Politecnico di Bari

Politecnico di Bari

Dipartimento di Ingegneria Elettrica e dell'Informazione

Corso di Laurea Triennale in Ingegneria dei Sistemi Medicali

DIPARTIMENTO DI INGEGNERIA ELETTRICA E DELL'INFORMAZIONE

Bioinformatics and Big Data Analytics

Outliers Detection

Eng. Nicola **Altini**, *Ph.D. Student*
Eng. Giacomo Donato **Cascarano**, *Ph.D. Student*
Prof. *Eng*. Vitoantonio **Bevilacqua**, *Ph.D.*

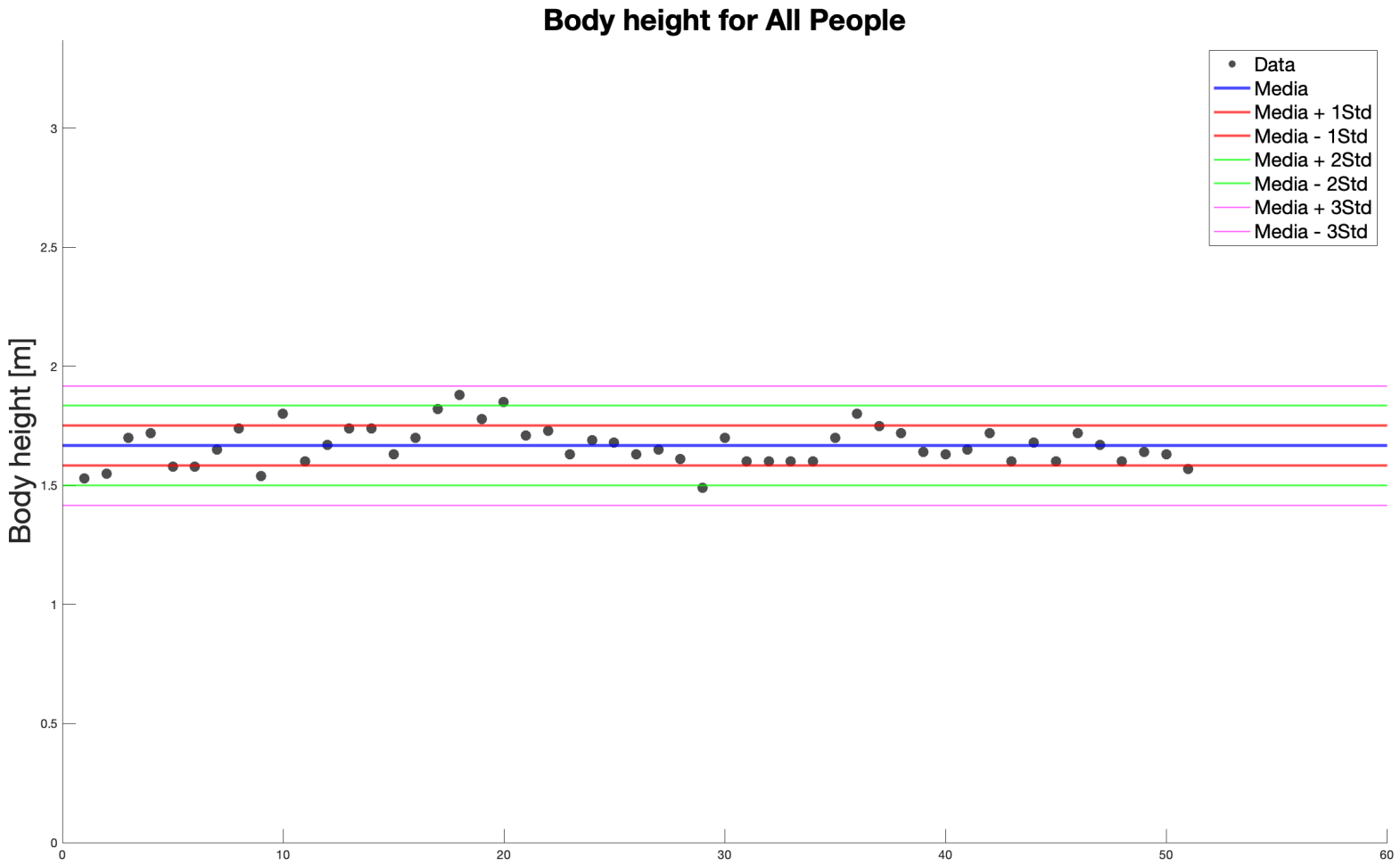
INDUSTRIAL INFORMATICS LAB
LABORATORIO DI INFORMATICA INDUSTRIALE

Anno Accademico 2019/2020

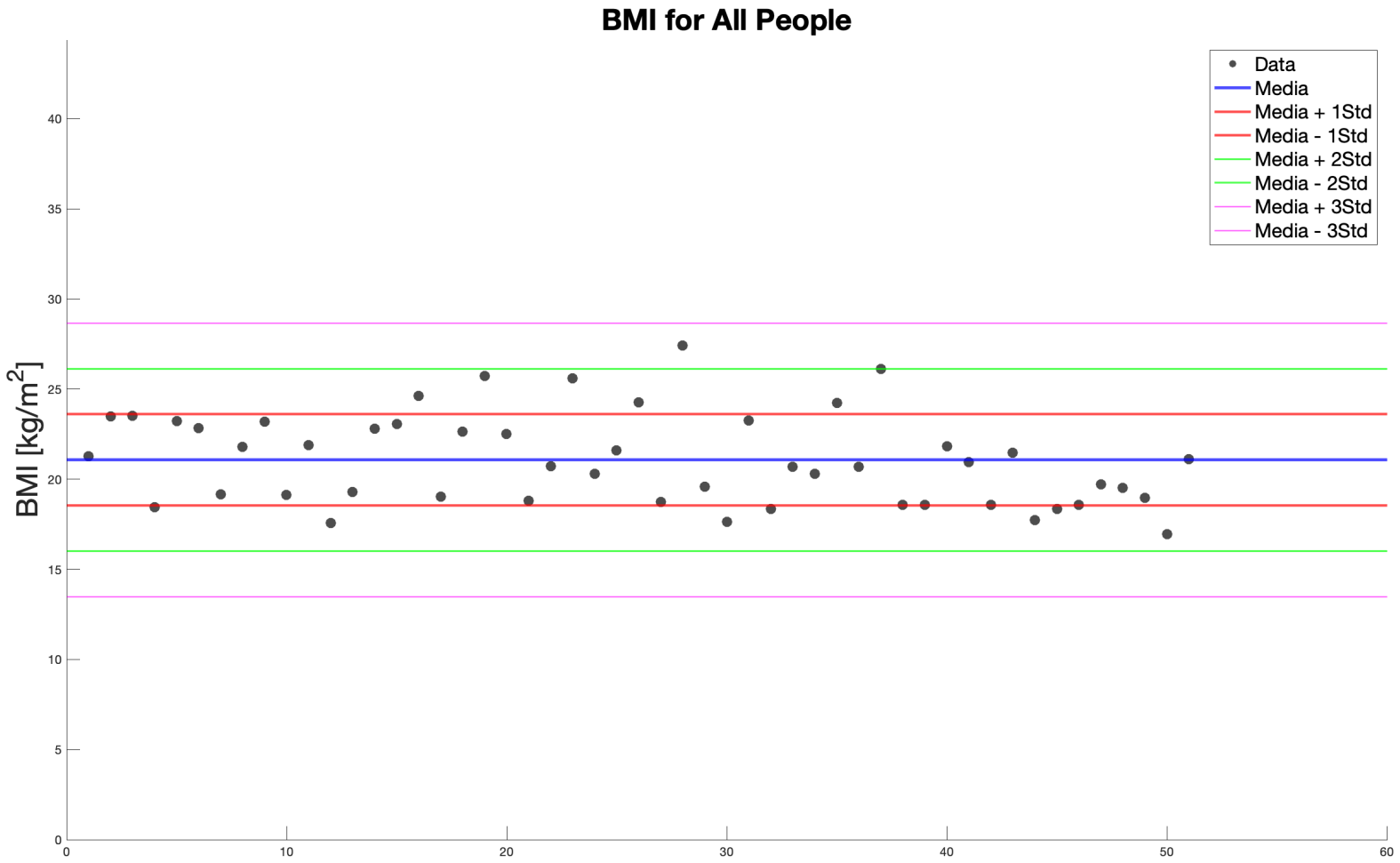
