Basic Business Statistics
11th Edition

Chapter 3

Numerical Descriptive Measures

Learning Objectives

In this chapter, you learn:

- To describe the properties of central tendency, variation, and shape in numerical data
- To calculate descriptive summary measures for a population
- To calculate descriptive summary measures for a frequency distribution
- To construct and interpret a boxplot
- To calculate the covariance and the 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 Mean

- The arithmetic mean (often just called “mean”) is the most common measure of central tendency

Pronounced x-bar

- For a sample of size n:

The i^{th} value

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

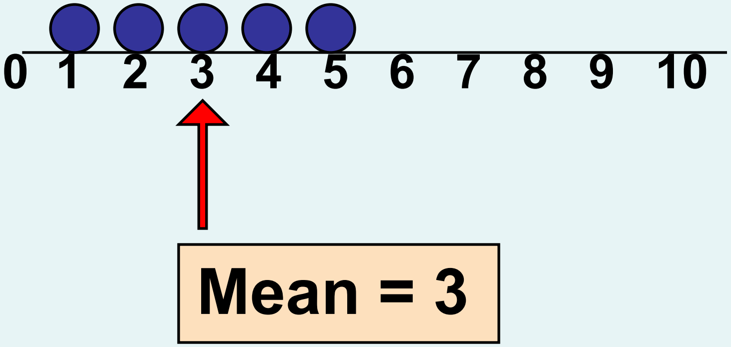
Sample size

Observed values

Measures of Central Tendency: The Mean

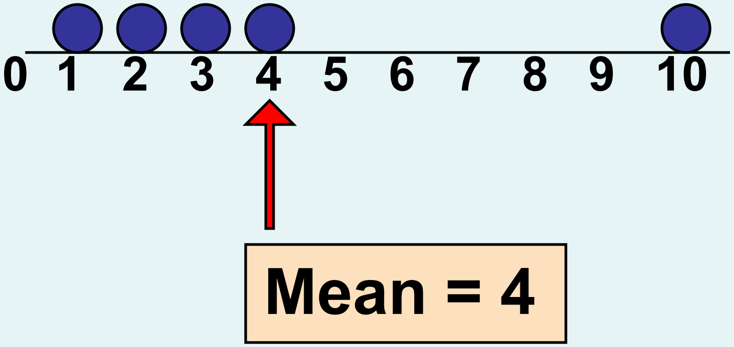
(continued)

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



Mean = 3

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

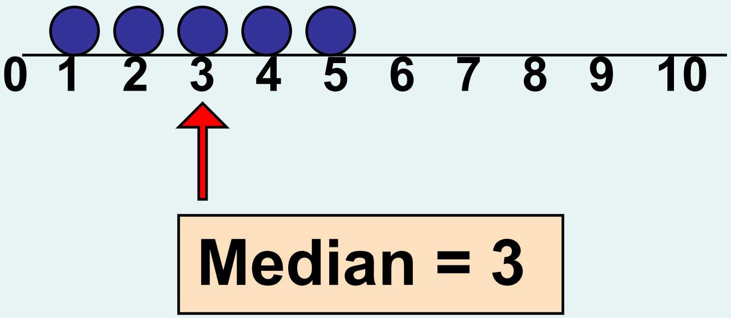


Mean = 4

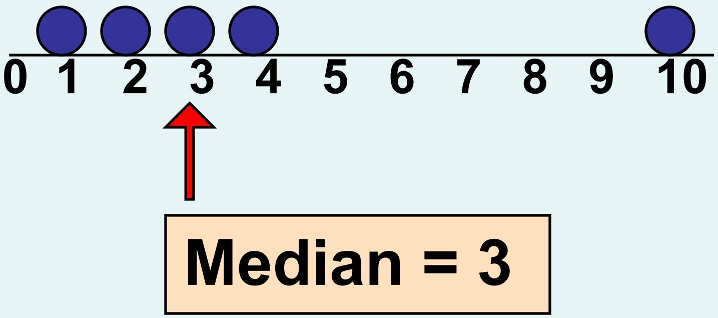
$$\frac{1+2+3+4+10}{5}=\frac{20}{5}=4$$

Measures of Central Tendency: The Median

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



Median = 3



Median = 3

- Not affected by extreme values

Measures of Central Tendency: Locating the Median

- The location of the median when the values are in numerical order (smallest to largest):

$$\boxed{\text{Median position} = \frac{n+1}{2} \text{ position in the ordered data}}$$

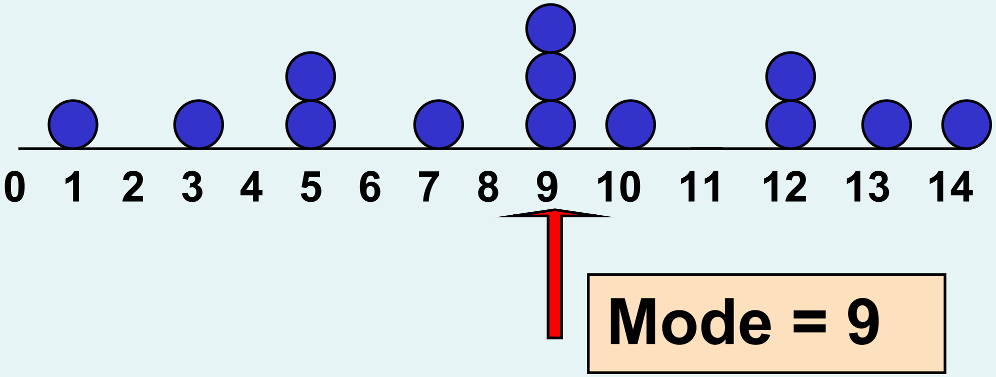
- If the number of values is odd, the median is the middle number

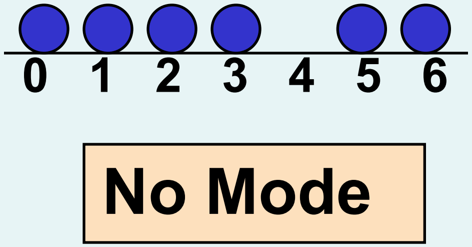
- If the number of values is even, the median is the average of the two middle numbers

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 (nominal) data
- There may 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

