

# Measures of Central Tendency

A measure of central tendency is a value that represents a typical, or central, entry of a data set. The three most commonly used measures of central tendency are the mean, the median, and the mode.



# Population and sample sizes

Population  
(N)




Sample  
(n)

# Mean ( $\mu$ or $\bar{x}$ )

The **mean** of a data set is the sum of the data entries divided by the number of entries.

Population mean:  $\mu = \frac{\sum_{i=1}^N x_i}{N}$     Sample mean:  $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$



“mu”                      “x-bar”

**Example:** the following are the ages of all seven employees of a small company:

53      42      60      57      51      44      57

Calculate the mean.

$$\mu = \frac{\sum x}{N} = \frac{364}{7} \quad \text{Add the ages and divide by 7.}$$
$$= 52$$

The mean age of the employees is 52 years.

# Median ( $\tilde{\mu}$ or $\tilde{x}$ )

The **median** of a data set is the value that lies in the middle of the data when the data set is ordered.

**Example:** calculate the median age of the seven employees.

53      42      60      57      51      44      57

To find the median, sort the data.

42      44      51      53      57      57      60

The median age of the employees is 53 years.

# Mode

The **mode** of a data set is the data entry or category that occurs with the greatest frequency. If no entry is repeated, the data set has no mode. If two entries occur with the same greatest frequency, each entry is a mode and the data set is called **bimodal**.

**Example:** find the mode of the ages of the seven employees.

53

42

60

57

51

44

57

The mode is 57 because it occurs the most times.

An **outlier** is a datum that is far from the other in the data set.

# Comparing the Mean, Median and Mode

**Example:** A 29-year-old employee joins the company and the ages of the employees are now:

53    42    60    57    51    44    57    23

- Recalculate the mean, the median, and the mode.
- Which measure of central tendency was affected when this new age was added?

Mean = 48.4    The mean takes every value into account, but is affected by the outlier.

Median = 52    The median and mode are not very influenced by extreme values.

Mode = 57

# Weighted Mean

A **weighted mean** is the mean of a data set whose entries have varying weights. A weighted mean is given by

$$\bar{x} = \frac{\sum x \cdot w}{\sum w}$$

where  $w$  is the weight of each entry  $x$ .



**Example:** grades in a statistics class are weighted as follows.

Tests are worth 50% of the grade, homework is worth 30% of the grade and the final is worth 20% of the grade.

A student receives a total of 80 points on tests, 100 points on homework, and 85 points on his final. What is his current grade?

Begin by organizing the data in a table.

Source	Score, $x$	Weight, $w$	$x w$
Tests	80	0.50	40
Homework	100	0.30	30
Final	85	0.20	17

$$\bar{x} = \frac{\sum x \cdot w}{\sum w} = \frac{87}{1} = 87$$

The student's current grade is 87.

<b>Provides:</b>	<b>Nominal</b>	<b>Ordinal</b>	<b>Interval</b>	<b>Ratio</b>
The "order" of values is known		✓	✓	✓
"Counts," aka "Frequency of Distribution"	✓	✓	✓	✓
Mode	✓	✓	✓	✓
Median		✓	✓	✓
Mean			✓	✓
Can quantify the difference between each value			✓	✓
Can add or subtract values			✓	✓
Can multiple and divide values				✓
Has "true zero"				✓

# Measures of variation



# Range

The **range** of a data set is the difference between the maximum and minimum data entries in the set.

$$\text{Range} = (\text{Maximum data entry}) - (\text{Minimum data entry})$$

Example:

The following data are the closing prices for a certain stock on ten successive Fridays. Find the range.

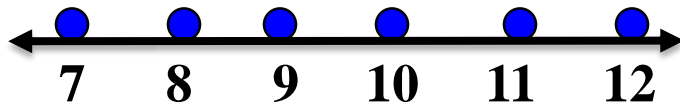
Stock	56	56	57	58	61	63	63	67	67	67
-------	----	----	----	----	----	----	----	----	----	----

The range is  $67 - 56 = 11$ .

