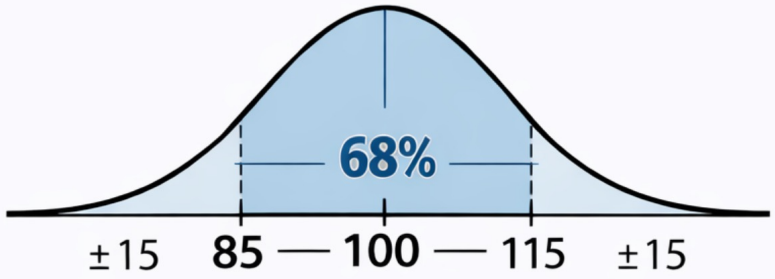
The CV as a measure of variability around the mean, i.e., the CV tells us how risky it is to use the mean as a predictor of individual observations.

CV = 0.15

Mean = 100

SD = 15

68% within 85 to 115



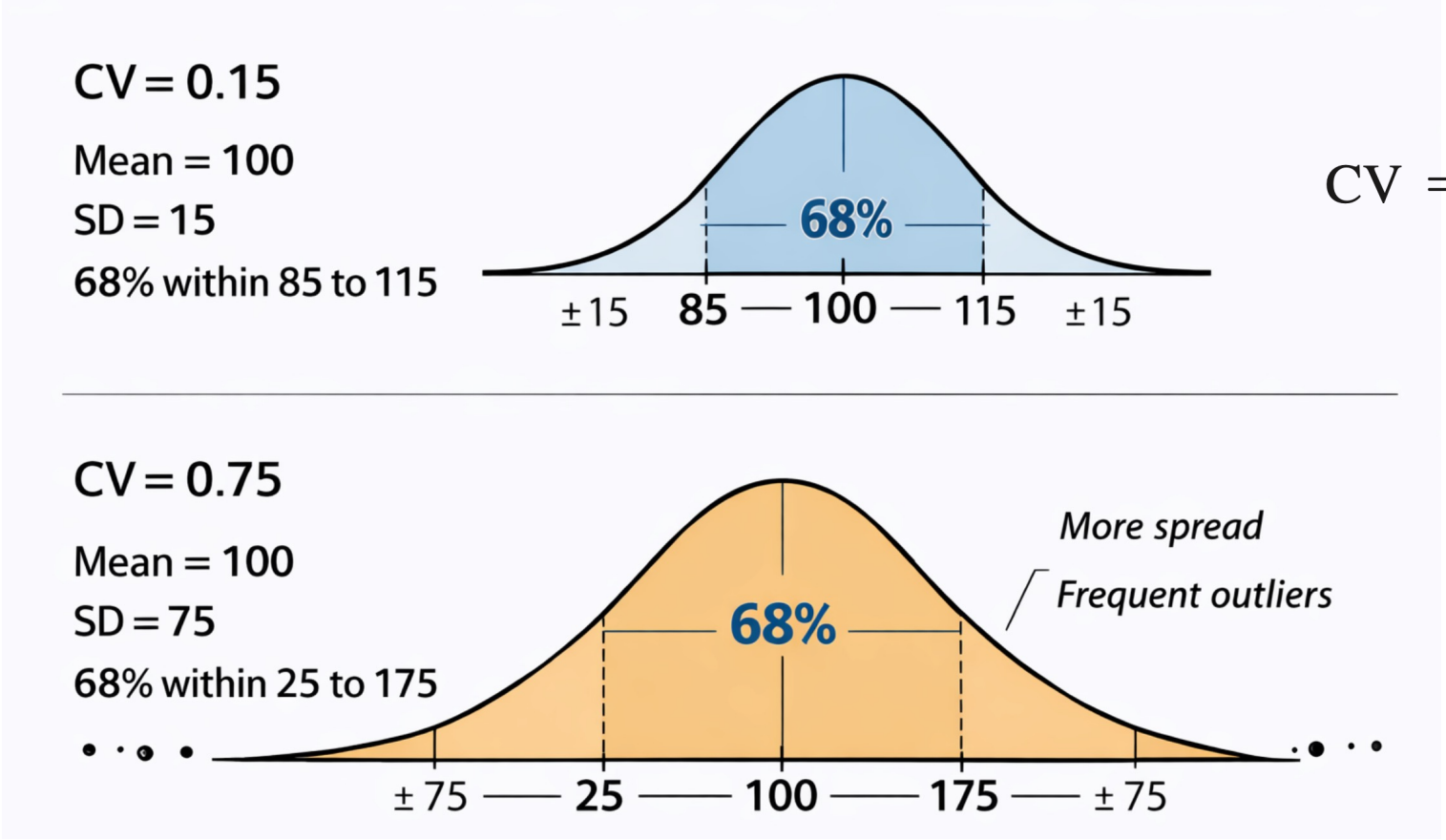
$$CV = \frac{s}{\bar{Y}} \times 100\%$$

Saying that a CV of 0.15 means that typical deviations are about 15% of the mean means that the standard deviation equals 0.15 × mean.

If the data are approximately normally distributed, about two-thirds (~68% to be exact) of the observations will fall within one standard deviation of the mean, that is, within15 units of the mean (i.e., between 85 and 115).

A percentage such as 15% of the mean describe the scale of variation, not exact probabilities. Normality converts this scale into specific frequencies; non-normal distributions change the frequencies, not the relative amount of variation.

The CV as a measure of variability around the mean, i.e., the CV tells us how risky it is to use the mean as a predictor of individual observations.



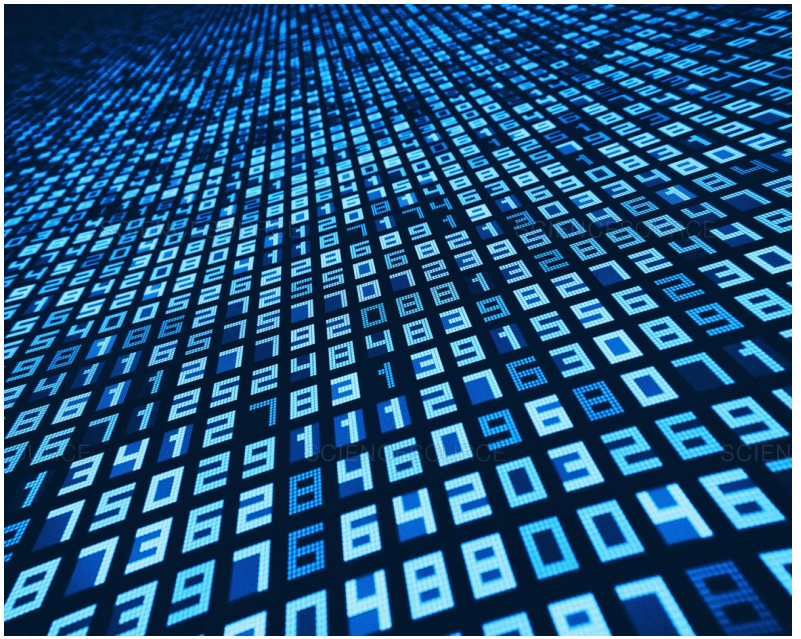
$$CV = \frac{s}{\bar{Y}} \times 100\%$$

In both examples above the mean is identical, but the relative spread of the data, and therefore the variability associated with using the mean as a predictor, is dramatically different.

Describing data

Samples and populations are often made of lots of individual (observational) units and their associated information (observations, variables).

We need to be able to describe samples by summary statistics (mean, median, variance, etc) so that these summaries can serve as an estimate of the same summaries for their statistical populations.



How do measures of location (mean versus median) and spread (standard deviation versus interquartile range) compare?

