Statistics for Managers Using Microsoft® Excel 5th Edition

Chapter 3

Numerical Descriptive Measures

Learning Objectives

In this chapter, you will learn:

- To describe the properties of central tendency, variation and shape in numerical data
- To calculate descriptive summary measures for a population
- To construct and interpret a box-and-whisker plot
- To describe the covariance and coefficient of correlation

Summary Definitions

- The **central tendency** is the extent to which all the data values group around a typical or central value.
- The **variation** is the amount of dispersion, or scattering, of values
- The **shape** is the pattern of the distribution of values from the lowest value to the highest value.

Measures of Central Tendency
The Arithmetic Mean

- The arithmetic mean (mean) is the most common measure of central tendency

For a sample of size n:

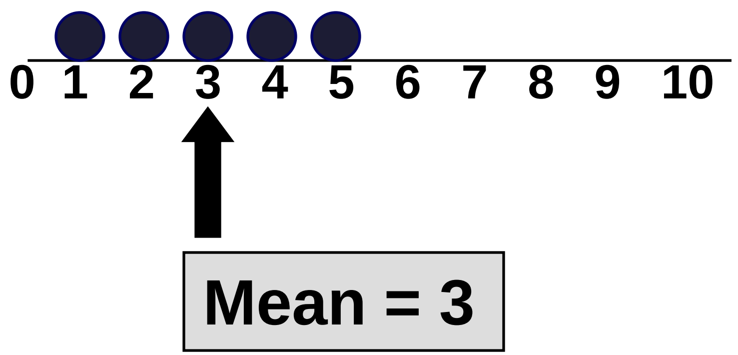
$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Sample size

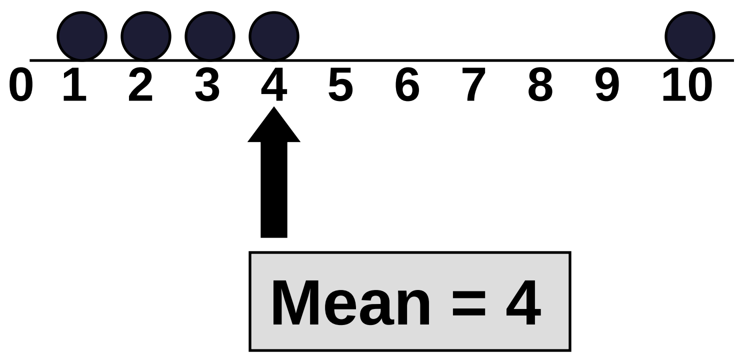
Observed values

Measures of Central Tendency
The Arithmetic Mean

- The most common measure of central tendency
- Mean = sum of values divided by the number of values
- Affected by extreme values (outliers)



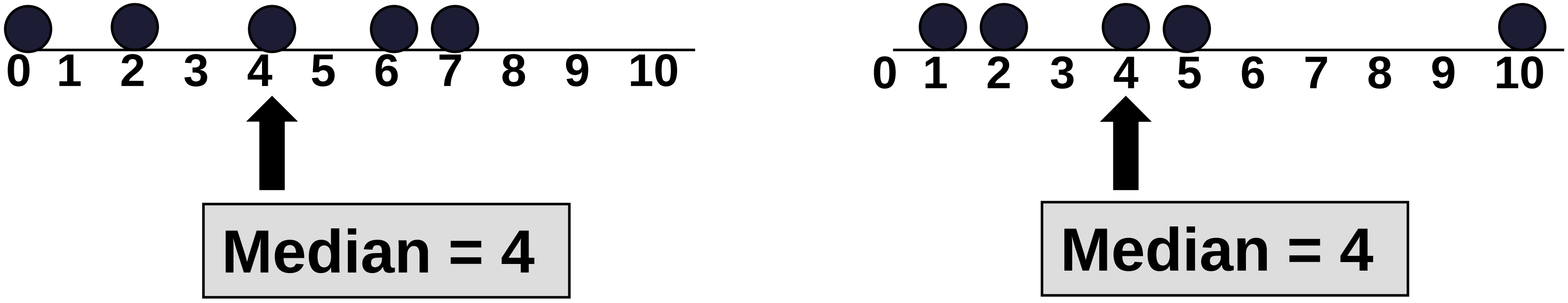
$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$



$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

Measures of Central Tendency
The Median

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



- Not affected by extreme values

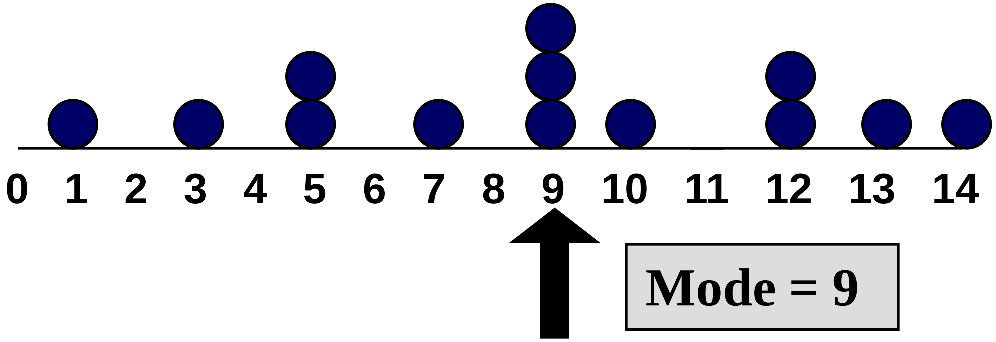
Measures of Central Tendency Locating the Median

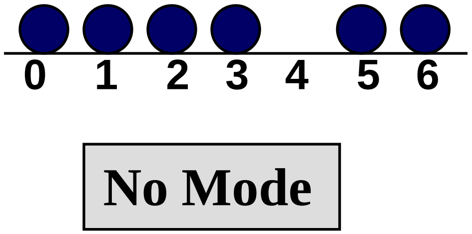
- The median of an ordered set of data is located at the $\frac{n+1}{2}$ ranked value.

- If the number of values is odd, the median is the middle number.
- If the number of values is even, the median is the average of the two middle numbers.
- Note that $\frac{n+1}{2}$ is NOT the value of the median, only the position of the median in the ranked data.

Measures of Central Tendency The Mode

- Value that occurs most often
- Not affected by extreme values
- Used for either numerical or categorical data
- There may be no mode
- There may be several modes





Measures of Central Tendency Review Example

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

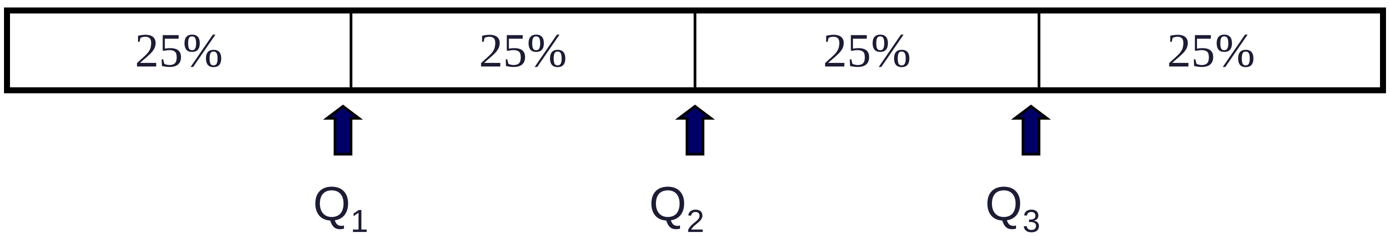
- **Mean:** ($3,000,000/5)
  = **$600,000**
- **Median:** middle value of ranked data
  = **$300,000**
- **Mode:** most frequent value
  = **$100,000**

Measures of Central Tendency Which Measure to Choose?