apulian bioengineering company

Mean and Standard Deviation in MATLAB

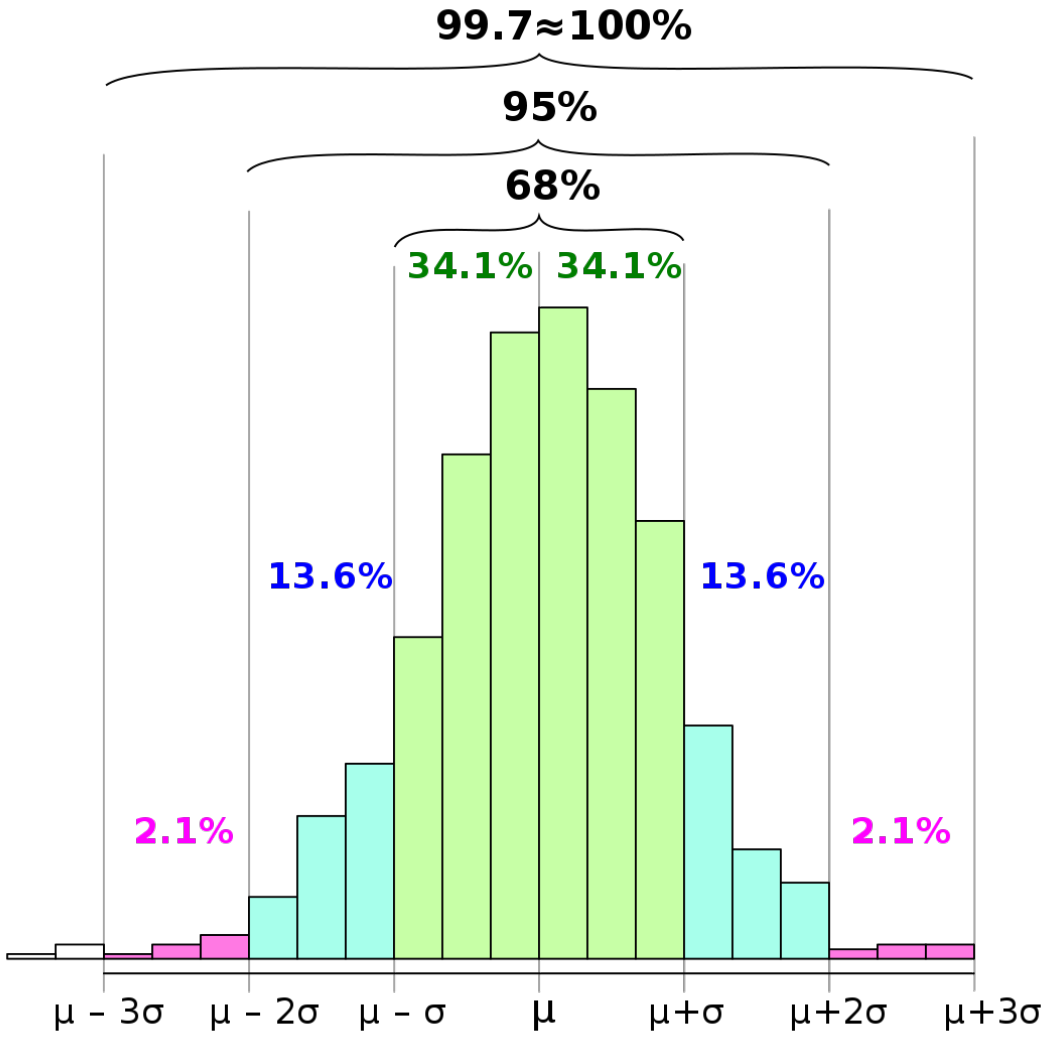
Mean and Standard Deviation in MATLAB



Mean and Standard Deviation in MATLAB



Outliers detection with standard deviation



Outliers detection with standard deviation

- The **68–95–99.7 rule**, also known as the empirical rule, is a shorthand used to remember the percentage of values that lie within a band around the mean in a normal distribution with a width of two, four and six standard deviations, respectively; more accurately, 68.27%, 95.45% and 99.73% of the values lie within one, two and three standard deviations of the mean, respectively.
