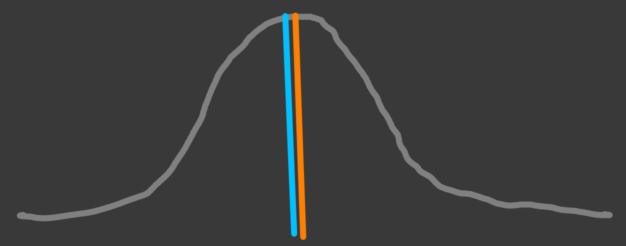
Statistics Data Science

in a normal distribution



they overlap but in other distribution they differe

Estimates of Location / measure of central tendency

Mean : The sum of all values divided by the number of values.

Median : The value such that one-half of the data lies above and below.
Synonym : 50th percentile

aka → middle data point in Sorted data

for odd count ⇒ $m = \frac{m_{-1} + m_{+1}}{2}$

Estimates of Variability / measure of dispersion or spread

Deviations : The difference between the observed values and the estimate of location. $= O - E$
Synonyms : errors, residuals

$O = X$, $E = \bar{X}_{population} / \bar{x}_{mean}$

Variance : The sum of squared deviations from the mean divided by n – 1 where n is the number of data values.
Synonym : mean-squared-error

$$s^2 = \frac{\sum_{i=1}^{n} (O_i - E)^2}{n-1}$$

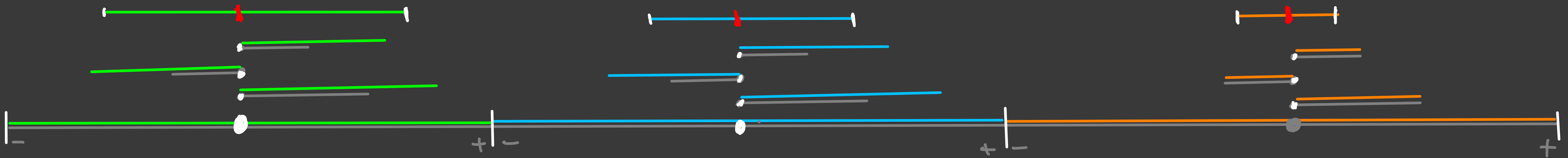
Standard deviation : The square root of the variance. $S = \sqrt{var}$

Mean absolute deviation : The mean of the absolute values of the deviations from the mean.

Synonyms : L1-norm, Manhattan norm

$$L1 = \frac{\sum_{i=1}^{n} abs(O-E)}{n}$$

$$abs(-Z/+Z) \Rightarrow +Z$$



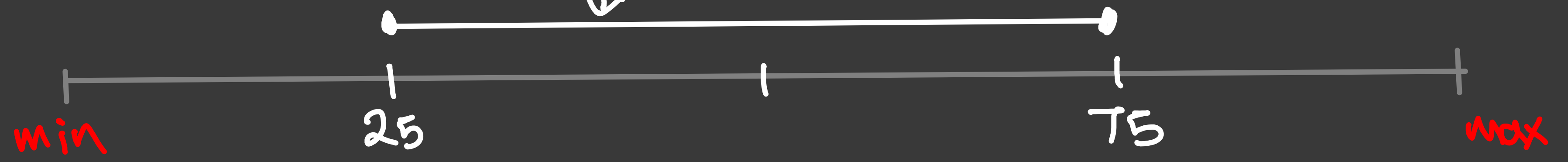
Range : The difference between the largest and the smallest value in a data set.

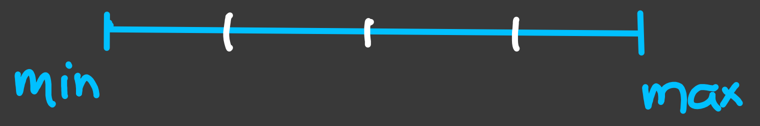
Interquartile range : The difference between the 75th percentile and the 25th percentile.

Synonym : IQR

* after sorting data 25% lie left of point X

Range = max − min





*More robust metrics include mean absolute deviation, median absolute deviation from the median, and percentiles (quantiles).

Robust: Resistant toward outliers

mid: median of set/data of $O_{1\dots n}$ observation

Median absolute deviation from the median :
The median of the absolute values of the deviations from the median.

$$MAD = \text{median}(\,\text{abs}(O_1 - \text{mid}),\ \text{abs}(O_2 - \text{mid}),\ \dots)$$

Percentile : The value such that P percent of the values take on this value or less and (100–P) percent take on this value or more.

