CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Descriptive statistics (handout)

Descriptive statistics

- Quantitative measures
  – a small number of values that summarise the main features of data
- Two categories
  – measures of central tendency (centre)
  e.g., mean, median, mode
  – measures of dispersion (spread)
  e.g., variance, standard deviation, IQR (interquartile range), range

Mean

- Most commonly called the “average”
  – add up the values for each case and divide by the total number of cases

Class -- IQs of 13 Students

102
128
131
98
140
93
110
115
109
89
106
119
97

Sum=(102+128+131+98+140+93+110
+115+109+89+106+119+97)=1437

Mean=1437/13=110.54

Mean

- Most commonly called the “average”
  – add up the values for each case and divide by the total number of cases

Sum=(102+128+131+98+140+93+110
+115+109+89+106+119+97)=1437
Mean=1437/13=110.54

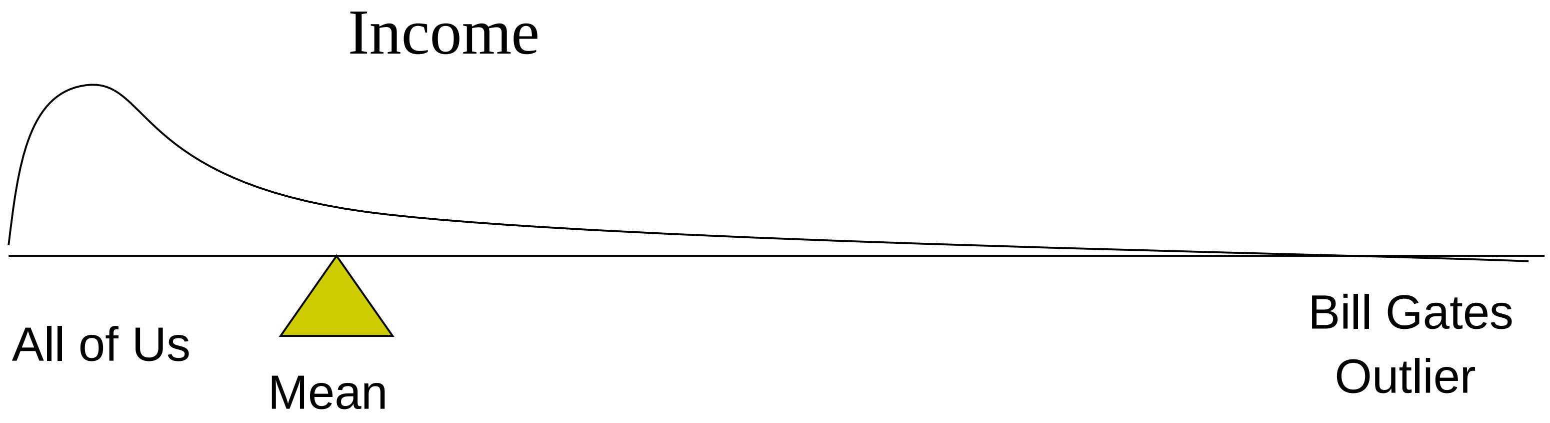
- The mean is the “balance point”
  – e.g., measure of common experience

Class -- IQs of 13 Students

102
128
131
98
140
93
110
115
109
89
106
119
97

Mean

- Means can be badly affected by outliers
  – data points with extreme values unlike the rest
  – outliers can make the mean a bad measure of central tendency or common experience



Median

- The middle value
  – when a variable’s values are ranked in order;
  – the point that divides a distribution into two equal halves
- When data are listed in order, the median is the point at which 50% of the cases are above and 50% below it
- The 50th percentile!

Median

89

93

97

98

102

106

109

110

115

119

128

131

140

Median = 109

(six cases above, six below)

Class -- IQs of 13 Students

102

128

131

98

140

93

110

115

109

89

106

119

97

Median

~~89~~

93

97

98

102

106

109

110

115

119

128

131

140

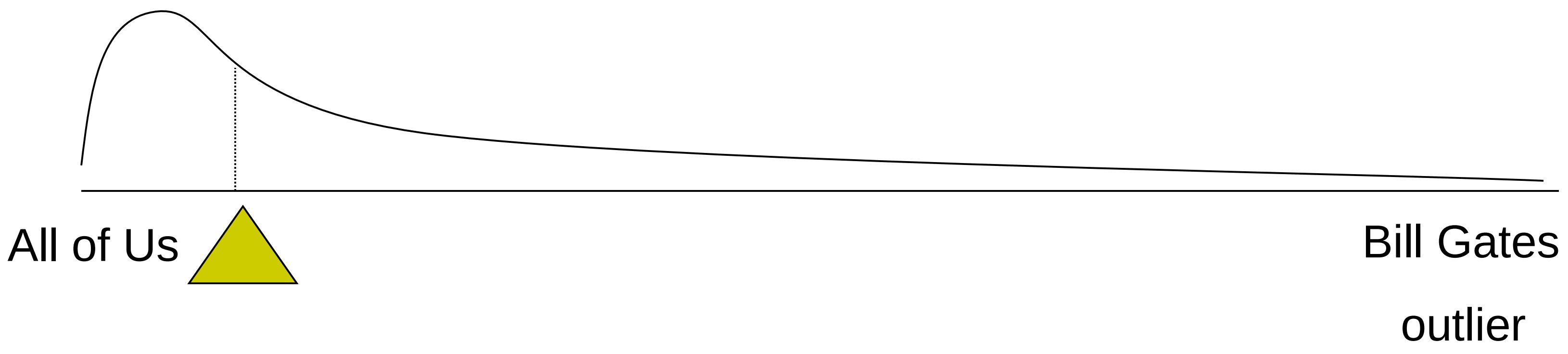
Median = 109.5

109 + 110 = 219/2 = 109.5

(six cases above, six below)

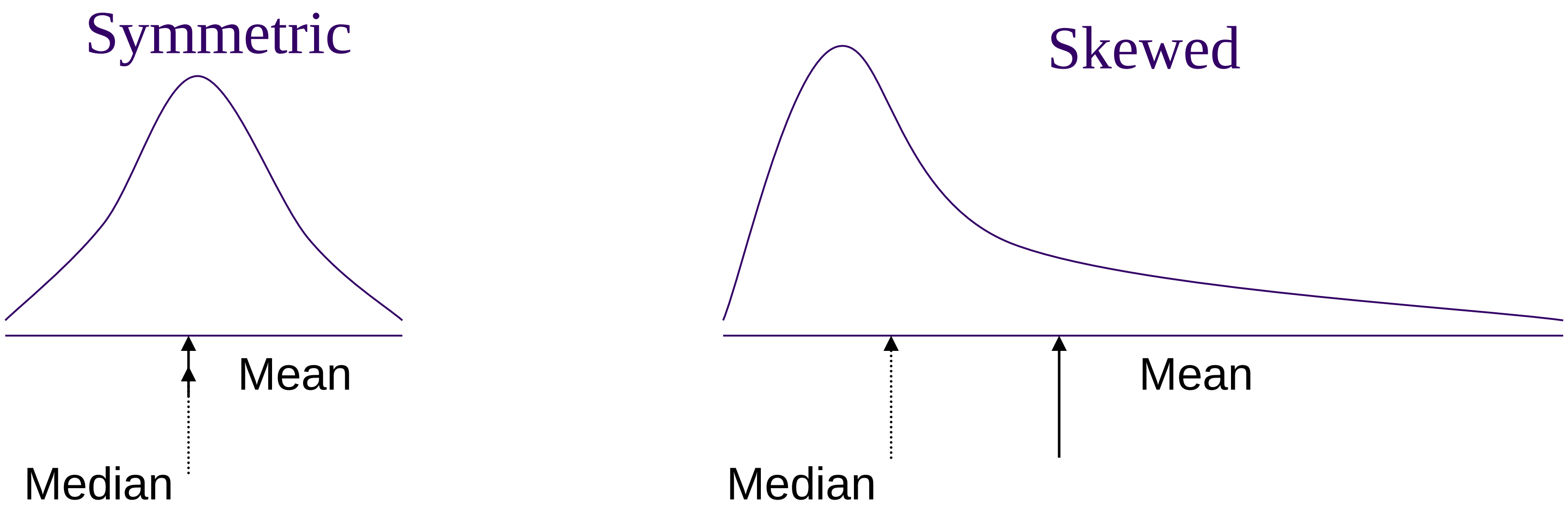
Median

- The median is unaffected by outliers
  – a better measure of central tendency
  – better describe the "typical person" than the mean when data are skewed



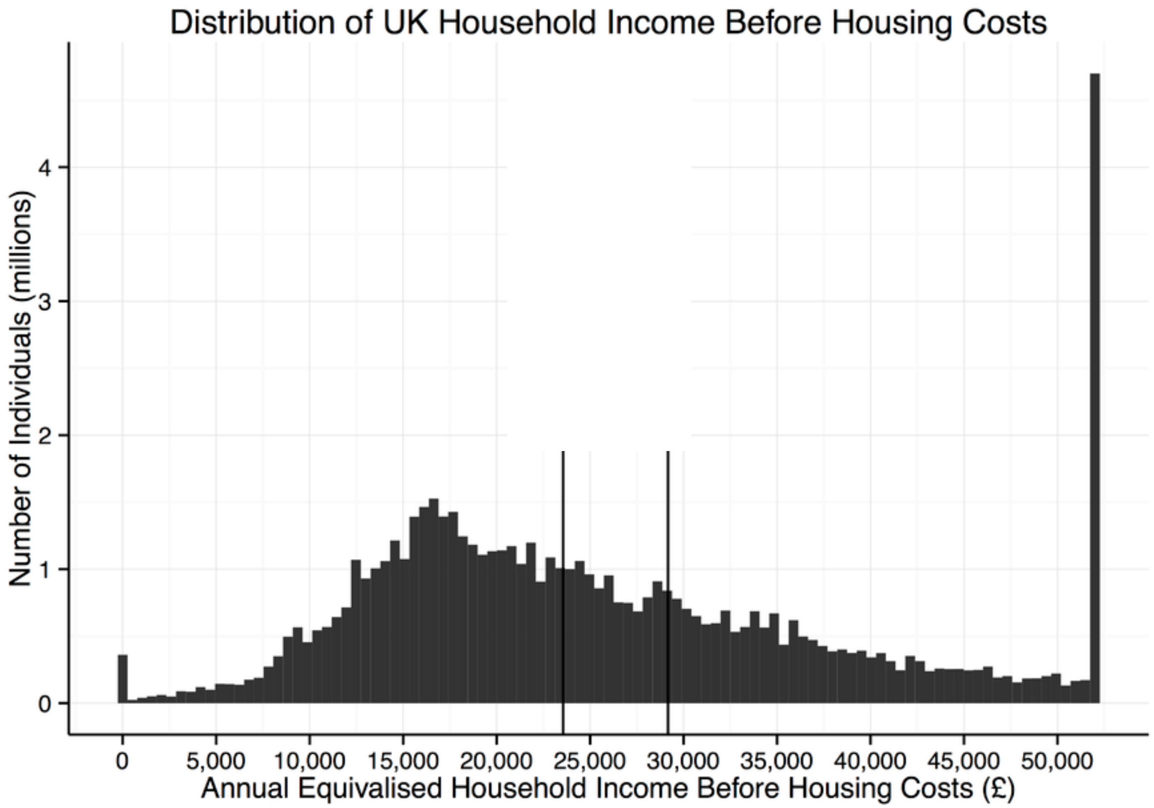
Median

- If the recorded values for a variable form a symmetric distribution, the median and mean are identical
- In skewed data, the mean lies further toward the skew than the median



