MATH1324

Numeric Descriptive Measures

www.rmit.edu.au

RMIT
UNIVERSITY

Objective/Goals

- Describe the properties of central tendency, variation, and shape in numerical data
- Construct and interpret a boxplot
- Calculate the covariance and the coefficient of correlation

But first, some definitions

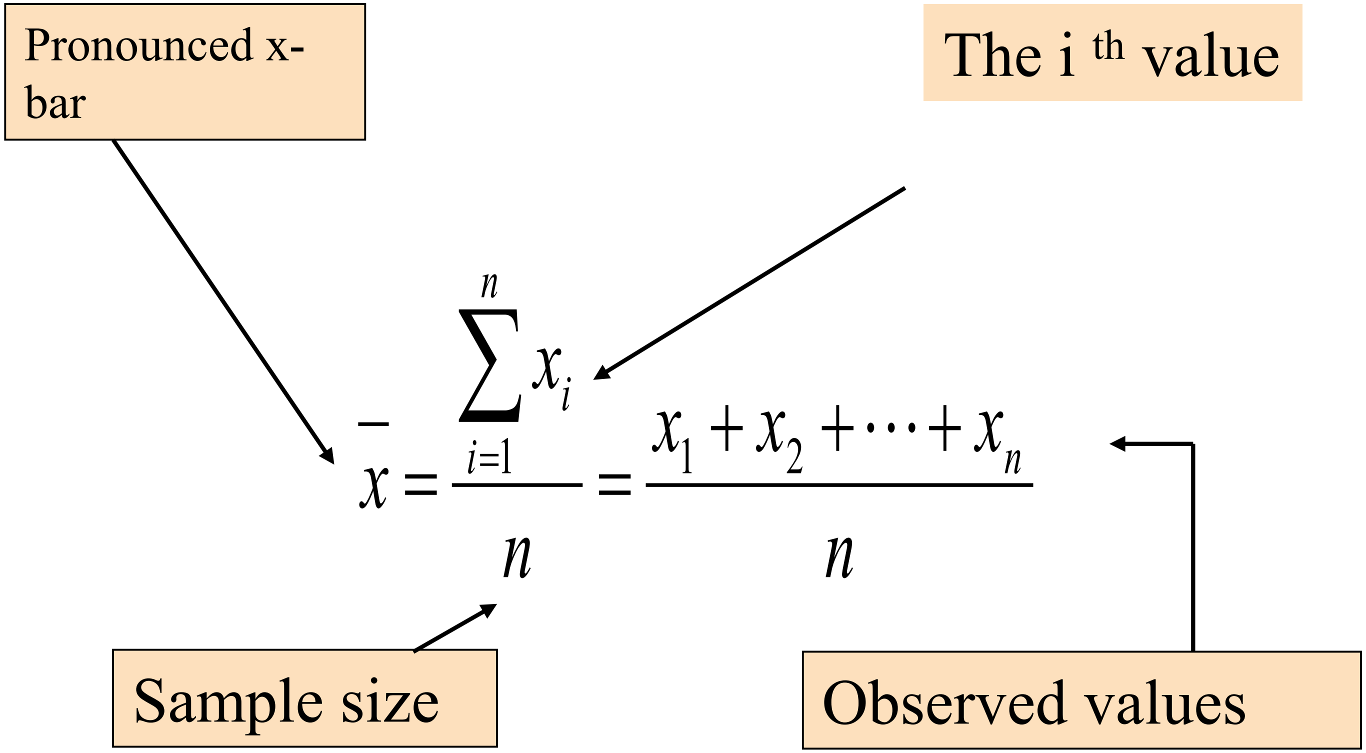
- The **central tendency** is the extent to which the values of a numerical variable group around a typical or central value.

- The **variation** is the amount of dispersion or scattering away from a central value that the values of a numerical variable show.

- The **shape** is the pattern of the distribution of values from the lowest value to the highest value

The Mean

- The arithmetic mean (often just called the “mean”) is the most common measure of central tendency. For a sample size of n:



Pronounced x-bar

The i^{th} value

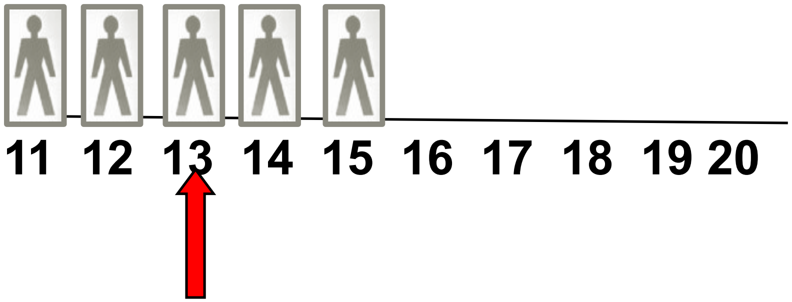
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Sample size

Observed values

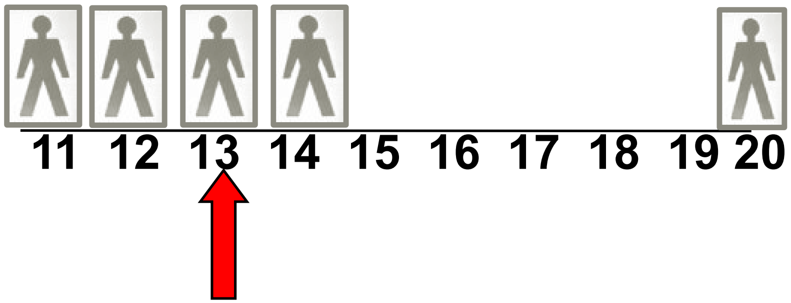
The Mean

- Mean = sum of values divided by the number of values
- Affected by extreme values (outliers)



Mean = 13

$$\frac{11+12+13+14+15}{5}=\frac{65}{5}=13$$

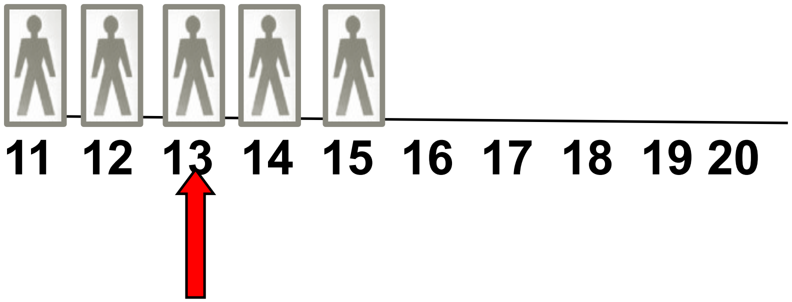


Mean = 14

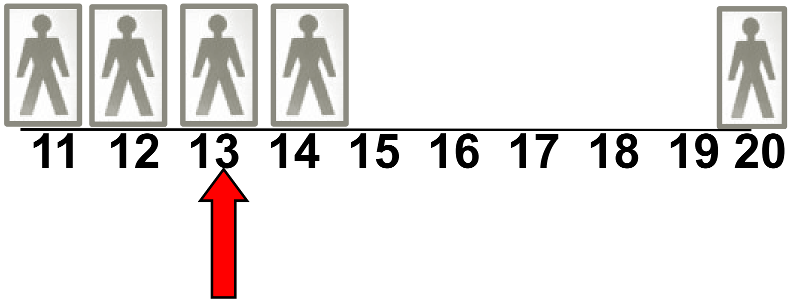
$$\frac{11+12+13+14+20}{5}=\frac{70}{5}=14$$

The Median

- In an ordered array, the median is the “middle” number (50% above, 50% below)
- Less sensitive to extreme values



