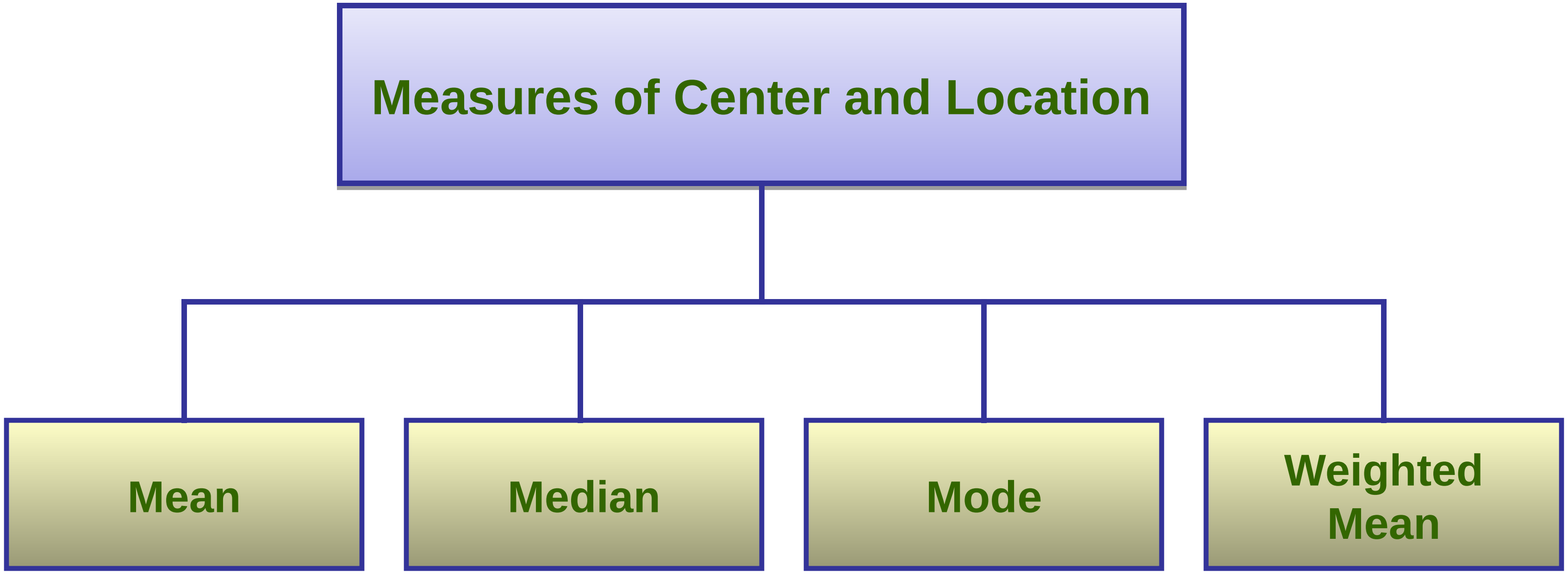
Lecture 3

chapter 3

Describing Data Using Numerical Measures

3.1 Measures of Center and Location



Parameter and Statistic

- Parameter
  - A measure computed from the entire population
  - As long as the population does not change, the value of the parameter will not change
- Statistic
  - A measure computed from a sample that has been selected from a population
  - The value of the statistic will depend on which sample is selected.

Population Mean

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

μ - Population mean
N - Population size
x_i - i^{th} individual value of variable x

- The average for all values in the population computed by dividing the sum of all values by the population size

Sample Mean

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

$\bar{x}$ - Sample mean
n - Sample size

- The average for all values in the sample computed by dividing the sum of all sample values by the sample size

Median

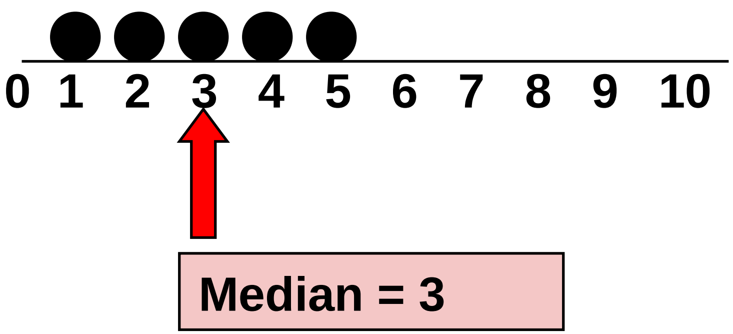
- The median is a center value that divides a data array into two halves (*Md*)
- Data Array
  - Data that have been arranged in numerical order
- Median Index

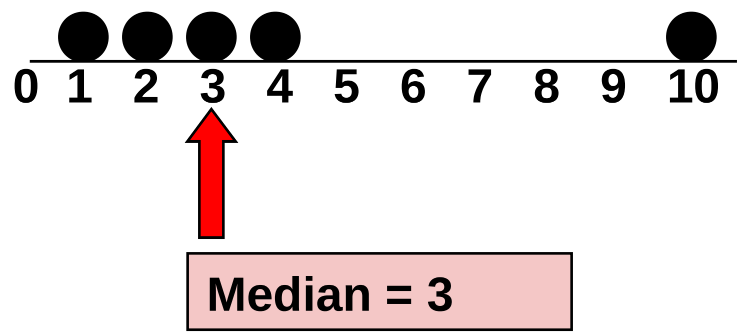
$$\boxed{i = \frac{1}{2}n}$$

  - i = The index of the point in the data set corresponding to the median value
  - n = Sample size