- The **mean** is generally used, unless extreme values (outliers) exist.
- Then **median** is often used, since the median is not sensitive to extreme values. For example, median home prices may be reported for a region; it is less sensitive to outliers.

Quartile Measures

- Quartiles split the ranked data into 4 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
- Q_2 is the same as the median (50% are smaller, 50% are larger)
- Only 25% of the values are greater than the third quartile

Quartile Measures
Locating Quartiles

Find a quartile by determining the value in the appropriate position in the ranked data, where

First quartile position: $\mathbf{Q_1 = (n+1)/4}$ **ranked value**

Second quartile position: $\mathbf{Q_2 = (n+1)/2}$ **ranked value**

Third quartile position: $\mathbf{Q_3 = 3(n+1)/4}$ **ranked value**

where **n** is the number of observed values

Quartile Measures Guidelines

- Rule 1: If the result is a whole number, then the quartile is equal to that ranked value.

- Rule 2: If the result is a fraction half (2.5, 3.5, etc), then the quartile is equal to the average of the corresponding ranked values.

- Rule 3: If the result is neither a whole number or a fractional half, you round the result to the nearest integer and select that ranked value.

Quartile Measures
Locating the First Quartile

- Example: Find the first quartile

Sample Data in Ordered Array:	11	12	13	16	16	17	18	21	22

First, note that n = 9.

Q_1 = is in the **(9+1)/4 = 2.5 ranked value** of the ranked data, so use the value half way between the 2^{nd} and 3^{rd} ranked values,

so **Q_1 = 12.5**

Q_1 and Q_3 are measures of non-central location
Q_2 = median, a measure of central tendency

Measures of Central Tendency The Geometric Mean

- Geometric mean
  - Used to measure the rate of change of a variable over time

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

- Geometric mean rate of return
  - Measures the status of an investment over time

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

  - Where R_i is the rate of return in time period i

Measures of Central Tendency
The Geometric Mean

An investment of $100,000 declined to $50,000 at the end of year one and rebounded to $100,000 at end of year two:

$$\underbrace{X_1 = \$100{,}000 \quad X_2 = \$50{,}000}_{\text{50\% decrease}} \quad X_3 = \$100{,}000$$

$$X_1 = \$100{,}000 \quad \underbrace{X_2 = \$50{,}000 \quad X_3 = \$100{,}000}_{\text{100\% increase}}$$

50% decrease

100% increase

The overall two-year return is zero, since it started and ended at the same level.

Measures of Central Tendency
The Geometric Mean

Use the 1-year returns to compute the arithmetic mean and the geometric mean:

Arithmetic mean rate of return:

$$\overline{X} = \frac{(-.5)+(1)}{2} = .25$$

Misleading result

Geometric mean rate of return:

$$\begin{aligned}\overline{R}_G &= [(1+R_1)\times(1+R_2)\times\cdots\times(1+R_n)]^{1/n} - 1\\ &= [(1+(-.5))\times(1+(1))]^{1/2} - 1\\ &= [(.50)\times(2)]^{1/2} - 1 = 1^{1/2} - 1 = 0\%\end{aligned}$$

More accurate result

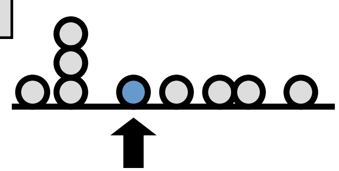
Measures of Central Tendency Summary

Central Tendency

Arithmetic Mean

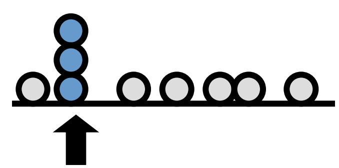
$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Median



Middle value in the ordered array

Mode



Most frequently observed value

Geometric Mean

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

Measures of Variation

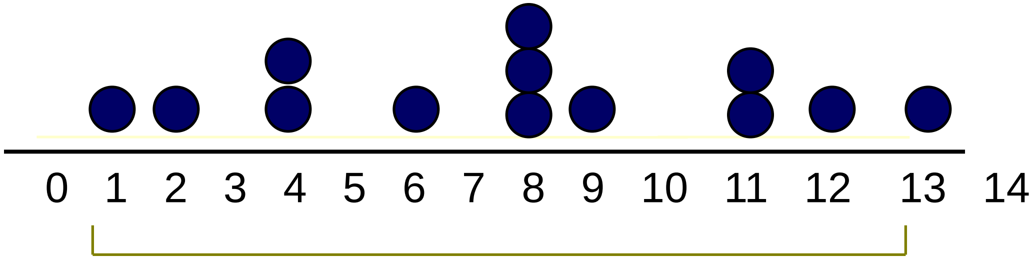
- Variation measures the spread, or dispersion, of values in a data set.
  - Range
  - Interquartile Range
  - Variance
  - Standard Deviation
  - Coefficient of Variation

Measures of Variation
Range

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

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

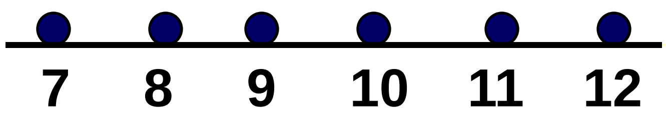
Example:



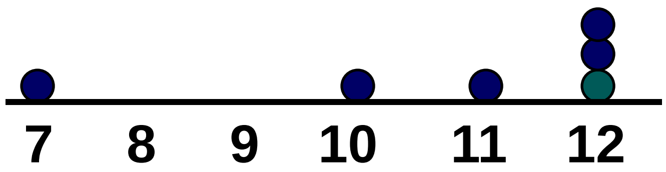
Range = 13 - 1 = 12

Measures of Variation
Disadvantages of the Range

- Ignores the way in which data are distributed



Range = 12 - 7 = 5



Range = 12 - 7 = 5

- Sensitive to outliers

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,**5**

Range = 5 - 1 = 4

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,**120**

Range = 120 - 1 = 119

Measures of Variation
Interquartile Range

- Problems caused by outliers can be eliminated by using the **interquartile range.**

