Topic 3: Numerical Descriptive Measures

Dr. Lei Pan

School of Accounting, Economics and Finance
Curtin University

Introduction

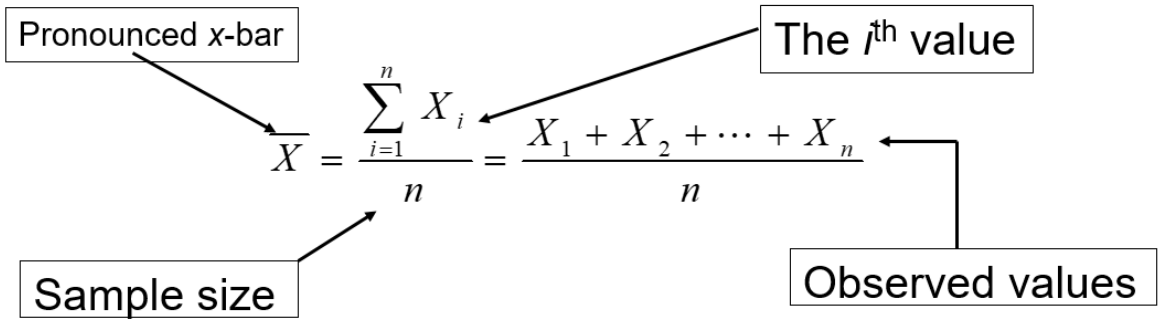
- This chapter discusses various numerical measures that can be used to summarise and describe numerical data.
- These numerical measures not only can be used to summarise a particular sample or population but will also enable the sample or population to be compared with others.
- These numerical measures are precise, objectively determined and easy to manipulate, interpret and compare.

Measures of Central Tendency

- Many data sets have a distinct **central tendency**, with the data values grouped or clustered around a central point.
- The three most important measures of central tendency are:
  - Mean
  - Median
  - Mode
- Each of these measures has advantages and disadvantages.

Mean (1 of 2)

- The arithmetic mean (often just called the "mean") is the most common measure of central tendency.
- To calculate the arithmetic mean, add all values of a variable in a data set and then divide the sum by the number of variable values in the data set.
- The **sample mean** is the mean calculated from sample data.

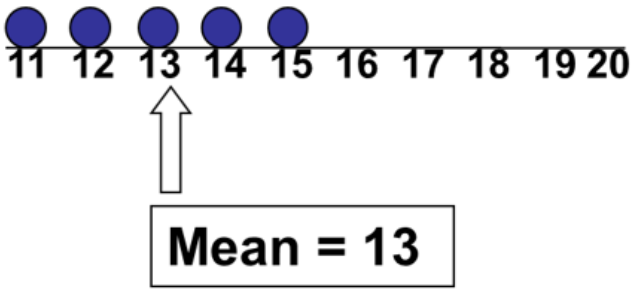


Curtin University

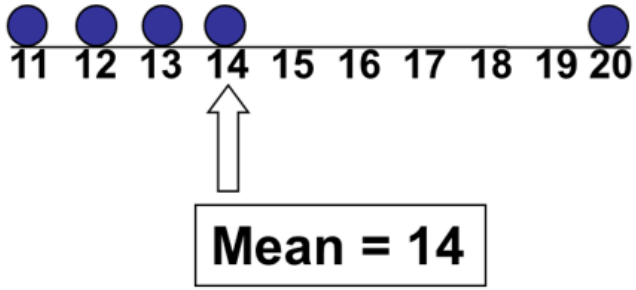
50 YEARS OF INNOVATION

Mean (2 of 2)

- Affected by extreme values (outliers).



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



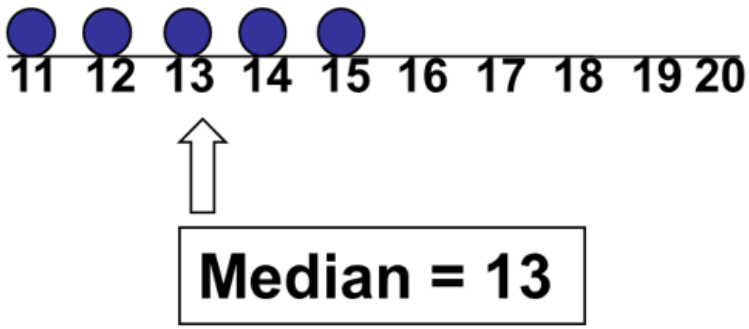
$$\frac{11+12+13+14+20}{5}=\frac{70}{5}=14$$

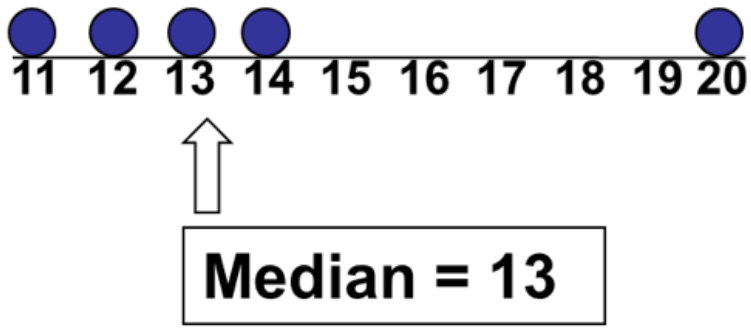
Median (1 of 2)

- The **median** is the middle value in a set of data that has been ordered from lowest to highest value.
- To calculate, first order the values from smallest to largest, and adhere to the following rules:
  - Rule 1: If the number of values in the data set is odd, the median is the middle ranked value
  - Rule 2: If the number of values in the data set is even, the median is the mean (average) of the two middle ranked values

Median (2 of 2)

- Less sensitive than the mean to extreme values



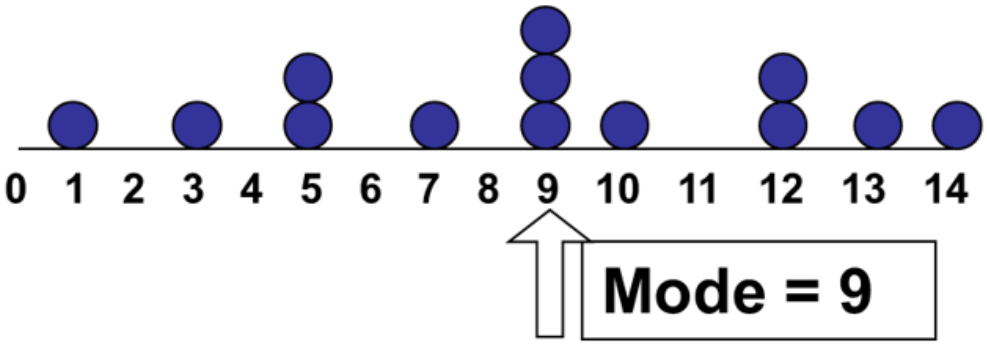


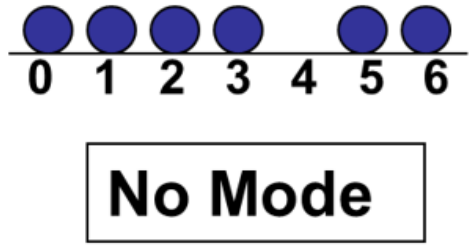
Mode (1 of 2)

- The **mode** is the value in a data set that appears most frequently.
- Not affected by extreme value.
- Used for either numerical or categorical data.

Mode (2 of 2)

- Unlike mean and median, there may be no unique (single) mode for a given data set.
- A set of data can also have no mode if no data value occurs more than once.





Measures of Central Tendency: Review Example

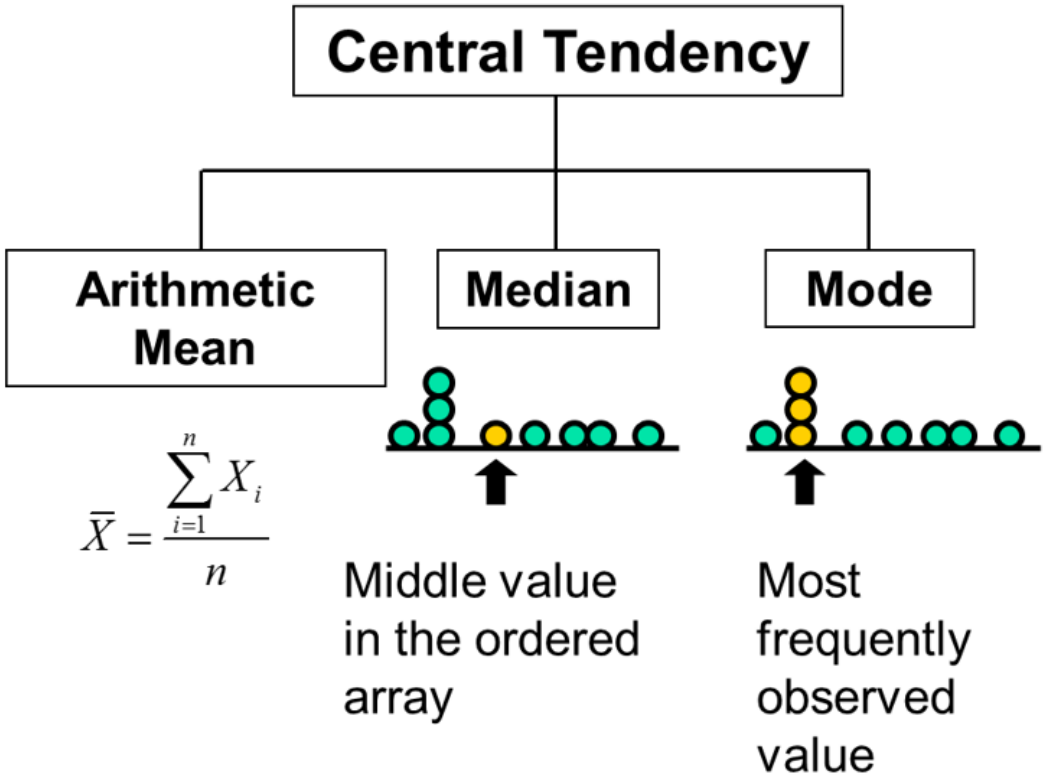
House Prices:
$2,000,000
$ 500,000
$ 300,000
$ 100,000
$ 100,000
Sum $3,000,000

