Quantitative Methods – I

A.Y. 2021-22

Practice 3

Lorenzo Cavallo

For any clarification/meeting: cavallo@istat.it

THEME #1

Quartiles of a data set

Quartiles

In statistics, a quartile divides the number of data points into four parts, or quarters. The data must be ordered from smallest to largest to compute quartiles.

The first quartile (Q1) is defined as the middle number between the smallest number (minimum) and the median of the data set.

The second quartile (Q2) is the median of a data set.

The third quartile (Q3) is the middle value between the median and the highest value (maximum) of the data set.

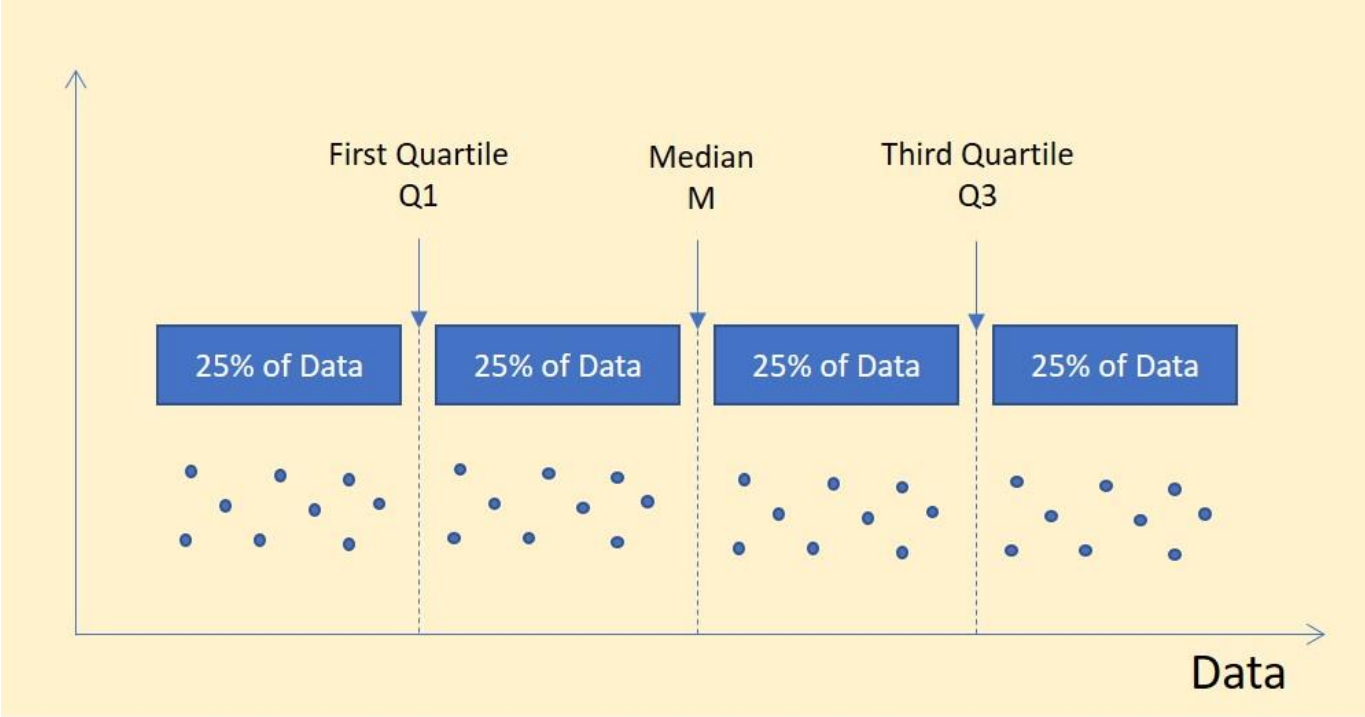
Symbol	Names	Definition
Q_1	**first quartile** **lower quartile** **25th percentile**	splits off the lowest 25% of data from the highest 75%
Q_2	**second quartile** **median** **50th percentile**	cuts data set in half
Q_3	**third quartile** **upper quartile** **75th percentile**	splits off the highest 25% of data from the lowest 75%