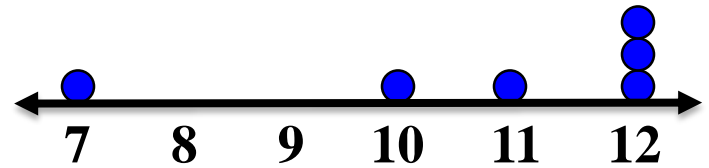
# Range

Ignores the way in which data are distributed.

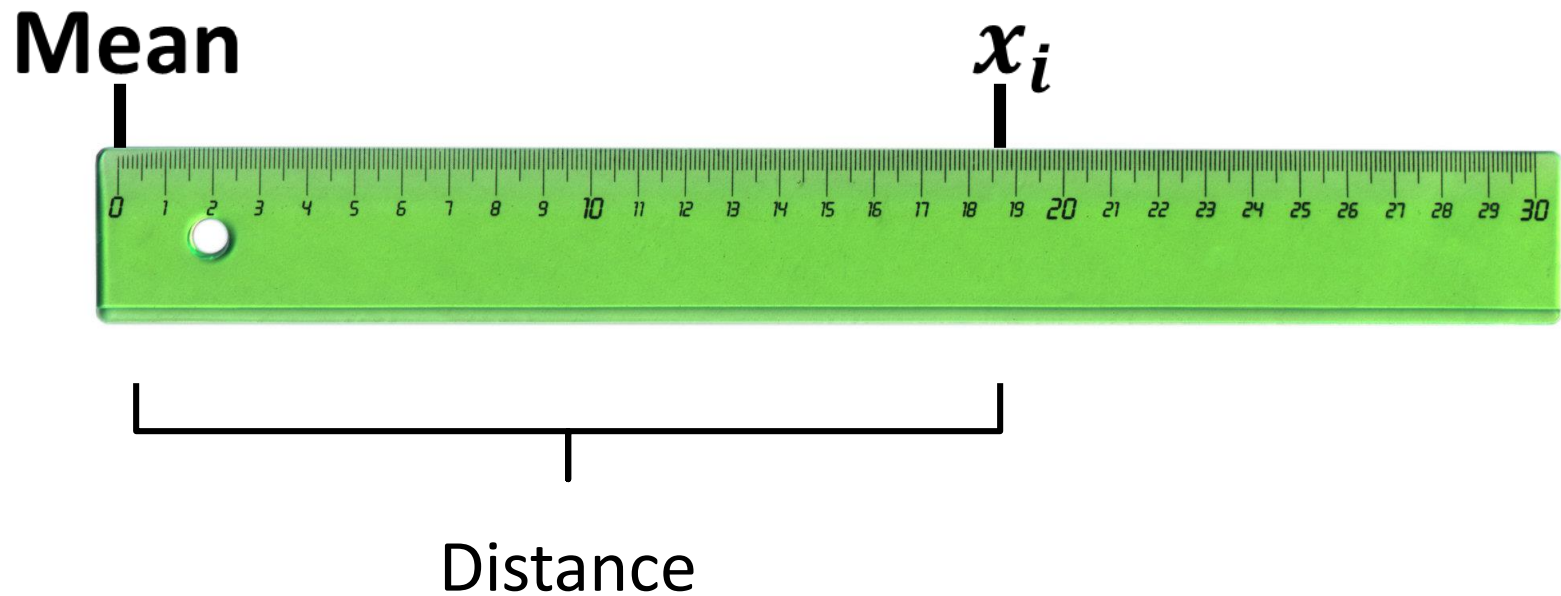
$$\text{Range} = 12 - 7 = 5$$



$$\text{Range} = 12 - 7 = 5$$



# Intuitive idea



# Population variance and standard deviation

The **population variance** of a population data set is

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

The **population standard deviation** of a population data set is

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



# Sample variance and standard deviation

The **sample variance** of a sample data set is

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

The **sample standard deviation** of a sample data set is

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

# Example

A statistic class with five students took a test with these test scores: 92, 95, 83, 76, 54.