- Mean: $\$3,000,000/5 = \$600,000$
- Median: middle value of ranked data = \$300,000
- Mode: most frequent value = \$100,000

Which Measure to Choose?

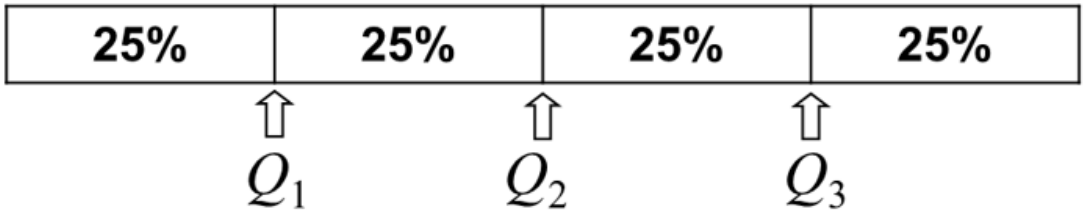
- The **mean** is generally used, unless extreme values (outliers) exist.
- The **median** is often used, since the median is not sensitive to extreme values.
  - e.g. median home prices may be reported for a region because it is less sensitive to outliers
- In many situations it makes sense to report both the **mean** and the **median**.

Summary


Central Tendency
Arithmetic Mean
Median
Mode
$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$
Middle value in the ordered array
Most frequently observed value

Quartiles (1 of 2)

- Quartiles split the ranked data into four segments, with an equal number of values per segment.



- The **first quartile**, Q_1, is the value for which 25% of the observations are smaller and 75% are larger.
- The **second quartile**, Q_2, is the same as the median (50% are smaller, 50% are larger).
- Only 25% of the observations are greater than the **third (or upper) quartile**, Q_3.

Quartiles (2 of 2)

- Similar to the median, the quartile can be found by determining the value in the appropriate position in the ranked data, where:
- First quartile position: $Q_1 = (n+1)/4$
- Second quartile position: $Q_2 = (n+1)/2$
- Third quartile position: $Q_3 = 3(n+1)/4$
- Note: n is the number of observed values (sample size)

Calculation Rules for Quartiles

- When calculating the ranked position use the following rules
  - If the result is a whole number then it is the ranked position to use.
  - If the result is a fractional half (e.g. 2.5, 7.5, 8.5, etc.) then average the two corresponding data values.
  - If the result is not a whole number or a fractional half then round the result to the nearest integer to find the ranked position.

You Do It

Find the first quartile of the following ordered array.

$$11, 12, 13, 16, 16, 17, 18, 21, 22$$

Geometric Mean and Geometric Mean Rate of Return

- The **geometric mean** is used to measure the average rate of change of a variable over n period of time.

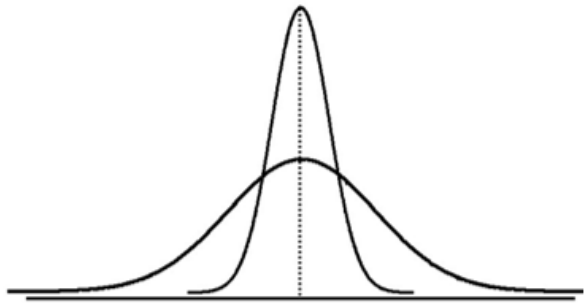
$$\bar{X}_G = (X_1 \times X_2 \times ... \times X_n)^{1/n}$$

- The **geometric rate of return** is used to measure the average return on an investment over time.

$$\bar{R}_G = [(1 + R_1) \times (1 + R_2) \times ... \times (1 + R_n)]^{1/n}$$

Measures of Variation

- Measures of variation give information on the **spread** or **variability** or **dispersion** of the data values.
- These measures include:
  - Range
  - Interquartile range
  - Variance
  - Standard deviation
  - Coefficient of variation

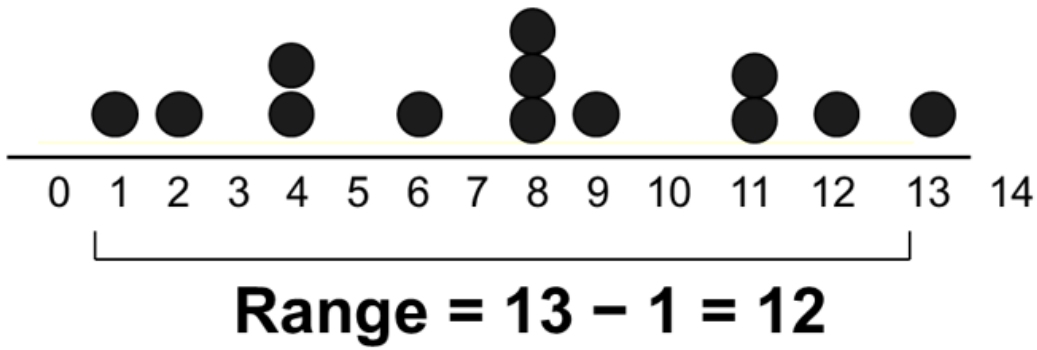


Same center, different variation

Range (1 of 2)

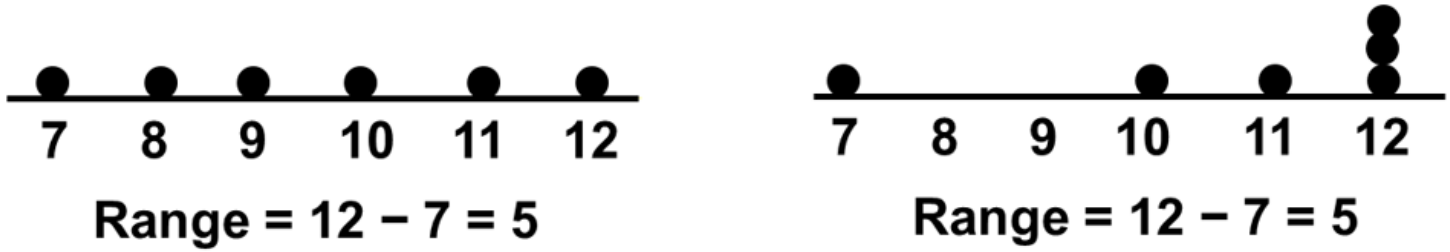
- The range is the simplest measure of variation.
- The range shows the difference between the largest and the smallest values in a set of data.

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



Range (2 of 2)

- Disadvantages include:
  - The range ignores the distribution of the data.