Median = 13



Median = 13

Locating the Median

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

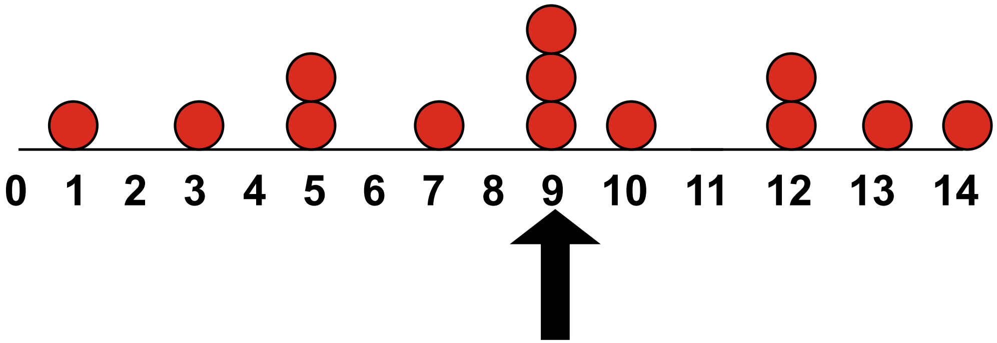
$$\text{Median position} = \frac{n+1}{2} \text{ position in the ordered data}$$

- 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

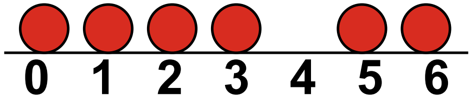
Note that is not the value of the median, only the position of the median in the ranked data

Mode

- 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

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**

Example

The number of quarts of milk purchased by 25 households:

0 0 1 1 1 1 1 2 2 2 2 2 2 2 2 2 3 3
3 3 3 4 4 4 5

- **Mean?**

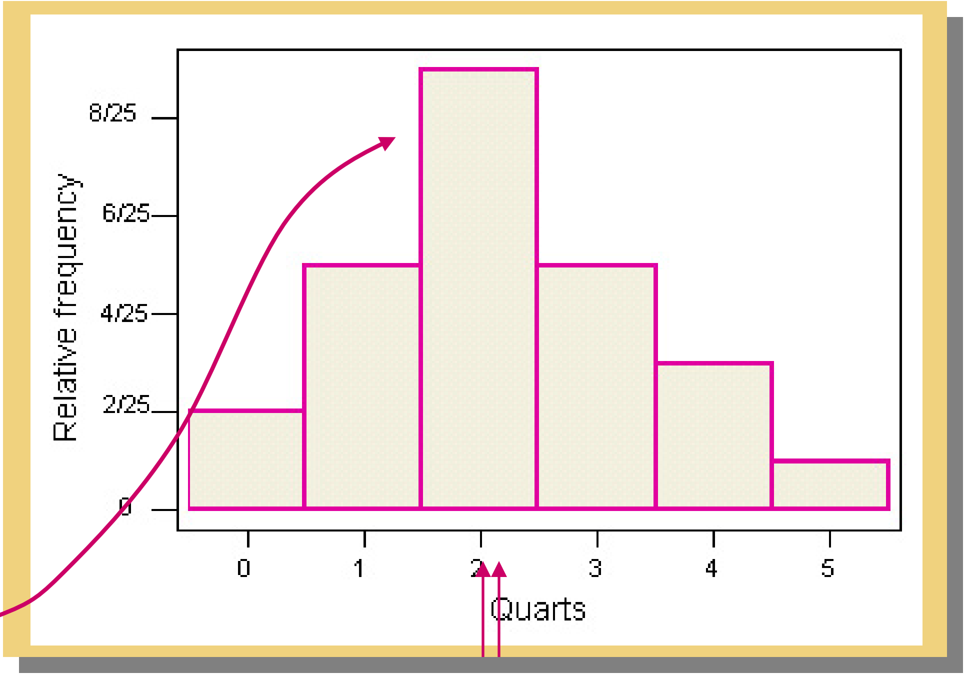
$$\bar{x} = \frac{\sum x_i}{n} = \frac{55}{25} = 2.2$$

- **Median?**

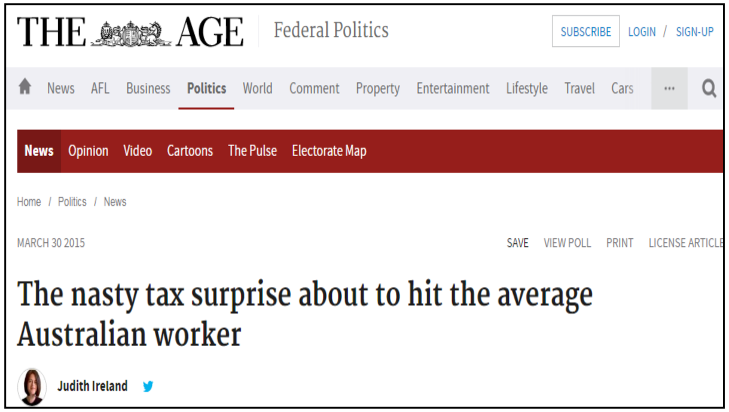
$$m = 2$$

- **Mode? (Highest peak)**

$$\text{mode} = 2$$



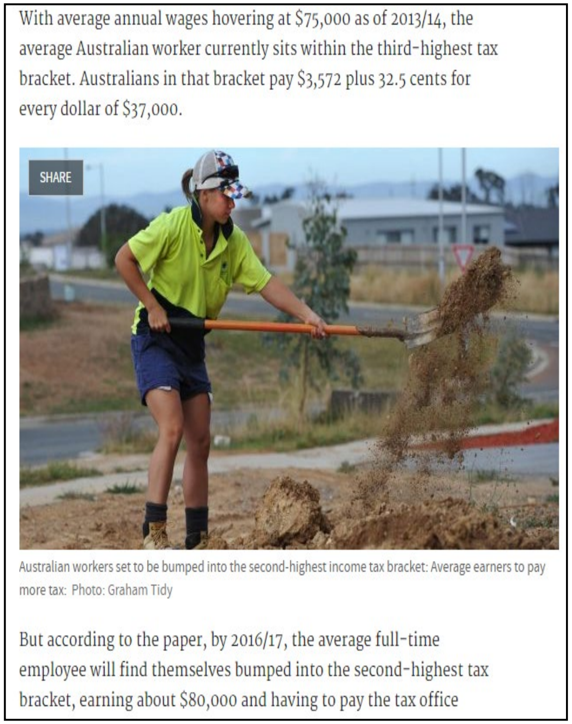
A real example
(Journalists should study statistics)



MARCH 30 2015

The nasty tax surprise about to hit the average Australian worker

Judith Ireland



With average annual wages hovering at $75,000 as of 2013/14, the average Australian worker currently sits within the third-highest tax bracket. Australians in that bracket pay $3,572 plus 32.5 cents for every dollar of $37,000.

Australian workers set to be bumped into the second-highest income tax bracket: Average earners to pay more tax: Photo: Graham Tidy

But according to the paper, by 2016/17, the average full-time employee will find themselves bumped into the second-highest tax bracket, earning about $80,000 and having to pay the tax office

"AWE statistics represent average gross (before tax) earnings of employees and **do not relate to average award rates or to the earnings of the 'average person'**. AWE estimates are derived by dividing estimates of weekly **total earnings by estimates of the number of employees**."