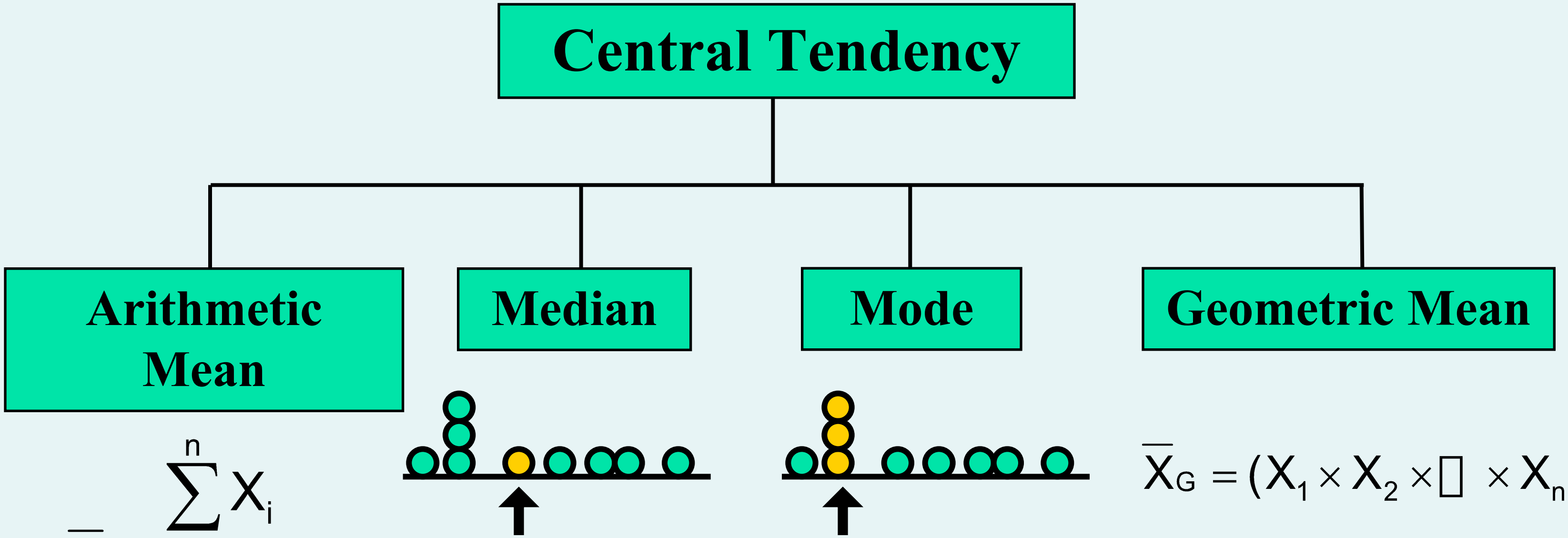
Sum **$3,000,000**

- **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.
- The **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.
- In some situations it makes sense to report both the **mean** and the **median**.

Measures of Central Tendency: Summary



Central Tendency

Arithmetic Mean

Median

Mode

Geometric Mean

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

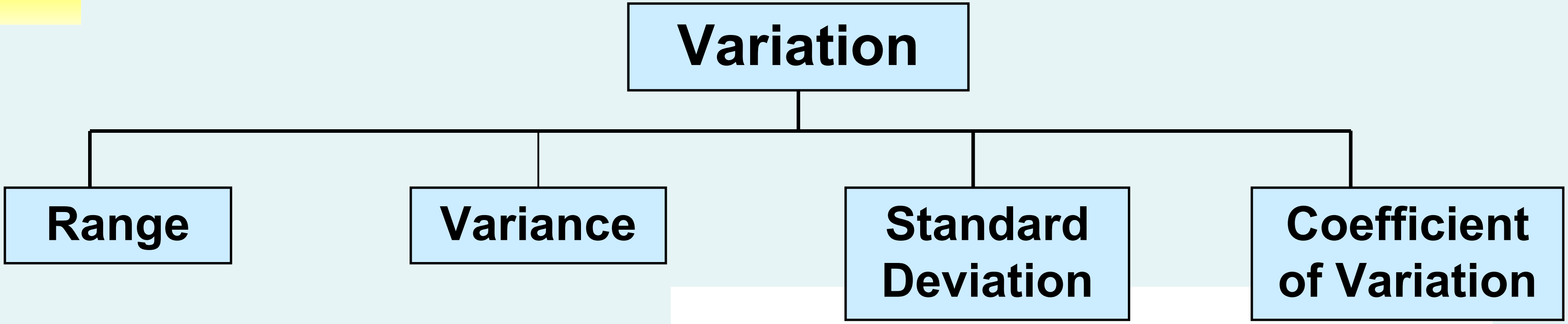
Middle value in the ordered array

Most frequently observed value

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

Rate of change of a variable over time

Measures of Variation



Variation

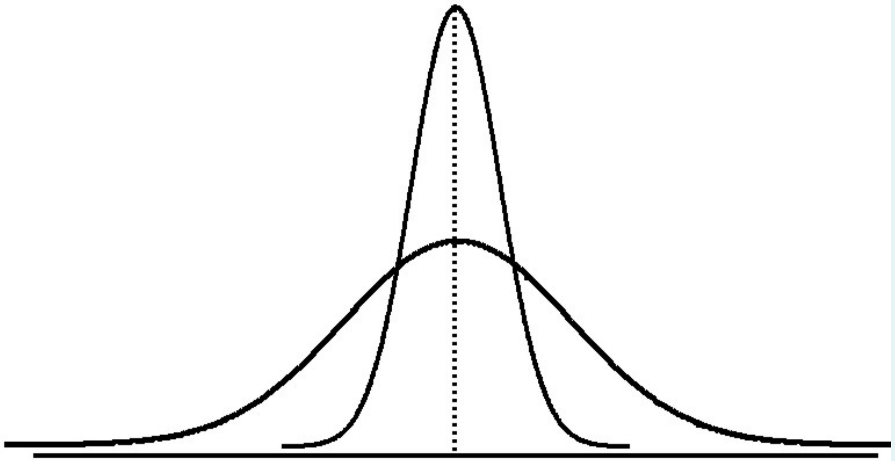
Range

Variance

Standard Deviation

Coefficient of Variation

- Measures of variation give information on the **spread** or **variability** or **dispersion** of the data values.



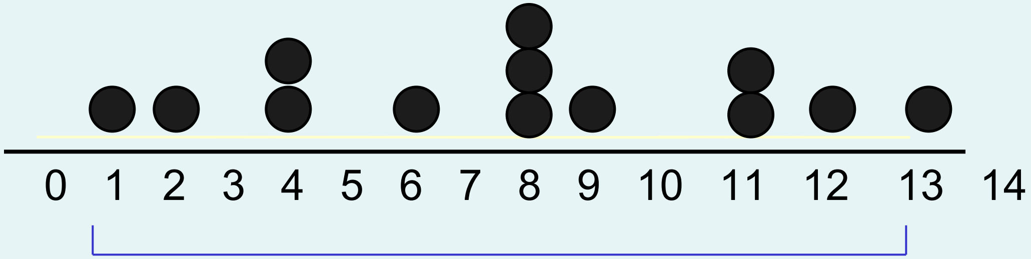
Same center, different variation

Measures of Variation: The Range

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

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

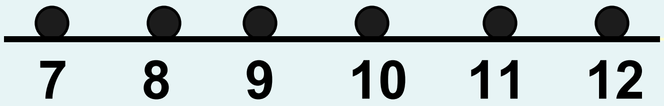
Example:



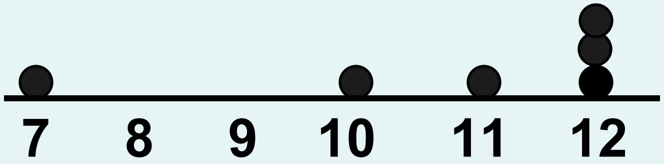
Range = 13 - 1 = 12

Measures of Variation: Why The Range Can Be Misleading

- 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: The Variance

- Average (approximately) of squared deviations of values from the mean
  - Sample variance:

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

Where $\bar{X}$ = arithmetic mean

n = sample size

X_i = ith value of the variable X

Measures of Variation: The Standard Deviation

- 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

  - Sample standard deviation:

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

Measures of Variation: The 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: Sample Standard Deviation: Calculation Example

Sample Data (X_i) :

10	**12**	**14**	**15**	**17**	**18**	**18**	**24**

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

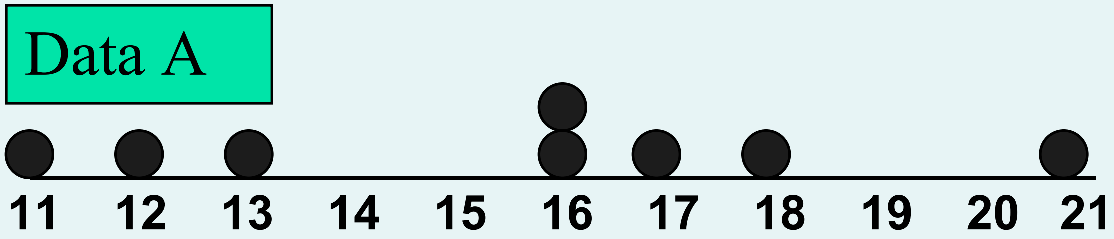
$$S = \sqrt{\frac{(10-\bar{X})^2 + (12-\bar{X})^2 + (14-\bar{X})^2 + \square + (24-\bar{X})^2}{n-1}}$$

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

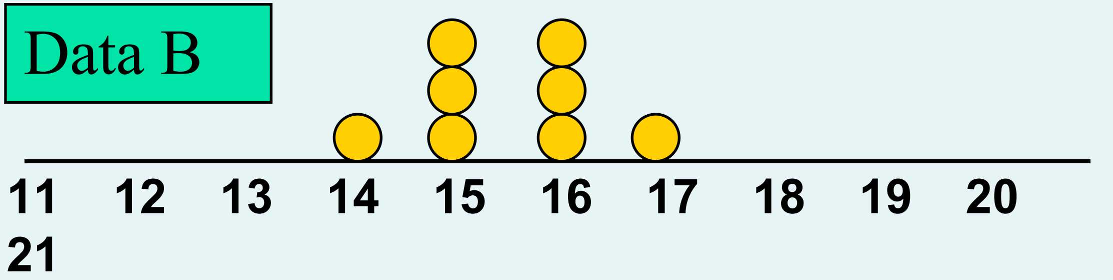
$$= \sqrt{\frac{130}{7}} = \boxed{4.3095}$$

→ A measure of the "average" scatter around the mean

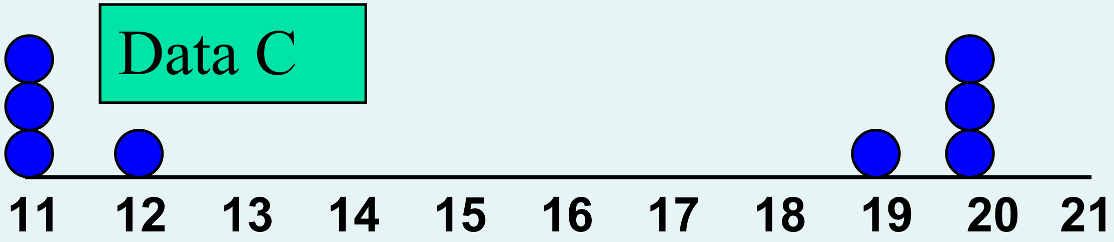
Measures of Variation: Comparing Standard Deviations


Data A

Mean = 15.5
S = 3.338

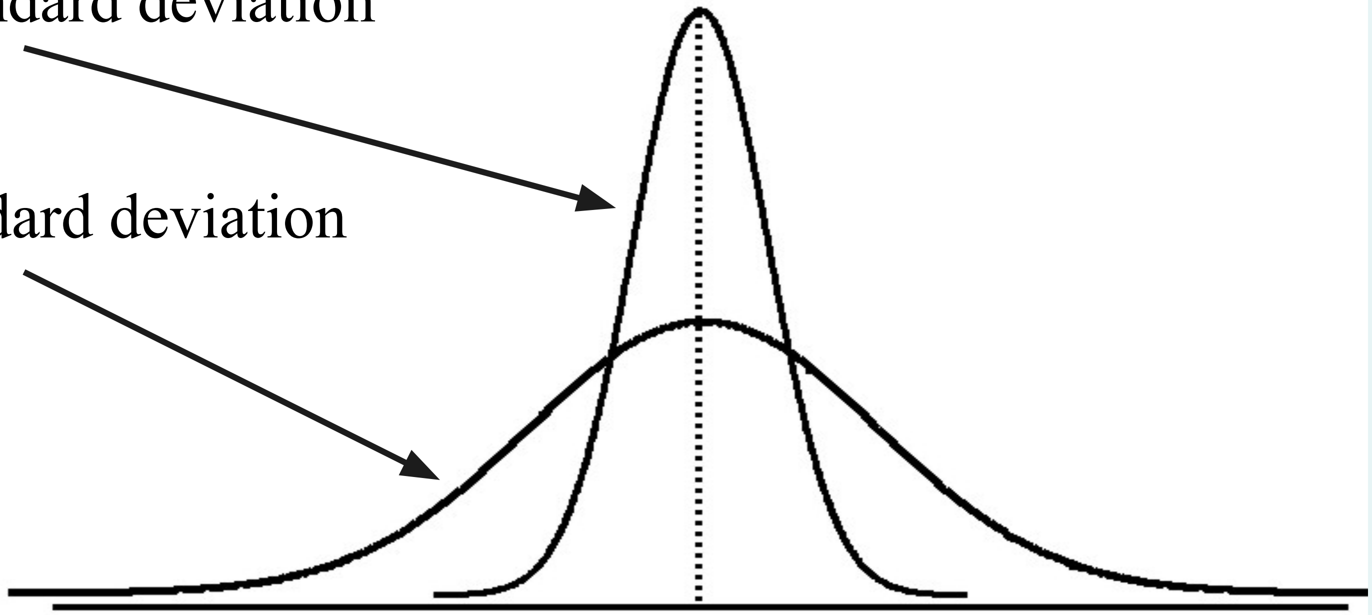

Data B

Mean = 15.5
S = 0.926


Data C

Mean = 15.5
S = 4.570

Measures of Variation: Comparing Standard Deviations



Smaller standard deviation

Larger standard deviation

Measures of 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.