  - Like the mean, the range is sensitive to outliers.

$$\mathbf{1}, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 4, \mathbf{5}$$

Range = 5 − 1 = 4

$$\mathbf{1}, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 4, \mathbf{120}$$

Range = 120 − 1 = 119

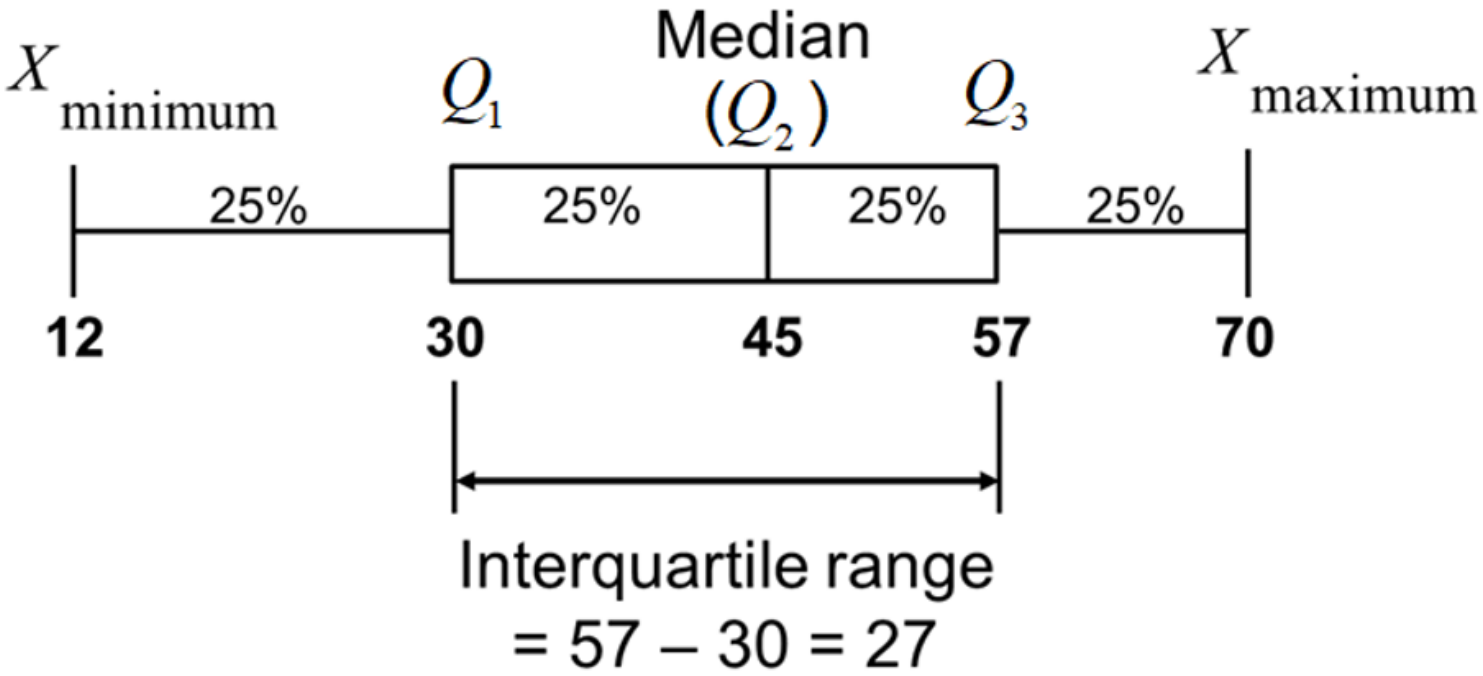
Interquartile Range (IQR)

- The **interquartile range** is the difference between the third and first quartiles in a set of data.

$$\text{IQR} = Q_3 - Q_1$$

- Like the median, Q_1 and Q_3, the IQR is a **resistant measure** (resistant to the presence of extreme values).
- The IQR is also called the midspread because it covers the middle 50% of the data.
- Use of the IQR eliminates outlier problems, as high- and low-valued observations are removed from calculations.

Calculating the Interquartile Range



Variance

- **Variance** measures average scatter around mean.
- Units are squared.

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

Where

$\overline{X}$ = arithmetic mean

n = sample size

X_i = i^{th} value of the variable X

You Do It

Which of the following is not a measure of location?

A. Mean
B. Median
C. Variance
D. Mode

Standard Deviation (1 of 2)

- Most commonly used measure of variation.
- Shows variation about the mean.
- Is the square root of the variance.
- Has the same units as the original data.

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

Standard Deviation (2 of 2)

Steps for Computing Standard Deviation

1. Compute the difference between each value and the mean.
2. Square each difference.
3. Add the squared differences.
4. Divide this total by $n - 1$ to get the sample variance.
5. Take the square root of the sample variance to get the sample standard deviation.

Calculation Example

Sample Data (X_i): **10 12 14 15 17 18 18 24**

$n = 8$ Mean $= \bar{X} = 16$

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

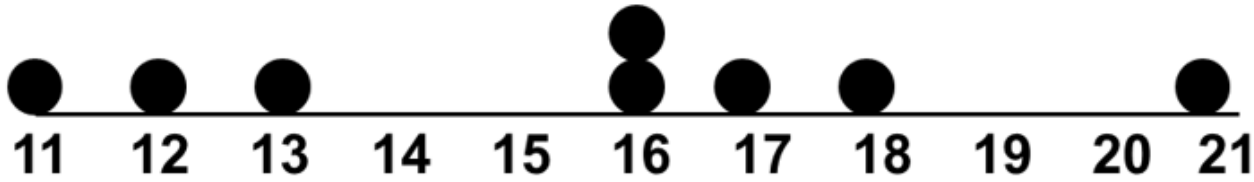
$$= \sqrt{\frac{(10-16)^2 + (12-16)^2 + (14-16)^2 + \cdots + (24-16)^2}{8-1}}$$

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

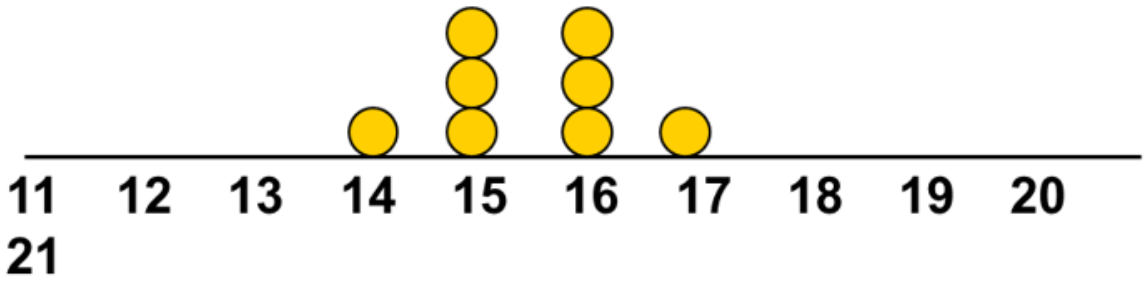
A measure of the "average" scatter around the mean

Comparing Standard Deviation (1 of 2)

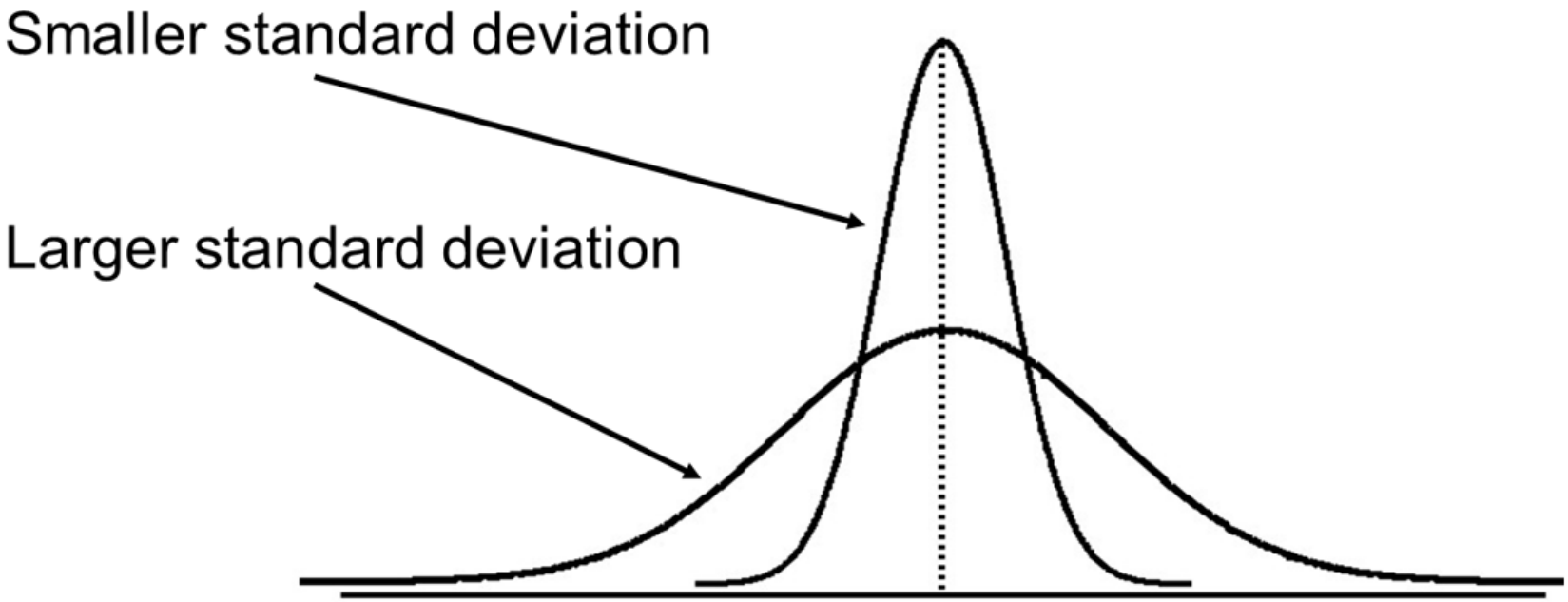
Data A (Mean = 15.5, S = 3.338)



Data B (Mean = 15.5, S = 0.926)



Comparing Standard Deviation (2 of 2)



Variance and Standard Deviation

- Advantages
  - Each value in the data set is used in the calculation.
  - Values far from the mean are given extra weight as deviations from the mean are squared.
- Disadvantages
  - Sensitive to extreme values (outliers).
  - Measures of absolute variation not relative variation.