- In the case of normally distributed data, the **three sigma rule** means that roughly 1 in 22 observations will differ by twice the standard deviation or more from the mean, and 1 in 370 will deviate by three times the standard deviation.

Outliers detection with standard deviation

- 68.27%, 95.45% and 99.73% of the values lie within one, two and three standard deviations of the mean, respectively.

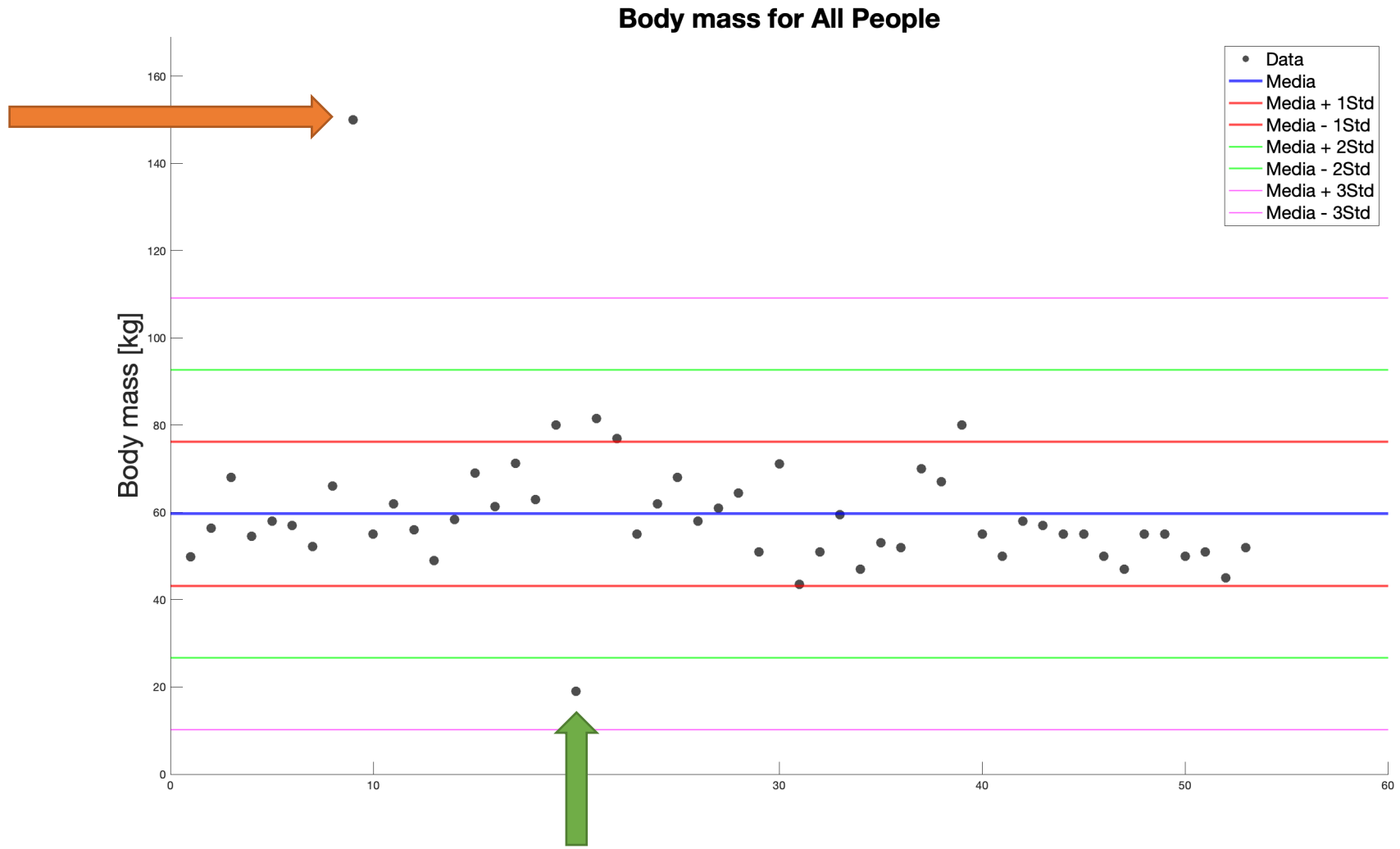
$$\Pr(\mu - 1\sigma < X < \mu + 1\sigma) \cong 0.6827$$