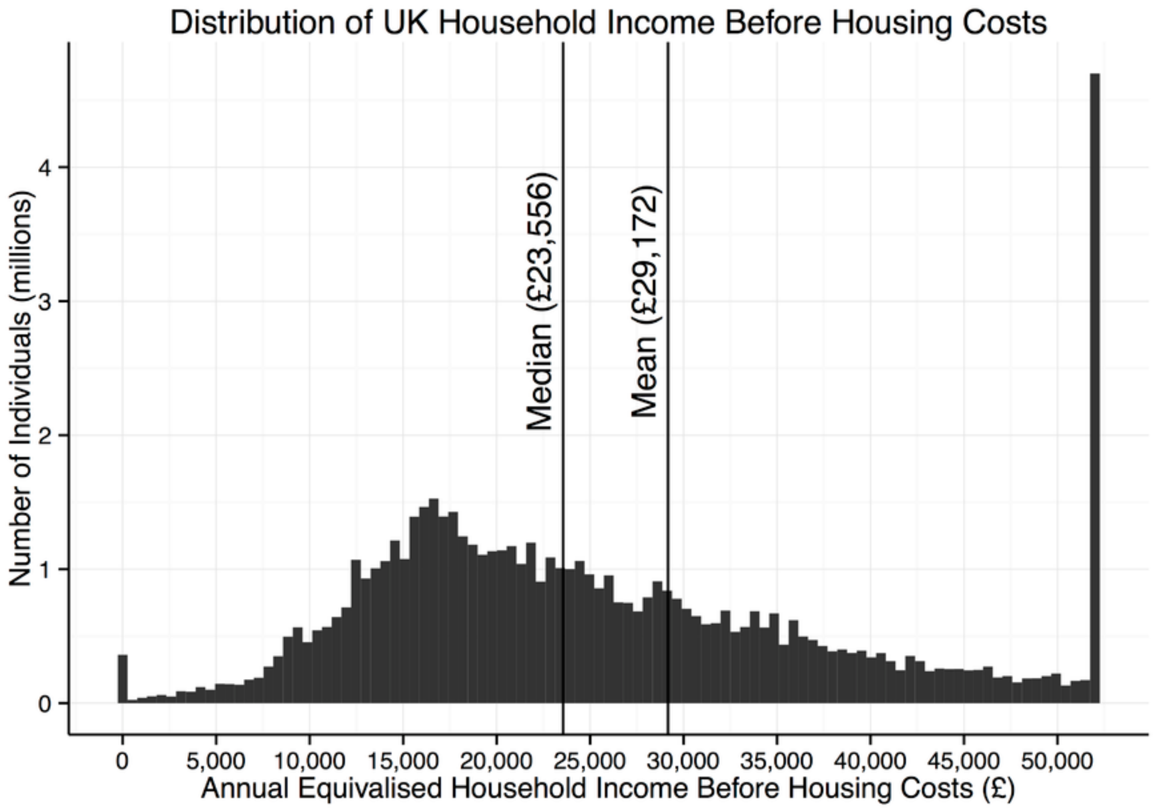
Income in the UK

- Data from the Households Below Average Income (HBAI) report from the Department of Work and Pensions 2013/14



Income in the UK

- Data from the Households Below Average Income (HBAI) report from the Department of Work and Pensions 2013/14



Mode

- The most common data point is called the mode
  – e.g., the IQ scores for a class
  80 87 89 93 93 96 97 98 102 103 105 106 109 109 109 110 111 115 119 120 127 128 131 131 140 162

Mode

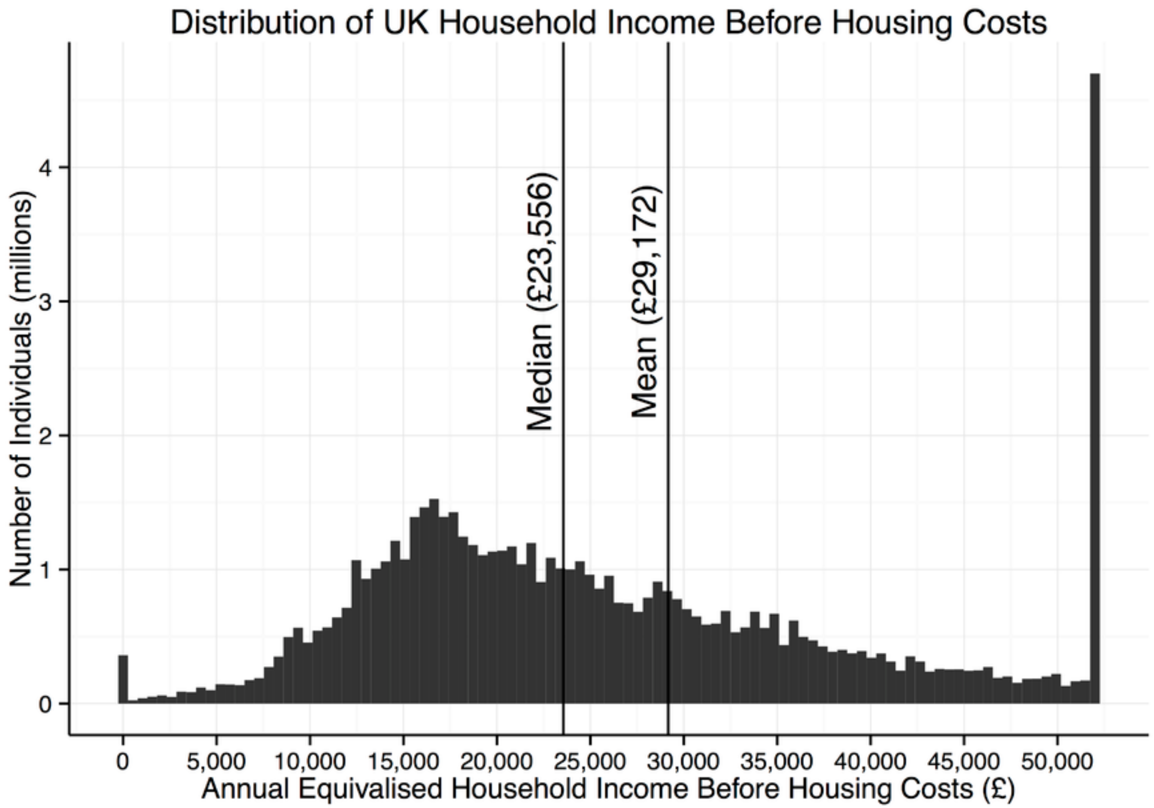
- The most common data point is called the mode
  – e.g., the IQ scores for a class
  80 87 89 93 93 96 97 98 102 103 105 106 109 109 109 110
  111 115 119 120 127 128 131 131 140 162

  mode!!

- BTW, it is possible to have more than one mode!

Mode

- It may not be at the centre of a distribution
- It may give you the most likely experience rather than the “typical” or “central” experience
- Where is the most likely UK household income?



Range

- The spread, or the distance, between the lowest and highest values of a variable
  – subtract its lowest value from its highest value

