Co-financed by the Connecting Europe Facility of the European Union

Describing Data: Numerical

Data Driven Healthcare

Module B

Prof. Paola Cerchiello – University of Pavia

Goals

- 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

Topics (1 of 2)

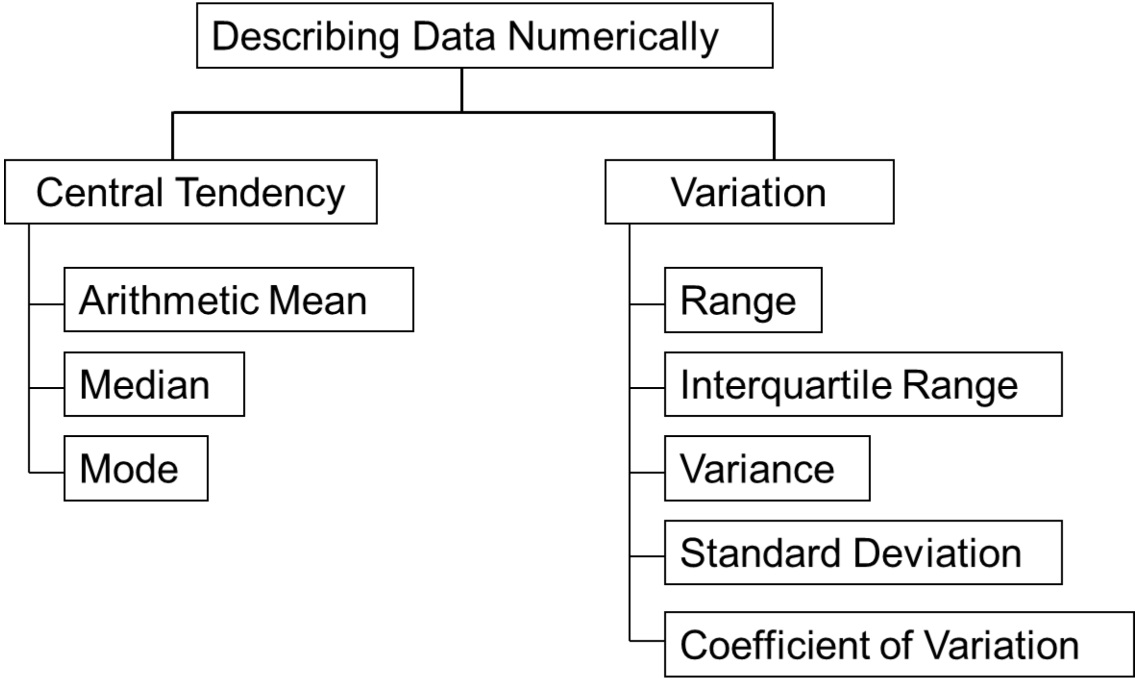
- Measures of central tendency, variation, and shape
  - Mean, median, mode, geometric mean
  - Quartiles
  - Range, interquartile range, variance and standard deviation, coefficient of variation
  - Symmetric and skewed distributions

- 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



Measures of Central Tendency

Overview

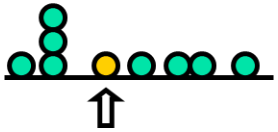
Central Tendency

- **Mean**

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

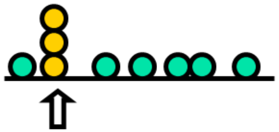
Arithmetic average

- **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*:

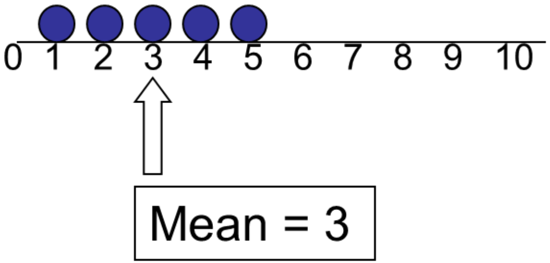
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Observed values

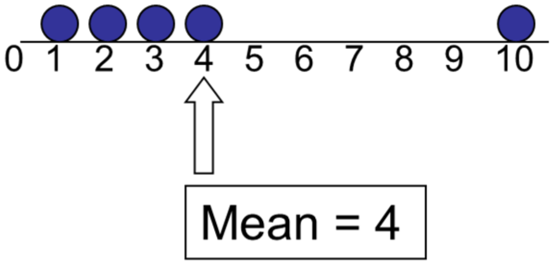
Sample size

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)