Disarming fish (protection against predation)

Threespine stickleback
(*Gasterosteus aculeatus*)

Plate Genotypes Ectodysplasin (Eda) locus (3rd generation)

MM (marine)

Mm (hybrid)

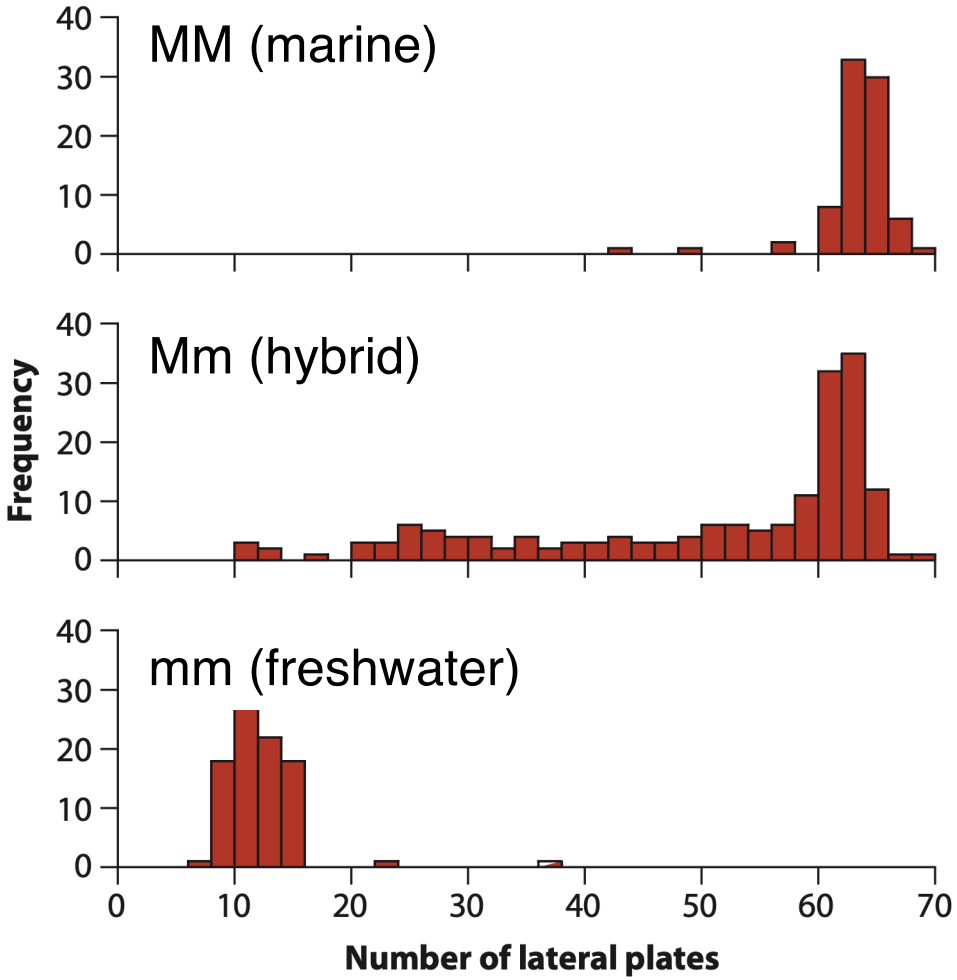
mm (freshwater)

Variation is at the heart of biology! The diversity among individuals and species forms the foundation of biological science.

How do measures of location (mean versus median) and spread (standard deviation versus interquartile range) compare?

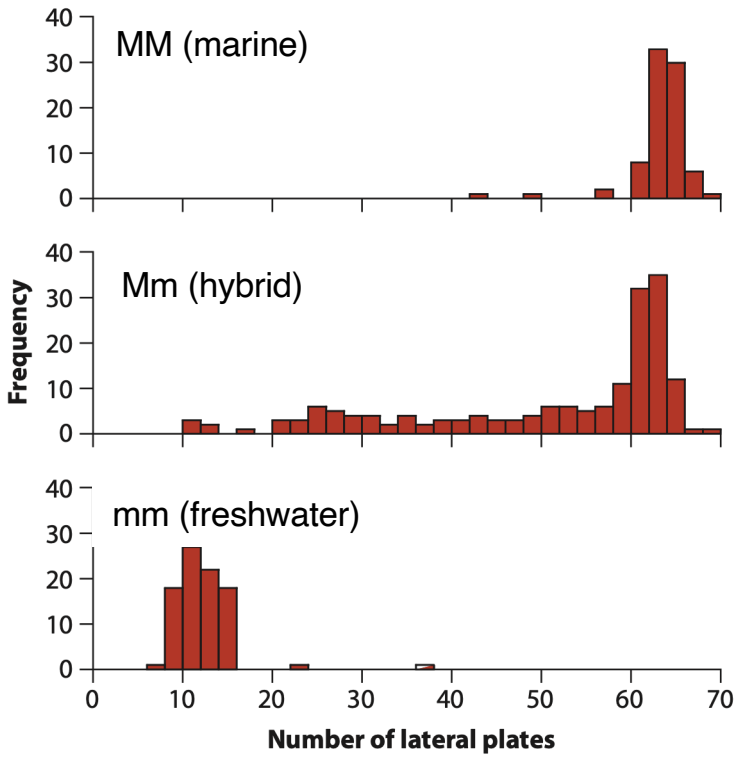
Disarming fish (protection against predation)

Variation is at the heart of biology! The diversity among individuals and species forms the foundation of biological science



Which distribution is more asymmetric?

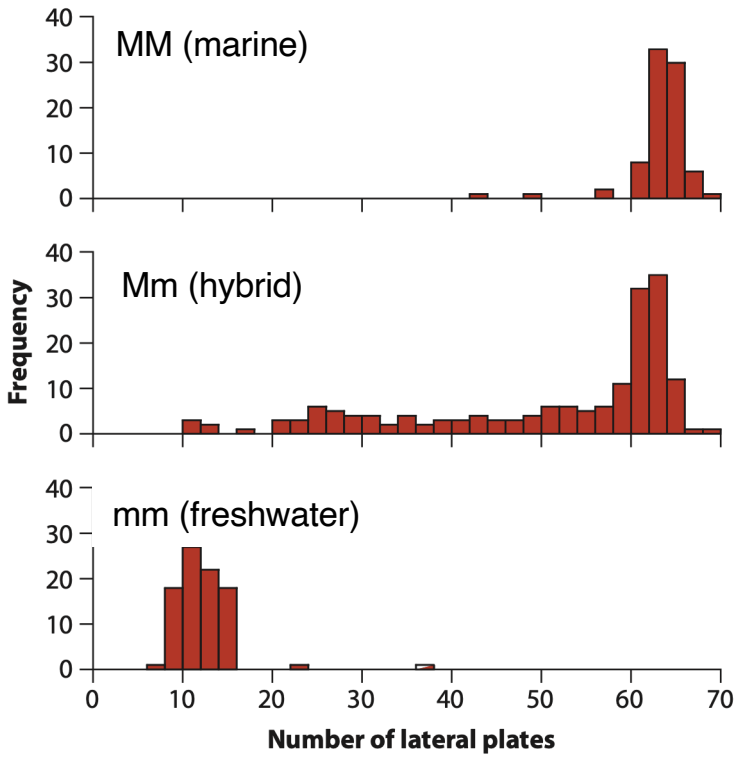
How do measures of location (mean versus median) and spread (standard deviation versus interquartile range) compare?