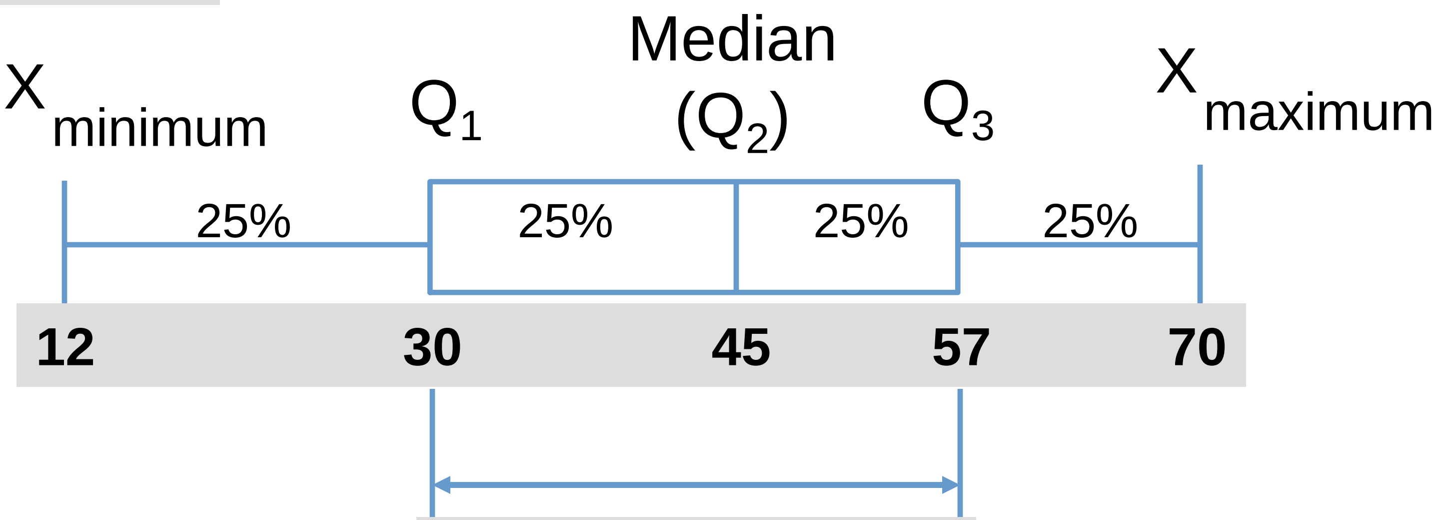
- The IQR can eliminate some high and low values and calculate the range from the remaining values.

- Interquartile range = 3rd quartile – 1st quartile

$$= Q_3 - Q_1$$

Measures of Variation
Interquartile Range

Example:



Interquartile range
= 57 – 30 = 27

Measures of Variation
Variance

- The **variance** is the average (approximately) of squared deviations of values from the mean.

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

Where $\bar{X}$ = arithmetic mean

n = sample size

X_i = i[th] value of the variable X

Measures of Variation
Standard Deviation

- Most commonly used measure of variation
- Shows variation about the mean
- Has the same units as the original data

Sample standard deviation: $$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n - 1}}$$

Measures of Variation
Standard Deviation

Steps for Computing Standard Deviation

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

Measures of Variation
Standard Deviation

Sample

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

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

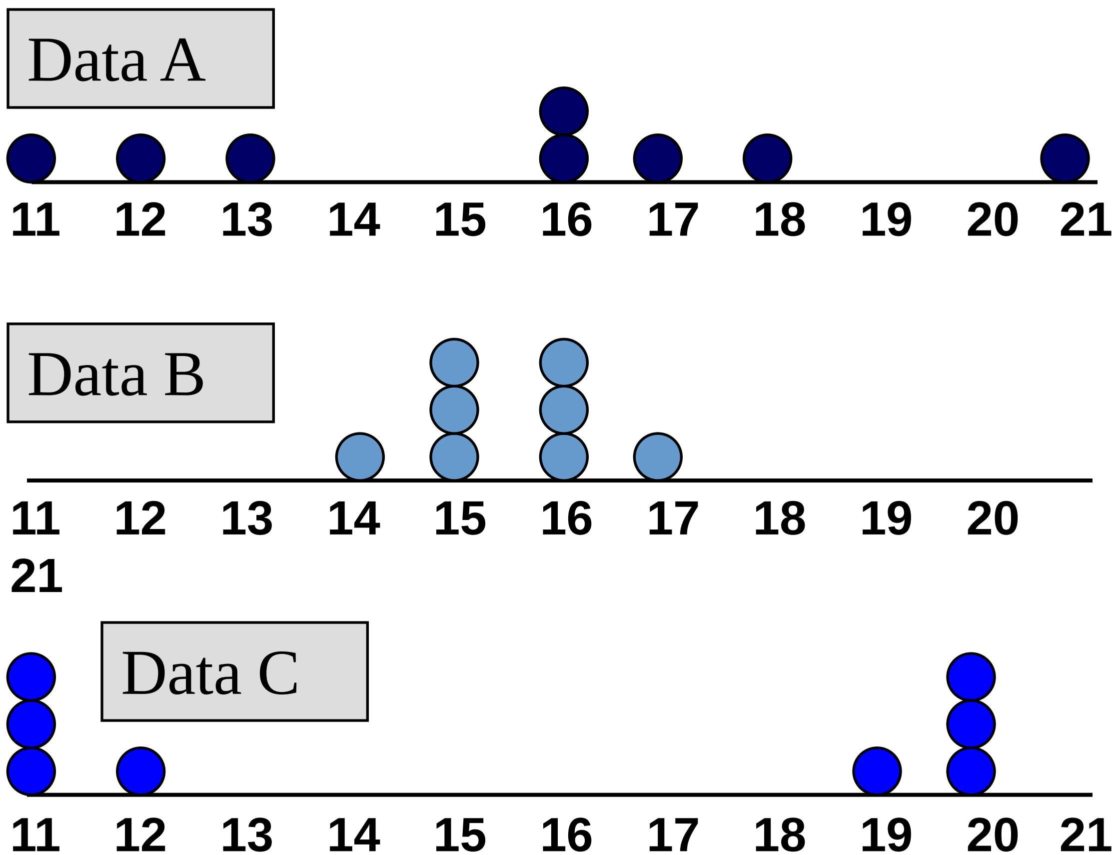
$$S = \sqrt{\frac{(10-\overline{X})^2 + (12-\overline{X})^2 + (14-\overline{X})^2 + \cdots + (24-\overline{X})^2}{n-1}}$$