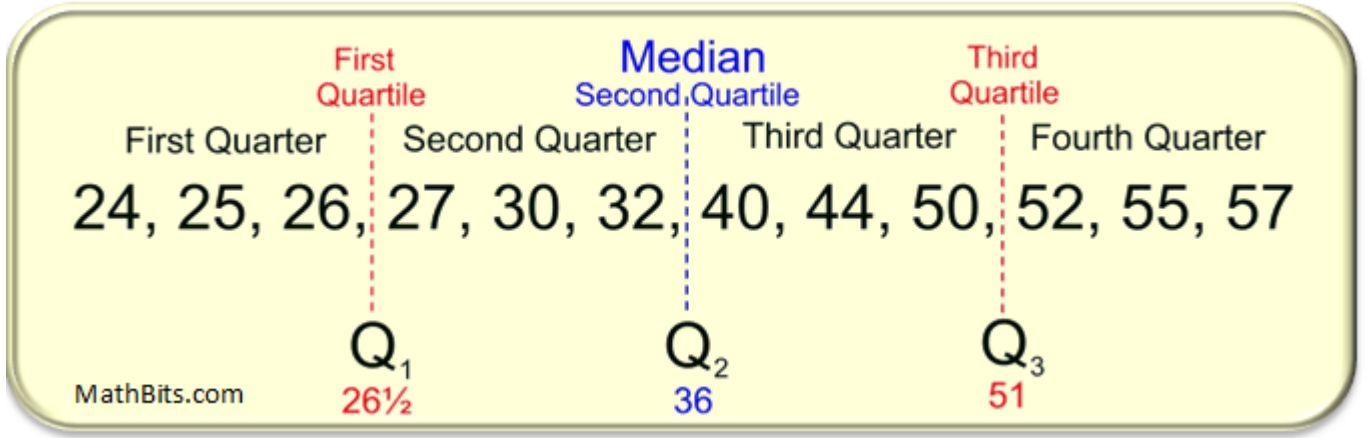
Computing methods

Use the median to divide the ordered data set into two-halves.

- If there is an odd number of data points in the original ordered data set, **do not include the median** in either half.
- If there is an even number of data points in the original ordered data set, split this data set exactly in half.

The lower quartile value is the median of the lower half of the data. The upper quartile value is the median of the upper half of the data.


First Quartile
Q1
Median
M
Third Quartile
Q3
25% of Data
25% of Data
25% of Data
25% of Data
Data


First
Quartile
Median
Second Quartile
Third
Quartile
First Quarter
Second Quarter
Third Quarter
Fourth Quarter
24, 25, 26, 27, 30, 32, 40, 44, 50, 52, 55, 57
Q1
26½
Q2
36
Q3
51
MathBits.com

Finding the Quartiles of a Data Set

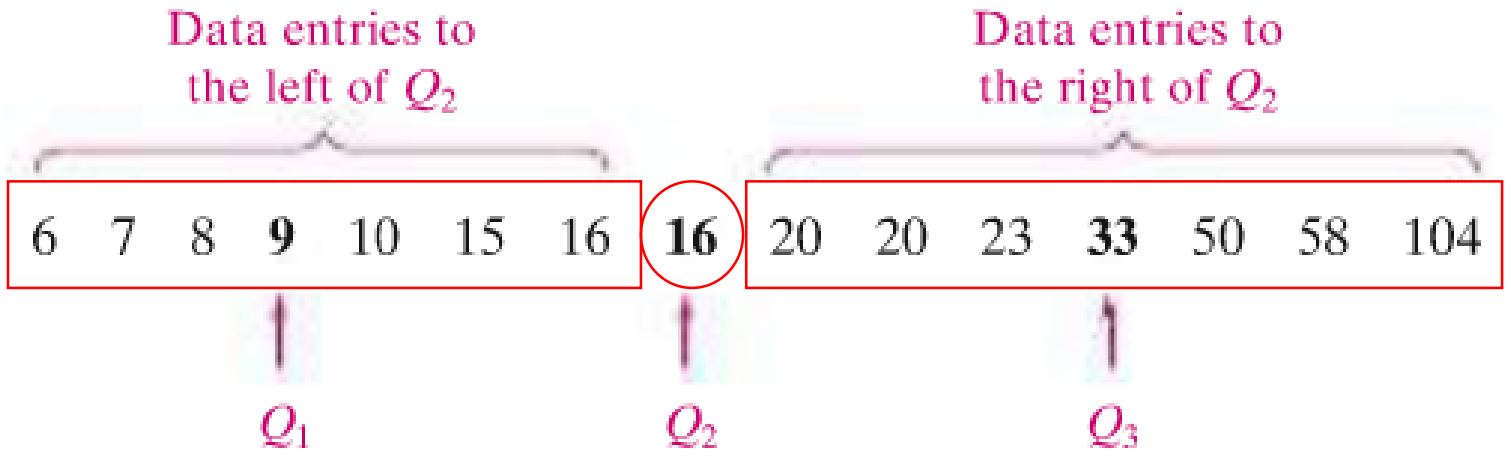
The number of nuclear power plants in the top 15 nuclear power-producing countries in the world are listed. Find the first, second, and third quartiles of the data set. What do you observe? *(Source: International Atomic Energy Agency)*