Range=140-89=51

Class -- IQs of 13 Students

102
128
131
98
140
93
110
115
109
89
106
119
97

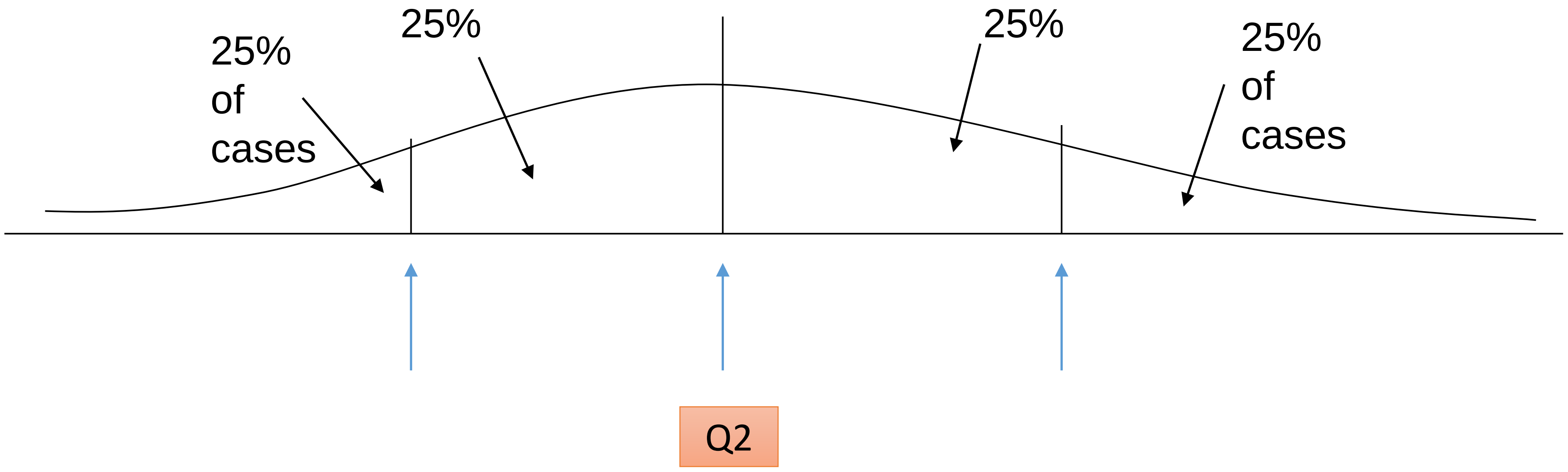
Range

- Tallest vs. shortest

Range=251-57=194cm

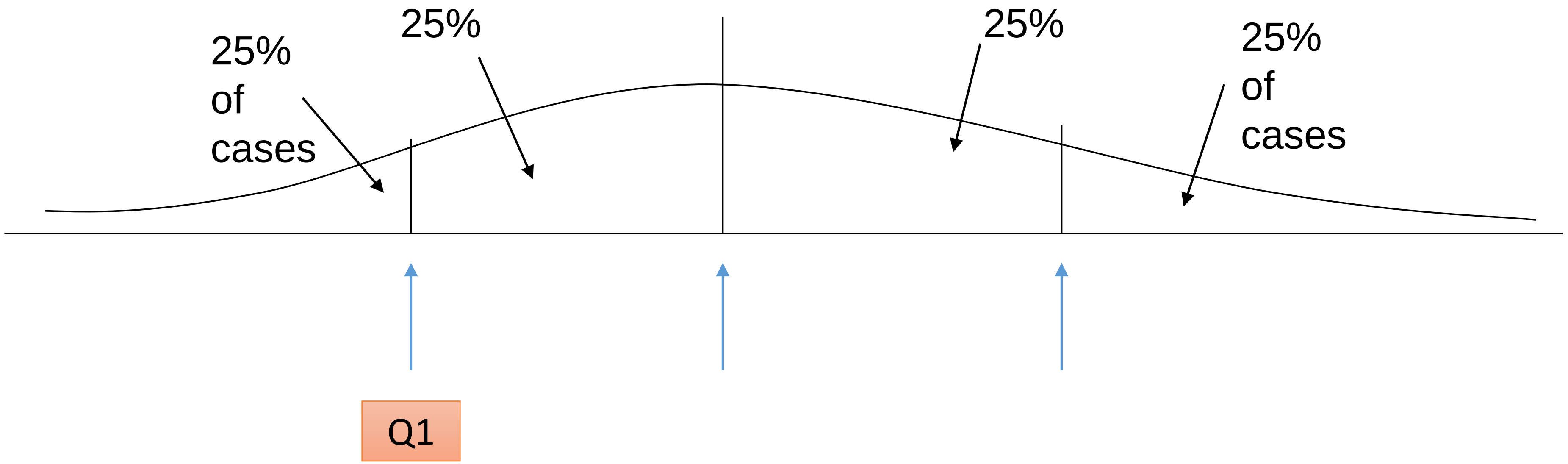
Interquartile range

- A quartile is the value that marks one of the divisions that breaks a ***rank-ordered*** data set into four equal parts
  – (Q2) the median is a quartile and divides the cases in half



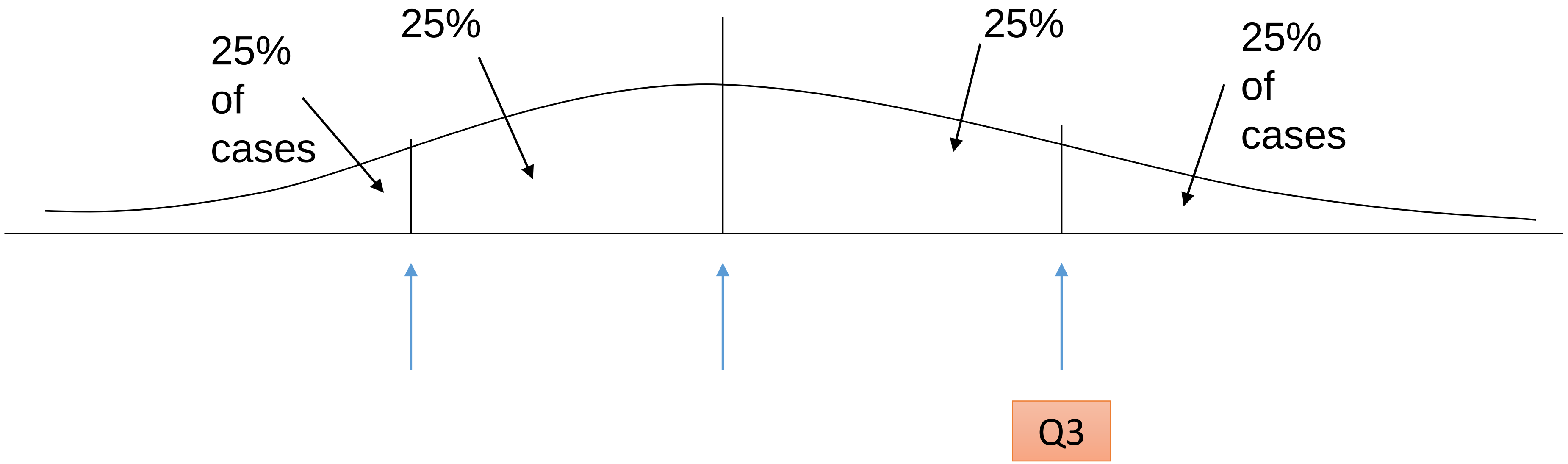
Interquartile range

- A quartile is the value that marks one of the divisions that breaks a series of values into four equal parts
  – (Q1) 25th percentile is a quartile that divides the first ¼ of cases from the latter ¾



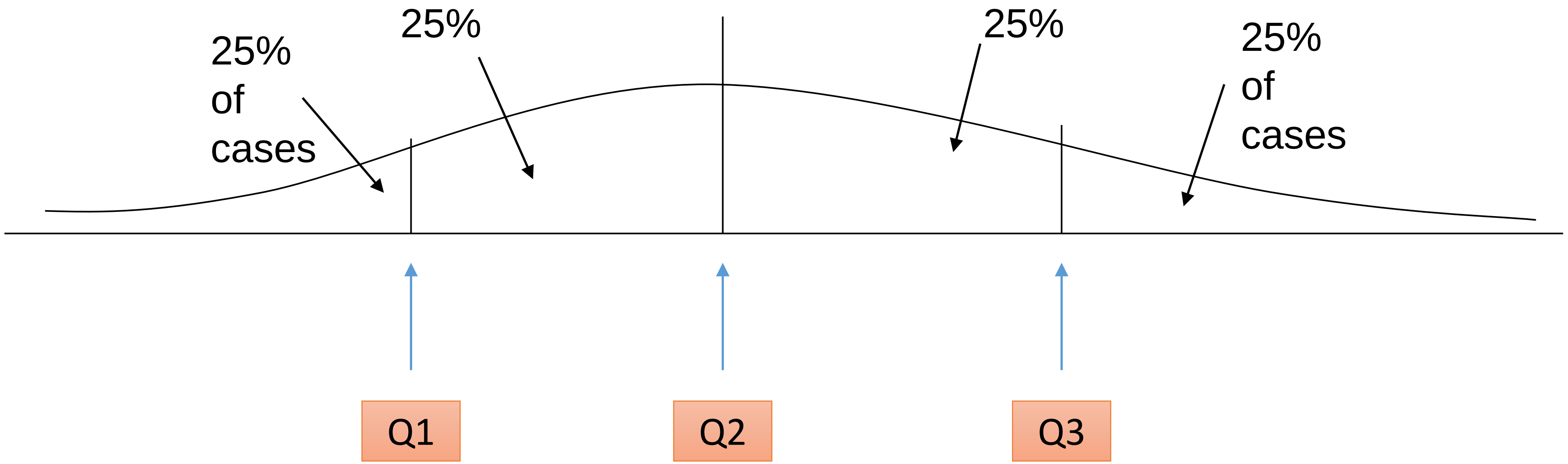
Interquartile range

- A quartile is the value that marks one of the divisions that breaks a series of values into four equal parts
  – (Q3) 75th percentile is a quartile that divides the first ¾ of cases from the latter ¼



Interquartile range

- A quartile is the value that marks one of the divisions that breaks a series of values into four equal parts
  – The interquartile range is the distance or range between the 25th percentile and the 75th percentile.



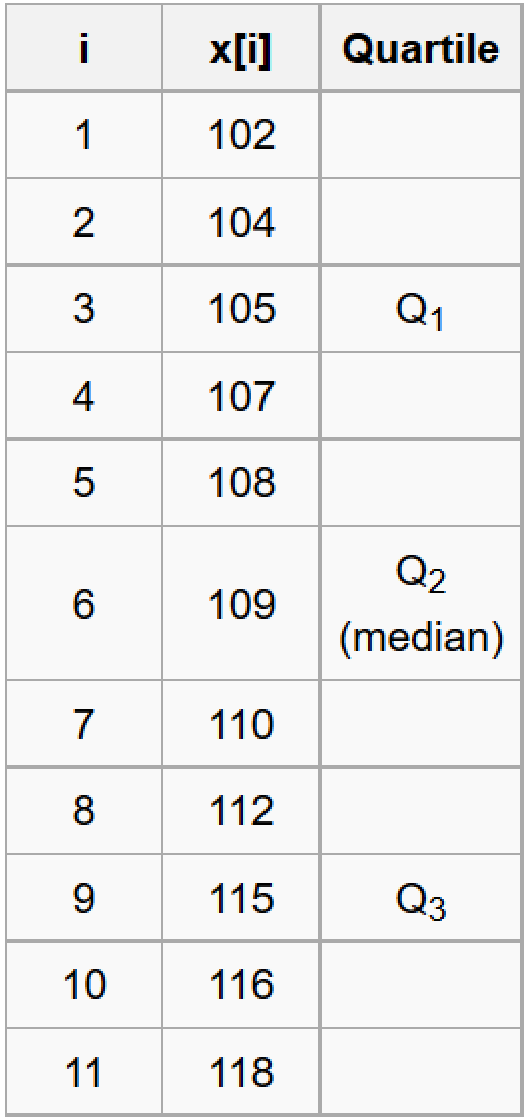
Interquartile range

- Below, what is the interquartile range (IQR)?
- Data
  109 115 118 116 102 107
  112 104 105 110 108

