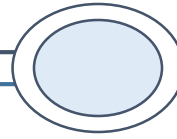


Exploration Data Analysis with Descriptive Statistics



IT525: Data Science

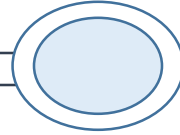
Department of Computer Science

Faculty of Science

Srinakharinwirot University

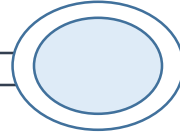
Semester 2/2017

Types of Data



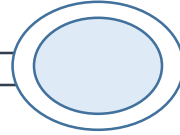
- Types of data can be classified according to the following three criteria.
 - How data are organized.
 - Structured, semi-structured, and unstructured
 - Whether data are quantifiable
 - Qualitative and quantitative (discrete and continuous)
 - Levels of measurement
 - Nominal, ordinal, interval, and ratio.
- Specifying type of data is important in data science because it determines type of operations that can be performed.

Types of Data According to Data Organizations



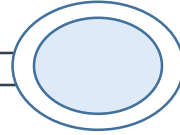
- **Unstructured data** – unorganized data or data in textual natural free form.
 - Example: Email content, textual documents that do not followed any standard format.
- **Semi-structured data** – textual data in some specific format
 - Example: Emails, Web page content, and research articles
- **Structured data** – data that is organized in a tabular format (rows and columns) where rows corresponding to observations and columns to features or characteristics
 - Example: tabular data, data in comma-separated value (csv) format, and data in databases
- **Notes**: data in unstructured and semi-structured formats normally must be pre-processed and transformed into a format that is suitable for further analyses, such as a tabular format.

Types of Data Based on Quantifiability



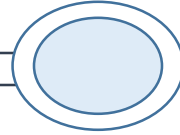
- **Qualitative data** – data that can be described using natural categories, natural languages, or numbers, but basic mathematical operations such as addition/subtraction CANNOT be performed on them.
 - Examples: gender, brand name, and **postal code**.
 - Postal code is mostly represented as a number, but basic mathematical operations cannot be performed on it
- **Quantitative data** – data that can be described using numbers, and some basic mathematical procedures such as addition/subtraction can be performed on them.
 - **Discrete value** – data that can be counted. It can only exist on certain values.
 - Examples:
 - **Continuous value** – data that can be measured. It exists on an infinite range of values.
 - Examples:

Types of Data Based on Levels of Measurement



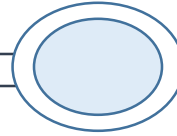
- **Nominal data** – data that are described by names or categories.
 - Example: Categories, genders, and nationalities.
 - Mathematical operations allowed: Equality and set membership
 - Measure of central tendency: Mode
- **Ordinal data** -- data that can ordered and ranked, but differences between data points have no meaning.
 - Example: Likert scales, ratings, and preferences,
 - Mathematical operations allowed: Ordering and comparison
 - Measure of central tendency: Median

Types of Data Based on Levels of Measurement



- **Interval data** -- Differences between data points are meaningful, but data have no absolute or natural zero.
 - Example: Time and temperature in Celsius and Fahrenheit
 - Mathematical operations allowed: Addition and subtraction
 - Measure of central tendency: Arithmetic mean
- **Ratio data** -- Data that have absolute or natural zero.
 - Example: Weight, bank balance, temperature in Kelvin
 - Mathematical operations allowed: Multiplication and division
 - Measure of central tendency: Geometric mean

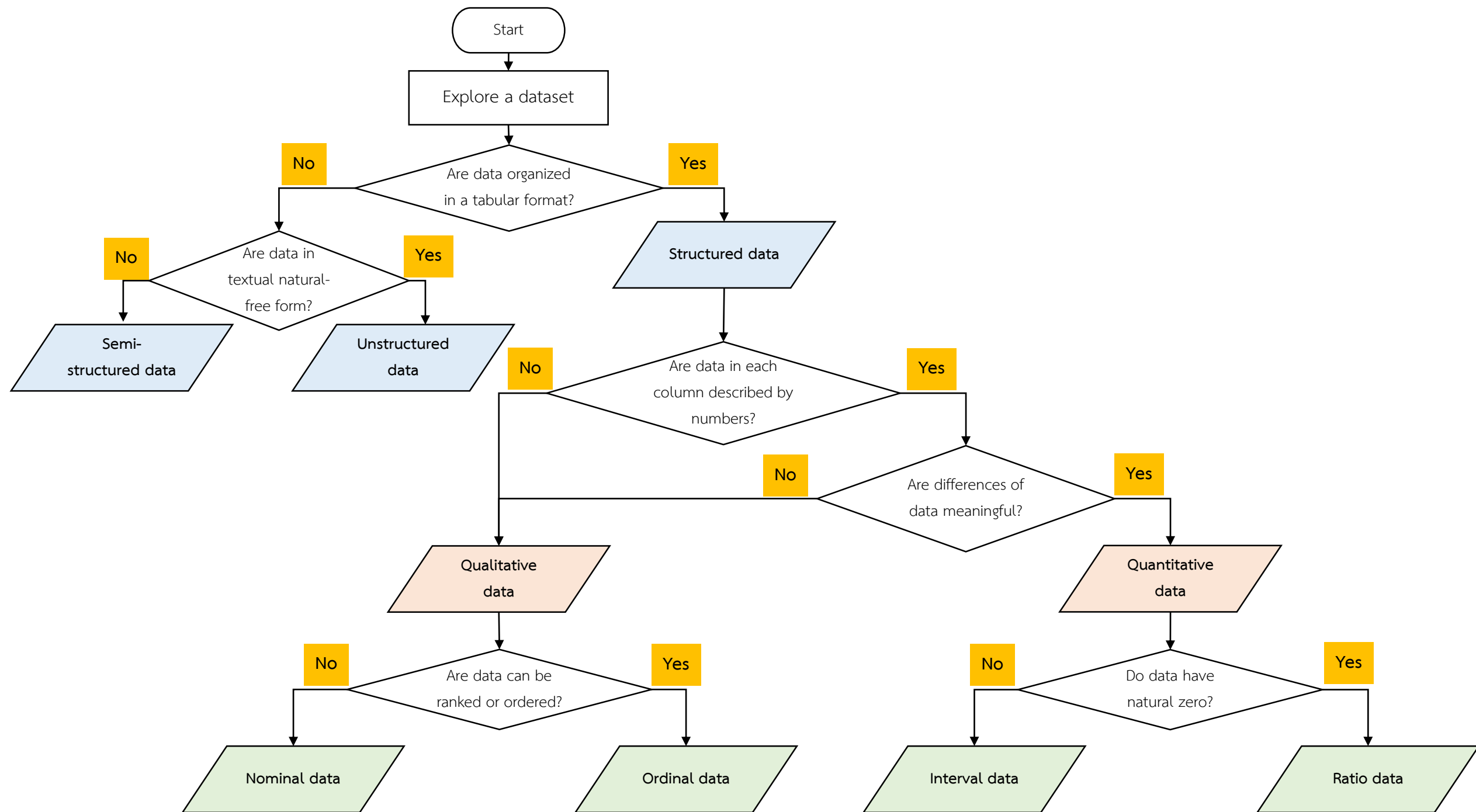
Types of Data According to Levels of Measurement



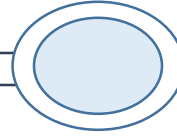
Types of data	Description	Mathematics Operations Allowed	Measure of central tendency	Examples
Nominal level	Data are described by names or categories.	Equality and set membership	Mode	Countries, gender, nationality.
Ordinal level	Data can ordered and ranked, but differences between data points have no meaning.	Ordering and comparison	Median	Likert scales, rating, preference
Interval level	Differences between data points are meaningful, but data have no absolute or natural zero.	Addition and subtraction	Arithmetic mean	Temperature in Celsius and Fahrenheit
Ratio level	Data have natural zero.	Multiplication and division	Geometric mean	Temperature in Kelvin, weight, bank balance.

Notes:

- Mathematics and statistics operations that are allowed to be performed in the lower levels of measurement can be done on data in the higher levels of measurement.
- Although arithmetic mean, calculated as the sum of n number divided by n , contains a division but the division is not performed directly on the interval data.
- Geometric mean is computed as the n -th root of all products of n numbers.
- Temperatures in Celsius and Fahrenheit can be converted to zero Kelvin, but their values become negative, which does not make sense for zero heat.



Example



Unstructured data

Donald Trump's twitter message about North Korea leader on Jan 2, 2018:

"North Korean Leader Kim Jong Un just stated that the "Nuclear Button is on his desk at all times." Will someone from his depleted and food starved regime please inform him that I too have a Nuclear Button, but it is a much bigger & more powerful one than his, and my Button works!"

Name of coffee shop	Monthly revenue	Postal Code	Average monthly customers	Continent of coffee origin	Customer rating	Survey Date
Rabika	300,000.00	10230	6,000	Ethiopia	Low	20/01/2018
Starbucks	1,800,000.00	10110	15,000	Africa	High	15/01/2018
Intanin	450,000.00	10230	5,500	Thailand	Moderate	10/01/2018

Types of data – the whole dataset is a structured data.

- Name of coffee shop – qualitative/nominal
- Monthly revenue – quantitative/continuous/ratio
- Postal code – qualitative/nominal
- Average monthly customers – quantitative/discrete/ratio
- Continent of coffee origin – qualitative/nominal
- Customer rating – qualitative/ordinal
- Survey date – quantitative/interval

หมายเหตุ: ข้อมูลในตารางเป็นข้อมูลสมมติ

Example: Read a sample dataset in csv format with column names from book url.

```
# Import pandas package
import pandas as pd

# Read a csv file hosted on a Web page.
url = 'https://raw.githubusercontent.com/sinanuozdemir/principles_of_data_science/master/data/chapter_2/drinks.csv'

# Assign data to a variable as a DataFrame object.
drinks = pd.read_csv(url)

# Display the top 10 rows.
drinks.head(10)

# Show a summary statistics for a column of qualitative data
drinks['continent'].describe()

# Show a summary statistics for a column of quantitative data
drinks['beer_servings'].describe()
```

Try to specify type of data in each column

Example: Read *iris* dataset in csv format without column names from UCI machine learning repository

```
# Import pandas package
import pandas as pd

# Read a csv file hosted on a Web page.
url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data'

# Set column names, and assign data to a variable as a DataFrame object.
column_names = ['sepal length', 'sepal width', 'petal length', 'petal width', 'class']
iris_dataset = pd.read_csv(url, sep=',', header=None, names = column_names)

# Change column names
new_column_names = ['sepal_length', 'sepal_width', 'petal_length', 'petal_width', 'class']
iris_dataset.columns = new_column_names

# Show column names
iris_dataset.columns

# Display the top 5 rows
iris_dataset.head()

# Display the top 10 rows.
iris_dataset.head(10)

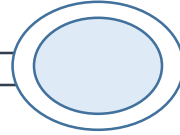
# Display the bottom 20 rows
iris_dataset.tail(20)

# Show a summary statistics for a column of qualitative data
iris_dataset['class'].describe()

# Show a summary statistics for a column of quantitative data
iris_dataset['petal_width'].describe()
```

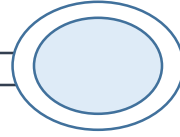
Try to specify type of data in each column.

Population and Sample



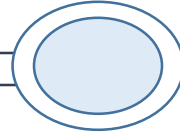
- **Population** – All items of interest for a particular decision or investigation
 - Example: All married drivers over the age of 25 in the U.S., or *all* subscribers to Netflix
- **Sample** – a subset of a population
 - Example: A list of individuals who rented a comedy from Netflix last year.
 - A good sample should accurately reflect all members in a population
 - Whether a sample is a true representative of a population depends on how the sample data are intended to be used.
 - Example: A six-student sample of 3 male and 3 female students is drawn from a class of 60 students in which there are equal number of male and female students.

Population and Sample



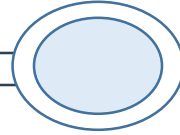
- Why using a sample rather than a population?
 - Most populations are too large to deal with effectively or practically.
 - All Internet users
 - Some population are infinite or uncountable.
 - Number of germs in a patient's body
 - Some population are unavailable during the data collection process.
 - No contact information or phone numbers of some alumni
- As a result, most data used in studies are sample data.

Population and Sample



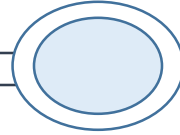
- **Sampling** – a process of obtaining a sample from a population
- **The purpose of sampling** – to obtain sufficient information to draw a valid inference about a population.
 - **Example**: marketing researchers use sampling to measure customer perceptions on new or existing goods and services.

Four Main Sampling Techniques



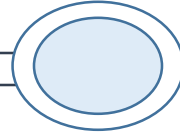
- **Random sampling** — subjects are selected by using chance methods or random numbers.
 - Example: Number each subject in the population, put numbered cards in a bowl, mix them together, and select as many cards as needed.
- **Systematic sampling** — subjects are selected by every k^{th} member of the population where k is a counting number.
 - Example: Need 50 out of 2,000 subjects in the population to be a sample. Compute $k = 2,000/50 = 40$. Number each subject in the population, select the first subject randomly, the next subjects are selected by adding the number of current subject with the value of k .

