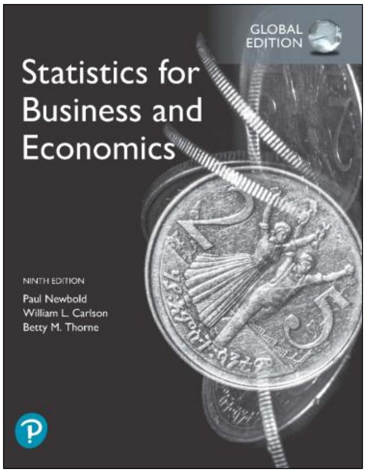
Statistics for Business and Economics

Ninth Edition, Global Edition



Chapter 2

Describing Data: Numerical

Chapter Goals

After completing this chapter, you should be able to:

- Compute and interpret the mean, median, and mode for a set of data
- Find the range, variance, standard deviation, and coefficient of variation and know what these values mean
- Apply the empirical rule to describe the variation of population values around the mean
- Explain the weighted mean and when to use it
- Explain how a least squares regression line estimates a linear relationship between two variables

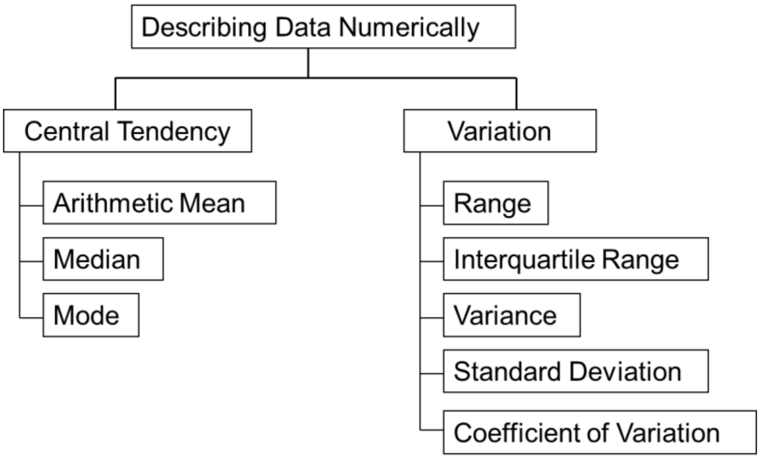
Chapter Topics (1 of 2)

- Measures of central tendency, variation, and shape
  - Mean, median, mode, geometric mean
  - Quartiles
  - Range, interquartile range, variance and standard deviation, coefficient of variation
  - Symmetric and skewed distributions
- Population summary measures
  - Mean, variance, and standard deviation
  - The empirical rule and Chebyshev’s Theorem

Chapter Topics (2 of 2)

- Five number summary and box-and-whisker plots
- Covariance and coefficient of correlation
- Pitfalls in numerical descriptive measures and ethical considerations

Describing Data Numerically



Describing Data Numerically

Central Tendency

Arithmetic Mean

Median

Mode

Variation

Range

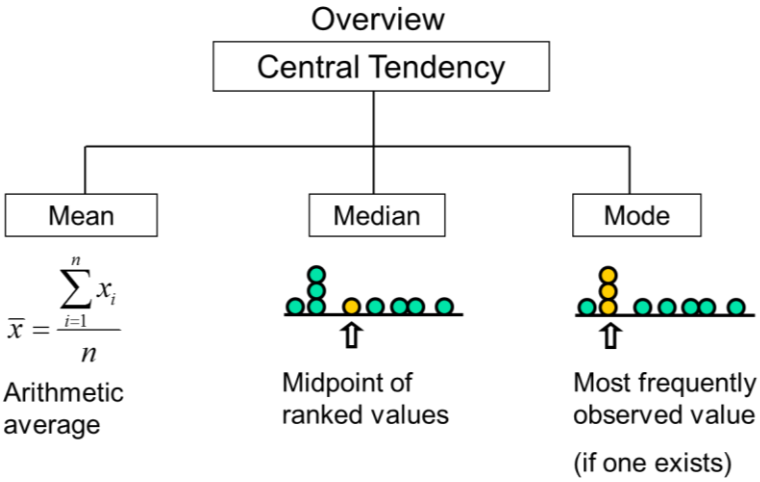
Interquartile Range

Variance

Standard Deviation

Coefficient of Variation

Section 2.1 Measures of Central Tendency



Overview

Central Tendency

Mean

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

Arithmetic average

Median

Midpoint of ranked values

Mode

Most frequently observed value

(if one exists)

Arithmetic Mean (1 of 2)

- The arithmetic mean (mean) is the most common measure of central tendency
  - For a population of *N* values:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

Population values

Population size

  - For a sample of size *n*:

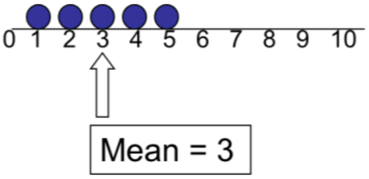
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Observed values

Sample size

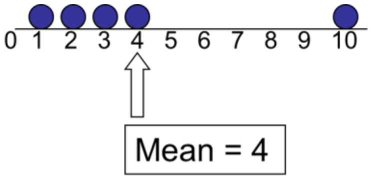
Arithmetic Mean (2 of 2)

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



Mean = 3

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

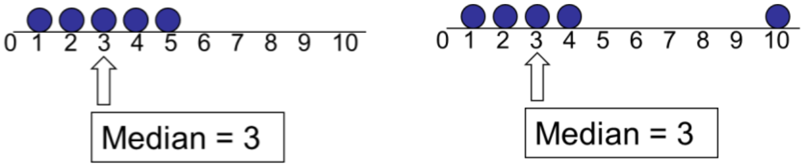


Mean = 4

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

Median

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



- Not affected by extreme values

Finding the Median

- The location of the median:

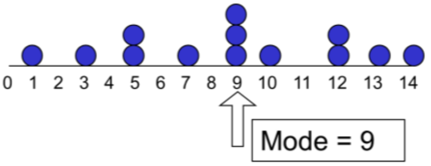
$$\text{Median position} = \left(\frac{n+1}{2}\right)^{\text{th}} \text{position in the ordered data}$$

  - 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

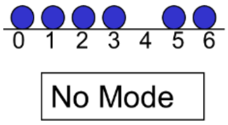
- Note that $\frac{n+1}{2}$ is not the value of the median, only the position of the median in the ranked data

Mode

- A measure of central tendency
- 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



Mode = 9



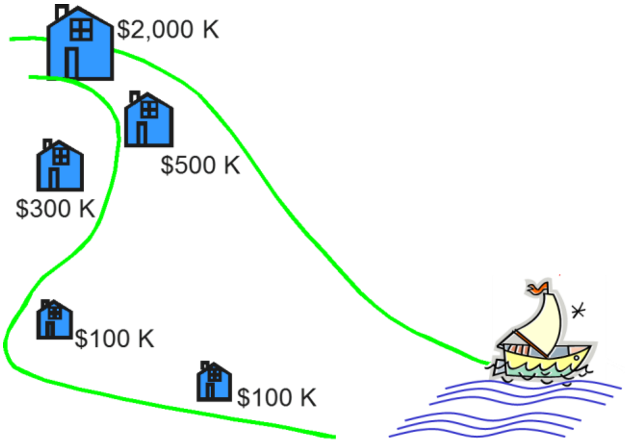
No Mode

Review Example

- Five houses on a hill by the beach

House Prices:

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



Review Example: Summary Statistics

House Prices:

$$
\begin{aligned}
&\$2{,}000{,}000\\
&500{,}000\\
&300{,}000\\
&100{,}000\\
&100{,}000\\
\hline
\text{Sum } &3{,}000{,}000
\end{aligned}
$$

- **Mean:** $\left(\dfrac{\$3{,}000{,}000}{5}\right)$

  = **$600,000**
- **Median:** middle value of ranked data

  = **$300,000**
- **Mode:** most frequent value

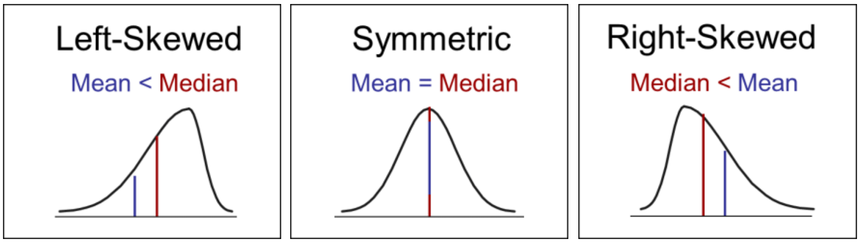
  = **$100,000**

Which Measure of Location Is the "Best"?

- **Mean** is generally used, unless extreme values (outliers) exist …
- Then **median** is often used, since the median is not sensitive to extreme values.
  - – Example: Median home prices may be reported for a region – less sensitive to outliers

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

Geometric Mean

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

$$\bar{x}_g = \sqrt[n]{(x_1 \times x_2 \times \cdots \times x_n)} = (x_1 \times x_2 \times \cdots \times x_n)^{\frac{1}{n}}$$

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

$$\bar{r}_g = (x_1 \times x_2 \times \ldots \times x_n)^{\frac{1}{n}} - 1$$

  - Where x_i is the rate of return in time period *i*

Example (1 of 2)

An investment of \$100,000 rose to \$150,000 at the end of year one and increased to \$180,000 at end of year two:

$$X_1 = \$100,000 \quad X_2 = \$150,000 \quad X_3 = \$180,000$$

50% increase 20% increase

What is the mean percentage return over time?