Australian Bureau of Statistics (ABS) 2014, *Average Weekly Earnings, Australia*, May 2014, cat. no. 6302.0

Weekly Household Gross Income 2013/14

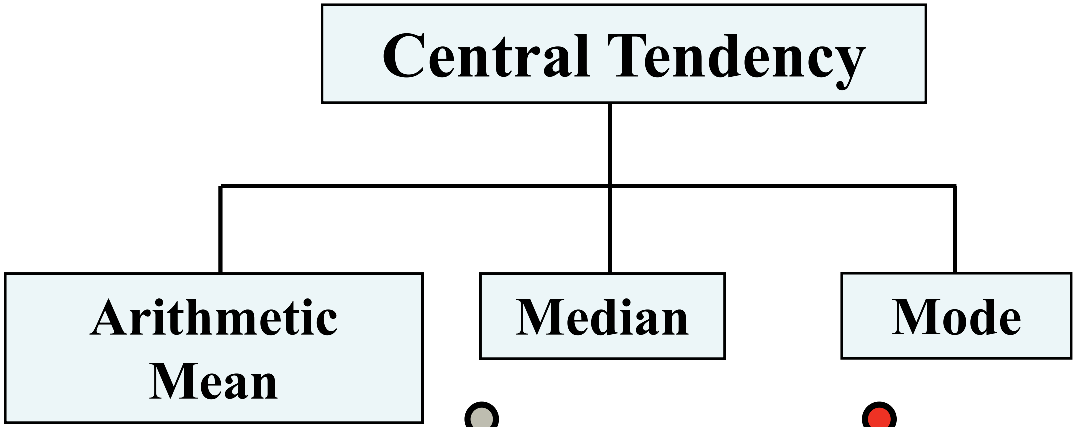
	Week	Year
Mean Gross Income	2,063	107,276
Median Gross Income	1,548	80,496

Australian Bureau of Statistics (ABS) 2014, *Household Income and Wealth, Australia*, 2015, cat. no. 6523.0

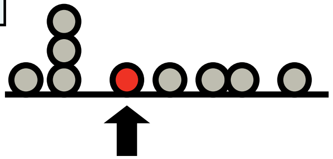
Which should you use?

- 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

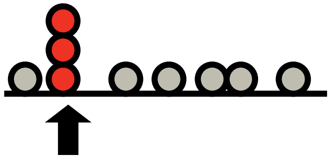
Summary of Central Tendency



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

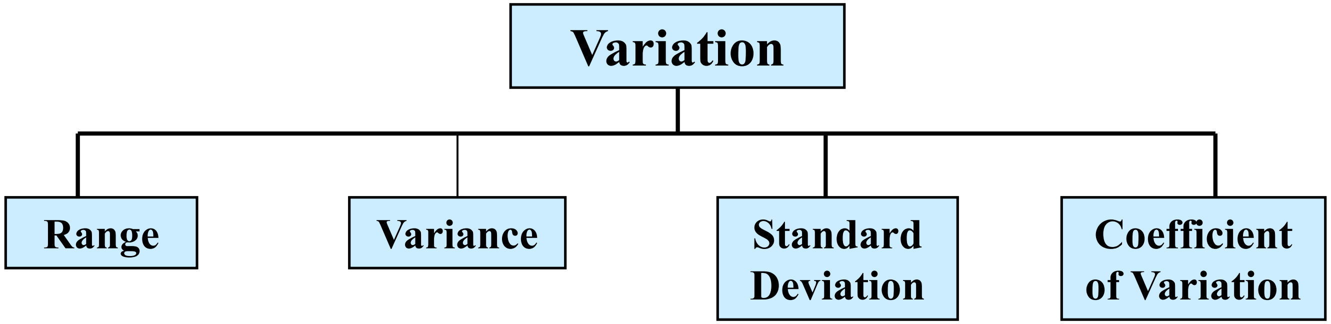


Middle value in the ordered array



Most frequently observed value

Measures of Variation



Variation

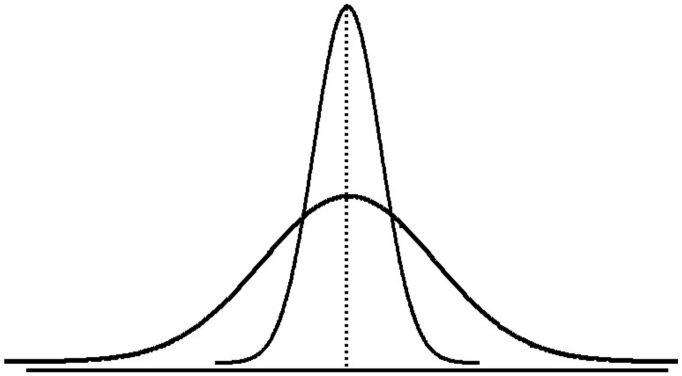
Range

Variance

Standard Deviation

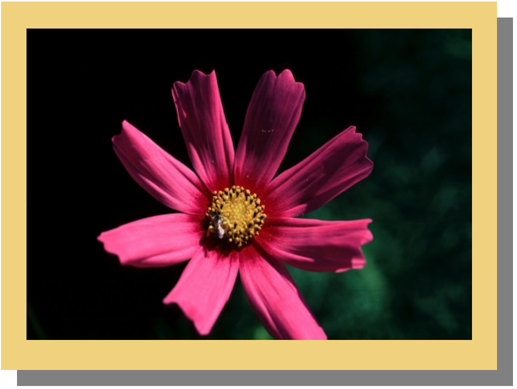
Coefficient of Variation

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



Same center, Different variation

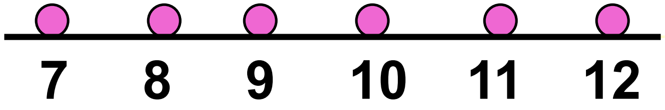
The Range



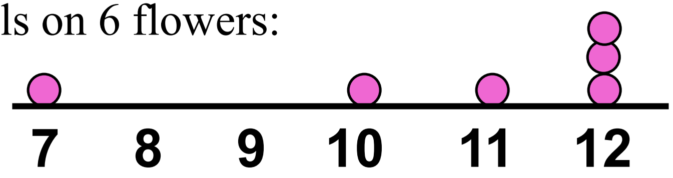
- Simplest measure of variation
- Difference between largest and smallest values

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

- Doesn't account for distribution
- **Example:** A botanist records the number of petals on 6 flowers:



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**

Sample Variance

- Average squared deviations from the mean, defined as

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

where $\overline{X}$ = arithmetic mean

n = sample size

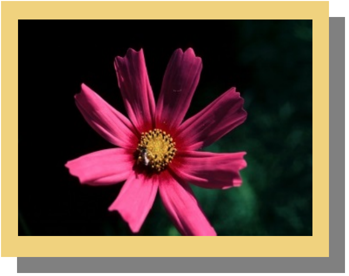
X_i = ith value of the variable X

Sample 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

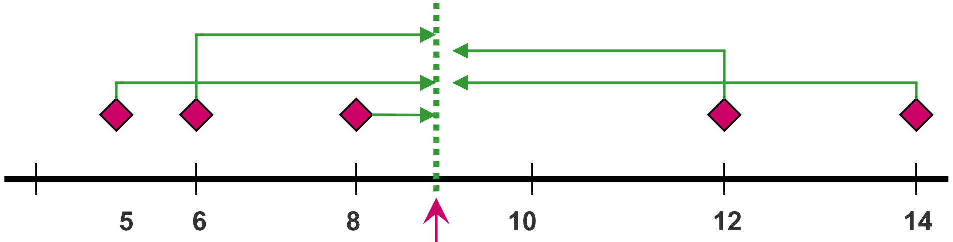
$$\boxed{S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n - 1}}}$$

The Variance

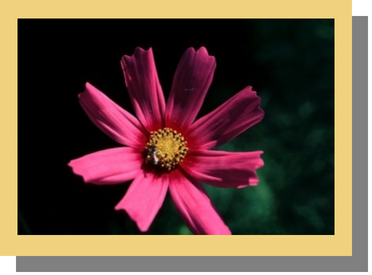


- The **variance** is measure of variability that uses all the measurements. It measures the average deviation of the measurements about their mean.
- **Flower petals: 5, 12, 6, 8, 14**

$$\boxed{\bar{x} = \frac{45}{5} = 9}$$



Calculate the Sample Variance



	x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
	5	-4	16
	12	3	9
	6	-3	9
	8	-1	1
	14	5	25
Sum	45	0	60