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

$$= \sqrt{\frac{126}{7}} = \boxed{4.2426}$$

A measure of the "average" scatter around the mean

Measures of Variation
Comparing Standard Deviation

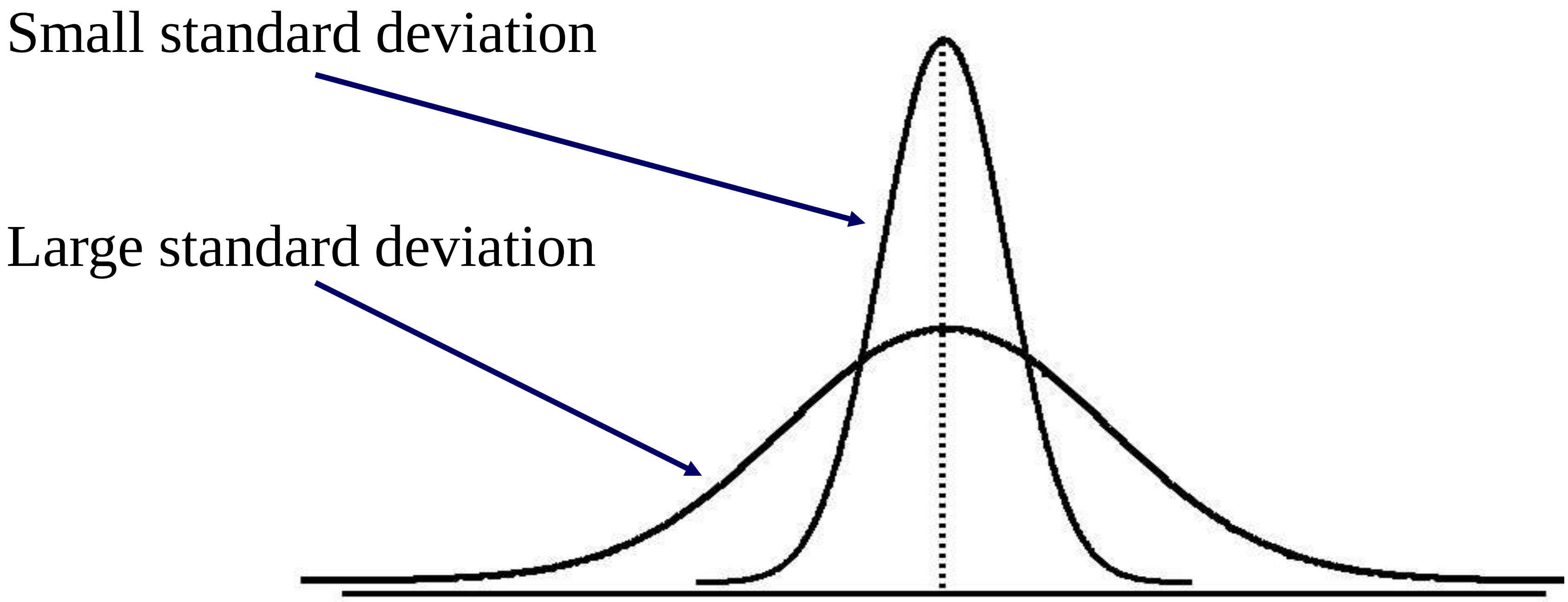


Mean = 15.5
S = 3.338

Mean = 15.5
S = 0.926

Mean = 15.5
S = 4.570

Measures of Variation Comparing Standard Deviation



Measures of Variation Summary Characteristics

- The more the data are spread out, the greater the range, interquartile range, variance, and standard deviation.
- The more the data are concentrated, the smaller the range, interquartile 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 is the standard deviation divided by the mean, multiplied by 100.
- It is always expressed as a percentage. (%)
- It shows variation relative to mean.
- The CV can be used to compare two or more sets of data measured in different units.

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

Coefficient of Variation

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

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

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

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

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

Locating Extreme Outliers Z-Score

- To compute the Z-**score** of a data value, subtract the mean and divide by the standard deviation.

- The Z-score is the number of standard deviations a data value is from the mean.

- A data value is considered an extreme outlier if its Z-score is less than -3.0 or greater than +3.0.

- The larger the absolute value of the Z-score, the farther the data value is from the mean.

Locating Extreme Outliers Z-Score

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

where X represents the data value

$\bar{X}$ is the sample mean

S is the sample standard deviation

Locating Extreme Outliers Z-Score

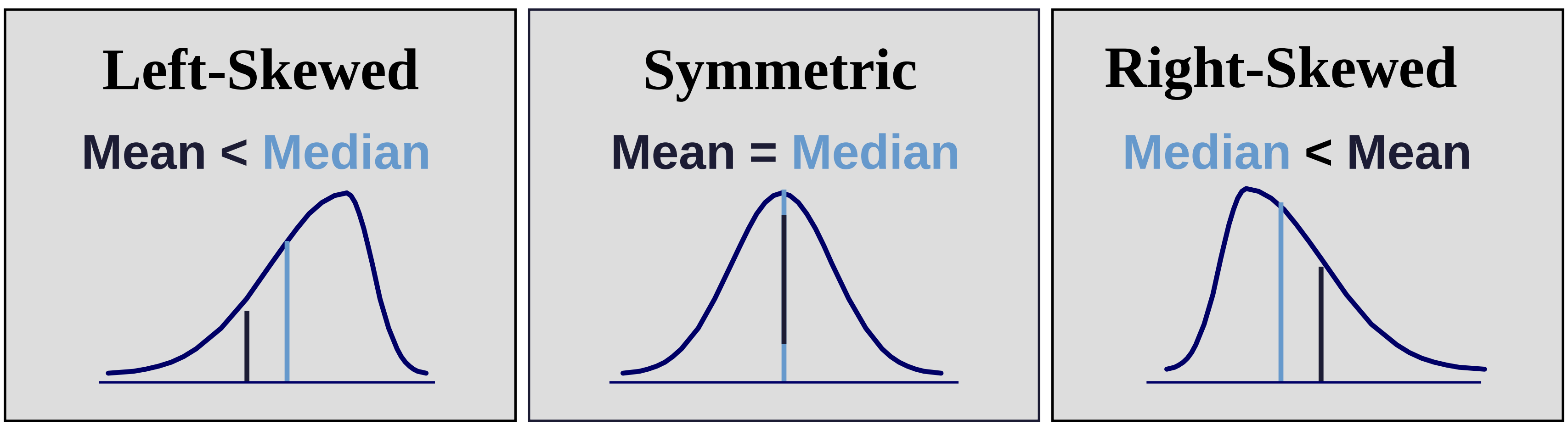
- Suppose the mean math SAT score is 490, with a standard deviation of 100.
- Compute the z-score for a test score of 620.

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

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

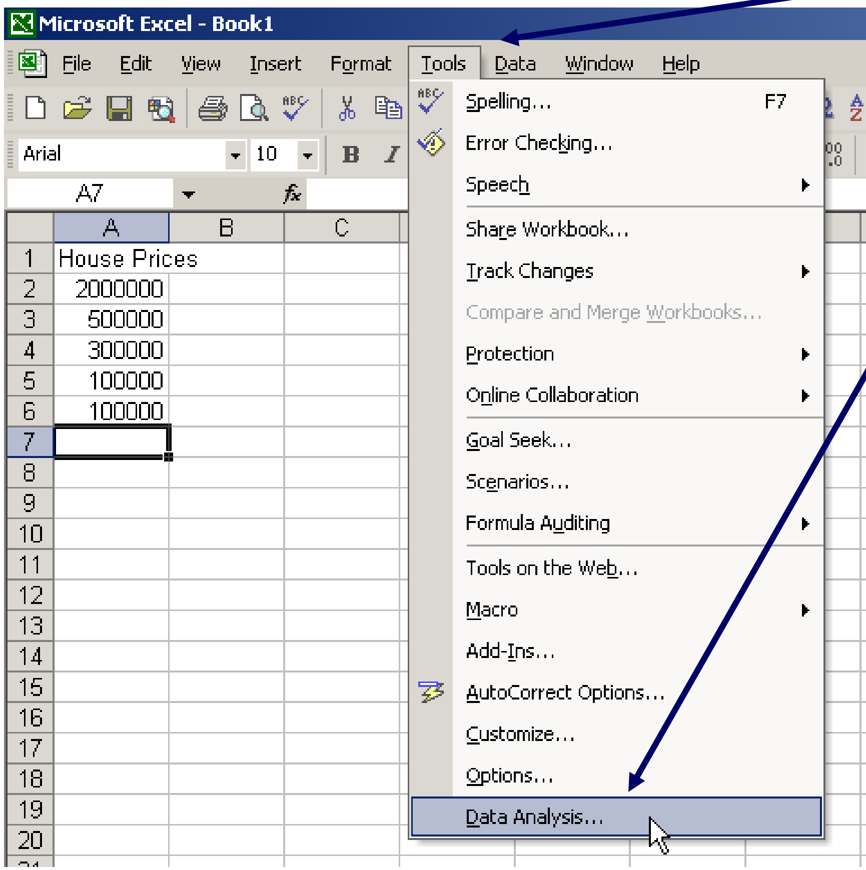
Shape of a Distribution

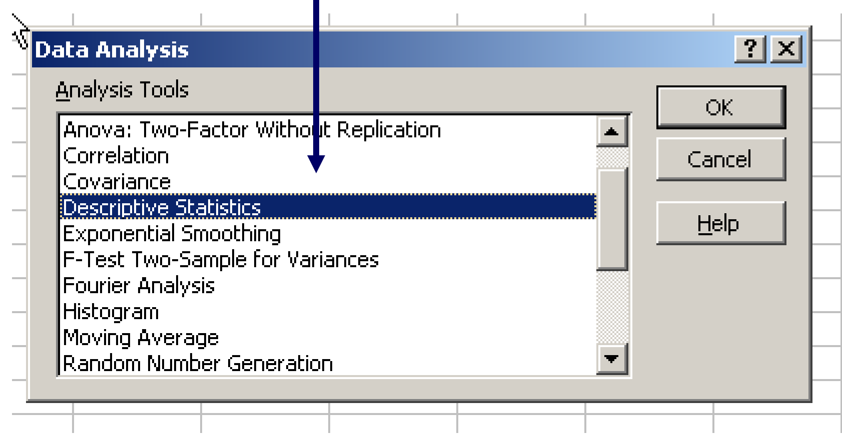
- Describes how data are distributed
- Measures of shape
  - Symmetric or skewed