7 20 16 6 58 9 20 50 23 33 8 10 15 16 104

Solution

First, order the data set and find the median Q_2. The first quartile Q_1 is the median of the data entries to the left of Q_2. The third quartile Q_3 is the median of the data entries to the right of Q_2.

$$n = 15 \rightarrow Median = \frac{15+1}{2} = 8th\ place$$



There is an odd number of data points in the original ordered data set, **do not include the median** in either half.

$$n_1 = n_2 = 7$$

$$Median = \frac{7+1}{2} = 4th\ place$$

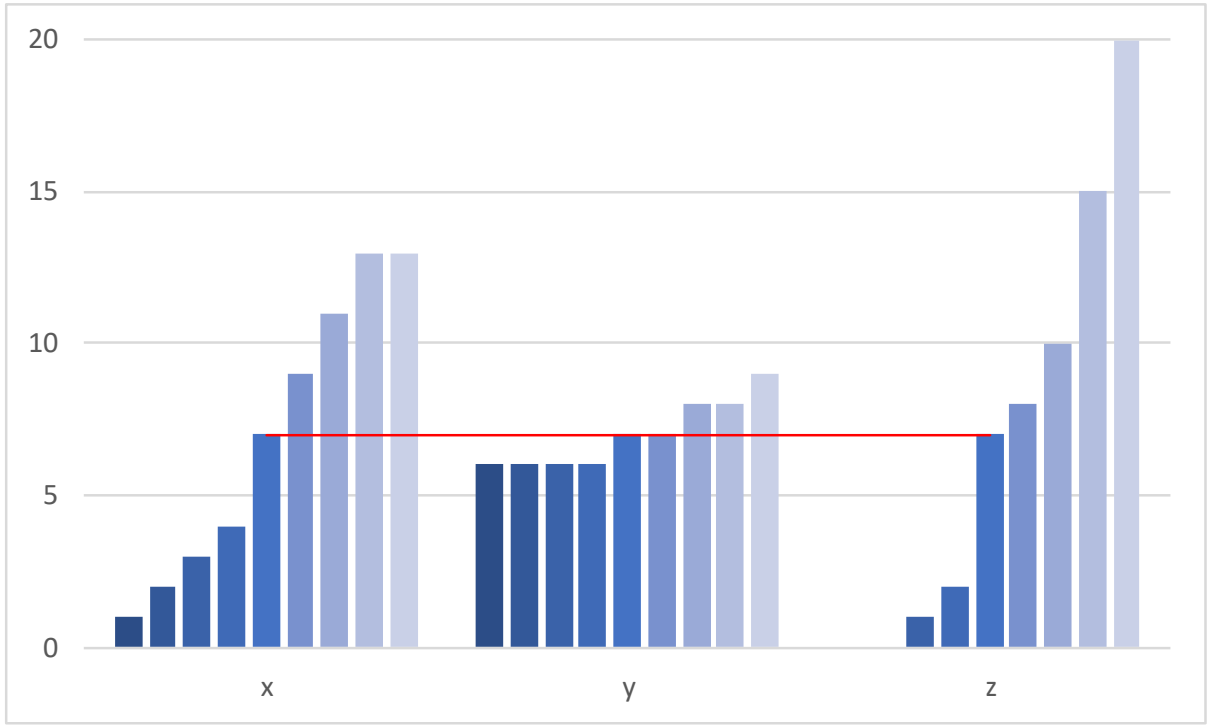
THEME #2

Measures of dispersion

x	y	z
1	6	0
2	6	0
3	6	1
4	6	2
7	7	7
9	7	8
11	8	10
13	8	15
13	9	20
63	63	63

Mean= 7

Median= 7



Same mean, same Median…
but different distribution

The measures of position
are not enough

Measures of dispersion

Range
It is obtained by taking the difference between the largest and the smallest values in a data set.