Use the Definition Formula:

$$s^2 = \frac{\sum(x_i - \overline{x})^2}{n-1}$$

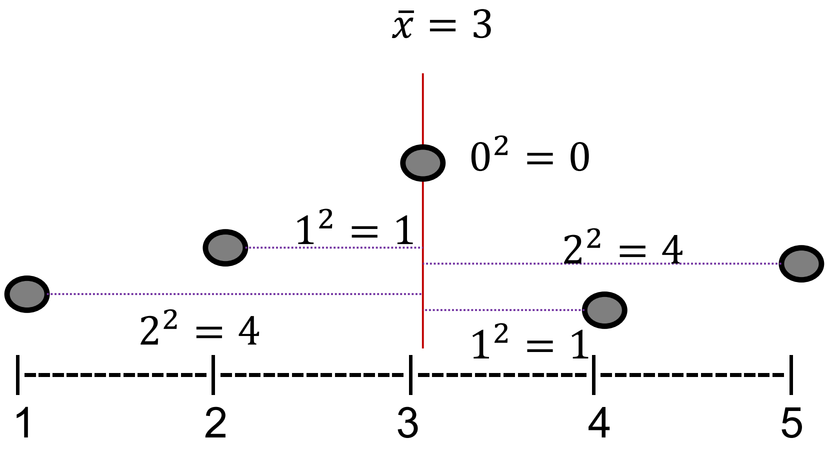
$$= \frac{60}{4} = 15$$

$$s = \sqrt{s^2} = \sqrt{15} = 3.87$$

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

Toy Example: 1, 2, 3, 4, 5



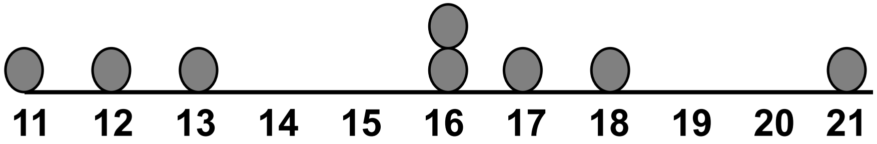
$$S^2 = \frac{4+1+0+1+4}{5-1} = \frac{10}{4} = 2.5$$

$$S = \sqrt{2.5} \approx 1.58$$

A measure of "Average" scatter about the mean

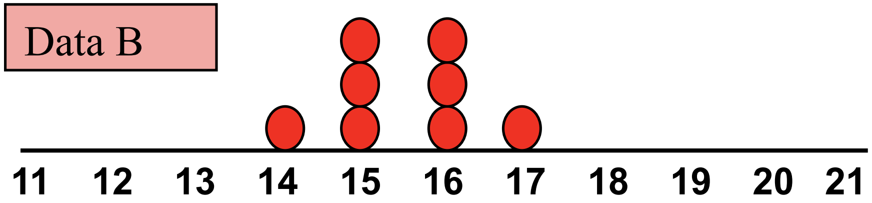
Comparing Standard Deviations

Data A



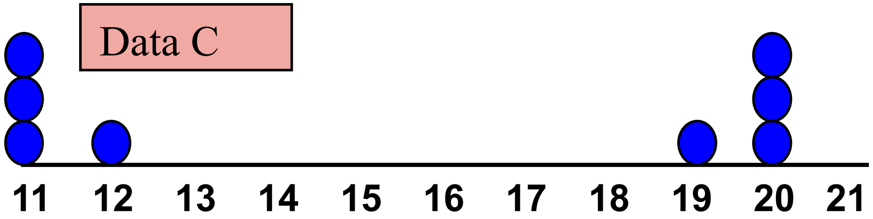
Mean = 15.5
S = 3.338

Data B

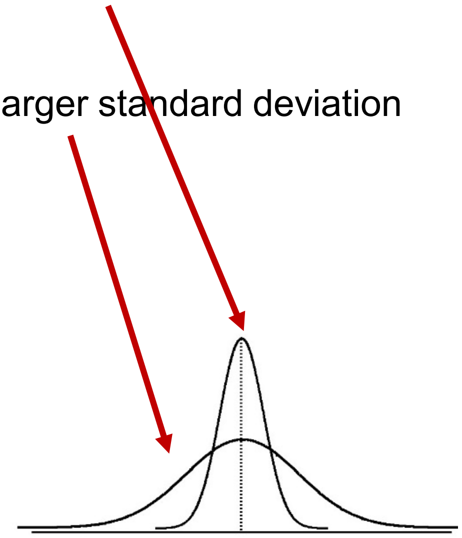


Mean = 15.5
S = 0.926

Data C



Mean = 15.5
S = 4.567



Smaller standard deviation

Larger standard deviation

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

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

Coefficient of Variation

- **Stock A:**
  - Average price last year = $50
  - Standard deviation = $5
- **Stock B:**
  - Average price last year = $100
  - Standard deviation = $5
- **Stock C:**
  - Average price last year = $8
  - Standard deviation = $2

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

Same SD, but one has lower CV

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

One has lower SD but higher CV

$$CV_C = \left(\frac{S}{\bar{X}}\right) \cdot 100\% = \frac{\$2}{\$8} \cdot 100\% = 25\%$$

Covariance

- The covariance measures the strength of the linear relationship between two numerical variables (X & Y)

- The sample covariance:

$$\boxed{\operatorname{cov}(X,Y)=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{n-1}}$$

  - Concerned with the strength of the relationship
  - No causal effect is implied

Correlation

- Covariance between two variables:
  - $\text{cov}(X,Y) > 0$ X and Y tend to move in the same direction
  - $\text{cov}(X,Y) < 0$ X and Y tend to move in opposite directions
  - $\text{cov}(X,Y) = 0$ X and Y are uncorrelated
- It is not possible to determine the relative strength of the relationship from the size of the covariance
  - For this, we use the coefficient of correlation:

$$\boxed{r = \frac{\text{cov}(X,Y)}{S_X S_Y}}$$

  - We’ll cover this in detail when we reach regression

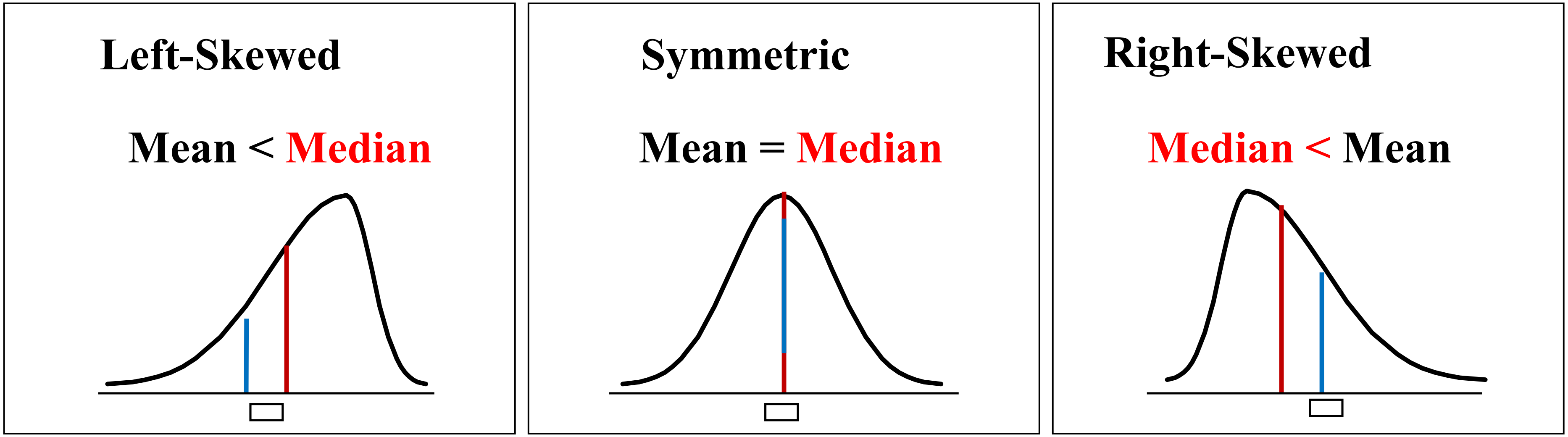
Summary of Variation Measures

- 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.