General Descriptive Stats Using Microsoft Excel

1. Select Tools.
2. Select Data Analysis.
3. Select Descriptive Statistics and click OK.



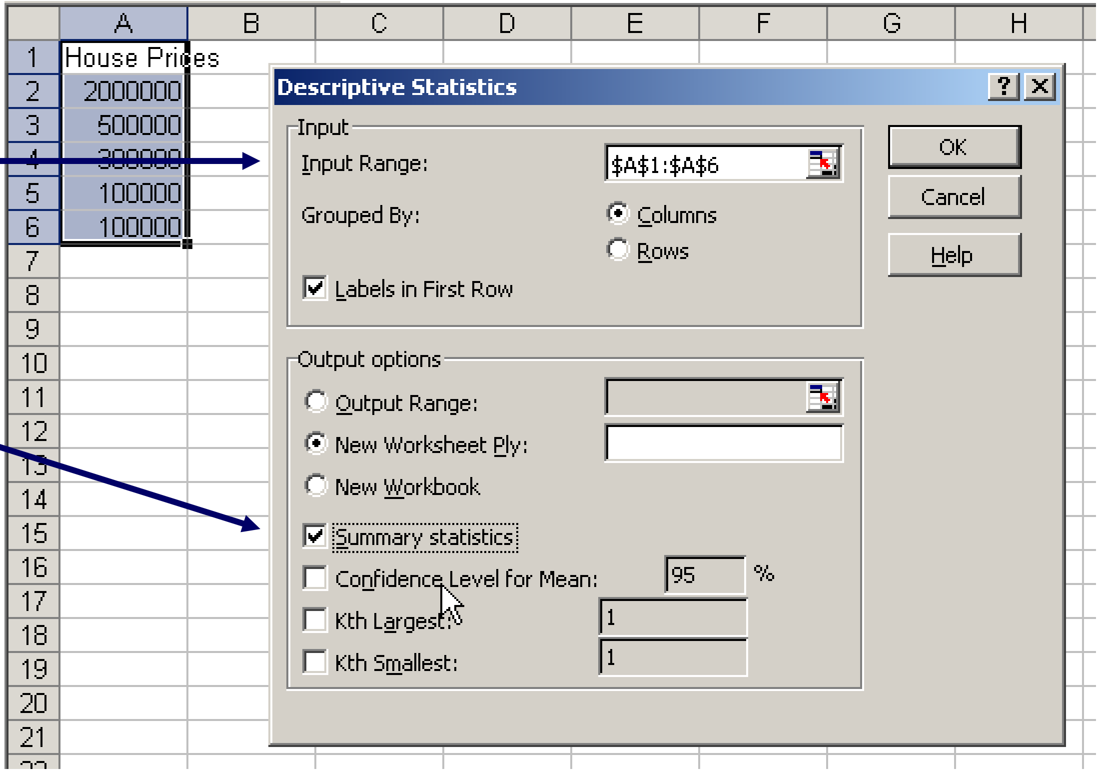


General Descriptive Stats Using Microsoft Excel

4. Enter the cell range.

5. Check the Summary Statistics box.

6. Click OK



General Descriptive Stats Using Microsoft Excel

Microsoft Excel descriptive statistics output, using the house price data:

House Prices:

$2,000,000
500,000
300,000
100,000
100,000

	A	B
1	*House Prices*	
2		
3	Mean	600000
4	Standard Error	357770.8764
5	Median	300000
6	Mode	100000
7	Standard Deviation	800000
8	Sample Variance	6.4E+11
9	Kurtosis	4.130126953
10	Skewness	2.006835938
11	Range	1900000
12	Minimum	100000
13	Maximum	2000000
14	Sum	3000000
15	Count	5
16		

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.

Population Mean

- 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^{th} value of the variable X

Population Variance

- The population variance is the average of 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$^{\text{th}}$ value of the variable X

Population Standard Deviation

- The **population standard deviation** is the most commonly used measure of variation.
- It has the same units as the original data.

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

Where μ = population mean

N = population size

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

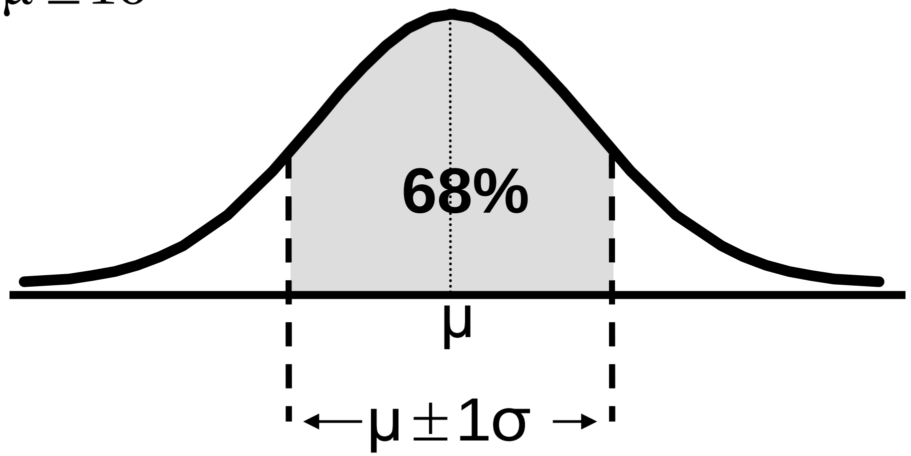
Sample statistics versus population parameters

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

The Empirical Rule

- The **empirical rule** approximates the variation of data in bell-shaped distributions.