Four Main Sampling Techniques



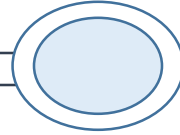
- **Stratified sampling** — Divide the population into subgroups/strata based on some characteristic relevant to the study, and then randomly select subjects from each subgroup.
 - Example: A study on finding differences in opinion regarding student activities between Mathematic department and Physics department. Students from both departments are randomly selected to use in the sample.
- **Cluster sampling** — Divide the population into sections or clusters based on some means such as geographic area, then select one or more clusters (randomly), and next uses all members in the chosen clusters as the members of the sample.
 - Used when the population is large or when it involves subjects residing in a large geographic area.
 - Example: A study to evaluate people's satisfaction on public transportation in Bangkok. Divide the population based on city, randomly select one or more cities as clusters, and use people residing in those selected cities as a sample.

Other Sampling Techniques



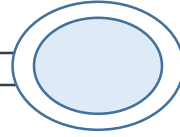
- **Convenience sampling** — subjects are selected based on the convenience of the researcher.
 - Example: Select subjects entering a shopping mall for interviews to find what stores they will be patronizing.
- **Volunteer sampling or self-selected sampling** — The subjects or respondents decide for themselves if they wish to be included in the sample.
 - Example: A radio station asks a question regarding a situation and then asks people to call in to give their answers.

Descriptive Statistics



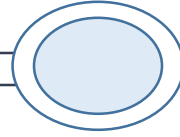
- **Measures of central tendency** – find the center of the distribution or the most typical case.
 - Mean, median, mode, and midrange
- **Measures of variation or dispersion** – determine the spread of the data values.
 - Range, variance, and standard deviation.
- **Measures of shape** – determine the shape of data distribution
 - Skewness and peakedness
- **Measures of position** – tells the position of a specific data value within the data set or its relative position in comparison with other data values.
 - Percentiles, deciles, and quartiles. – also called norms.

Measures of Central Tendency



- **Measures of central tendency** – the estimates of a single value that represents the “centering” of a set of data.
 - The value that represents the majority of data.
- Four statistical measures
 - Arithmetic Mean
 - Median
 - Mode
 - Midrange

Arithmetic Mean



- Arithmetic Mean (or Average) – the sum of the observations divided by the number of observations.
- Calculating the Mean:

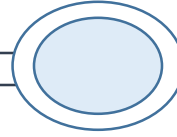
- The mean of a population (μ) of N observations

$$\mu = \frac{\sum_{i=1}^N x_i}{N}$$

- The mean of a sample \bar{x} of n observations

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

Arithmetic Mean: Example



- Example: Compute the mean of the following sample data set.

- 110, 76, 29, 38, 105, 31

$$\bar{X} = \frac{\sum X}{n} = \frac{110 + 76 + 29 + 38 + 105 + 31}{6} = \frac{389}{6} = 64.8$$

```
>>> mylist = [110, 76, 29, 38, 105, 31]
>>> mymean = sum(mylist)/len(mylist)
>>> print(mymean)
64.83333333333333
```

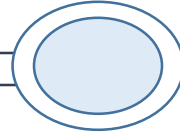
```
>>> import numpy as np
>>> mylist = [110, 76, 29, 38, 105, 31]
>>> mymean = np.mean(mylist)
>>> print(mymean)
64.833333333333329
```

```
>>> import pandas as pd
>>> mylist = [['data', [110, 76, 29, 38, 105, 31]]]
>>> df = pd.DataFrame.from_items(mylist)
>>> mymean = df['data'].mean()
>>> print(mymean)
64.83333333333333
```

References:

- Creating a data frame object from a list in Pandas at <http://pbpython.com/pandas-list-dict.html>

Arithmetic Mean



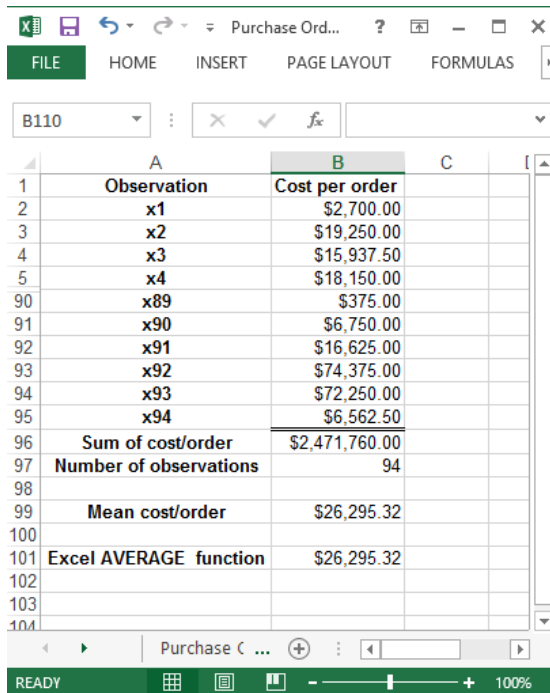
- A property of the mean – the sum of the deviations of each observation from the mean is zero.

$$\sum_i (x_i - \bar{x}) = 0$$

- Interpretation:
 - The sum of the deviations above the mean *is equal to* the sum of the deviations below the mean.
 - It does NOT imply that half the data lie above or below the mean.
 - The mean value is unique for every set of data.
 - The mean can be affected by outliers.

Arithmetic Mean

- **Example #01:** Computing the Mean Cost per Order using *Purchase Orders.xlsx*



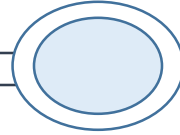
The screenshot shows an Excel spreadsheet with the following data:

Observation	Cost per order
x1	\$2,700.00
x2	\$19,250.00
x3	\$15,937.50
x4	\$18,150.00
x89	\$375.00
x90	\$6,750.00
x91	\$16,625.00
x92	\$74,375.00
x93	\$72,250.00
x94	\$6,562.50
Sum of cost/order	\$2,471,760.00
Number of observations	94
Mean cost/order	\$26,295.32
Excel AVERAGE function	\$26,295.32

```
>>> import os
>>> os.getcwd()
'C:\\Users\\MyWindows10\\
>>> path = os.chdir('C:\\Users\\MyWindows10\\Desktop\\lecture02\\datasets\\Data_Files')
>>> os.getcwd()
'C:\\Users\\MyWindows10\\Desktop\\lecture02\\datasets\\Data_Files'
>>> filename = 'Purchase Orders.xlsx'
>>> df = pd.read_excel(filename, sheetname='Data', skiprows=2)
>>> df.columns
Index(['Supplier ', 'Order No.', 'Item No.', 'Item Description', 'Item Cost',
      'Quantity', 'Cost per order', 'A/P Terms (Months)', 'Order Date',
      'Arrival Date'],
      dtype='object')
>>> df['Cost per order'].mean()
26295.31914893617
```

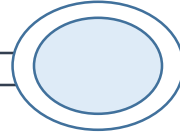
Datasets available at
<http://media.pearsoncmg.com/ph/bp/bridgepages/teamsite/evans/>

Median



- **Median** – specifies the middle value when the data are arranged from the least to the greatest.
 - Therefore, half of the data are below the median, and the remaining half of the data are above the median
- Computing the Median:
 - For an odd number of observations
 - The median = the middle of the sorted numbers
 - For an even number of observations
 - The median = the mean of the two middle numbers
- The median is NOT affected by outliers.

Median



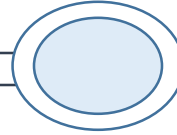
- To find a median in a data set,
 - Sort the data values in ascending order.
 - Count the number of data values, n , in the data set.
 - Determine the data value to be used as the median by:
 - If n is odd, select the middle data value as the median.

$$\text{Median} = \text{data value at } \text{floor}(n/2) + 1$$

- If n is even, compute the mean of the two middle values by adding them and divided the sum by 2.

$$\text{Median} = \frac{\text{data value at } \frac{n}{2} + \text{data value at } (\frac{n}{2} + 1)}{2}$$

Median: Example



- Example: 177, 153, 122, 141, 189, 155, 162, 165, 149, 157, 240

- Step 1: Sort the data values in ascending order.

- 122, 141, 149, 153, 155, 157, 162, 165, 177, 189, 240

- Step 2: Count the number of data values.

- $n = 11$ (odd number)

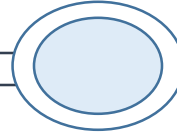
- Step 3: Determine the number of be used as the median.

- Since n is odd, the median is the data value at the position $\text{floor}(11/2) + 1 = 5 + 1 = 6$
- The median is 157, which is the 6th position in the data set.

```
>>> odd = [177, 153, 122, 141, 189, 155, 162, 165, 149, 157, 240]
>>> n = len(odd)
>>> position = n//2 + 1
>>> odd.sort(reverse=False)
>>> median = odd[position-1]
157
```

Note: List index starts at zero.

Median: Example

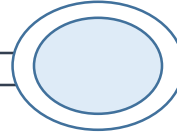


- Example: 684, 764, 656, 702, 856, 1133, 1132, 1303
 - Step 1: Sort the data values in ascending order.
 - 656, 684, 702, 764, 856, 1132, 1133, 1303
 - Step 2: Count the number of data values.
 - $n = 8$ (even number)
 - Step 3: Determine the number of be used as the median.
 - Since n is even
 - The data value at the position $n/2 = 8/2 = 4$, which is 764.
 - The data value at the position $n/2+1 = 4+1 = 5$, which is 856
 - The median is $(764+856)/2 = 1620/2 = 810$

```
>>> even = [684, 764, 656, 702, 856, 1133, 1132, 1303]
>>> n=len(even)
>>> even.sort()
>>> print((even[n//2-1]+even[n//2])/2)
810
```

Note: List index starts at zero.

Median

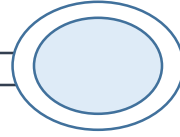


- **Example #02:** Finding the Median Cost per Order using *Purchase Orders.xlsx*

	A	B	C	D	E
1	Rank	Cost per order			
2	1	\$68.75			
3	2	\$82.50			
44	43	\$13,650.00			
45	44	\$14,910.00			
46	45	\$14,910.00			
47	46	\$15,087.50			
48	47	\$15,562.50		\$15,562.50	
49	48	\$15,750.00		\$15,750.00	
50	49	\$15,937.50	Average	\$15,656.25	
51	50	\$16,276.75			
93	92	\$110,000.00			
94	93	\$121,000.00			
95	94	\$127,500.00			
96					
97					
98					
99					

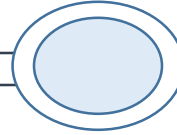
```
>>> import os
>>> os.getcwd()
>>> path = os.chdir('C:\\Users\\MyWindows10\\Desktop\\lecture02\\datasets\\Data_Files')
>>> os.getcwd()
>>> filename = 'Purchase Orders.xlsx'
>>> wb = pd.read_excel(filename, sheetname=None, skiprows=2)
>>> wb.keys() # Show all sheet name
odict_keys(['Data', 'Mean', 'Median', 'Variance', 'z-scores', 'Descriptive Statistics'])
>>> df = wb['Data']
>>> df.columns
Index(['Supplier ', 'Order No.', 'Item No.', 'Item Description', 'Item Cost',
      'Quantity', 'Cost per order', 'A/P Terms (Months)', 'Order Date',
      'Arrival Date'],
      dtype='object')
>>> df['Cost per order'].median()
15656.25
```

Mode



- **Mode** – the observation that occurs most often in a dataset.
 - A data set can have:
 - All values are mode --- their frequency of occurrence are equal.
 - one mode – called unimodal
 - two modes – called bimodal
 - more than two modes – called multimodal
- When a data set has more than one mode, each value is used as the mode.

Mode for Numeric Values: Example

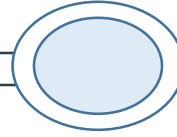


- Example: 18.0, 14.0, 34.5, **10**, 11.3, **10**, 12.4, **10**
 - The mode is 10 which occurs 3 times.
- Example: 'fish', **cat**, **cat**, 'dog', 'chicken'
 - The mode is 'cat' which occurs 2 times.
- Example: **104**, **104**, **104**, **104**, **104**, 107, **109**, **109**, **109**, 110, **109**, 111, 112, 111, **109**
 - The modes are 104 and 109 because each of them occurs the most frequent. 5 times.
- Example: 649, 789, 642, 613, 610, 600
 - All values are mode because each value occurs only once.

```
>>> import statistics
>>> numeric_data = [18.0, 14.0, 34.5, 10, 11.3, 10, 12.4, 10]
>>> statistics.mode(num_data)
10.0
>>> categorical_data = ['fish', 'cat', 'cat', 'dog', 'chicken']
>>> statistics.mode(categorical_data)
'cat'
```

```
>>> from collections import Counter
>>> def find_mode(lst):
    counter = Counter(lst)
    _, val = counter.most_common(1)[0]
    return [x for x, y in counter.items() if y == val]
>>> data = [104, 104, 104, 104, 104, 107, 109, 109, 109, 110,
109, 111, 112, 111, 109]
>>> find_mode(data)
```

Mode for Grouped Data: Example

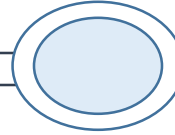


- The mode for grouped data, called *the modal class*, is the class with the highest frequency.
 - Example: Find the modal class for the following frequency distribution.