$$\Pr(\mu - 2\sigma < X < \mu + 2\sigma) \cong 0.9545$$

$$\Pr(\mu - 3\sigma < X < \mu + 3\sigma) \cong 0.9973$$

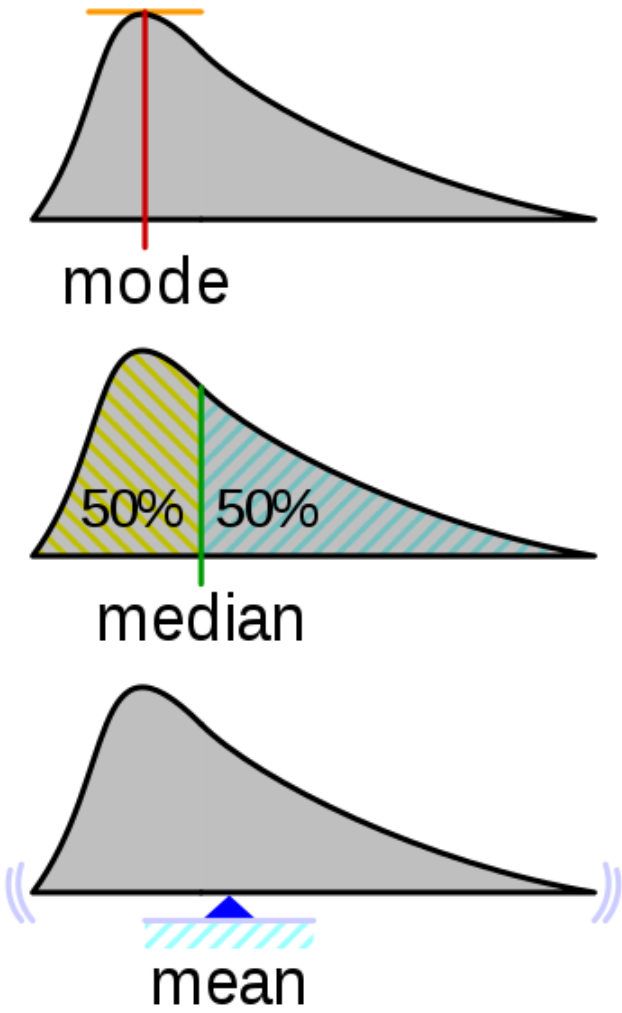
- Rule for outlier detection with standard deviation:
- If $data < \mu - 3\sigma$ or $data > \mu + 3\sigma$, then $data$ is considered an outlier.

Outliers detection with standard deviation



Median

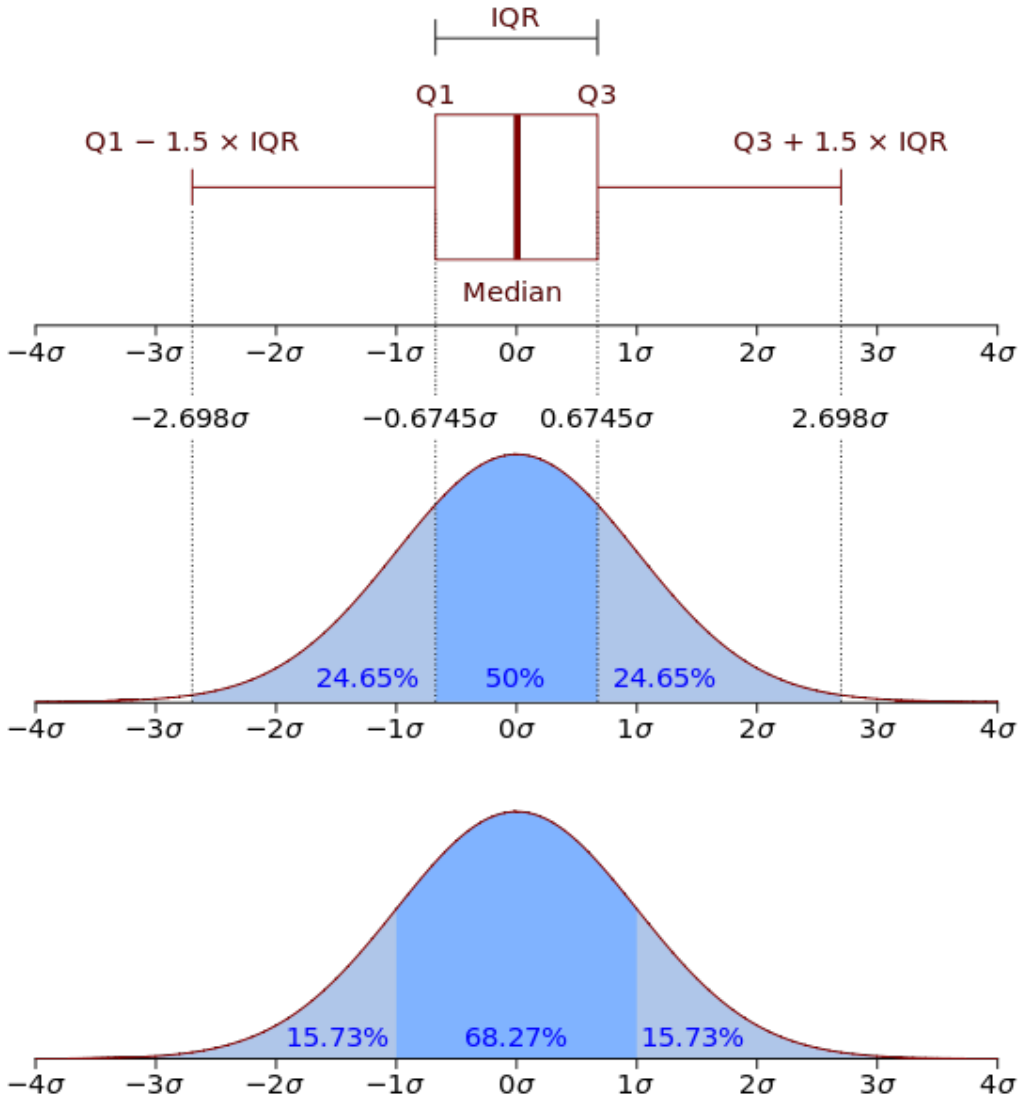
- The median is the value separating the higher half from the lower half of a data sample, a population or a probability distribution. For a data set, it may be thought of as the "middle" value.
- The basic advantage of the median in describing data compared to the mean (often simply described as the "average") is that it is not skewed so much by a small proportion of extremely large or small values, and so it may give a better idea of a "typical" value.
- For example, in understanding statistics like household income or assets, which vary greatly, the mean may be skewed by a small number of extremely high or low values.