Median

- In an ordered array (lowest to highest), the median is the “middle” number, i.e., the number that splits the distribution in half numerically
  - 50% of the data is above the median, 50% is below
  - Represented as *Md*
- The median is <u>not</u> affected by extreme values





Computing the Median

- Step 1: Collect the sample data
- Step 2: Sort data from smallest to largest
- Step 3: Calculate the median index
  - If *i* is not an integer, round up to next highest integer
  - If *i* is an integer, the median is the average of the values in position i and *i* + 1
- Step 4: Find the median

Median Example

Data array: 4 4 5 5 9 11 12 14 16 19 22 23 24

Note that n = 13

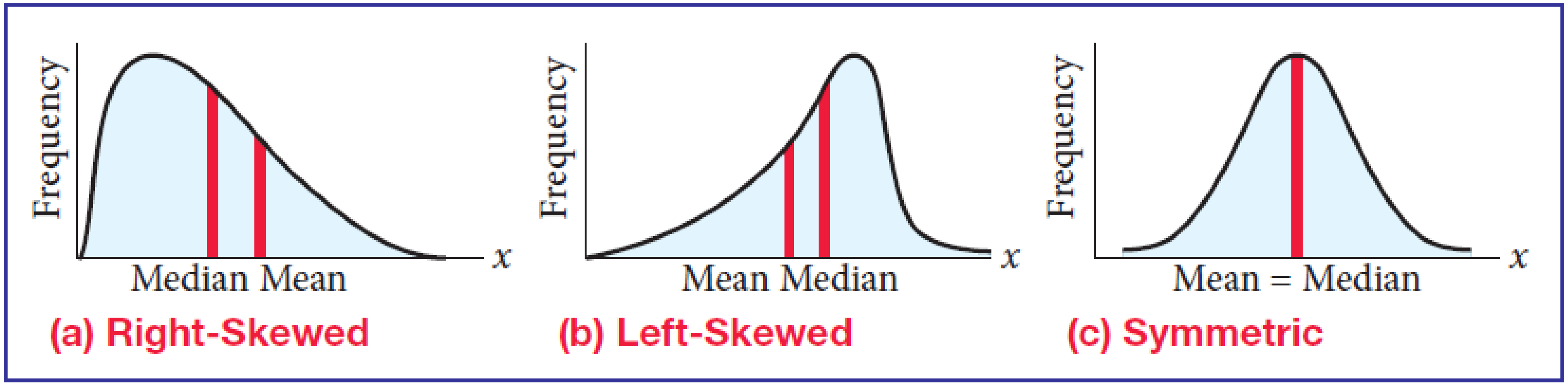
Find the median index $i = (1/2)(13) = 6.5$

Since 6.5 is not an integer, round up to 7

The median is the value in the 7th position: $\boldsymbol{M_d = 12}$

Skewed and Symmetric Distributions

- Symmetric Data
  - Data sets whose values are evenly spread around the center.
- Skewed Data
  - Data sets that are not symmetric



Mode

- The value in a data set that occurs most frequently
- Is not affected by extreme values
- Can be used for both quantitative and qualitative data
- Can have more than one mode, or no mode
- Distribution with two modes - bimodal

The "Best" Measure

- **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
- **Mode** is good for determining more likely to occur

Weighted Mean

- The mean value of data values that have been weighted according to their relative importance

Weighted Mean for a Population

$$\mu_W = \frac{\sum w_i x_i}{\sum w_i}$$

Weighted Mean for a Sample

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i}$$

w_i - The weight of the i^{th} data value

x_i - The i^{th} data value

Calculating Weighted Mean

- Step 1: Collect the desired data and determine the weight to be assigned to each data value
- Step 2: Multiply each weight by the data value and sum these
- Step 3: Sum weights for all values
- Step 4: Compute the weighted mean

Percentiles and Quartiles

Percentiles

The p^{th} percentile in a data array:

- p% are less than or equal to this value
- $(100 - p)$% are greater than or equal to this value

(where $0 \le p \le 100$)

- 50^{th} percentile is the median

Quartiles

1st quartile = 25th percentile

2nd quartile = 50th percentile

Also the median

3rd quartile = 75th percentile

Calculating Percentiles

- Step 1: Sort the data in order from the lowest to highest value.
- Step 2: Determine the percentile location index:

$$\boxed{i = \frac{p}{100}(n)}$$

- Step 3: If *i* is not an integer, then round to next highest integer. The p^{th} percentile is located at the rounded index position. If *i* is an integer, the p^{th} percentile is the average of the values at location index positions *i* and *i* + 1.