Summary Characteristics

- The more the data are spread out, the greater the range, variance, and standard deviation.
- The more the data are concentrated, the smaller the range, variance, and standard deviation.
- If the values are all the same (no variation), all these measures will be zero.
- None of these measures are ever negative.

Coefficient of Variation

- The **coefficient of variation** measures relative variation.
- Shows variation relative to mean.
- It can be used to compare two or more sets of data measured in different units.
- It is always expressed as percentage (%).

$$\text{CV} = (\frac{S}{\bar{X}}) \cdot 100\%$$

Example 1

- Stock A:
  - Average price last year = \$50; standard deviation=\$5

$$\text{CV} = (\frac{S}{\bar{X}} \cdot 100\%) = \frac{\$5}{\$50} \cdot 100\% = 10\%$$

- Stock B:
  - Average price last year = \$100; standard deviation=\$5

$$\text{CV} = (\frac{S}{\bar{X}} \cdot 100\%) = \frac{\$5}{\$100} \cdot 100\% = 5\%$$

- Both stocks have the same standard deviation, but Stock B is less variable relative to its price.

Example 2

- Stock A:
  - Average price last year = \$50; standard deviation=\$5

$$\text{CV} = (\frac{S}{\bar{X}} \cdot 100\%) = \frac{\$5}{\$50} \cdot 100\% = 10\%$$

- Stock C:
  - Average price last year = \$8; standard deviation=\$2

$$\text{CV} = (\frac{S}{\bar{X}} \cdot 100\%) = \frac{\$2}{\$8} \cdot 100\% = 25\%$$

- Stock C has much smaller standard deviation but a much higher coefficient of variation.

Locating Extreme Outliers: Z-Score

- The Z score is the difference between a given observation and the mean, divided by the standard deviation.

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

- The Z-score is the number of standard deviations a data value is from the mean.
- A data value is considered an extreme outlier if its Z-score is less than -3.0 or greater than +3.0.
- The larger the absolute value of the Z-score, the farther the data value is from the mean.
- A negative Z score would indicate that a value is below the mean.

Example

- Suppose the mean math SAT score is 490, with a standard deviation of 100.
- Compute the Z-score for a test score of 620.

$$Z = \frac{X - \bar{X}}{S} = \frac{620 - 490}{100} = \frac{130}{100} = 1.3$$

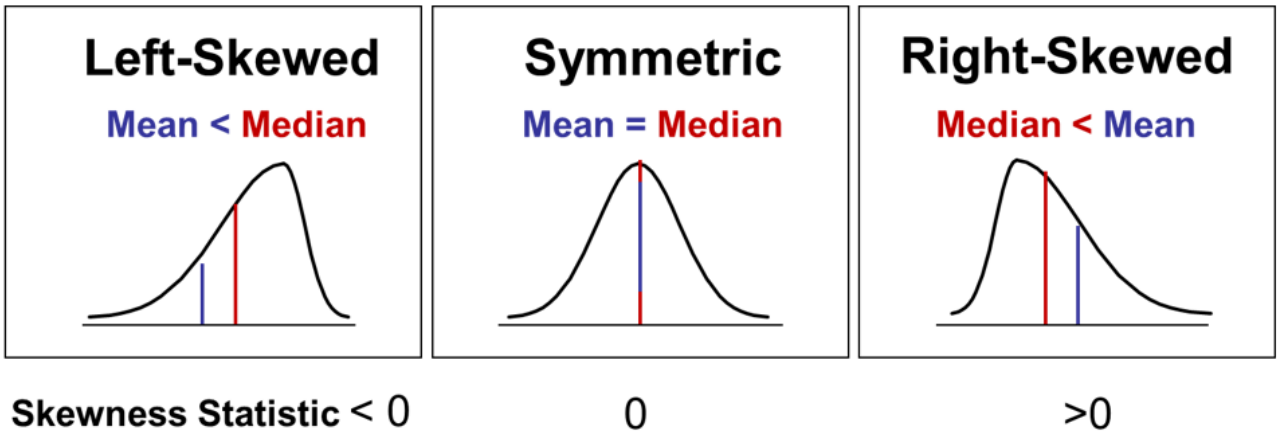
- A score of 620 is 1.3 standard deviations above the mean and would not be considered an outlier.

Shape of a Distribution

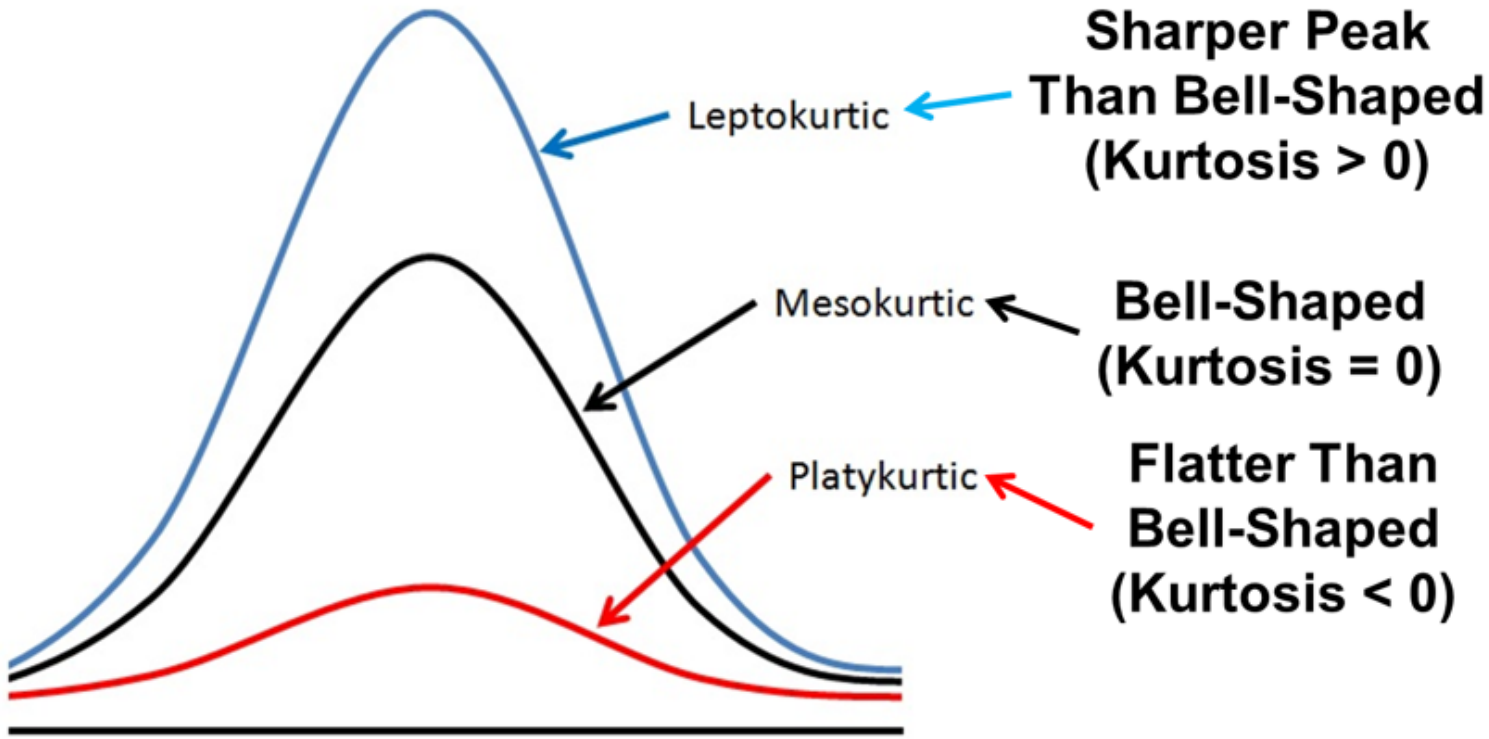
- Describes how data are distributed.
- Two useful shape related statistics are:
  - Skewness: measures the extent to which data values are not symmetrical
  - Kurtosis: measures the peakedness of the curve of the distribution, that is, how sharply the curve rises approaching the center of the distribution

Skewness

- A distribution is **symmetrical** if the lower and upper halves of the graph are mirror images of each other.
- If the distribution is not symmetrical, it may be **skewed**.
- Distribution can be **skewed to the right**, or **positively skewed**.
- Distribution can be **skewed to the left**, or **negatively skewed**.