Interquartile range

- Below, what is the IQR?
- Data
  109 115 118 116 102 107
  112 104 105 110 108

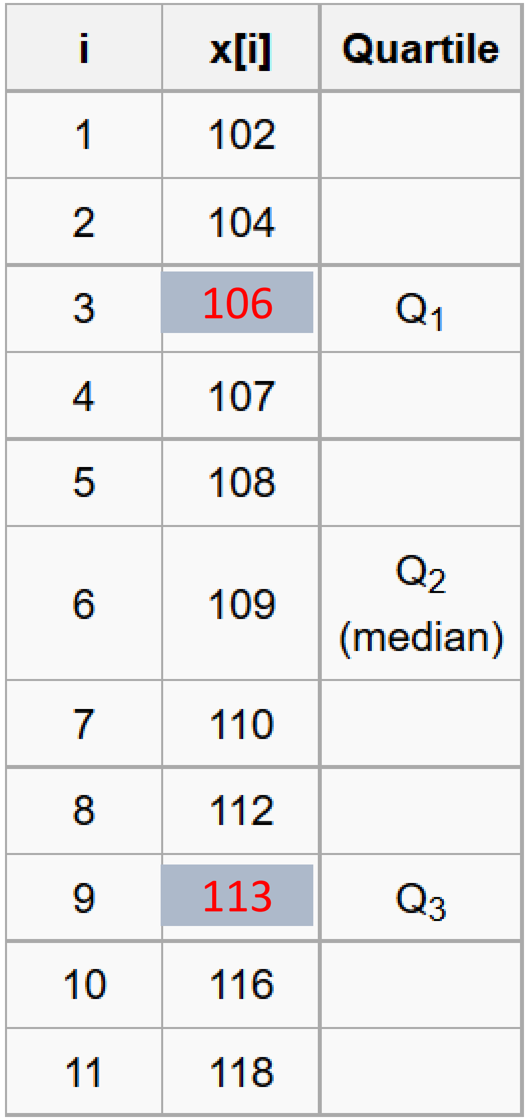
- IQR=115-105=10
- (note we just define the markers, Q1 and Q3 in a rough way, but we can make more precise calculations to find Q1 and Q3…)



i	x[i]	Quartile
1	102	
2	104	
3	105	Q_1
4	107	
5	108	
6	109	Q_2 (median)
7	110	
8	112	
9	115	Q_3
10	116	
11	118	

Interquartile range

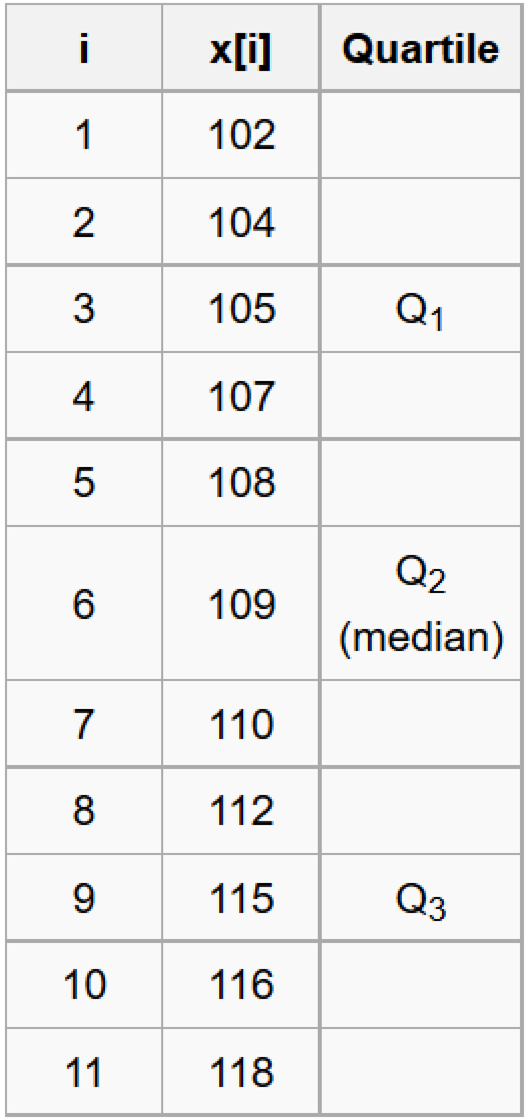
- Below, what is the IQR?
- Data
  109 113 118 116 102 107
  112 104 106 110 108
- IQR=113-106=7 (less spread)



i	x[i]	Quartile
1	102	
2	104	
3	106	Q_1
4	107	
5	108	
6	109	Q_2 (median)
7	110	
8	112	
9	113	Q_3
10	116	
11	118	

Interquartile range

- The IQR is often used to find outliers in data
  – observations that fall below Q1-1.5(IQR) or above Q3+1.5(IQR)
  – any outliers in our data?



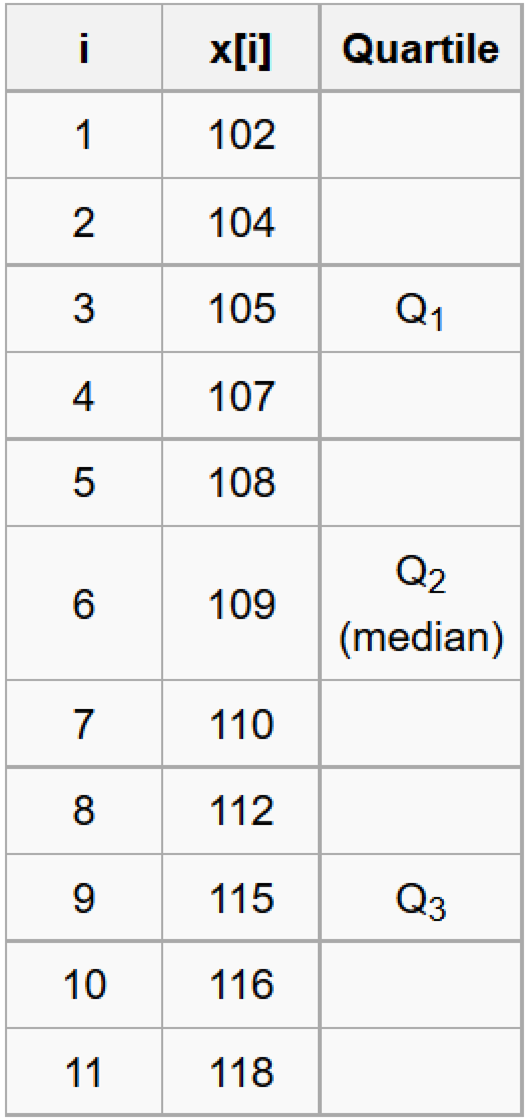
i	x[i]	Quartile
1	102	
2	104	
3	105	Q_1
4	107	
5	108	
6	109	Q_2 (median)
7	110	
8	112	
9	115	Q_3
10	116	
11	118	

Interquartile range

- The IQR is often used to find outliers in data
  – observations that fall below Q1-1.5(IQR) or above Q3+1.5(IQR)

- Q1-1.5(IQR) = 105-1.5*10=90
  Q3+1.5(IQR) = 115+1.5*10=130

- Outliers: beyond [90, 130]



i	x[i]	Quartile
1	102	
2	104	
3	105	Q_1
4	107	
5	108	
6	109	Q_2 (median)
7	110	
8	112	
9	115	Q_3
10	116	
11	118	

Interquartile range

- Exercise: height (cm) of 7 people
- Question 1: what is the IQR?
- Question 2: are there any outliers?

145, 183, 188, 185, 187, 180, 160

Interquartile range

- Exercise: height (cm) of 7 people
- Question 1: what is the IQR?
- Question 2: are there any outliers?

145, 183, 188, 185, 187, 180, 160

Rank in order

145, 160, 180, 183, 185, 187, 188

50%

50%

Q2 (second quartile)

Interquartile range

- Exercise: height (cm) of 7 people
- Question 1: what is the IQR?
- Question 2: are there any outliers?

145, 160, 180, 183, 185, 187, 188

Let’s define Q1 and Q3 in a more precise way!
Q1 (25th percentile): 7(i.e., total number of values) * 0.25 = 1.75 ≈ 2
Q3 (75th percentile): 7(i.e., total number of values) * 0.75 = 5.25 ≈ 5

Interquartile range

- Exercise: height (cm) of 7 people
- Question 1: what is the IQR?
- Question 2: are there any outliers?

145, 160, 180, 183, 185, 187, 188

Q1

Q3

Q1 (25th percentile): 7(i.e., total number of values) * 0.25 = 1.75 ≈ 2
Q3 (75th percentile): 7(i.e., total number of values) * 0.75 = 5.25 ≈ 5
Now,
Q1=(160+180)/2=170   Q3=(185+187)/2=186

Interquartile range

- Exercise: height (cm) of 7 people
- Question 1: what is the IQR?
- Question 2: are there any outliers?

145, 160, 180, 183, 185, 187, 188

Q1

Q3

Q1=(160+180)/2=170 Q3=(185+187)/2=186

IQR = Q3-Q1=186-170=16

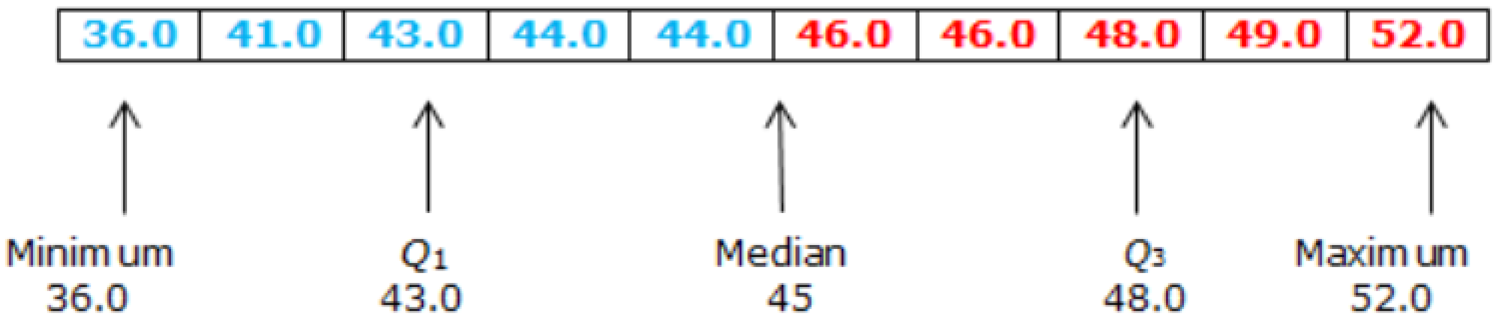
Q1-1.5(IQR)=170-1.5*16=170-24=146

Q3+1.5(IQR)=186+1.5*16=186+24=210

Anyone outside the range [146, 210]? So, we can detect the outlier…

Interquartile range

- Map the values onto a “box”



