Basics of Statistics

Summary Measures

- **Describing Data Numerically**
  - **Central Tendency**
    - Arithmetic Mean
    - Median
    - Mode
  - **Variation**
    - Range
    - Variance
    - Standard Deviation

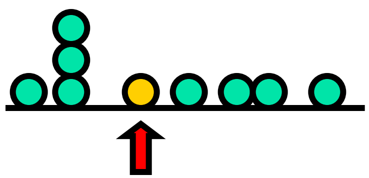
Measures of Central Tendency

Overview

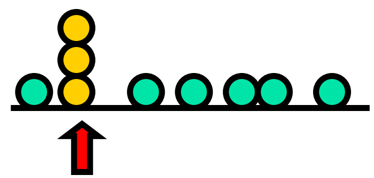
Central Tendency

Arithmetic Mean | **Median** | **Mode**

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



Midpoint of ranked values



Most frequently observed value

Arithmetic Mean

- The arithmetic mean (mean) is the most common measure of central tendency
  - For a sample of size n:

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

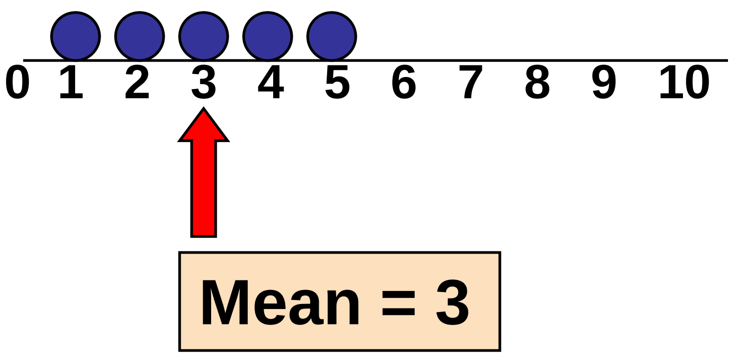
Sample size

Observed values

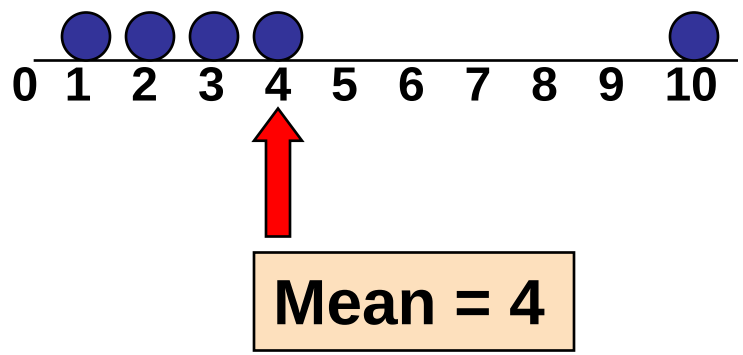
Arithmetic Mean

(continued)

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



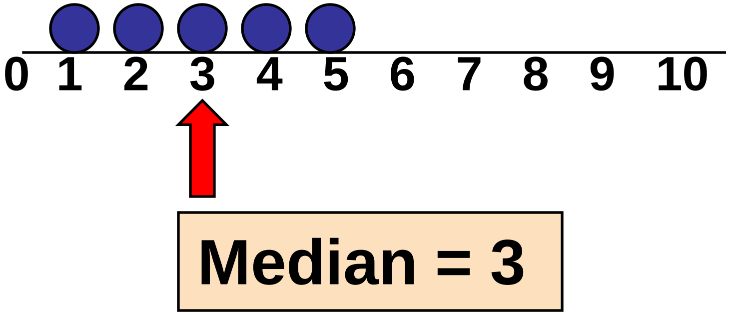
$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