Kurtosis


Leptokurtic
Sharper Peak
Than Bell-Shaped
(Kurtosis > 0)
Mesokurtic
Bell-Shaped
(Kurtosis = 0)
Platykurtic
Flatter Than
Bell-Shaped
(Kurtosis < 0)

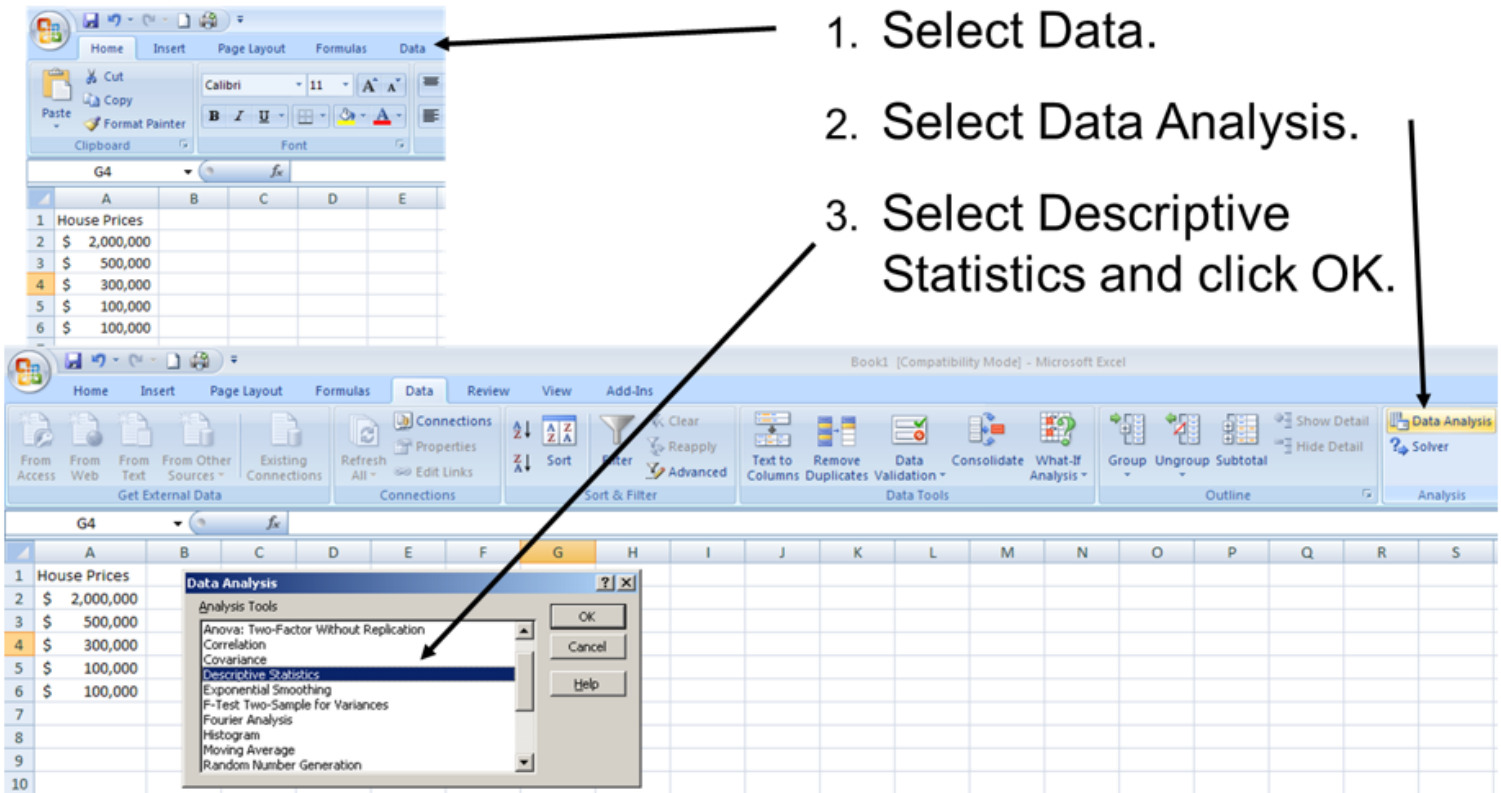
General Descriptive Stats Using Microsoft Excel

House Prices		Descriptive Statistics		
$ 2,000,000		Mean	$ 600,000	=AVERAGE(A2:A6)
$ 500,000		Standard Error	$ 357,770.88	=D6/SQRT(D14)
$ 300,000		Median	$ 300,000	=MEDIAN(A2:A6)
$ 100,000		Mode	$ 100,000.00	=MODE(A2:A6)
$ 100,000		Standard Deviation	$ 800,000	=STDEV(A2:A6)
		Sample Variance	640,000,000,000	=VAR(A2:A6)
		Kurtosis	4.1301	=KURT(A2:A6)
		Skewness	2.0068	=SKEW(A2:A6)
		Range	$ 1,900,000	=D12 - D11
		Minimum	$ 100,000	=MIN(A2:A6)
		Maximum	$ 2,000,000	=MAX(A2:A6)
		Sum	$ 3,000,000	=SUM(A2:A6)
		Count	5	=COUNT(A2:A6)

Curtin University

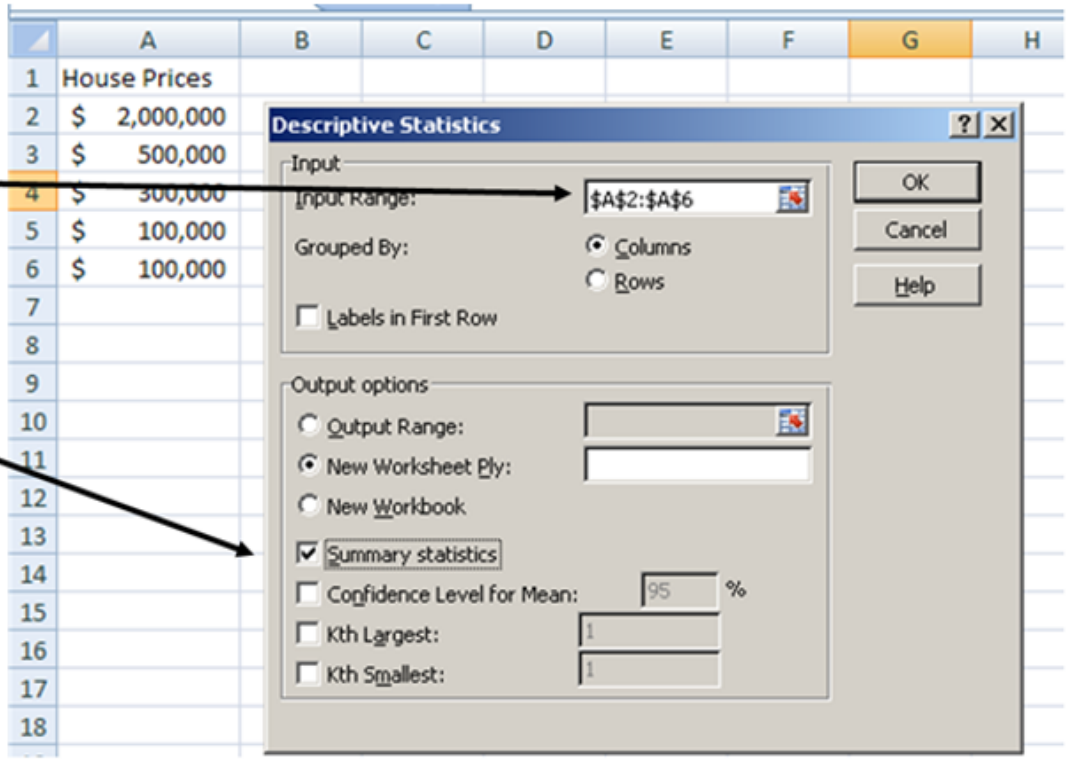
50 YEARS OF INNOVATION

General Descriptive Stats Using Data Analysis ToolPak (1 of 2)



General Descriptive Stats Using Data Analysis ToolPak (2 of 2)

4. Enter the cell range.
5. Check the Summary Statistics box.
6. Click OK



Excel Output

House Prices	
Mean	600000
Standard Error	357770.8764
Median	300000
Mode	100000
Standard Deviation	800000
Sample Variance	640,000,000,000
Kurtosis	4.1301
Skewness	2.0068
Range	1900000
Minimum	100000
Maximum	2000000
Sum	3000000
Count	5

Numerical Descriptive Measures for a Population

- Descriptive statistics discussed previously described a sample, not the population.
- Summary measures describing a population, called **parameters**, are denoted with Greek letters.
- Important population parameters are the population mean, variance, and standard deviation.

The mean μ

- The **population mean** is the sum of the values in the population divided by the population size, N.

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{X_1 + X_2 + \cdots + X_N}{N}$$

Where

μ = population mean

N = population size

$X_i = i^{\text{th}}$ value of the variable X

Population Variance

- **Population variance** is the average of the squared deviations of values from the mean.