Example (2 of 2)

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

Arithmetic mean rate of return:

$$\overline{X} = \frac{(50\%)+(20\%)}{2} \boxed{= 35\%}$$

Misleading result

Geometric mean rate of return:

$$\overline{r}_g = (x_1 \times x_2)^{\frac{1}{n}} - 1$$

$$= \left[(50)\times(20)\right]^{\frac{1}{2}} - 1$$

$$= (1000)^{\frac{1}{2}} - 1 = 31.623 - 1 = \boxed{30.623\%}$$

Accurate result

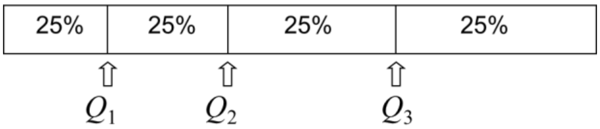
Percentiles and Quartiles

- Percentiles and Quartiles indicate the position of a value relative to the entire set of data
- Generally used to describe large data sets
- Example: An IQ score at the 90th percentile means that 10% of the population has a higher IQ score and 90% have a lower IQ score.

*P*th percentile = value located in the $\left(\frac{P}{100}\right)(n+1)^{\text{th}}$ ordered position

Quartiles (1 of 2)

- Quartiles split the ranked data into 4 segments with an equal number of values per segment (note that the widths of the segments may be different)



- 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 observations are greater than the third quartile

Quartile Formulas

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

First quartile position: $Q_1 = 0.25(n+1)$

Second quartile position: $Q_2 = 0.50(n+1)$
(the median position)

Third quartile position: $Q_3 = 0.75(n+1)$

where *n* is the number of observed values

Quartiles (2 of 2)

- Example: Find the first quartile

Sample Ranked Data: $11 \;\; 12 \;\; 13 \;\; 16 \;\; 16 \;\; 17 \;\; 18 \;\; 21 \;\; 22$

$(n=9)$ ⇧

$Q_1 =$ is in the $0.25(9+1) = 2.5$ position of the ranked data

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

so $Q_1 = 12.5$

Five-Number Summary

The **five-number summary** refers to five descriptive measures:

minimum

first quartile

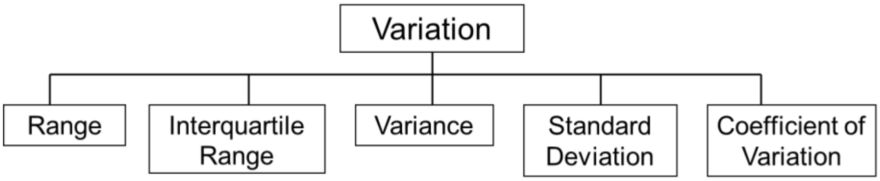
median

third quartile

maximum

$$\text{minimum} < Q_1 < \text{median} < Q_3 < \text{maximum}$$

Section 2.2 Measures of Variability



Variation

Range

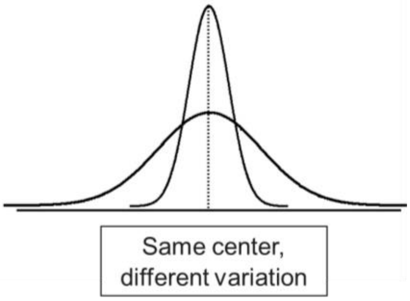
Interquartile Range

Variance

Standard Deviation

Coefficient of Variation

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

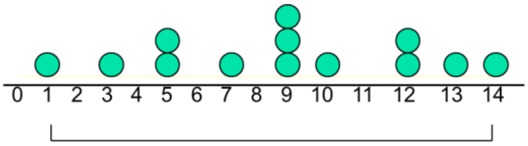


Range

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

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

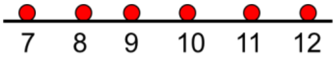
Example:



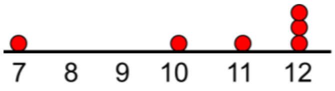
Range = 14 − 1 = 13

Disadvantages of the Range

- Ignores the way in which data are distributed



Range = 12 − 7 = 5



Range = 12 − 7 = 5

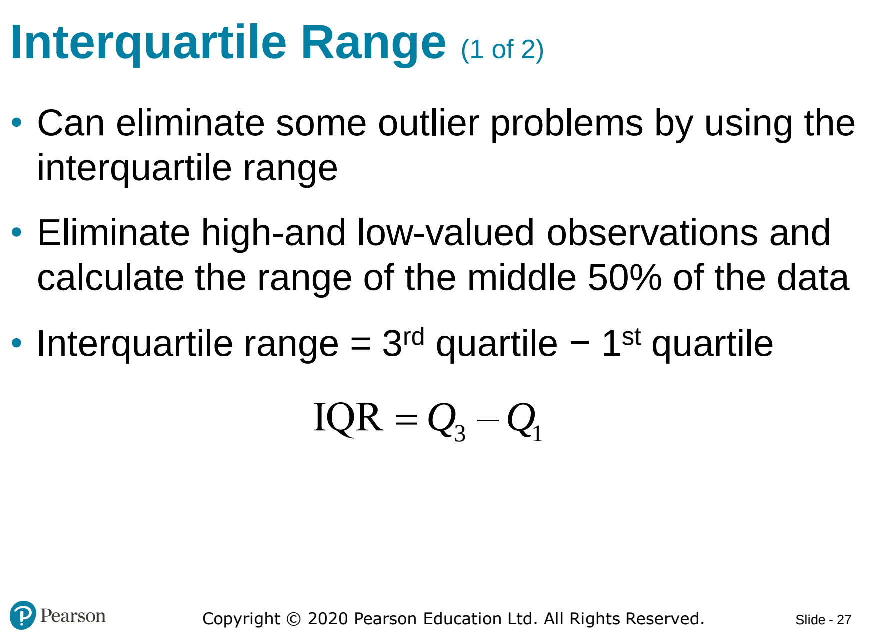
- Sensitive to outliers

$$1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,5$$

Range = 5 − 1 = 4

$$1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,120$$

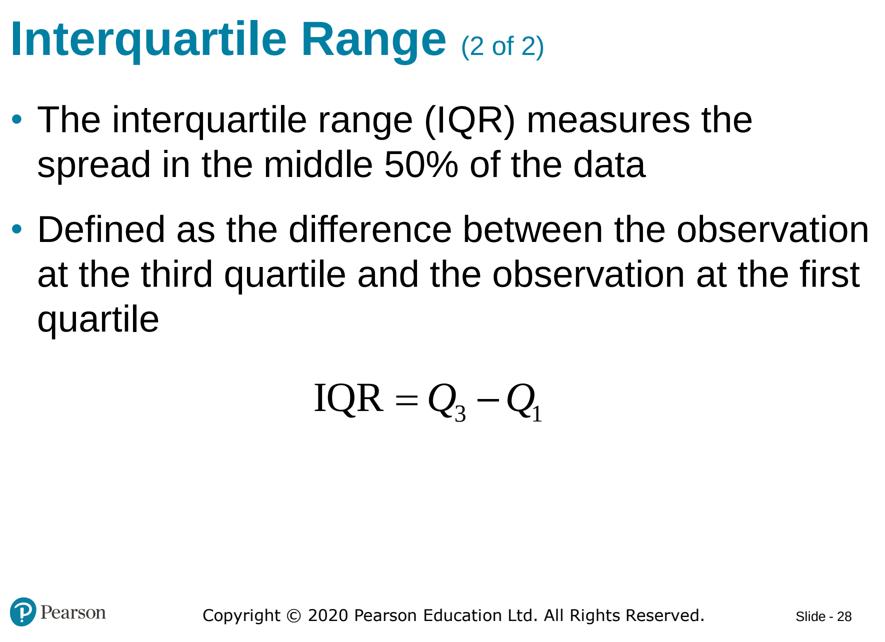
Range = 120 − 1 = 119



Interquartile Range (1 of 2)

- Can eliminate some outlier problems by using the interquartile range
- Eliminate high-and low-valued observations and calculate the range of the middle 50% of the data
- Interquartile range = 3^{rd} quartile − 1^{st} quartile

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



Interquartile Range (2 of 2)

- The interquartile range (IQR) measures the spread in the middle 50% of the data
- Defined as the difference between the observation at the third quartile and the observation at the first quartile

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

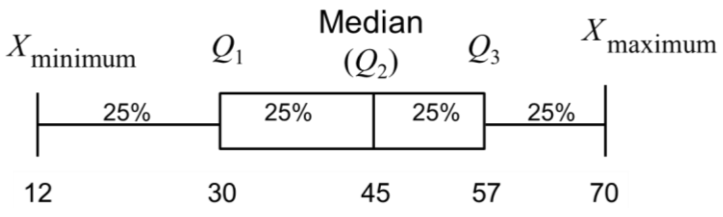
Box-and-Whisker Plot (1 of 2)