Measures of Variation: The Coefficient of Variation

- Measures relative variation
- Always in percentage (%)
- Shows variation relative to mean
- Can be used to compare the variability of two or more sets of data measured in different units

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

Measures of Variation: Comparing Coefficients of Variation

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

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

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

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

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

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

where X represents the data value

$\overline{X}$ 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.

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

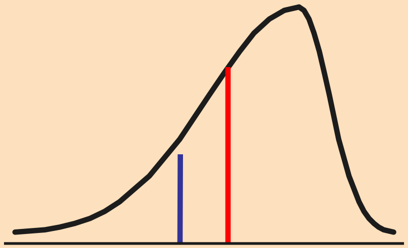
A score of 620 is 1.3 standard deviations above the mean and would not be considered an outlier.

Shape of a Distribution

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

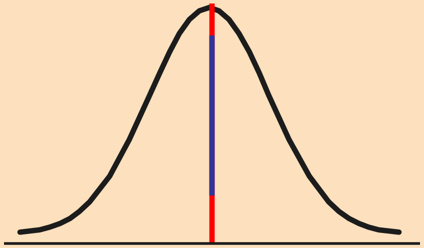
Left-Skewed

Mean < Median



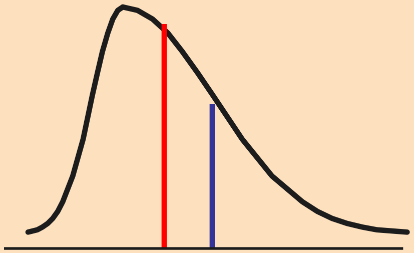
Symmetric

Mean = Median



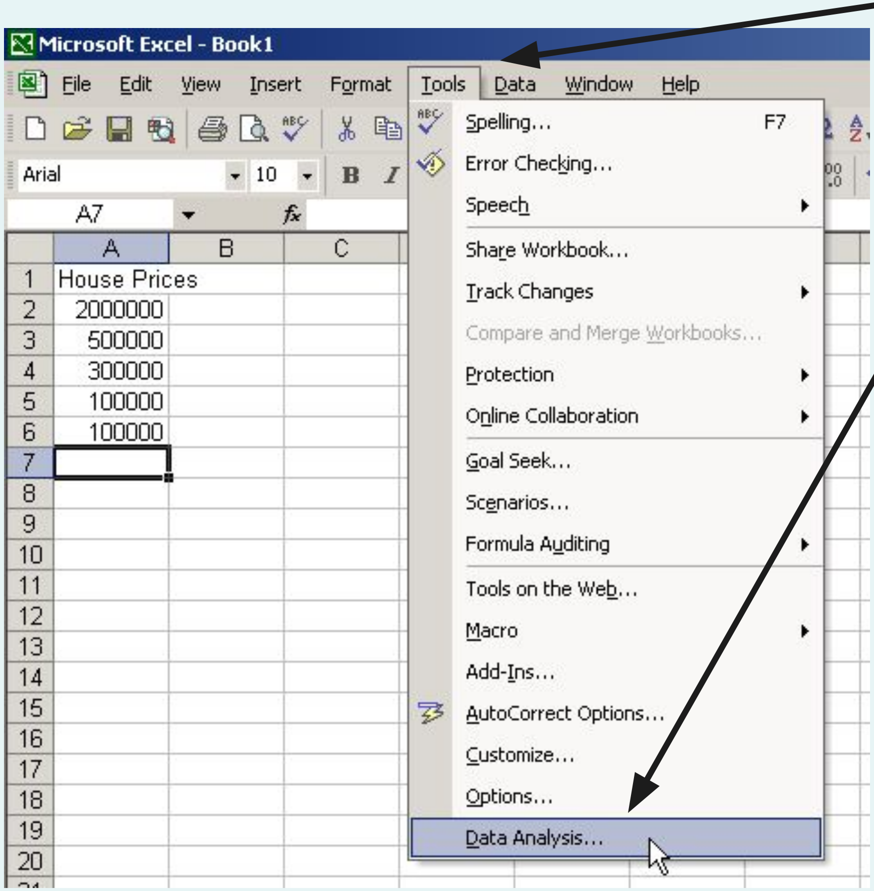
Right-Skewed

Median < Mean

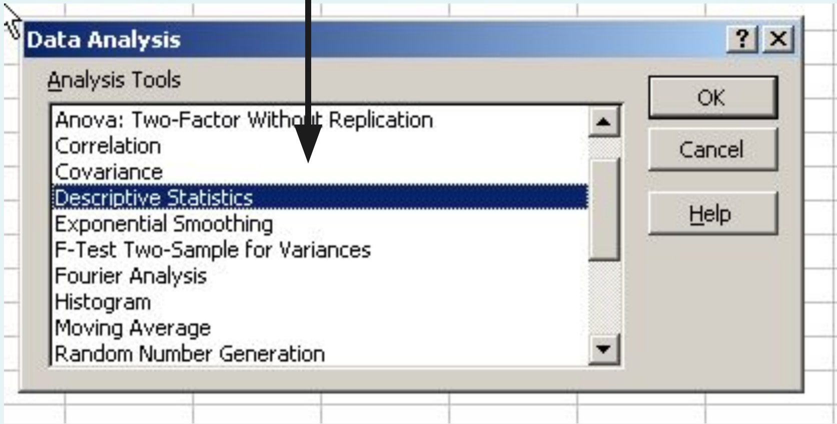


General Descriptive Stats Using Microsoft Excel

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



	A	B	C
1	House Prices		
2	2000000		
3	500000		
4	300000		
5	100000		
6	100000		
7			

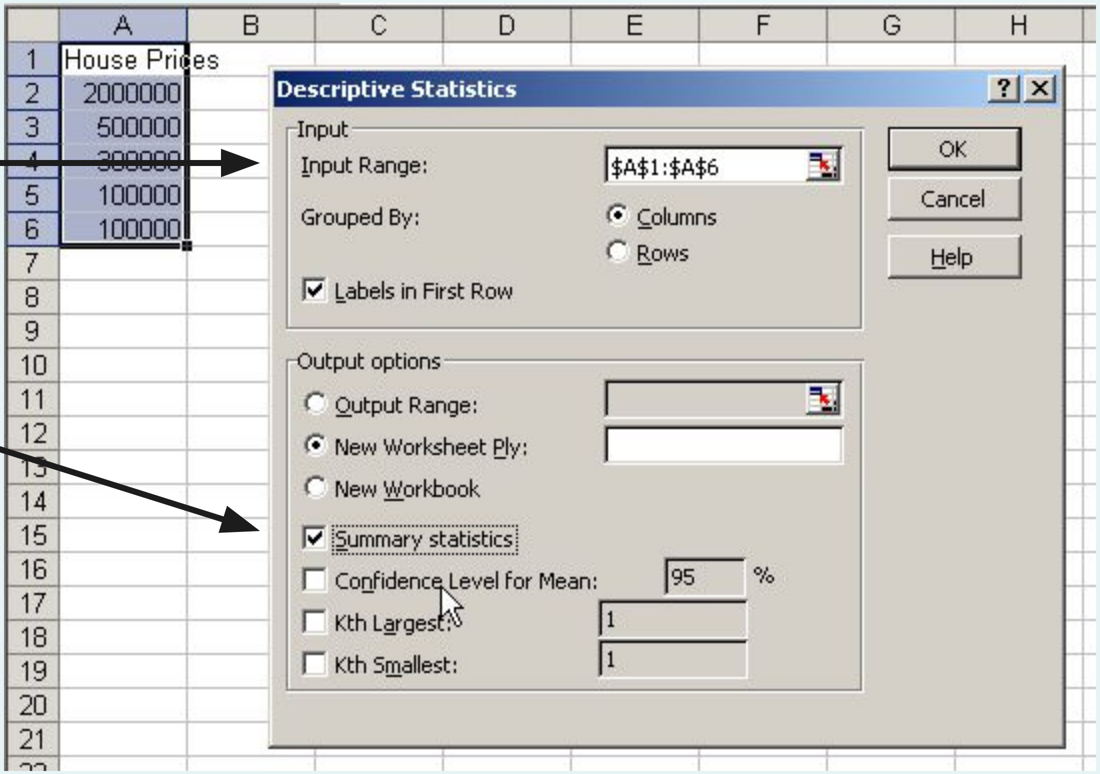


General Descriptive Stats Using Microsoft Excel

4. Enter the cell range.

5. Check the Summary Statistics box.

6. Click OK



House Prices
2000000
500000
300000