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

Where

μ = population mean

N = population size

$X_i = i^{th}$ value of the variable X

Population Standard Deviation

- Most commonly used measure of variation
- Shows variation about the mean
- Is the square root of the population variance
- Has the same units as the original data

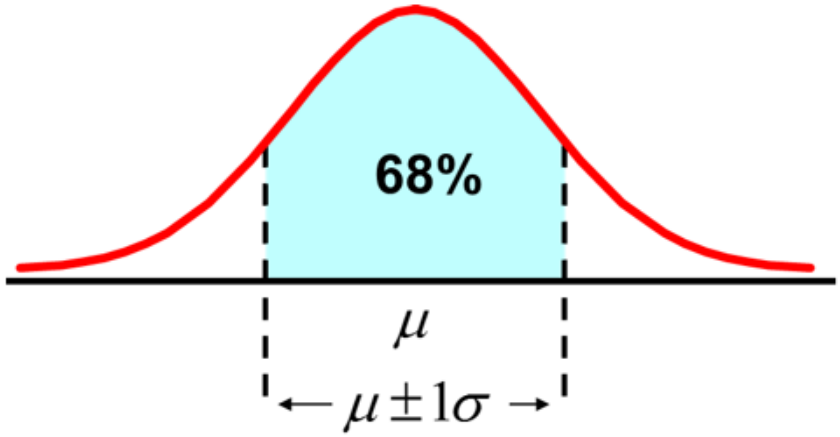
$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

Sample Statistics vs. Population Parameters

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

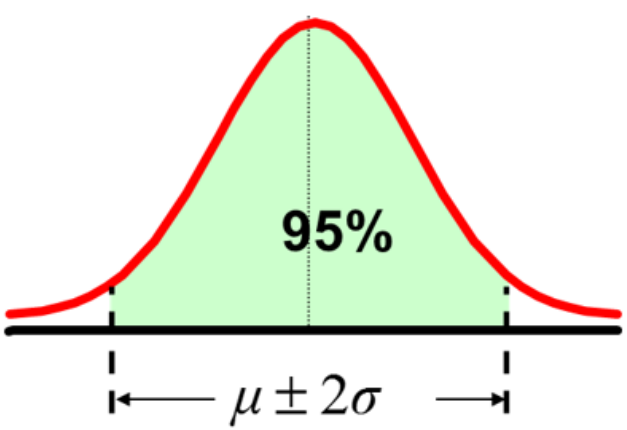
Empirical Rule (1 of 2)

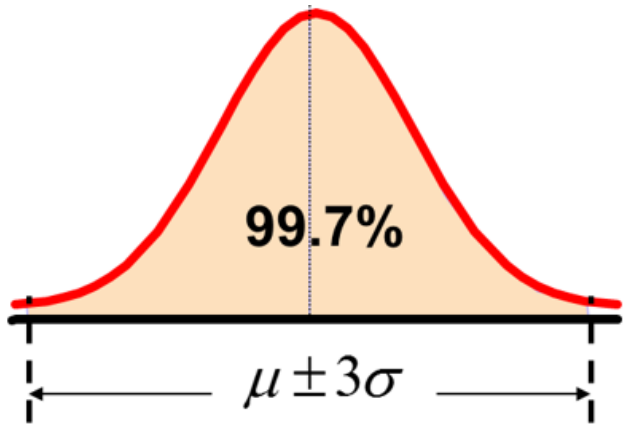
- The empirical rule approximates the variation of data in a bell-shaped distribution.
- If the data distribution is approximately bell-shaped, then the interval $\mu \pm 1\sigma$ contains about 68% of the values in the population.



Empirical Rule (2 of 2)

- The interval $\mu \pm 2\sigma$ contains about 95% of the values in the population.
- The interval $\mu \pm 3\sigma$ contains about 99.7% of the values in the population.





Using the Empirical Rule

- Suppose that the variable Math SAT scores is bell-shaped with a mean of 500 and a standard deviation of 90. Then,
  - Approximately 68% of all test takers scored between 410 and 590, (500 ± 90).
  - Approximately 95% of all test takers scored between 320 and 680, (500 ± 180).
  - Approximately 99.7% of all test takers scored between 230 and 770, (500 ± 270).

The Chebyshev Rule

- The **Chebyshev rule** gives lower bounds of the distribution of data values in terms of standard deviations from the mean for any distribution.
- It states that, for all data sets, population or sample, the percentage of values within k standard deviations of the mean must be at least:

$$[1 - (\frac{1}{k})^2] \cdot 100\%$$

Example

At least	Within
$\left(1-\frac{1}{2^2}\right)\times 100\% = 75\%$	$k = 2(\mu \pm 2\sigma)$
$\left(1-\frac{1}{3^2}\right)\times 100\% = 88.89\%$	$k = 3(\mu \pm 3\sigma)$

Approximating the Mean

- Sometimes only a frequency distribution is available, not the raw data.
- Use the midpoint of a class interval to approximate the values in that class.

$$\overline{X} = \frac{\sum_{j=1}^{c} m_j f_j}{n}$$

Where

n = number of values or sample size

c = number of classes in the frequency distribution

m_j = midpoint of the j^{th} class

f_j = number of values in the j^{th} class

Approximating the Standard Deviation

- Assume that all values within each class interval are located at the midpoint of the class.

$$S = \sqrt{\frac{\sum_{j=1}^{c} (m_j - \overline{X})^2 f_j}{n-1}}$$

The Five Number Summary

- The five-number summary is numerical data summarised by quartiles.
- The five numbers that help describe the center, spread and shape of data are:
  - X_{smallest}
  - Q_1
  - Median
  - Q_3
  - X_{largest}

Relationships Among the Five-Number Summary and Distribution Shape

Left-Skewed	Symmetric	Right-Skewed
$\text{Median} - X_{\text{smallest}}$	$\text{Median} - X_{\text{smallest}}$	$\text{Median} - X_{\text{smallest}}$
$>$	$\approx$	$<$
$X_{\text{largest}} - \text{Median}$	$X_{\text{largest}} - \text{Median}$	$X_{\text{largest}} - \text{Median}$
$Q_1 - X_{\text{smallest}}$	$Q_1 - X_{\text{smallest}}$	$Q_1 - X_{\text{smallest}}$
$>$	$\approx$	$<$
$X_{\text{largest}} - Q_3$	$X_{\text{largest}} - Q_3$	$X_{\text{largest}} - Q_3$
$\text{Median} - Q_1$	$\text{Median} - Q_1$	$\text{Median} - Q_1$
$>$	$\approx$	$<$
$Q_3 - \text{Median}$	$Q_3 - \text{Median}$	$Q_3 - \text{Median}$

Box-and-Whisker Plot

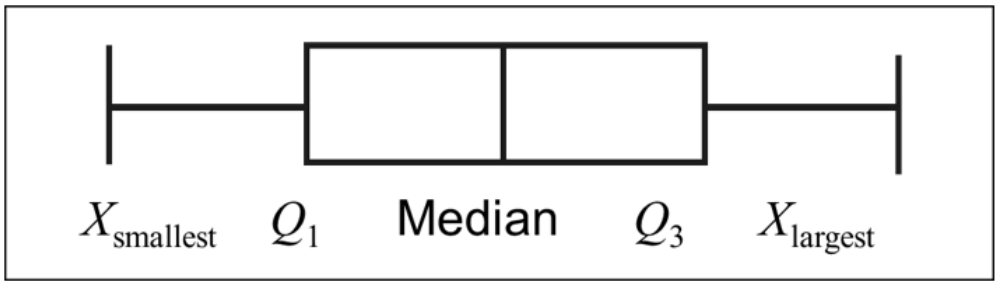
- A **box-and-whisker plot** (also known as a **boxplot**) is a graphical display of data using the five-number summary.

$$X_{\text{smallest}} \text{--} Q_1 \text{--} \text{Median} \text{--} Q_3 \text{--} X_{\text{largest}}$$

- It shows the range, interquartile range and quartiles.
- The vertical line drawn within the box represents the median.
- The vertical line at the left side of the box represents Q_1.
- The vertical line at the right side of the box represents Q_3.

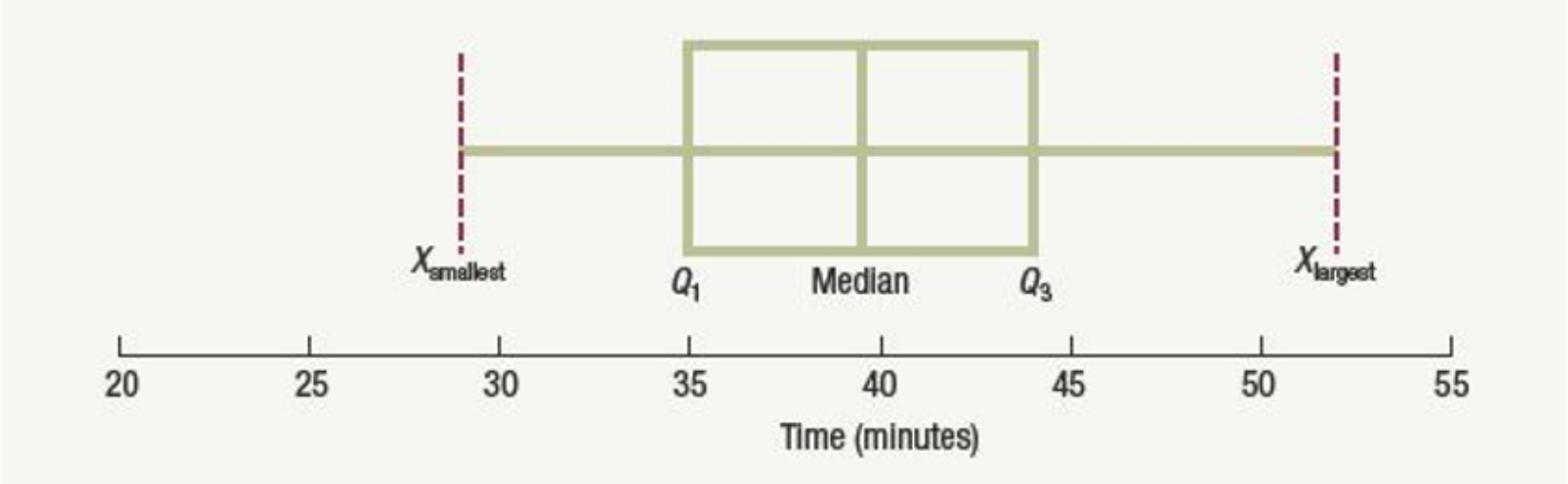
Shape of Boxplots

- If data are symmetric around the median then the box and central line are centered between the endpoints.