Class	Frequency
5.5-10.5	1
10.5-15.5	2
15.5-20.5	3
20.5-25.5	5
25.5-30.5	4
30.5-35.5	3
35.5-40.5	2

The modal class is 20.5-25.5.

Mode: Example

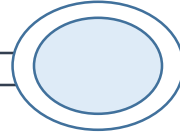


● Example #03: Finding the Mode using *Purchase Orders.xlsx*

	A	B	C	D	E	F	G	H	I	J
1	Purchase Orders									
2										
3	Supplier	Order No.	Item No.	Item Description	Item Cost	Quantity	Cost per order	A/P Terms (Months)	Order Date	Arrival Date
4	Manley Valve	C1111	9955	Door Decal	\$ 0.55	125	\$ 68.75	30	11/05/11	11/10/11
5	Manley Valve	A9876	9955	Door Decal	\$ 0.55	150	\$ 82.50	30	11/01/11	11/06/11
6	Hulkey Fasteners	C8989	9966	Hatch Decal	\$ 0.75	500	\$ 375.00	30	08/25/11	08/31/11
87	Hulkey Fasteners	D3232	1122	Airframe fasteners	\$ 4.25	17,000	\$ 72,250.00	30	10/11/11	10/19/11
88	Hulkey Fasteners	D2121	1122	Airframe fasteners	\$ 4.25	17,500	\$ 74,375.00	30	10/25/11	11/03/11
89	Hulkey Fasteners	C3232	1122	Airframe fasteners	\$ 4.25	18,000	\$ 76,500.00	30	10/01/11	10/08/11
90	Steelpin Inc.	C0467	8008	Machined Valve	\$ 645.00	120	\$ 77,400.00	30	10/28/11	11/04/11
91	Manley Valve	C3333	8148	Machined Valve	\$ 655.50	125	\$ 81,937.50	30	10/10/11	10/17/11
92	Hulkey Fasteners	C1212	1122	Airframe fasteners	\$ 4.25	19,500	\$ 82,875.00	30	08/05/11	08/13/11
93	Steelpin Inc.	B3222	8008	Machined Valve	\$ 645.00	150	\$ 96,750.00	30	10/15/11	10/26/11
94	Alum Sheeting	A0446	5417	Control Panel	\$ 255.00	406	\$ 103,530.00	30	09/01/11	09/10/11
95	Durrable Products	A1456	5454	Control Panel	\$ 220.00	500	\$ 110,000.00	45	10/15/11	10/20/11
96	Durrable Products	A1344	5454	Control Panel	\$ 220.00	550	\$ 121,000.00	45	10/09/11	10/14/11
97	Alum Sheeting	A0433	5417	Control Panel	\$ 255.00	500	\$ 127,500.00	30	10/20/11	10/27/11
98										
99										
100							Mode (Single)	30		
101										
102										
103										

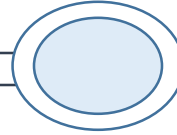
```
>>> import os
>>> os.getcwd()
>>> path = os.chdir('C:\\Users\\MyWindows10\\Desktop\\lecture02\\datasets\\Data_Files')
>>> os.getcwd()
>>> filename = 'Purchase Orders.xlsx'
>>> sheet = pd.read_excel(filename, sheetname='Data', skiprows=2)
>>> sheet.columns
Index(['Supplier ', 'Order No.', 'Item No.', 'Item Description', 'Item Cost',
      'Quantity', 'Cost per order', 'A/P Terms (Months)', 'Order Date',
      'Arrival Date'],
      dtype='object')
>>> from collections import Counter
>>> def find_mode(lst):
    counter = Counter(lst)
    _, val = counter.most_common(1)[0]
    return [x for x, y in counter.items() if y == val]
>>> find_mode(sheet['A/P Terms (Months)'])
[30]
```


Midrange



- **Midrange** – the average of the greatest and the least values in the data set.
 - Caution: extreme values can distort the midrange value.
 - Often used for only small sample sizes (because it uses only two values of data)
- Computation:
 - $\text{Midrange} = (\text{largest value} + \text{smallest value})/2$

Midrange



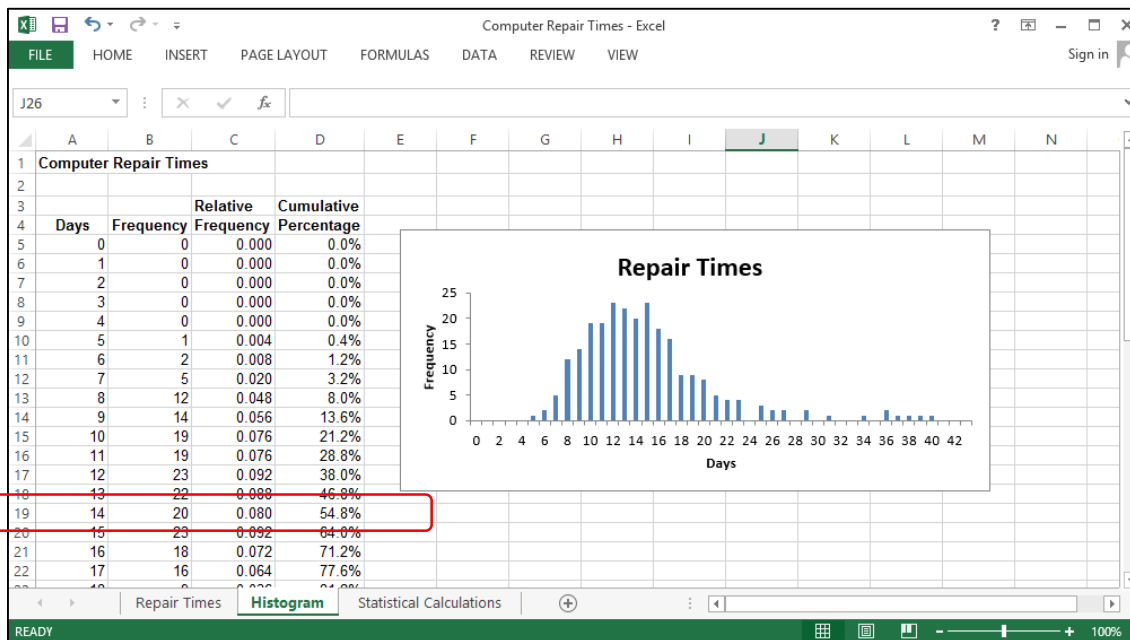
• Example #04: Computing the Midrange using *Purchase Orders.xlsx*

Supplier	Order No.	Item No.	Item Description	Item Cost	Quantity	Cost per order	A/P Terms (Months)	Order Date	Arrival Date
Manley Valve	C1111	9955	Door Decal	\$ 0.55	125	\$ 68.75	30	11/05/11	11/10/11
Manley Valve	A9876	9955	Door Decal	\$ 0.55	150	\$ 82.50	30	11/01/11	11/06/11
Hulkey Fasteners	C8989	9966	Hatch Decal	\$ 0.75	500	\$ 375.00	30	08/25/11	08/31/11
Hulkey Fasteners	D3232	1122	Airframe fasteners	\$ 4.25	17,000	\$ 72,250.00	30	10/11/11	10/19/11
Hulkey Fasteners	D2121	1122	Airframe fasteners	\$ 4.25	17,500	\$ 74,375.00	30	10/25/11	11/03/11
Hulkey Fasteners	C3232	1122	Airframe fasteners	\$ 4.25	18,000	\$ 76,500.00	30	10/01/11	10/08/11
Steelpin Inc.	C0467	8008	Machined Valve	\$ 645.00	120	\$ 77,400.00	30	10/28/11	11/04/11
Manley Valve	C3333	8148	Machined Valve	\$ 655.50	125	\$ 81,937.50	30	10/10/11	10/17/11
Hulkey Fasteners	C1212	1122	Airframe fasteners	\$ 4.25	19,500	\$ 82,875.00	30	08/05/11	08/13/11
Steelpin Inc.	B3222	8008	Machined Valve	\$ 645.00	150	\$ 96,750.00	30	10/15/11	10/26/11
Alum Sheeting	A0446	5417	Control Panel	\$ 255.00	406	\$ 103,530.00	30	09/01/11	09/10/11
Durable Products	A1456	5454	Control Panel	\$ 220.00	500	\$ 110,000.00	45	10/15/11	10/20/11
Durable Products	A1344	5454	Control Panel	\$ 220.00	550	\$ 121,000.00	45	10/09/11	10/14/11
Alum Sheeting	A0433	5417	Control Panel	\$ 255.00	500	\$ 127,500.00	30	10/20/11	10/27/11
MIN						\$ 68.75			
MAX						\$ 127,500.00			
MIDRANGE						63784.375			

```
>>> import os
>>> os.getcwd()
>>> path = os.chdir('C:\\Users\\MyWindows10\\Desktop\\lecture02\\datasets\\Data_Files')
>>> os.getcwd()
>>> filename = 'Purchase Orders.xlsx'
>>> sheet = pd.read_excel(filename, sheetname='Data', skiprows=2)
>>> sheet.columns
Index(['Supplier ', 'Order No.', 'Item No.', 'Item Description', 'Item Cost',
      'Quantity', 'Cost per order', 'A/P Terms (Months)', 'Order Date',
      'Arrival Date'],
      dtype='object')
min1 = min(sheet['Cost per order'])
max1 = max(sheet['Cost per order'])
min2 = sheet['Cost per order'].min()
max2 = sheet['Cost per order'].max()
midrange = (max1 + min1)/2
if min1 == min2 and max1 == max2:
    print(midrange)
```

Case Study: Computer Repair Time

- **Business Case:** A national electronic retailer want to decide a duration guarantee on computer repair time (*Computer Repair Times.xlsx*).



- Based on a sample data, the mean, median, and mode of the repair times are 14.9, 14, and 15 days, respectively.
- Look at the data in the cumulative frequency column on the next slide, half of the computer repairs are finished after 14 days – therefore, the retailer should NOT quote a 14-day guarantee.
- The longest repair time is 6 weeks – too long
- When looking at the distribution, 90% of the repair time are finished within 3 weeks – reasonable

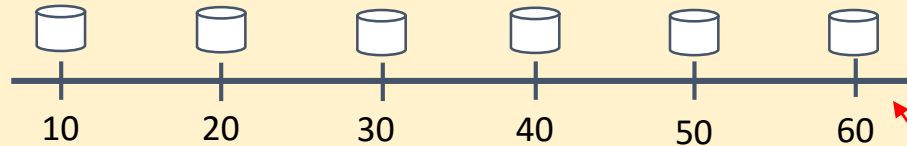
Case Study: Paint Durable Testing

- Business Case:** To examine how long each brand of paint between brand A and brand B will last longer before fading by using 6 gallons of each paint for testing.

Brand A	Brand B
10	35
60	45
50	30
30	35
40	40
20	25

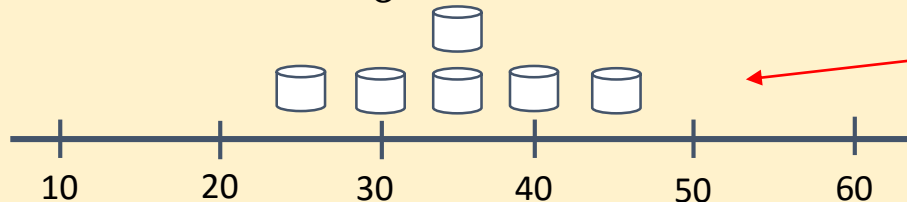
Unit: months

$$\mu_A = \frac{(10 + 60 + 50 + 30 + 40 + 20)}{6} = 35 \text{ months}$$



When using only the mean values, they are the same, 35 months.

$$\mu_B = \frac{(35 + 45 + 30 + 35 + 40 + 25)}{6} = 35 \text{ months}$$

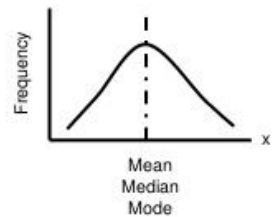


However, when looking at the variation of data values, brand B is more consistent than brand A.

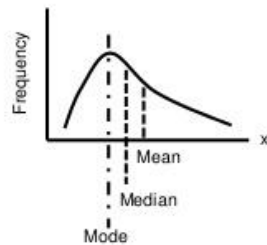
Distribution Shapes

- Data values can be distributed in many different shapes.