- A box-and-whisker plot is a graph that describes the shape of a distribution
- Created from the five-number summary: the minimum value, Q_1, the median, Q_3, and the maximum
- The inner box shows the range from Q_1 to Q_3, with a line drawn at the median
- Two "whiskers" extend from the box. One whisker is the line from Q_1 to the minimum, the other is the line from Q_3 to the maximum value

Box-and-Whisker Plot (2 of 2)

The plot can be oriented horizontally or vertically

Example:



Population Variance

- Average of squared deviations of values from the mean
  - Population variance: $$\sigma^2 = \frac{\sum_{i=1}^{N}\left(x_i - \mu\right)^2}{N}$$

Where μ = population mean

N = population size

$x_i = i^{\text{th}}$ value of the variable x

Sample Variance

- Average (approximately) of squared deviations of values from the mean
  - Sample variance: $$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

Where $\overline{x}$ = arithmetic mean

n = sample size

$x_i = i^{\text{th}}$ value of the variable x

Population Standard Deviation

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

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

Sample Standard Deviation

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

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

Calculation Example: Sample Standard Deviation

Sample Data $\left(x_i\right)$:

10	12	14	15	17	18	18	24

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

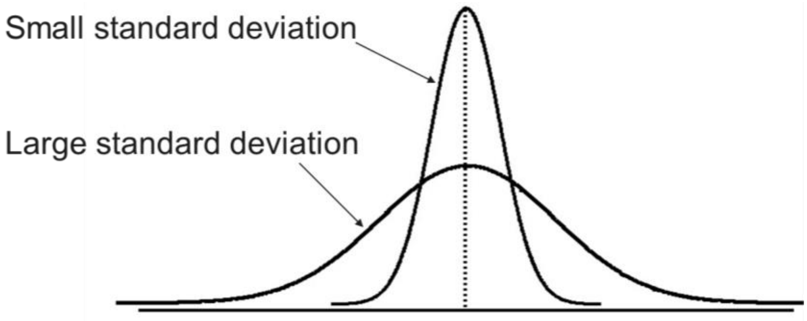
$$s = \sqrt{\frac{(10-\bar{x})^2 + (12-\bar{x})^2 + (14-\bar{x})^2 + \cdots + (24-\bar{x})^2}{n-1}}$$

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

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

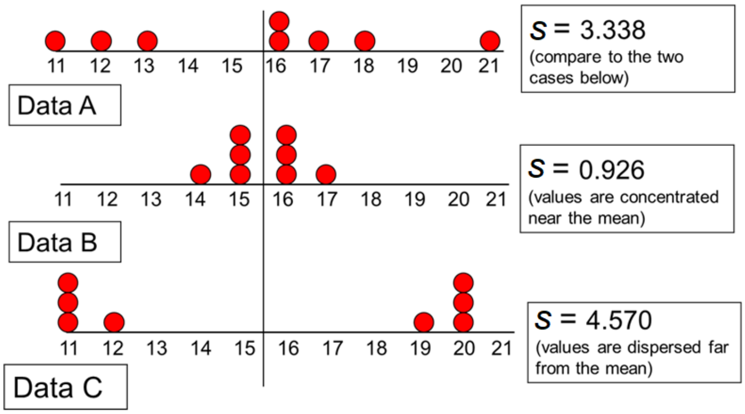
A measure of the "average" scatter around the mean

Measuring Variation



Comparing Standard Deviations

Mean = 15.5 for each data set



s = 3.338
(compare to the two cases below)

s = 0.926
(values are concentrated near the mean)

s = 4.570
(values are dispersed far from the mean)

Advantages of Variance and Standard Deviation

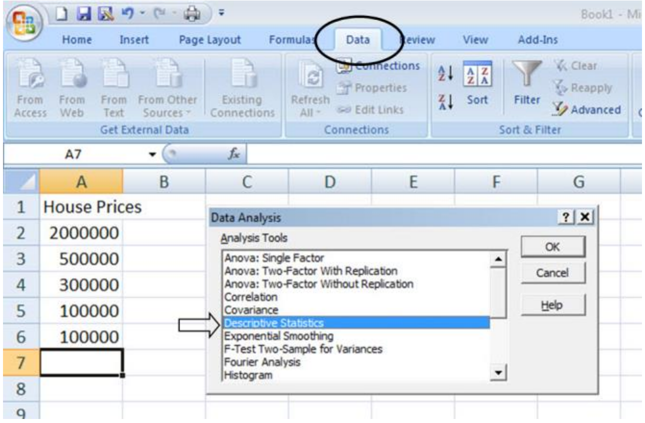
- Each value in the data set is used in the calculation
- Values far from the mean are given extra weight (because deviations from the mean are squared)