100000
100000

Descriptive Statistics

Input

Input Range: $A$1:$A$6

Grouped By: ● Columns ○ Rows

☑ Labels in First Row

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

☑ Summary statistics

□ Confidence Level for Mean: 95 %

□ Kth Largest: 1

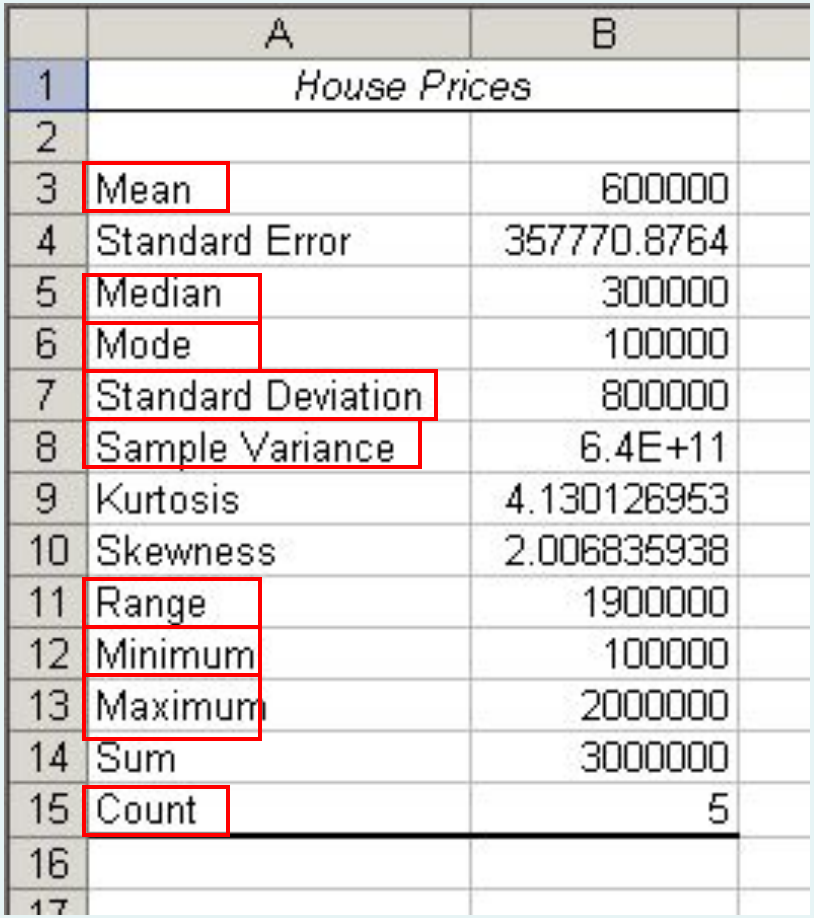
□ Kth Smallest: 1

Excel output

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		
17		

Minitab Output

```
Descriptive Statistics: House Price

                  Total
Variable          Count     Mean  SE Mean   StDev      Variance        Sum  Minimum
House Price           5   600000   357771  800000  6.40000E+11    3000000   100000

                                          N for
Variable           Median  Maximum     Range    Mode   Skewness  Kurtosis
House Price        300000  2000000   1900000  100000       2.01      4.13
```

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.

Numerical Descriptive Measures for a Population: The mean μ

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

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

Where

μ = population mean

N = population size

X_i = ith value of the variable X

Numerical Descriptive Measures For A Population: The Variance σ^2

- Average of squared deviations of values from the mean

  - Population variance:

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

Where μ = population mean

N = population size

X_i = ith value of the variable X

Numerical Descriptive Measures For A Population: The Standard Deviation σ

- 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

  - Population standard deviation:

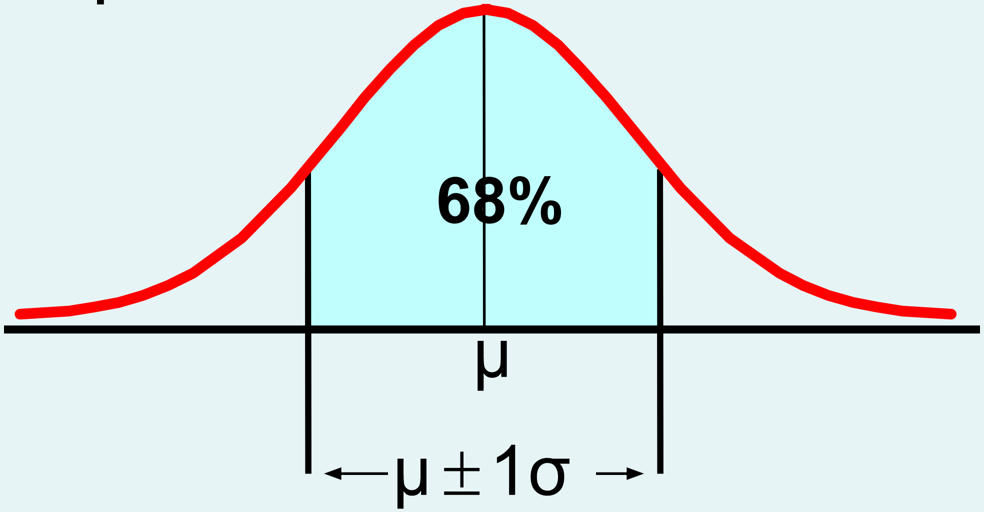
$$\boxed{\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}}$$