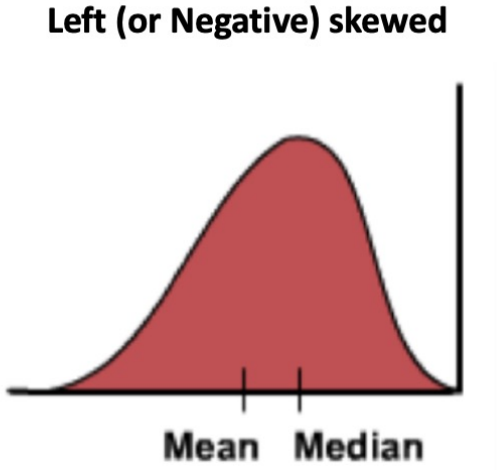
Which distribution is more asymmetric?

Genotype	*n*	Mean	Median	Standard deviation	Interquartile range
MM	82	62.8	63	3.4	2
Mm	174	50.4	59	15.1	21
mm	88	11.7	11	3.6	3

How do measures of location (mean versus median) and spread (standard deviation versus interquartile range) compare?

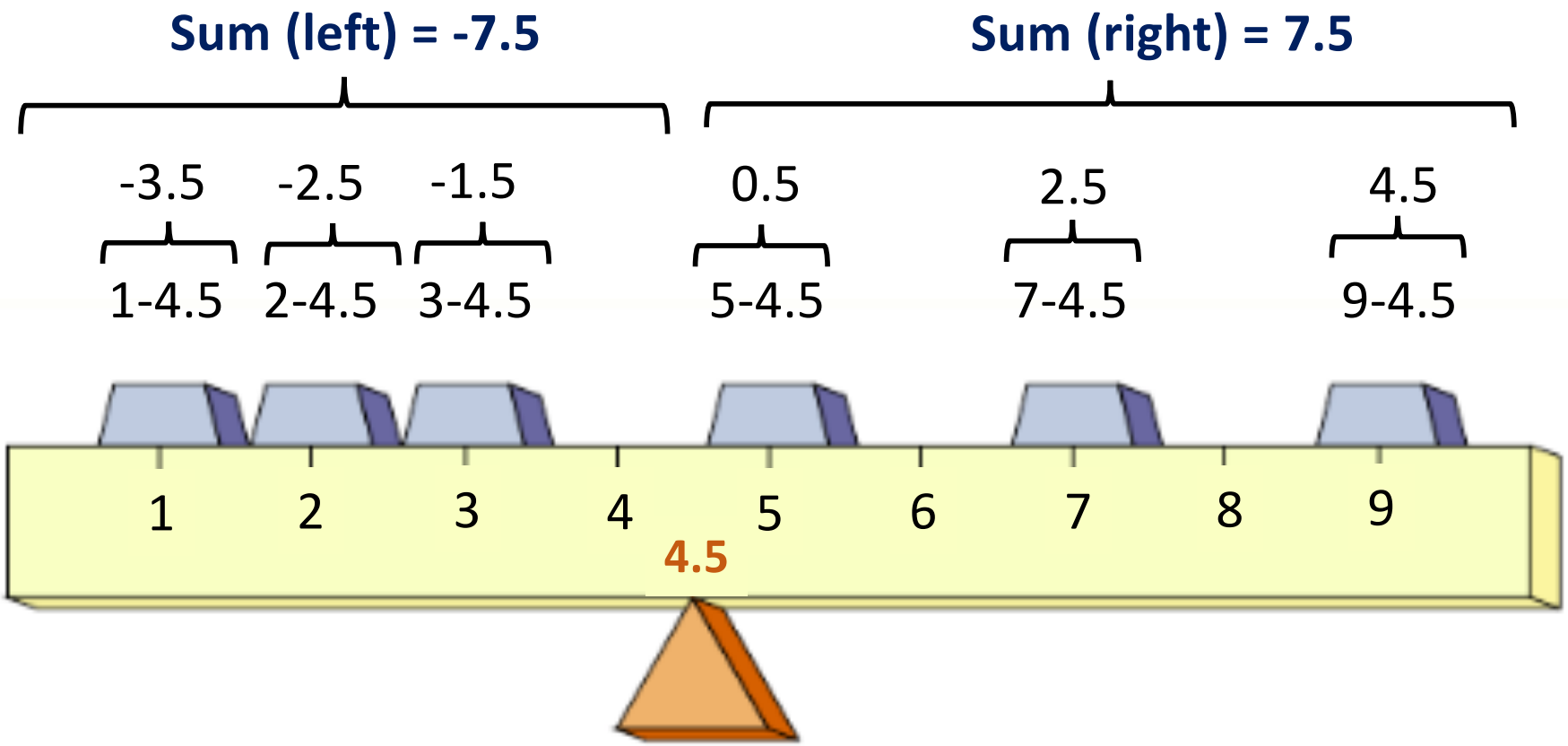


Which distribution is more asymmetric?



Genotype	*n*	Mean	Median	Standard deviation	Interquartile range
MM	82	62.8	63	3.4	2
Mm	174	50.4	59	15.1	21
mm	88	11.7	11	3.6	3

The mean can be understood as the center of gravity of a distribution: the point at which the values on either side balance each other



$\overline{X}=4.5$

Sum (left) + Sum (right) = -7.5 + 7.5 = 0

Recall from lecture 5: the sum of deviations from the mean is always zero, making the mean the 'center of gravity' of a distribution.

Quantities needed to calculate the standard deviation and variance of snake undulation rate ($\overline{Y} = 1.375\ Hz$).

Observations (Y_i)	Deviations $(Y_i - \overline{Y})$	Squared deviations $(Y_i - \overline{Y})^2$
0.9	−0.475	0.225625
1.2	−0.175	0.030625
1.2	−0.175	0.030625
1.3	−0.075	0.005625
1.4	0.025	0.000625
1.4	0.025	0.000625
1.6	0.225	0.050625
2.0	0.625	0.390625
Sum	0.000	0.735

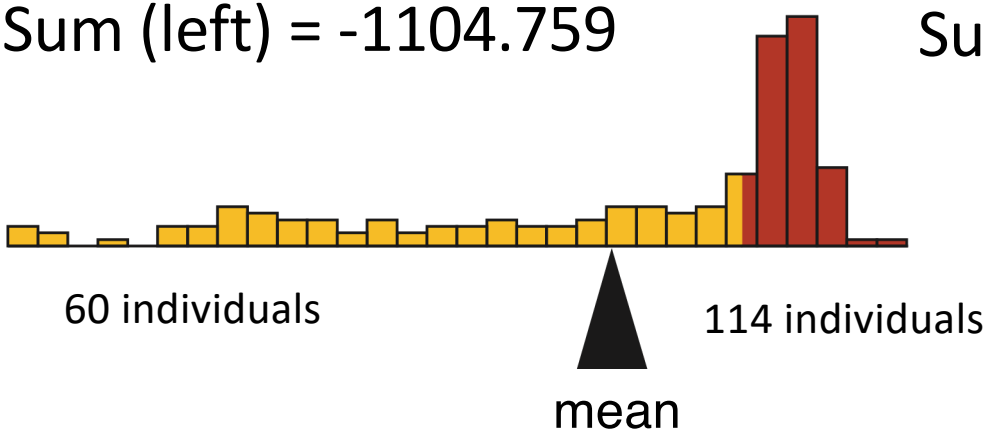
Mean ($\bar{Y}$) versus Median (referred as to Q_2)