Quartiles

- A quartile is a type of quantile which divides the number of data points into four more or less equal parts, or quarters.
- The first quartile (Q1) is defined as the middle number between the smallest number and the median of the data set. It is also known as the lower quartile or the 25th empirical quartile and it marks where 25% of the data is below or to the left of it (if data is ordered on a timeline from smallest to largest).
- The second quartile (Q2) is the median of the data and 50% of the data lies below this point.
- The third quartile (Q3) is the middle value between the median and the highest value of the data set. It is also known as the upper quartile or the 75th empirical quartile and 75% of the data lies below this point.

Quartiles


IQR
Q1
Q3
Q1 − 1.5 × IQR
Q3 + 1.5 × IQR
Median
−4σ −3σ −2σ −1σ 0σ 1σ 2σ 3σ 4σ
−2.698σ
−0.6745σ
0.6745σ
2.698σ
24.65%
50%
24.65%
−4σ −3σ −2σ −1σ 0σ 1σ 2σ 3σ 4σ
15.73%
68.27%
15.73%
−4σ −3σ −2σ −1σ 0σ 1σ 2σ 3σ 4σ

Box plot

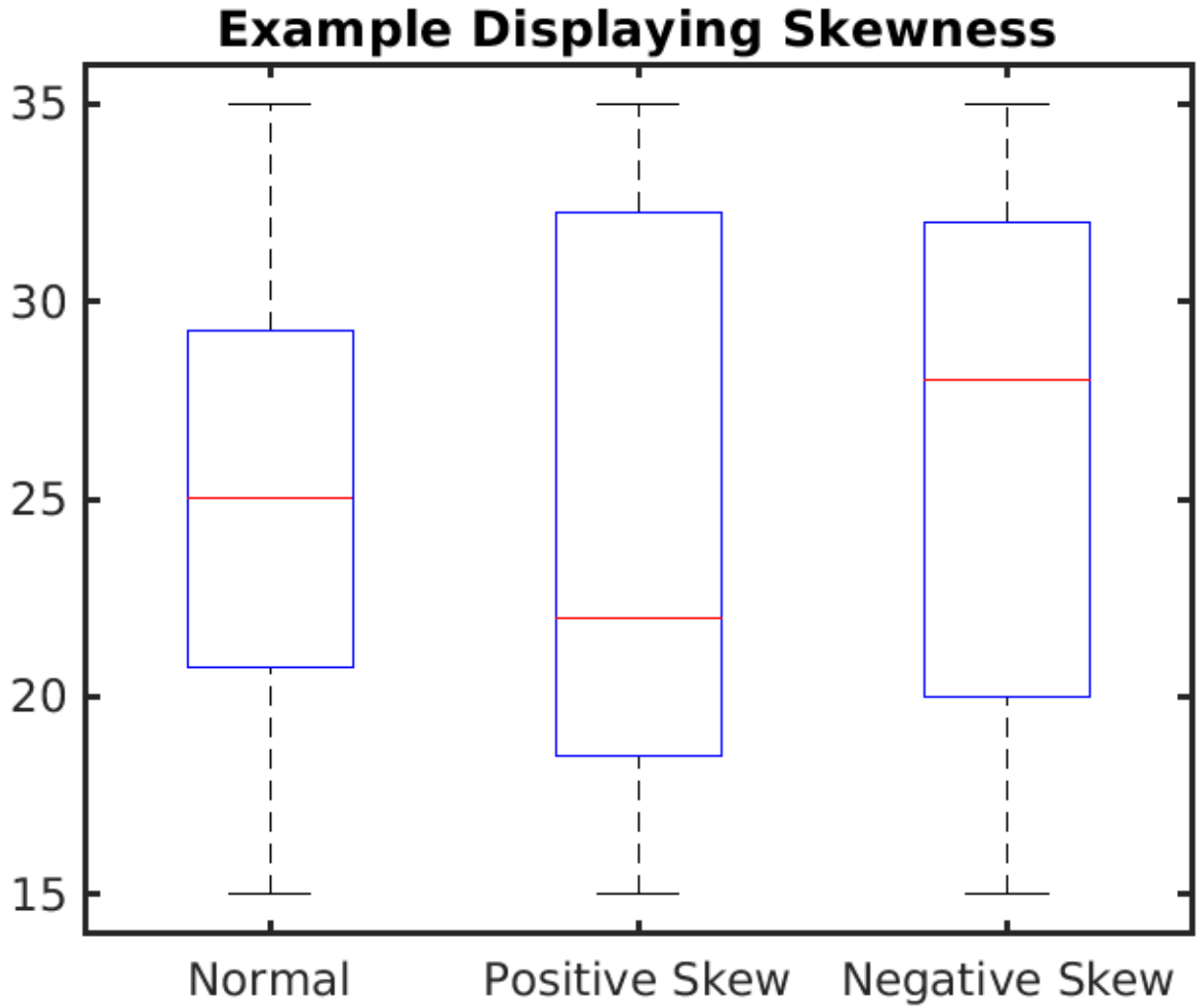
- Box plots are a tool for graphically depicting groups of numerical data through their quartiles. Box plots may also have lines extending from the boxes (whiskers) indicating variability outside the upper and lower quartiles, hence the terms box-and-whisker plot and box-and-whisker diagram. Outliers may be plotted as individual points.
- Box plots are non-parametric: they display variation in samples of a statistical population without making any assumptions of the underlying statistical distribution

