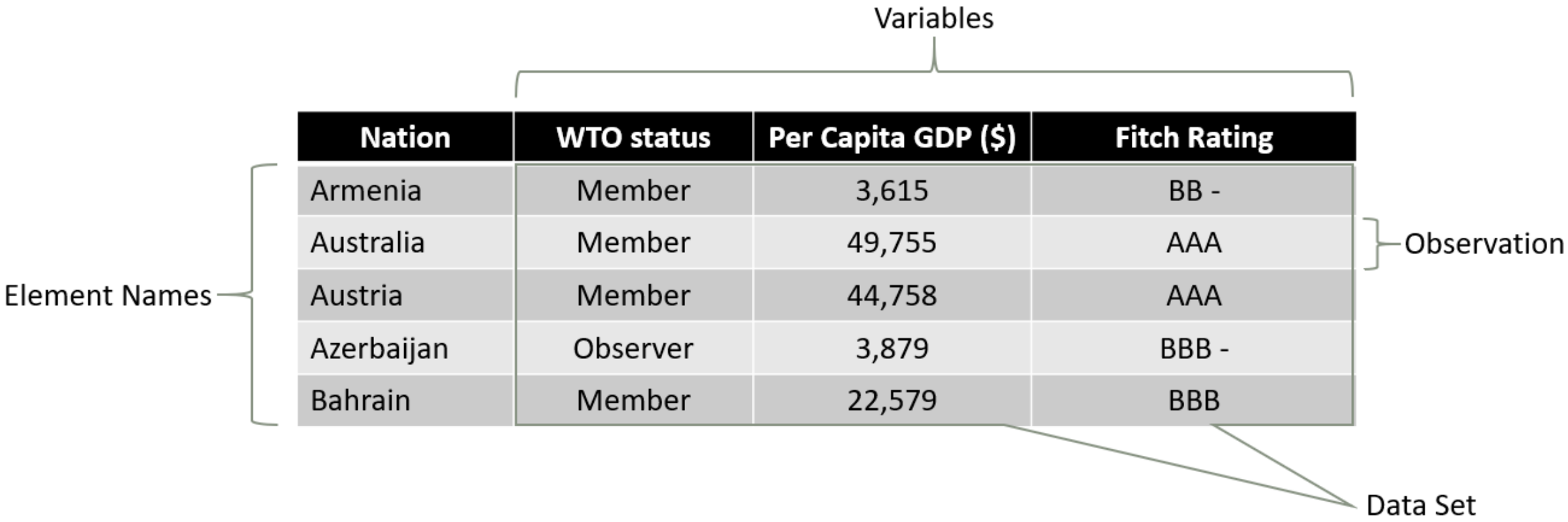
Data and Data Sets

- Data are the facts and figures collected, analyzed, and summarized for presentation and interpretation.
- All the data collected in a particular study are referred to as the data set for the study.

Elements, Variables, and Observations

- Elements are the entities on which data are collected.
- A variable is a characteristic of interest for the elements.
- The set of measurements obtained for a particular element is called an observation.
- A data set with *n* elements contains *n* observations.
- The total number of data values in a complete data set is the number of elements multiplied by the number of variables.

Data, Data Sets, Elements, Variables, and Observations



Scales of Measurement

- Scales of measurement include
  - Nominal
  - Ordinal
  - Interval
  - Ratio
- The scale determines the amount of information contained in the data.
- The scale indicates the data summarization and statistical analyses that are most appropriate.

Scales of Measurement

- **Nominal**
  - Data are labels or names used to identify an attribute of the element.
  - A nonnumeric label or numeric code may be used.
  - Example: The WTO status category for the nations in the previous example is classified using nonnumerical labels—"member" and "observer."
  - Alternatively, a numeric code could be used for the WTO status variable by letting 1 denote a member nation and 2 denote an observer nation.

Scales of Measurement

- **Ordinal**
  - The data have the properties of nominal data and the order or rank of the data is meaningful.
  - A nonnumeric label or numeric code may be used.
  - Example: The nonnumeric rating labels from AAA to F used for Fitch rating. These can be rank ordered from best credit rating AAA to poorest credit rating F.
  - Numerical code can also be used—Class rank of a student in school.

Scales of Measurement

- **Interval**
  - The data have the properties of ordinal data, and the interval between observations is expressed in terms of a fixed unit of measure.
  - Interval data are always numeric.
  - Example: Melissa has a SAT score of 1985, while Kevin has a SAT score of 1880. Melissa scored 105 points more than Kevin.

Scales of Measurement

- **Ratio**
  - The data have all the properties of interval data, and the ratio of two values is meaningful.
  - Variables such as distance, height, weight, and time use the ratio scale.
  - This scale must contain a zero value that indicates that nothing exists for the variable at the zero point.
  - Example: Melissa's college record shows 36 credit hours earned, while Kevin's record shows 72 credit hours earned. Kevin has twice as many credit hours earned as Melissa.

Scales of Measurement

- Data can be further classified as being categorical or quantitative.
- The statistical analysis that is appropriate depends on whether the data for the variable is categorical or quantitative.
- In general, there are more alternatives for statistical analysis when the data are quantitative.

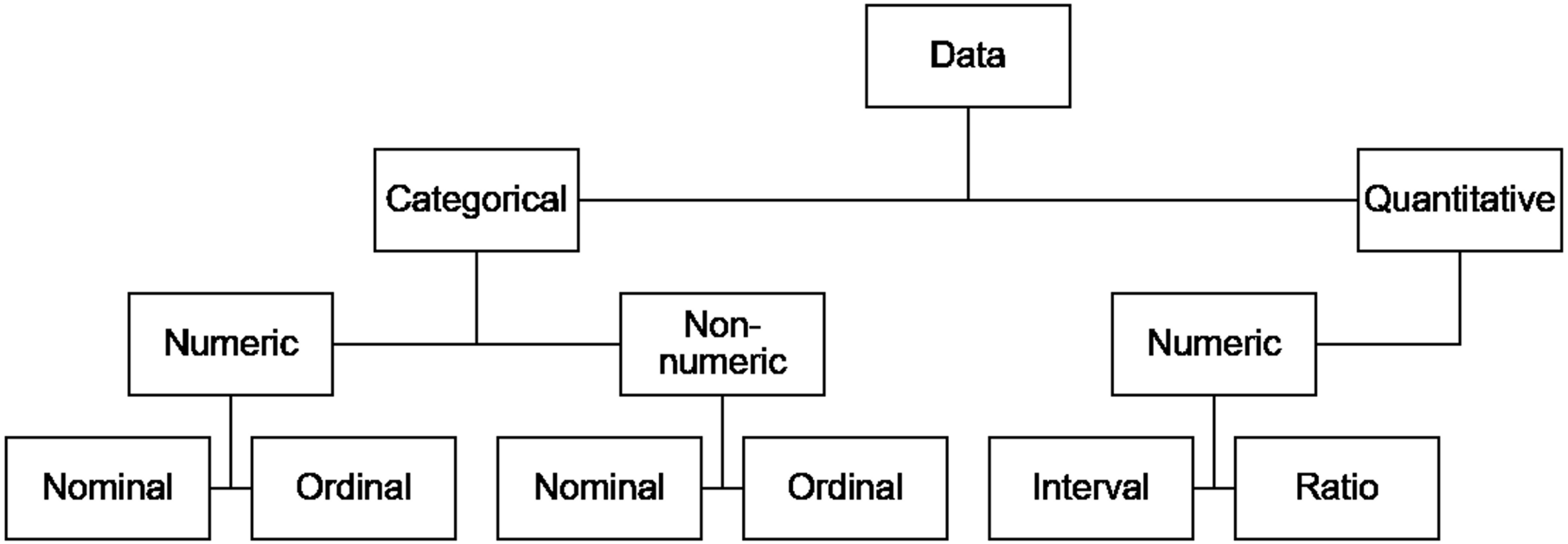
Categorical Data

- Labels or names used to identify an attribute of each element
- Often referred to as qualitative data
- Use either the nominal or ordinal scale of measurement
- Can be either numeric or nonnumeric
- Appropriate statistical analysis is rather limited

Quantitative Data

- Quantitative data indicate how many or how much:
  - discrete, if measuring how many
  - continuous, if measuring how much
- Quantitative data are always numeric.
- Ordinary arithmetic operations are meaningful for quantitative data.

Scales of Measurement



Cross-sectional Data

- Cross-sectional data are collected at the same or approximately the same point in time.

  Example:

  Data detailing different variables like status, Per capita GDP, Fitch rating for 60 different WTO nations at the same point in time.

Time Series Data

- Time series data are collected over several time periods.

  Example:

  U.S average price per gallon of conventional regular gasoline between 2012 and 2018

  Graphs of time series help analysts understand:

  - what happened in the past,
  - identify any trends over time, and
  - project future values for the time series.

Data Sources

- Existing Sources
  - Internal company records—almost any department
  - Business database services—Dow Jones & Co.
  - Government agencies—U.S. Department of Labor
  - Industry associations—U.S. Travel Association
  - Special-interest organizations—Graduate Management Admission Council (GMAT)
  - Internet—more and more firms

Data Sources

- Data Available from Internal Company Records

Record	Some of the Data Available
Employee records	Name, address, social security number
Production records	Part number, quantity produced, direct labor cost, material cost
Inventory records	Part number, quantity in stock, reorder level, economic order quantity
Sales records	Product number, sales volume, sales volume by region
Credit records	Customer name, credit limit, accounts receivable balance
Customer profile	Age, gender, income, household size

Data Sources

- Data Available from Selected Government Agencies

Government **Agency**	Some of the Data Available
Census Bureau	Population data, number of households, household income
Federal Reserve Board	Data on money supply, exchange rates, discount rates
Office of Mgmt. & Budget	Data on revenue, expenditures, debt of federal government
Department of Commerce	Data on business activity, value of shipments, profit by industry
Bureau of Labor Statistics	Customer spending, unemployment rate, hourly earnings, safety record
DATA.GOV	More than 150,000 data sets including agriculture, consumer education, health, and manufacturing data

Data Sources

- Statistical Studies—Observational
  - In observational studies, no attempt is made to control or influence the variables of interest.
    - A survey is a good example.
  - An example of an observational study is researchers observing a randomly selected group of customers that enter a Walmart Supercenter to collect data on variables such as time spent in the store, gender of the customer, and the amount spent.

Data Sources

- Statistical Studies—Experimental
  - In experimental studies, the variable of interest is first identified. Then one or more variables are identified and controlled so that data can be obtained about how they influence the variable of interest.
  - The largest experimental study ever conducted is believed to be the 1954 Public Health Service experiment for the Salk polio vaccine. Nearly two million U.S. children (grades 1 through 3) were selected.

Data Acquisition Considerations

- Time Requirement
  - Searching for information can be time consuming.
  - Information may no longer be useful by the time it is available.
- Cost of Acquisition
  - Organizations often charge for information even when it is not their primary business activity.
- Data Errors
  - Using any data that happen to be available or were acquired with little care can lead to misleading information.

Descriptive Statistics

- Most of the statistical information in newspapers, magazines, company reports, and other publications consist of data that are summarized and presented in a form that is easy to understand.

- Such summaries of data, which may be tabular, graphical, or numerical, are referred to as descriptive statistics.

Descriptive Statistics – Example: Auto Repair

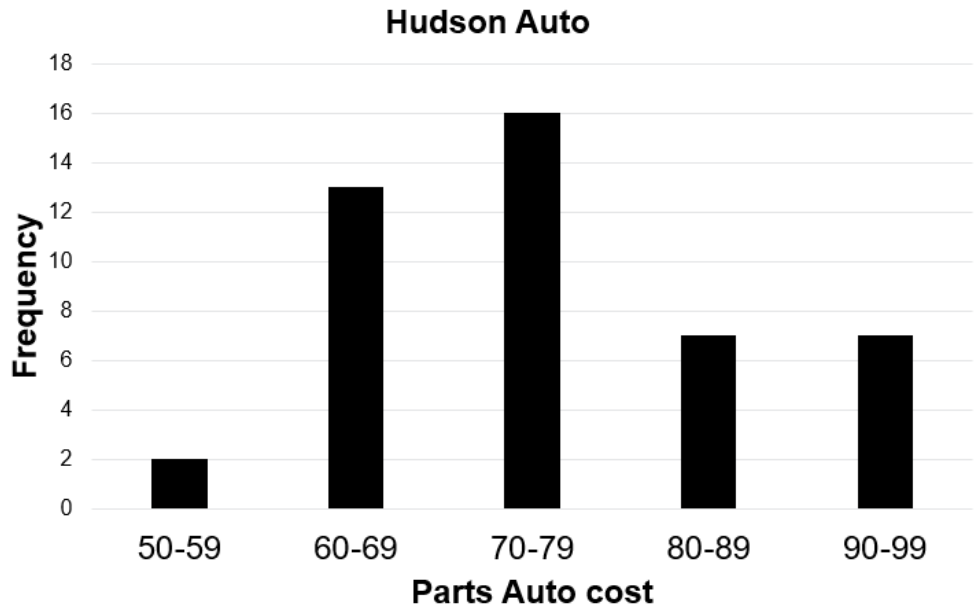
The manager of Hudson Auto would like to have a better understanding of the cost of parts used in the engine tune-ups performed in her shop. She examines 50 customer invoices for tune-ups. The costs of parts, rounded to the nearest dollar, are listed on the next slide.