Using Microsoft Excel

- Descriptive Statistics can be obtained from Microsoft® Excel
  - Select:
    data/data analysis/descriptive statistics
  - Enter details in dialog box

Using Excel (1 of 2)

- Select data/data analysis/descriptive statistics



	A
1	House Prices
2	2000000
3	500000
4	300000
5	100000
6	100000
7	
8	
9	

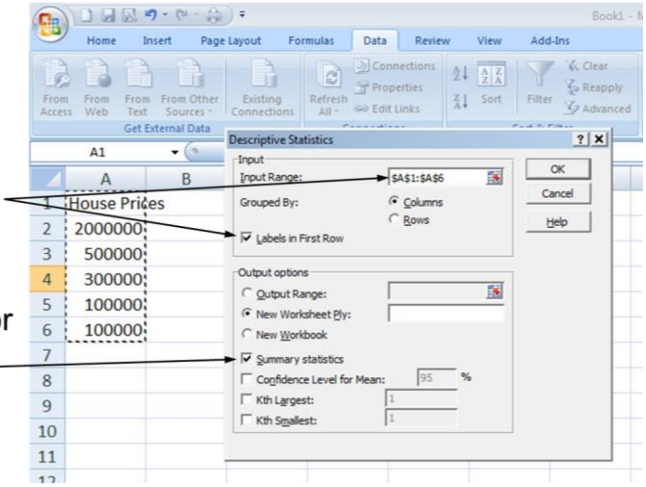
Data Analysis

Analysis Tools

Anova: Single Factor
Anova: Two-Factor With Replication
Anova: Two-Factor Without Replication
Correlation
Covariance
Descriptive Statistics
Exponential Smoothing
F-Test Two-Sample for Variances
Fourier Analysis
Histogram

Using Excel (2 of 2)

- Enter input range details
- Check box for summary statistics
- Click OK



	A	B
1	House Prices	
2	2000000	
3	500000	
4	300000	
5	100000	
6	100000	
7		
8		
9		
10		
11		

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

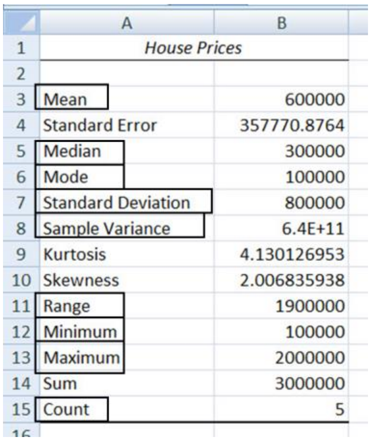
Excel output

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

Coefficient of Variation

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

Population coefficient of variation:

$$CV = \left(\frac{\sigma}{\mu}\right)\cdot 100\%$$

Sample coefficient of variation:

$$CV = \left(\frac{s}{\overline{x}}\right)\cdot 100\%$$

Comparing 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

Chebychev’s Theorem (1 of 2)

- For any population with mean μ and standard deviation σ, and $k > 1$, the percentage of observations that fall within the interval

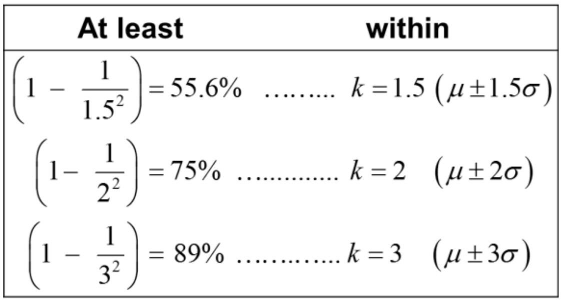
$$[\mu + k\sigma]$$

Is at least

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

Chebychev’s Theorem (2 of 2)

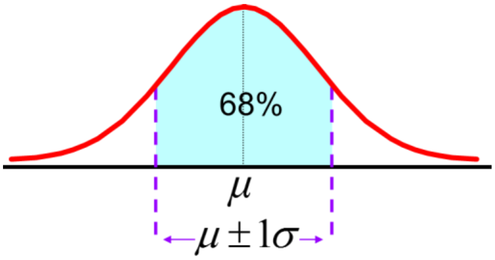
- Regardless of how the data are distributed, at least $\left(1-\frac{1}{k^2}\right)$ of the values will fall within k standard deviations of the mean (for $k > 1$)
  - Examples:



At least	within
$\left(1 - \frac{1}{1.5^2}\right) = 55.6\%$	$k = 1.5\ (\mu \pm 1.5\sigma)$
$\left(1 - \frac{1}{2^2}\right) = 75\%$	$k = 2\ (\mu \pm 2\sigma)$
$\left(1 - \frac{1}{3^2}\right) = 89\%$	$k = 3\ (\mu \pm 3\sigma)$