Mm (hybrid) - most asymmetric distribution

Mean is the "centre of gravity";
it balances sums

Sum (left) = -1104.759

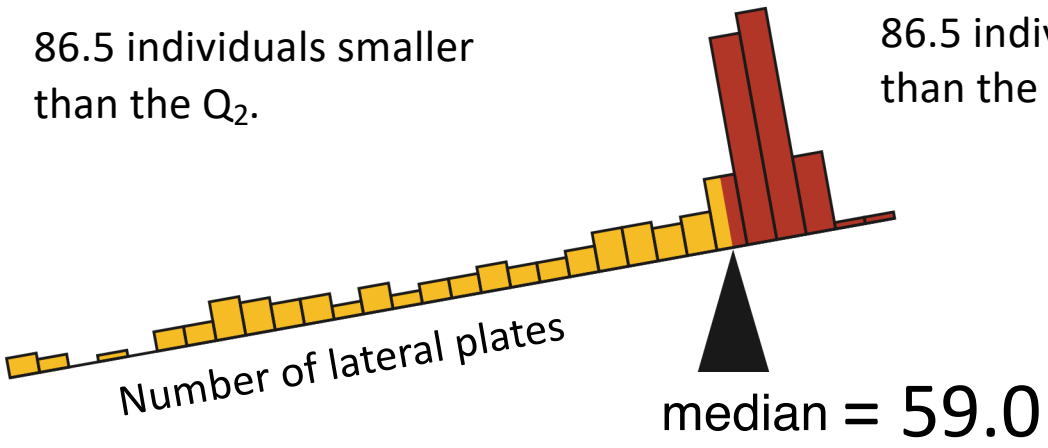
Sum (right) = 1104.759



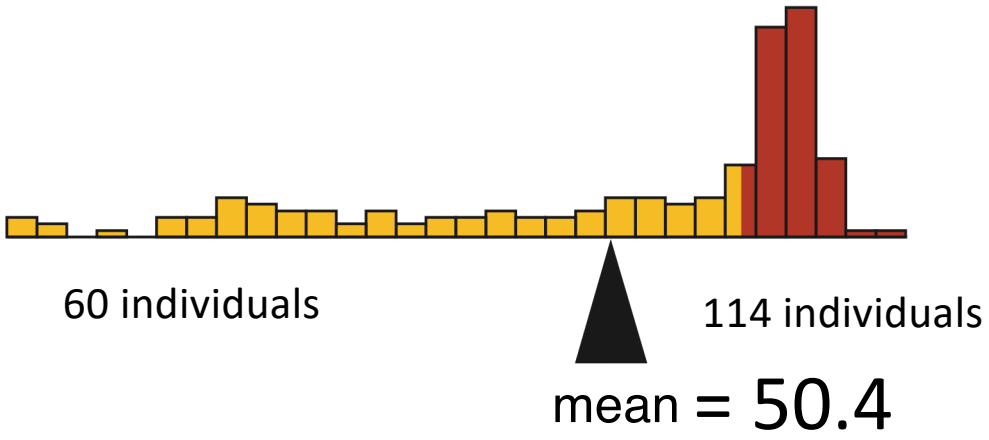
$\overline{Y}$=50.4

Mean ($\bar{Y}$) versus Median (referred as to Q_2)