Shapes of Distributions

- Describes how data are distributed
- Two useful shape related statistics are:
  - Skewness
    - Measures the extent to which data values are not symmetrical
  - Kurtosis
    - Kurtosis affects the “peakedness” of the curve of the distribution—that is, how sharply the curve rises approaching the modal value

Skewness



Left-Skewed

Mean < Median

Symmetric

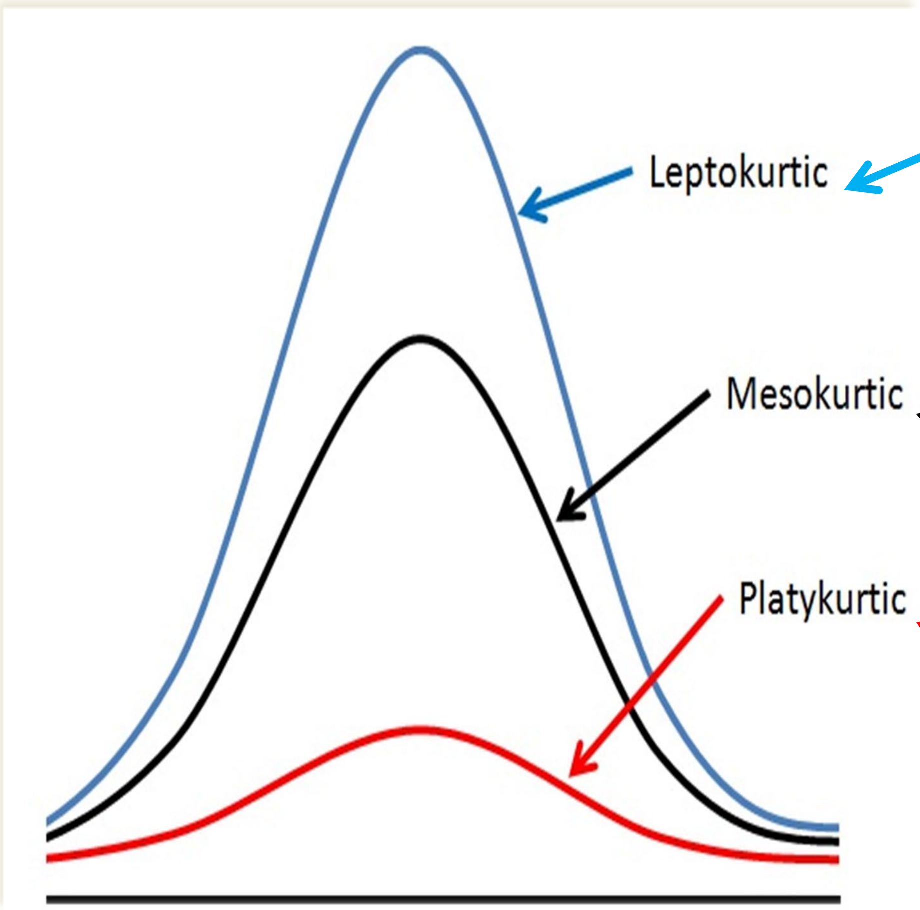
Mean = Median

Right-Skewed

Median < Mean

Skewness Statistic	< 0	0	>0

Kurtosis

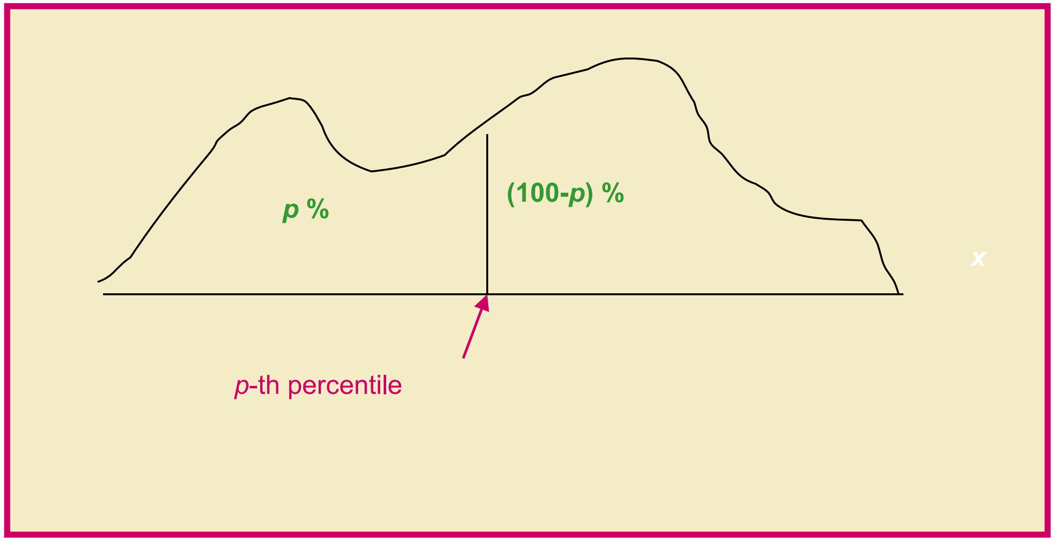


Sharper Peak
Than Bell-Shaped
(Kurtosis > 0)

Bell-Shaped
(Kurtosis = 0)

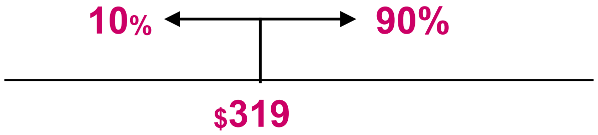
Flatter Than
Bell-Shaped
(Kurtosis < 0)

- The pth percentile is a value so that **roughly p% of the data are smaller and (100-p)% of the data are larger.**



Examples

- 90% of all men (16 and older) earn more than $319 per week.



$319 is the 10th percentile.