- Sample of Parts Cost ($) for 50 Tune-ups

91	78	93	57	75	52	99	80	97	62
71	69	72	89	66	75	79	75	72	76
104	74	62	68	97	105	77	65	80	109
85	97	88	68	83	68	71	69	67	74
62	82	98	101	79	105	79	69	62	73

Descriptive Statistics – Example: Auto Repair

Parts Cost ($)	Frequency	Percent Frequency
50 to 59	2	4%
60 to 69	13	26%
70 to 79	16	32%
80 to 89	7	14%
90 to 99	7	14%
100 to 109	5	10%
TOTAL	50	100%



Numerical Descriptive Statistics

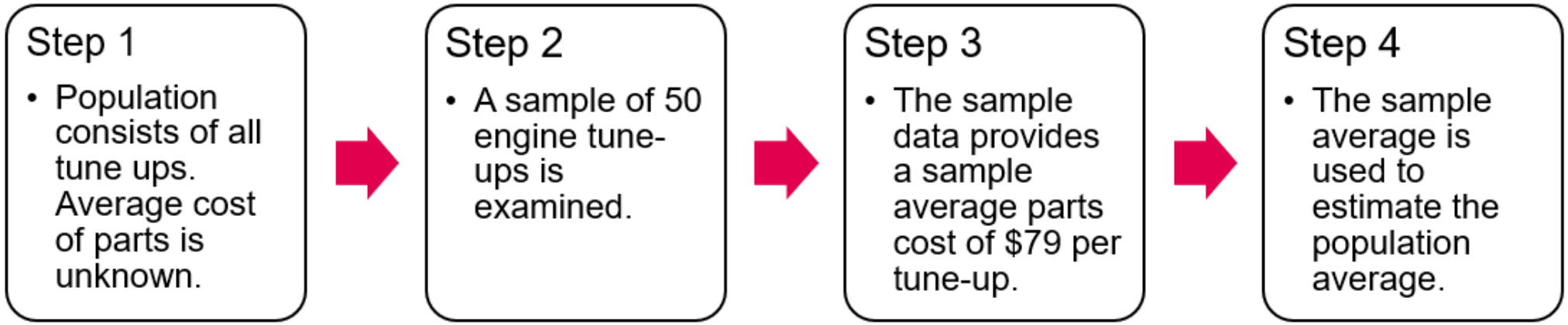
- The most common numerical descriptive statistic is the mean (or average).
- The mean demonstrates a measure of the central tendency, or central location, of the data for a variable.
- Hudson’s mean cost of parts, based on the 50 tune-ups studied, is $79 (found by summing up the 50 cost values and then dividing by 50).

Statistical Inference

- Population: The set of all elements of interest in a particular study.
- Sample: A subset of the population.
- Statistical inference: The process of using data obtained from a sample to make estimates and test hypotheses about the characteristics of a population.
- Census: Collecting data for the entire population.
- Sample survey: Collecting data for a sample.

Process of Statistical Inference

Example: Auto Repair



Analytics

Scientific process of transforming data into insight for making better decisions.

- Descriptive analysis—Analytical techniques that describe what happened in the past.
- Predictive analysis
  - Analytical techniques that use models constructed from past data to predict future.
  - Helps assess the impact the impact of one variable on another
- Prescriptive analysis—Analytical techniques that yield a best course of action to take.

Big Data and Data Warehousing

- Organizations obtain large amounts of data on a daily basis by means of magnetic card readers, bar code scanners, point-of-sale terminals, and touchscreen monitors. Large and complex data sets are known as big data.
- Walmart captures data on 20 to 30 million transactions per day.
- Visa processes 6,800 payment transactions per second.
- Capturing, storing, and maintaining the data, referred to as data warehousing, is a significant undertaking.

Data Mining

- Analysis of the data in the warehouse might aid in decisions that will lead to new strategies and higher profits for the organization.
- Using a combination of procedures from statistics, mathematics, and computer science, analysts “mine the data” to convert it into useful information.
- The most effective data mining systems use automated procedures to extract information from the data prompted by only general or even vague queries by the user.

Data Mining Applications

- The major applications of data mining have been made by companies with a strong consumer focus such as retail, financial, and communication firms.
- Data mining is used to identify related products that customers who have already purchased a specific product are also likely to purchase (and then pop-ups are used to draw attention to those related products).
- As another example, data mining is used to identify customers who should receive special discount offers based on their past purchasing volumes.

Data Mining Requirements

- Statistical methodology such as multiple regression, logistic regression, and correlation are heavily used.
- Also needed are computer science technologies involving artificial intelligence and machine learning.
- A significant investment in time and money is required as well.

Data Mining Model Reliability

- Finding a statistical model that works well for a particular sample of data does not necessarily mean that it can be reliably applied to other data.
- With the enormous amount of data available, the data set can be partitioned into a training set (for model development) and a test set (for validating the model).
- There is, however, a danger of over fitting the model to the point that misleading associations and conclusions appear to exist.
- Careful interpretation of results and extensive testing is important.

Descriptive Statistics: Tabular and Graphical Displays

- Summarizing Data for a Categorical Variable
  - Categorical data use labels or names to identify categories of like items.
- Summarizing Data for a Quantitative Variable
  - Quantitative data are numerical values that indicate how much or how many.

Summarizing Data for a Categorical Variable

- Frequency Distribution
- Relative Frequency Distribution
- Percent Frequency Distribution
- Bar Chart
- Pie Chart

Frequency Distribution

- A frequency distribution is a tabular summary of data showing the number (frequency) of observations in each of several non-overlapping categories or classes.
- The objective is to provide insights about the data that cannot be quickly obtained by looking only at the original data.

Frequency Distribution: Example

- Soft drink purchasers were asked to select one among the five popular soft drinks: Coca-Cola, Diet Coke, Dr. Pepper, Pepsi, and Sprite.
- Soft drinks selected by a sample of 20 purchasers are:

Coca-Cola	Pepsi	Dr. Pepper
Diet Coke	Dr. Pepper	Dr. Pepper
Dr. Pepper	Pepsi	Pepsi
Pepsi	Coca-Cola	Diet Coke
Pepsi	Diet Coke	Dr. Pepper
Pepsi	Pepsi	Sprite
Pepsi	Pepsi	

Rating	Frequency
Coca-Cola	2
Diet Coke	3
Dr. Pepper	5
Pepsi	9
Sprite	1
Total	20

Relative and Percent Frequency Distribution

- The relative frequency of a class is the fraction or proportion of the total number of data items belonging to a class.

$$\text{Relative frequency of a class} = \frac{\text{Frequency of the class}}{n}$$

  - A relative frequency distribution is a tabular summary of a set of data showing the relative frequency for each class.

  - The percent frequency of a class is the relative frequency multiplied by 100.

  - A percent frequency distribution is a tabular summary of a set of data showing the percent frequency for each class.

Relative and Percent Frequency Distribution

Rating	Relative Frequency	Percent Frequency
Coca-Cola	.38	38
Diet Coke	.16	16
Dr. Pepper	.10	10
Pepsi	.26	26
Sprite	.10	10
Total	1.00	100

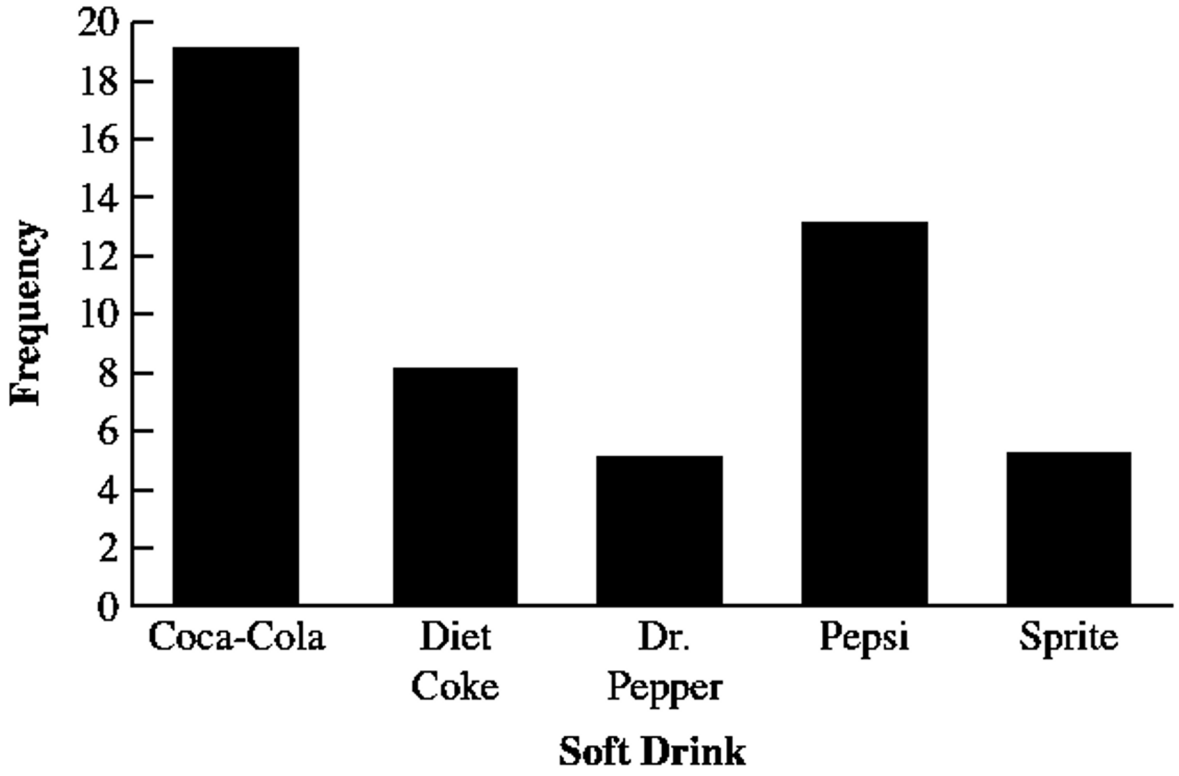
.38(100) = 38

5/50 = 0.10

Bar Chart

- A bar chart is a graphical display for depicting categorical data.
- On one axis (usually the horizontal axis), we specify the labels that are used for each of the classes.
- A frequency, relative frequency, or percent frequency scale can be used for the other axis (usually the vertical axis).
- Using a bar of fixed width drawn above each class label, we extend the height appropriately.
- The bars are separated to emphasize the fact that each class is separate.

Bar Chart of Soft Drink Purchases



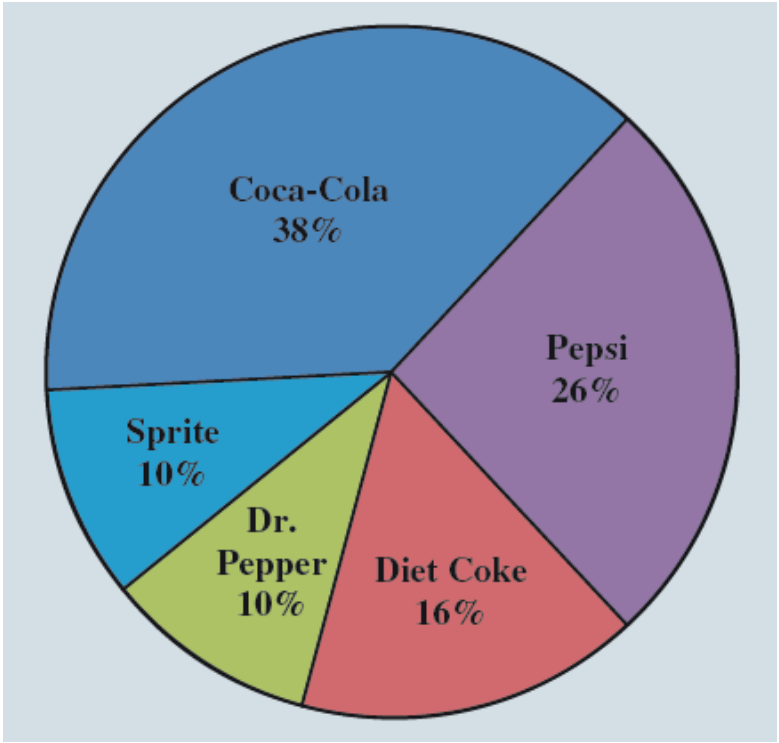
Pie Chart

- The pie chart is a commonly used graphical display for presenting relative frequency and percent frequency distributions for categorical data.
- First draw a circle; then use the relative frequencies to subdivide the circle into sectors that correspond to the relative frequency for each class.
- Since there are 360 degrees in a circle, a class with a relative frequency of .25 would consume $.25(360) = 90$ degrees of the circle.

Pie Chart of Soft Drink Purchases

Inferences from the pie chart:

- Over one-third of the customers surveyed prefer Coca-Cola.
- The second preference is for Pepsi with 26% of the customers opting for it.
- Only 10% of the customers opt for Sprite or Dr. Pepper.



Summarizing Data for a Quantitative Variable

- Frequency Distribution
- Relative Frequency and Percent Frequency Distributions
- Dot Plot
- Histogram
- Cumulative Distributions
- Stem-and-Leaf Display

Frequency Distribution

Example

Sanderson and Clifford, a small public accounting firm, wants to determine time in days required to complete year-end audits. It takes a sample of 20 clients.

Year-End Audit Time (in Days)

12	14	19	18
15	15	18	17
20	27	22	23
22	21	33	28
14	18	16	13

Frequency Distribution

The three steps necessary to define the classes for a frequency distribution with quantitative data are:

- **Step 1:** Determine the number of non-overlapping classes.
- **Step 2:** Determine the width of each class.
- **Step 3:** Determine the class limits.

Frequency Distribution

Guidelines for Determining the Number of Classes

- Use between 5 and 20 classes.
- Data sets with a larger number of elements usually require a larger number of classes.
- Smaller data sets usually require fewer classes.
- The goal is to use enough classes to show the variation in the data, but not so many classes that some contain only a few data items.
- Use classes of equal width.

$$\text{Approximate Class Width} = \frac{\text{Largest data value} - \text{Smallest data value}}{\text{Number of classes}}$$

- Making the classes the same width reduces the chance of inappropriate interpretations.

Frequency Distribution

Note on Number of Classes and Class Width

- In practice, the number of classes and the appropriate class width are determined by trial and error.
- Once a possible number of classes is chosen, the appropriate class width is found.
- The process can be repeated for a different number of classes.
- Ultimately, the analyst uses judgment to determine the combination of the number of classes and class width that provides the best frequency distribution for summarizing the data.