Range=Largest value−Smallest value

Interquartile Range
The difference between the third and the first quartiles

IQR = Q3 − Q1

Variance and Standard Deviation
The variance is the squared deviation of a variable from its mean.
The standard deviation is obtained by taking the positive square root of the variance.

$$\sigma^2 = \frac{\sum (x-\mu)^2}{N} \quad \text{and} \quad s^2 = \frac{\sum (x-\bar{x})^2}{n-1}$$

$$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}} \quad \text{and} \quad s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$$

Coefficient of Variation
The coefficient of variation, denoted by CV, expresses standard deviation as a percentage of the mean.

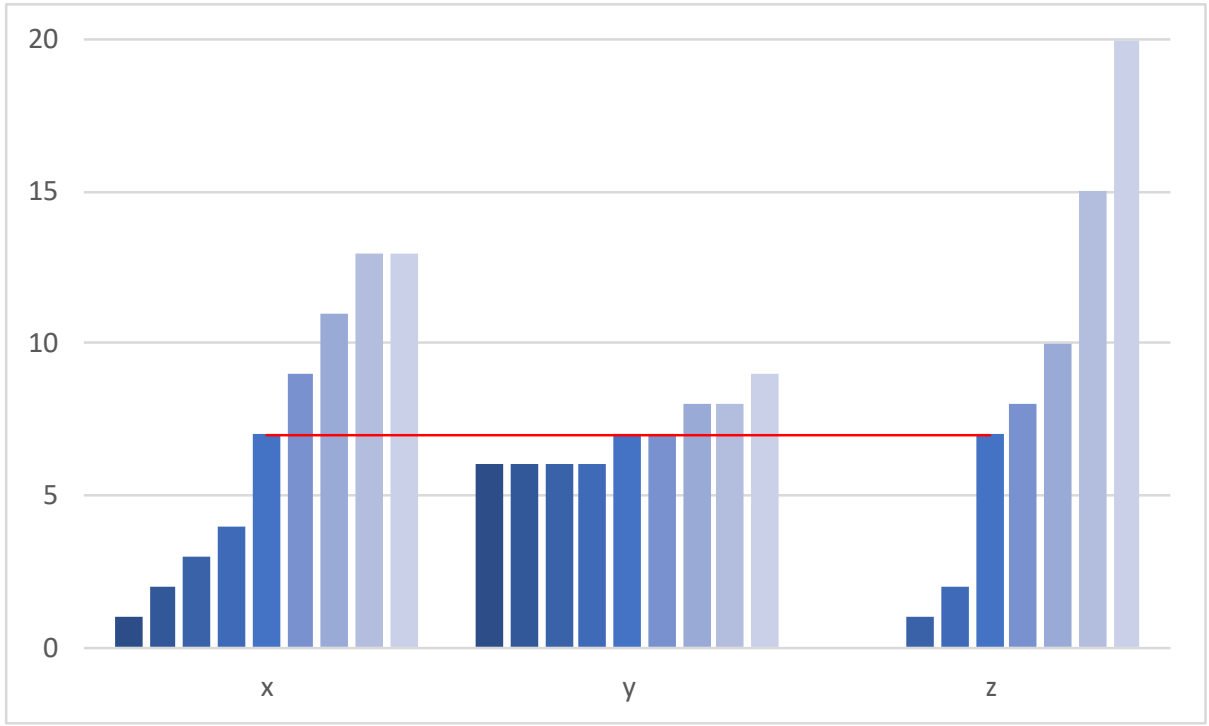
For population data : $CV = \frac{\sigma}{\mu} \times 100\%$