36.0	41.0	43.0	44.0	44.0	46.0	46.0	48.0	49.0	52.0

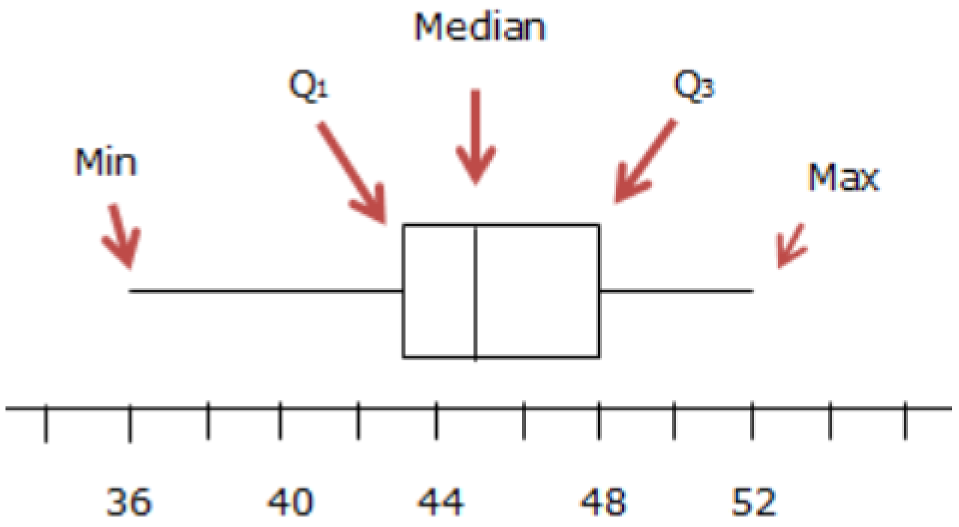
Minim um
36.0

Q_1
43.0

Median
45

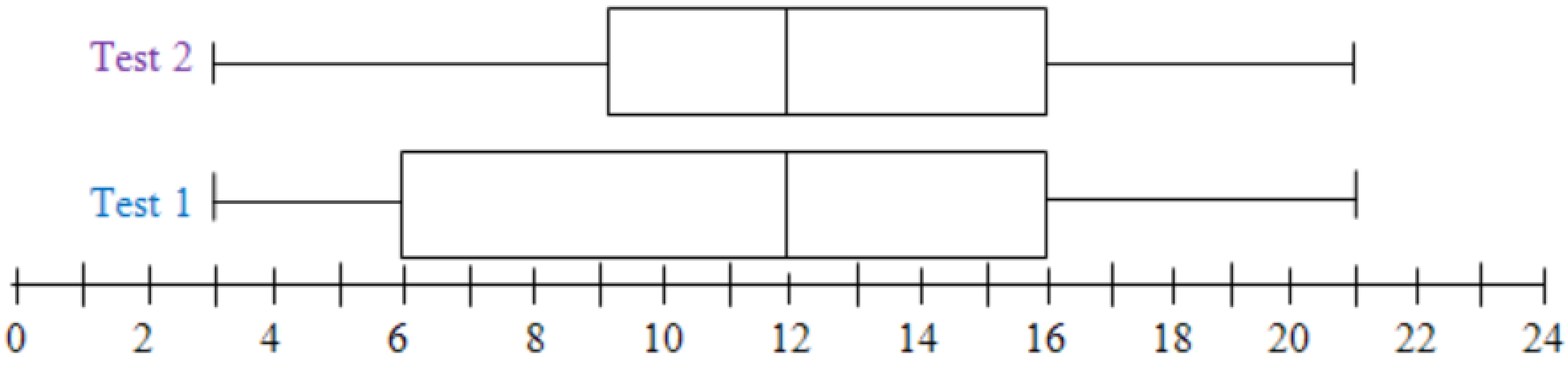
Q_3
48.0

Maxim um
52.0



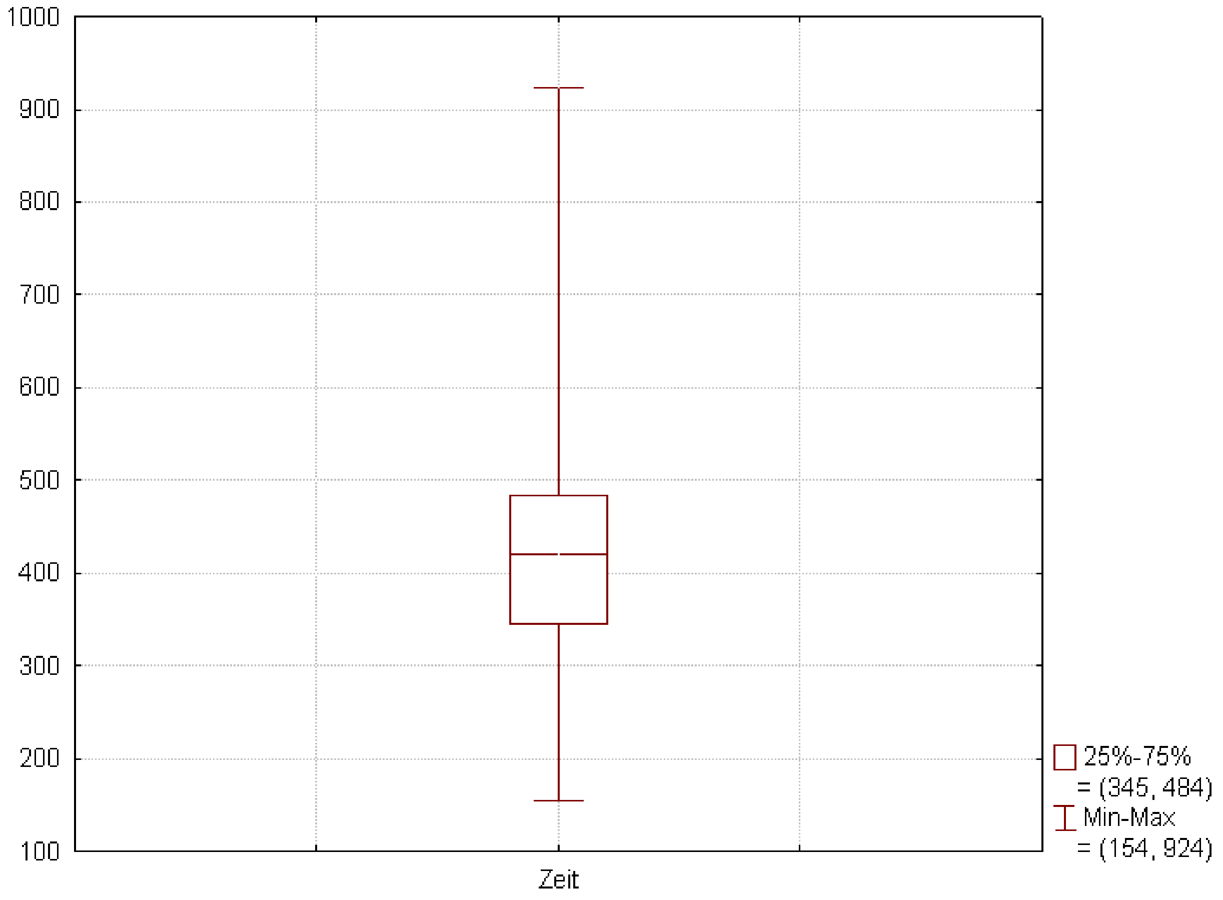
Interquartile range

- Compare two data sets
  – Suppose that plots below represent quiz scores out of 25 points for Quiz 1 and Quiz 2 for the same class
  – What do these plots show about how the class did on test #2 compared to test #1?



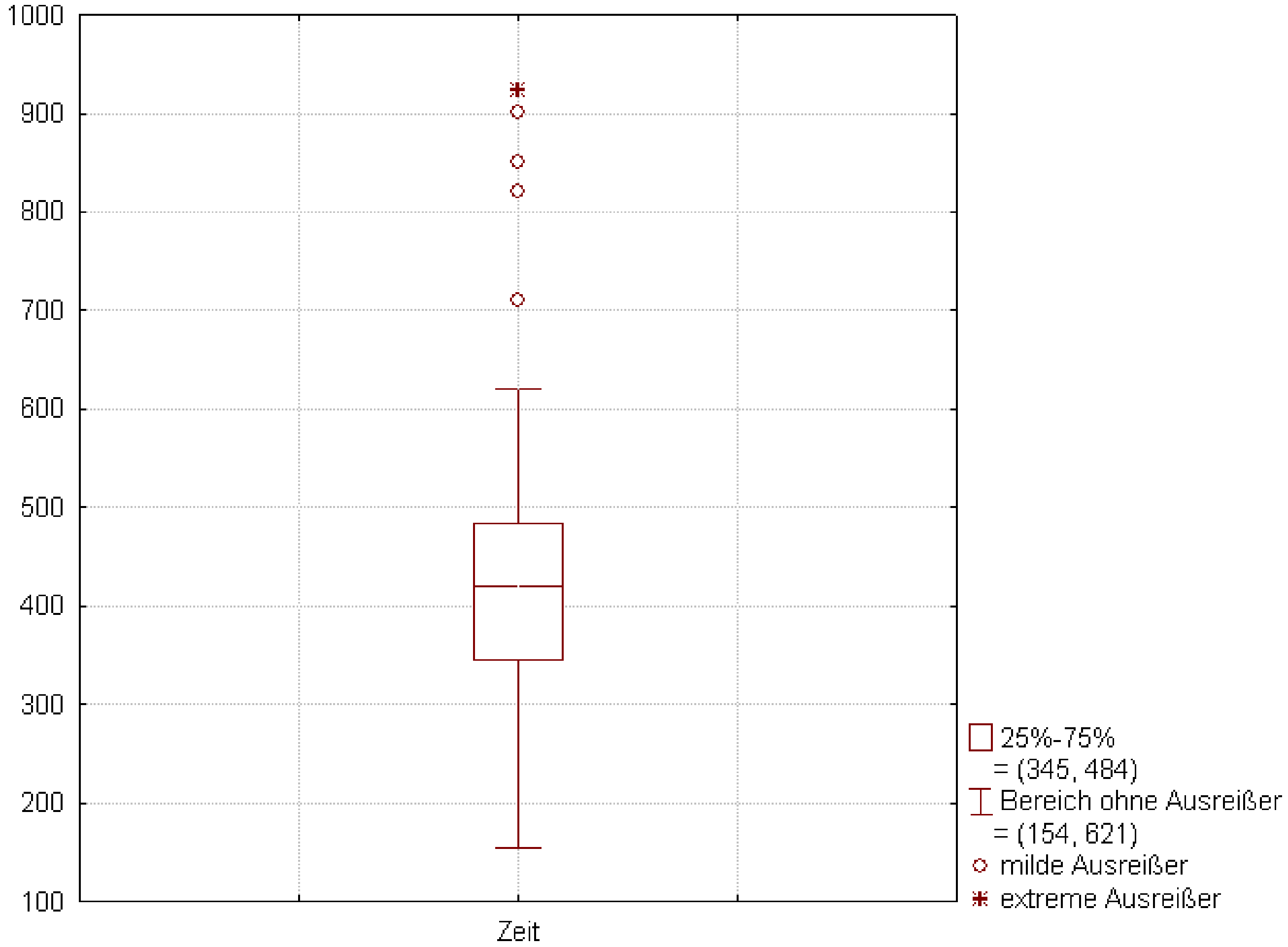
Box plot (or box-and-whisker plot)