Mm (hybrid) - most asymmetric distribution

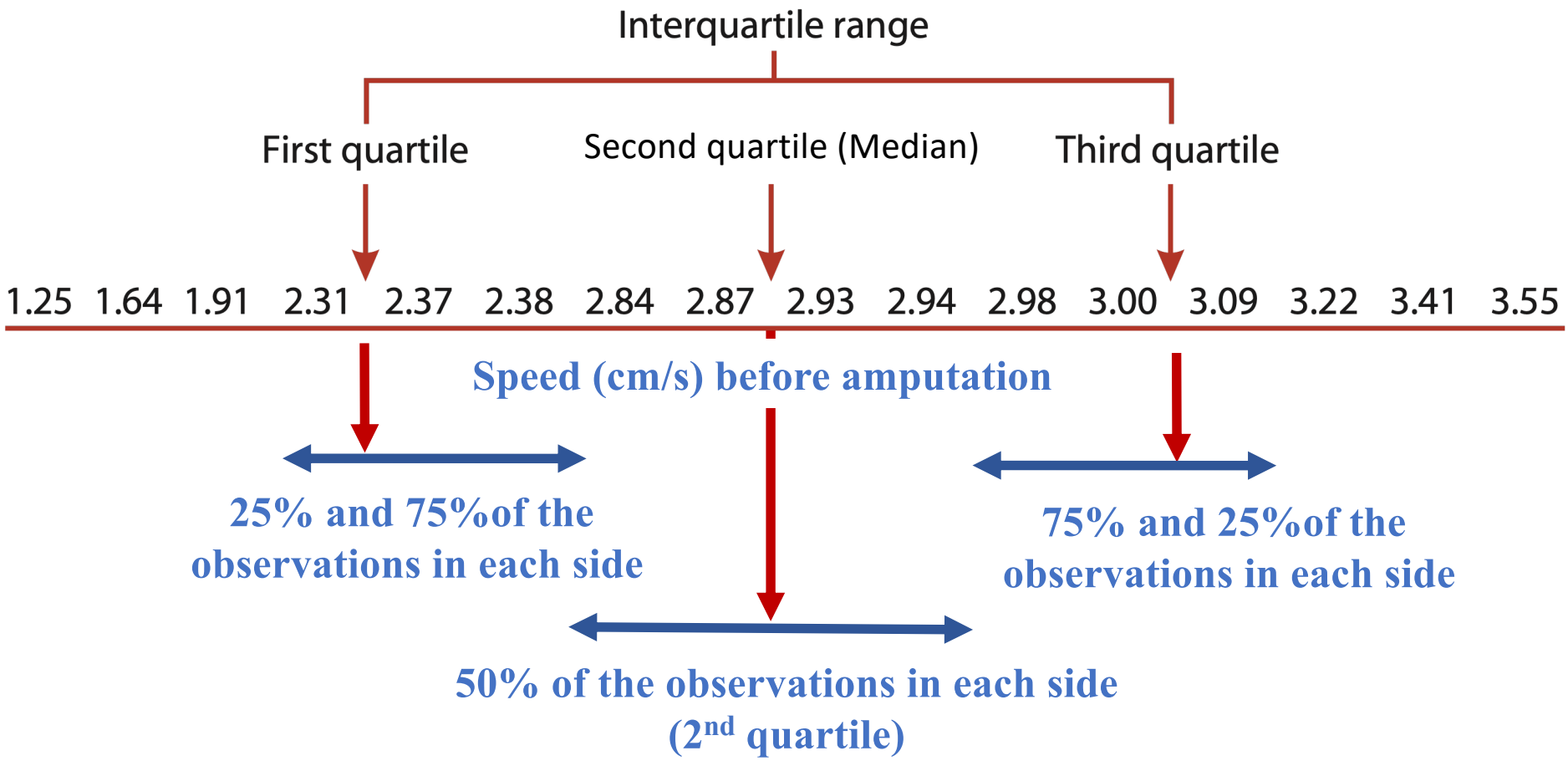


Median is the middle of the distribution;
it balances frequencies



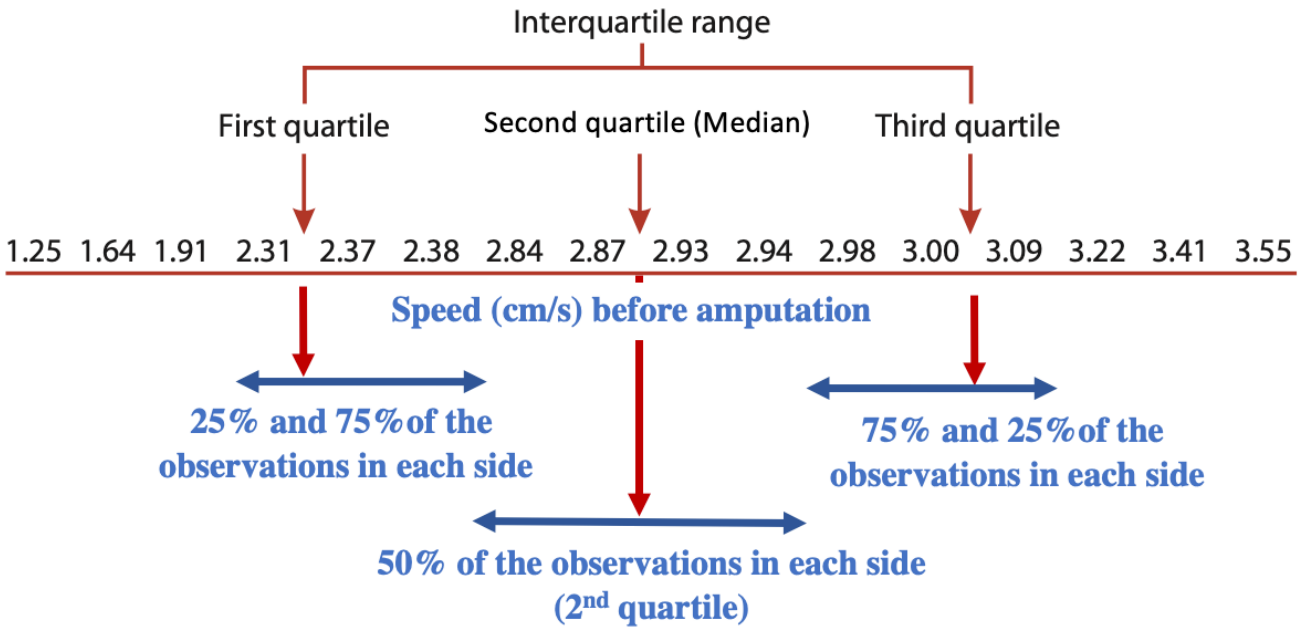
Mean is the "centre of gravity";
it balances sums

Interquartile range: the corresponding measure of spread for the median



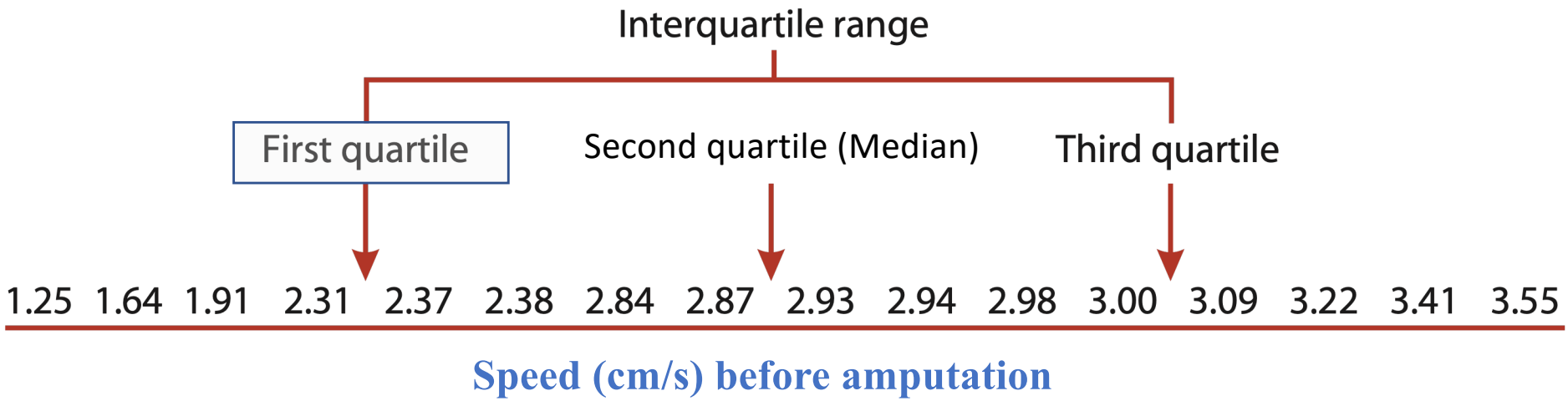
Remember: The corresponding metric of ***spread*** for the mean is the ***standard deviation***

Interquartile range: the corresponding measure of spread for the median



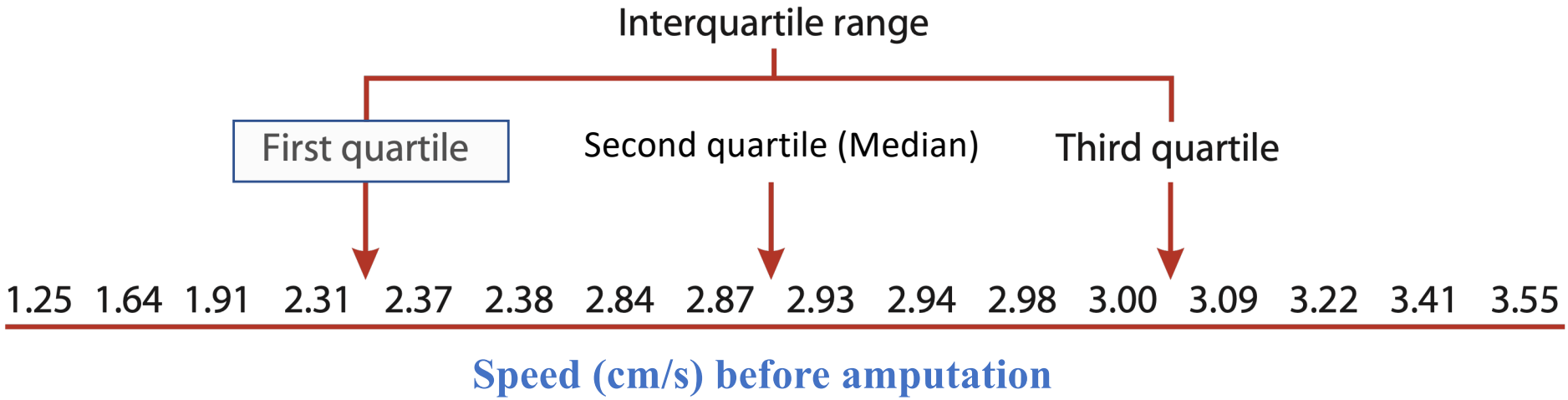
It's important to understand what the first, second (median), and third quartiles represent, and how the median differs from the mean. While I don't expect you to calculate these values by hand, working through some examples helps develop numeracy and a better intuition for what these summaries actually mean.

Interquartile range: the corresponding measure of spread for the median



While the calculation of the median (2nd quartile or Q2) differs depending on whether the number of observations is odd or even, the calculations for the 1st (Q1) and 3rd (Q3) quartiles remain the same.