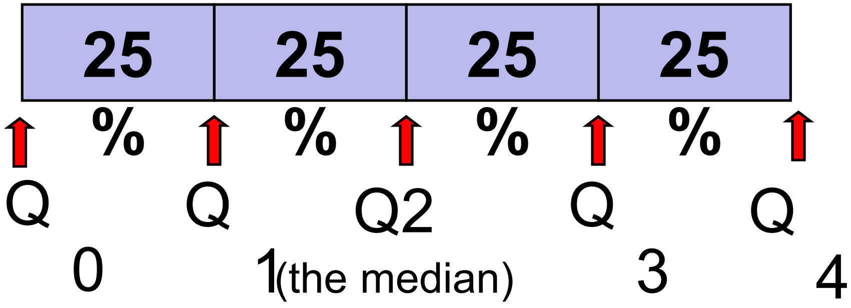
50th Percentile $\equiv$ Median (Q_2)

25th Percentile $\equiv$ Lower Quartile (Q_1)

75th Percentile $\equiv$ Upper Quartile (Q_3)

Quartiles

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



25 %

25 %

25 %

25 %

Q 0

Q 1

Q2 (the median)

Q 3

Q 4

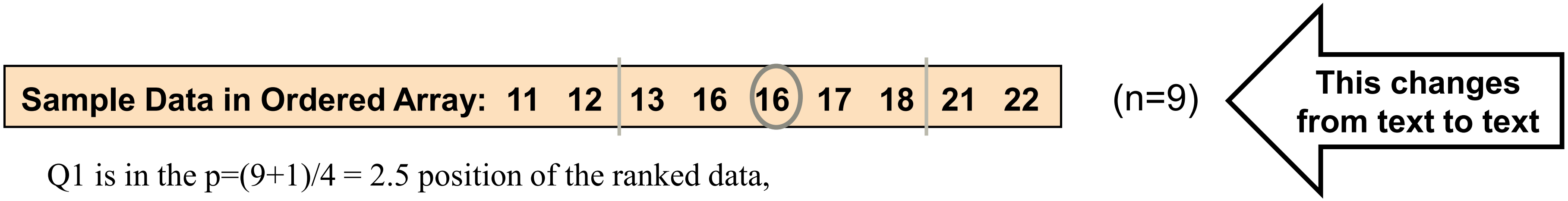
- The first quartile, Q1, is the value for which 25% of the observations are smaller and 75% are larger
- Q2 is the same as the median (50% of the observations are smaller and 50% are larger)
- Only 25% of the observations are greater than the third quartile Q3
- Q0 is the minimum and Q4 is the maximum.

Calculating Q1 and Q3

- To find Q1 & Q3, we first determine the values in the appropriate position in the ranked data, where
  - First quartile Q1 position: p= {(n+1)/4} = 0.25(n+1)
    - Q1 = {(n+1)/4}th ranked data value
  - Third quartile Q3 position: p= {3(n+1)/4} =0.75(n+1)
    - Q3 = {3(n+1)/4}th ranked data value
  - where n is the number of observed values

Calculating Q1 and Q3

- 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 we use linear interpolation on the two adjacent ranks to calculate Q1 and Q3:



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

(n=9)

This changes from text to text

Q1 is in the $p=(9+1)/4 = 2.5$ position of the ranked data,

so $Q1 = x(2)+0.5*(x(3) - x(2)) = 12+0.5(13-12)=12.5$

Q2 is in the $p=(9+1)/2 = 5$th position of the ranked data,

so $Q2 = \text{median} = x(5) = 16$

Q3 is in the $p=3(9+1)/4 = 7.5$ position of the ranked data,

so $Q3 = X(7)+0.5*(X(8) - X(7)) = 18+0.5(21-18) = 19.5$

Note: if $p=2.25$ for Q1 and 7.75 for Q3, then the fraction in front of the brackets (currently 0.5) would be replaced by 0.25 and 0.75, respectively.

Measures of Variation Interquartile Range

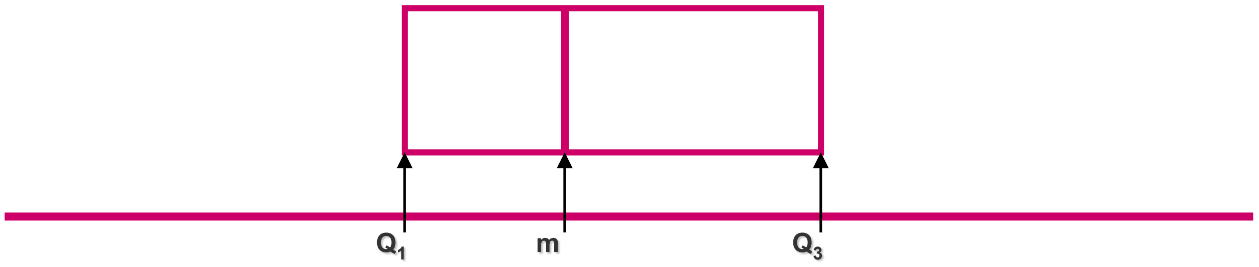
- The IQR is $Q_3 – Q_1$ and measures the spread in the middle 50% of the data
- The IQR is also called the midspread because it covers the middle 50% of the data
- The IQR is a measure of variability that is not influenced by outliers or extreme values
- Measures like Q_1, Q_3, and IQR that are not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

Constructing a Box Plot

✓Calculate Q_1, the median, Q_3 and IQR.

✓Draw a horizontal line to represent the scale of measurement.

✓Draw a box using Q_1, the median, Q_3.

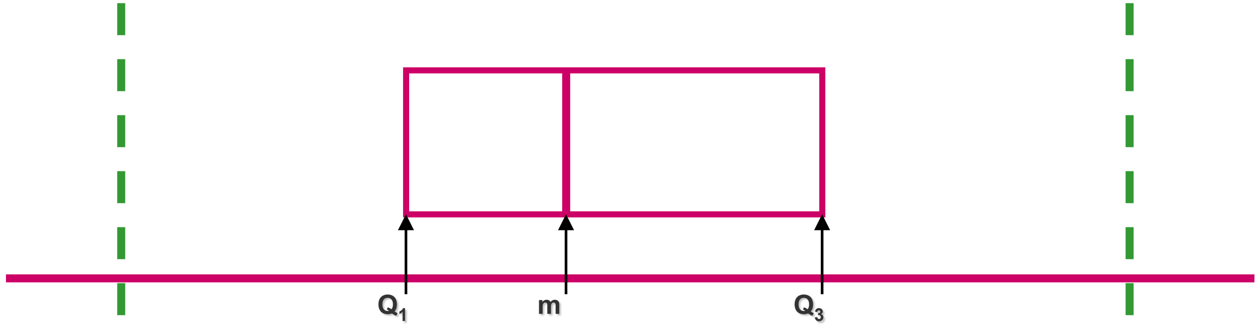


Constructing a Box Plot

✓Isolate outliers by calculating

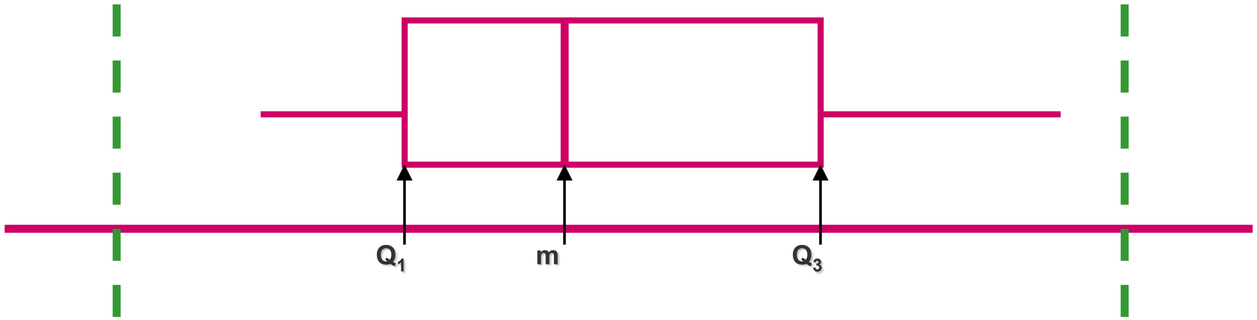
- ✓Lower fence: Q_1-1.5 IQR
- ✓Upper fence: Q_3+1.5 IQR

✓Measurements beyond the upper or lower fence are outliers and are marked (*).