Frequency Distribution

Guidelines for Determining the Class Limits

- Class limits must be chosen so that each data item belongs to one and only one class.
- The lower class limit identifies the smallest possible data value assigned to the class.
- The upper class limit identifies the largest possible data value assigned to the class.
- The appropriate values for the class limits depend on the level of accuracy of the data.
- An open-end class requires only a lower class limit or an upper class limit.

Frequency Distribution

Class Midpoint

- In some cases, we want to know the midpoints of the classes in a frequency distribution for quantitative data.

- The class midpoint is the value halfway between the lower and upper class limits.

Frequency Distribution: Example – Sanderson and Clifford

- If we choose five classes:

$$\text{Approximate Class Width} = \frac{(33-12)}{5} = 4.2 \cong 4$$

Time in days	Frequency
10 to 14	4
15 to 19	8
20 to 24	5
25 to 29	2
30 to 34	1
Total	20

Relative and Percent Frequency Distribution: Example – Sanderson and Clifford

Audit time **(in days)**	Relative Frequency	Percent Frequency
10 to 14	.20 (4/20)	20 (0.2 × 100)
15 to 19	.40	40
20 to 25	.25	25
25 to 29	.10	10
30 to 34	.05	5
	Total 1.00	100

Insights obtained from the Percent Frequency Distribution:

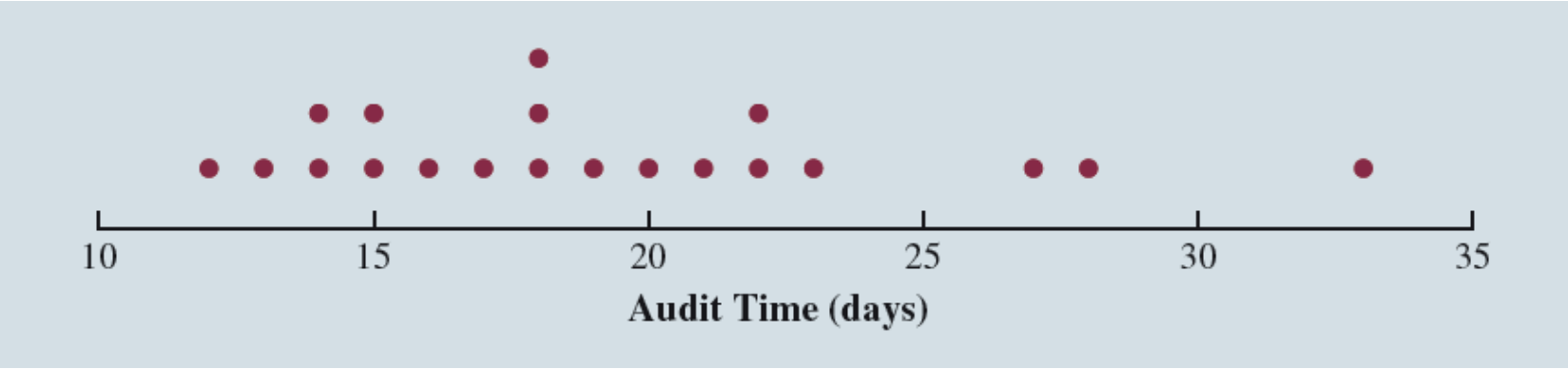
40% of the audits required from 15 to 19 days.

Another 25% of the audits required 20 to 25 days.

Only 5% of the audits required more than 30 days.

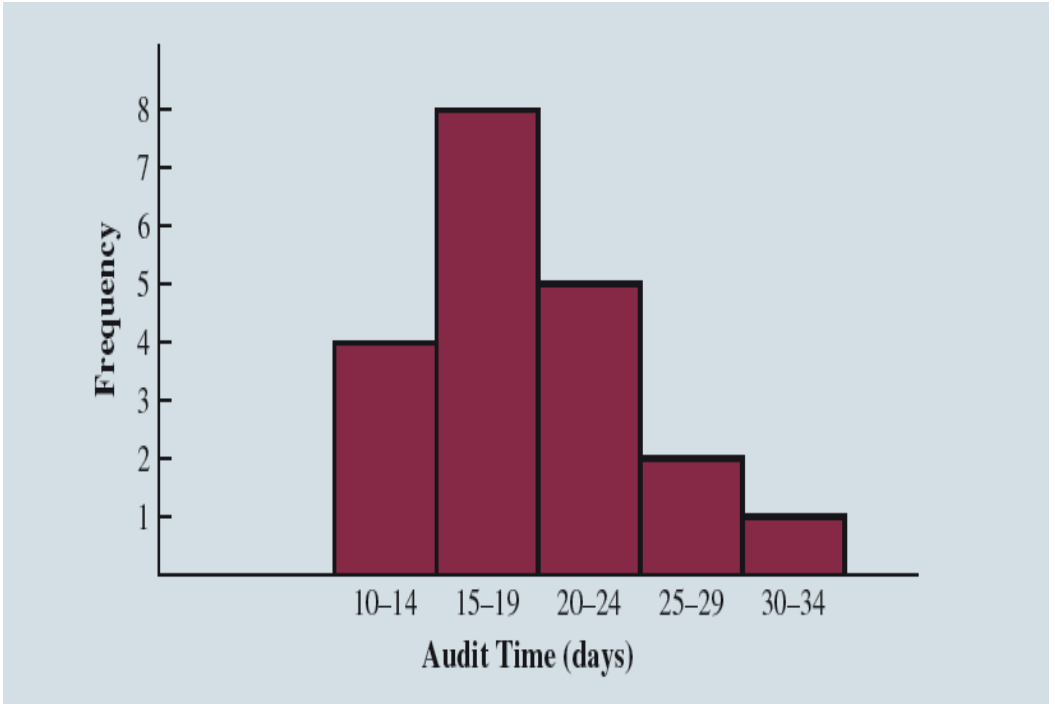
Dot Plot

- One of the simplest graphical summaries of data is a dot plot.
- A horizontal axis shows the range of data values.
- Then each data value is represented by a dot placed above the axis.



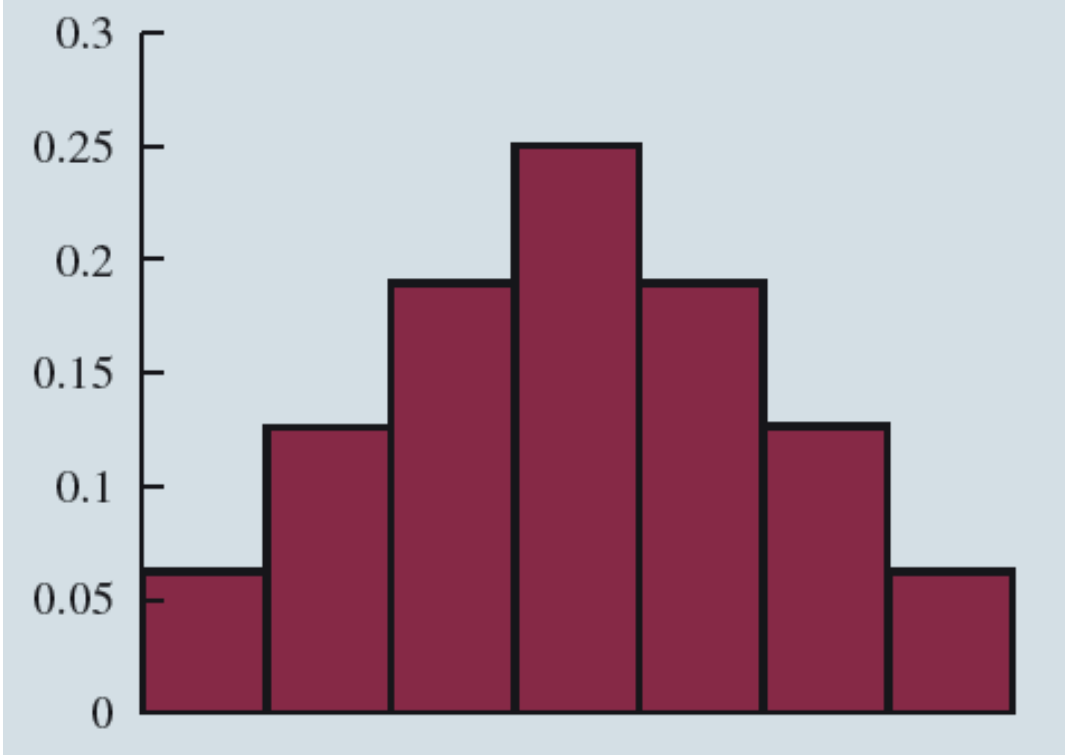
Histogram

- Another common graphical display of quantitative data is a histogram.
- The variable of interest is placed on the horizontal axis.
- A rectangle is drawn above each class interval with its height corresponding to the interval's frequency, relative frequency, or percent frequency.
- Unlike a bar graph, a histogram has no natural separation between rectangles of adjacent classes.



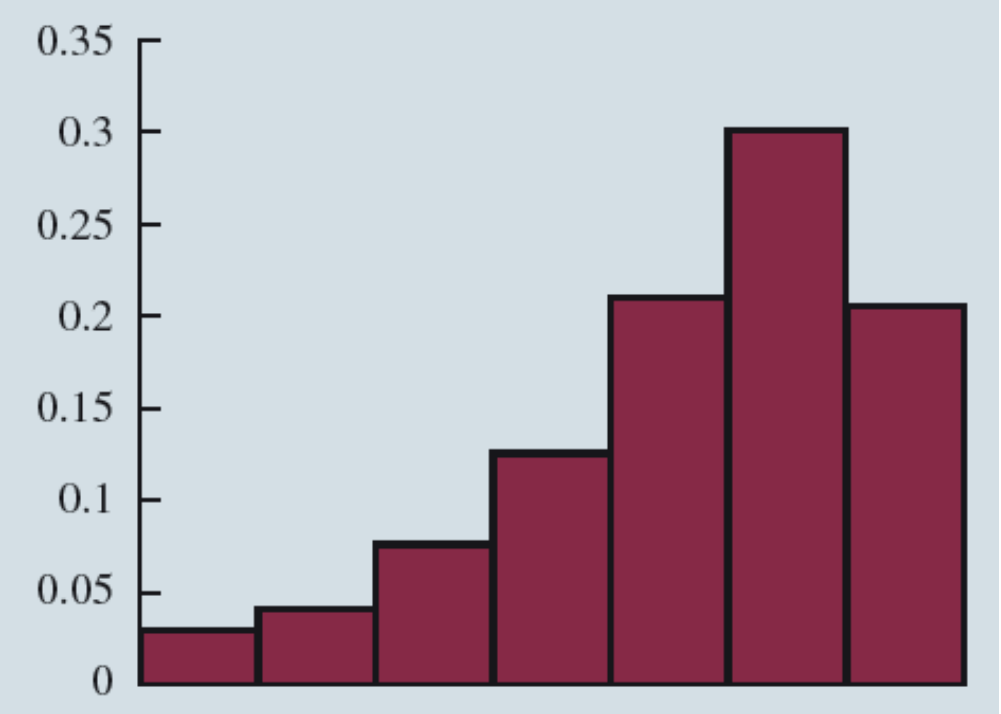
Histograms Expose Skewness

- Symmetric
  - Left tail is the mirror image of the right tail.
  - **Example:** Heights of People



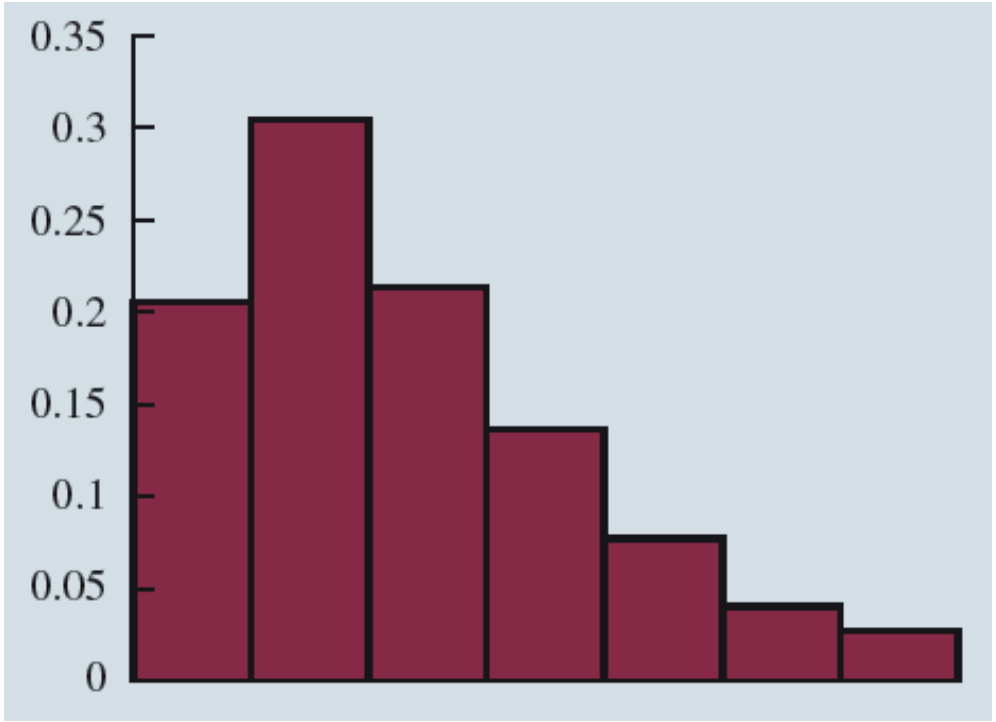
Histograms Expose Skewness

- Moderately Skewed Left
  - A longer tail to the left
  - **Example:** Exam Scores