Box plot

- **Median** (**Q2** / 50th Percentile): the middle value of the dataset.
- **First quartile** (**Q1** / 25th Percentile): is also known as the lower quartile qn(0.25) and is the middle value between the smallest number (not the minimum) and the median of the dataset.
- **Third quartile** (**Q3** / 75th Percentile): is also known as the upper quartile qn(0.75) and is the middle value between the largest number (not the maximum) and the median of the dataset.
- **Interquartile Range** (**IQR**): is the distance between the upper and lower quartile.

$$IQR = Q3 - Q1$$

Skewness example



Basic Statistics in MATLAB

- **Mean value of array**

```
M = mean(A)
```

- **Variance**

```
V = var(A)
```

- **Standard deviation**

```
S = std(A)
```

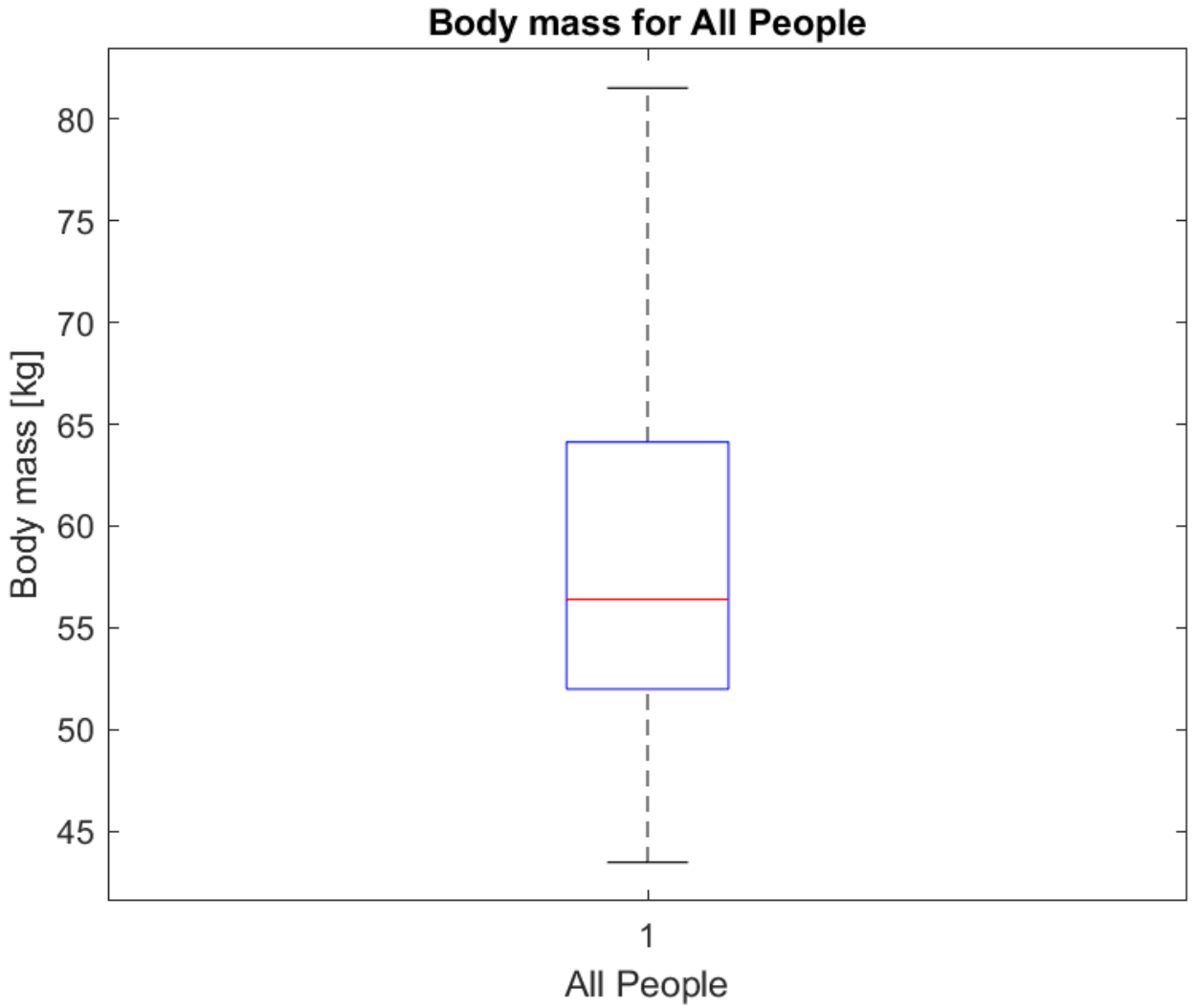
- **Minimum or maximum elements of an array**

```
M = min(A)
M = max(A)
```