Find the population variance and standard deviation for this class.

1) Find the mean:  $\mu = \frac{92+95+83+76+54}{5} = 80$

2) Find the deviation from the mean:

$$92-80=12 \quad 95-80=15 \quad 83-80=3$$

$$76-80=-4 \quad 54-80=-26$$

3) Square the deviation from the mean:

$$(12)^2=144 \quad (15)^2=225 \quad (3)^2=9$$

$$(-4)^2=16 \quad (-26)^2=676$$

4) Find the sum of the squares:

$$144 + 225 + 9 + 16 + 676 = 1070$$

5) Divide the sum of squares by the number of items

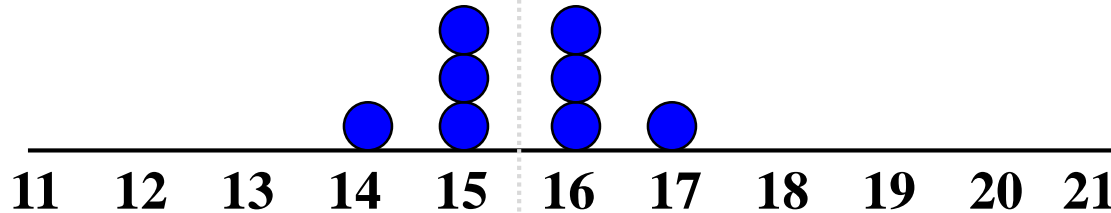
$$\sigma^2 = \frac{1070}{5} = 214 \text{ points}^2$$

6) Find the square root of the variance

$$\sigma = \sqrt{214} = 14.63 \text{ points}$$

# Comparing Standard Deviations

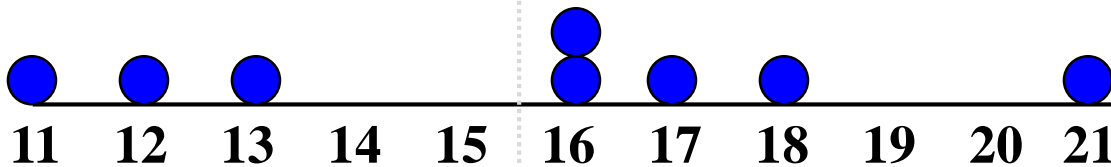
Data A



Mean = 15.5

$S = 0.9258$

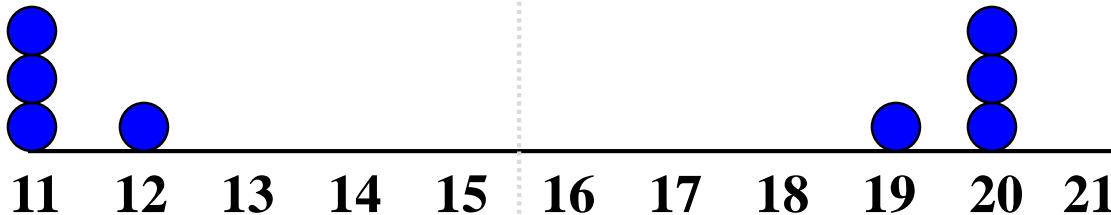
Data B



Mean = 15.5

$S = 3.338$

Data C



Mean = 15.5

$S = 4.57$

# Coefficient of variation

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

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

# Example



1000 ml pack

$$s = 15 \text{ ml}$$
$$\bar{x} = 1005 \text{ ml}$$

$$CV = \frac{s}{\bar{x}} 100\% = \frac{15 \text{ ml}}{1005 \text{ ml}} 100\% = 1.49\%$$