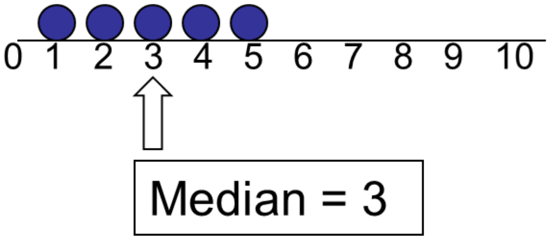
$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

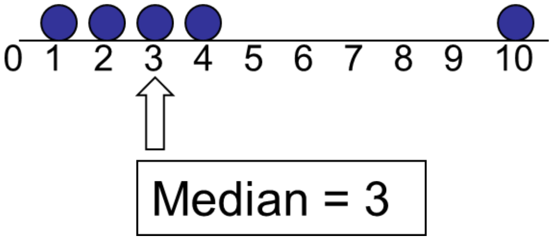


$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

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:

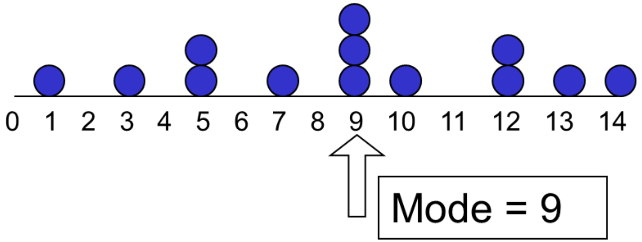
$$\text{Median position} = \left(\frac{n+1}{2}\right)^{\text{th}} \text{ position in the ordered data}$$

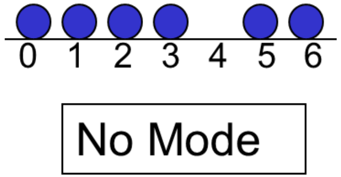
  - If the number of values is odd, the median is the middle number
  - If the number of values is even, the median is the average of the two middle numbers

- 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

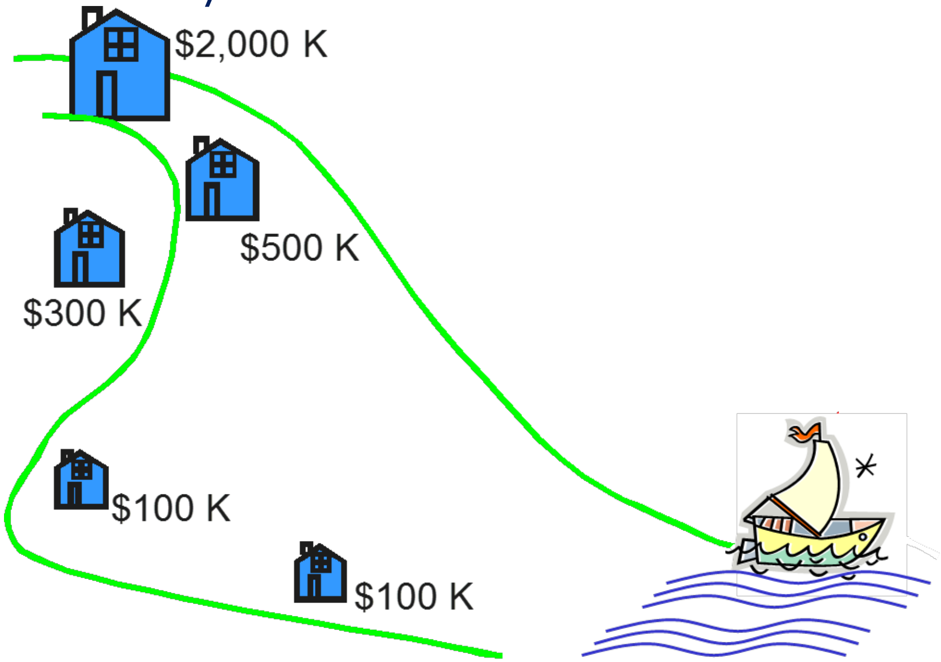




Review Example

- Five houses on a hill by the beach

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



Review Example: Summary Statistics

House Prices:

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

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

  = **$600,000**

- **Median:** middle value of ranked data

  = **$300,000**

- **Mode:** most frequent value

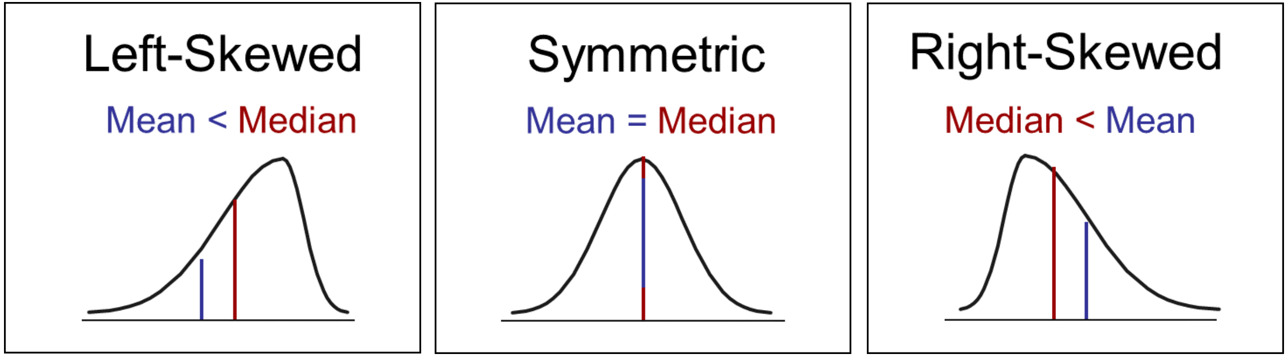
  = **$100,000**

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



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 I Q score at the 90th percentile means that 10% of the population has a higher I Q score and 90% have a lower I Q score.

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

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

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 $Q_1 = 12.5$

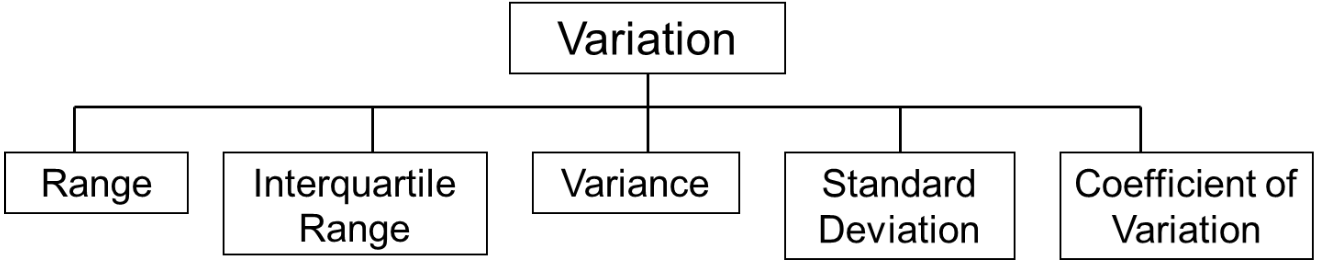
Five-Number Summary

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

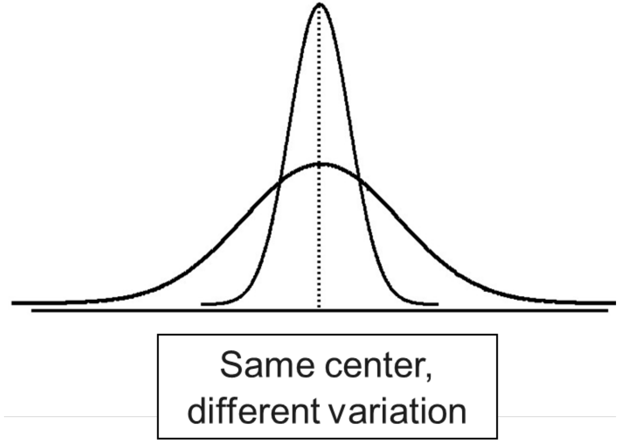
- minimum
- first quartile
- median
- third quartile
- maximum

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

Measures of Variability



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

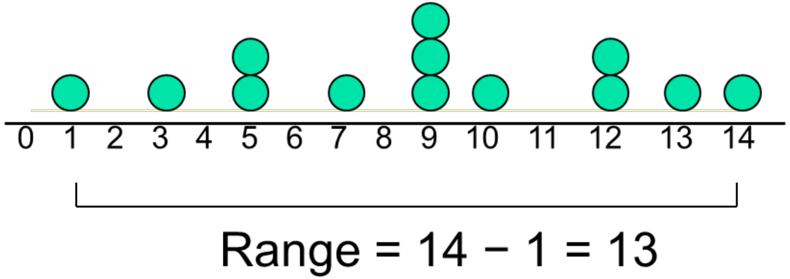


Range

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

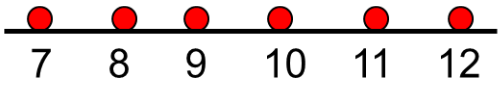
$$\text{Range} = X_{\text{largest}} - X_{\text{smallest}}$$