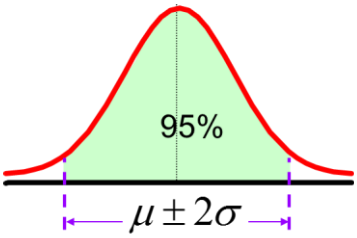
The Empirical Rule (1 of 2)

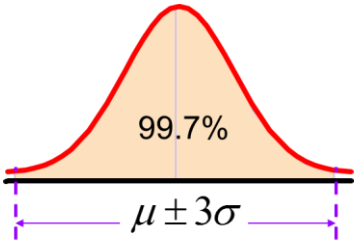
- If the data distribution is bell-shaped, then the interval:
- $\mu \pm 1\sigma$ contains about 68% of the values in the population or the sample



The Empirical Rule (2 of 2)

- $\mu \pm 2\sigma$ contains about 95% of the values in the population or the sample
- $\mu \pm 3\sigma$ contains almost all (about 99.7%) of the values in the population or the sample





z-Score (1 of 3)

A z-score shows the position of a value relative to the mean of the distribution.

- indicates the number of standard deviations a value is from the mean.
  - A z-score greater than zero indicates that the value is greater than the mean
  - a z-score less than zero indicates that the value is less than the mean
  - a z-score of zero indicates that the value is equal to the mean.

z-Score (2 of 3)

- If the data set is the entire population of data and the population mean, μ, and the population standard deviation, σ, are known, then for each value, x_i, the z-score associated with x_i is

$$z = \frac{x_i - \mu}{\sigma}$$

z-Score (3 of 3)

- If intelligence is measured for a population using an IQ score, where the mean IQ score is 100 and the standard deviation is 15, what is the z-score for an IQ of 121?

$$z = \frac{x_i - \mu}{\sigma} = \frac{121 - 100}{15} = 1.4$$

A score of 121 is 1.4 standard deviations above the mean.

Section 2.3 Weighted Mean and Measures of Grouped Data

- The weighted mean of a set of data is

$$\bar{x} = \frac{\sum_{\mathrm{i}=1}^{n} w_i x_i}{n} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{n}$$

  - Where w_i is the weight of the i^{th} observation and $n = \sum w_i$
- Use when data is already grouped into *n* classes, with w_i values in the i^{th} class

Approximations for Grouped Data (1 of 2)

Suppose data are grouped into *K* classes, with frequencies $f_1, f_2, \ldots, f_K,$ and the midpoints of the classes are $m_1, m_2, \ldots, m_K$

- For a sample of *n* observations, the mean is

$$\overline{x} = \frac{\sum_{i=1}^{K} f_i m_i}{n} \qquad \text{where} \quad n = \sum_{i=1}^{K} f_i$$

Pearson

Approximations for Grouped Data (2 of 2)

Suppose data are grouped into *K* classes, with frequencies $f_1, f_2, \ldots, f_K,$ and the midpoints of the classes are $m_1, m_2, \ldots, m_K$

- For a sample of *n* observations, the variance is

$$s^2 = \frac{\sum_{i=1}^{K} f_i (m_i - \overline{x})^2}{n-1}$$

Pearson

Section 2.4 Measures of Relationships Between Variables

Two measures of the relationship between variable are

- Covariance
  - a measure of the direction of a linear relationship between two variables
- Correlation Coefficient
  - a measure of both the direction and the strength of a linear relationship between two variables

Covariance

- The covariance measures the strength of the linear relationship between two variables
- The population covariance:

$$\mathrm{Cov}(x,y)=\sigma_{xy}=\frac{\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y)}{N}$$

- The sample covariance:

$$\mathrm{Cov}(x,y)=s_{xy}=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{n-1}$$

  - Only concerned with the strength of the relationship
  - No causal effect is implied

Interpreting Covariance

- **Covariance** between two variables:

$\text{Cov}(x, y) > 0 \rightarrow$ *x* and *y* tend to move in the same direction

$\text{Cov}(x, y) < 0 \rightarrow$ *x* and *y* tend to move in opposite directions

$\text{Cov}(x, y) = 0 \rightarrow$ *x* and *y* are independent

Coefficient of Correlation

- Measures the relative strength of the linear relationship between two variables