- Figure. Boxplot with whiskers from minimum to maximum



Box plot (or box-and-whisker plot)

- Figure. Same Boxplot with whiskers with maximum 1.5 IQR



Box plot (or box-and-whisker plot)

- Figure. with whiskers with maximum 1.5 IQR

145, 160, 180, 183, 185, 187, 188

Q1 Q2 Q3

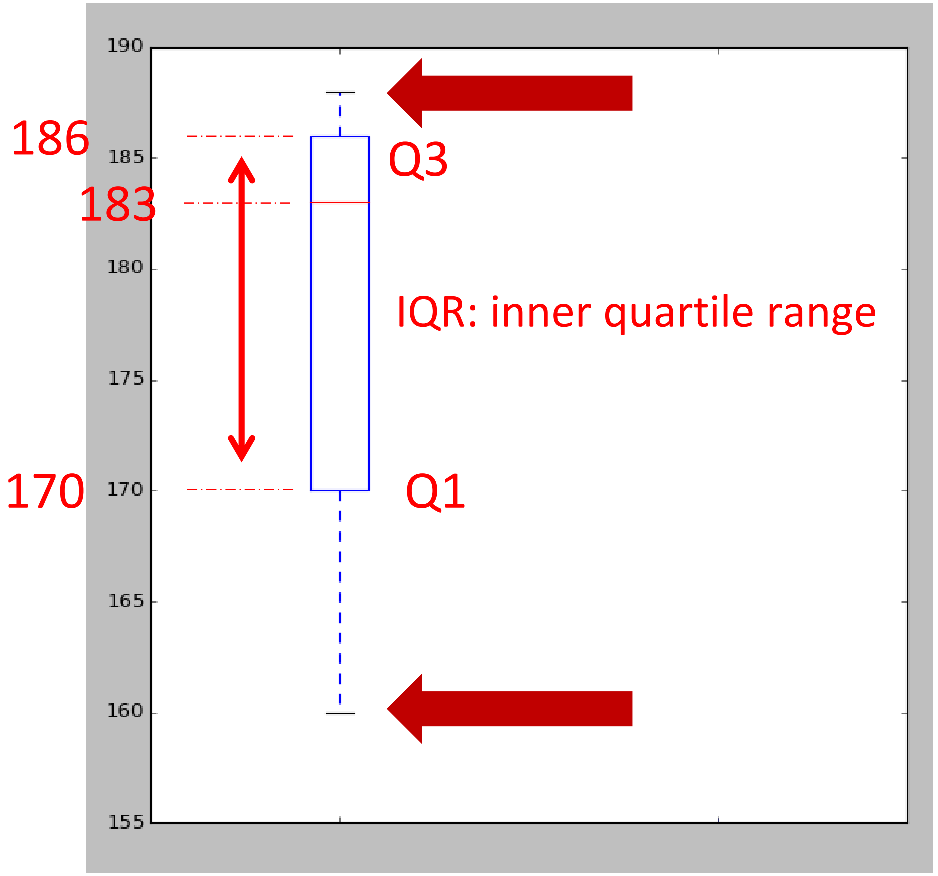
Q1=(160+180)/2=170 Q3=(185+187)/2=186

IQR = Q3-Q1=186-170=16

Q1-1.5(IQR)=170-1.5*16=170-24=146

Q3+1.5(IQR)=186+1.5*16=186+24=210

Note: it plots the extremes as 160 and 188, why?



186

183

170

Q3

Q1

IQR: inner quartile range

Variance

- A measure of the spread of the values on a variable – (a number that at first seems complex to calculate…)
- Calculating variance starts with a “deviation” – a deviation is the distance away from the mean of a case’ score

If the average person’s car costs £20,000, my deviation from the mean is - £14,000!

6K - 20K = -14K

Variance

- Example: Class A – IQ scores
- The deviation of 102 from 110.54?

- Deviation of 115?

Class A--IQs of 13 Students

102

128

131

98

140

93

110

115

109

89

106

119

97

Mean = *110.54*

Variance

- Example: Class A – IQ scores
- The deviation of 102 from 110.54?
  102 - 110.54 = **-8.54**
- Deviation of 115?
  115 - 110.54 = **4.46**

Class A--IQs of 13 Students

102
128
131
98
140
93
110
115
109
89
106
119
97

Mean = *110.54*

Variance

- Example: Class A – IQ scores
- The deviation of 102 from 110.54?
  102 - 110.54 = **-8.54**
- Deviation of 115?
  115 - 110.54 = **4.46**
- We want to add these to get total deviations, but if we were to do that, we would get zero every time. Why?
- We need a way to eliminate neg. signs

Class A--IQs of 13 Students

102
128
131
98
140
93
110
115
109
89
106
119
97

Mean = *110.54*

Variance

- Squaring the deviations will eliminate negative signs… – a deviation squared
- Back to the IQ examples:
  $(102 - 110.54)^2 = (-8.54)^2 = 72.93$
  $(115 - 110.54)^2 = (4.46)^2 = 19.89$
- If you were to add all the squared deviations together, you'd get what we call the "Sum of Squares (SS)"

Variance

Class A, sum of squares:

$(102 – 110.54)^2 + (115 – 110.54)^2 +$
$(128 – 110.54)^2 + (109 – 110.54)^2 +$
$(131 – 110.54)^2 + (89 – 110.54)^2 +$
$(98 – 110.54)^2 + (106 – 110.54)^2 +$
$(140 – 110.54)^2 + (119 – 110.54)^2 +$
$(93 – 110.54)^2 + (97 – 110.54)^2 +$
$(110 – 110.54)^2 =$ SS = 2825.39

Class A--IQs of 13 Students

102	115
128	109
131	89
98	106
140	119
93	97
110	

Mean = *110.54*

Variance

- The last step…
- The approximate average sum of squares (SS) is the variance
  SS/(n) = variance for a population
  SS/(n-1) = variance for a sample (a random sample drawn from some large parent population)
  n: total number of values
- For Class A, Variance = 2825.39/(n-1)=2825.39/12=235.45
  How helpful is that???

Standard deviation

- To convert variance into something of meaning, let's create standard deviation
- The square root of the variance reveals the average deviation of the observations from the mean

standard deviation =sqrt(variance)

Standard deviation

- Class A, the deviation is sqrt(235.45) = **15.34**
- The average of persons' deviation from the mean IQ (of 110.54) is 15.34 IQ points
- Review:
  1. mean, and deviation
  2. deviation squared
  3. sum of squares
  4. variance
  5. standard deviation