Example:

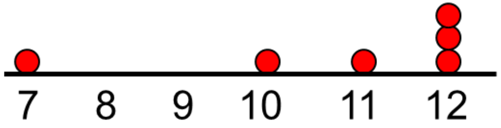


Disadvantages of the Range

- 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 = 3rd quartile – 1st 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

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

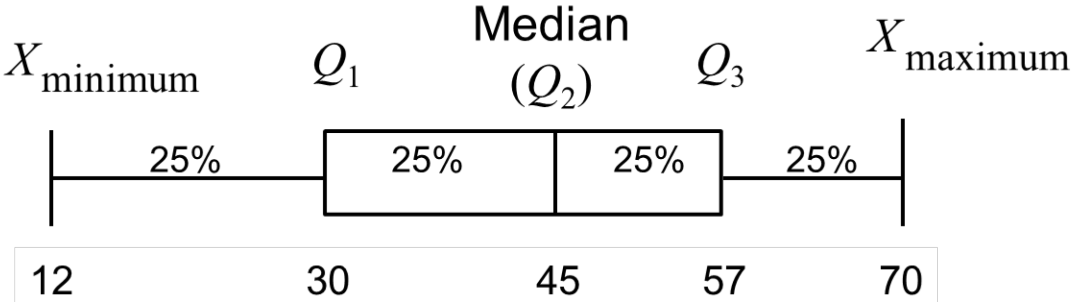
Box-and-Whisker Plot (1 of 2)

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

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:

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

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 (x_i):

10	12	14	15	17	18	18	24

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

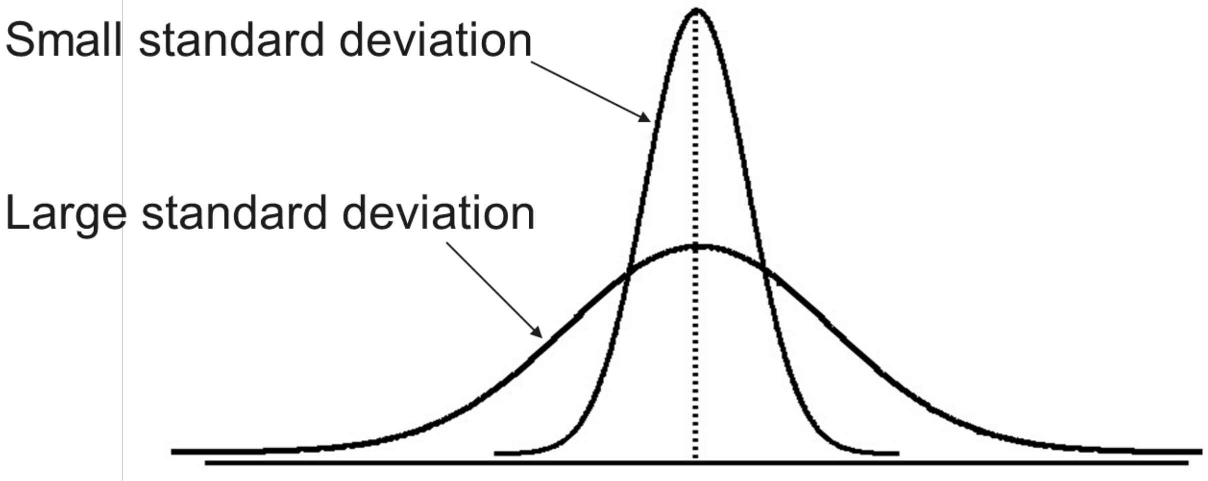
$$s = \sqrt{\frac{(10-\bar{x})^2 + (12-\bar{x})^2 + (14-\bar{x})^2 + \cdots + (24-\bar{x})^2}{n-1}}$$