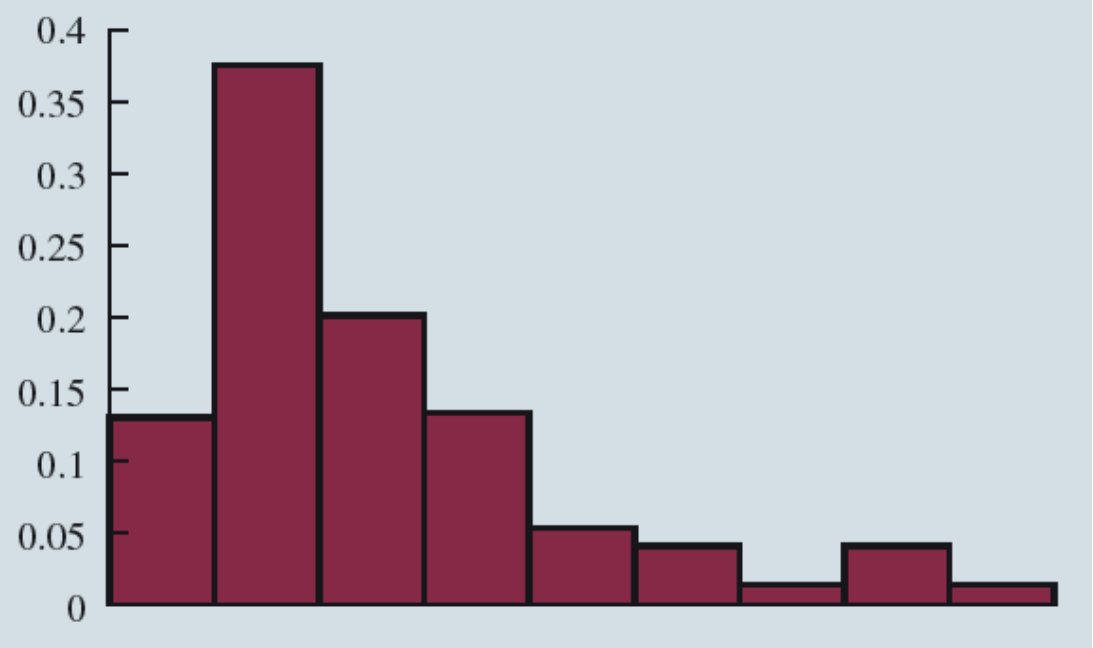
Histograms Expose Skewness

- Moderately Right Skewed
  - A longer tail to the right
  - **Example:** Housing Values



Histograms Expose Skewness

- Highly Skewed Right
  - A very long tail to the right
  - **Example:** Executive Salaries



Cumulative Distribution

- Cumulative frequency distribution: Shows the *number* of items with values less than or equal to the upper limit of each class.
- Cumulative relative frequency distribution: Shows the *proportion* of items with values less than or equal to the upper limit of each class.
- Cumulative percent frequency distribution: Shows the *percentage* of items with values less than or equal to the upper limit of each class.

Cumulative Distribution

- The last entry in a cumulative frequency distribution always equals the total number of observations.
- The last entry in a cumulative relative frequency distribution always equals 1.00.
- The last entry in a cumulative percent frequency distribution always equals 100.

Cumulative Distribution: Example – Sanderson and Clifford

Audit time (Days)	Cumulative Frequency	Cumulative Relative Frequency	Cumulative **Percent Frequency**
Less than or equal to 14	4	.20	20
Less than or equal to 19	12	.60	60
Less than or equal to 24	17	.85	85
Less than or equal to 29	19	.95	95
Less than or equal to 34	20	1.00	100

Summarizing Data of Two Variables Using Tables

- Thus far, we have focused on methods that are used to summarize the data for one variable at a time.
- Often a decision maker is interested in tabular and graphical methods that will help illustrate the relationship between two variables.

Crosstabulation

- A crosstabulation is a tabular summary of data for two variables.
- Crosstabulation can be used when:
  - one variable is categorical and the other is quantitative,
  - both variables are categorical, or
  - both variables are quantitative.
- The left and top margin labels define the classes for the two variables.

Crosstabulation: Example – Restaurant Review

Crosstabulation of quality rating and meal price data for 300 Los Angeles restaurants is given here.

Insights gained from preceding crosstabulation:

- Greatest number of restaurants in the sample (64) have a very good rating and meal prices in the $20 to 29 range.
- Only 2 restaurants have an excellent rating and meal prices in the range of $10 to 19.

Converting the entries in the table into row percentages or column percentages can provide additional insight about the relationship between the two variables.

Crosstabulation: Row Percentages: Example – Restaurant Review

	Meal Price				
Quality Rating	$10 to 19	$20 to 29	$30 to 39	$40 to 49	Total
Good	50	47.6	2.4	0	100
Very Good	22.7	42.7	30.6	4	100
Excellent	3	21.2	42.4	33.4	100

- Good restaurants charging a meal price of $10 to 19 divided by total number of good restaurants:

$$\frac{42}{84} \times 100 = 50\%$$

Crosstabulation: Simpson's Paradox

- Data in two or more crosstabulations are often aggregated to produce a summary crosstabulation.
- We must be careful in drawing conclusions about the relationship between the two variables in the aggregated crosstabulation.
- In some cases the conclusions based upon an aggregated crosstabulation can be completely reversed if we look at the unaggregated data. The reversal of conclusions based on aggregate and unaggregated data is called Simpson's paradox.

Crosstabulation: Simpson's Paradox – Example

- Data in two or more crosstabulations are often aggregated to produce a summary crosstabulation.
- We must be careful in drawing conclusions about the relationship between the two variables in the aggregated crosstabulation.
- In some cases the conclusions based upon an aggregated crosstabulation can be completely reversed if we look at the unaggregated data. The reversal of conclusions based on aggregate and unaggregated data is called Simpson's paradox.

Crosstabulation: Simpson's Paradox – Example

"Kendall is the better judge as he has a higher percentage of verdicts upheld!" But realize that the conclusion or interpretation may be reversed depending upon whether you are viewing unaggregated or aggregated crosstabulation data.

	Judge		
Verdict	Luckett	Kendall	Total
Upheld	129 (86%)	110 (88%)	239
Reversed	21 (14%)	15 (12%)	36
Total (%)	150 (100%)	125 (100%)	275

	Judge Luckett		
Verdict	Common Pleas	Municipal Court	Total
Upheld	29 (91%)	100 (85%)	129
Reversed	3 (9%)	18 (15%)	21
Total (%)	32 (100%)	118 (100%)	150

	Judge Kendall		
Verdict	Common Pleas	Municipal Court	Total
Upheld	90 (90%)	20 (80%)	110
Reversed	10 (10%)	5 (20%)	15
Total (%)	100 (100%)	25 (100%)	125

Summarizing Data for Two Variables Using Graphical Displays

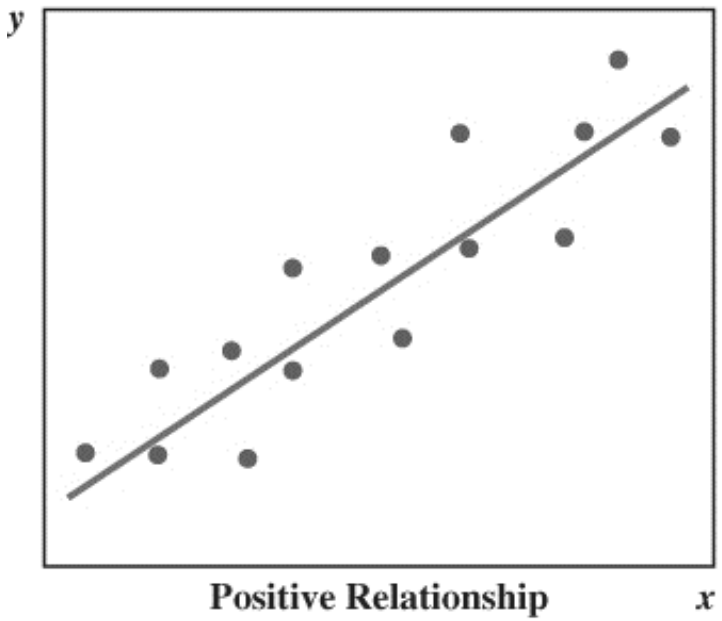
- In most cases, a graphical display is more useful than a table for recognizing patterns and trends.
- Displaying data in creative ways can lead to powerful insights.
- Scatter diagrams and trendlines are useful in exploring the relationship between two variables.

Scatter Diagram and Trendline

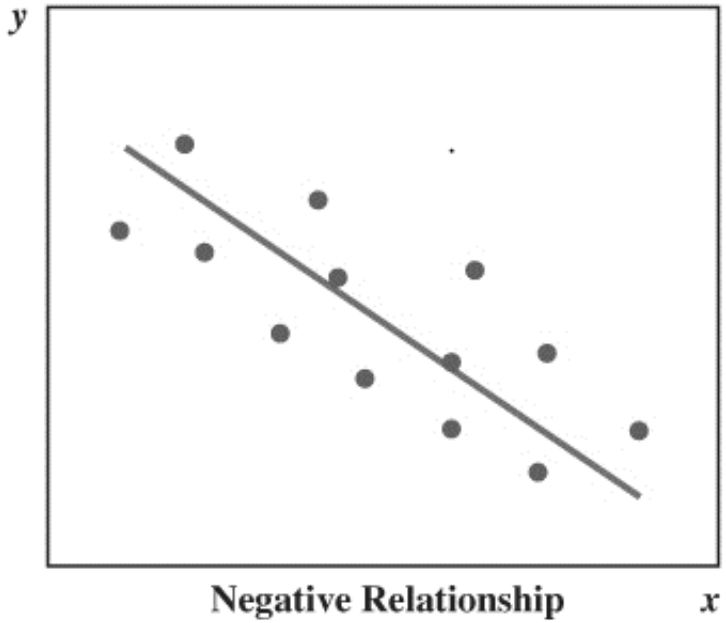
A scatter diagram is a graphical presentation of the relationship between two quantitative variables.

- One variable is shown on the horizontal axis and the other variable is shown on the vertical axis.
- The general pattern of the plotted points suggests the overall relationship between the variables.
- A trendline provides an approximation of the relationship.

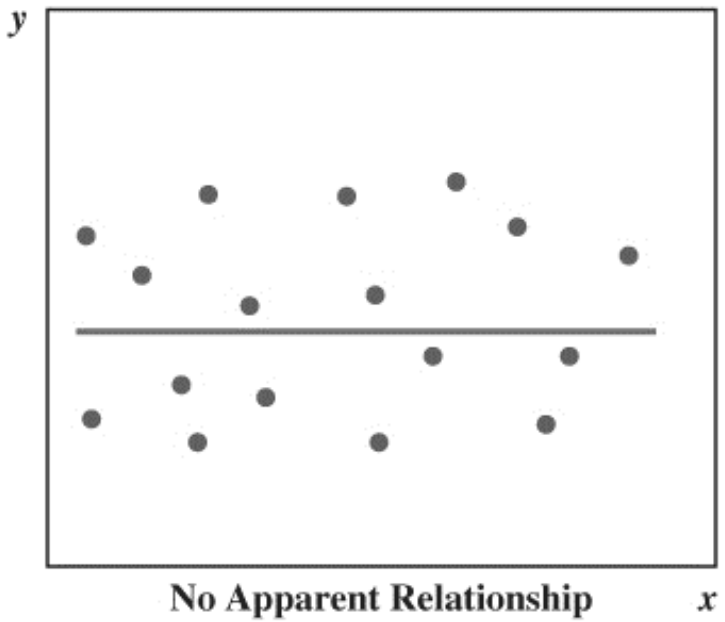
Scatter Diagram and Trendline



Positive Relationship



Negative Relationship

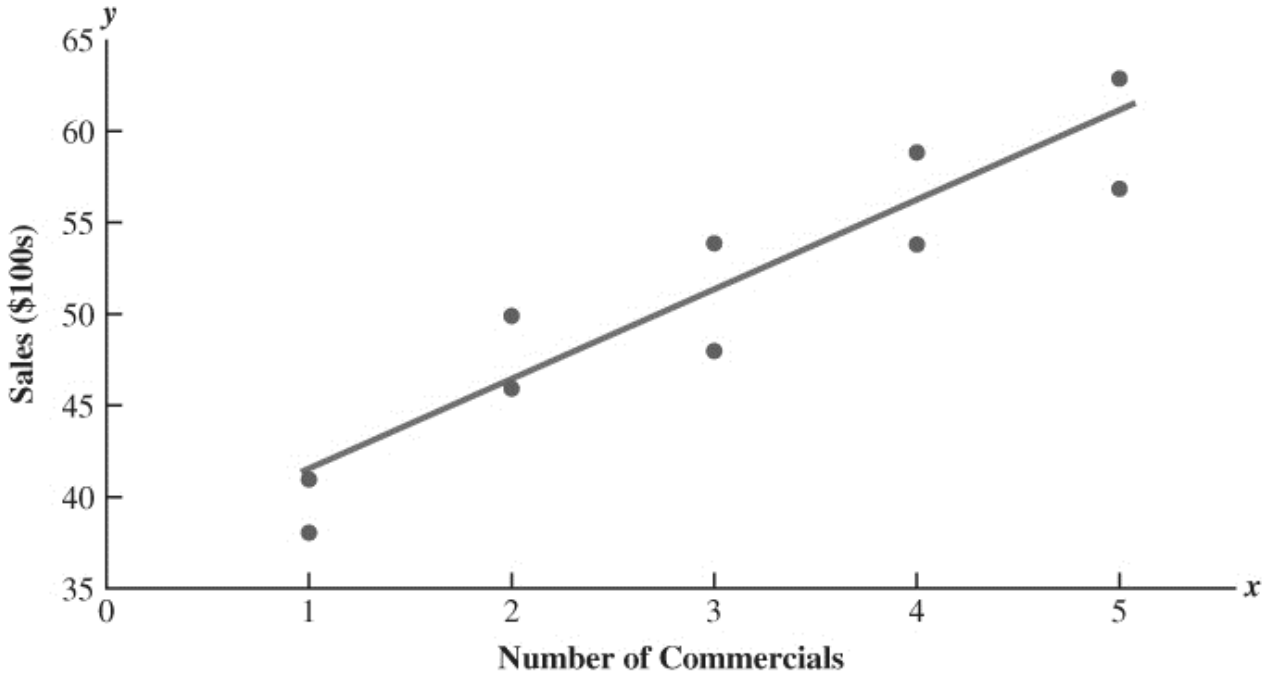


No Apparent Relationship

Scatter Diagram and Trendline: Example

An electronics store in San Francisco wants to analyze the relationship between sales and advertising. Sample data for ten weeks with sales in hundreds of dollars is shown below:

Week	Number of Commercials (x)	Sales ($100s)(y)
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46



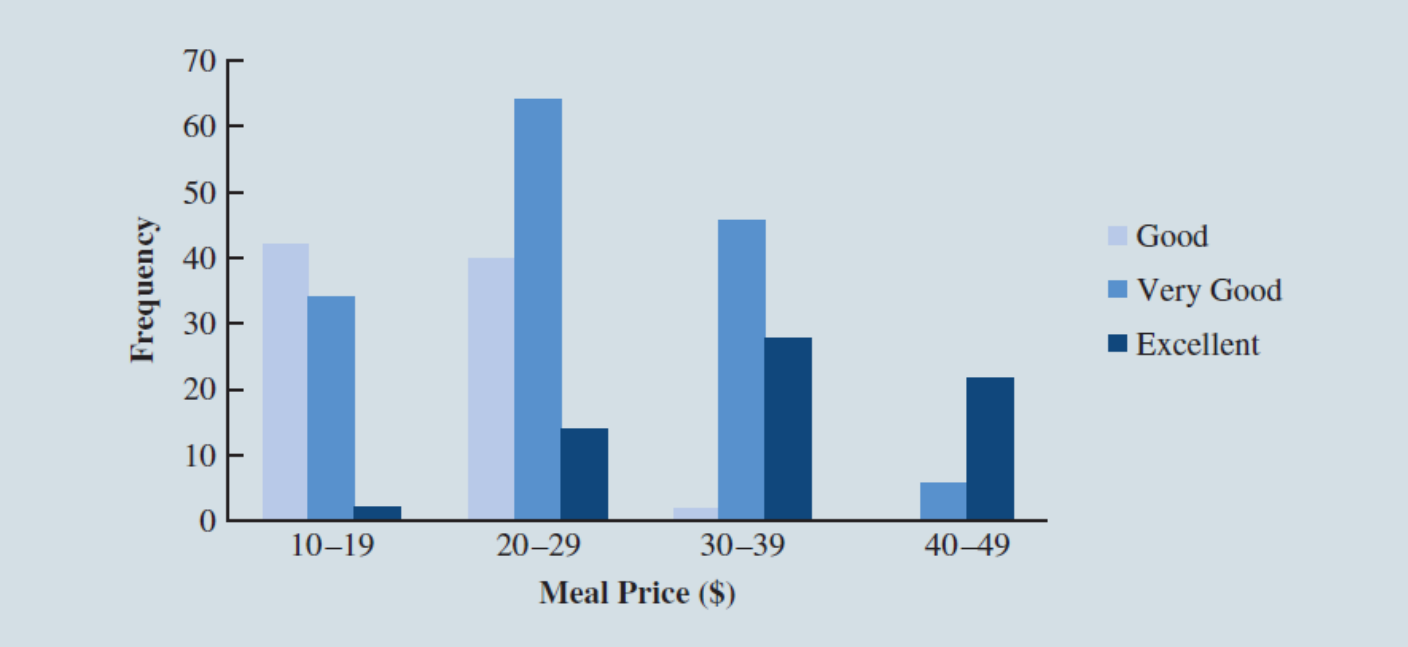
Scatter Diagram and Trendline: Example

Insights gained from the San Francisco Electronics Store scatter diagram:

- The scatter diagram indicates a positive relationship between the number of commercials and sales.
- Higher sales are associated with greater number of commercials.
- The relationship is not perfect; all plotted points in the scatter diagram are not on a straight line.

Side-by-Side bar Chart

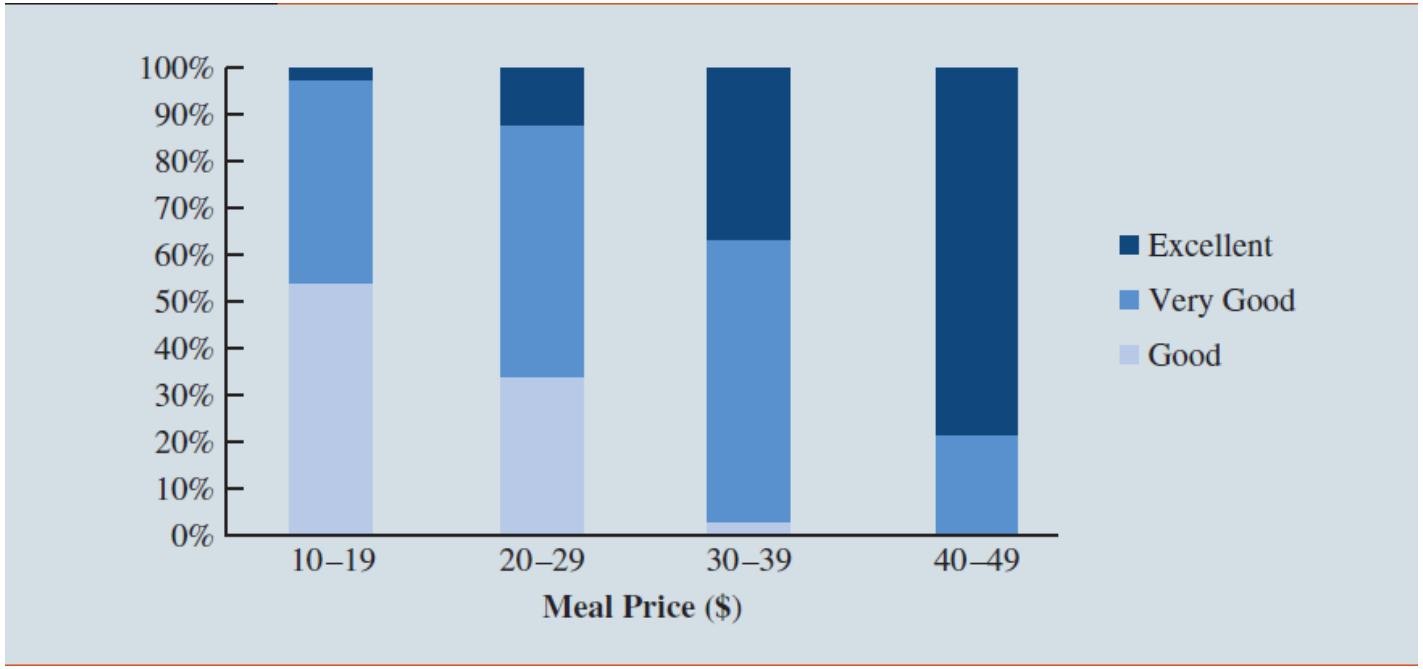
- A side-by-side bar chart is a graphical display for depicting multiple bar charts on the same display.
- Each cluster of bars represents one value of the first variable.
- Each bar within a cluster represents one value of the second variable.



BUSINESS SCHOOL | RWTH AACHEN UNIVERSITY

Stacked Bar Chart

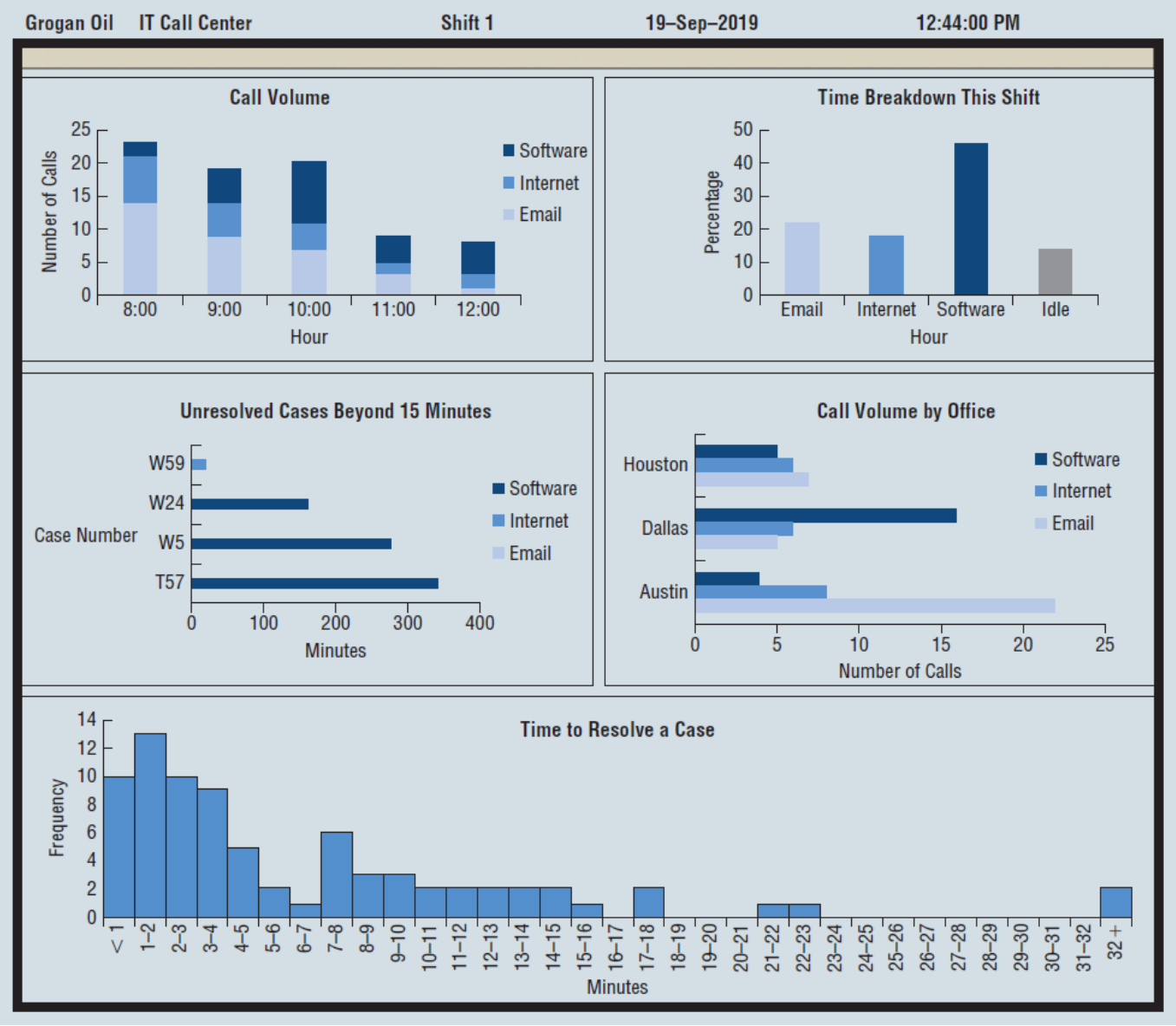
- A stacked bar chart is another way to display and compare two variables on the same display.
- It is a bar chart in which each bar is broken into rectangular segments of a different color.
- If percentage frequencies are displayed, all bars will be of the same height (or length), extending to the 100% mark.



Data Dashboard

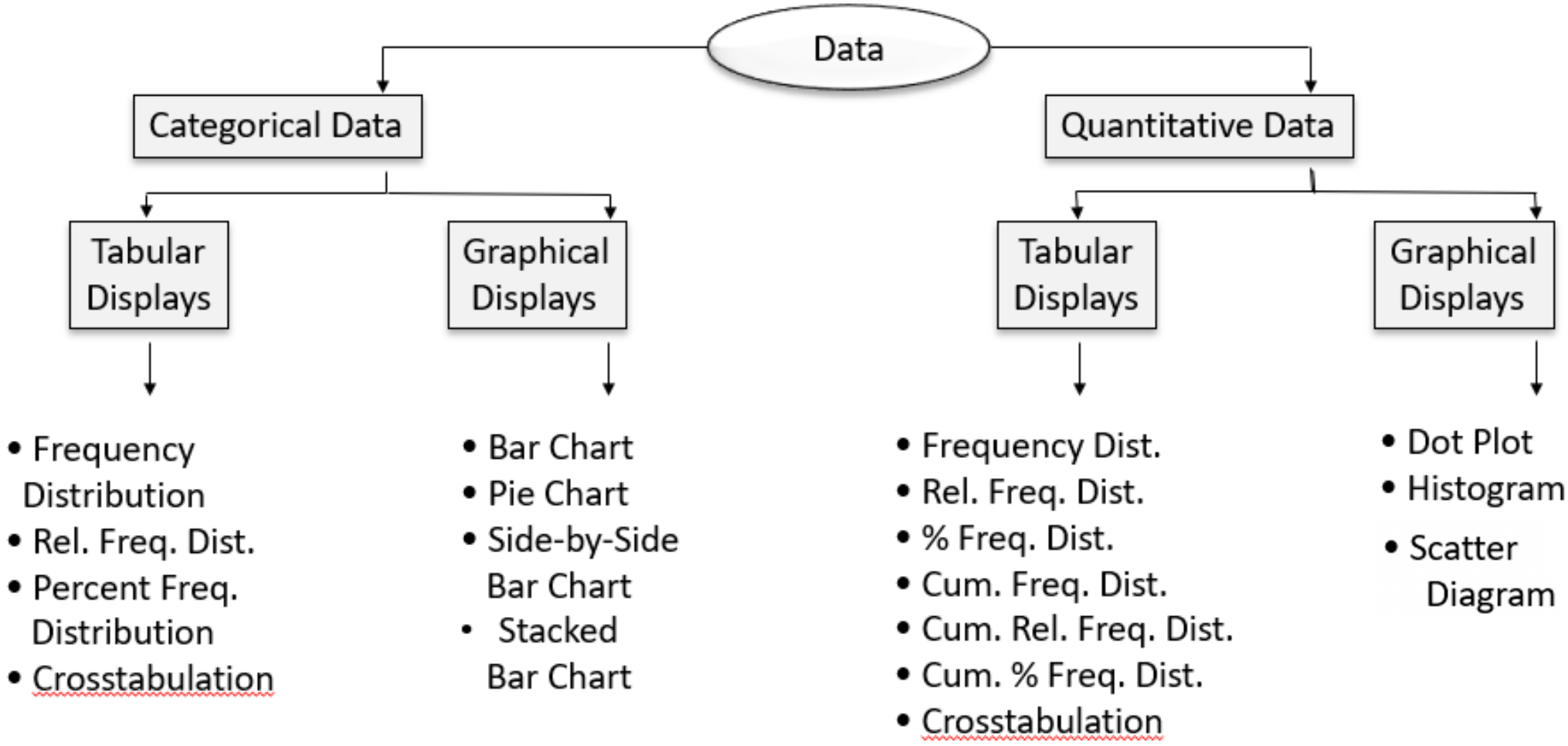
- A data dashboard is a widely used data visualization tool.
- It organizes and presents key performance indicators used to monitor an organization or process.
- It provides timely, summary information that is easy to read, understand, and interpret.
- Guidelines:
  - Minimize the need for screen scrolling.
  - Avoid unnecessary use of color or 3D.
  - Use borders between charts to improve readability.