Interquartile range: the corresponding measure of spread for the median – first quartile Q_1



$$\text{Positioning } Q_1: j = 0.25n = (0.25)(16) = 4$$

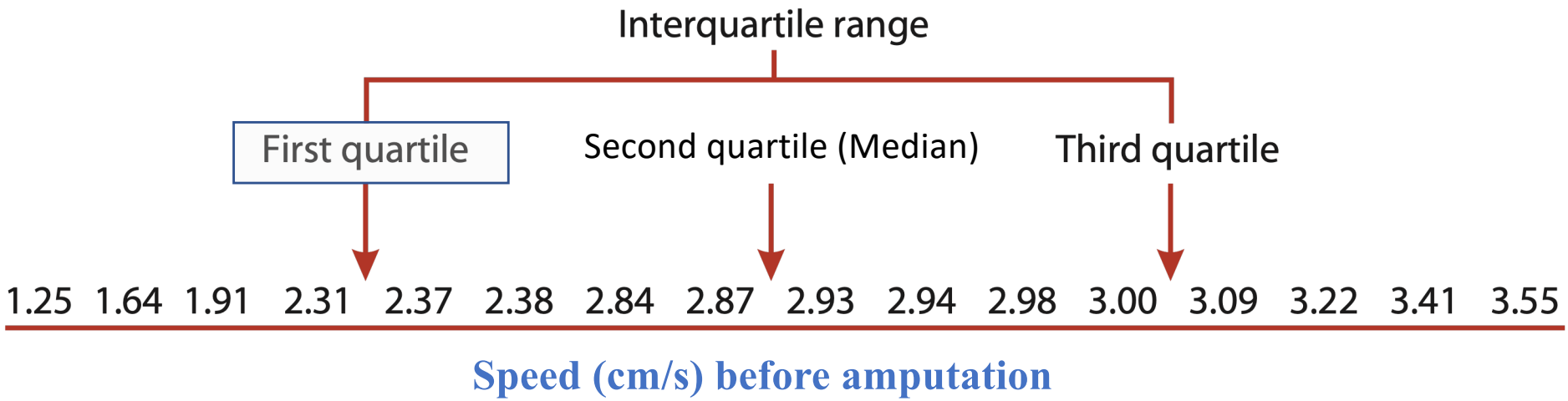
where *n* is the number of observations in the data (i.e., 16 male spiders).

Because here *j* is an integer (i.e., whole number, not a fraction), then the 1st quartile is the average of

$Y_{(j)}$ and $Y_{(j+1)}$ = $Y_{(4)}$ and $Y_{(4+1)}$ =(2.31 + 2.37) / 2 = 2.340 cm/s

First quartile (Q_1) = 2.340 cm/s

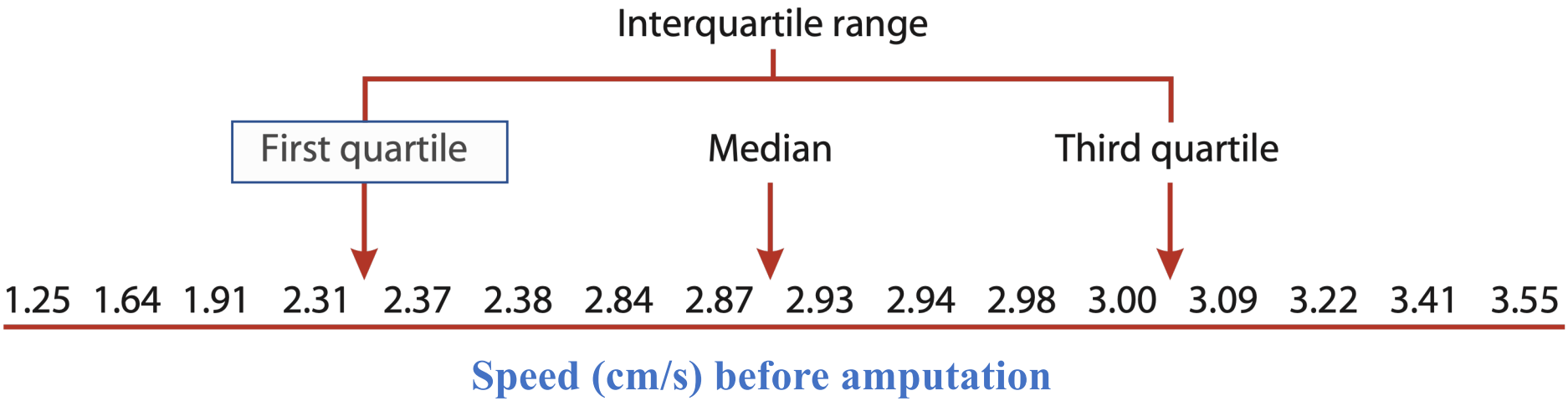
Interquartile range: the corresponding measure of spread for the median – first quartile Q_1



This is not exactly the default rule in R, but the values are very similar. There are several different rules for calculating quartiles.

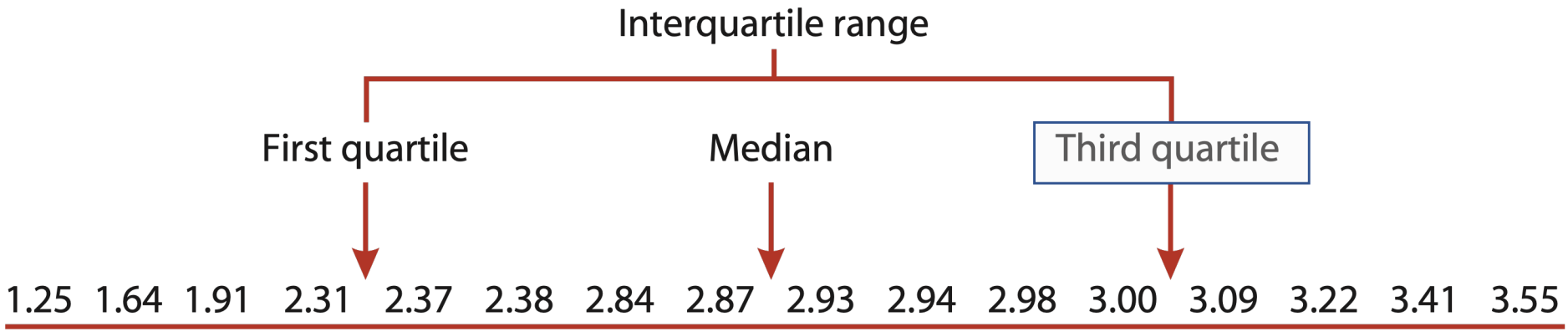
First quartile (Q_1) = 2.340 cm/s

Interquartile range: the corresponding measure of spread for the median



If *j* was not an integer, round *j* (e.g., say *j* was 4.32 then round *j* = 4). We would then have picked the 4th value in the ranked distribution (i.e., 2.31 cm/s)

Interquartile range: the corresponding measure of spread for the median

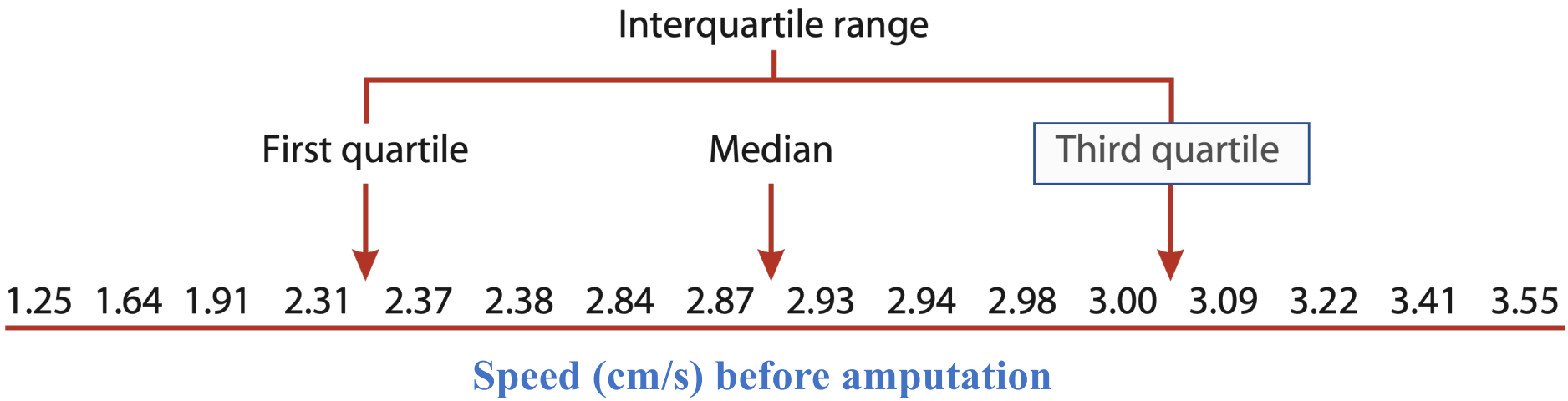


$$\text{Positioning } Q_3\text{: } j = 0.75n = (0.75)(16) = 12$$

where n is the number of observations. If j is an integer (*whole number, not a fraction*), then the 3rd quartile is the average of $Y_{(j)}$ and $Y_{(j+1)}$ = $Y_{(12)}$ and $Y_{(12+1)}$ =(3.00 + 3.09) / 2 = 3.045 cm/s