50 ml pack

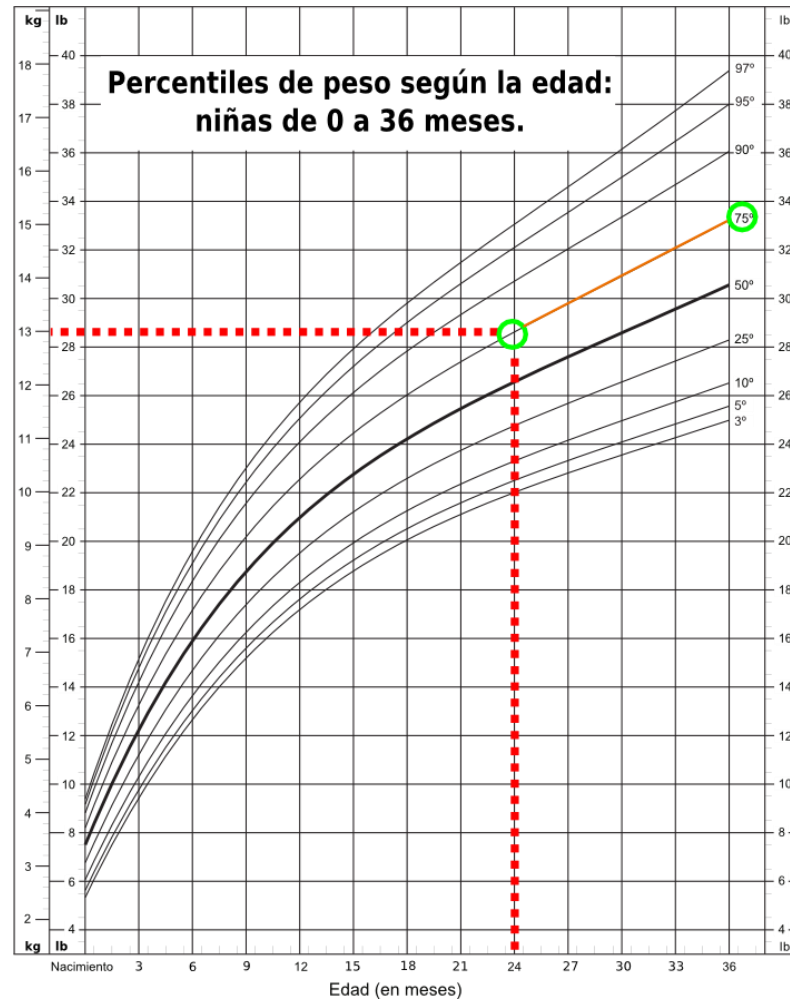
$$s = 3 \text{ ml}$$
$$\bar{x} = 45 \text{ ml}$$

$$CV = \frac{s}{\bar{x}} 100\% = \frac{3 \text{ ml}}{45 \text{ ml}} 100\% = 6.67\%$$

# Position measurement



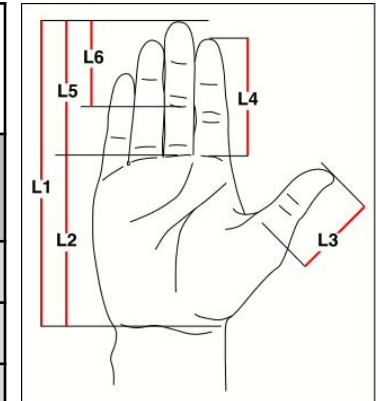
## Examples



## Hand anthropometry of non-disabled individuals

(Sources: DTI, 2002; Ergonomics for Schools, 2008; RoyMech, 2008)

Dimension	Gender	5th percentile (mm)	50th percentile (mm)	95th percentile (mm)
Hand length (L1)	Male	173-175	178-189	205-209
	Female	159-160	167-174	189-191
Palm length (L2)	Male	98	107	116
	Female	89	97	105
Thumb length (L3)	Male	44	51	58
	Female	40	47	53
Thumb breadth	Male	11-12	23	26-27
	Female	10-14	20-21	24
Index finger length (L4)	Male	64	72	79
	Female	60	67	74
Hand breadth	Male	78	87	95
	Female	69	76	83-85



# Order Statistics

$x_3$ : represents the 3rd observed datum in sample.