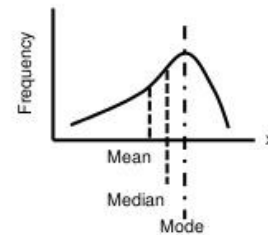
The shape of the frequency distribution



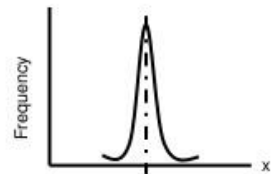
(a) Symmetrical shape



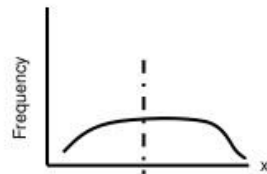
(b) Skewed to the right
(positively skewed)



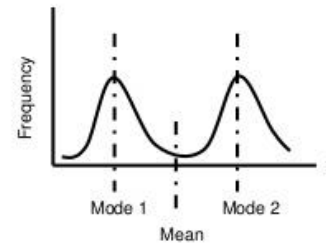
(c) Skewed to the left
(negatively skewed)



(d) Steep Shape



(e) Flat Shape



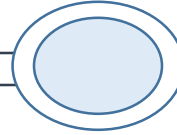
(f) Bimodal or Multimodal

Common Shapes of Frequency Distribution

53

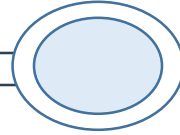
The three most important
shapes of frequency distribution

Distribution Shapes



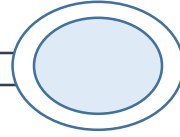
- **Positively-skewed** distribution or **right-skewed** distribution
 - The majority of the data values fall to the left of the mean, or **the tail is to the right side.**
 - The mean is to the right of the median, and the mode is to the left of the median.
 - Example: the exam scores when most students did poorly.
- **Symmetric** distribution
 - The data values are evenly distributed on both sides of the mean.
 - When the data is unimodal, the mean, median, and mode are the same at the center of the distribution.
 - Example: IQ scores and heights of adult males.
- **Negatively-skewed** distribution or **left-skewed** distribution
 - The majority of the data values fall to the right of the mean, or ***the tail is to the left side.***
 - The mean is to the left of the median, and the mode is to the right of the median.
 - Example: The exam scores when most students scores very high.

Measures of Dispersion



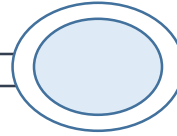
- **Measures of dispersion** – The degree of variation in the data or the numerical spread (or compactness) of the data
- Seven statistical measures
 - Range
 - Interquartile Range
 - Variance
 - Standard Deviation
 - Chebyshev's Theorem and the Empirical Rules
 - Standardized Values
 - Coefficient of Variation

Range

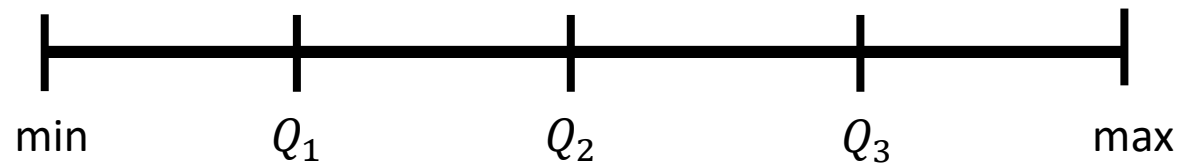


- Range – the difference between the maximum value and the minimum value in the data set.
 - Range = maximum value – minimum value
- Issues
 - Can be affected by outliers – often only used for a very small data set.

Interquartile Range

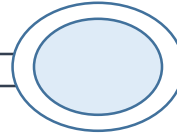


- Interquartile Range (IQR) – the difference between the first and the third quartiles, $Q_3 - Q_1$.
 - Also known as “Midspread”
 - The sorted data set is divided into four parts, and only the middle 50% of the data are used for computation.



- IQR is NOT affected by extreme values or outliers

Interquartile Range: Example

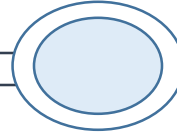


- Computing the Interquartile Range in the 'Repair Time (Days)' column *Computer Repair Times.xlsx* data file using Python **pandas**

Sample	Repair Time (Days)
1	18
2	15
3	17
4	9
5	37
6	15
7	8
8	29
9	10
10	14
11	17
12	12
13	13
14	12

```
>>> import os
>>> import pandas as pd
>>> os.chdir('C:\\Users\\MyWindows10\\Desktop\\lecture02\\datasets\\Data_Files')
>>> filename1 = 'Computer Repair Times.xlsx'
>>> sheet1 = pd.read_excel(filename, sheetname='Data', skiprows=2)
>>> sheet1.columns
Index(['Sample', 'Repair Time (Days)'], dtype='object')
>>> q1, q3 = sheet1['Repair Time (Days)'].quantile([0.25, 0.75])
>>> iqr = q3 - q1
>>> print(iqr)
6
```

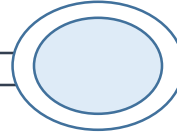
Interquartile Range: Example



- Computing the Interquartile Range in the ‘Cost per order’ column of *Purchase Orders.xlsx* data file using Python **numpy**

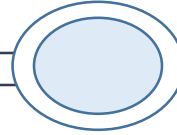
```
>>> import os
>>> import numpy as np
>>> os.chdir('C:\\Users\\MyWindows10\\Desktop\\lecture02\\datasets\\Data_Files')
>>> filename = 'Purchase Orders.xlsx'
>>> sheet = pd.read_excel(filename, sheetname='Data', skiprows=2)
>>> q1, q3 = np.percentile(sheet['Cost per order'], [25, 75])
>>> iqr = q3 - q1
>>> print(q1)
6757.8125
>>> print(q3)
27593.75
>>> print(iqr)
20835.9375
```

Variance



- **Variance** – the average of the squared deviations of the observations from the mean
 - The computation involves all the data.
- Two types of variance
 - The variance of a population
$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$
 - The variance of a sample
$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$
- **Notes:**
 - The larger the variance, the more the data are spread out from the mean, and the more variability can be expected in the observations.
 - An interpretation of a variance value may be difficult for practical business applications because the dimension of the variance is expressed as the square of the observations, according to the formula

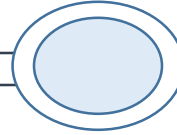
Variance: Example



- Computing the Variance in 'Cost per order' column using *Purchase Orders.xlsx*.

```
>>> import os
>>> import numpy as np
>>> os.chdir('C:\\Users\\MyWindows10\\Desktop\\lecture02\\datasets\\Data_Files')
>>> filename = 'Purchase Orders.xlsx'
>>> sheet = pd.read_excel(filename, sheetname='Data', skiprows=2)
>>> population_variance = np.var(sheet['Cost per order'], ddof=0)
>>> sample_variance = np.var(sheet['Cost per order'], ddof=1)
>>> print(population_variance)
881120163.4646332
>>> print(sample_variance)
890594573.824468
```

Standard Deviation



- Standard Deviation – the square root of the variance

- Two types of standard deviation

- Standard deviation of a population

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

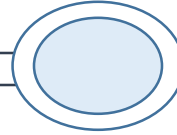
- Standard deviation of a sample

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

- Notes:

- The Standard Deviation is easier to interpret than the Variance
 - Its units of measure are the same as those of the data – see the formula.

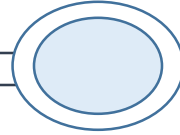
Standard Deviation: Example



- Computing the Standard Deviation in the 'Cost per order' column using *Purchase Orders.xlsx*.

```
>>> import os
>>> import numpy as np
>>> os.chdir('C:\\Users\\MyWindows10\\Desktop\\lecture02\\datasets\\Data_Files')
>>> filename = 'Purchase Orders.xlsx'
>>> sheet = pd.read_excel(filename, sheetname='Data', skiprows=2)
>>> population_std = np.std(sheet['Cost per order'], ddof=0)
>>> sample_std = np.std(sheet['Cost per order'], ddof=1)
>>> print(population_std)
29683.668295287112
>>> print(sample_std)
29842.83119652805
```

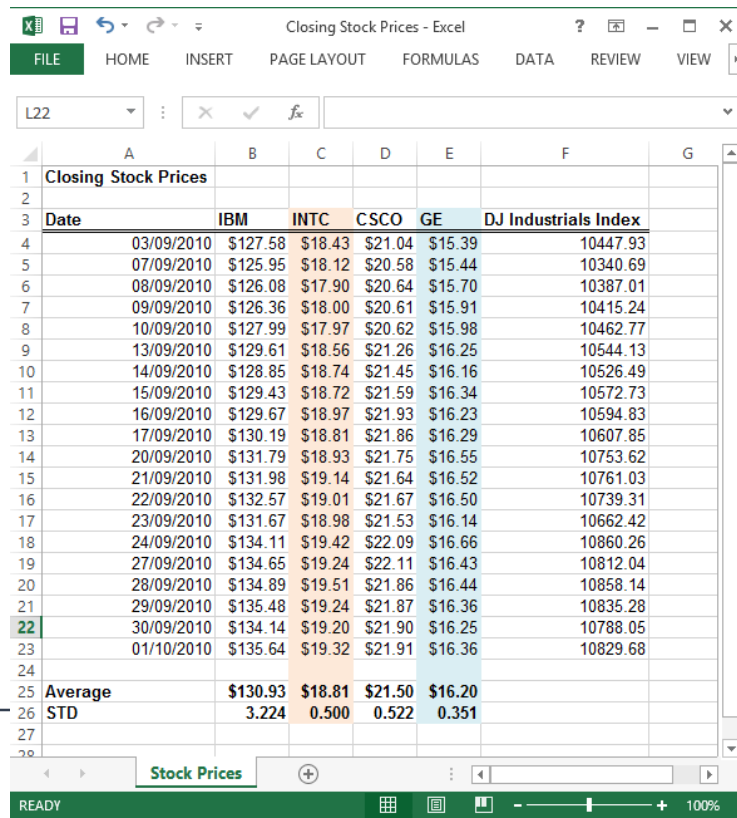
Standard Deviation: Case Study



- **Business Case**: Using standard Deviation (SD) to measure risk in stock prices.
 - Associates standard deviation with variations in stock prices
 - Measures the tendency of a fund's monthly returns that vary from the long-term average.
 - The higher the SD is, the higher the risk.
 - Two mutual funds
 - Mutual fund #01: Averaged return is 11% with SD of 10%
 - Mutual fund #02: Averaged return is 14% with SD of 20%
 - A larger the standard deviation
 - While a greater potential of a higher return exists, there is also greater risk of realizing a lower return.
 - Mutual fund #02 is riskier than mutual fund #01

Standard Deviation: Case Study

- Business Case: Evaluate whether INTC or GE has a lower risk based on standard deviation, *Closing Stock Prices.xlsx*



Date	IBM	INTC	CSCO	GE	DJ Industrials Index
03/09/2010	\$127.58	\$18.43	\$21.04	\$15.39	10447.93
07/09/2010	\$125.95	\$18.12	\$20.58	\$15.44	10340.69
08/09/2010	\$126.08	\$17.90	\$20.64	\$15.70	10387.01
09/09/2010	\$126.36	\$18.00	\$20.61	\$15.91	10415.24
10/09/2010	\$127.99	\$17.97	\$20.62	\$15.98	10462.77
13/09/2010	\$129.61	\$18.56	\$21.26	\$16.25	10544.13
14/09/2010	\$128.85	\$18.74	\$21.45	\$16.16	10526.49
15/09/2010	\$129.43	\$18.72	\$21.59	\$16.34	10572.73
16/09/2010	\$129.67	\$18.97	\$21.93	\$16.23	10594.83
17/09/2010	\$130.19	\$18.81	\$21.86	\$16.29	10607.85
20/09/2010	\$131.79	\$18.93	\$21.75	\$16.55	10753.62
21/09/2010	\$131.98	\$19.14	\$21.64	\$16.52	10761.03
22/09/2010	\$132.57	\$19.01	\$21.67	\$16.50	10739.31
23/09/2010	\$131.67	\$18.98	\$21.53	\$16.14	10662.42
24/09/2010	\$134.11	\$19.42	\$22.09	\$16.66	10860.26
27/09/2010	\$134.65	\$19.24	\$22.11	\$16.43	10812.04
28/09/2010	\$134.89	\$19.51	\$21.86	\$16.44	10858.14
29/09/2010	\$135.48	\$19.24	\$21.87	\$16.36	10835.28
30/09/2010	\$134.14	\$19.20	\$21.90	\$16.25	10788.05
01/10/2010	\$135.64	\$19.32	\$21.91	\$16.36	10829.68
Average	\$130.93	\$18.81	\$21.50	\$16.20	
STD	3.224	0.500	0.522	0.351	

Average and SD of the daily closing stock price

INTC:

MEAN = \$18.81

SD = \$0.50

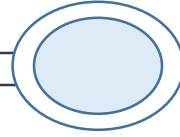
GE:

MEAN = \$16.19

SD = \$0.35

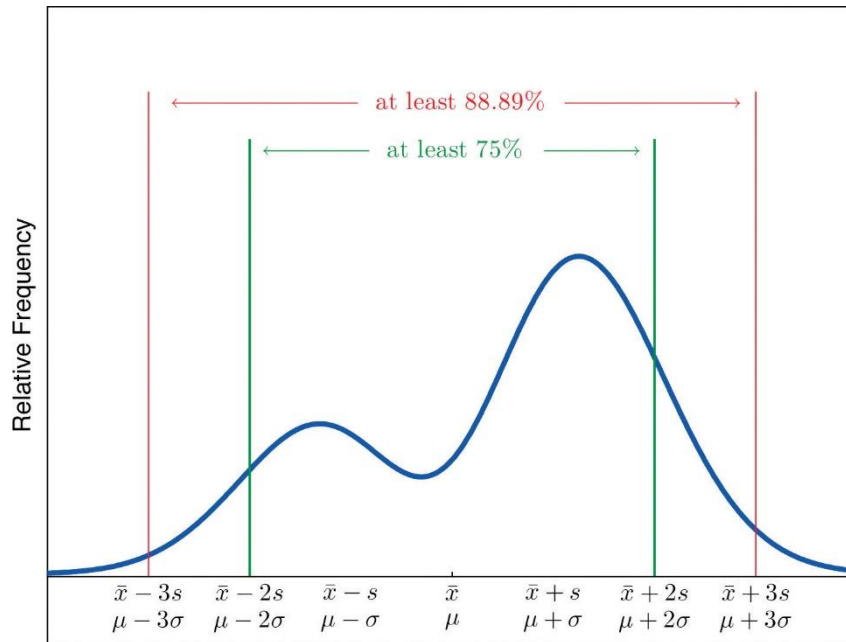
- The variability (SD) of GE closing stock prices is less than that of INTC.
- GE has a lower risk than INTC.

Chebyshev's Theorem



- Chebyshev's Theorem – For any numerical data set,
 - There are at least $1 - \frac{1}{k^2}$ of the data lie within k standard deviations of the mean.
 - That is, the proportion of data is in the interval with end-points $\bar{x} \pm ks$ for samples and $\mu \pm k\sigma$ for population, where $k > 1$

Chebyshev's Theorem



$$k = 2,$$

$$1 - \frac{1}{2^2} = \frac{3}{4} = 75\%$$

$$k = 3,$$

$$1 - \frac{1}{3^2} = \frac{8}{9} = 89\%$$

Chebyshev's Theorem: Example

- Applying Chebyshev's Theorem using Excel file, *Purchase Orders.xlsx*

	A	B	C	D	E	F	G
	Supplier	Order No.	Item No.	Item Description	Item Cost	Quantity	Cost per order
4	Alum Sheeting	A0223	4224	Bolt-nut package	\$ 3.95	4,500	\$ 17,775.00
5	Alum Sheeting	A0433	5417	Control Panel	\$ 255.00	500	\$ 127,500.00
6	Alum Sheeting	A0443	1243	Airframe fasteners	\$ 4.25	10,000	\$ 42,500.00
7	Alum Sheeting	A0446	5417	Control Panel	\$ 255.00	406	\$ 103,530.00
8	Alum Sheeting	B0247	1243	Airframe fasteners	\$ 4.25	9,000	\$ 38,250.00
9	Alum Sheeting	B0447	5634	Side Panel	\$ 185.00	150	\$ 27,750.00
10	Alum Sheeting	B0479	5634	Side Panel	\$ 185.00	140	\$ 25,900.00
11	Alum Sheeting	B0567	1243	Airframe fasteners	\$ 4.25	10,500	\$ 44,625.00
12	Durable Products	A1234	9399	Gasket	\$ 3.65	1,250	\$ 4,562.50
13	Durable Products	A1235	9399	Gasket	\$ 3.65	1,450	\$ 5,292.50
14	Durable Products	A1344	5454	Control Panel	\$ 220.00	550	\$ 121,000.00
15	Durable Products	A1345	9399	Gasket	\$ 3.65	1,470	\$ 5,365.50
16	Durable Products	A1346	9399	Gasket	\$ 3.65	1,985	\$ 7,245.25
17	Durable Products	A1456	5454	Control Panel	\$ 220.00	500	\$ 110,000.00
18	Durable Products	A1457	4569	Bolt-nut package	\$ 3.50	3,900	\$ 13,650.00
19	Durable Products	A1567	1369	Airframe fasteners	\$ 4.20	15,000	\$ 63,000.00
20	Durable Products	B1234	7258	Pressure Gauge	\$ 90.00	100	\$ 9,000.00
21	Durable Products	B1345	7258	Pressure Gauge	\$ 90.00	120	\$ 10,800.00
22	Durable Products	B1468	1369	Airframe fasteners	\$ 4.20	14,000	\$ 58,800.00
23	Durable Products	B1589	5275	Shielded Cable/ft.	\$ 1.00	25,000	\$ 25,000.00
24	Durable Products	B1666	1369	Airframe fasteners	\$ 4.20	10,000	\$ 42,000.00
25	Fast-Tie Aerospace	B2333	6321	O-Ring	\$ 2.45	1,300	\$ 3,185.00
26	Fast-Tie Aerospace	B2345	6321	O-Ring	\$ 2.45	1,200	\$ 2,940.00
27	Fast-Tie Aerospace	B2356	6321	O-Ring	\$ 2.45	2,500	\$ 6,125.00
28	Fast-Tie Aerospace	B2367	6321	O-Ring	\$ 2.45	1,250	\$ 3,062.50
29	Fast-Tie Aerospace	B2378	6321	O-Ring	\$ 2.45	1,500	\$ 3,675.00
30	Fast-Tie Aerospace	B2498	5689	Side Panel	\$ 175.00	150	\$ 26,250.00

In Cost per Order column,

$$n = 94$$

$$\bar{x} = 26,295.32$$

$$SD = 29,842.8312$$

For two standard deviation ($k = 2$), The Interval

$$\text{Beginning interval} = \bar{x} - ks$$

$$= 26,295.32 - 2 * 29,842.8312$$

$$= -33,390.3424$$

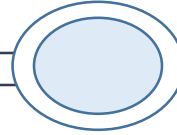
$$\text{Ending interval} = \bar{x} + ks$$

$$= 26,295.32 + 2 * 29,842.8312$$

$$= 85,980.9824$$

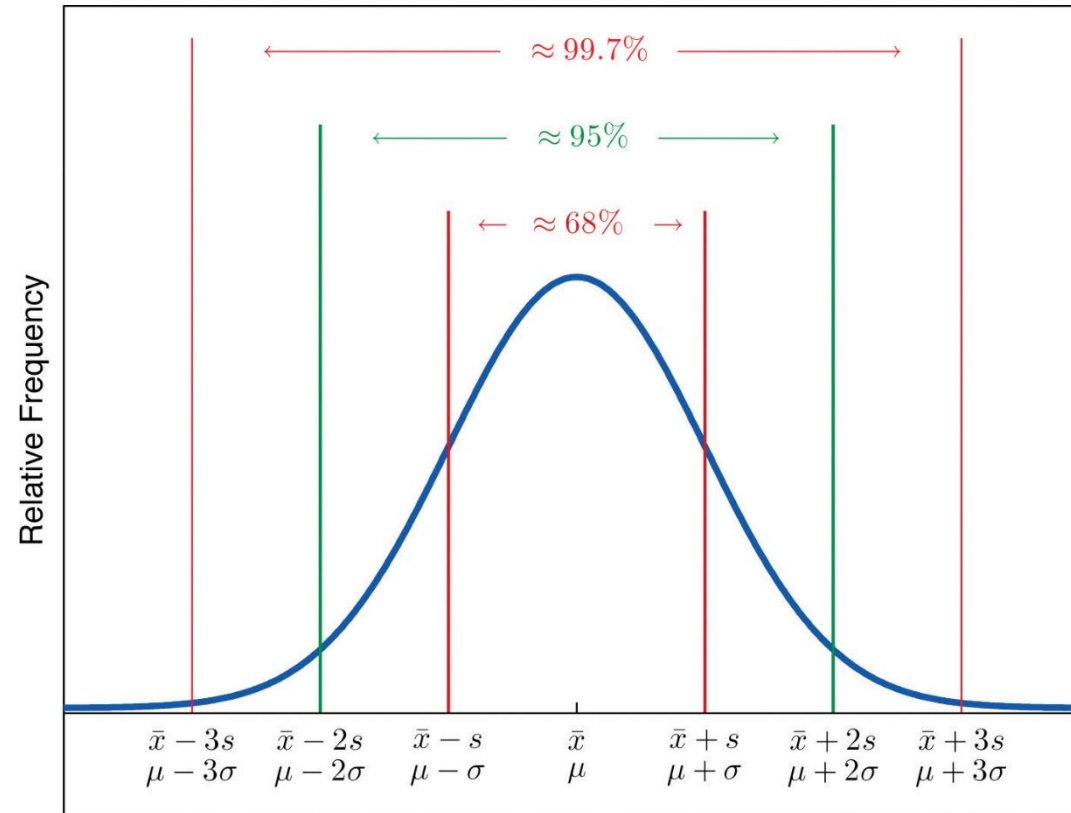
- Look at data in Excel file, sort data based on column G, "Cost per order".
- There are data from row #4 to row #92 or 89 rows out of 94 rows or 94.68% of data lie within this interval. (at least 75% according to the theorem).

Empirical Rules

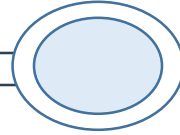


- Empirical Rules – when data is normally-distributed, the percentage of data are generally much higher than the value that Chebyshev's theorem specifies.
 - Rule #1: Approximately 68% of the observations will fall within one standard deviation of the mean, or between $\bar{x} \pm 1s$
 - Rule #2: Approximately 95% of the observations will fall within two standard deviations of the mean, or between $\bar{x} \pm 2s$
 - Rule #3: Approximately 99.7% of the observations will fall within three standard deviations of the mean, or between $\bar{x} \pm 3s$
- Note:
 - Chebyshev's theorem is applied to any type of data distributions, but Empirical rules are only applied to the normal distribution.

Empirical Rules

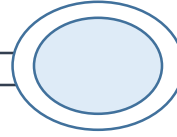


Empirical Rules: Case Study



- **Business Case**: A retailer knows that on average, an order is delivered by standard ground transportation in 8 days ($\bar{x} = 8$) with a standard deviation of 1 day ($s = 1$)
 - Using the empirical rule #2, the retailer can tell a customer with 95% confidence that their package should arrive within 6 to 10 days.
 - Rule #2, $k = 2 \rightarrow 8 \pm 1 * 2 = (6, 10)$

Empirical Rules: Case Study

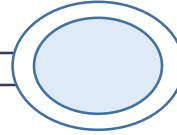


- **Business Case:** Process Capability Index (C_p) – measures how well a manufacturing process can achieve the specifications.

$$C_p = \frac{\text{upper specification} - \text{lower specification}}{\text{total variation}}$$

- **Example:**
 - A specification for a dimension of manufacturing parts is $5 \pm 0.2 \text{ cm}$ or between 4.80 and 5.20. In other words, a part with dimension outside this range would be classified as defective.
 - To measure the manufacturing process, usually, each manufactured part will be measured its dimension, compute the total variation using the third empirical rule (Rule #3), and then compare the result to specification by dividing the specification range by the total variation.

Empirical Rules: Case Study



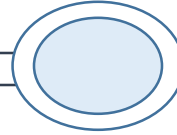
- **Business Case:** Applying Empirical Rules to Measure the Capability of a Manufacturing Process

- Excel data file – **Manufacturing Measurements.xlsx**
- Computation:
 - Suppose that the part specification is 5.00 ± 0.2 centimeters
 - The specification variation = $5.2 - 4.8 = 0.70$
 - Compute mean and standard deviation for the whole data
 - Mean = 4.99
 - Standard Deviation = 0.117
 - Calculate total variation of manufactured parts using the Rule #3 or $\bar{x} \pm 3s$
 - $\bar{x} - 3s = 4.99 - 3 * 0.117 = 4.64$
 - $\bar{x} + 3s = 4.99 + 3 * 0.117 = 5.34$
 - Total variation = $5.34 - 4.64 = 0.70$
 - Compute C_p
 - $C_p = \frac{5.20 - 4.80}{0.70} = 0.57$

```
>>> import os
>>> import pandas as pd
>>> os.chdir('C:\\Users\\MyWindows10\\Desktop\\lecture02\\datasets\\Data_Files')
>>> filename = 'Manufacturing Measurements.xlsx'
>>> df = pd.read_excel(filename, sheetname='Data', skiprows=2, header=None)
>>> dataset = pd.concat(df[0]+ df[1]+ df[2]+ df[3]+ df[4]+ df[5]+ df[6]+ df[7])
>>> mean = dataset.mean()
>>> std = dataset.std(ddof=1) #sample std
```

$C_p < 1.0$ is NOT good because it means that the variation in the process is wider than the specification limits. In practice, many manufacturers want to have C_p values of at least 1.5

Empirical Rules: Case Study

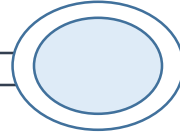


- Further analysis,
 - From the data, there are 3 parts which are below 4.8 (lower limit) and other 5 parts exceed 5.2 (upper limit). The total number of parts is 200. Therefore, $(3+5) = 8$ out of 200 or 4% are defective, and 96% are acceptable.