Data Dashboard: Example



BUSINESS SCHOOL | RWTH AACHEN UNIVERSITY

Tabular and Graphic Display: Summary



Numerical Measures

- If the measures are computed for data from a sample, they are called sample statistics.
- If the measures are computed for data from a population, they are called population parameters.
- A sample statistic is referred to as the point estimator of the corresponding population parameter.

Measures of Location

- Mean
- Median
- Mode
- Weighted Mean
- Geometric Mean
- Percentiles
- Quartiles

Mean

Perhaps the most important measure of location is the mean.

- The mean provides a measure of central location.
- The mean of a data set is the average of all the data values.
- The sample mean $\bar{x}$ is the point estimator of the population mean μ.

Sample and Population Mean $\bar{x}$

$$\bar{x} = \frac{\sum x_i}{n}$$

Where:

$\sum x_i$ = sum of the values of the n observations

n = number of observations in the sample

$$\mu = \frac{\sum x_i}{n}$$

Where:

$\sum x_i$ = sum of the values of the n observations

n = number of observations in the population

Sample Mean: Example – Monthly Starting Salary

A placement office wants to know the average starting salary of business graduates. Monthly starting salaries for a sample of 12 business school graduates is provided here.

Graduate	Monthly Starting Salary ($)	Graduate	Monthly Starting Salary ($)
1	5850	7	5890
2	5950	8	6130
3	6050	9	5940
4	5880	10	6325
5	5755	11	5920
6	5710	12	5880

$$\bar{x} = \frac{\sum x_i}{n} = \frac{71{,}280}{12} = 5{,}940$$

Median

- The median of a data set is the value in the middle when the data items are arranged in ascending order.
- Whenever a data set has extreme values, median is the preferred measure of central location.
- The median is the measure of location most often reported for annual income and property value data.
- A few extremely large incomes or property values can inflate the mean.

Median

For an odd number of observations:

7 observations

26	18	27	12	14	27	19

12	14	18	19	26	27	27

In ascending order

Median is the middle value

Median = 19

Median

For an even number of observations: 8 observations

26	18	27	12	14	27	19	30

12	14	18	19	26	27	27	30

In ascending order

Median is the average of the middle two values.

$$\text{Median} = \frac{(19+26)}{2} = 22.5$$

Median

Example: Monthly Starting Salary Averaging the 6th and 7th data values:

Note: The data is in ascending order.

5,710	5,755
5,850	5,880
5,880	5,890
5,920	5,940
5,950	6,050
6,130	6,325

$$\text{Median} = \frac{(5{,}890 + 5{,}920)}{2} = 5{,}905$$

Trimmed Mean

- Another measure sometimes used when extreme values are present is the trimmed mean.
- It is obtained by deleting a percentage of the smallest and largest values from a data set and then computing the mean of the remaining values.
- For example, the 5% trimmed mean is obtained by removing the smallest 5% and the largest 5% of the data values and then computing the mean of the remaining values.

Mode

- The mode of a data set is the value that occurs with greatest frequency.
- The greatest frequency can occur at two or more different values.
- If the data have exactly two modes, the data are bimodal.
- If the data have more than two modes, the data are multimodal.

Mode

Example: Monthly Starting Salary

The only monthly starting salary that occurs more than once is $5,880.

Mode = 5,880

Note: The data is in ascending order.

5,710	5,755
5,850	5,880
5,880	5,890
5,920	5,940
5,950	6,050
6,130	6,325

Weighted Mean

- In some instances the mean is computed by giving each observation a weight that reflects its relative importance.
- The choice of weights depends on the application.
- The weights might be the number of credit hours earned for each grade, as in GPA.
- In other weighted mean computations, quantities such as pounds, dollars, or volume are frequently used.

Weighted Mean

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i}$$

where: x_i = value of observation i

w_i = weight for observation i

Numerator: sum of the weighted data values

Denominator: sum of the weights

If data is from a population, μ replaces $\bar{x}$.

Weighted Mean

Example: Purchase of Raw Material

Consider the following sample of five purchases of a raw material over a period of three months:

Purchase	Cost per Pound ($)	Number of Pounds
1	3.00	1200
2	3.4	500
3	2.8	2750
4	2.9	1000
5	3.25	800

Weighted Mean

Example: Purchase of raw material

Purchase	Cost per Pound ($) x_i	Number of Pounds w_i	$w_i x_i$
1	3.00	1200	3600
2	3.4	500	1700
3	2.8	2750	7700
4	2.9	1000	2900
5	255755	800	2600

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i} = \frac{18{,}500}{6{,}250} = 2.96 = \$2.96$$

FYI, equally weighted (simple) mean = $3.07

Geometric Mean

- The geometric mean is calculated by finding the *n*th root of the product of *n* values.

$$\bar{x}_g = \sqrt[n]{(x_1)(x_2)...(x_n)}$$
$$= \left[(x_1)(x_2)...(x_n)\right]^{1/n}$$

- It is often used in analyzing growth rates in financial data (where using the arithmetic mean will provide misleading results).
- It should be applied anytime you want to determine the mean rate of change over several successive periods (be it years, quarters, weeks, . . .).
- Other common applications include changes in populations of species, crop yields, pollution levels, and birth and death rates.

Geometric Mean: Example – Mutual Fund

Year	Annual Return %	Growth Factor
1	−22.1	0.779
2	28.7	1.287
3	10.9	1.109
4	4.9	1.049
5	15.8	1.158
6	5.5	1.055
7	−37.0	0.630
8	26.5	1.265
9	15.1	1.151
10	2.1	1.021

$$\bar{x}_g = \sqrt[10]{(0.779)(1.287)(1.109)(1.049)(1.158)(1.055)(0.630)(1.265)(1.151)(1.021)}$$

$$= \sqrt[10]{1.334493}$$

$$= 1.029275$$

Average growth rate per period is $(1.029275 - 1)(100) = 2.9\%$

BUSINESS SCHOOL | RWTH AACHEN UNIVERSITY

Percentiles

- A percentile provides information about how the data are spread over the interval from the smallest value to the largest value.
- Admission test scores for colleges and universities are frequently reported in terms of percentiles.
- The *p*th percentile of a data set is a value such that at least *p*% of the items take on this value or less and at least $(100 - p)\%$ of the items take on this value or more.

Percentiles

Arrange the data in ascending order.

Compute L_p, the location of the pth percentile.

$$L_p = \left(\frac{p}{100}\right)(n+1)$$

80th Percentile

Example: Monthly Starting Salary $L_p = (p/100)(n+1) = (80/100)(12+1) = 10.4$

(the 10th value plus .4 times the difference between the 11th and 10th values)

$$80\text{th Percentile} = 6{,}050 + 0.4(6{,}130 - 6{,}050) = 6{,}082$$

5,710	5,755
5,850	5,880
5,880	5,890
5,920	5,940
5,950	6,050
6,130	6,325

80th Percentile

Example: Monthly Starting Salary

At least 80% of the items take on a value of 6082 or less.
10/12 = .833 or 83%

At least 20% of the items take on a value of 6082 or more.
2/12 = .167 or 16.7%

5,710	5,755
5,850	5,880
5,880	5,890
5,920	5,940
5,950	6,050
6,130	6,325

Quartiles

Quartiles are specific percentiles.

- First Quartile = 25th Percentile
- Second Quartile = 50th Percentile = Median
- Third Quartile = 75th Percentile

Third Quartile = 80th Percentile

Example: Monthly Starting Salary

$$L_p = (p/100)(n+1) = (75/100)(12+1) = 9.75$$

(the 9th value plus .75 times the difference between the 10th and 9th values)

$$\text{Third quartile} = 5{,}950 + .75(6{,}050 - 5{,}950) = 6{,}025$$

5,710	5,755
5,850	5,880
5,880	5,890
5,920	5,940
5,950	6,050
6,130	6,325

Measures of Variability

- It is often desirable to consider measures of variability (dispersion), as well as measures of location.
- For example, in choosing supplier A or supplier B we might consider not only the average delivery time for each but also the variability in delivery time for each.

Measures of Variability

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

Range

- The range of a data set is the difference between the largest and smallest data values.

$$\text{Range} = \text{Largest value} - \text{Smallest value}$$

- It is the simplest measure of variability.
- It is very sensitive to the smallest and largest data values.

Range

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

$$\text{Range} = 6{,}325 - 5{,}710 = 615$$

5,710	5,755
5,850	5,880
5,880	5,890
5,920	5,940
5,950	6,050
6,130	6,325

Interquartile Range

- The interquartile range of a data set is the difference between the third quartile and the first quartile.
- It is the range for the middle 50% of the data.
- It overcomes the sensitivity to extreme data values.

Interquartile Range

Example: Monthly Starting Salary

- 3rd Quartile $(Q_3) = 6{,}000$
- 1st Quartile $(Q_1) = 5{,}865$
- $\text{IQR} = Q_3 - Q_1 = 6{,}000 - 5{,}865 = 135$

5,710	5,755
5,850	5,880
5,880	5,890
5,920	5,940
5,950	6,050
6,130	6,325

Variance

- The variance is a measure of variability that utilizes all the data.
- It is based on the difference between the value of each observation (x_i) and the mean ($\overline{X}$ for a sample, μ for a population).
- The variance is useful in comparing the variability of two or more variables.

Variance

- The variance is the average of the squared differences between each data value and the mean.
- The variance is computed as follows:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

for a sample

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

for a population

Bessel's correction

Standard Deviation

- The standard deviation of a data set is the positive square root of the variance.
- It is measured in the same units as the data, making it more easily interpreted than the variance.

The standard deviation is computed as follows:

- For a sample

$$s = \sqrt{s^2}$$

- For a population

$$\sigma = \sqrt{\sigma^2}$$

Coefficient of Variation

- The coefficient of variation indicates how large the standard deviation is in relation to the mean.

The coefficient of variation is computed as follows:

$$\left[\frac{s}{\bar{x}} \times 100\right]\%$$

for a sample

$$\left[\frac{\sigma}{\mu} \times 100\right]\%$$

for a population

Sample Variance, Standard Deviation and Coefficient of Variation

Example: Monthly Starting Salary

Variance

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = 27,440.91$$

Standard Deviation

$$s = \sqrt{s^2} = \sqrt{27,440.91} = 165.65$$

Coefficient of Variation

$$\left[\frac{s}{\bar{x}} \times 100\right]\% = \left[\frac{165.65}{3,940} \times 100\right]\% = 4.2\%$$

Measures of Distribution Shape, Relative Location, and Detecting Outliers

- Distribution Shape
- z-Scores
- Chebyshev's Theorem
- Empirical Rule
- Detecting Outliers

Distribution Shape: Skewness

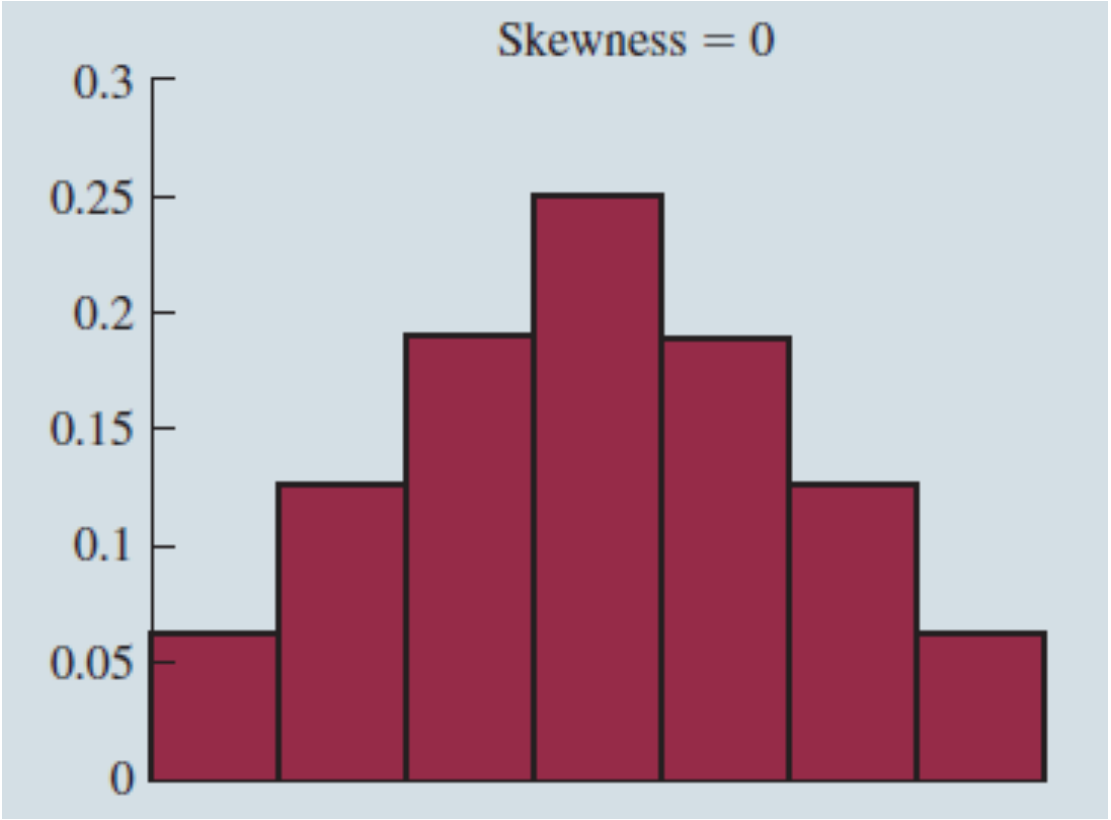
- An important measure of the shape of a distribution is called skewness.
- The formula for the skewness of sample data is

$$\text{Skewness} = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^3$$

- Skewness can be easily computed using statistical software.