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

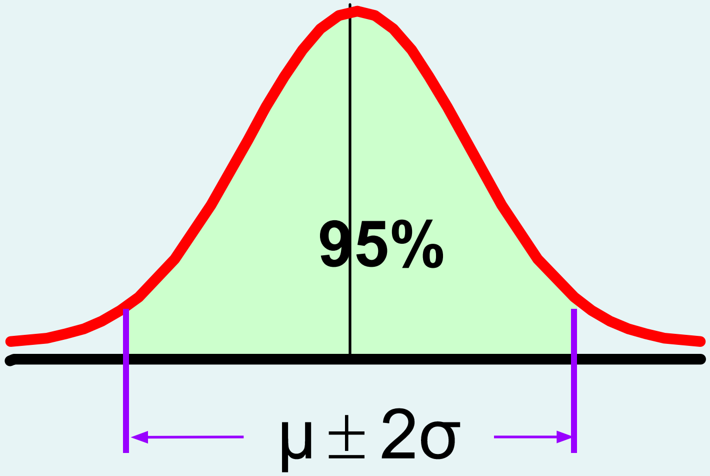
- The empirical rule approximates the variation of data in a bell-shaped distribution
- Approximately 68% of the data in a bell shaped distribution is within 1 standard deviation of the mean or $\mu \pm 1\sigma$

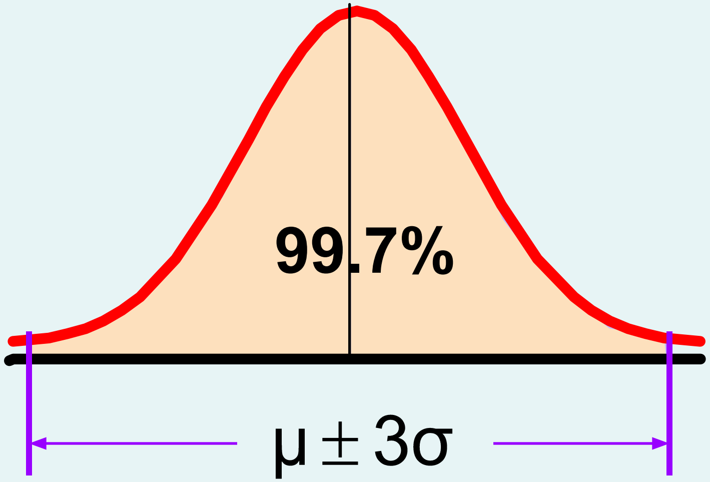


The Empirical Rule

- Approximately 95% of the data in a bell-shaped distribution lies within two standard deviations of the mean, or μ ± 2σ

- Approximately 99.7% of the data in a bell-shaped distribution lies within three standard deviations of the mean, or μ ± 3σ





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).

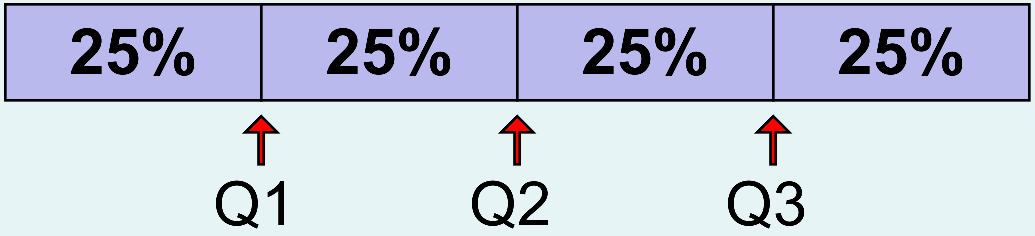
Chebyshev Rule

- Regardless of how the data are distributed, at least $(1 - 1/k^2) \times 100\%$ of the values will fall within k standard deviations of the mean (for $k > 1$)
  - Examples:

At least	within
$(1 - 1/2^2) \times 100\% = 75\%$	$k=2$ $(\mu \pm 2\sigma)$
$(1 - 1/3^2) \times 100\% = 89\%$	$k=3$ $(\mu \pm 3\sigma)$

Quartile Measures

- Quartiles split the ranked data into 4 segments with an equal number of values per segment



25% 25% 25% 25%

Q1 Q2 Q3

- 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% of the observations are smaller and 50% are larger)
- Only 25% of the observations 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: $Q_1 = (n+1)/4$ ranked value

Second quartile position: $Q_2 = (n+1)/2$ ranked value

Third quartile position: $Q_3 = 3(n+1)/4$ ranked value

where n is the number of observed values

Quartile Measures: Calculation Rules

- 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.

Quartile Measures: Locating Quartiles

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

(n = 9)

Q_1 is in the (9+1)/4 = 2.5 position of the ranked data

so use the value half way between the 2nd and 3rd values,

so **Q_1 = 12.5**

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

Quartile Measures
Calculating The Quartiles: Example

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

(n = 9)

Q_1 is in the (9+1)/4 = 2.5 position of the ranked data,

so **Q_1 = (12+13)/2 = 12.5**

Q_2 is in the (9+1)/2 = 5th position of the ranked data,

so **Q_2 = median = 16**

Q_3 is in the 3(9+1)/4 = 7.5 position of the ranked data,

so **Q_3 = (18+21)/2 = 19.5**

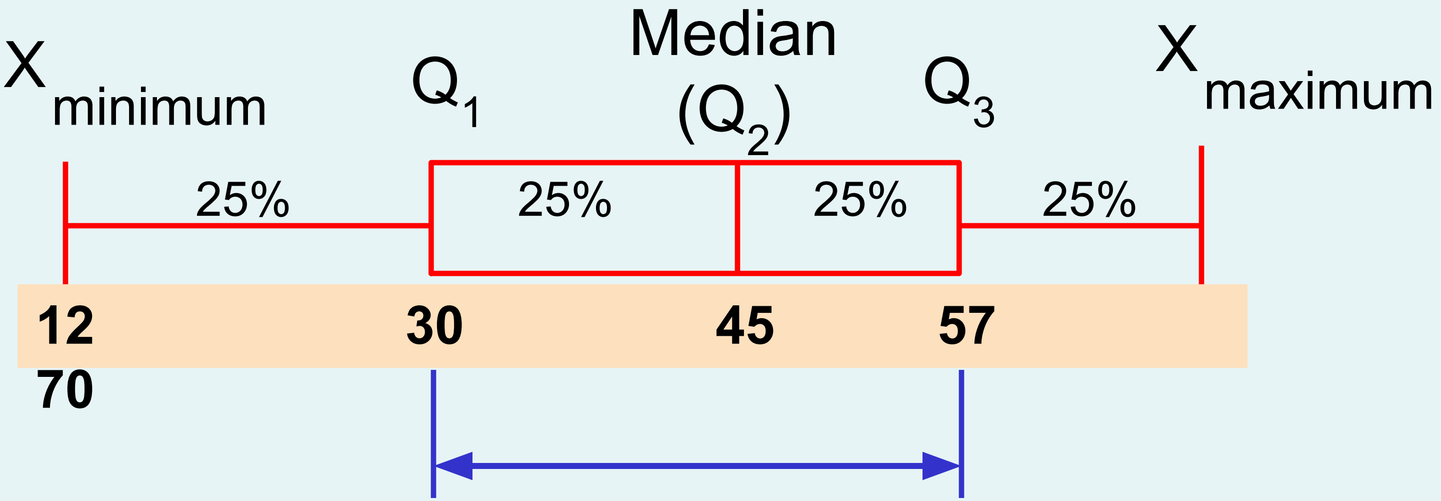
Q_1 and Q_3 are measures of non-central location
Q_2 = median, is a measure of central tendency