For sample data : $CV = \frac{s}{\bar{x}} \times 100\%$

x	y	z
1	6	0
2	6	0
3	6	1
4	6	2
7	7	7
9	7	8
11	8	10
13	8	15
13	9	20
63	63	63

Mean= 7

Median= 7



Range(x)= 13-1= 12

Range(y)= 9-6= 3

Range(z)= 20-0= 20

Variance(x)= 19,78 St.dev(x) 4,45

Variance(y)= 1,11 St.dev(y) 1,05

Variance(z)= 44,67 St.dev(z) 6,68

Ex. 9

Calculate range, median and interquartile range of following values:

14 3 5 4 6 8 12 5 3 7 18

Rank the data

3
3
4
5
5
6
7
8
12
14
18

Median

n=11

(n+1)/2=6th position

Median= 6

For the first and the third quartile there is an odd number of data points in the original ordered data set, **do not include the median** in either half.

First Quartile

Below the median (6) we have n=5 values

(n+1)/2=3rd position

Q1= 4

Third Quartile

Above the median (6) we have n=5 values

(n+1)/2=3rd position

Q3= 12

Range

Min= 3
Max= 18

Range=Largest value−Smallest value

Range= 18 - 3 = 15

Interquartile Range

IQR = Q3 − Q1

IQR = 12 – 4 = 8

3.83 The following data represent the differences (in seconds) between each winner's time of Belmont Stakes horse racing for the years 1999–2011 and the best time of 1973.

3.80 7.20 2.80 5.71 4.26 3.50 4.75 3.81 4.74 5.65 3.54 7.57 6.88

a. Compute the range, variance, and standard deviation for these data.

Rank the data

2.80
3.50
3.54
3.80
3.81
4.26
4.74
4.75
5.65
5.71
6.88
7.20
7.57

Calculate:

Min= 2.80
Max= 7.57

Range= 7.57 – 2.80 = 4.77

Calculate the mean

$$\mu = \frac{\sum x}{N} = \frac{64.21}{13} = 4.94$$

3.83 The following data represent the differences (in seconds) between each winner's time of Belmont Stakes horse racing for the years 1999–2011 and the best time of 1973.

3.80 7.20 2.80 5.71 4.26 3.50 4.75 3.81 4.74 5.65 3.54 7.57 6.88

a. Compute the range, variance, and standard deviation for these data.

$$\mu = \frac{\sum x}{N} = \frac{64.21}{13} = 4.94$$

x	x-μ	$(x-\mu)^2$
2.80	-2.14	4.58
3.50	-1.44	2.07
3.54	-1.40	1.96
3.80	-1.14	1.30
3.81	-1.13	1.28
4.26	-0.68	0.46
4.74	-0.20	0.04
4.75	-0.19	0.04
5.65	0.71	0.51
5.71	0.77	0.59
6.88	1.94	3.77
7.20	2.26	5.11
7.57	2.63	6.92
		28.61

Calculate the variance and the standard deviation

$$\sigma^2 = \sum \frac{(x-\mu)^2}{N} = \frac{28.61}{13} = 2.20$$

$$\sigma = \sqrt{2.20} = 1.48$$

Exercise 10

Let the following unitary distribution of the character X be given:

2 4 2 2 4 2 0 4 0 2 4 1 6

Calculate the variance and the standard deviation.

x_i	n_i	$x_i n_i$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 n_i$
0	2	0	12.25	24.5
2	5	10	2.25	11.25
4	4	16	0.25	1
16	1	16	156.25	156.25
	12	**42**		**193**

Calculate the mean

$$\bar{x} = \sum \frac{x_i\, n_i}{n}$$

Calculate the variance and the standard deviation

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

$$s^2 = \frac{1}{n}\sum_{i=1}^{k} (x_i - \bar{x})^2 n_i = \frac{193}{12} = \mathbf{16.08\bar{3}}.$$