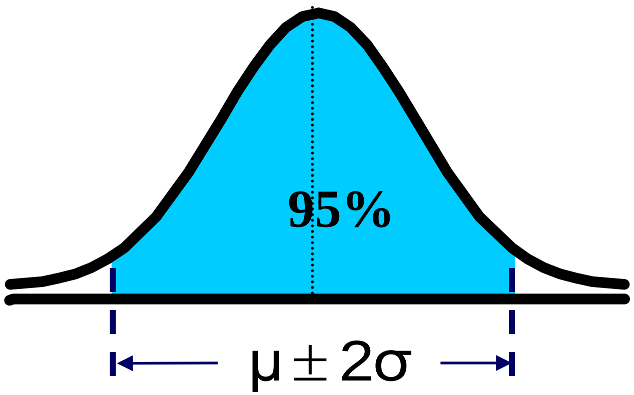
Approximately 68% of the data in a bell-shaped distribution lies within one standard deviation of the mean, or $\mu \pm 1\sigma$

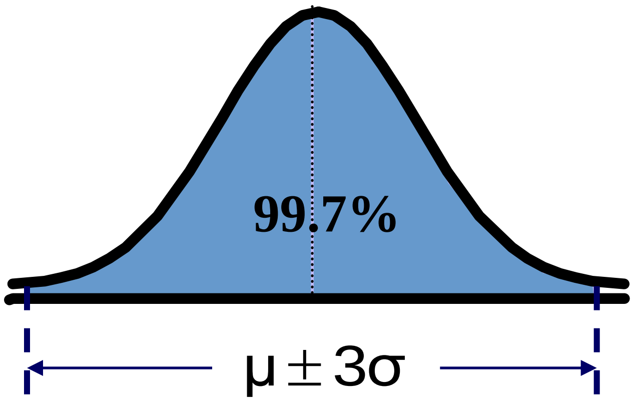


The Empirical Rule

- Approximately 95% of the data in a bell-shaped distribution lies within two standard deviation of the mean, or $\mu \pm 2\sigma$

- Approximately 99.7% of the data in a bell-shaped distribution lies within three standard deviation of the mean, or $\mu \pm 3\sigma$





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, :

  - 68% of all test takers scored between 410 and 590 (500 +/- 90).
  - 95% of all test takers scored between 320 and 680 (500 +/- 180).
  - 99.7% of all test takers scored between 230 and 770 (500 +/- 270).

Chebyshev Rule

- Regardless of how the data are distributed (symmetric or skewed), at least $(1 - 1/k^2)$ of the values will fall within k standard deviations of the mean (for k > 1)

- Examples:

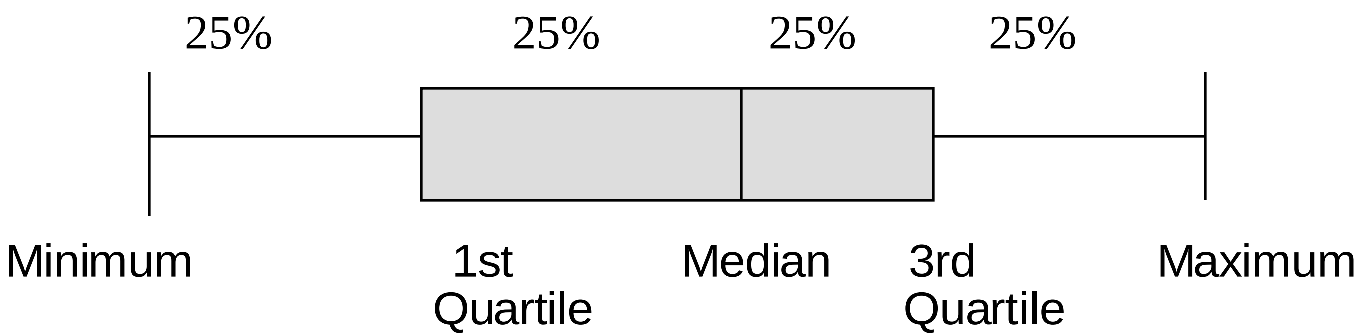
	At least	within
k=2	$(1 - 1/2^2) = 75\%$ ………...	$(\mu \pm 2\sigma)$
k=3	$(1 - 1/3^2) = 89\%$ ……….	$(\mu \pm 3\sigma)$

Exploratory Data Analysis The Five Number Summary

- The five numbers that describe the spread of data are:
  - Minimum
  - First Quartile (Q_1)
  - Median (Q_2)
  - Third Quartile (Q_3)
  - Maximum

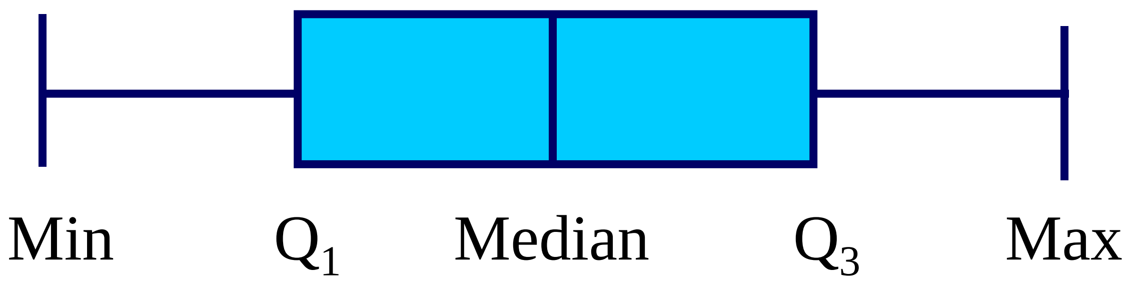
Exploratory Data Analysis
The Box-and-Whisker Plot

- The Box-and-Whisker Plot is a graphical display of the five number summary.



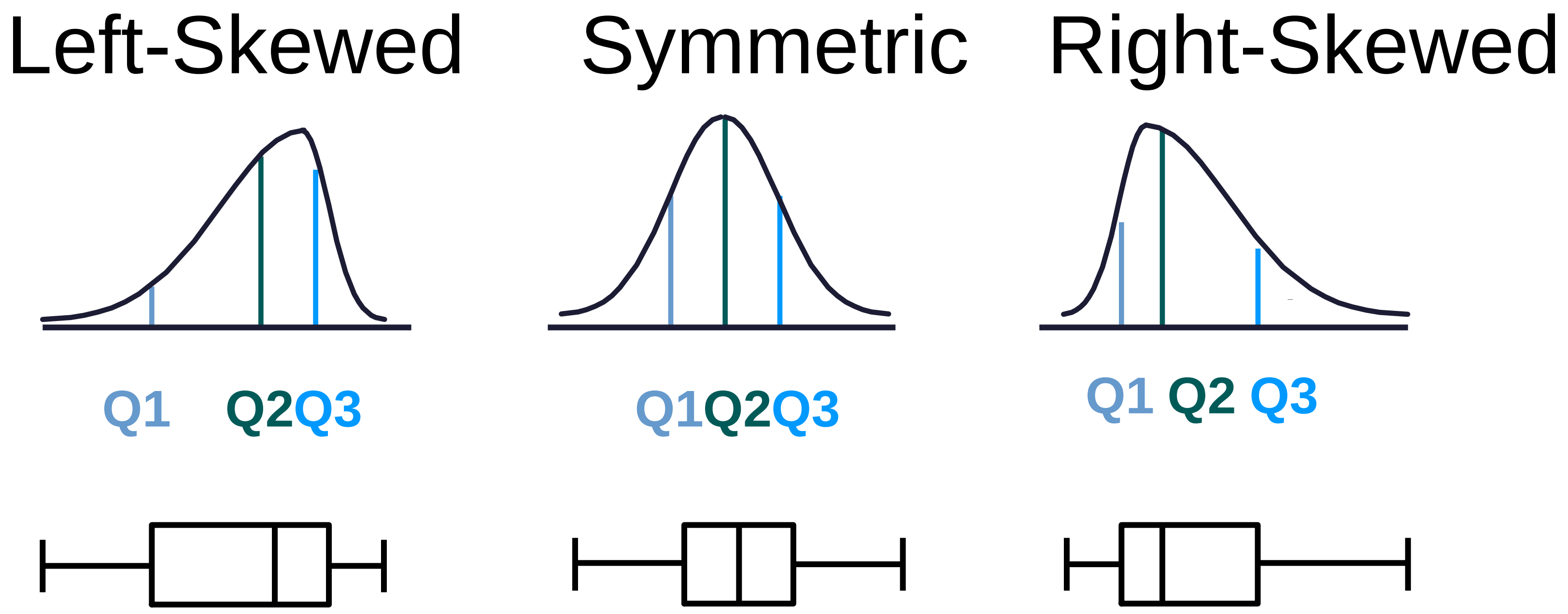
Exploratory Data Analysis The Box-and-Whisker Plot

- The Box and central line are centered between the endpoints if data are symmetric around the median.