$x_{(3)}$ : represents the third ordered value, when the observations are ordered from the smallest to largest.

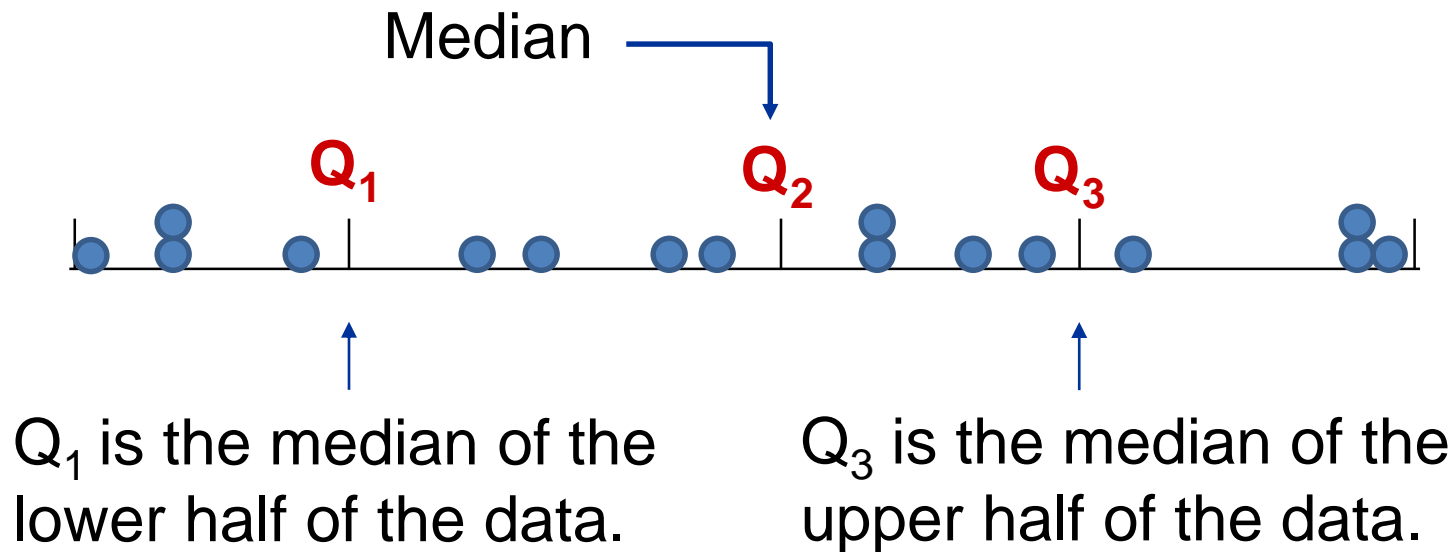
→ The 3rd order statistic.

**Example:** for the data 2, 6, -1, 8, 0, -1, 8, 6. Find the order statistics.

-1	-1	0	2	6	6	8	8
↓	↓	↓	↓	↓	↓	↓	↓
$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$

# Quartiles

The three **quartiles**,  $Q_1$ ,  $Q_2$ , and  $Q_3$ , approximately divide an **ordered data** set into four equal parts.



# Deciles and Percentiles

**Deciles:** approximately divide an **ordered data** set into 10 parts. There are 9 deciles:  $D_1, D_2, \dots, D_9$ .

**Percentiles:** approximately divide an **ordered data** set into 100 parts. There are 99 percentiles:  $P_1, P_2, P_3 \dots P_{99}$ .

**Quartiles, deciles** and **percentiles** are useful if one have a large number of observations.

**Example:** A test score at the 80th percentile  $P_{80}$  (or at the 8<sup>th</sup> decil  $D_8$ ), indicates that the test score is greater than 80% of all other test scores and less than or equal to 20% of the scores.