Class A--IQs of 13 Students

102
128
131
98
140
93
110

115
109
89
106
119
97

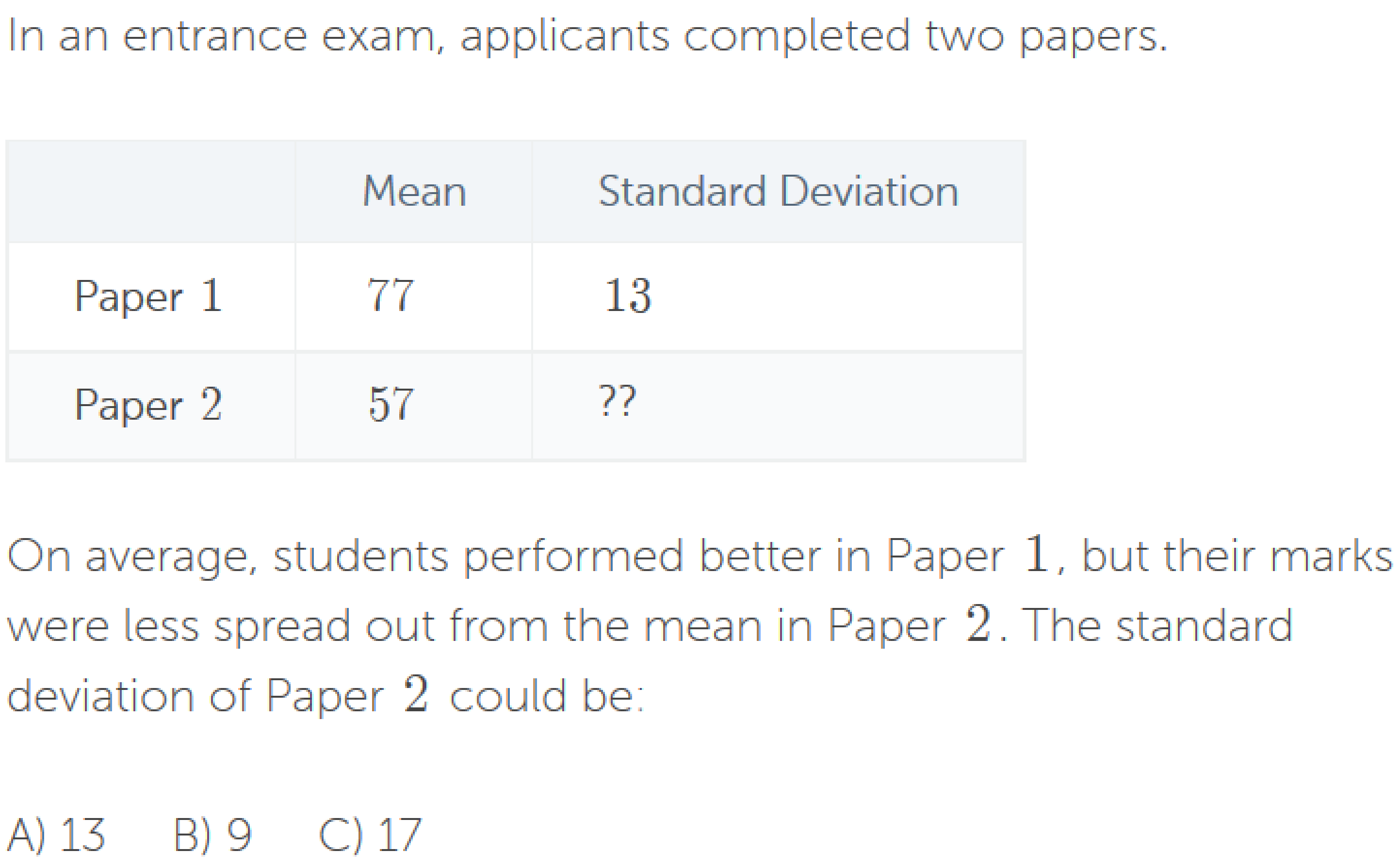
Mean = *110.54*

Understanding the standard deviation

- A large standard deviation indicates that the data points can spread far from the mean
- A small standard deviation indicates that they are clustered closely around the mean
- Example:
  data1=[0, 0, 14, 14]
  data2=[0, 6, 8, 14]
  data3=[6, 6, 8, 8]
  same mean=7
  different standard deviation=7, 5, 1

Understanding the standard deviation

- Exercise



In an entrance exam, applicants completed two papers.

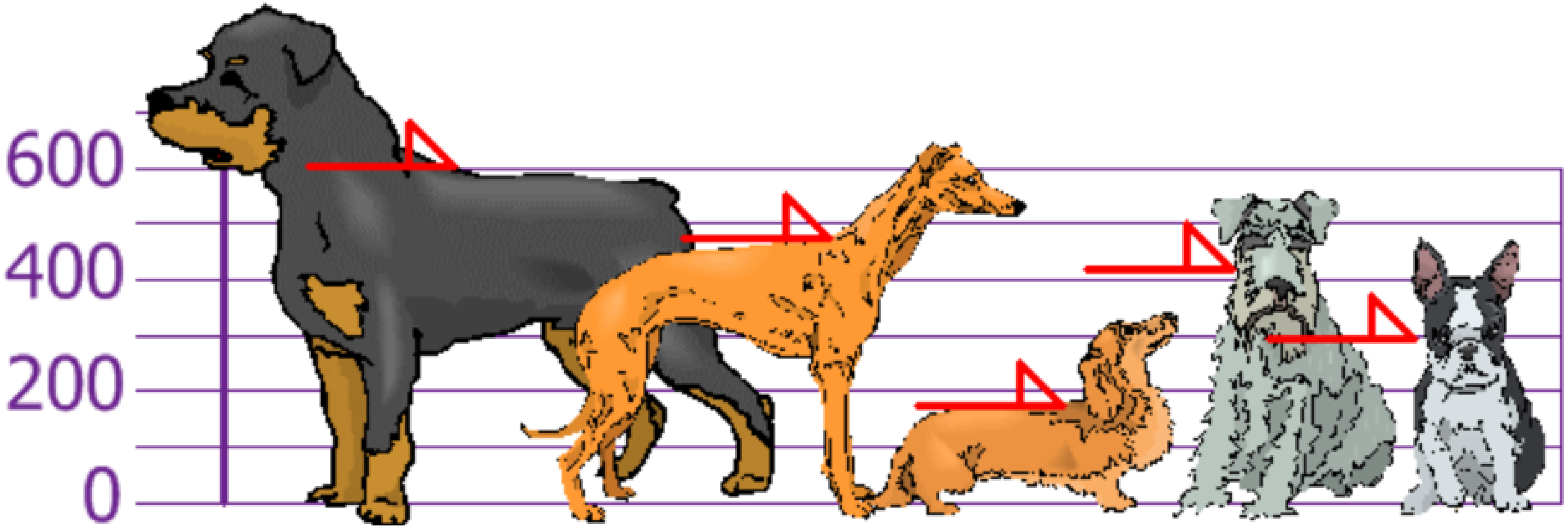
	Mean	Standard Deviation
Paper 1	77	13
Paper 2	57	??

On average, students performed better in Paper 1, but their marks were less spread out from the mean in Paper 2. The standard deviation of Paper 2 could be:

A) 13 B) 9 C) 17

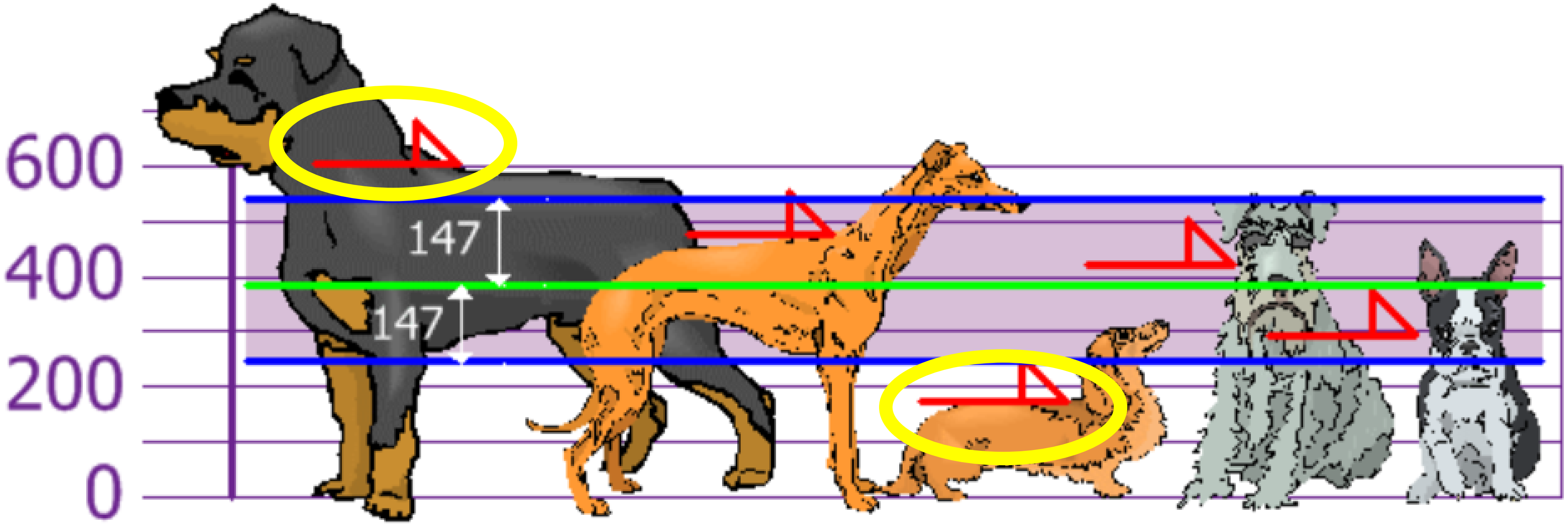
Understanding the standard deviation

- A measure of how spread out numbers are (how far from the "normal"…)
- Example: heights of dogs (in millimetres)



Understanding the standard deviation

- Example: heights of dogs (in millimetres)
  standard deviation = 147mm
- We have a "standard" way of knowing what is normal, and what is extra large or extra small



Uses for standard deviation

- Example:
  A class of students took a math test. Their teacher found that the mean score on the test was an 85%. She then calculated the standard deviation of the other test scores and found a very small standard deviation which suggested that most students scored very close to 85%.

Uses for standard deviation

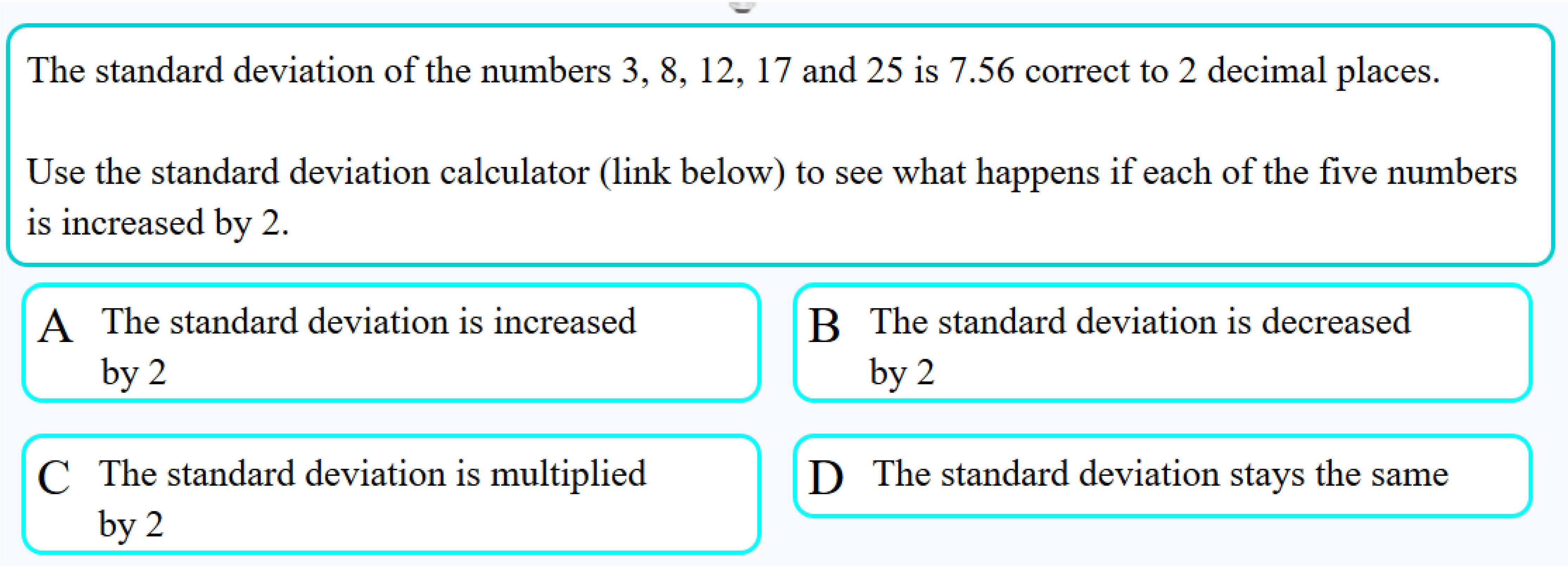
- Example:
  A class of students took a test in Language Arts. The teacher determines that the mean grade on the exam is a 65%. She is concerned that this is very low, so she determines the standard deviation to see if it seems that most students scored close to the mean, or not. The teacher finds that the standard deviation is high. After closely examining all of the tests, the teacher is able to determine that several students with very low scores were the outliers that pulled down the mean of the entire class's scores.