The screenshot shows an Excel spreadsheet with the following data and summary statistics:

	A	B	C	D	E	F	G	H	I	J	K
1	Manufacturing Measurements										
2											
3	5.21	5.87	4.85	4.95	5.07	4.96	4.96	5.11	Mean	4.99	
4	5.02	5.33	4.82	4.86	4.82	4.96	5.06	5.11	Standard deviation	0.117	
5	4.90	5.11	5.02	5.13	5.03	4.94	4.86	5.08			
6	5.00	5.07	4.90	4.95	4.85	5.19	4.96	5.03	Mean - 3*Stdev	4.640	
7	5.16	4.93	4.73	5.22	4.89	4.91	4.99	4.94	Mean + 3*Stdev	5.340	
8	5.03	4.99	5.04	4.81	4.82	5.01	4.94	4.88	Total variation	0.700	
9	4.96	5.04	5.07	4.91	5.18	4.93	5.06	4.91			
10	5.04	5.14	4.81	4.95	5.02	5.05	4.95	4.86	Lower Specification	4.8	
11	4.98	5.09	5.04	4.94	5.05	4.96	5.02	4.89	Upper Specification	5.2	
12	5.07	5.06	5.03	4.81	4.88	4.92	5.01	4.91	Specification range	0.4	
13	5.02	4.85	5.01	5.11	5.08	4.95	5.04	4.87			
14	5.08	4.93	5.14	4.81	4.98	5.08	5.01	4.93	Cp	0.57	
15	4.85	5.04	5.12	4.97	5.02	4.97	5.02	5.14			
16	4.90	5.09	4.89	5.07	4.99	5.04	5.03	4.87			
17	4.97	5.07	4.91	5.03	5.02	4.94	5.18	4.98			
18	5.09	4.99	4.97	4.81	5.03	4.98	5.08	4.88			
19	4.89	5.01	4.98	4.95	5.02	5.03	5.14	4.88			
20	4.87	4.88	5.01	4.89	5.07	5.05	4.92	5.01			
21	5.01	4.93	5.01	5.08	4.95	4.91	4.97	4.93			
22	4.97	5.10	5.09	4.93	4.95	5.09	4.92	4.93			
23	4.76	4.94	4.93	4.99	4.94	5.21	5.14	4.99			
24	4.94	4.88	5.04	4.94	5.12	4.87	4.92	4.91			
25	4.92	4.89	5.11	5.13	5.08	5.02	5.03	4.96			
26	4.91	4.89	5.07	5.02	4.91	4.81	4.98	4.78			
27	4.96	5.02	5.13	5.13	4.92	4.98	4.89	4.88			

Standardized Values



- **Standardized Value (z-score)** – a relative measure of the distance of an observation from the mean, independent of the units of measurement.

- Computation:

- Subtract the sample mean from the i th observation, x_i
- Divide the result by the sample standard deviation.

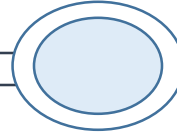
- The z-Score for the i th observation in a data set

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

- Interpretation of z-Score

- $z_i < 0 \rightarrow x_i$ lies to the left of the mean
- $z_i > 0 \rightarrow x_i$ lies to the right of the mean

Standardized Values

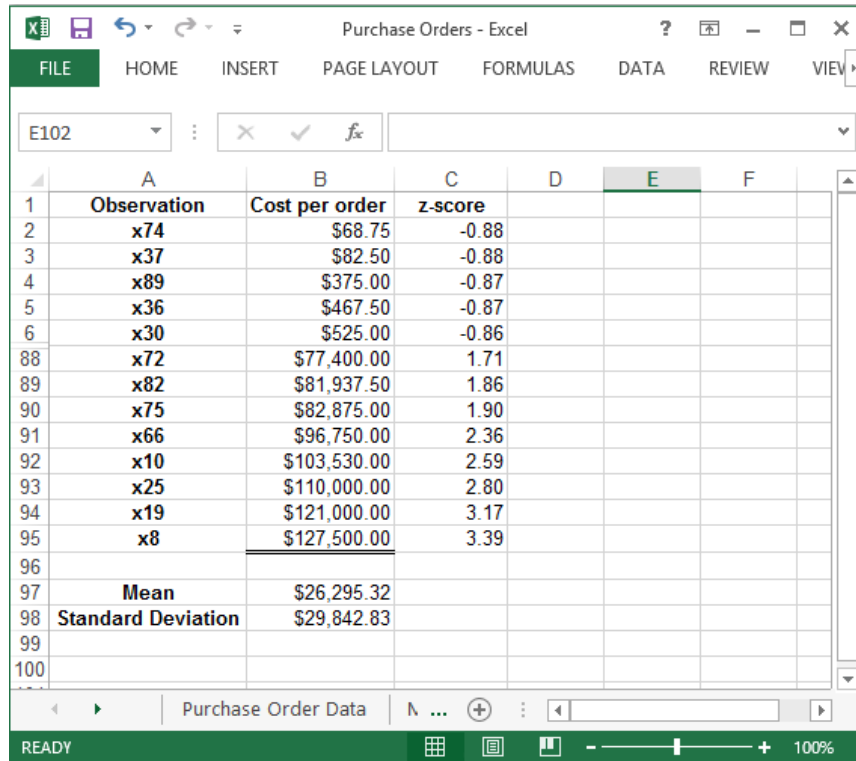


- What does the **Z**-score means?
 - Dividing the distance from the mean by the standard deviation is like scaling the distance to express it in units of standard deviation
 - Like normalizing the value.
 - Therefore,
 - **Z**-score = 1 → the observation is one standard deviation to the right of the mean.
 - **Z**-score = -1.5 → the observation is 1.5 standard deviation to the left of the mean
- **Benefit of Z-score** – It allows us to compare two data values in different data sets that have different means and standard deviations.
 - *Data values with the same Z-score means that the observations have the same relative distance from their respective means.*

When we look at the normal distribution graph, values on the x-axis such as $\bar{x} - 2s, \bar{x} - 1s, \bar{x} + 0s, \bar{x} + 1s, \bar{x} + 2s$ and so on. The number -2, -1, 0, 1, 2 are the z-Score values.

Standardized Values: Example

- Computing z-Score using *Purchase Orders.xlsx*

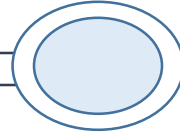


	A	B	C
1	Observation	Cost per order	z-score
2	x74	\$68.75	-0.88
3	x37	\$82.50	-0.88
4	x89	\$375.00	-0.87
5	x36	\$467.50	-0.87
6	x30	\$525.00	-0.86
88	x72	\$77,400.00	1.71
89	x82	\$81,937.50	1.86
90	x75	\$82,875.00	1.90
91	x66	\$96,750.00	2.36
92	x10	\$103,530.00	2.59
93	x25	\$110,000.00	2.80
94	x19	\$121,000.00	3.17
95	x8	\$127,500.00	3.39
96			
97	Mean	\$26,295.32	
98	Standard Deviation	\$29,842.83	
99			
100			

```
>>> import os
>>> import pandas as pd
>>> from scipy.stats import zscore
>>> os.chdir('C:\\Users\\MyWindows10\\Desktop\\lecture02\\datasets\\Data_Files')
>>> filename = 'Purchase Orders.xlsx'
>>> df = pd.read_excel(filename, sheetname='Data', skiprows=2, header=None)
# Compute z-score for all columns of data
>>> df.apply(zscore)

# Select only numeric column
>>> import numpy as np
>>> numeric_col = df.select_dtypes(include=[np.number]).columns
>>> df[numeric_col].apply(zscore)
```

Coefficient of Variation

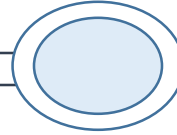


- **Coefficient of Variation (CV)** – provides a relative measure of the dispersion in data relative to the mean.

$$CV = \frac{\textit{standard deviation}}{\textit{mean}}$$

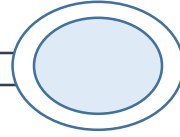
- Can be expressed in percentage by multiplying the value with 100.
- Useful for comparing the variability of two or more data sets when their scales (or mean values) differ.

Coefficient of Variation



- Standardized value vs. Coefficient of Variation
 - Standardized value (z-Score) tells a distance of a data value from its mean.
 - Its unit is the unit of measurement.
 - Compares the relative distances from the means of two data values in different data sets that have different means and different standard deviations – *which data value is farther or closer to its means than the other?*
 - Coefficient of Variation (CV) tells a variability of a data set.
 - It is unit-less
 - Compares the relative magnitude of the standard deviations of two or more different data sets – *which data set has higher or lower variations?*
 - Two standard deviations in different data sets cannot be compared with each other directly because their units of measurement may be different.

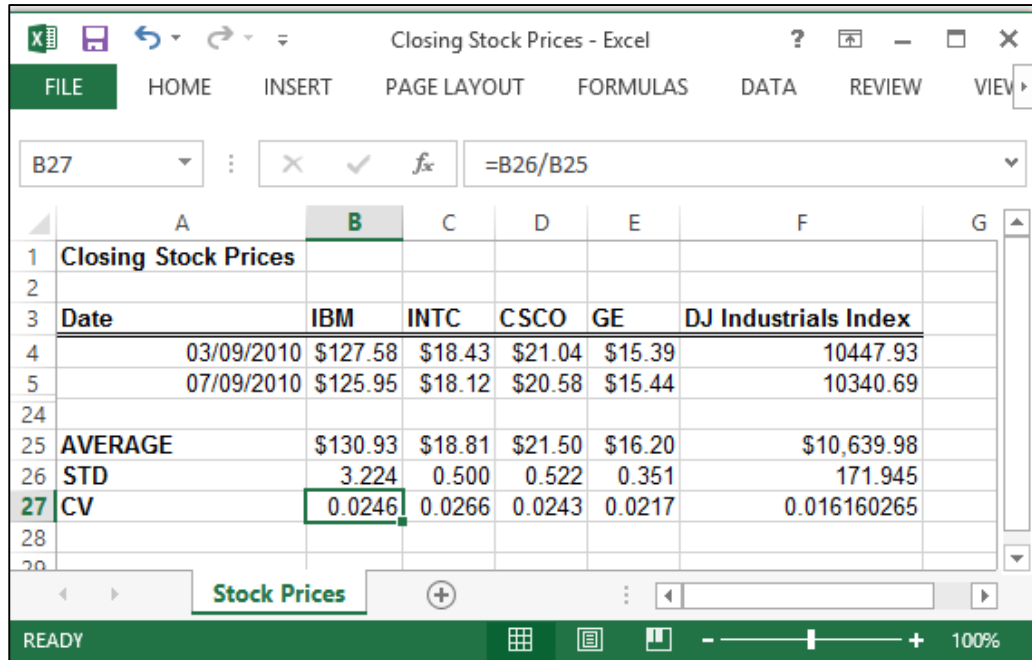
Coefficient of Variation



- **Business Example:** Give a relative measure of risk to return.
 - The smaller the CV, the smaller the relative risk is for the return provided.
- **Return-to-risk ratio:** the reciprocal of CV or $1/CV$
 - Easier to interpret than CV
 - The higher return-to-risk ratio, the better the fund is.

Coefficient of Variation: Example

- Applying the Coefficient of Variation using *Closing Stock Prices.xlsx*



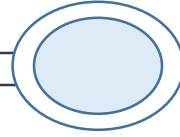
The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E	F	G
1	Closing Stock Prices						
2							
3	Date	IBM	INTC	CSCO	GE	DJ Industrials Index	
4	03/09/2010	\$127.58	\$18.43	\$21.04	\$15.39	10447.93	
5	07/09/2010	\$125.95	\$18.12	\$20.58	\$15.44	10340.69	
24							
25	AVERAGE	\$130.93	\$18.81	\$21.50	\$16.20	\$10,639.98	
26	STD	3.224	0.500	0.522	0.351	171.945	
27	CV	0.0246	0.0266	0.0243	0.0217	0.016160265	
28							
29							

The formula bar shows the formula for cell B27: $=B26/B25$.