Third quartile (Q_3) = 3.045 cm/s

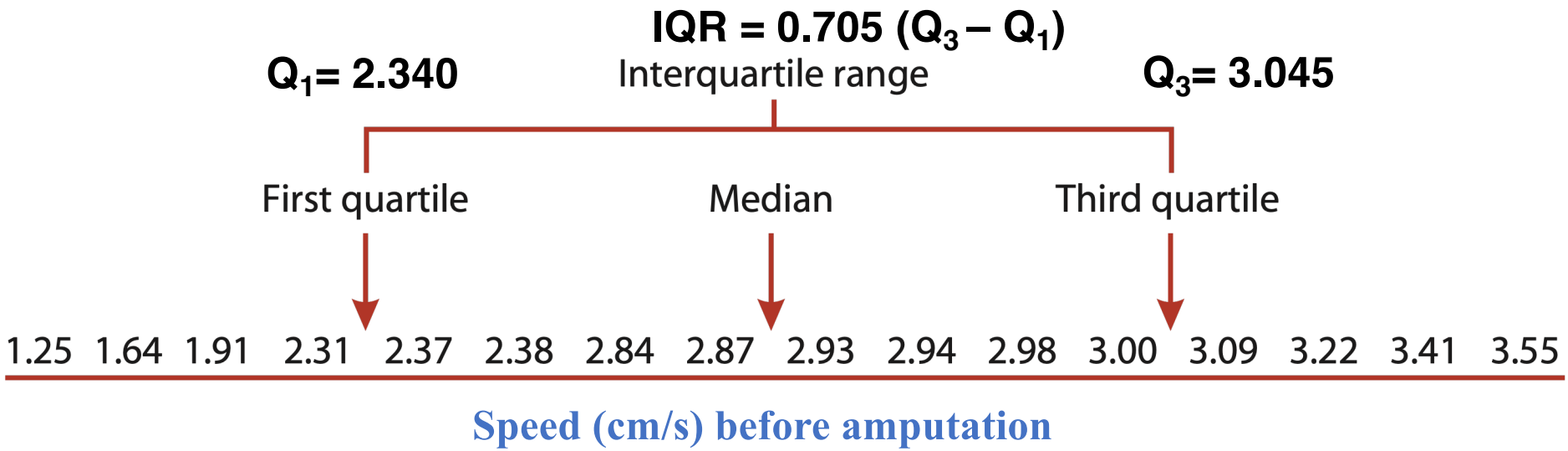
Interquartile range: the corresponding measure of spread for the median



$$\text{Positioning } Q_3\text{: } j = 0.75n = (0.75)(16) = 12$$

If *j* was not an integer (here 12), round *j* (e.g., say j was 12.72 then *j* = 13). We would then have picked the 13th value in the ranked distribution (i.e., 3.09 cm/s). The same applies for *j* in Q_1.

Interquartile range: the corresponding measure of spread for the median



The ***interquartile range*** (IQR) for the speed data before amputation is then $Q_3 - Q_1 = 3.045 - 2.340 = 0.705$ cm/s

Remember: the mean reflects all values in a distribution but is influenced by extreme values. The median, while not as representative of the entire distribution, is resistant to the influence of extreme values.

Mean is the "centre of gravity";
it balances sums

Median is the middle of the distribution;
it balances frequencies

$$Y = 53, 58, 62, 64, 68, 72, 73, 77, 86, 87, 88, 92$$

$$\bar{Y} = 73.3$$

$$Q_2 = 72.5$$

$$Y = 53, 58, 62, 64, 68, 72, 73, 77, 86, 87, 88, 192$$

$$\bar{Y} = 81.7$$

$$Q_2 = 72.5$$

Let’s take a “power break” – 1 minute

The median is the middle measure of a set of observations (distribution)

If the number of observations (*n*) is ***even***, then the median is calculated differently:

Tidarren (spider)

It gives an "arm" (or a pedipalp) for a female spider.

Running speed (cm/s) of male *Tidarren* spiders before and after voluntary amputation of one pedipalp.

Oxyopes salticus

Spider	Speed before	Speed after	Spider	Speed before	Speed after
1	1.25	2.40	9	2.98	3.70
2	2.94	3.50	10	3.55	4.70
3	2.38	4.49	11	2.84	4.94
4	3.09	3.17	12	1.64	5.06
5	3.41	5.26	13	3.22	3.22
6	3.00	3.22	14	2.87	3.52
7	2.31	2.32	15	2.37	5.45
8	2.93	3.31	16	1.91	3.40

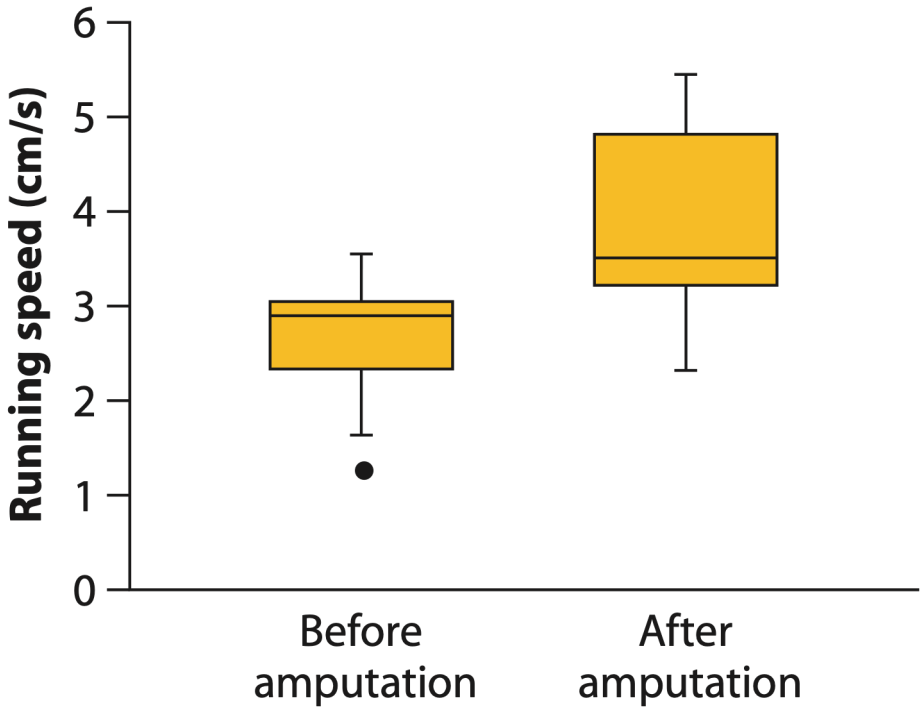
Representing data distributions by their quartiles: Boxplot (box-and-whisker plot)

What are the advantages of a box plot?

A boxplot quickly shows where most values lie (location) and how spread out the data are.