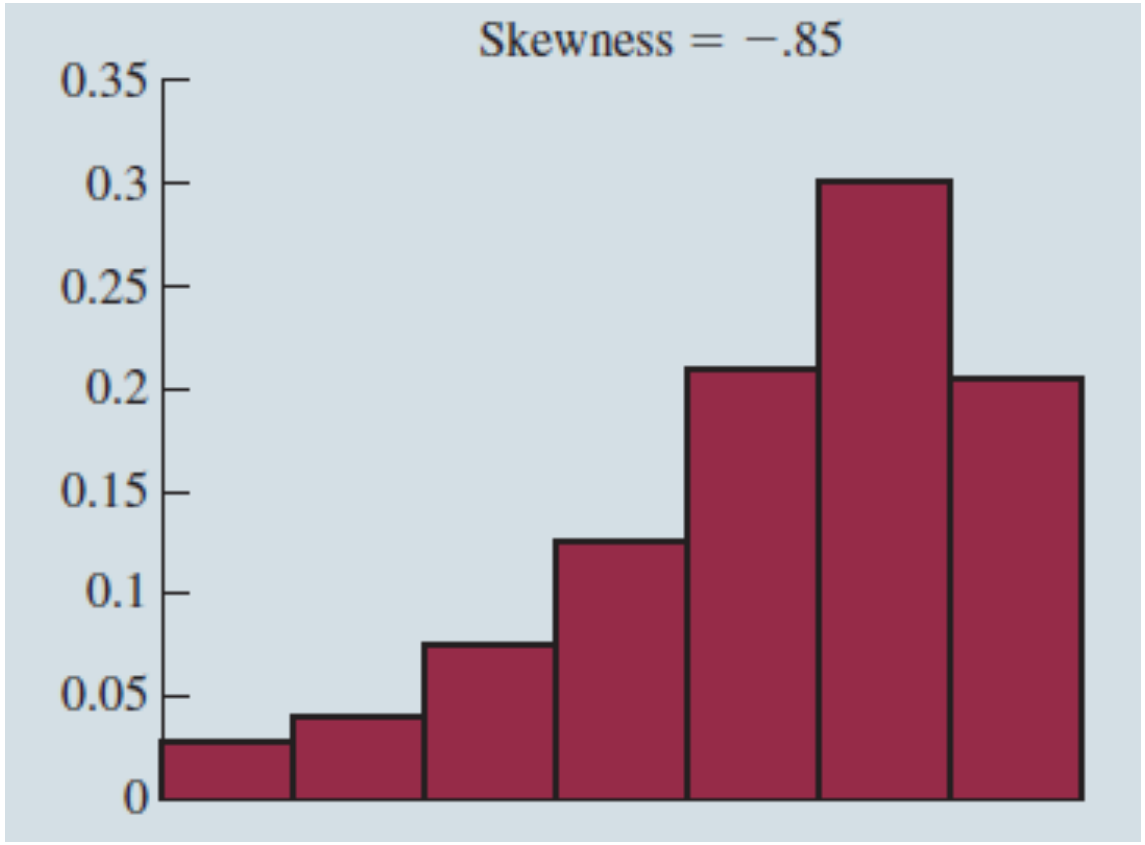
Distribution Shape: Skewness

- Symmetric (not skewed)
  - Skewness is zero.
  - Mean and median are equal.



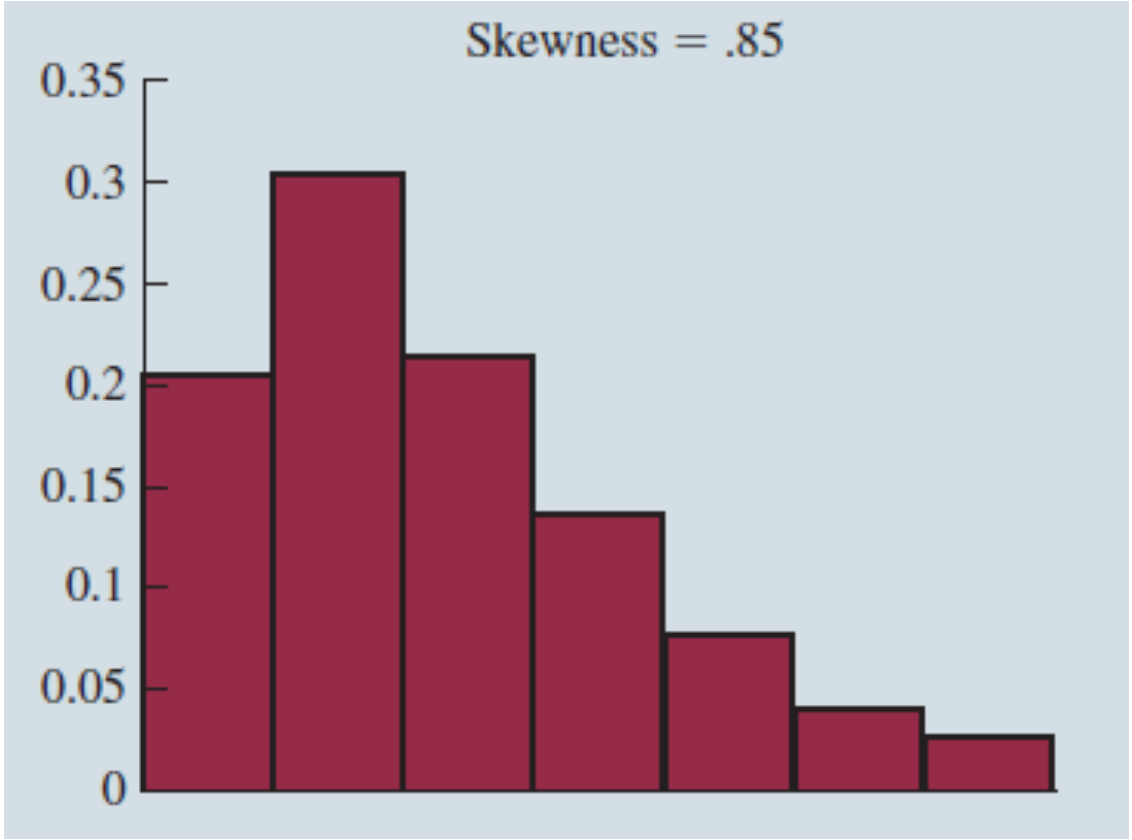
Distribution Shape: Skewness

- Moderately Skewed Left
  - Skewness is negative.
  - Mean will usually be less than the median.



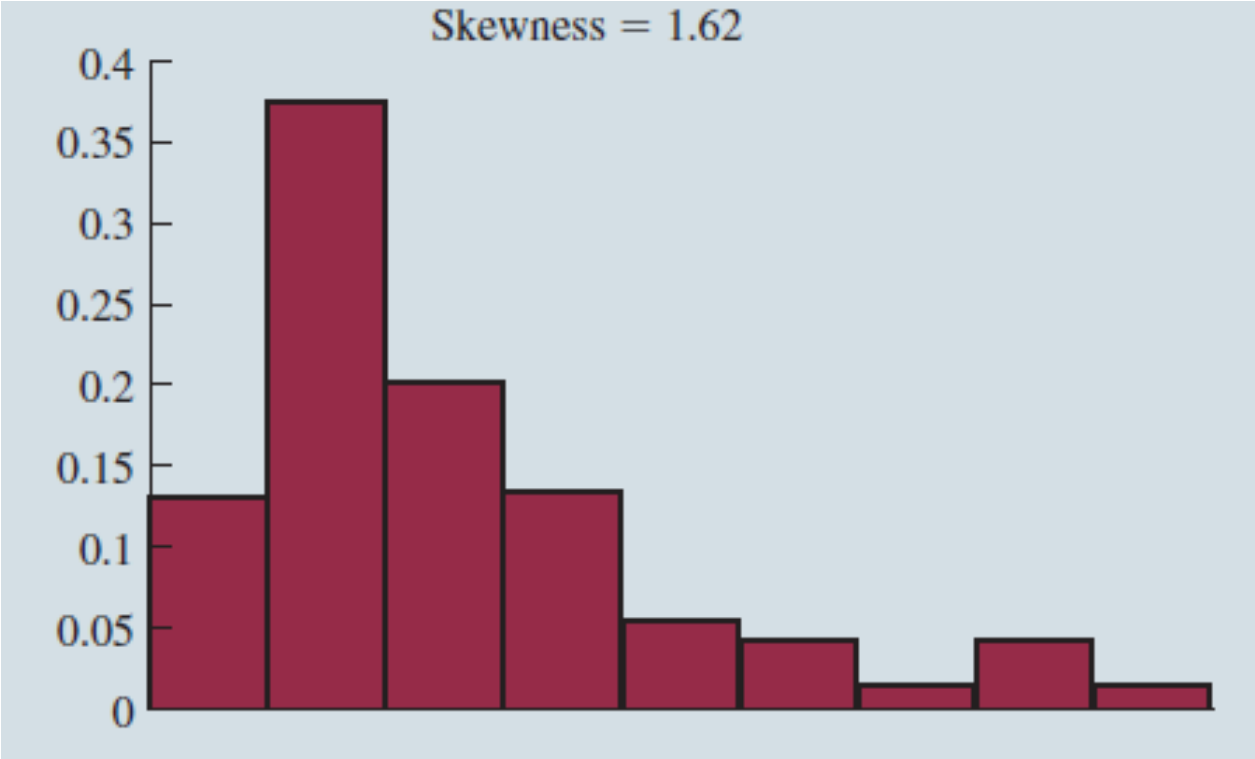
Distribution Shape: Skewness

- Moderately Skewed Right
  - Skewness is positive.
  - Mean will usually be more than the median.



Distribution Shape: Skewness

- Highly Skewed Right
  - Skewness is positive (often above 1.0).
  - Mean will usually be more than the median.



z-Scores

- The z-score is often called the standardized value.
- It denotes the number of standard deviations a data value x_i is from the mean.

$$z_i = \frac{x_i - \bar{x}}{s}$$

z-Scores

- An observation’s z-score is a measure of the relative location of the observation in a data set.
- A data value less than the sample mean will have a z-score less than zero.
- A data value greater than the sample mean will have a z-score greater than zero.
- A data value equal to the sample mean will have a z-score of zero.

z-Scores

Example: Class size data $$Z_i = \frac{x_i - \bar{x}}{s}$$

Number of students In class	Deviation about the Mean	Z score $\left(\frac{x_i - \bar{x}}{s}\right)$
46	2	2/8 = 0.25
54	10	10/8 = 1.25
42	−2	−2/8 = −0.25
46	2	2/8 = 0.25
32	−12	−12/8 = −1.5

Note: $\bar{x} = 44$ and $s = 8$ for the given data.

Chebyshev's Theorem

- At least $(1-1/z^2)$ of the items in <u>any</u> data set will be within *z* standard deviations of the mean, where z is any value greater than 1.

- Chebyshev's theorem requires $z > 1$, but *z* need not be an integer.

Chebyshev's Theorem

- At least 75% of the data values must be within $z = 2$ standard deviations of the mean.
- At least 89% of the data values must be within $z = 3$ standard deviations of the mean.
- At least 94% of the data values must be within $z = 4$ standard deviations of the mean.

Chebyshev's Theorem

Example: Midterm scores of students

Suppose the midterm test scores of 100 students in a course had a mean of 70 and a standard deviation of 5. We want to know the number of students having test scores between 60 and 80.

60 and 80 are 2 standard deviations below and above the mean respectively.

$$60 = 70 - 2(5) \longrightarrow s$$
$$80 = 70 + 2(5)$$
$$Z = 75\%$$

Chebyshev's Theorem

Example: Midterm scores of students

Number of students having test scores between 58 and 82:

$$(58-70)/5 = -2.4$$

$$(82-70)/5 = 2.4$$

$$z = 2.4$$

$$(1-1/z^2) = (1-1/(2.4)^2) = 0.826 = 82.6\%$$

Empirical Rule

When the data are believed to approximate a bell-shaped distribution:

- The empirical rule can be used to determine the percentage of data values that must be within a specified number of standard deviations of the mean.
- The empirical rule is based on the normal distribution

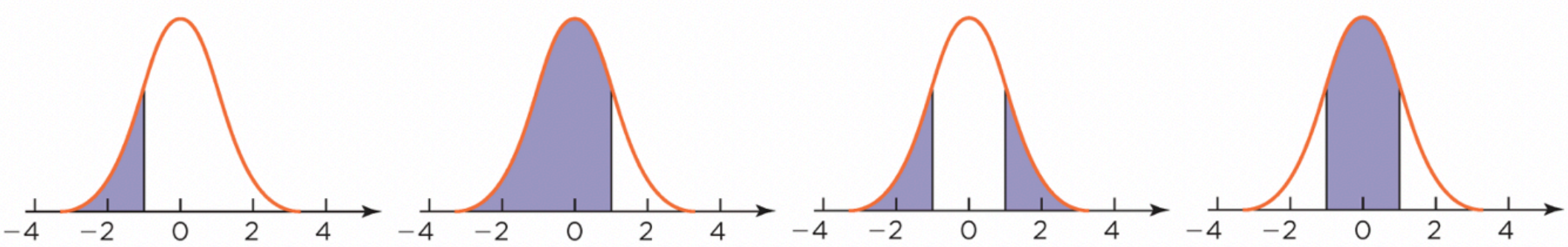
Empirical Rule

For data having a bell-shaped distribution:

- Approximately 68% of the data values will be within +/−1 standard deviation of its mean.
- Approximately 95% of the data values will be within +/ −2 standard deviations of its mean.
- Almost all (approximately 99.7%) of the data values will be within +/ −3 standard deviations of its mean.