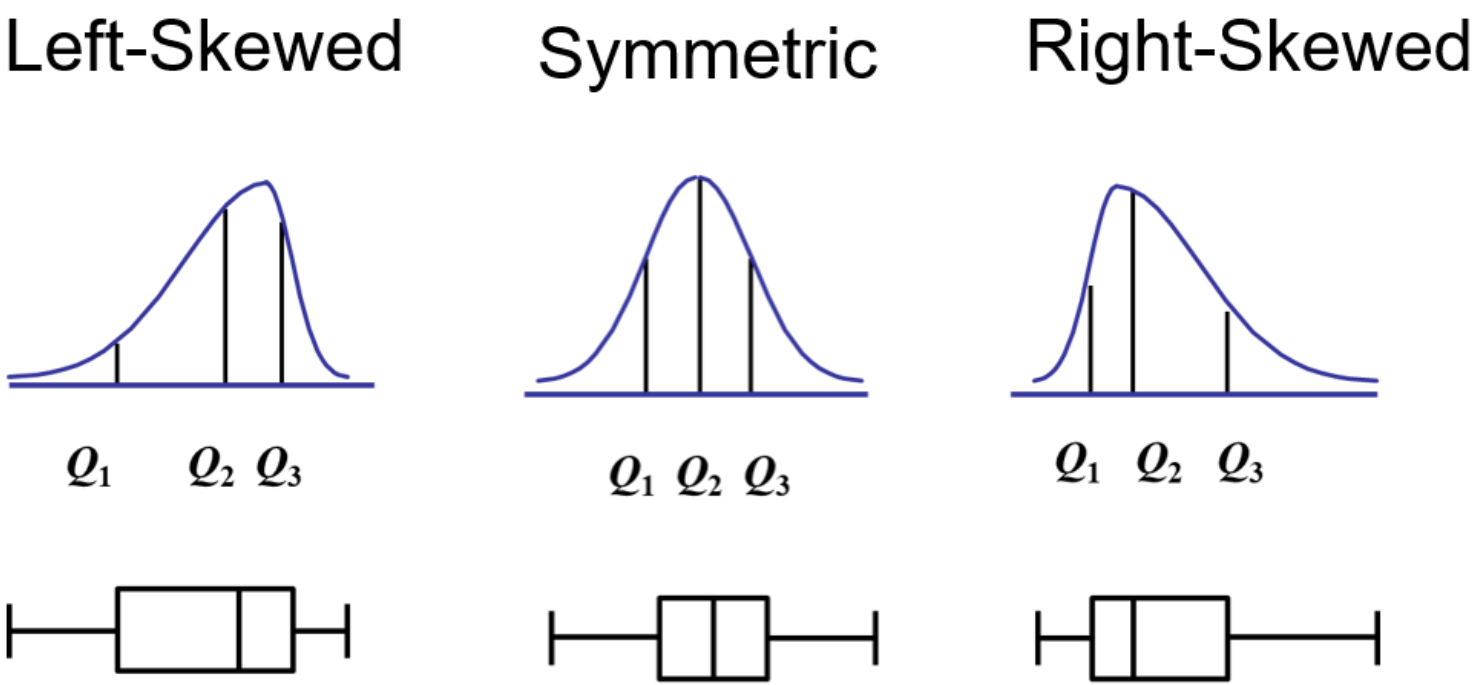
- A Boxplot can be shown in either a vertical or horizontal orientation.

Example



The box-and-whisker plot of the times to get ready in the above figure confirms a very slight right skewness since the right whisker is slightly longer than the left whisker

Distribution Shape and The Boxplot



Two Measures of the Relationship Between Two Numerical Variables

- Scatter plots allow you to visually examine the relationship between two numerical variables and now we will discuss two quantitative measures of such relationships.
- The Covariance
- The Coefficient of Correlation

Covariance

- The **covariance measures** the strength of the linear relationship between **two numerical variables**.
- It is only concerned with the direction of the relationship.
- No causal effect is implied.
- It is affected by units of measurement.

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

Interpreting Covariance

- **Covariance** between two variables:
  - $cov(X, Y) > 0 \rightarrow X$ and Y tend to move in the same direction
  - $cov(X, Y) < 0 \rightarrow X$ and Y tend to move in opposite directions
  - $cov(X, Y) = 0 \rightarrow X$ and Y are independent
- The covariance has a major flaw:
  - It is not possible to determine the relative strength of the relationship from the size of the covariance.

Coefficient of Correlation

- The coefficient of correlation measures the relative strength of a linear relationship between two numerical variables.
- Sample coefficient of correlation:

$$r = \frac{\operatorname{cov}(X,Y)}{S_X S_Y}$$

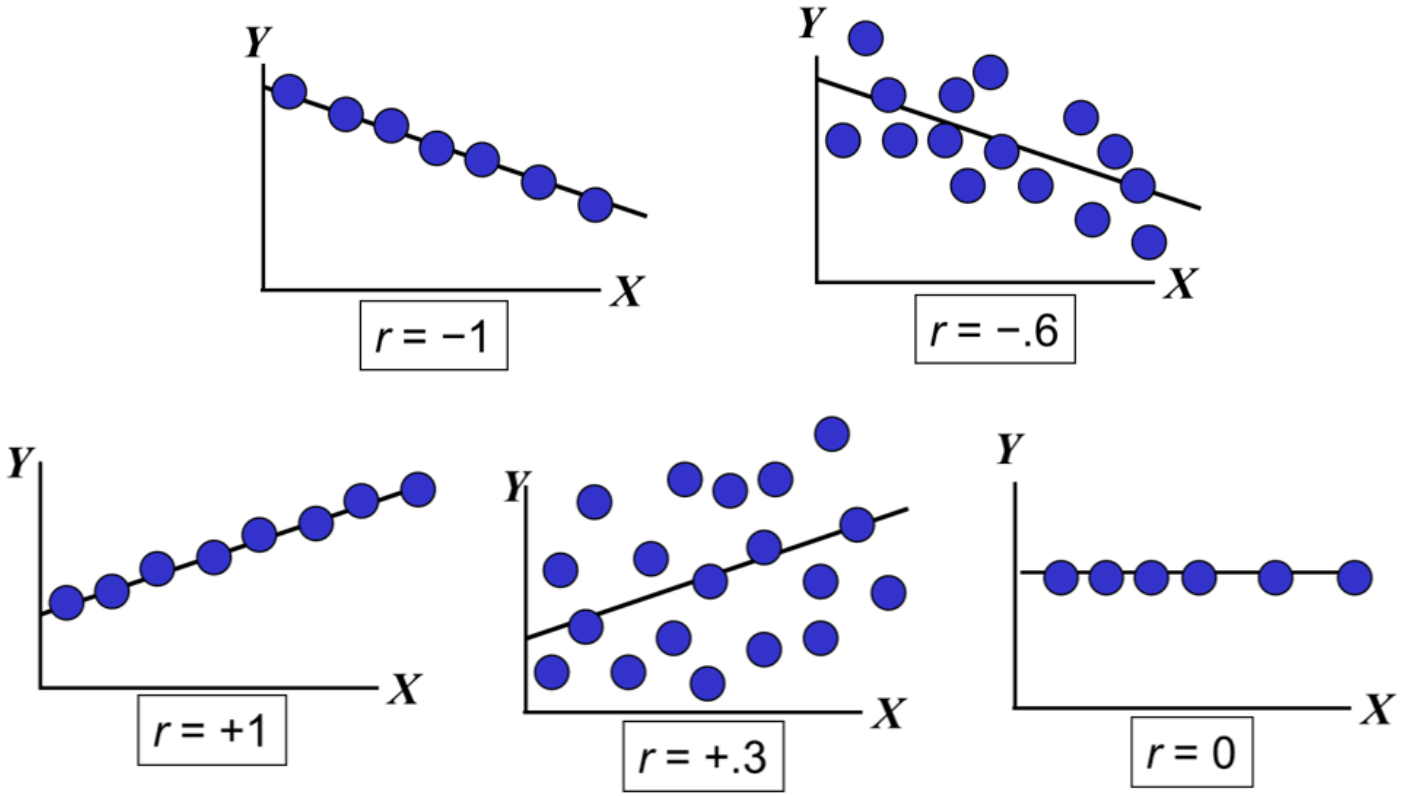
where

$$\operatorname{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1} \quad S_x = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}} \quad S_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1}}$$