Synonym : quantile

100% → 1.0
30% → 0.3
28% → 0.28

let x=15

$$\text{Percentile} = \frac{\text{count of O below X}}{\text{count of O total}} = \frac{\text{below count}(X)}{50} = \frac{14}{50} = 0.28$$

$$O_{1\dots50} = \text{Sorted}(1, 2, 3, \dots, 50)$$

50 percentile -> 50 % point in Observation are below some point X

↓

0.5

5 number summary

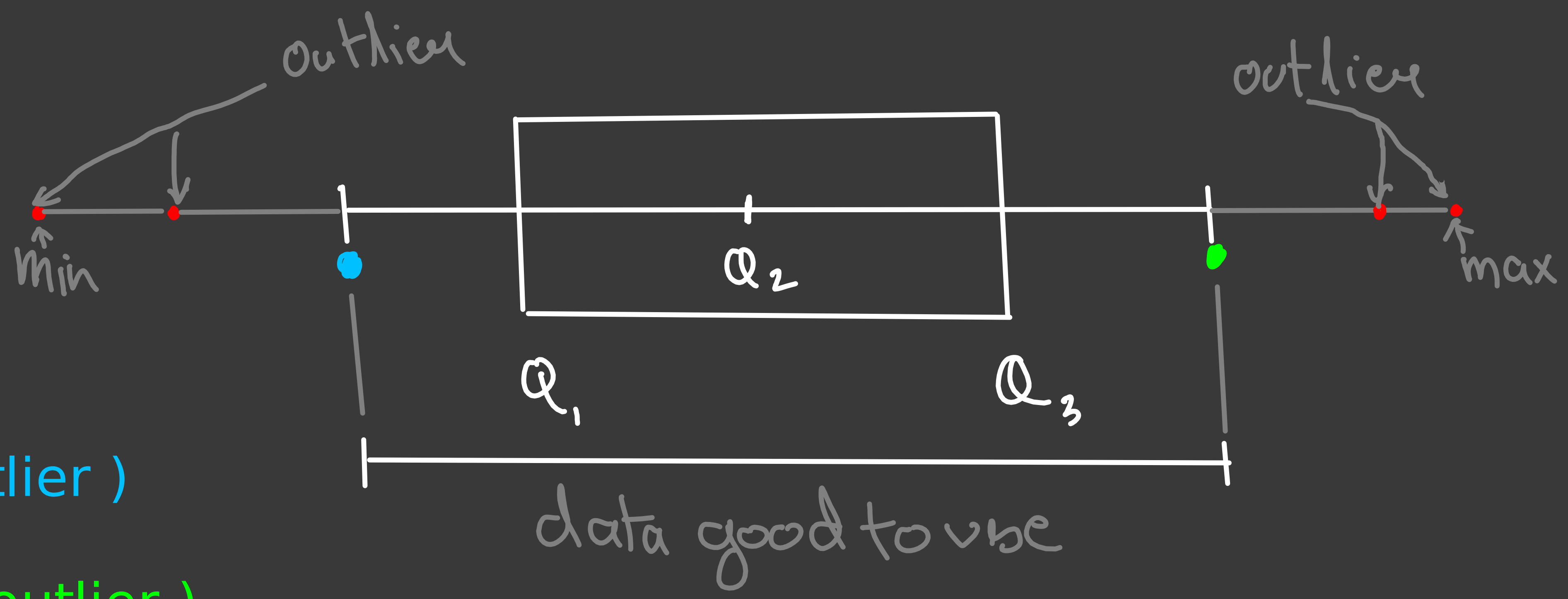
data after sort = D

1. min of
2. max of

D

3. First Quartile / Q1 / 25th percentile of
4. median / 50th percentile of
5. Third Quartile / Q3 / 75th percentile of

D



lower limit / fence (left (<--) this point consider outlier)

higher limit / fence (right (-->) this point consider outlier)

IRQ : is the difference in Q3 and Q1 $IRQ = Q3 - Q1$

$$\text{lower limit} = Q1 - (1.5 \times IRQ)$$

$$\text{higher limit} = Q3 + (1.5 \times IRQ)$$