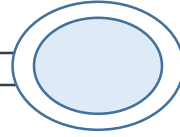
- Using only STD, IBM stock price is more risky than other stocks.
- Since the mean values of stocks are different (especially IBM's mean), using only STD does not provide meaningful information.
- Using CV, all stocks are not very different.
- DJ Industrials Index is less risky than other stocks.

Measures of Shape



- **Measures of shape** – describes the distribution of the data within a data set.
 - The distribution of data can be either symmetric (normal distribution) or asymmetric (skewed distribution)
 - Symmetric distribution
 - Two sides of the distribution (around the mean) are a mirror image of each other.
 - Normal distribution is the true symmetric distribution.
 - Asymmetric distribution or skewed distribution
 - Two sides of the distribution around the mean are NOT mirrored each other.

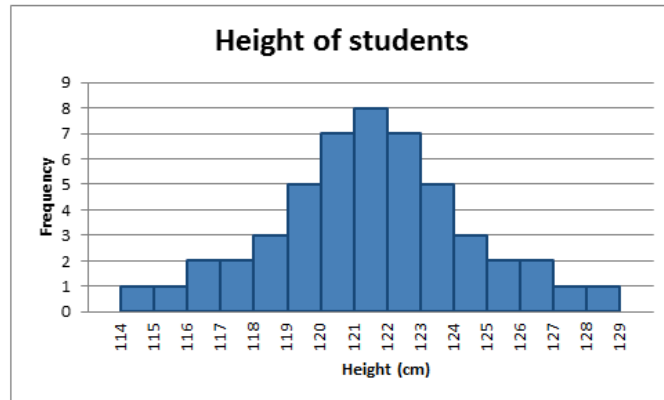
Measures of Shape



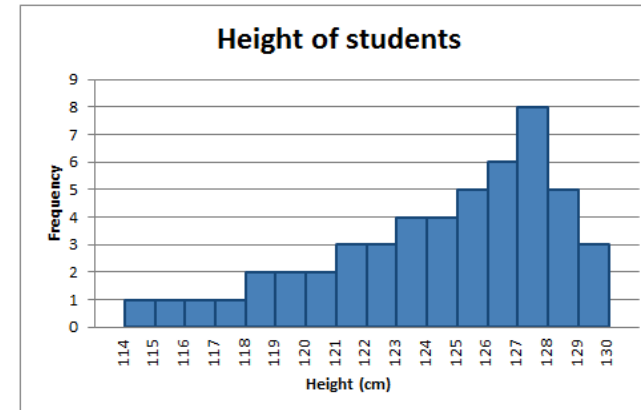
- Skewness – describes the lack of symmetry
 - The tendency for the data values to be concentrated (frequent) on one side and tailed off to the other side of the distribution of values.
- Two types of skewed distribution
 - Positively-skewed distribution
 - Concentrated on the left side, and tailed off to the right.
 - Negatively-skewed distribution
 - Concentrated on the right side, and tailed off to the left.

Measures of Shape

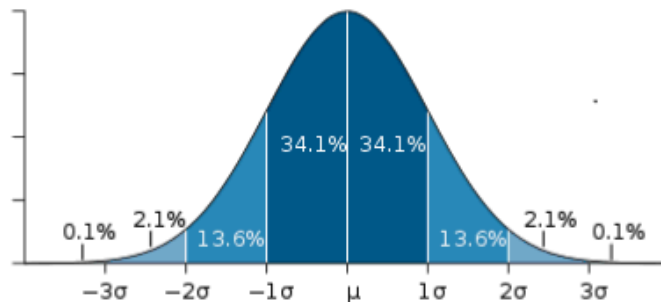
Symmetric Distribution



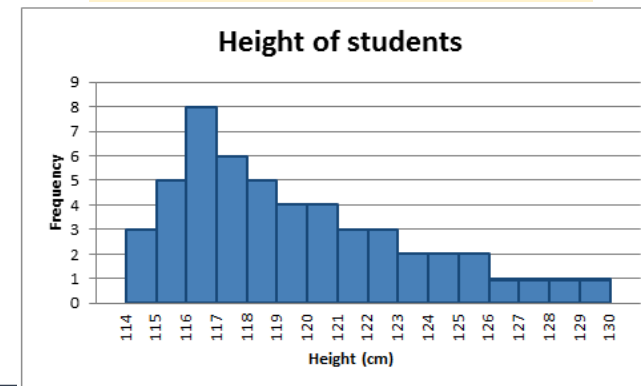
Negatively-Skewed Distribution



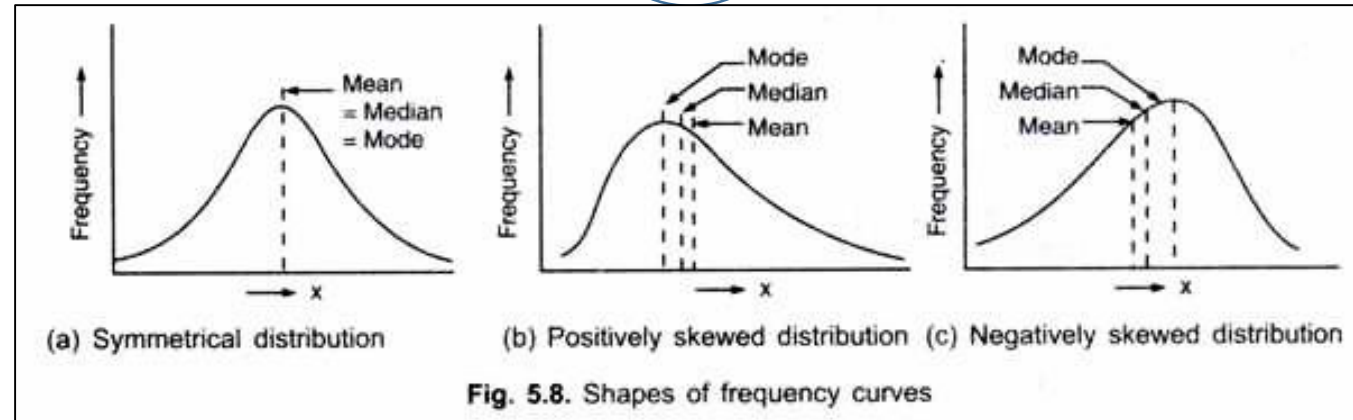
Normal Distribution



Positively-Skewed Distribution



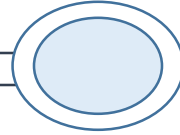
Measures of Shape



<http://cdn.yourarticlelibrary.com/wp-content/uploads/2015/06/image676.png>

- Two statistical measures of shape
 - The coefficient of skewness (CS) – measures the degree of asymmetry of observations around the mean.
 - The coefficient of kurtosis (CK) – measures the degree of kurtosis (peakness) of a population.

The Coefficient of Skewness (CS)



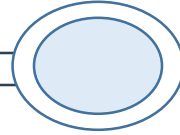
- The coefficient of Skewness (CS) for a population can be computed as:

$$CS = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^3}{\sigma^3}$$

- The coefficient of Skewness (CS) for a sample:

$$CS = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{s^3}$$

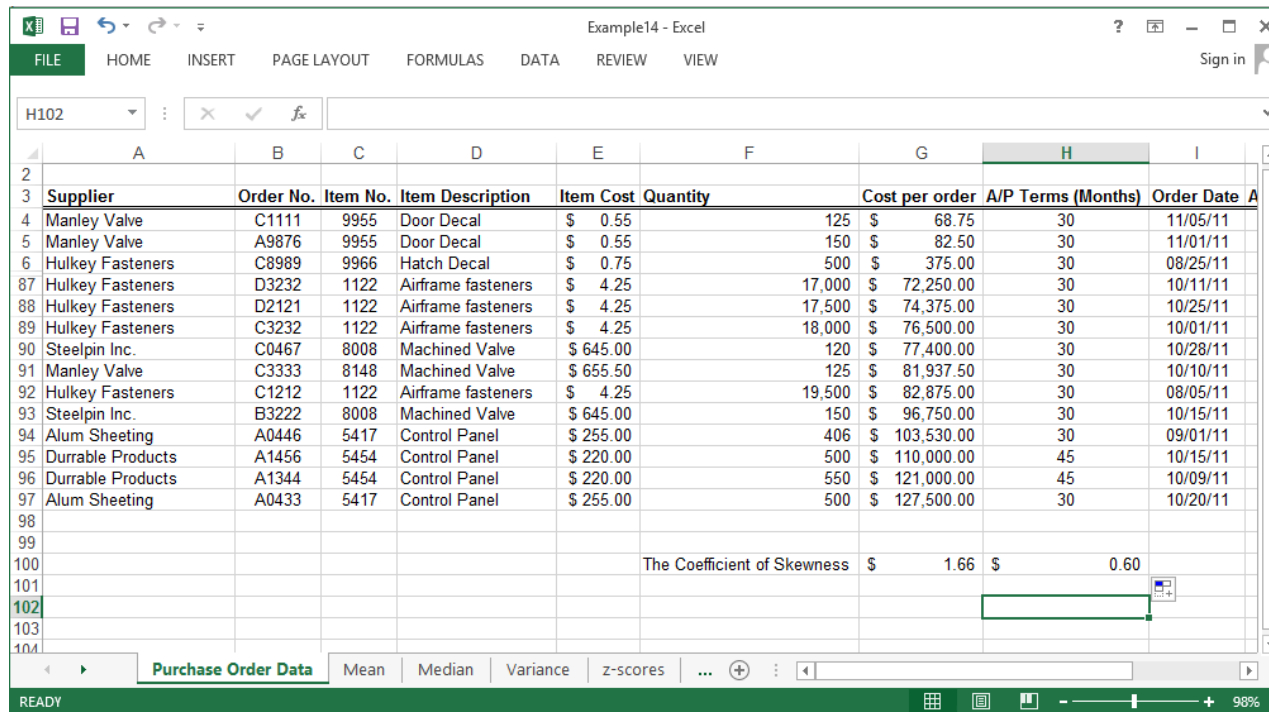
The Coefficient of Skewness (CS)



- What the value of the Coefficient of Skewness (CS) tells us:
 - $CS > 0$, the distribution of values is positively skewed.
 - $CS < 0$, the distribution of values is negatively skewed.
 - The closer a CS value is to zero, the less the degree of skewness.
 - $CS > 1$ or $CS < -1$, suggest a high degree of skewness.
 - $0.5 < CS < 1$ or $-1 < CS < -0.5$, suggest a moderate skewness
 - $-0.5 < CS < 0.5$, indicate relative symmetry.

The Coefficient of Skewness (CS): Example

- Measuring Skewness using *Purchase Orders.xlsx*



The screenshot shows an Excel spreadsheet with the following data:

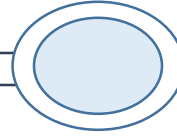
Supplier	Order No.	Item No.	Item Description	Item Cost	Quantity	Cost per order	A/P Terms (Months)	Order Date
Manley Valve	C1111	9955	Door Decal	\$ 0.55	125	\$ 68.75	30	11/05/11
Manley Valve	A9876	9955	Door Decal	\$ 0.55	150	\$ 82.50	30	11/01/11
Hulkey Fasteners	C8989	9966	Hatch Decal	\$ 0.75	500	\$ 375.00	30	08/25/11
Hulkey Fasteners	D3232	1122	Airframe fasteners	\$ 4.25	17,000	\$ 72,250.00	30	10/11/11
Hulkey Fasteners	D2121	1122	Airframe fasteners	\$ 4.25	17,500	\$ 74,375.00	30	10/25/11
Hulkey Fasteners	C3232	1122	Airframe fasteners	\$ 4.25	18,000	\$ 76,500.00	30	10/01/11
Steelpin Inc.	C0467	8008	Machined Valve	\$ 645.00	120	\$ 77,400.00	30	10/28/11
Manley Valve	C3333	8148	Machined Valve	\$ 655.50	125	\$ 81,937.50	30	10/10/11
Hulkey Fasteners	C1212	1122	Airframe fasteners	\$ 4.25	19,500	\$ 82,875.00	30	08/05/11
Steelpin Inc.	B3222	8008	Machined Valve	\$ 645.00	150	\$ 96,750.00	30	10/15/11
Alum Sheeting	A0446	5417	Control Panel	\$ 255.00	406	\$ 103,530.00	30	09/01/11
Durable Products	A1456	5454	Control Panel	\$ 220.00	500	\$ 110,000.00	45	10/15/11
Durable Products	A1344	5454	Control Panel	\$ 220.00	550	\$ 121,000.00	45	10/09/11
Alum Sheeting	A0433	5417	Control Panel	\$ 255.00	500	\$ 127,500.00	30	10/20/11
The Coefficient of Skewness						\$ 1.66	\$ 0.60	

CS of Cost per order = 1.66

CS of A/P Terms = 0.60

Both are positively skewed, and A/P Terms have a smaller skewness than Cost per order.

The Coefficient of Skewness (CS)



- The effects of skewness on the values of mode, median, and mean.