Quartile Measures: The Interquartile Range (IQR)

- The IQR is $Q_3 – Q_1$ and measures the spread in the middle 50% of the data
- The IQR is also called the midspread because it covers the middle 50% of the data
- The IQR is a measure of variability that is not influenced by outliers or extreme values
- Measures like Q_1, Q_3, and IQR that are not influenced by outliers are called resistant measures

Calculating The Interquartile Range

Example:



X_{minimum}

Q_1

Median (Q_2)

Q_3

X_{maximum}

25% 25% 25% 25%

12 30 45 57 70

Interquartile range
= 57 – 30 = 27

The Five Number Summary

The five numbers that help describe the center, spread and shape of data are:

- $X_{smallest}$
- First Quartile (Q_1)
- Median (Q_2)
- Third Quartile (Q_3)
- $X_{largest}$

Relationships among the five-number summary and distribution shape

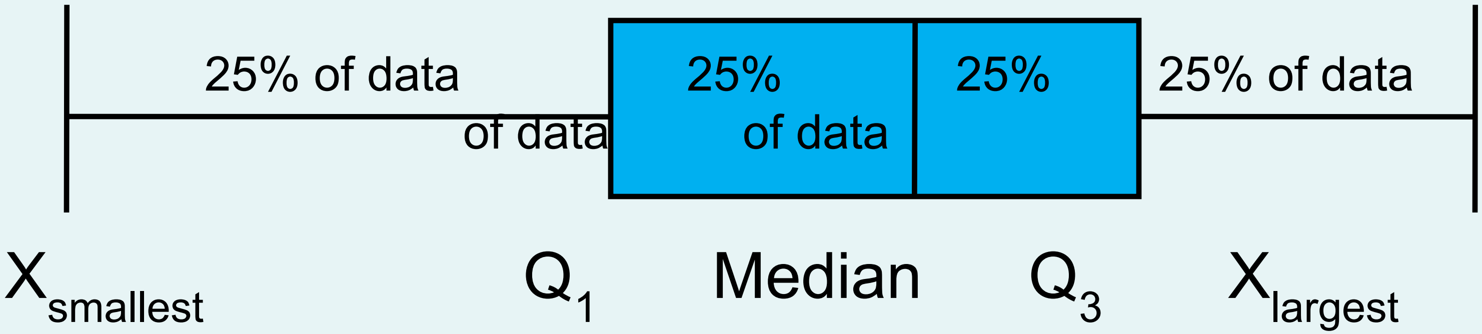
Left-Skewed	Symmetric	Right-Skewed
Median – $X_{smallest}$ $>$ $X_{largest}$ – Median	Median – $X_{smallest}$ $\approx$ $X_{largest}$ – Median	Median – $X_{smallest}$ $<$ $X_{largest}$ – Median
$Q_1 - X_{smallest}$ $>$ $X_{largest} - Q_3$	$Q_1 - X_{smallest}$ $\approx$ $X_{largest} - Q_3$	$Q_1 - X_{smallest}$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

Five Number Summary and The Boxplot

- The Boxplot: A Graphical display of the data based on the five-number summary:

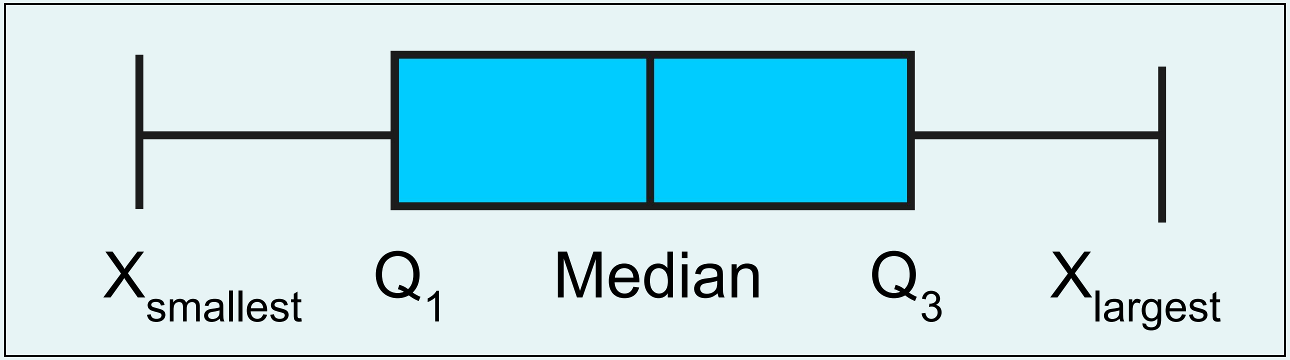
$X_{smallest}$ -- Q_1 -- Median -- Q_3 -- $X_{largest}$

Example:



Five Number Summary: Shape of Boxplots

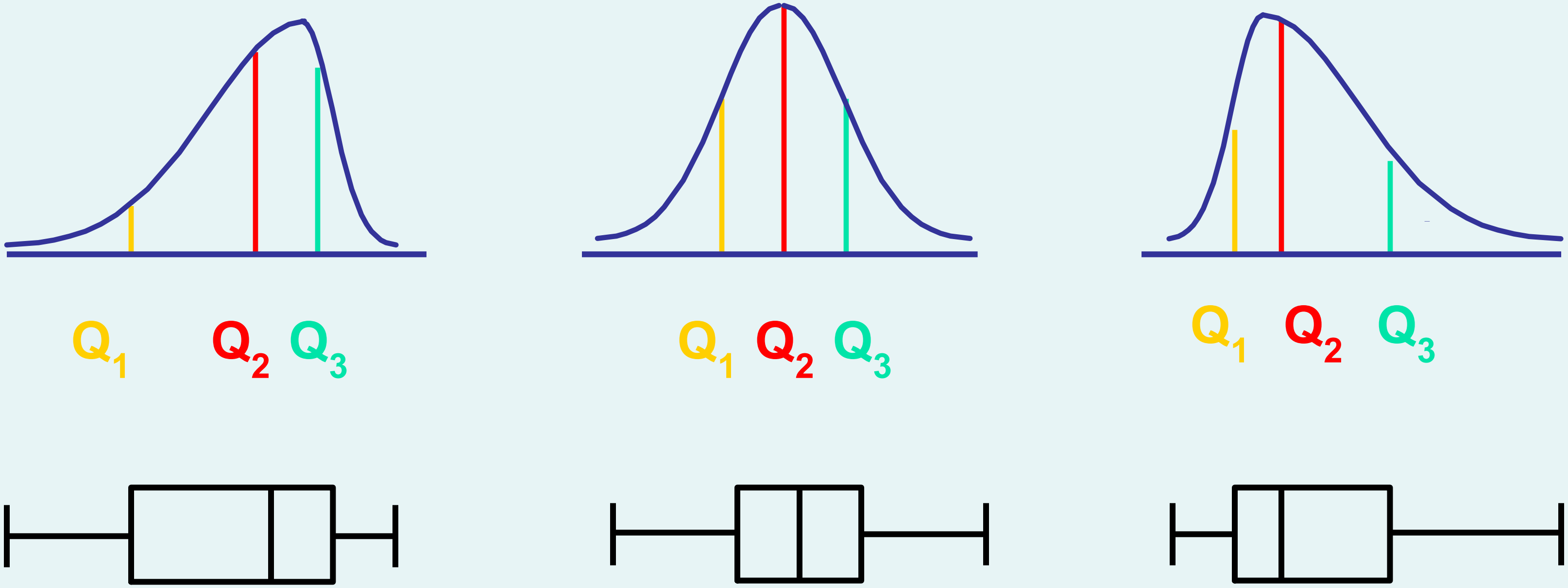
- If data are symmetric around the median then the box and central line are centered between the endpoints



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

Distribution Shape and The Boxplot

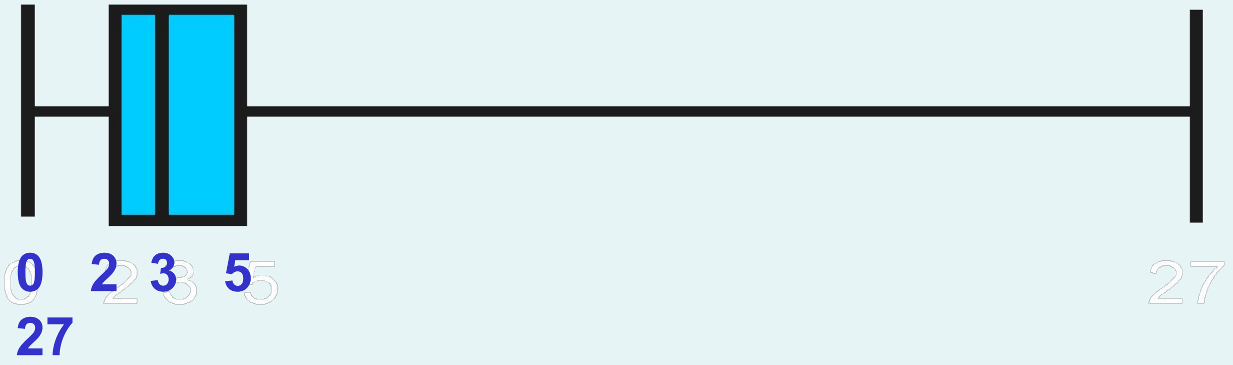
Left-Skewed Symmetric Right-Skewed



Boxplot Example

- Below is a Boxplot for the following data:

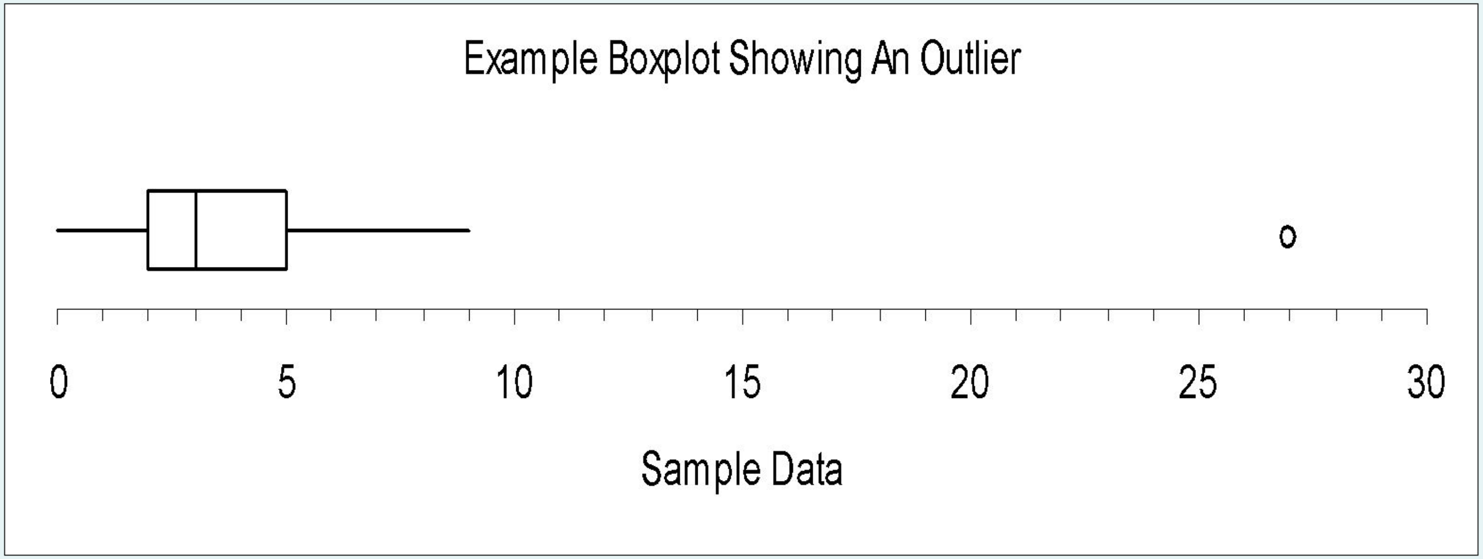
$X_{smallest}$		Q_1			Q_2			Q_3		$X_{largest}$
0	2	2	2	3	3	4	5	5	9	27



- The data are right skewed, as the plot depicts

Boxplot example showing an outlier

- The boxplot below of the same data shows the outlier value of 27 plotted separately
- A value is considered an outlier if it is more than 1.5 times the interquartile range below Q_1 or above Q_3



Chapter Summary

- Described measures of central tendency
  - Mean, median, mode, geometric mean
- Described measures of variation
  - Range, interquartile range, variance and standard deviation, coefficient of variation, Z-scores
- Illustrated shape of distribution
  - Symmetric, skewed
- Described data using the 5-number summary
  - Boxplots