Features of the Coefficient of Correlation

- The population coefficient of correlation is referred as ρ.
- The sample coefficient of correlation is referred to as r.
- Either ρ or r have the following features:
  - Unit free
  - Range between -1 and 1
  - The closer to -1, the stronger the negative linear relationship
  - The closer to 1, the stronger the positive linear relationship
  - The closer to 0, the weaker the linear relationship

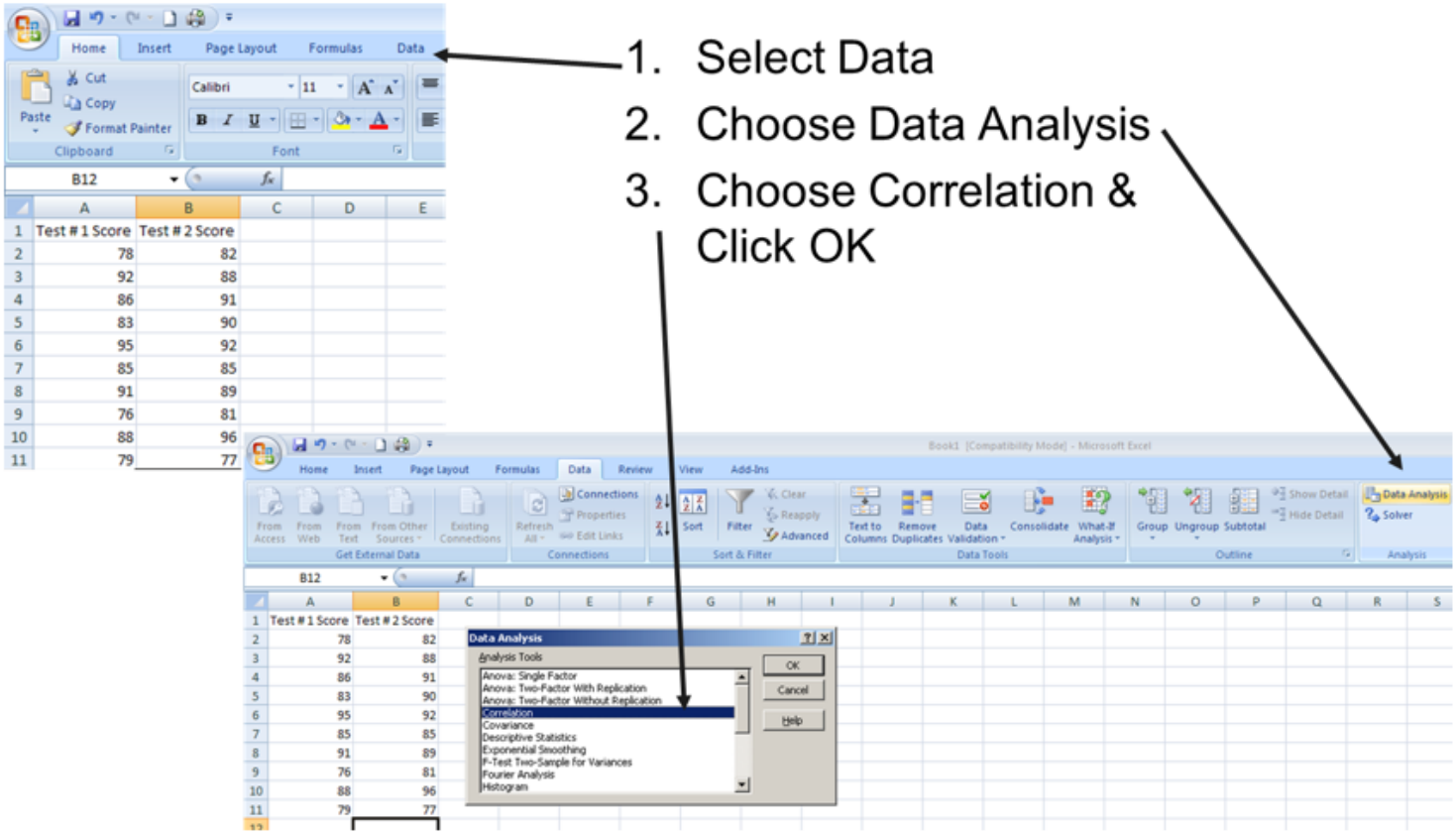
Scatter Plots of Sample Data with Various Coefficients of Correlation



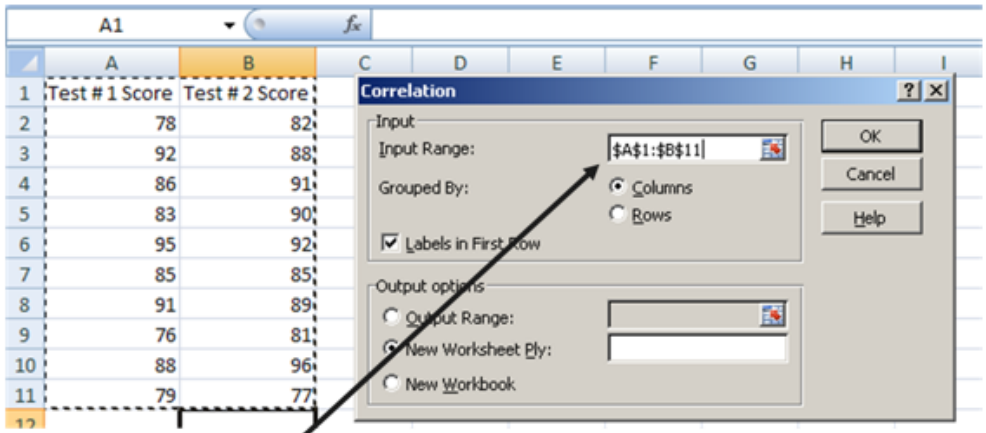
Coefficient of Correlation Using Microsoft Excel

Test #1 Score	Test #2 Score		**Correlation Coefficient**	
78	82		0.7332	=CORREL(A2:A11,B2:B11)
92	88			
86	91			
83	90			
95	92			
85	85			
91	89			
76	81			
88	96			
79	77			

Coefficient of Correlation Using Data Analysis ToolPak (1 of 3)



Coefficient of Correlation Using Data Analysis ToolPak (2 of 3)

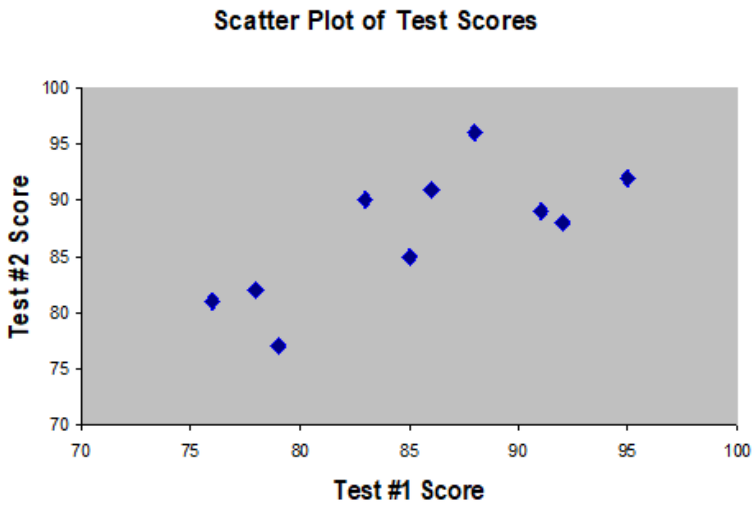


4. Input data range and select appropriate options
5. Click OK to get output

	A	B	C
1		*Test # 1 Score*	*Test # 2 Score*
2	Test # 1 Score	1	
3	Test # 2 Score	0.733243705	1
4			

Coefficient of Correlation Using Data Analysis ToolPak (3 of 3)

- $r = 0.733$
- There is a relatively strong positive linear relationship between test score #1 and test score #2.
- Students who scored high on the first test tended to score high on second test.



Pitfalls in Numerical Descriptive Measures

- While one's analysis is **objective**, one's interpretation is **subjective**.
- Be careful to avoid errors that may arise either in the objectivity of your analysis or in the subjectivity of your interpretation.
- Everyone sees the world from different perspectives.
- But findings should be presented in a fair, neutral and transparent manner.

Ethical Considerations

- Ethical considerations arise when you are deciding what results to include in a report.
- Both good and bad results should be included.
- In oral and written reports, results should be given in a fair and neutral manner.
- Unethical behavior includes:
  - Willfully choosing an inappropriate summary measure to distort facts.
  - Selectively failing to report pertinent findings.