$$s = \sqrt{s^2} \simeq \sqrt{16.083} = \mathbf{4.01}.$$

THEME #2

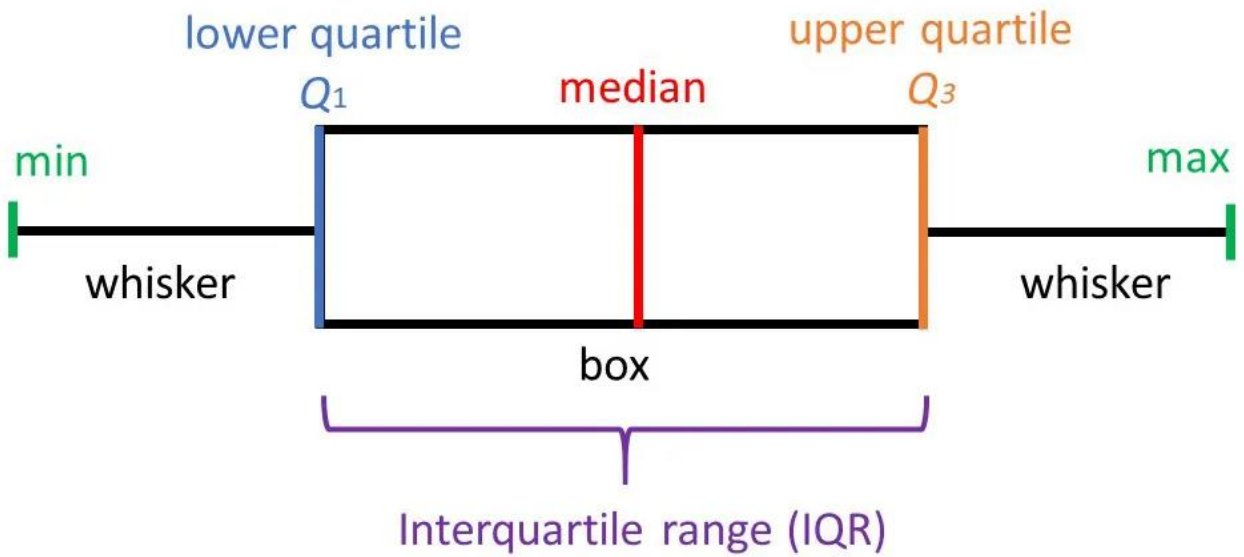
Box-Plot

Box-Whiskers Plot (or Box-Plot)

Comprehensive graphical representation of a distribution.

Indeed, it provides info on:

-position: median, Q1, and Q3 (lines)
-dispersion: interquartile range (box)
-shape of the distribution (whiskers)
-extreme values: outliers



bp.1 An employee of a computer store recorded the number of sales he made each month. In the past 12 months, he sold the following numbers of computers:

51, 20, 25, 39, 7, 44, 92, 41, 22, 6, 42, 18.

Make the box and whisker plots.

First, put the data in ascending order. Then find the median.
N=12
6, 7, 18, 20, 22, 25, 39, 41, 42, 44, 51, 92

Median position = (N+1)/2 = (12 + 1) / 2 = 6.5th value
Median = (sixth + seventh observations) / 2 = (25 + 39) / 2 = 32

There are six numbers below the median, namely: 6, 7, 18, 20, 22, 25.
Q1 position = the median of these six items = (6 + 1) / 2 = 3.5th value
Q1 = (third + fourth observations) / 2 = (18 + 20) / 2 = 19

There are six numbers above the median, namely: 39, 41, 42, 44, 51, 92.
Q3 position = the median of these six items = (6 + 1) / 2 = 3.5th value
Q3 = (third + fourth observations) / 2 = (42+44) / 2 = 43

bp.1 An employee of a computer store recorded the number of sales he made each month. In the past 12 months, he sold the following numbers of computers:

51, 20, 25, 39, 7, 44, 92, 41, 22, 6, 42, 18.

Make the box and whisker plots.