- A Box-and-Whisker plot can be shown in either vertical or horizontal format.

Exploratory Data Analysis
The Box-and-Whisker Plot



Sample Covariance

- The **sample covariance** measures the strength of the linear relationship between two numerical variables.

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

- The covariance is only concerned with the strength of the relationship.
- No causal effect is implied.

Sample Covariance

- **Covariance** between two random variables:

- $\text{cov}(X,Y) > 0$ X and Y tend to move in the same direction
- $\text{cov}(X,Y) < 0$ X and Y tend to move in opposite directions
- $\text{cov}(X,Y) = 0$ X and Y are independent

The Correlation Coefficient

- The **correlation coefficient** measures the relative strength of the *linear* relationship between two variables.

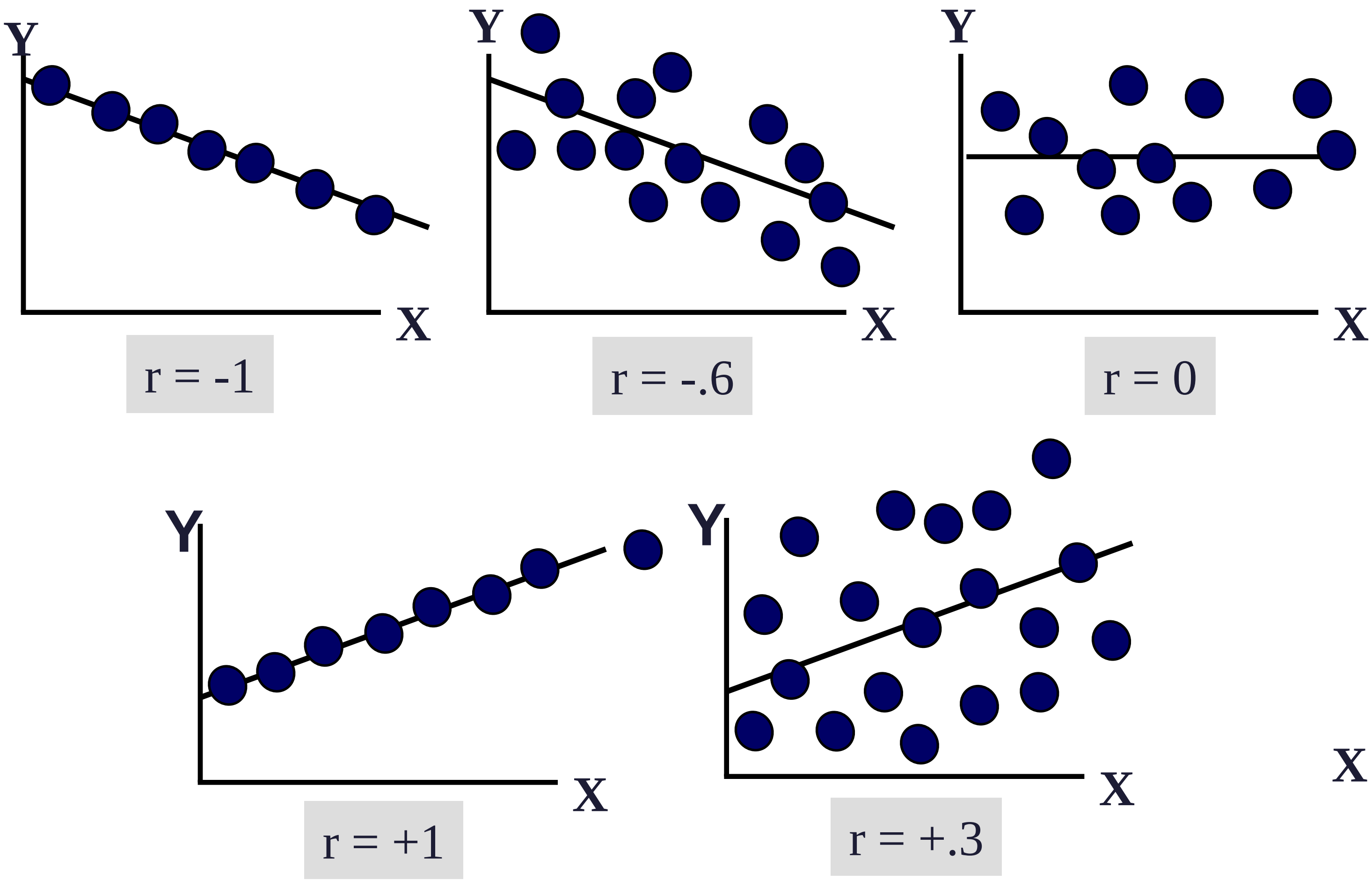
- Sample coefficient of correlation:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{\text{cov}(X,Y)}{S_X S_Y}$$

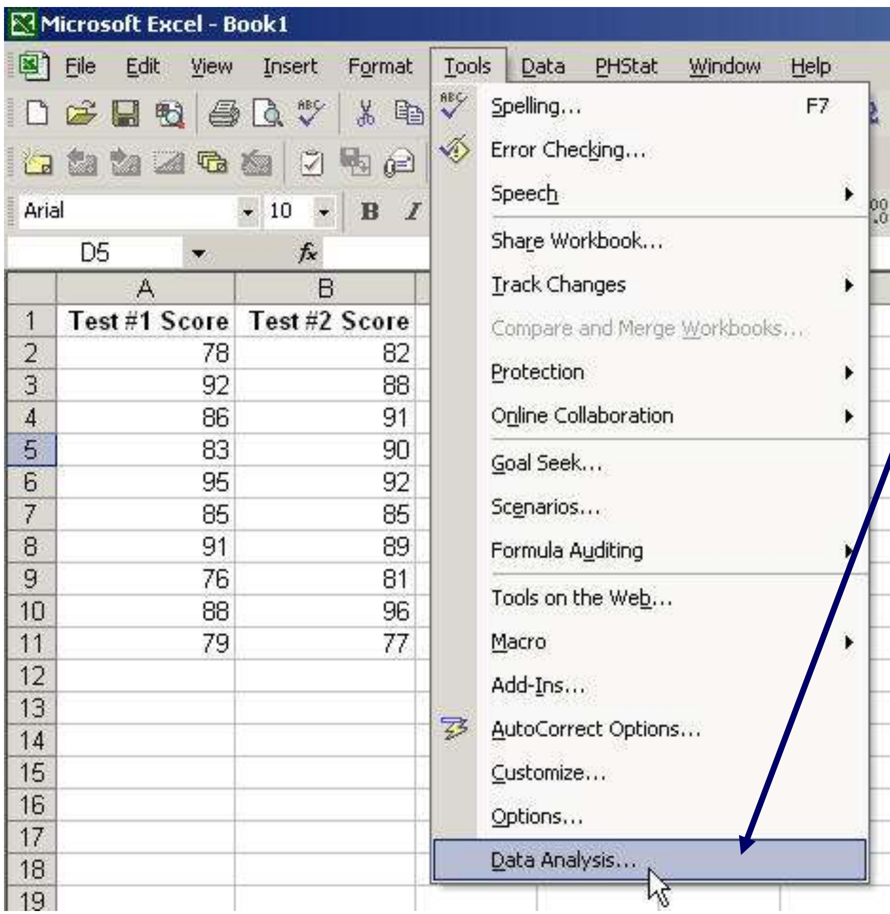
The Correlation Coefficient

- Unit free
- Ranges 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 any linear relationship