Percentile Example

- Find the 60th percentile in an ordered array of 19 values

36	40	42	46	51	56	62	65	71	74	78	82	84	87	88	90	92	95	97

- Percentile location index:

$$i = \frac{p}{100}(n) = \frac{60}{100}(19) = 11.4$$

Use value at 12th position

- 60th percentile equals 82

Quartile Example

- Find the 1st quartile in an ordered array of 19 values

36	40	42	46	51	56	62	65	71	74	78	82	84	87	88	90	92	95	97

- Quartile location index:

$$i = \frac{q}{100}(n) = \frac{25}{100}(19) = 4.75$$

Use value at 5th position

- 1st quartile Q_1 equals 51

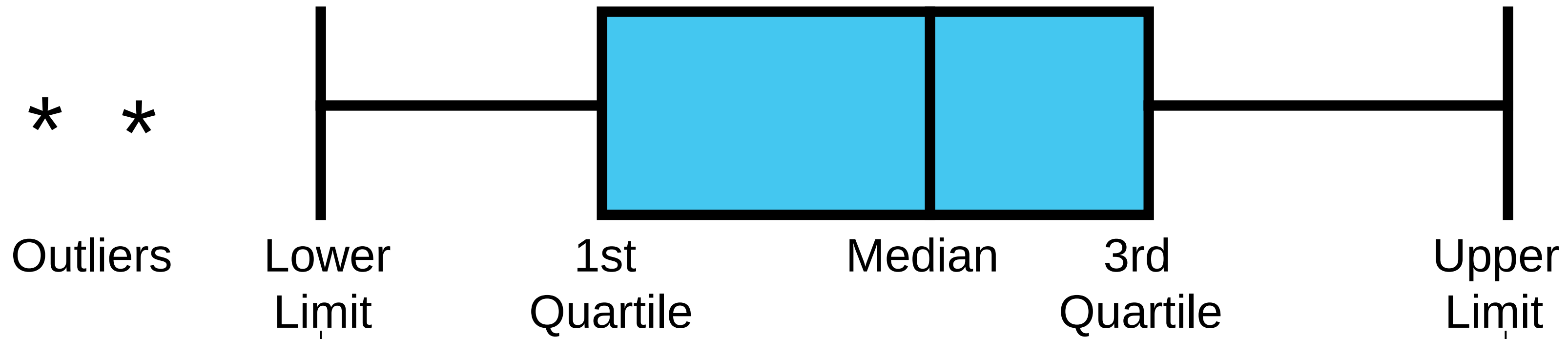
Box and Whisker Plot

- A graph that is composed of two parts: a box and the whiskers
- The box has a width that ranges from the first quartile ($Q1$) to the third quartile ($Q3$)
- A vertical line through the box is placed at the median.
- Limits are located at a value that is 1.5 multiplied by the difference between $Q1$ and $Q3$ below $Q1$ and above $Q3$.
- The whiskers extend to the left to the lowest value within the limits and to the right to the highest value within the limits.

Constructing a Box and Whisker Plot

- Step 1: Sort values from lowest to highest
- Step 2: Find Q_1, Q_2, Q_3
- Step 3: Draw the box so that the ends are at Q_1 and Q_3
- Step 4: Draw a vertical line through the median
- Step 5: Calculate the interquartile range ($IQR = Q_3 - Q_1$)
- Step 6: Extend dashed lines from each end to the highest and lowest values within the limits
- Step 7: Identify outliers with an asterisk (*)

Constructing a Box and Whisker Plot



The lower limit is $Q_1 - 1.5\,(Q_3 - Q_1)$

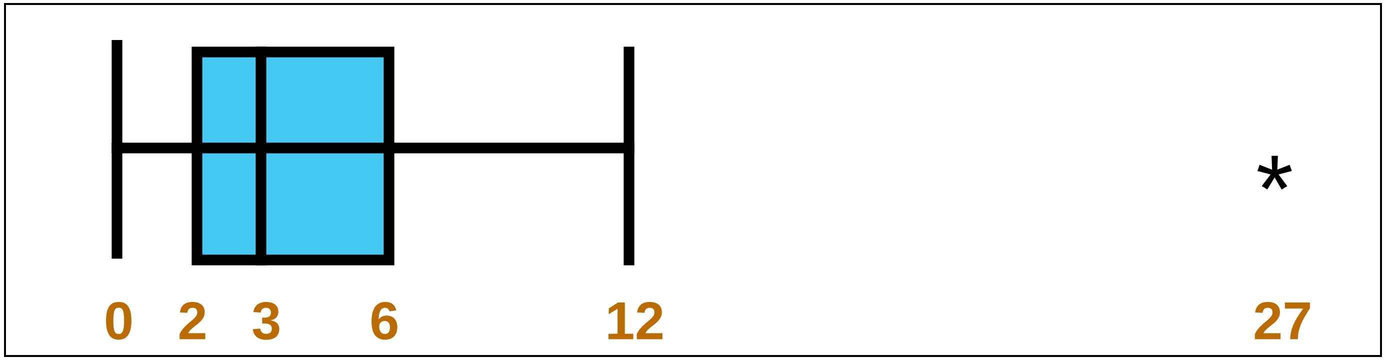
The upper limit is $Q_3 + 1.5\,(Q_3 - Q_1)$

- The center box extends from Q_1 to Q_3
- The line within the box is the median
- The whiskers extend to the smallest and largest values within the calculated limits
- Outliers are plotted outside the calculated limits

Box and Whisker Plot Example

- Below is a Box-and-Whisker plot for the following data:

Min		Q_1			Q_2			Q_3		Max
0	2	2	2	3	3	4	5	6	11	27



Upper limit = $Q_3 + 1.5\,(Q_3 - Q_1)$

$= 6 + 1.5\,(6 - 2) = 12$

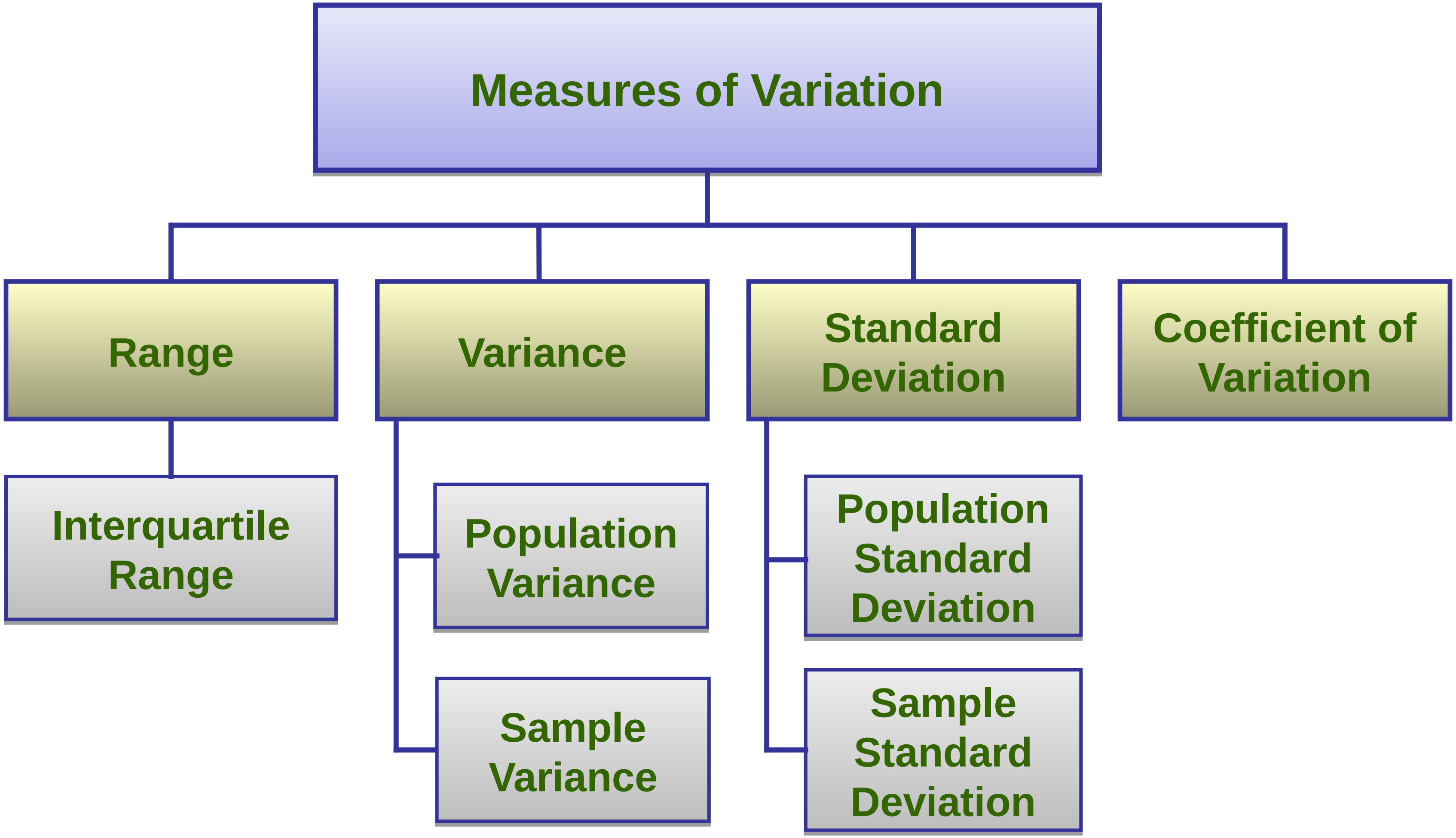
27 is above the upper limit so is shown as an outlier

- This data is right skewed, as the plot depicts

Descriptive Measures of the Center

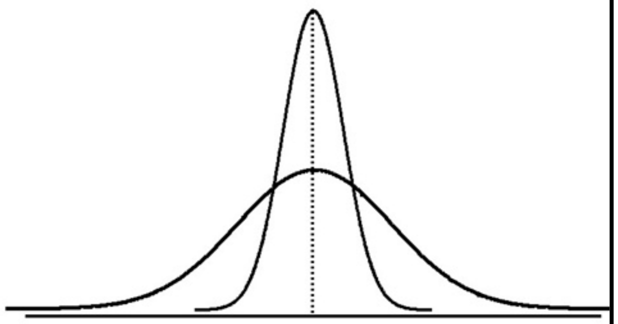
Descriptive Measure	Computation Method	Data Level	Advantages/ Disadvantages
Mean	Sum of values divided by the number of values	Ratio Interval	• Numerical center of the data • Sum of deviations from the mean is zero • Sensitive to extreme values
Median	Middle value for data that have been sorted	Ratio Interval Ordinal	• Not sensitive to extreme values • Computed only from the center values • Does not use information from all the data
Mode	Value(s) that occur most frequently in the data	Ratio Interval Ordinal Nominal	• May not reflect the center • May not exist • Might have multiple modes