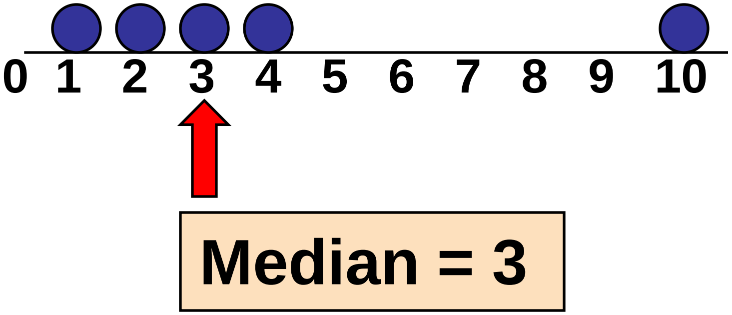


$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

Median

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





- Not affected by extreme values

Finding the Median

- The location of the median:

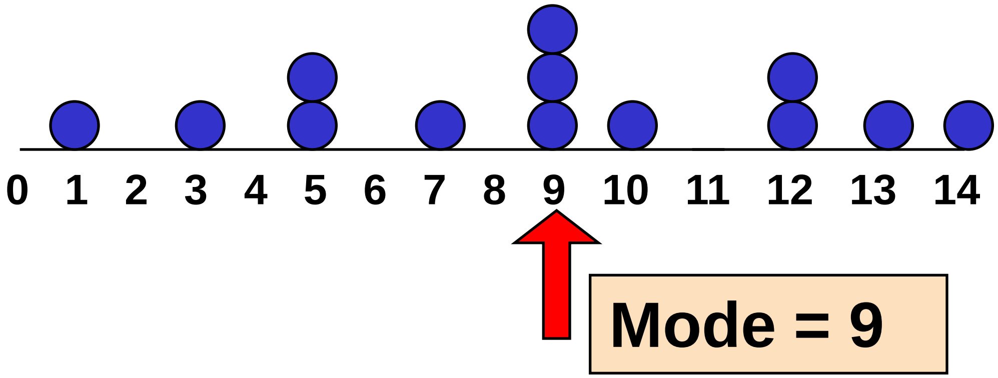
$$\text{Median position} = \frac{n+1}{2} \text{ position in the ordered data}$$

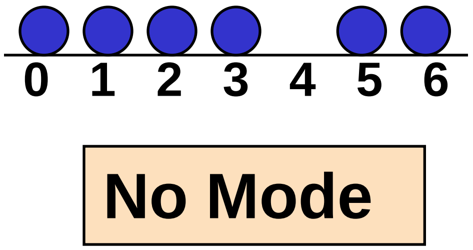
  - If the number of values is odd, the median is the middle number
  - If the number of values is even, the median is the average of the two middle numbers

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

Mode

- A measure of central tendency
- 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