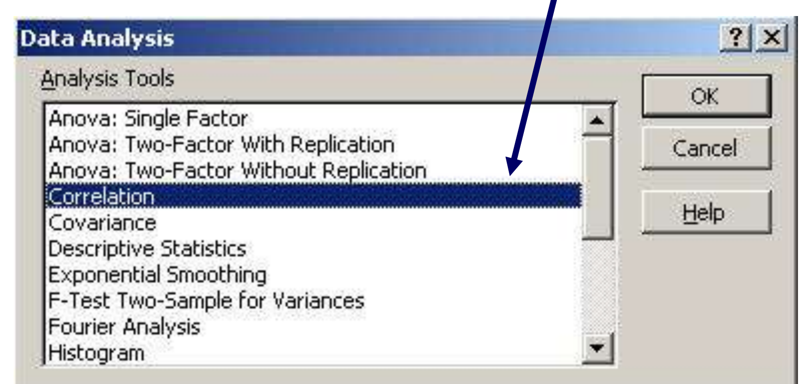
The Correlation Coefficient



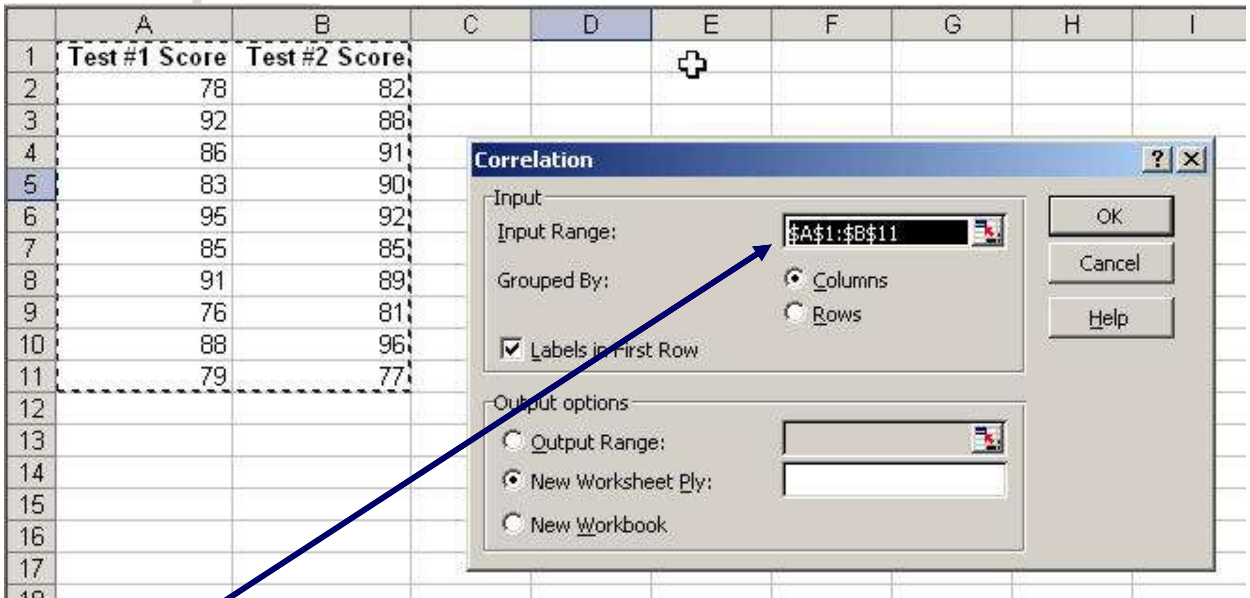
The Correlation Coefficient Using Microsoft Excel



1. Select Tools/Data Analysis
2. Choose Correlation from the selection menu
3. Click OK . . .



The Correlation Coefficient Using Microsoft Excel

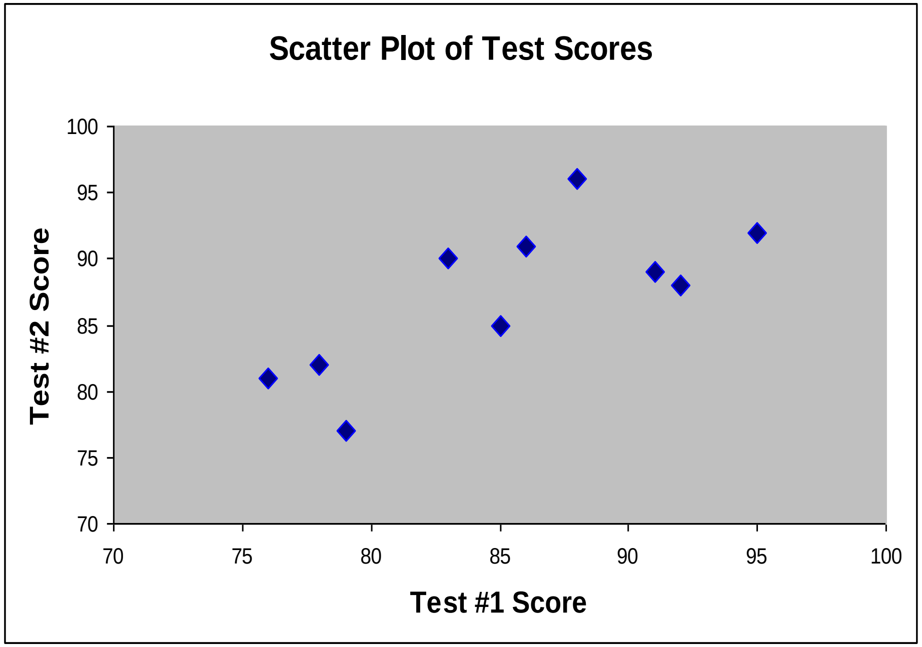


3. Input data range and select appropriate options
4. Click OK to get output

	Test #1 Score	Test #2 Score
Test #1 Score	1	
Test #2 Score	0.733243705	1

The Correlation Coefficient Using Microsoft Excel

- r = .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

- Data analysis is **objective**
  - Analysis should report the summary measures that best meet the assumptions about the data set.
- Data interpretation is **subjective**
  - Interpretation should be done in fair, neutral and clear manner.

Ethical Considerations

Numerical descriptive measures:

- Should document both good and bad results
- Should be presented in a fair, objective and neutral manner
- Should not use inappropriate summary measures to distort facts

Chapter Summary

In this chapter, we have

- Described measures of central tendency
  - Mean, median, mode, geometric mean
- Discussed quartiles
- Described measures of variation
  - Range, interquartile range, variance and standard deviation, coefficient of variation
- Illustrated shape of distribution
  - Symmetric, skewed, box-and-whisker plots

Chapter Summary

In this chapter, we have

- Discussed covariance and correlation coefficient.
- Addressed pitfalls in numerical descriptive measures and ethical considerations.