# Quantiles

- Quartiles, deciles and percentiles are also called **QUANTILES**.
- A general notation for the  **$p$ -quantile** is:

$$q_p, \text{ where } 0 < p < 1.$$

- $q_{0.1} = D_1, q_{0.2} = D_2, \dots, q_{0.9} = D_9$ , are deciles.
- $q_{0.25} = Q_1, q_{0.5} = Q_2, q_{0.75} = Q_3$ , are quartiles.
- $q_p = P_{100p}, p = 0.01, 0.02, \dots, 0.99$ , are percentiles.

# Calculating the sample $p$ -quantile

1. Order the data from smallest to largest.

$$x_{(1)}, x_{(2)}, \dots, x_{(k)}, x_{(k+1)}, \dots, x_{(n)}$$

2. Determine the product  $np$ .
  - If  $np$  is not an integer, round it up to the next integer, say  $k$ , find the corresponding ordered value, e. g.  $q_p = x_{(k)}$ .
  - If  $np$  is an integer, say  $k$ , calculate the average of the  $k$ -th and  $(k + 1)$ -st ordered values, e. g.  $q_p = (x_{(k)} + x_{(k+1)})/2$ .

# Finding Quartiles

**Example:** The quiz scores for 15 students is listed below. Find the first, second and third quartiles of the scores.

28 43 48 51 43 30 55 44 48 33 45 37 37 42 38

For  $Q_1$  we have:  $n(0.25) = 3.75 \approx 4$ .

For  $Q_2$  we have:  $n(0.5) = 7.50 \approx 8$ .

For  $Q_3$  we have:  $n(0.75) = 11.25 \approx 12$ .

**Then, in the ordered data:**

28 30 33 37 37 38 42 43 43 44 45 48 48 51 55

$q_{0.25} = Q_1 = x_{(4)}$        $q_{0.5} = Q_2 = x_{(8)}$        $q_{0.75} = Q_3 = x_{(12)}$

About one fourth of the students scores 37 or less; about one half score 43 or less; and about three fourths score 48 or less.



# Interquartile Range

The **interquartile range (IQR)** of a data set is the difference between the third and first quartiles.

$$\text{Interquartile range (IQR)} = Q_3 - Q_1.$$

## Example:

The quartiles for 15 quiz scores are listed below. Find the interquartile range.

$$Q_1 = 37$$

$$Q_2 = 43$$

$$Q_3 = 48$$

$$\begin{aligned}(\text{IQR}) &= Q_3 - Q_1 \\ &= 48 - 37 \\ &= 11\end{aligned}$$

The quiz scores in the middle portion of the data set vary by at most 11 points.