Empirical Rule

- Bell shaped distribution



z	$P(Z \leq -z)$	$P(Z \leq z)$	$P(\|Z\| > z)$	$P(\|Z\| \leq z)$
1	0.1587	0.8413	0.3173	**0.6827**
2	0.02275	0.97725	0.04550	**0.95450**
3	0.00135	0.99865	0.00270	**0.99730**

Detecting Outliers

- An outlier is an unusually small or unusually large value in a data set.
- A data value with a z-score less than −3 or greater than +3 might be considered an outlier.
- It might be:
  - an incorrectly recorded data value
  - a data value that was incorrectly included in the data set
  - a correctly recorded unusual data value that belongs in the data set

Outliers

Example: Class size data

Number of students In class	Deviation about the Mean	Z score $\left(\frac{x_i - \bar{x}}{s}\right)$
46	2	2/8 = 0.25
54	10	10/8 = 1.25
42	−2	−2/8 = −0.25
46	2	2/8 = 0.25
32	−12	−12/8 = −1.5

- −1.5 shows the fifth class size is farthest from the mean.
- No outliers are present as the z values are within the +/−3 guideline.

Five-Number Summaries and Boxplots

- Summary statistics and easy-to-draw graphs can be used to quickly summarize large quantities of data.
- Two tools that accomplish this are five-number summaries and boxplots.

Five-Number Summaries

- Smallest Value
- First Quartile
- Median
- Third Quartile
- Largest Value

Five-Number Summaries

Example: Monthly starting salary

- Lowest Value = 5,710
- Third Quartile = 6,025
- Median = 5905
- First Quartile = 5,857.5
- Largest Value = 6,325

Monthly Starting Salary ($)	
5,710	5,755
5,850	5,880
5,880	5,890
5,920	5,940
5,950	6,050
6,130	6,325

Boxplot

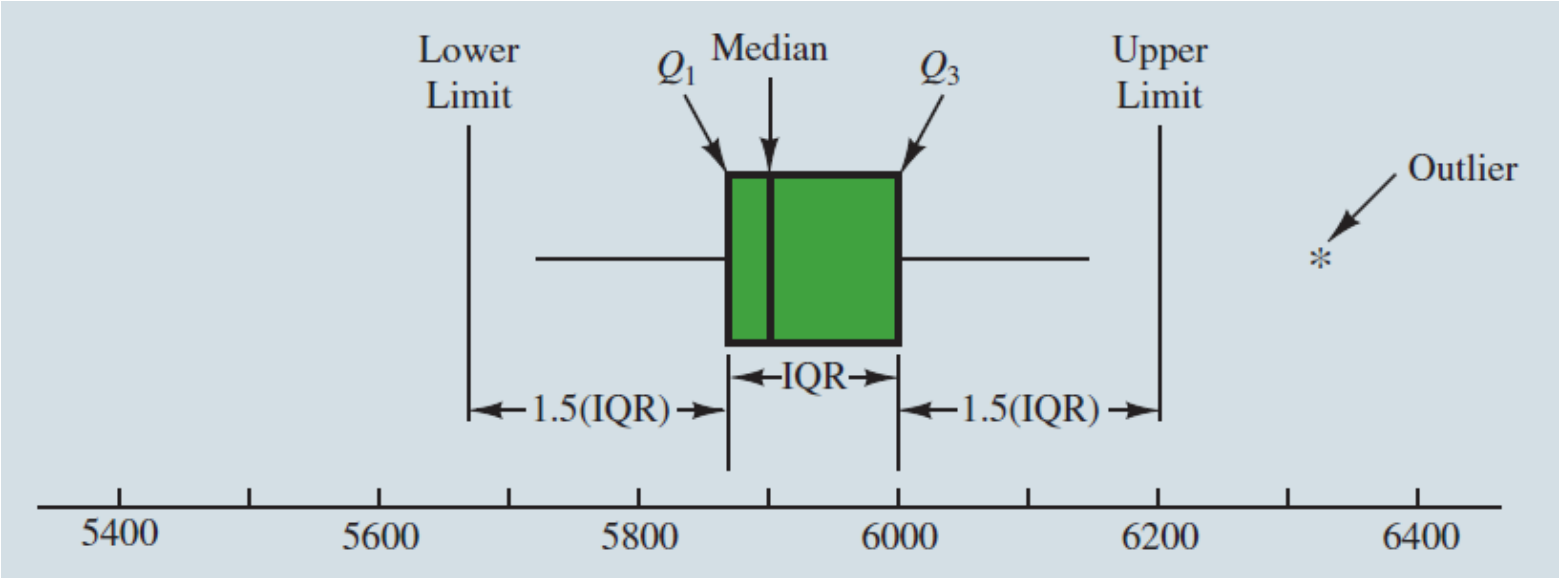
- A boxplot is a graphical summary of data that is based on a five-number summary.
- A key to the development of a boxplot is the computation of the median and the quartiles,

  Q_1 and Q_3.

- Boxplots provide another way to identify outliers.

Boxplot: Monthly starting salary

- A box is drawn with its ends located at the first and third quartiles.
- A vertical line is drawn in the box at the location of the median (second quartile).



- Limits are located using the interquartile range (IQR).
- Data outside these limits are considered outliers.
- The locations of each outlier are shown with the symbol.

Boxplot

Example: Monthly starting salary

- The lower limit is located 1.5(IQR) below Q_1.

$$\text{Lower Limit: } Q_1 - 1.5(\text{IQR}) = 5{,}857.5 - 1.5(167.5) = 5{,}606.25$$

- The upper limit is located 1.5(IQR) above Q_3.

$$\text{Upper Limit: } Q_3 + 1.5(\text{IQR}) = 6{,}025 + 1.5(167.5) = 6{,}276.25$$

- There is one outlier: 6,325.

Measures of Association Between Two Variables

- Thus far we have examined numerical methods used to summarize the data for one variable at a time.
- Often a manager or decision maker is interested in the relationship between two variables.
- Two descriptive measures of the relationship between two variables are covariance and correlation coefficient.

Covariance

- The covariance is a measure of the linear association between two variables.
- Positive values indicate a positive relationship.
- Negative values indicate a negative relationship.

Covariance

The covariance is computed as follows:

For samples:

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

For population:

$$\sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}$$

Correlation

- Correlation is a measure of linear association and not necessarily causation.
- Just because two variables are highly correlated, it does not mean that one variable is the cause of the other.

Correlation Coefficient

- The correlation coefficient is computed as follows:

For samples: $$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

For population: $$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Correlation Coefficient

- The coefficient can take on values between −1 and +1.
- Values near −1 indicate a strong negative linear relationship.

  Values near +1 indicate a strong positive linear relationship.
- The closer the correlation is to zero, the weaker the relationship.

Covariance and Correlation Coefficient

Example: San Francisco Electronics store

The store's manager wants to determine the relationship between the number of weekend television commercials shown and the sales at the store during the following week.

Covariance and Correlation Coefficient

Week	Number of Commercials	Sales ($100s)
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	49

Covariance and Correlation Coefficient

	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
	2	50	−1	−1	1
	5	57	2	6	12
	1	41	−2	−10	20
	3	54	0	3	0
	4	54	1	3	3
	1	38	−2	−13	26
	5	63	2	12	24
	3	48	0	−3	0
	4	59	1	8	8
	2	46	−1	−5	5
Totals	30	510	0	0	99

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{99}{10-1} = 11$$

Covariance and Correlation Coefficient

Example: San Francisco Electronics Store

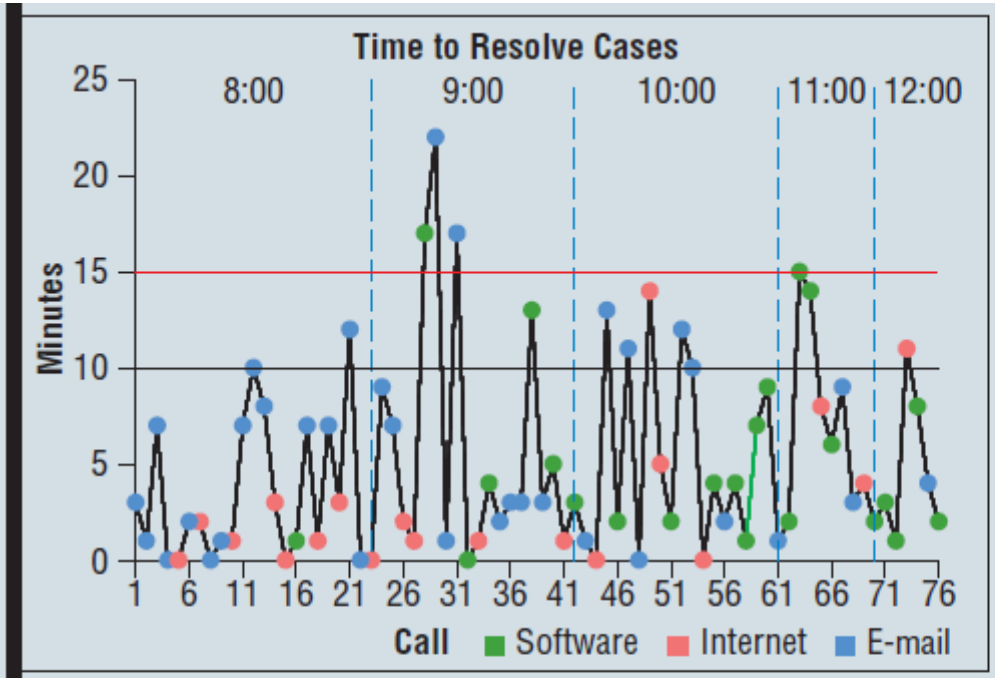
Sample Covariance $$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} = 99/9 = 11$$

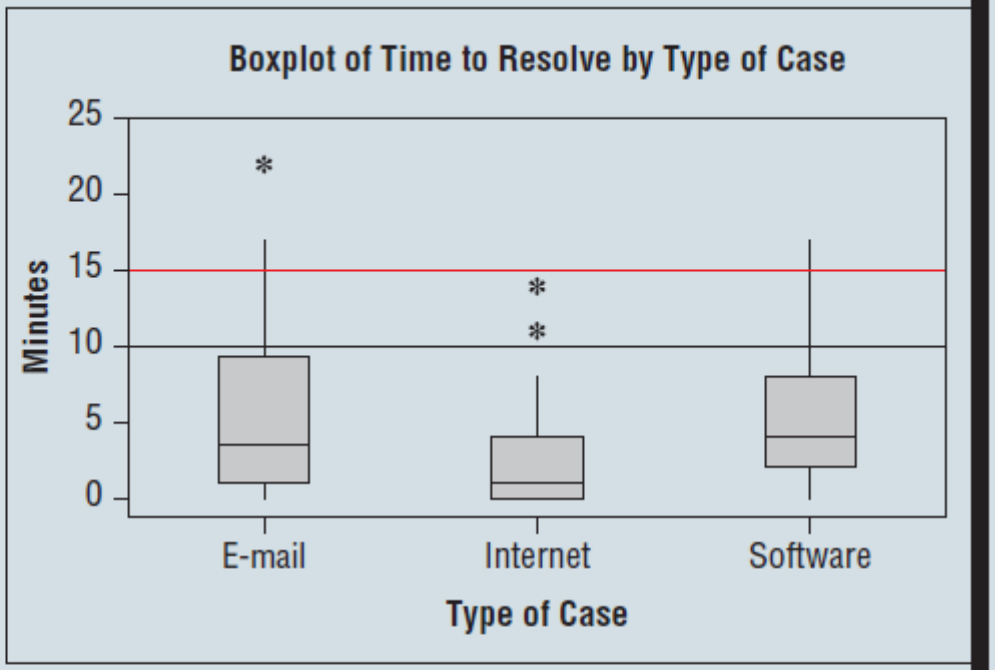
Sample Correlation Coefficient $$r_{xy} = \frac{s_{xy}}{s_x s_y} = 11/(1.49 \times 7.93) = 0.93$$

Data Dashboards: Adding Numerical Measures to Improve Effectiveness

- Data dashboards are not limited to graphical displays.
- The addition of numerical measures, such as the mean and standard deviation of KPIs, to a data dashboard is often critical.
- Dashboards are often interactive.
- Drilling down refers to functionality in interactive dashboards that allows the user to access information and analyses at increasingly detailed level.

Data Dashboards: Adding Numerical Measures to Improve Effectiveness





Summary Statistics—Resolved Cases

Type of Case	Cases	Mean	Median	Std. Dev
E-mail	34	5.8	3.5	5.4
Internet	19	3.0	1.0	4.0
Software	23	5.4	4.0	5.0

Hour	Cases	Mean	Median	Std. Dev
8:00	22	3.5	2.0	3.7
9:00	19	5.8	3.0	6.6
10:00	19	5.3	4.0	4.8
11:00	9	6.9	6.0	5.1
12:00	6	4.8	3.5	3.9

RWTH BUSINESS SCHOOL

Mathematics & Statistics
M.Sc. Data Analytics and Decision Science

Prof. Dr. Thomas S. Lontzek

AACSB ACCREDITED

BUSINESS SCHOOL | RWTH AACHEN UNIVERSITY