3.2 Measures of Variation



Variation

- A set of data exhibits variation if all the data are not the same value
- Measures of variation give information on the spread or variability
  - Smaller value – less variation
  - Larger value – more variation


Same center, different variation

Range

- A measure of variation that is computed by finding the difference between the maximum and minimum values in a data set

$$\boxed{R = \text{Maximum Value} - \text{Minimum Value}}$$

- Simplest measure of variation
- Is very sensitive to extreme values
- Ignores the data distribution

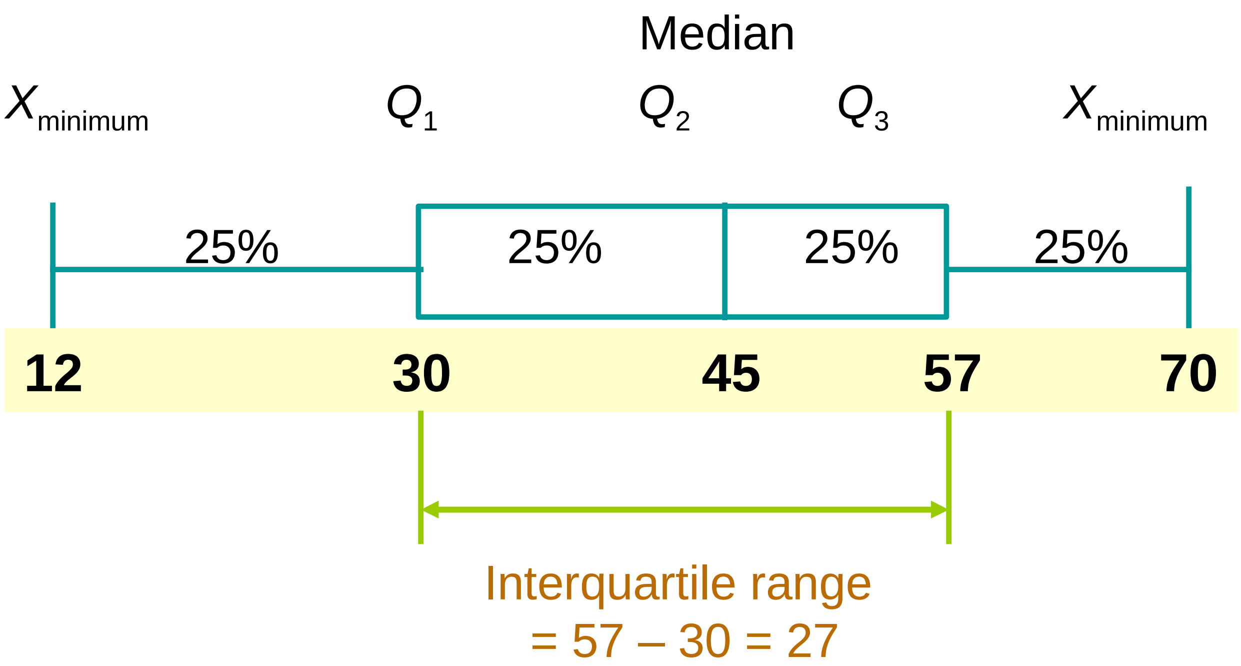
Interquartile Range

- A measure of variation that is determined by computing the difference between the third and first quartiles

- Eliminates outlier problems
- Eliminates some high- and low-valued observations

$$\boxed{\textit{Interquartile Range} = Q_3 - Q_1}$$

Interquartile Range Example



Population Variance

- The average of the squared distances of the data values from the mean.

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

μ - population mean, *N* – population size

- Shortcut formula:

$$\sigma^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{N}}{N}$$

Population Standard Deviation

- The most commonly used measure of variation
- The positive square root of the variance

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

- Has the same units as the original data

Sample Variance and Standard Deviation

- Sample data have been selected from the population
- Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

- Sample Standard Deviation

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

Computing Sample Variance and Standard Deviation

- Step 1: Select the sample and record the data for the variable of interest
- Step 2: Select expression for sample variance
- Step 3: Compute $\bar{x}$
- Step 4: Determine the sum of the squared deviations of each x value from $\bar{x}$
- Step 5: Compute the sample variance
- Step 6: Compute the sample standard deviation by taking the square root of the variance