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

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

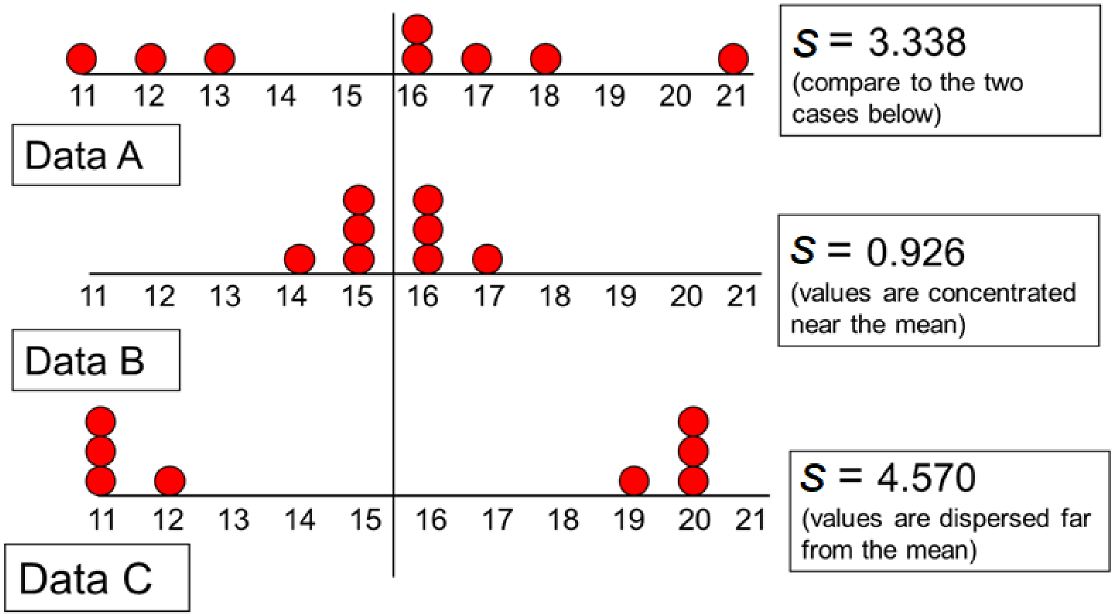
⇒ A measure of the "average" scatter around the mean

Measuring Variation



Comparing Standard Deviations

Mean = 15.5 for each data set



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

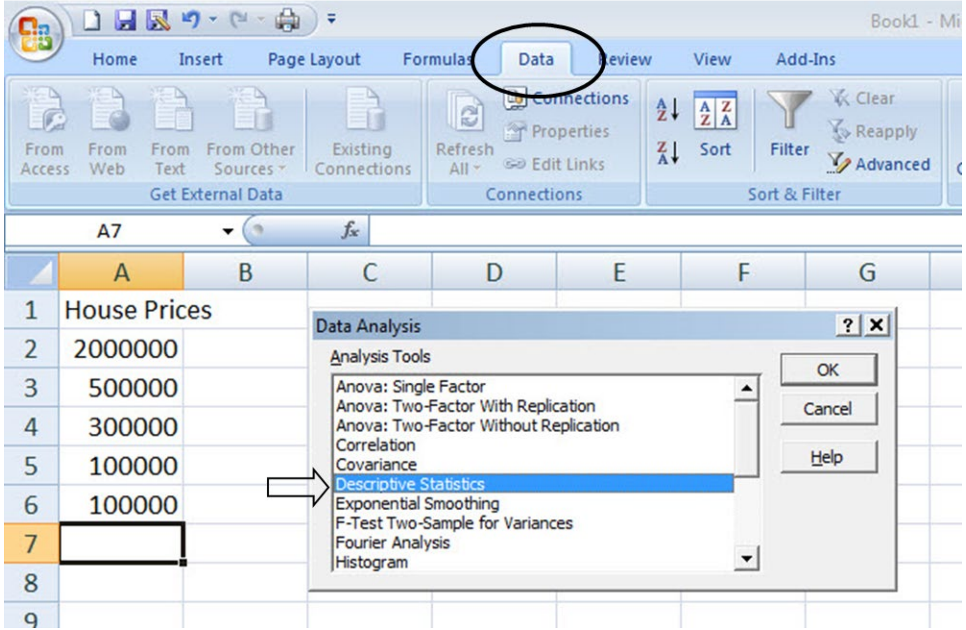
- Select:

data/data analysis/descriptive statistics

- Enter details in dialog box

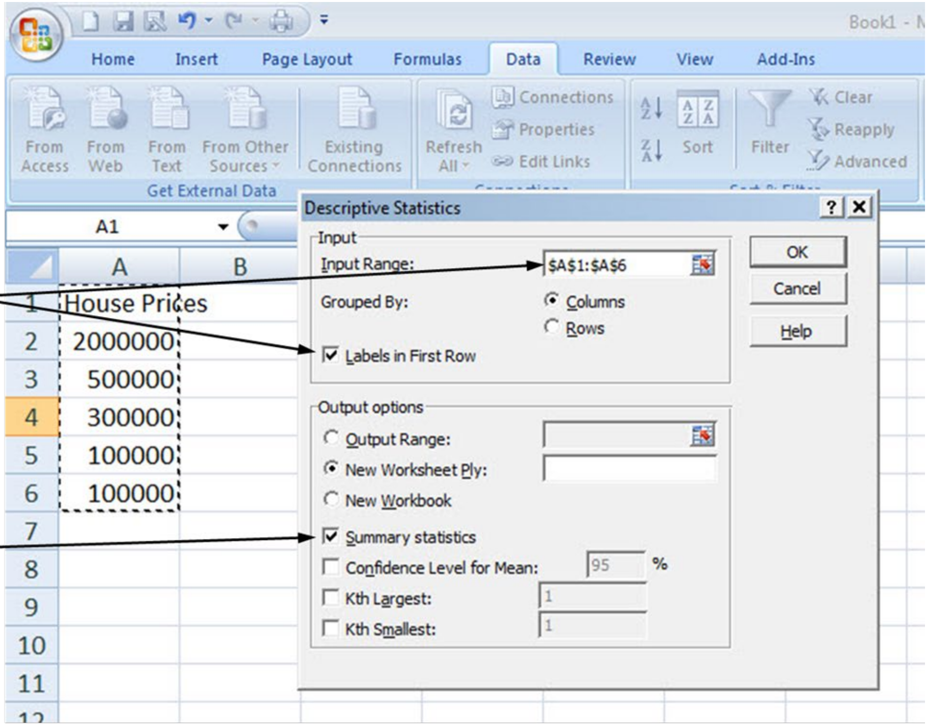
Using Excel (1 of 2)

- Select data/data analysis/descriptive statistics



Using Excel (2 of 2)

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



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		

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:

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

Comparing Coefficient of Variation

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

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

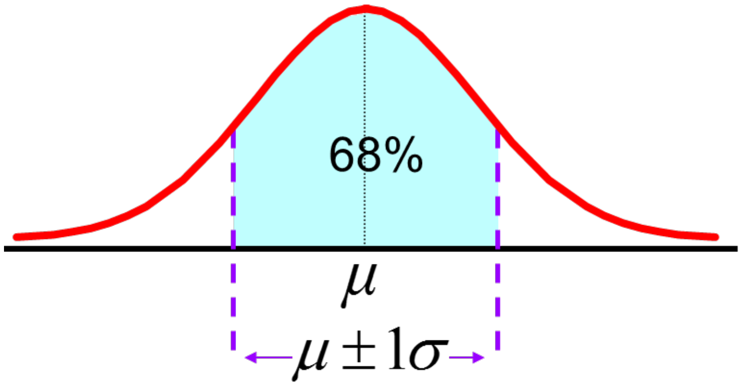
- Stock B:
  - Average price last year = $100
  - Standard deviation = $5

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

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

The Empirical Rule (1 of 2)

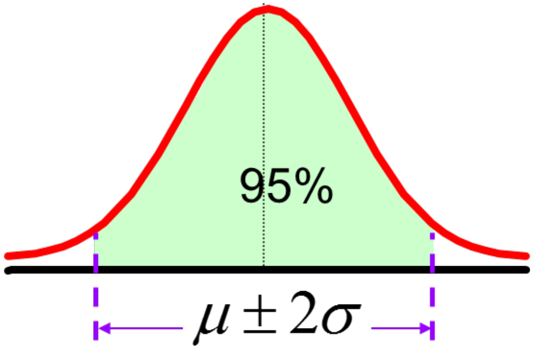
- If the data distribution is bell-shaped, then the interval:
- $\mu \pm 1\sigma$ contains about 68% of the values in the population or the sample

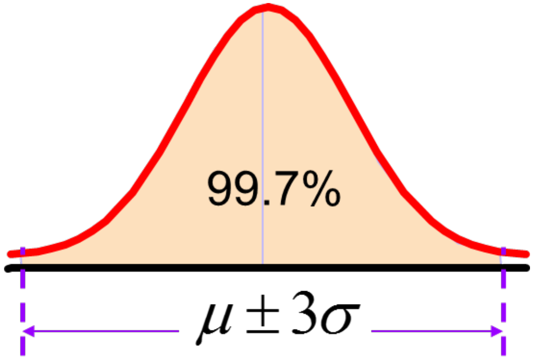


The Empirical Rule (2 of 2)

$\mu \pm 2\sigma$. 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.