Uses for standard deviation

- Example:
  An employer wants to determine if the salaries in one department seem fair for all employees, or if there is a great disparity. He finds the average of the salaries in that department and then calculates the variance, and then the standard deviation. The employer finds that the standard deviation is slightly higher than he expected, so he examines the data further and finds that while most employees fall within a similar pay bracket, three loyal employees who have been in the department for 20 years or more, far longer than the others, are making far more due to their longevity with the company. Doing the analysis helped the employer to understand the range of salaries of the people in the department.

Data analysis

- Question 1:



The standard deviation of the numbers 3, 8, 12, 17 and 25 is 7.56 correct to 2 decimal places.

Use the standard deviation calculator (link below) to see what happens if each of the five numbers is increased by 2.

A The standard deviation is increased by 2

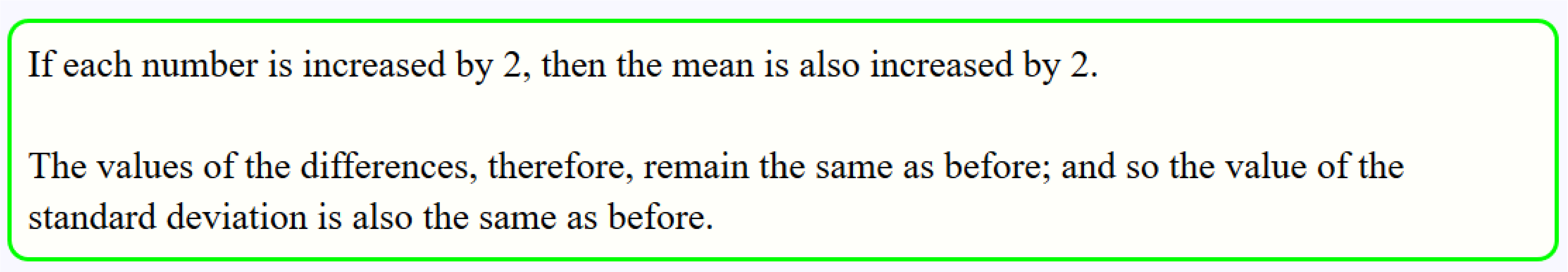
B The standard deviation is decreased by 2

C The standard deviation is multiplied by 2

D The standard deviation stays the same

Data analysis

- Question 1:

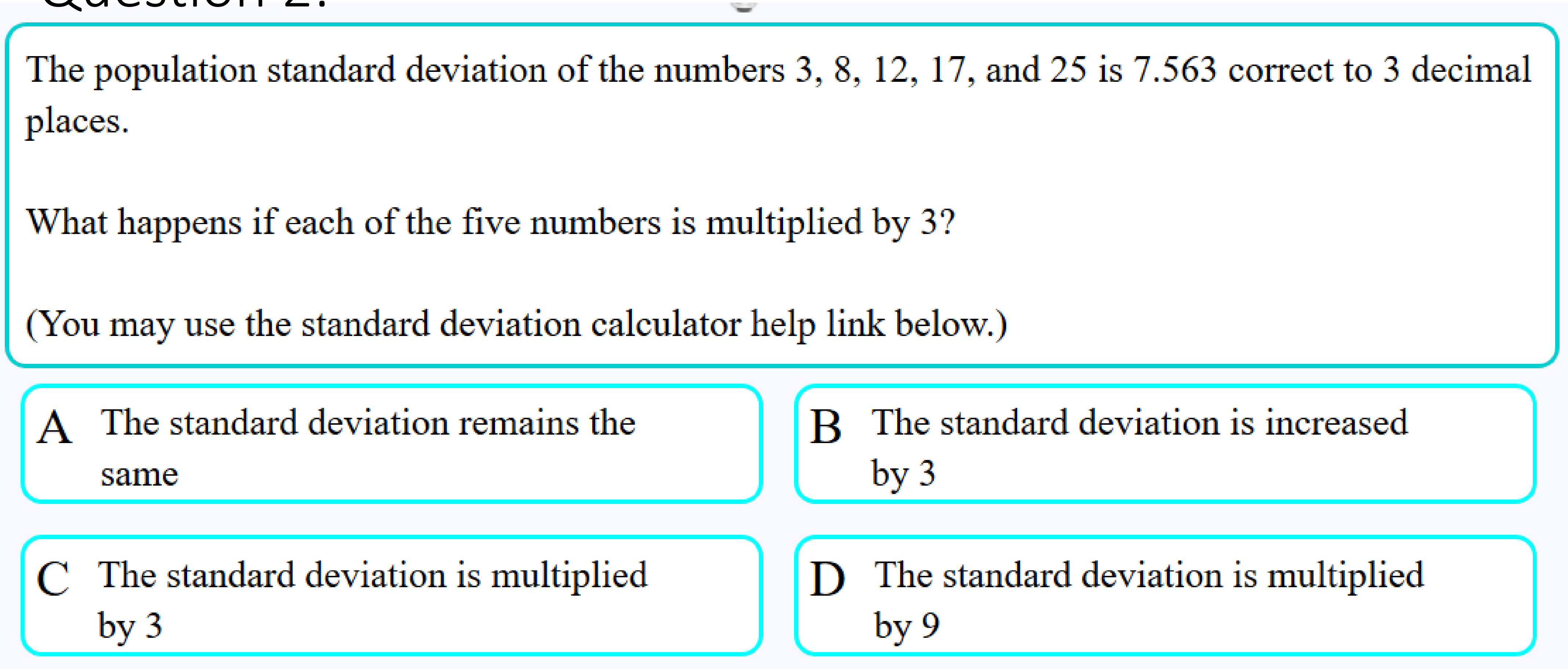


If each number is increased by 2, then the mean is also increased by 2.

The values of the differences, therefore, remain the same as before; and so the value of the standard deviation is also the same as before.

Data analysis

- Question 2:



The population standard deviation of the numbers 3, 8, 12, 17, and 25 is 7.563 correct to 3 decimal places.

What happens if each of the five numbers is multiplied by 3?

(You may use the standard deviation calculator help link below.)

A The standard deviation remains the same

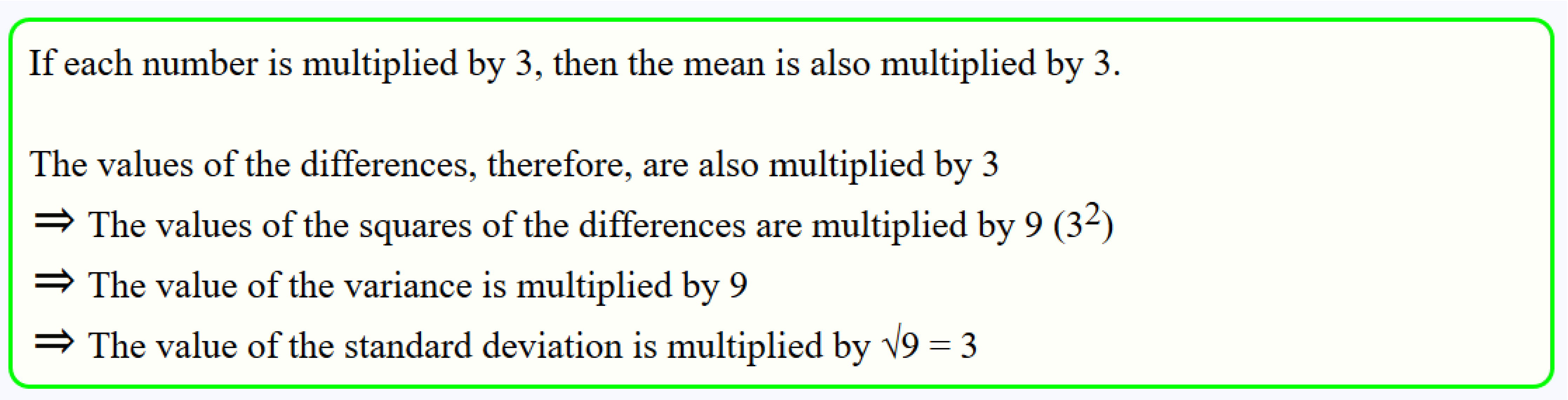
B The standard deviation is increased by 3

C The standard deviation is multiplied by 3

D The standard deviation is multiplied by 9

Data analysis

- Question 2:



If each number is multiplied by 3, then the mean is also multiplied by 3.

The values of the differences, therefore, are also multiplied by 3

⇒ The values of the squares of the differences are multiplied by 9 (3^2)

⇒ The value of the variance is multiplied by 9

⇒ The value of the standard deviation is multiplied by $\sqrt{9} = 3$

Data analysis

- Example (USA): most American employers issue raises based on percent of salary…
  If your budget went up by 5%, salaries can go up by 5%

- What happens with a percentage raise?

Data analysis

Acme Toilet Cleaning Services

Salary Pool: $200,000

Incomes:

President: $100K; Manager: 50K; Secretary: 40K; and Toilet Cleaner: 10K

Mean: $50K

Range: $90K

Variance: $1,050,000,000

Standard Deviation: $32.4K

These can be considered "measures of inequality"

Now, let's apply a 5% raise.

Data analysis

After a 5% raise, the pool of money increases by $10K to $210,000

Incomes:

President: $105K; Manager: 52.5K; Secretary: 42K; and Toilet Cleaner: 10.5K

Mean: $52.5K – went up by 5%

Range: $94.5K – went up by 5%

Variance: $1,157,625,000

Standard Deviation: $34K –went up by 5%

Measures of Inequality

The flat percentage raise increased inequality. The top earner got 50% of the new money. The bottom earner got 5% of the new money. Measures of inequality went up by 5%.

Data analysis

- Example (USA): most American employers issue raises based on percent of salary…

- The gap between the rich and poor expands
  This is why some progressive organisations give a percentage raise with a flat increase for lowest wage earners. For example, 5% or $1,000, whichever is greater.