Standard Deviation Calculation Example

Sample Data (x_i)	4	7	1	0	5	0	3	2	6	2

$n = 10 \qquad \bar{x} = 3$

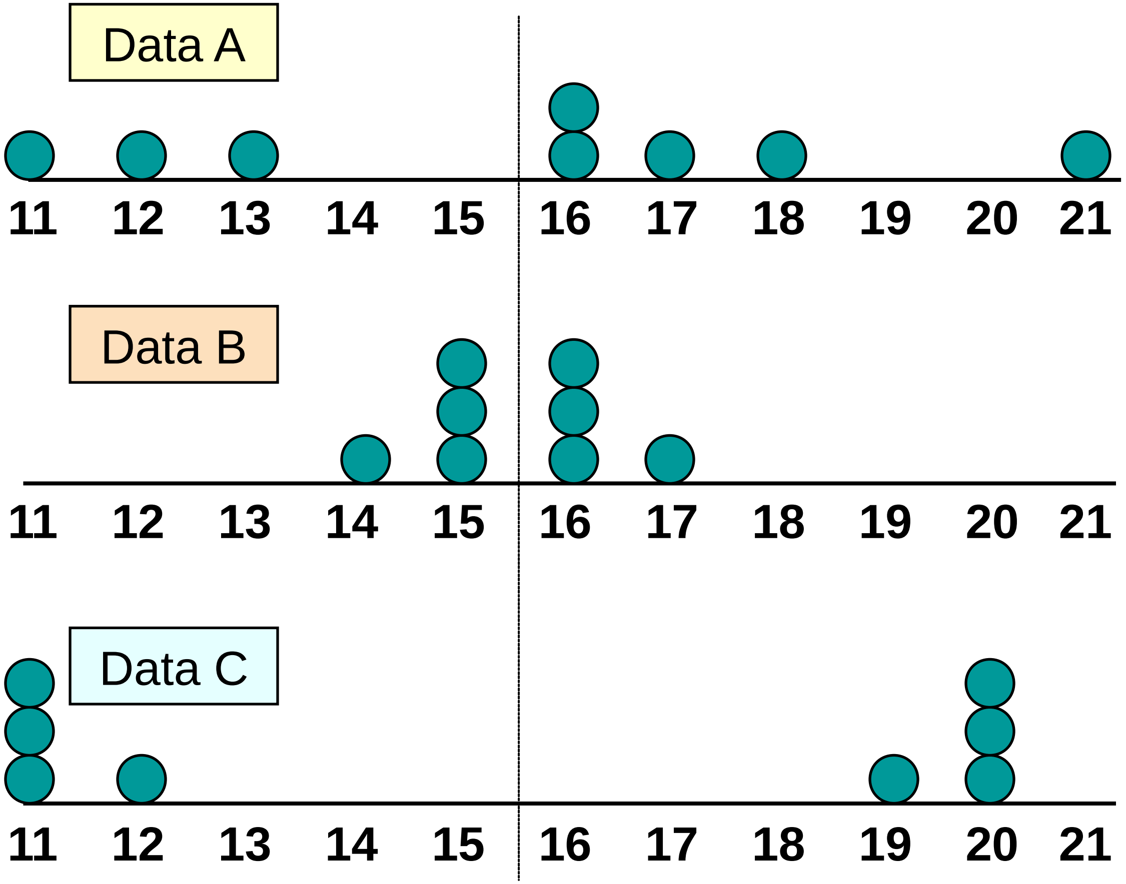
$$S = \sqrt{\frac{(4-\bar{x})^2+(7-\bar{x})^2+(1-\bar{x})^2+(0-\bar{x})^2+\ldots+(6-\bar{x})^2+(2-\bar{x})^2}{10-1}} =$$

$$= \sqrt{\frac{(4-3)^2+(7-3)^2+(1-3)^2+(0-3)^2+\ldots+(6-3)^2+(2-3)^2}{10-1}} =$$

$$= \sqrt{\frac{54}{9}} = 2.449$$

Comparing Standard Deviations

Same mean, but different standard deviations:



Mean = 15.5
s = 3.338

Mean = 15.5
s = 0.9258

Mean = 15.5
s = 4.57

3.3 Using the Mean and Standard Deviation Together

- Coefficient of Variation (*CV*)
  - The ratio of the standard deviation to the mean expressed as a percentage. The coefficient of variation is used to measure variation relative to the mean
  - Measures relative variation
  - Always expressed in percentage (%)
  - Shows variation relative to mean

Coefficient of Variation

- Is used to compare two or more sets of data measured in different units
- Population *CV*