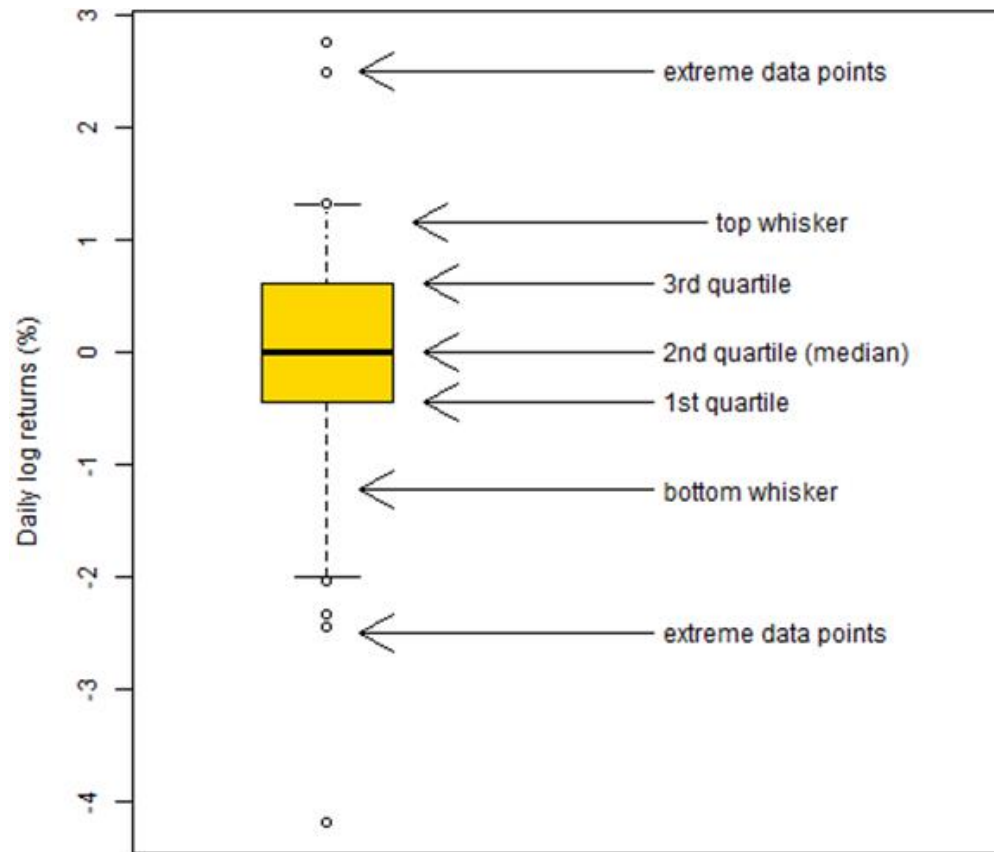
# Boxplot

A **boxplot** is an exploratory data analysis tool that highlights the important features of a data set.

The **five-number summary** is used to draw the graph.

- The minimum entry
- $Q_1$
- $Q_2$  (median)
- $Q_3$
- The maximum entry

# Parts of a boxplot



The maximum length of whisker from  $Q_1$  (or  $Q_3$ ) is 1.5 IQR. The whisker extends to the most extreme observation within 1.5 IQR units of  $Q_1$  (or  $Q_3$ ).

# Purpose of Boxplots

Boxplots let us to evaluate:



1. Variability
2. Outliers
3. Symmetry
4. Comparison of subpopulations

# Example

Use the data from the 15 quiz scores to draw a box-and-whisker plot.

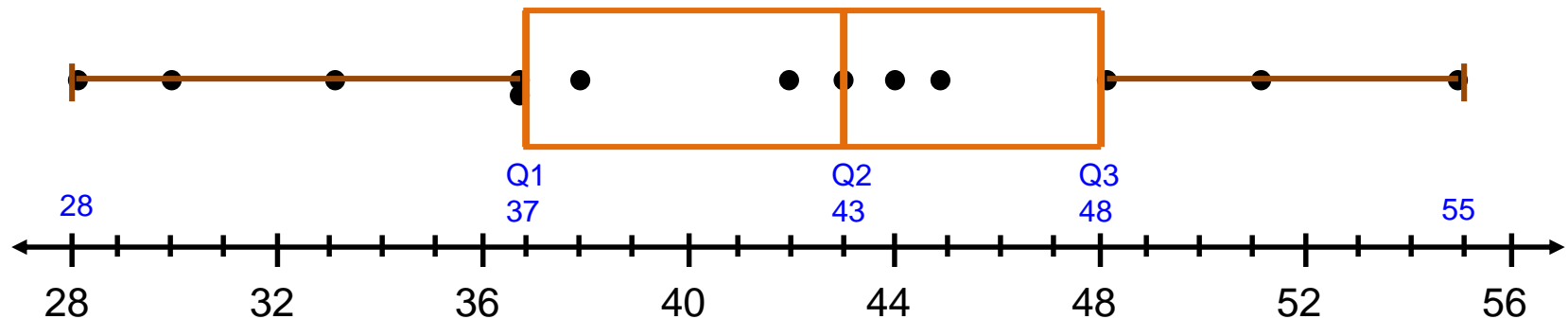
28 30 33 37 37 38 42 43 43 44 45 48 48 51 55