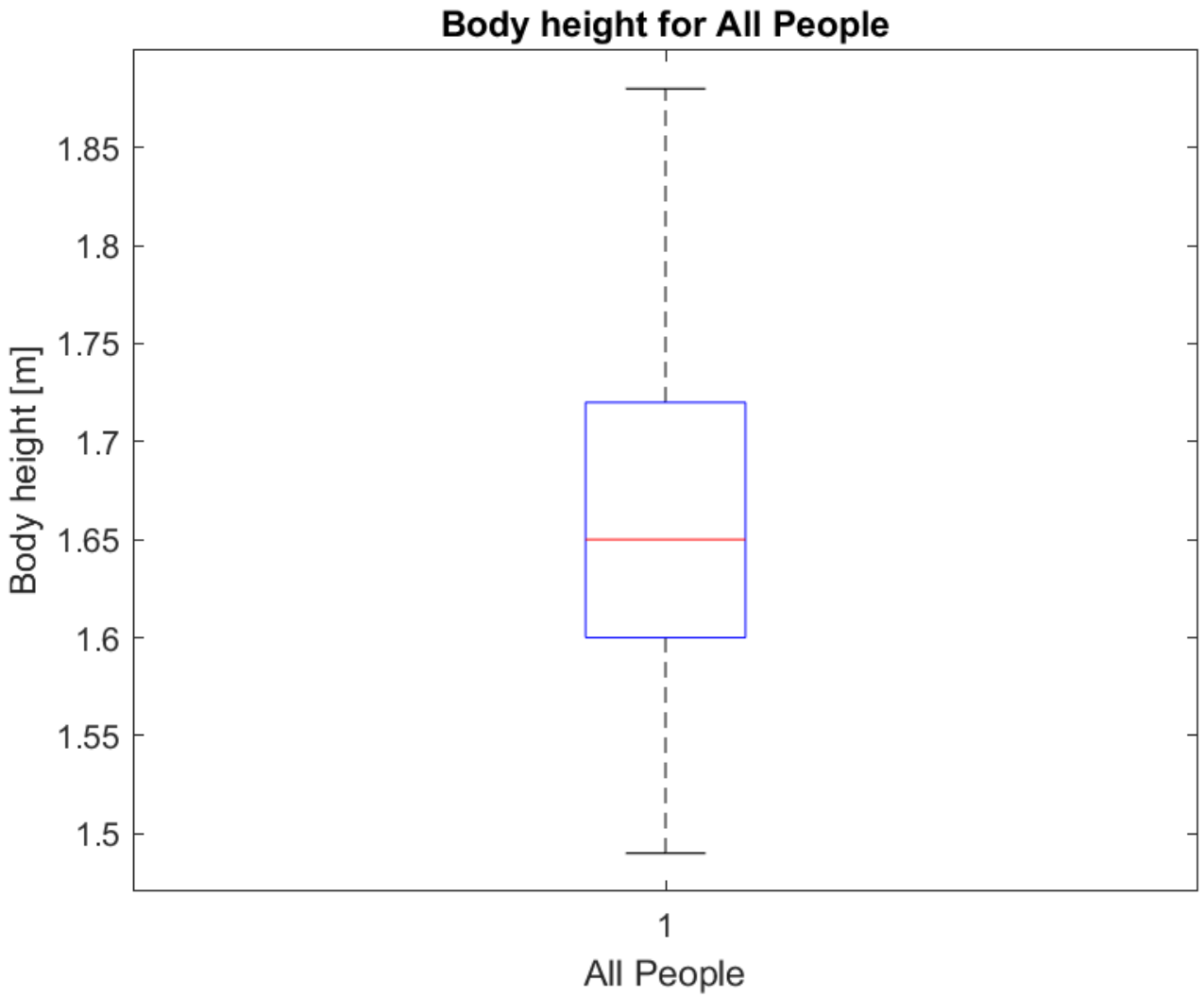
- **Boxplot of an array**

```
boxplot(X)
```