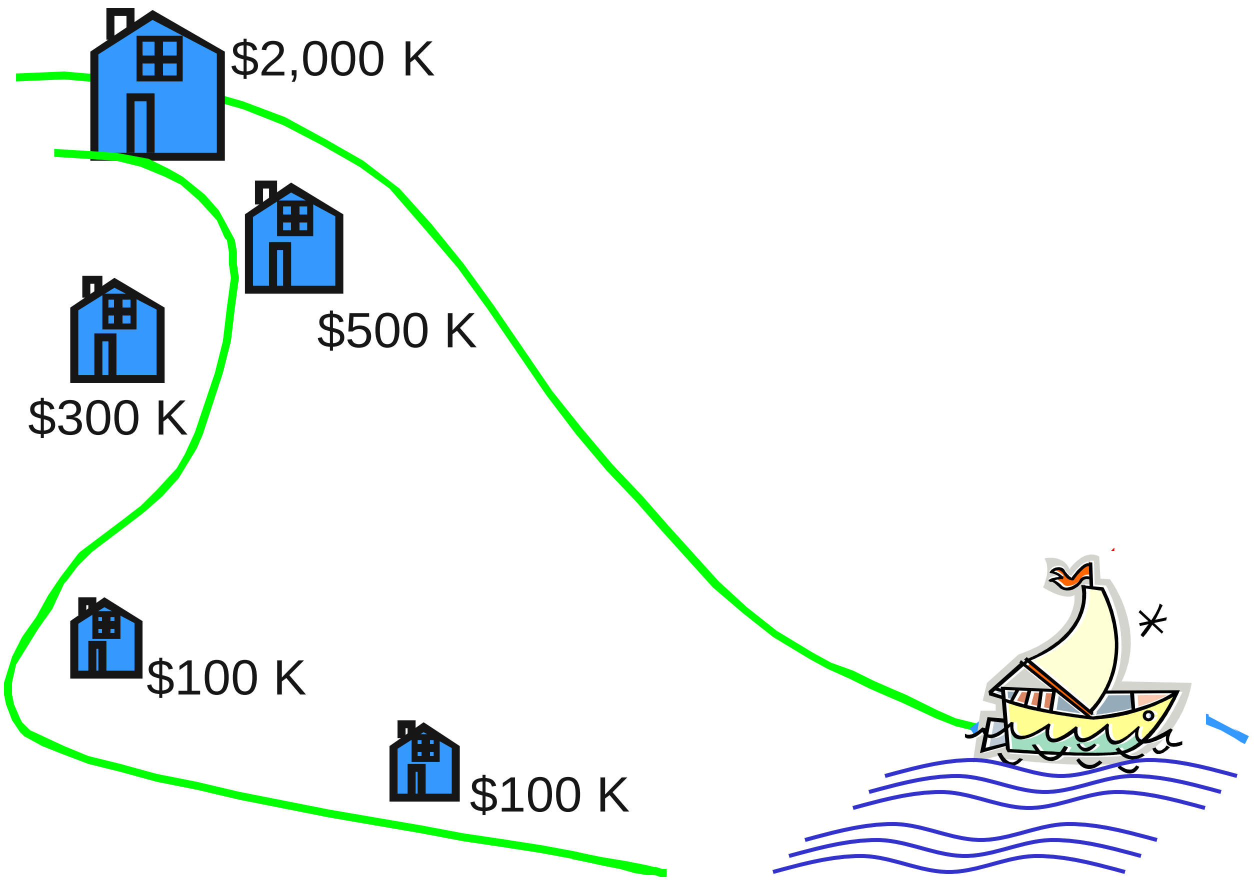


Review Example

- 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:
$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 of location is the “best”?

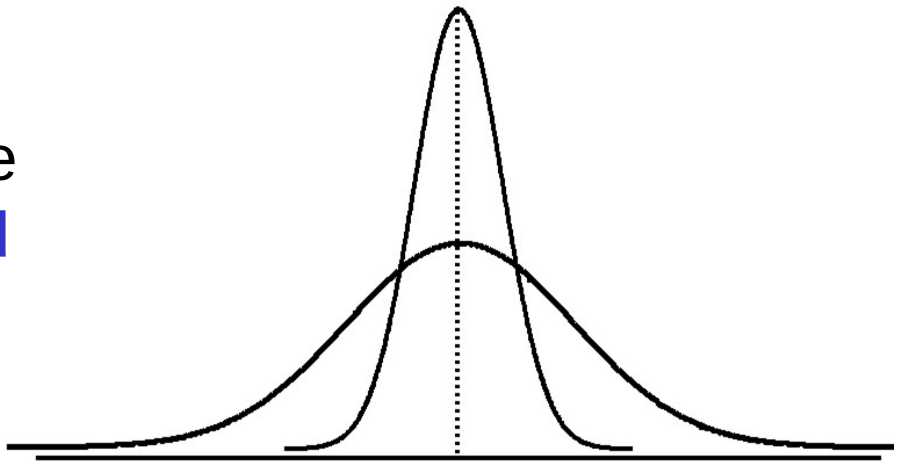
- **Mean** is generally used, unless extreme values (outliers) exist
- Then **median** is often used, since the median is not sensitive to extreme values.
  - Example: Median home prices may be reported for a region – less sensitive to outliers