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

- Sample *CV*

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

Comparing Coefficients of Variation

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

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

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

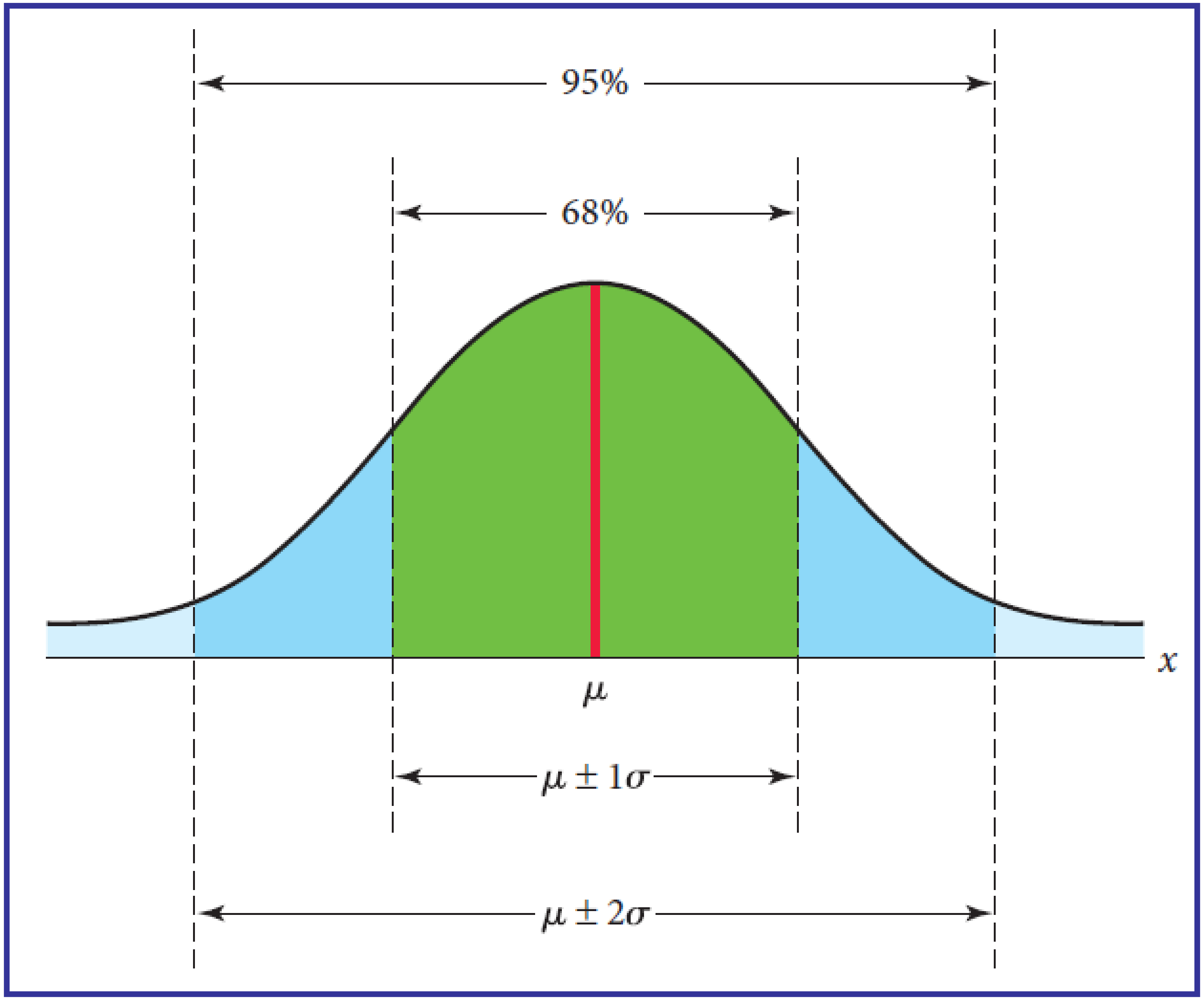
$$CV_A = \left(\frac{s}{\bar{x}}\right) * 100\% = \frac{\$5}{\$100} * 100\% = 5\%$$

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

The Empirical Rule

- If the data distribution is bell shaped, then the interval
  - $\mu \pm 1\sigma$ contains approximately 68% of the values
  - $\mu \pm 2\sigma$ contains approximately 95% of the values
  - $\mu \pm 3\sigma$ contains virtually all of the data values

The Empirical Rule



Tchebysheff's Theorem

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

$(1 - 1/1^2) =$ 0% .... $k=1$ $(\mu \pm 1\sigma)$
$(1 - 1/2^2) =$ 75% .... $k=2$ $(\mu \pm 2\sigma)$
$(1 - 1/3^2) =$ 89% .... $k=3$ $(\mu \pm 3\sigma)$

Standardized Data Values

- The number of standard deviations a value is from the mean
- Standardized data values are also referred to as *z* scores
  - Population z score

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

  - Sample z score

$$z = \frac{x - \bar{x}}{s}$$

x – data value
μ - population mean
σ – population standard deviation
$\bar{x}$ – sample mean
s – sample standard deviation

Converting Data to Standardized Values

- Step 1: Collect the population or sample values for the quantitative variable of interest.
- Step 2: Compute the population mean and standard deviation or the sample mean and standard deviation.
- Step 3: Convert the values to standardized z-values