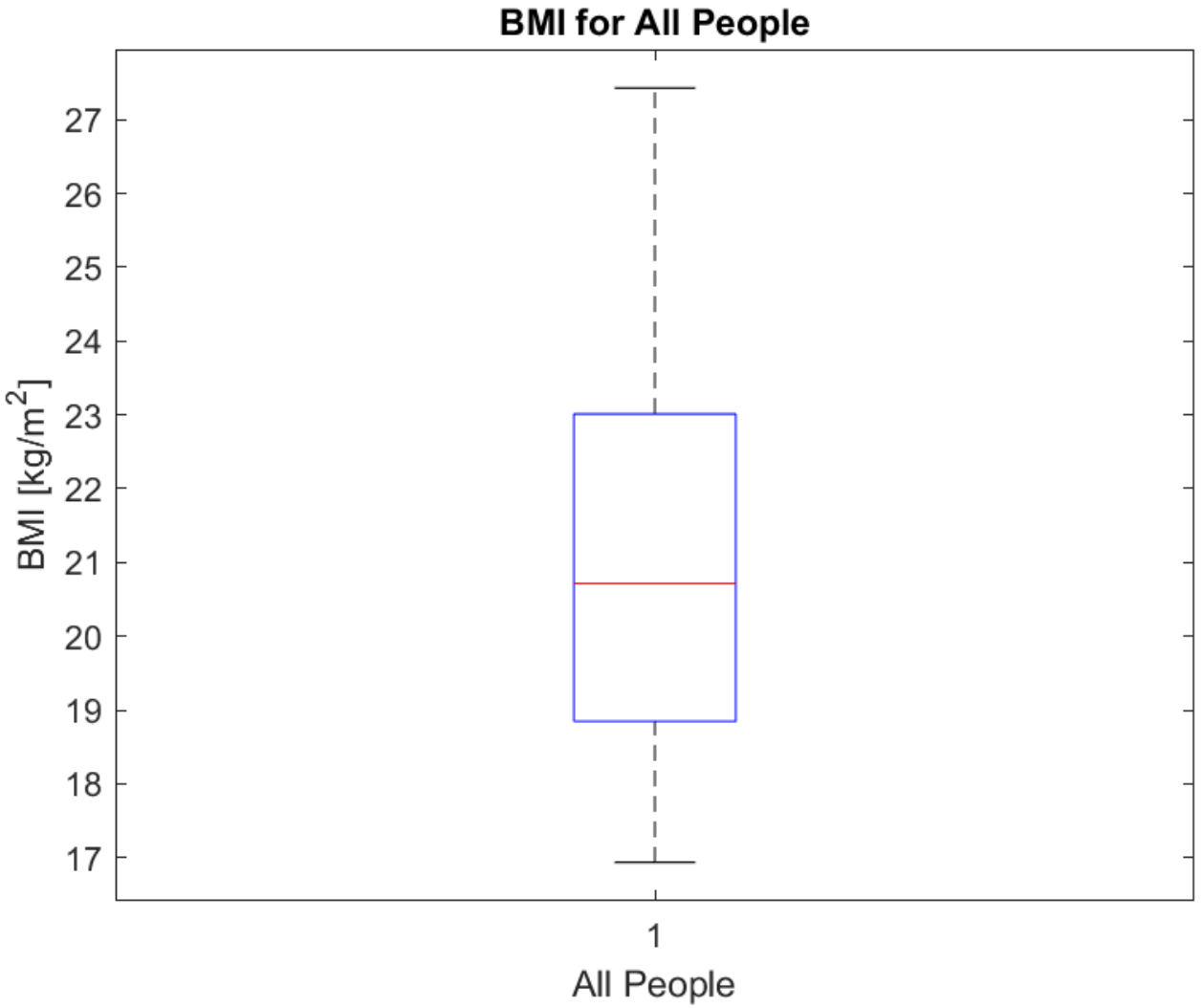
Boxplot in MATLAB



Boxplot in MATLAB



Boxplot in MATLAB



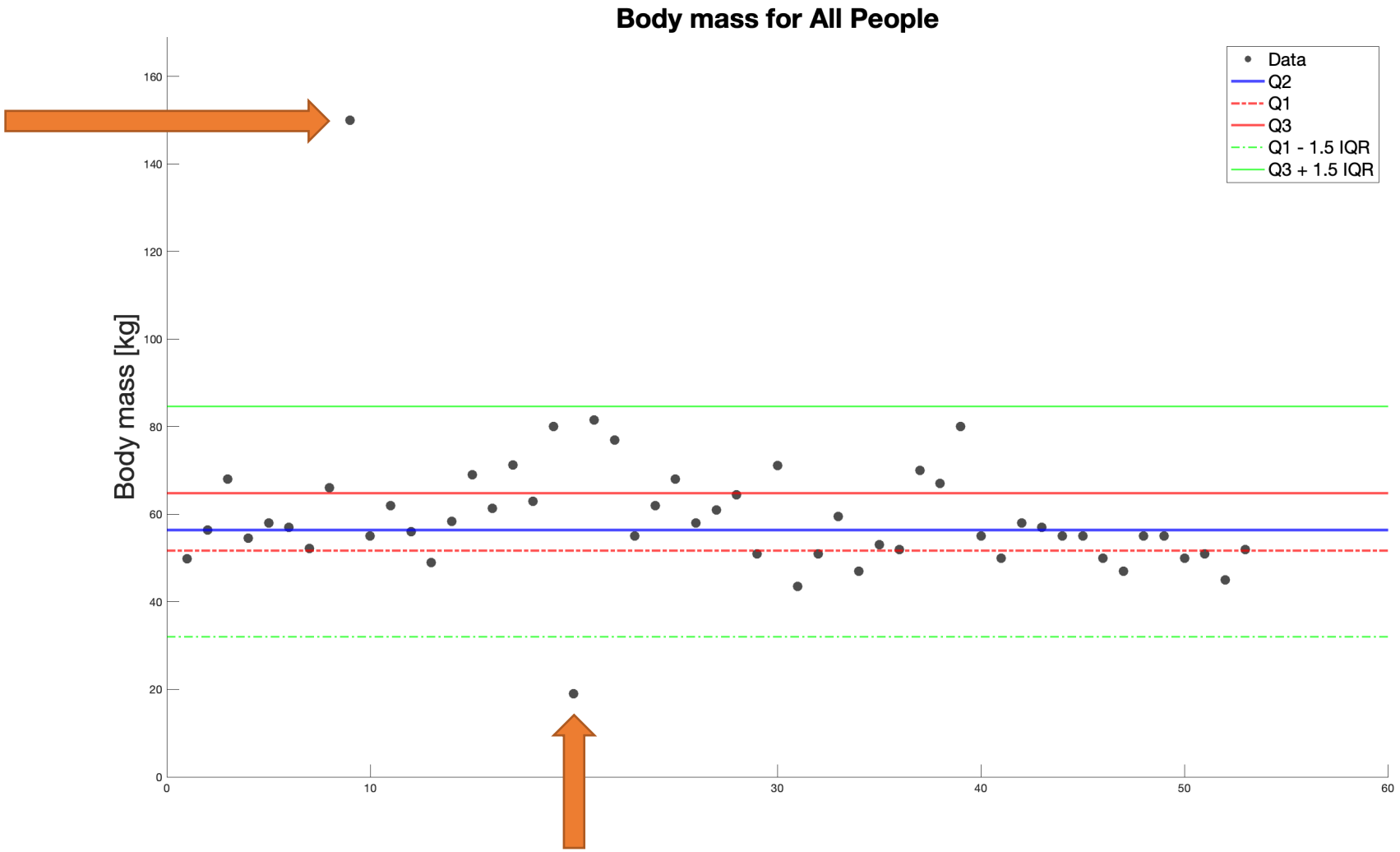
Outliers detection with quartiles

- The interquartile range (IQR) can be used to identify outliers in a data set.

$$IQR = Q3 - Q1$$

- Rule for outlier detection with IQR:
- If $data < Q1 - 1.5 * IQR$ or $data > Q3 + 1.5 * IQR$, then $data$ is considered an outlier.

Outliers detection with quartiles



References

- Wikipedia
  - Median.
    URL: https://en.wikipedia.org/wiki/Median
  - Quartile.
    URL: https://en.wikipedia.org/wiki/Quartile
  - Boxplot.
    URL: https://en.wikipedia.org/wiki/Box_plot