- **Case #01:** No skewness

- Mean = Median = Mode

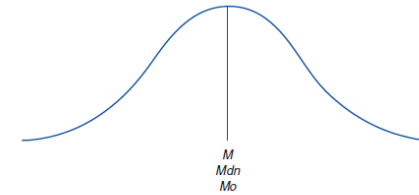
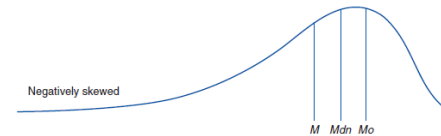


Figure 5.3 Position of Mean, Median, and Mode in Normally Distributed Data

- **Case #02:** Negatively-skewed

- Mean < Median < Mode



- **Case #03:** Positively-skewed

- Mode < Median < Mean

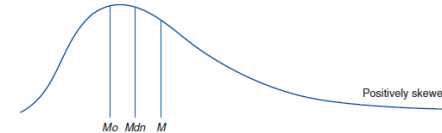
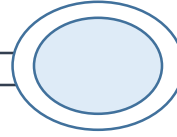


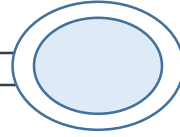
Figure 5.4 Position of Mode, Median, and Mean in Skewed Score Distributions

The Coefficient of Kurtosis (CK)



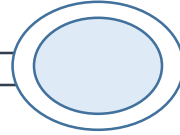
- **Kurtosis** – the peakedness (high, narrow) or flatness (short, flat-topped) of a histogram
 - The coefficient of kurtosis (CK) – measures the degree of kurtosis of the data set (population or sample)
 - CK for a population can be computed as:
 - $CK = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^4}{\sigma^4}$
 - CK for a sample can be calculated by:
 - $CK = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{s^4}$

The Coefficient of Kurtosis (CK)



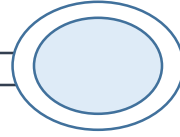
- Interpretation of the values of CK
 - The higher the values of CK is, the more peaked and less dispersion the distributions of data are
 - Distributions with values of $CK < 3$ are more flat with a wide degree of dispersion.
 - Distributions with values of $CK > 3$ are more peaked with less dispersion.

The Coefficient of Kurtosis (CK): Case Study



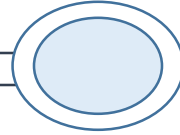
- **Business Example:** Using skewness and kurtosis can help give more information for evaluating risk than just using the standard deviation.
 - A distribution with both negatively-skewed and positively-skewed may have the same standard deviation.
 - However,
 - The negatively skewed (tail to the left) will have higher probabilities of larger returns (if the X-axis is corresponding to return rate).
 - The higher the kurtosis, the more area the histogram has in the tails rather than in the middle, which can indicate a greater potential for extreme and possibly catastrophic outcomes.

Measures of Association



- **Measures of Association** – measures the degree of association between two variables.
 - Two variables have a strong statistical relationship to each other if they appear to move together (direct or inverse).
 - A number of attendance at sports games are closely related to the win percentage of the team.
 - Ice cream sales have a strong relationship with temperature.
 - Scatter chart can be used to visually examine relationships between two variables.

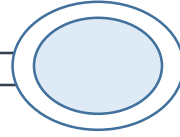
Measures of Association



- **Business Example** -- Understanding the relationships between company's internal factors and external measures is very important in making good business decisions.
 - Internal factors – product quality, employee training, and pricing factors.
 - External measures – profitability and customer satisfaction.
- If the relationship is a causal relationship and we can measure the strength of the relationship, we can use the information to make good business decisions.

Notes: Relationships between two variables can exist but they may NOT be a causal relationship such as CEOs' golf skills and their company performance. Therefore, we MUST be cautious in drawing inferences about casual relationship based on statistical relationship alone.

Covariance



- **Covariance** – a measure of the linear association between two variables, X and Y.

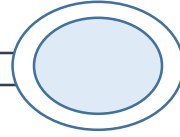
- Covariance of a population

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)}{N}$$

- Covariance of a sample

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

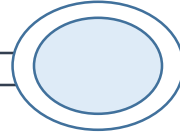
Covariance: Example



- Computing the Covariance between “Graduation %” and “Median SAT”
- using *Colleges and Universities.xlsx*.

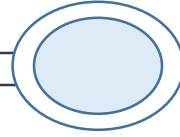
COVARIANCE.S = 263.3703231

Correlation



- **Correlation** – measures a linear relationship between two variables or how strongly two variables are related to each other.
 - Independent of the units of measurement.
 - Measured by the correlation coefficient known as “*Pearson product moment correlation coefficient*”

Correlation



- Compute the correlation between X and Y:

- Correlation coefficient for a population

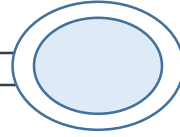
$$\rho_{xy} = \frac{cov(X, Y)}{\sigma_x \sigma_y}$$

- Correlation coefficient for a sample

$$r_{xy} = \frac{cov(X, Y)}{s_x s_y}$$

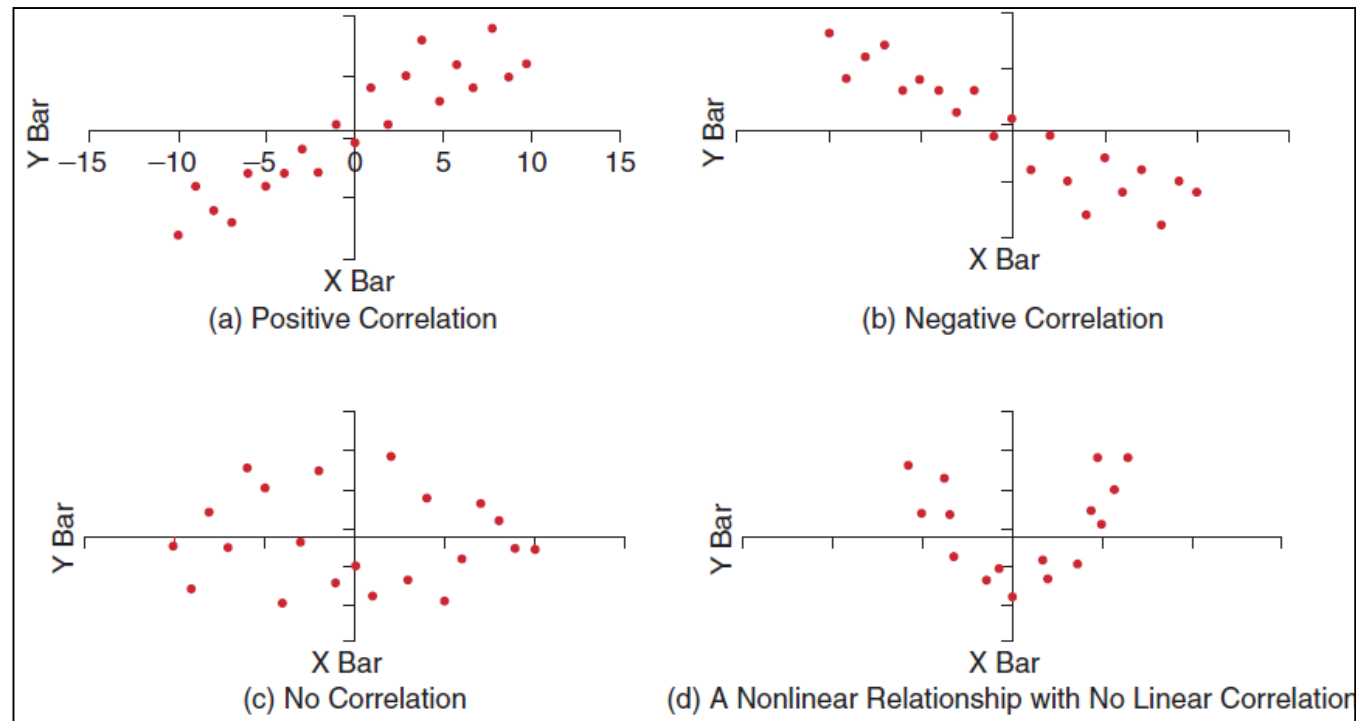
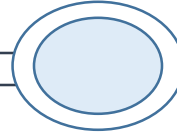
Dividing the covariance by the standard deviations is the way to scale the numerical value of the covariance to a number between -1 and 1

Correlation



- What the correlation value means:
 - Correlation = 0: there is no linear relationship between two variables.
 - Therefore, if one variable changes, we cannot reasonably predict what the other variable might do.
 - Correlation > 0: There is a direct linear relationship between two variables
 - While one variable increases (or decreases), the other variable also increases (or decreases).
 - Correlation < 0: There is an inverse direct relationship between two variables
 - While one variable increases, the other decreases.

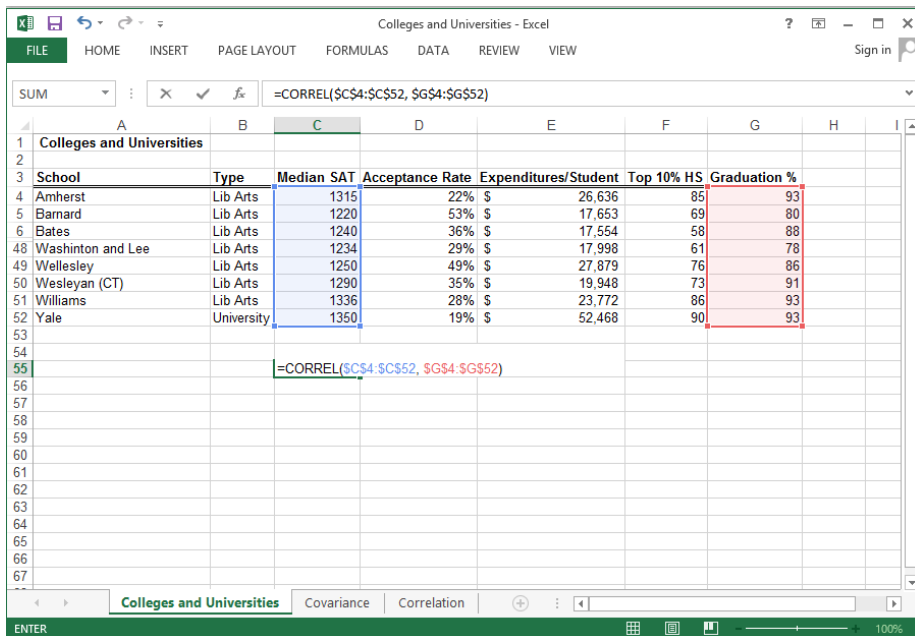
Correlation



The figures are from the textbook's *ppt* slides provided by Pearson representative.

Correlation: Example

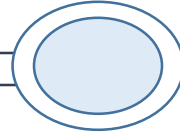
- Computing the Correlation Coefficient between “Graduation %” and “Median SAT” using *Colleges and Universities.xlsx*.



School	Type	Median SAT	Acceptance Rate	Expenditures/Student	Top 10% HS	Graduation %
Amherst	Lib Arts	1315	22%	\$ 26,636	85	93
Barnard	Lib Arts	1220	53%	\$ 17,653	69	80
Bates	Lib Arts	1240	36%	\$ 17,554	58	88
Washington and Lee	Lib Arts	1234	29%	\$ 17,998	61	78
Wellesley	Lib Arts	1250	49%	\$ 27,879	76	86
Wesleyan (CT)	Lib Arts	1290	35%	\$ 19,948	73	91
Williams	Lib Arts	1336	28%	\$ 23,772	86	93
Yale	University	1350	19%	\$ 52,468	90	93

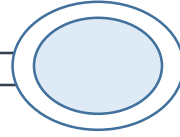
Correlation = 0.564146827

Outliers



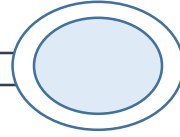
- **Outliers** – unusually large or small values in data.
 - Can causes some significant difference in the result from statistical analyses.
- Approaches to identifying outliers
 - **Manual inspection on data** for possible errors such as a misplaced decimal point.
 - **Histograms**
 - **Empirical rule and Z-score** – outliers are values that are more than three standard deviation from the mean.
 - **Interquartile range**
 - “Mild outlier” – between $1.5 \times \text{IQR}$ and $3 \times \text{IQR}$ to the left of $Q1$ or to the right of $Q3$
 - “Extreme outlier” -- more than $3 \times \text{IQR}$ from $Q1$ and $Q3$

Investigating Outliers: Example



- Inspecting outliers in the columns “Square Feet” and “Market Value” using *Home Market Value.xlsx*.

References



1. Allan G. Bluman, *Elementary Statistics: A Step by Step Approach*, Seventh Edition, McGraw-Hill International Edition, 2015.
2. Hector Cuesta, *Practical Data Analysis*, Packt Publishing Ltd., 2013.
3. James R. Evans, *Business Analytics*, Second Edition, Global Edition, Pearson Education Limited, 2017.
4. Sinan Ozdemir, *Principle of Data Science*, Packt Publishing, Ltd. 2016.