## Five-number summary

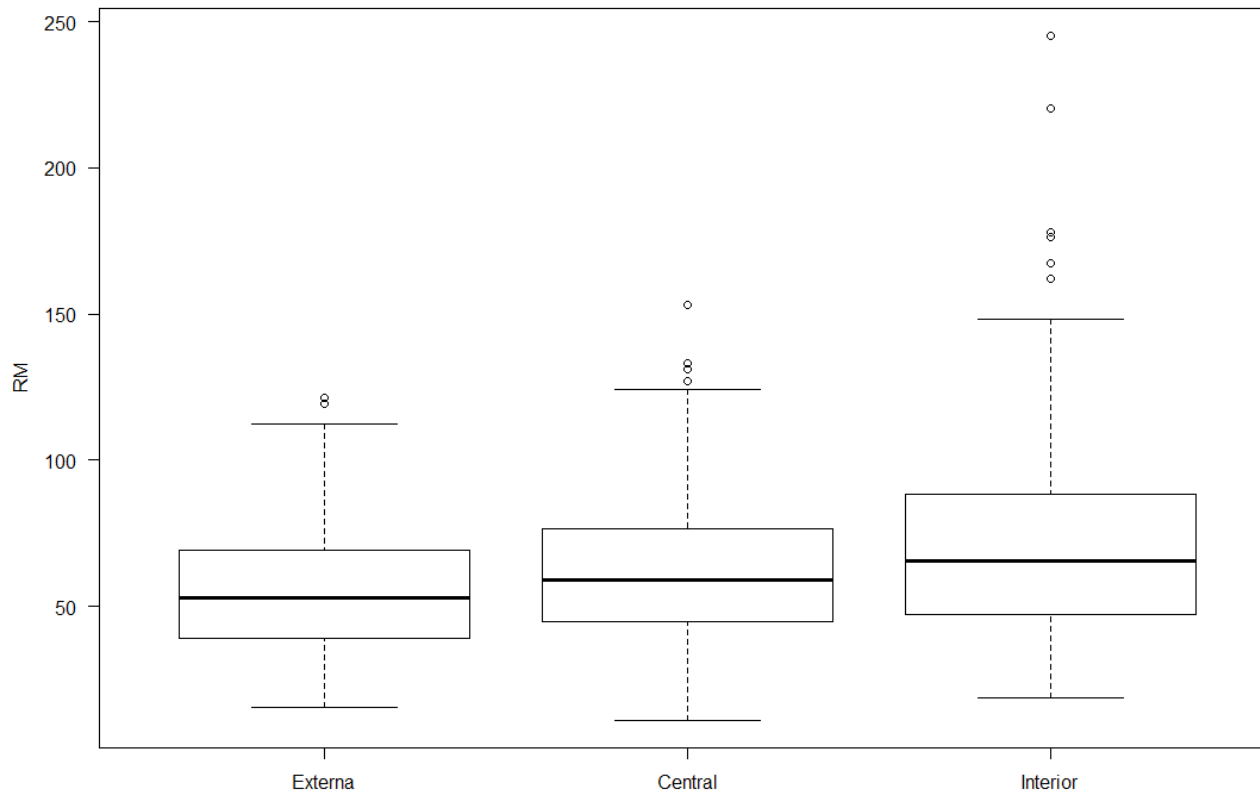
- The minimum entry 28
- $Q_1$  37
- $Q_2$  (median) 43
- $Q_3$  48
- The maximum entry 55

$$\text{IQR} = Q_3 - Q_1 = 11$$

$$\text{Max. length} = 1.5 \text{ IQR} = 16.5$$



# Example (comparison of subpopulations)



# Example (comparison of subpopulations)