Measures of Variation

Variation

- **Range**
- **Variance**
- **Standard Deviation**
- **Coefficient of Variation**

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



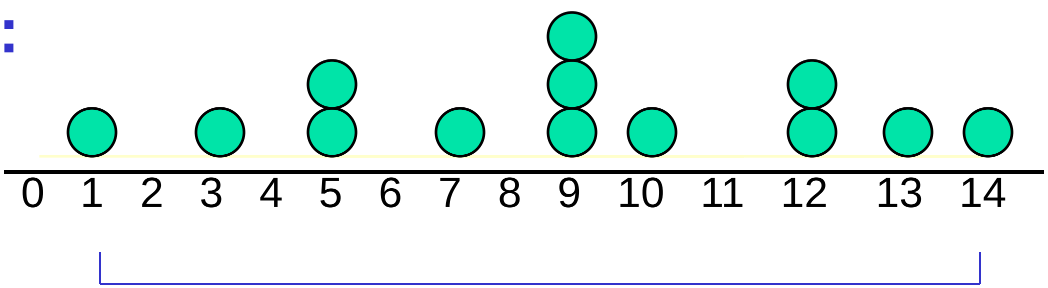
Same center, different variation

Range

- Simplest measure of variation
- Difference between the largest and the smallest values in a set of data:

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

Example:



Range = 14 - 1 = 13

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 $\overline{X}$ = mean

n = sample size

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

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:

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

Calculation Example: Sample Standard Deviation

Sample Data (X_i) :

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

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

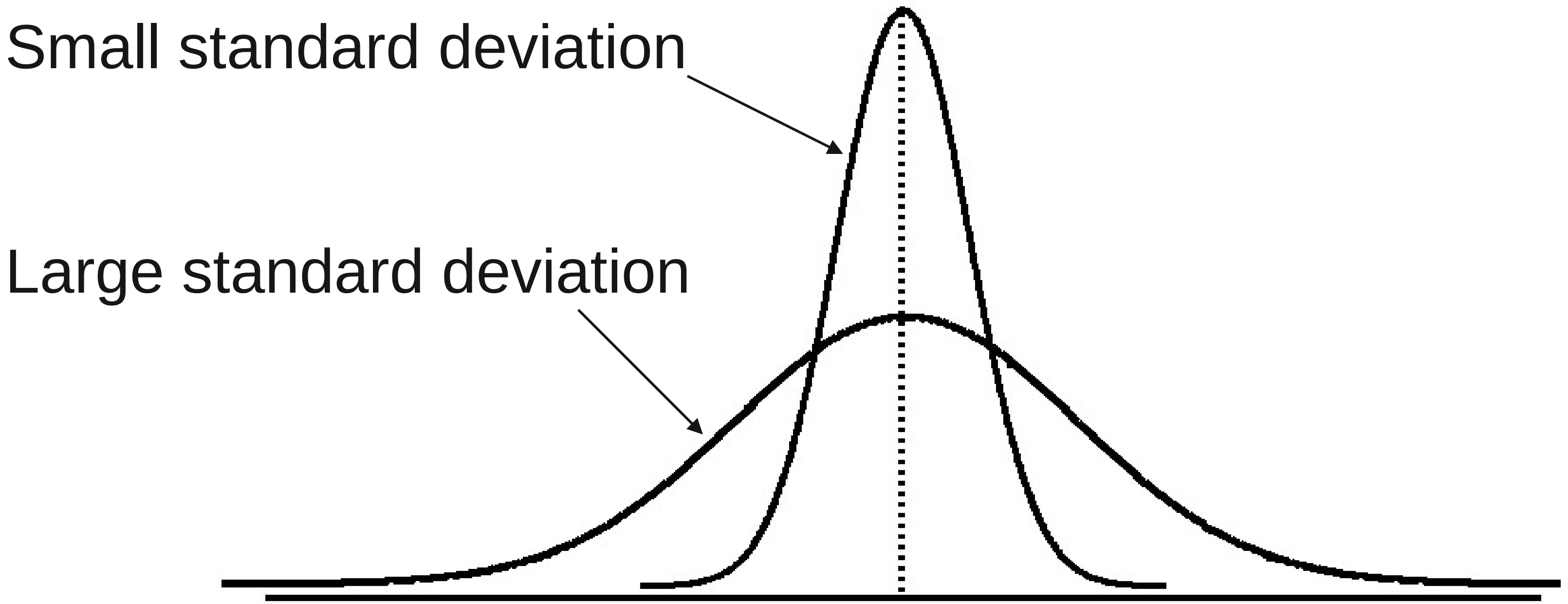
$$S = \sqrt{\frac{(10-\bar{X})^2 + (12-\bar{X})^2 + (14-\bar{X})^2 + \cdots + (24-\bar{X})^2}{n-1}}$$