- Population correlation coefficient:

$$\rho = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y}$$

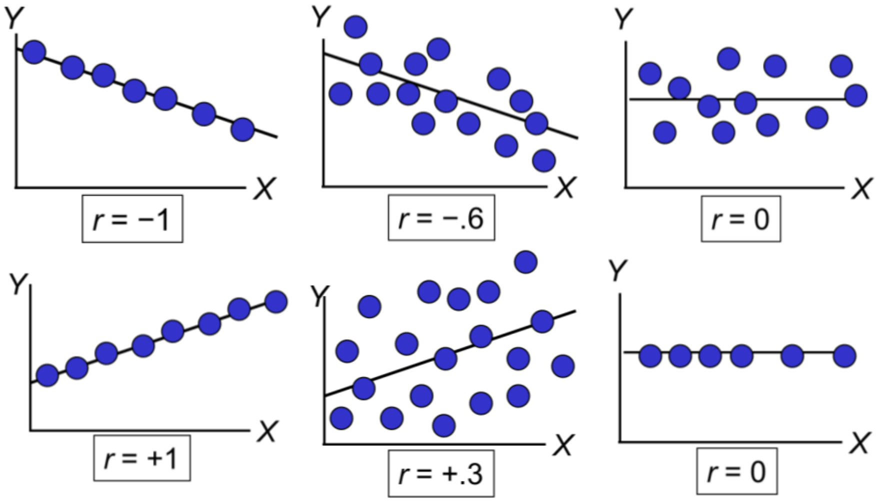
- Sample correlation coefficient:

$$r = \frac{\text{Cov}(x, y)}{s_x s_y}$$

Features of Correlation Coefficient, *r*

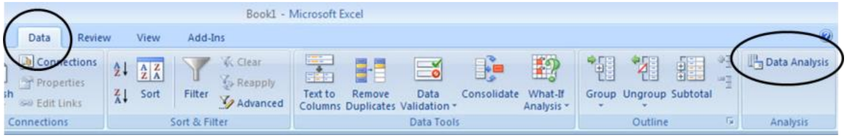
- Unit free
- Ranges between −1 and 1
- The closer to −1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker any positive linear relationship

Scatter Plots of Data with Various Correlation Coefficients

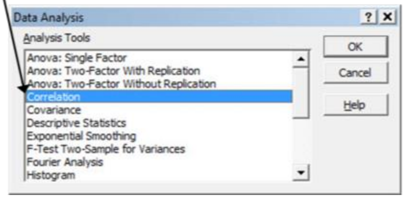


Using Excel to Find the Correlation Coefficient (1 of 2)

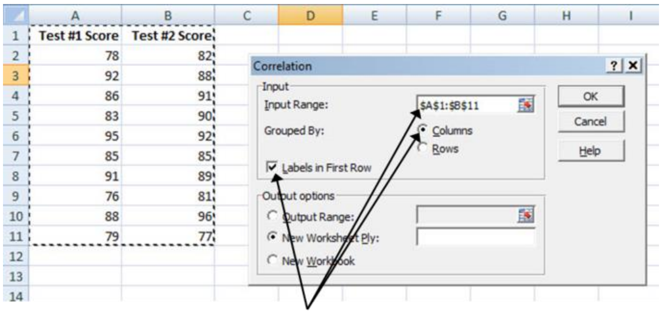
- Select Data/Data Analysis



- Choose Correlation from the selection menu
- Click OK . . .



Using Excel to Find the Correlation Coefficient (2 of 2)



	A	B
1	Test #1 Score	Test #2 Score
2	78	82
3	92	88
4	86	91
5	83	90
6	95	92
7	85	85
8	91	89
9	76	81
10	88	96
11	79	77
12		
13		
14		

Correlation

Input

Input Range: $A$1:$B$11

Grouped By: ● Columns ○ Rows

☑ Labels in First Row

Output options

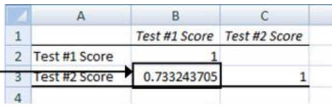
○ Output Range:

● New Worksheet Ply:

○ New Workbook

- Input data range and select appropriate options

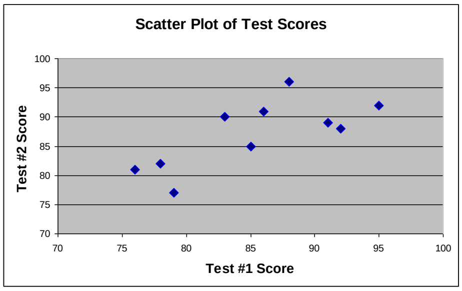
- Click OK to get output



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

Interpreting the Result

- *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

Pearson

Chapter Summary

- Described measures of central tendency
  - Mean, median, mode
- Illustrated the shape of the distribution
  - Symmetric, skewed
- Described measures of variation
  - Range, interquartile range, variance and standard deviation, coefficient of variation
- Discussed measures of grouped data
- Calculated measures of relationships between variables
  - covariance and correlation coefficient

Pearson