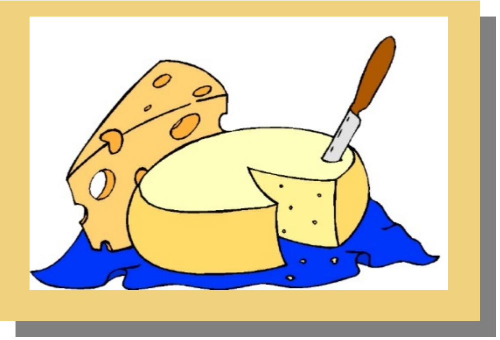


Constructing a Box Plot

✓Draw "whiskers" connecting the largest and smallest measurements that are NOT outliers to the box.



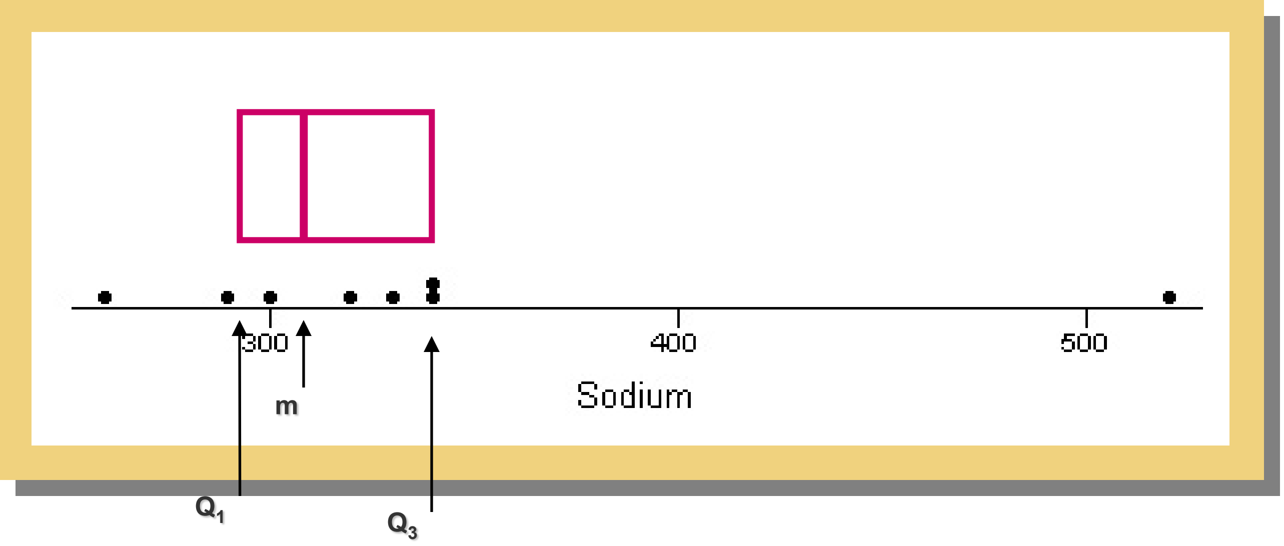
Example



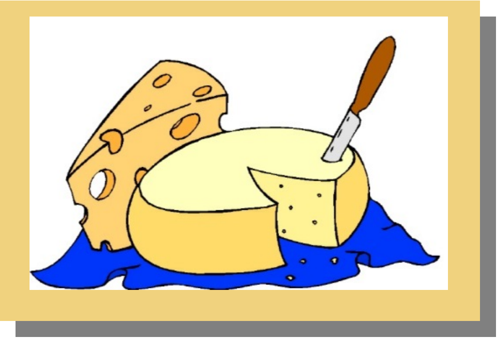
Amount of sodium in 8 brands of cheese:

260   290   300   320   330   340   340   520

$Q_1 = 292.5$ m = 325 $Q_3 = 340$



Example

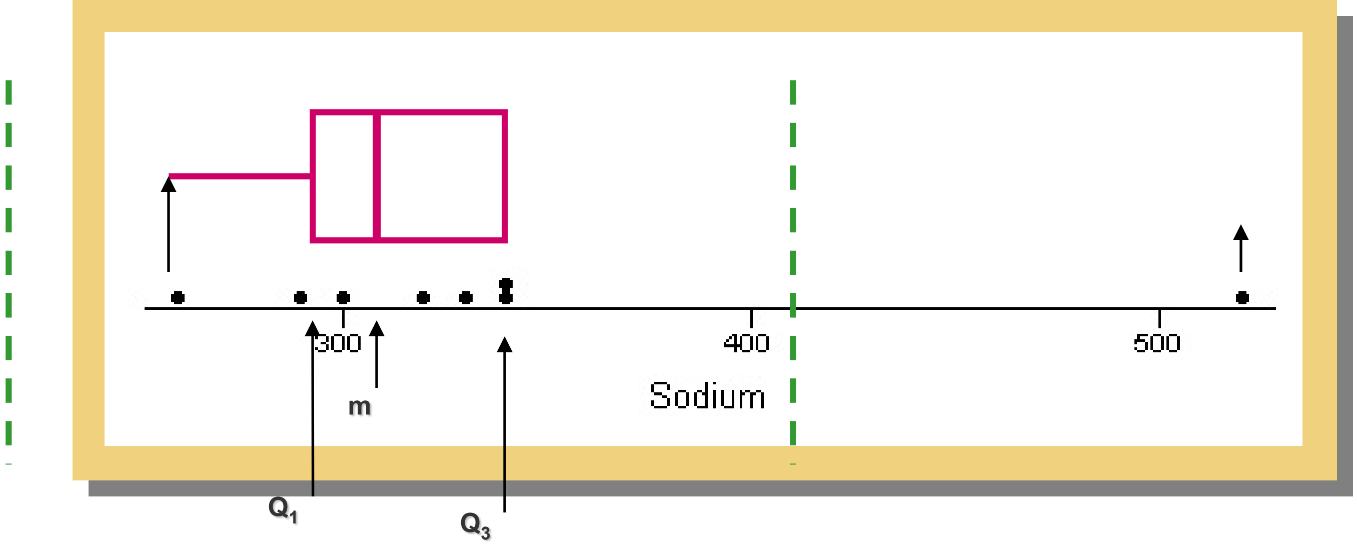


IQR = 340-292.5 = 47.5

Lower fence = 292.5-1.5(47.5) = 221.25

Upper fence = 340 + 1.5(47.5) = 411.25

Outlier: x = 520

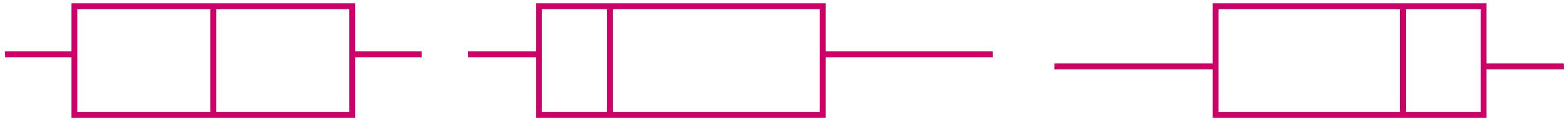


Interpreting Box Plots

✓Median line in center of box and whiskers of equal length—symmetric distribution