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

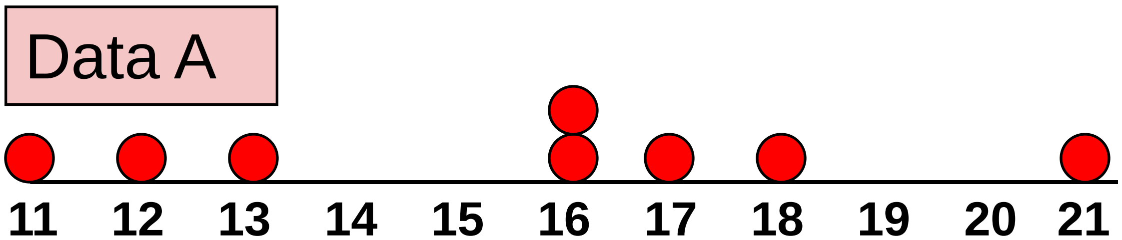
$$= \sqrt{\frac{130}{7}} = \boxed{4.3095}$$

A measure of the "average" scatter around the mean

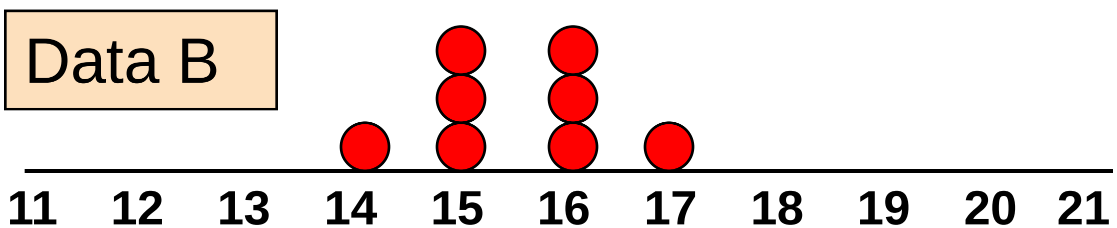
Measuring variation



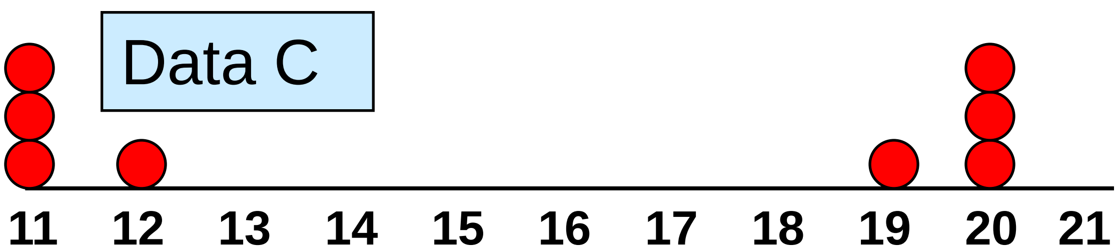
Comparing Standard Deviations



Mean = 15.5
S = 3.338



Mean = 15.5
S = 0.926



Mean = 15.5
S = 4.567

Example

- UP Crop Data on SPSS
- How to draw inferences from them?