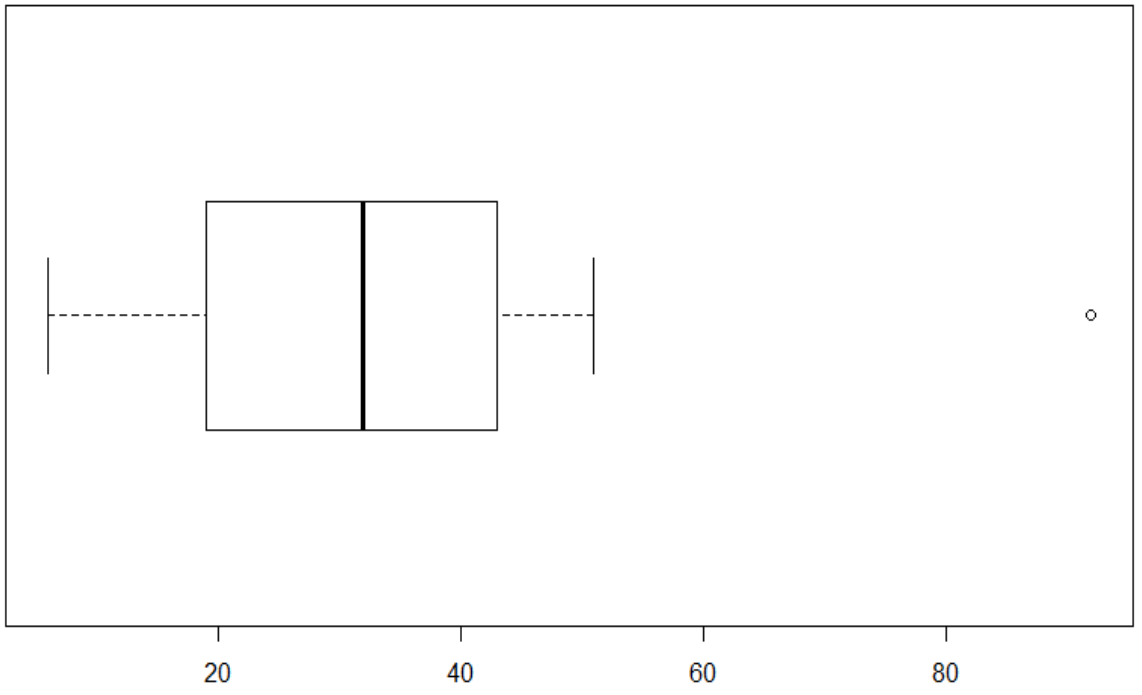
Median = 32
Q1 = 19
Q3 = 43

IQR= Q3 – Q1 = 43 - 19 = 24

Whiskers:
Upper = Q3 + 1.5 IQR = 43 + 1.5 · 24 = 43 + 36 = 79

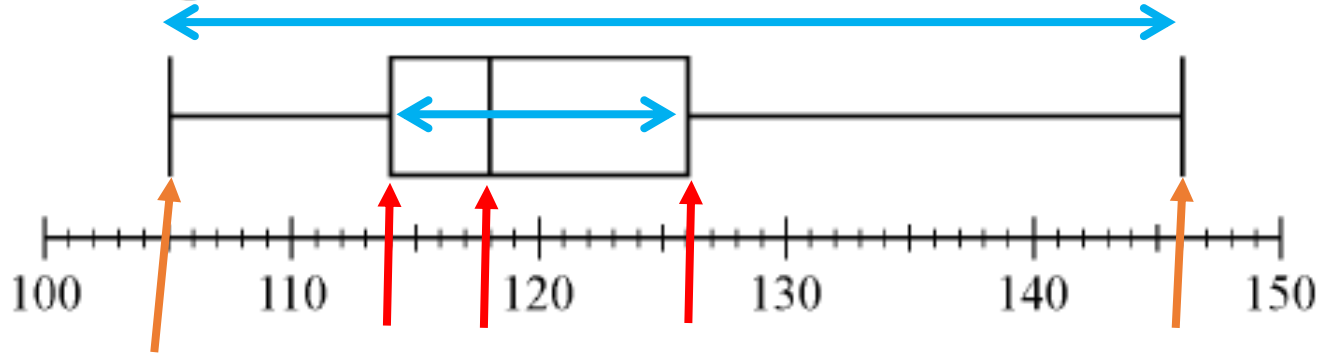
Lower = Q1 – 1.5 IQR = 19 – 36 = -17
(smaller than the minimum value)

1 Upper outlier (92)



bp.2

Eight hundred insects were weighed, and the resulting measurements, in milligrams, are summarized in the boxplot below.



(a) What are the range, the three quartiles, and the interquartile range of the measurements?

Solution
We have 800 data (n=800).

From the box-plot we can conclude that:
The minimum value is 105 and maximum value is 146
The **range** it is maximum minus smallest or 146−105=41
The **median** is 118.
First quartile is 114 and **third** is 126.
The **IQR (Interquartile Range)** is 126−114=12