✓Median line left of center and long right whisker—skewed right

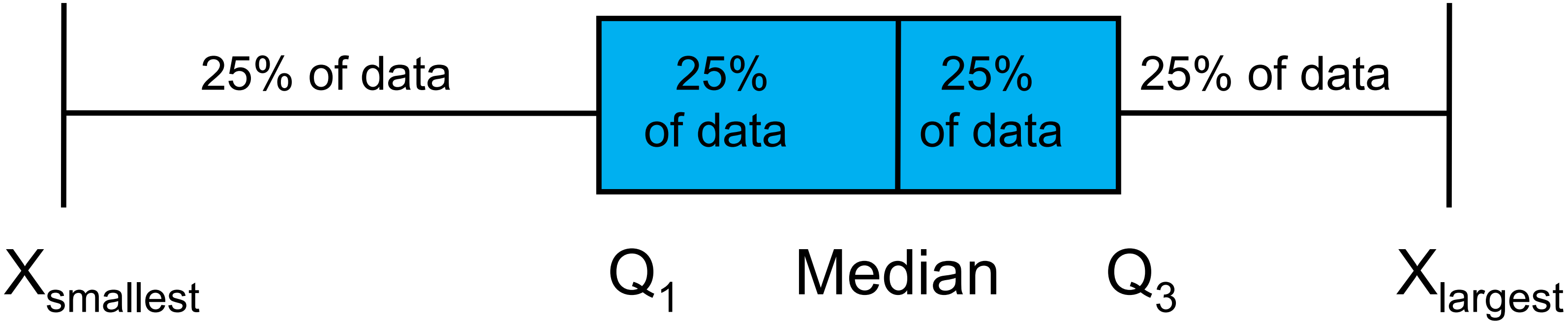
✓Median line right of center and long left whisker—skewed left



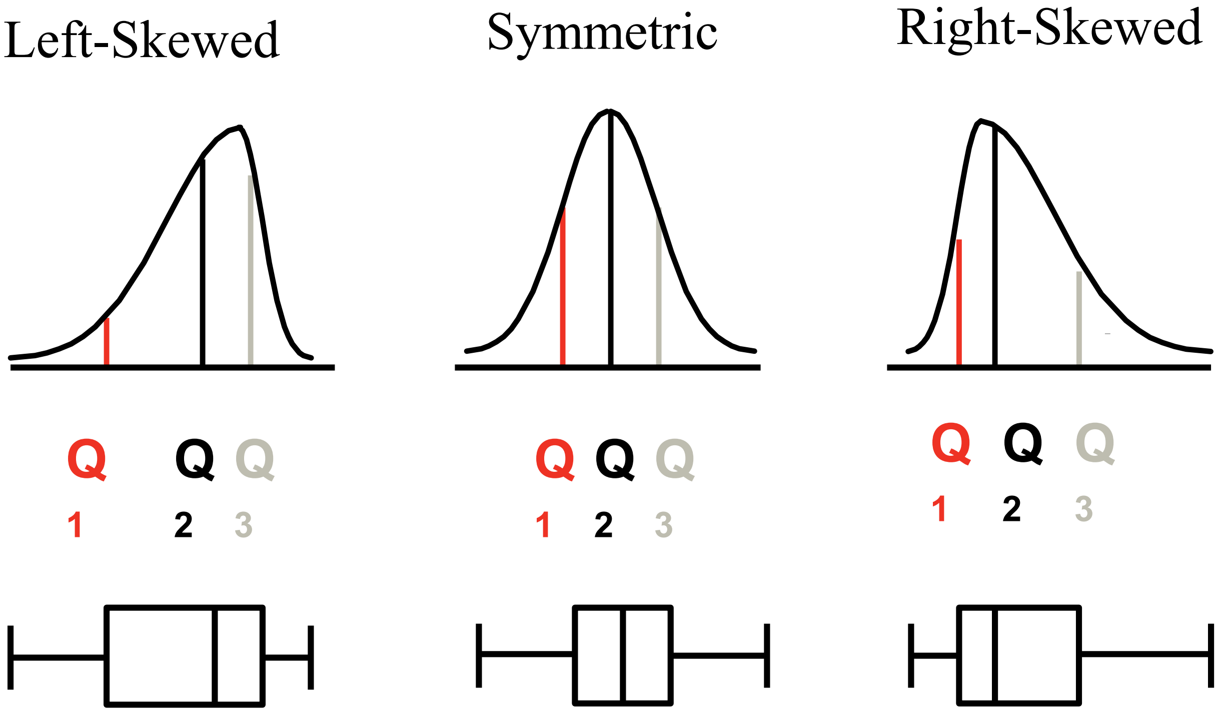
The 5-number summary and Boxplots

- The five numbers that help describe the center, spread and shape of data are:
  - Minimum Value (Q0)
  - First Quartile (Q1)
  - Median (Q2)
  - Third Quartile (Q3)
  - Maximum Value (Q4)

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



Boxplots



Left-Skewed

Symmetric

Right-Skewed

Identifying Outliers

- Small outliers: 1.5 X IQR

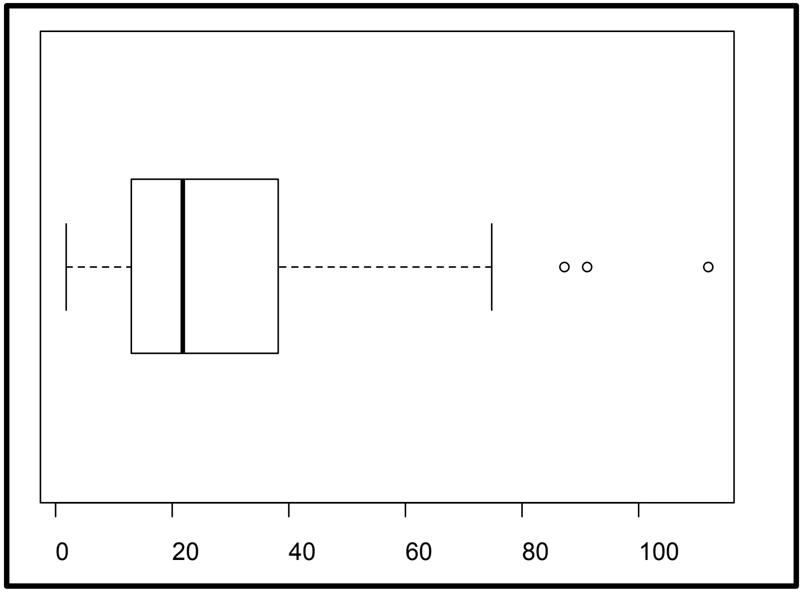
$$\text{Lower Limit}: \; Q_1 - 1.5 \times IQR$$

$$\text{Upper Limit}: \; Q_3 + 1.5 \times IQR$$

- Large outliers: 3 X IQR

$$\text{Lower Limit}: \; Q_1 - 3 \times IQR$$

$$\text{Upper Limit}: \; Q_3 + 3 \times IQR$$



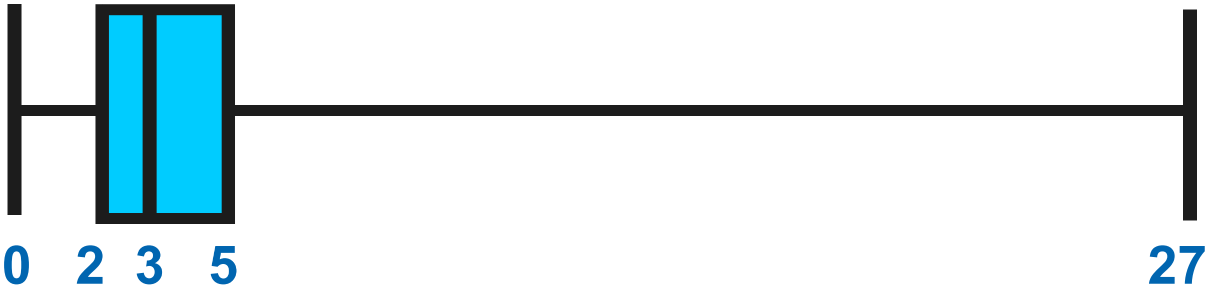
The whisker's lengths are determined by either Xmin/Xmax or limits of the small outliers, whichever closer to their corresponding side of the box.

Boxplot Example

- Below is a Boxplot for the following data:

$X_{smallest}$ Q_1 Q_2 / Median Q_3 $X_{largest}$

0 2 2 2 3 3 4 5 5 9 27



- The data are right skewed, as the plot depicts
  - Are there any outliers?