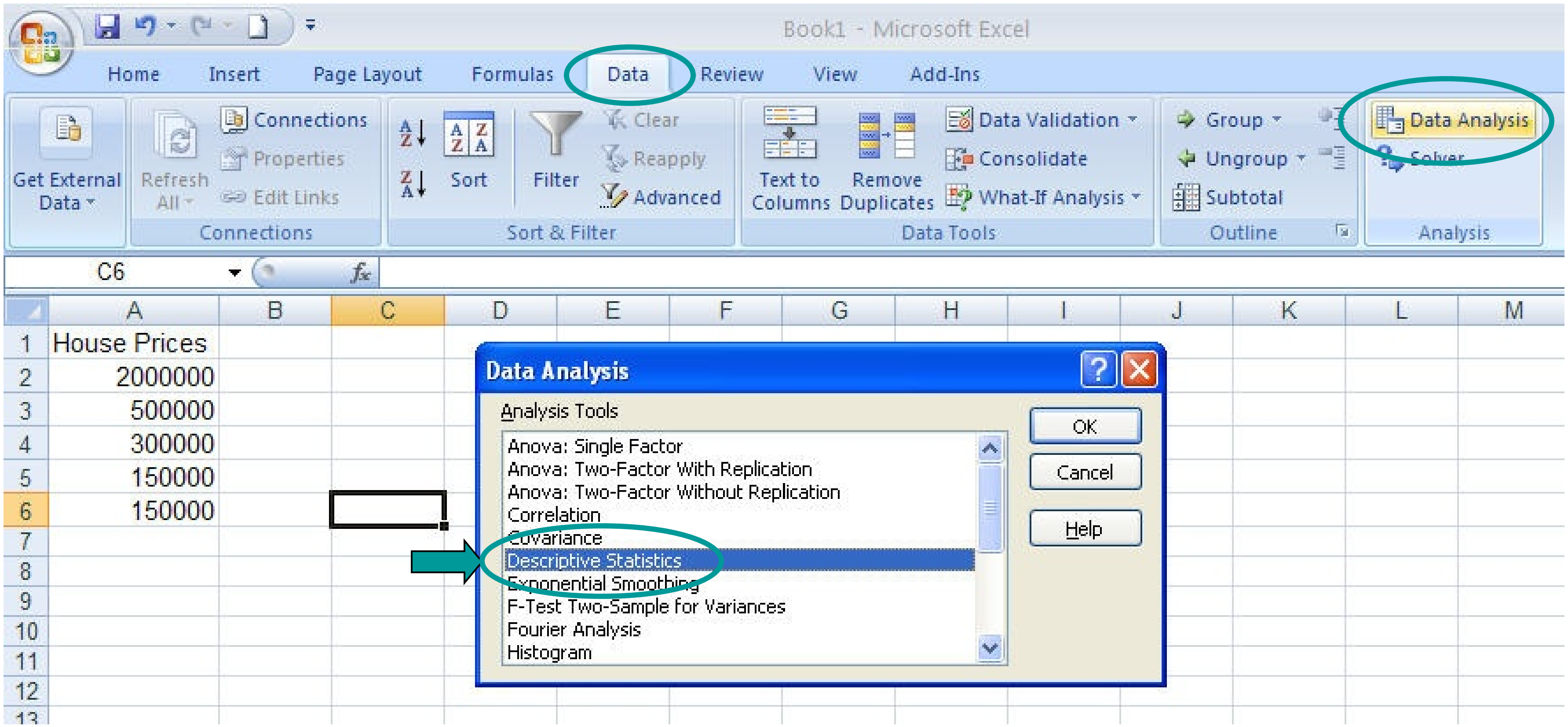
Standardized Value Calculation Example

- IQ scores in a large population have a bell-shaped distribution with mean $\mu = 100$ and standard deviation $\sigma = 15$
- Find the standardized score (z-score) for a person with an IQ of 121.

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

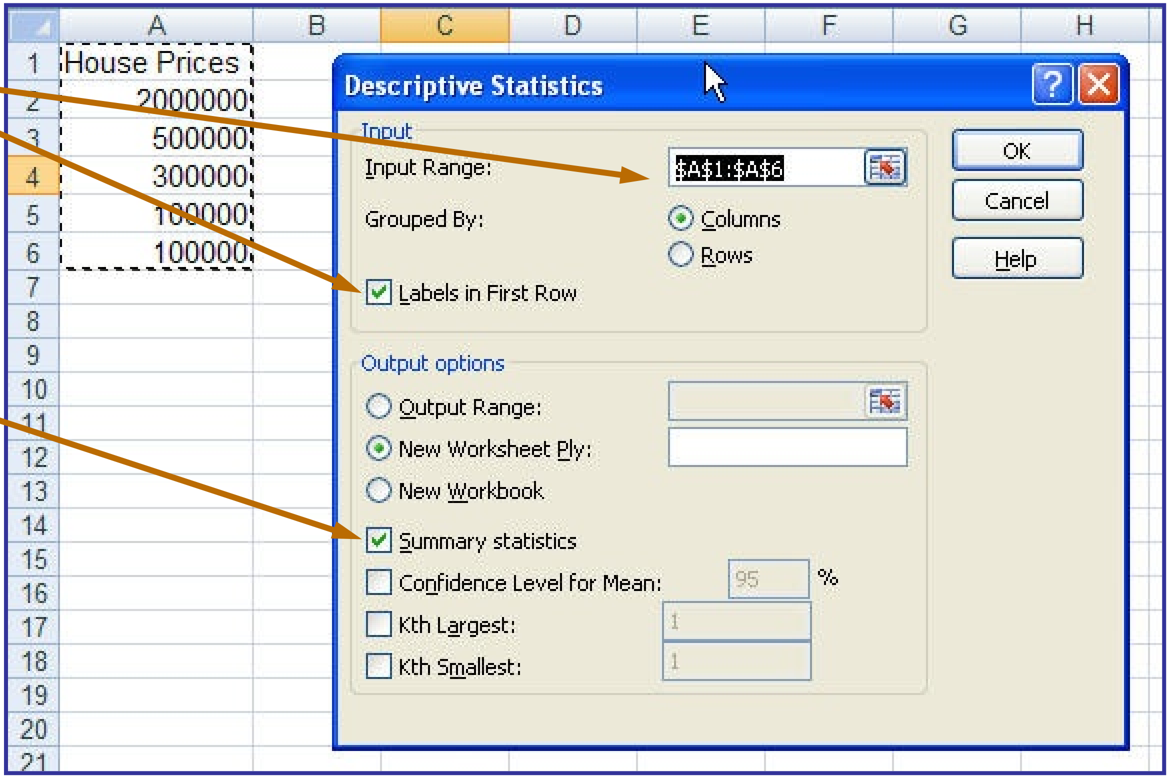
- Someone with an IQ of 121 is 1.4 standard deviations above the mean

How to Do It in Excel?



How to Do It in Excel?

- Enter dialog box details
- Check box for summary statistics
- Click OK



Descriptive Statistics Output

- Excel Output

	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		

Mean

Median

Mode

Standard Deviation

Variance

Range