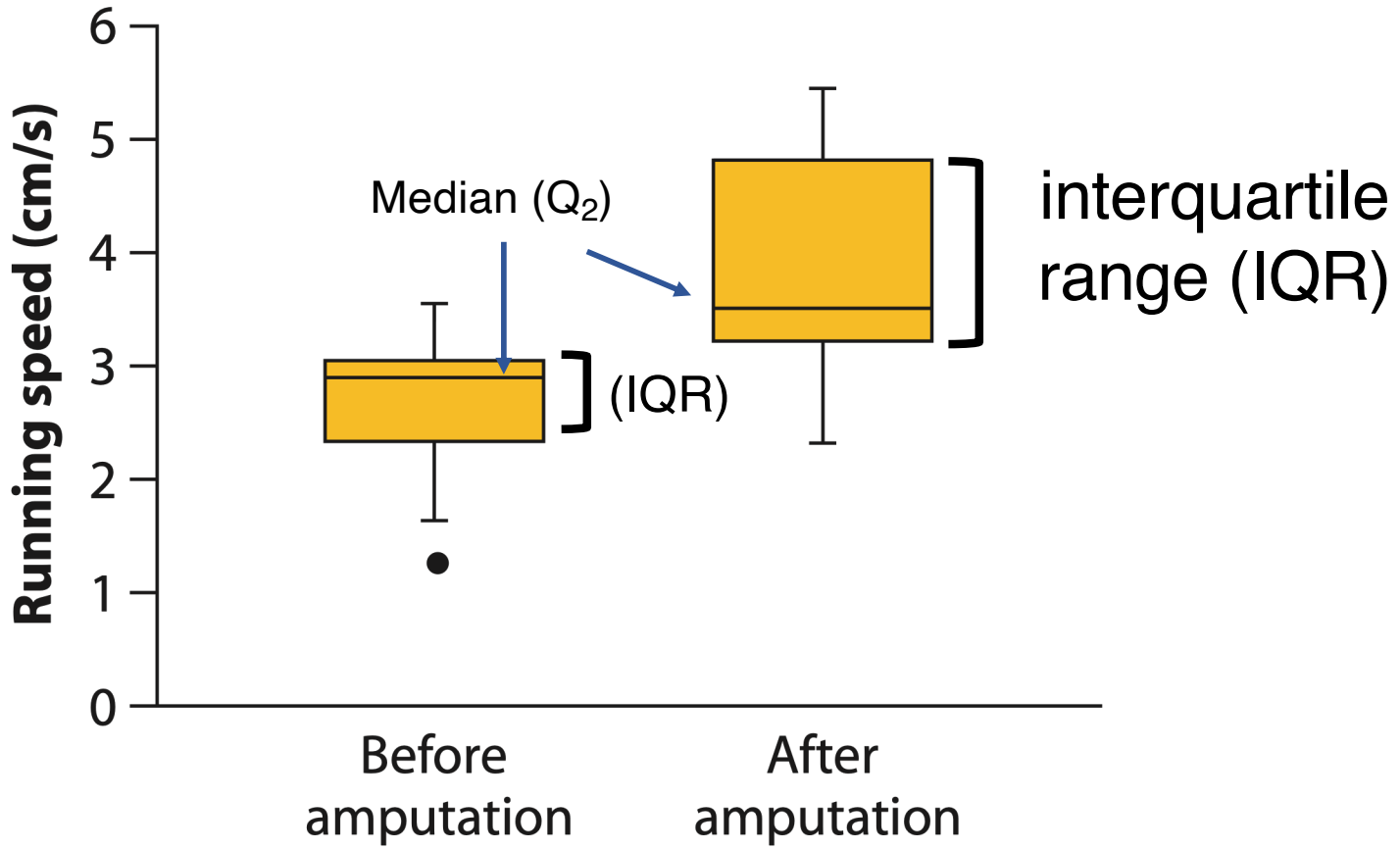
It helps reveal whether a distribution is symmetric or skewed.

It also stands out from many other graphs by clearly identifying potential outliers.



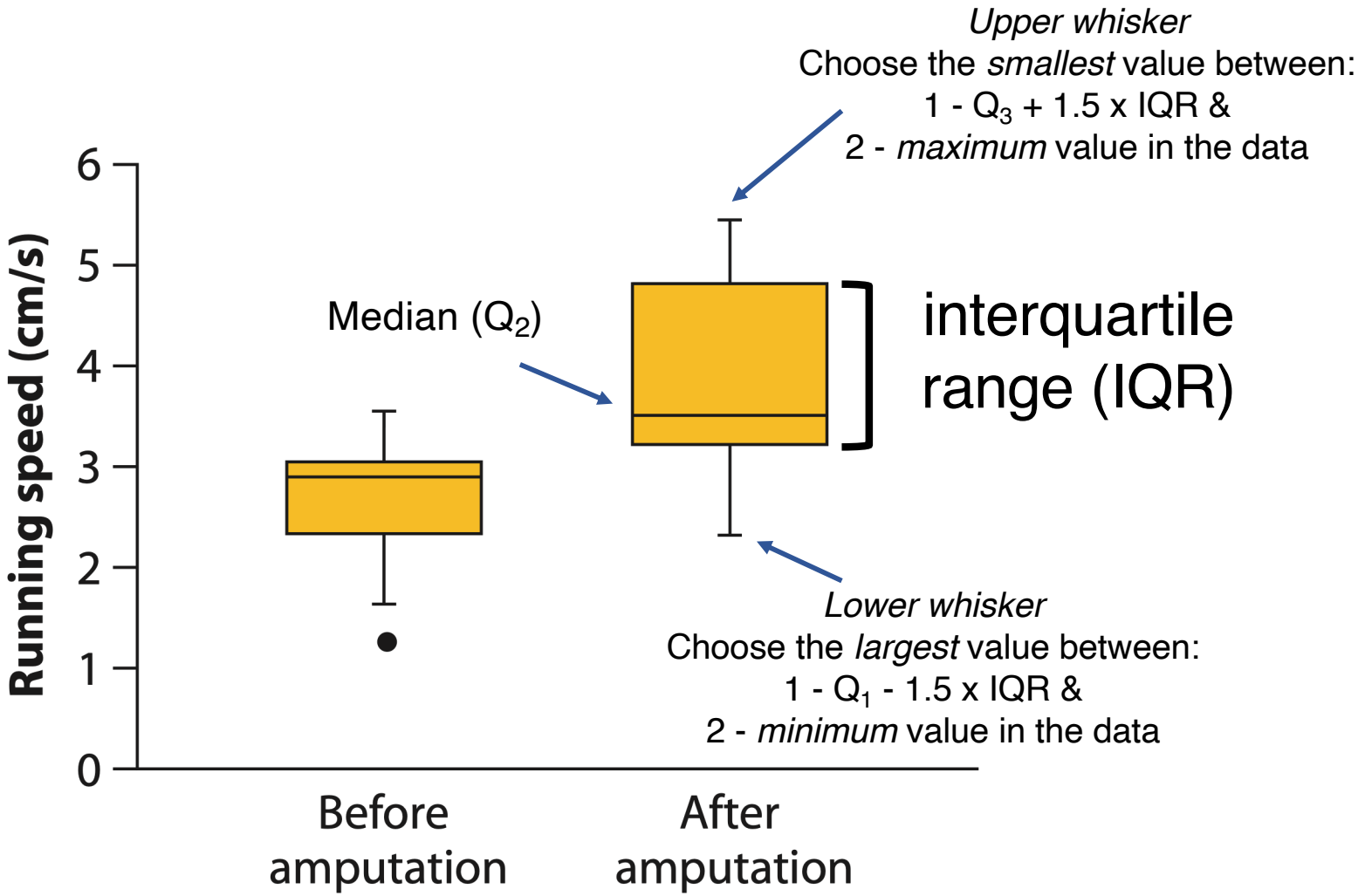
source - www.netmba.com/statistics/plot/box/

Representing data distributions by their quartiles: Boxplot (box-and-whisker plot)



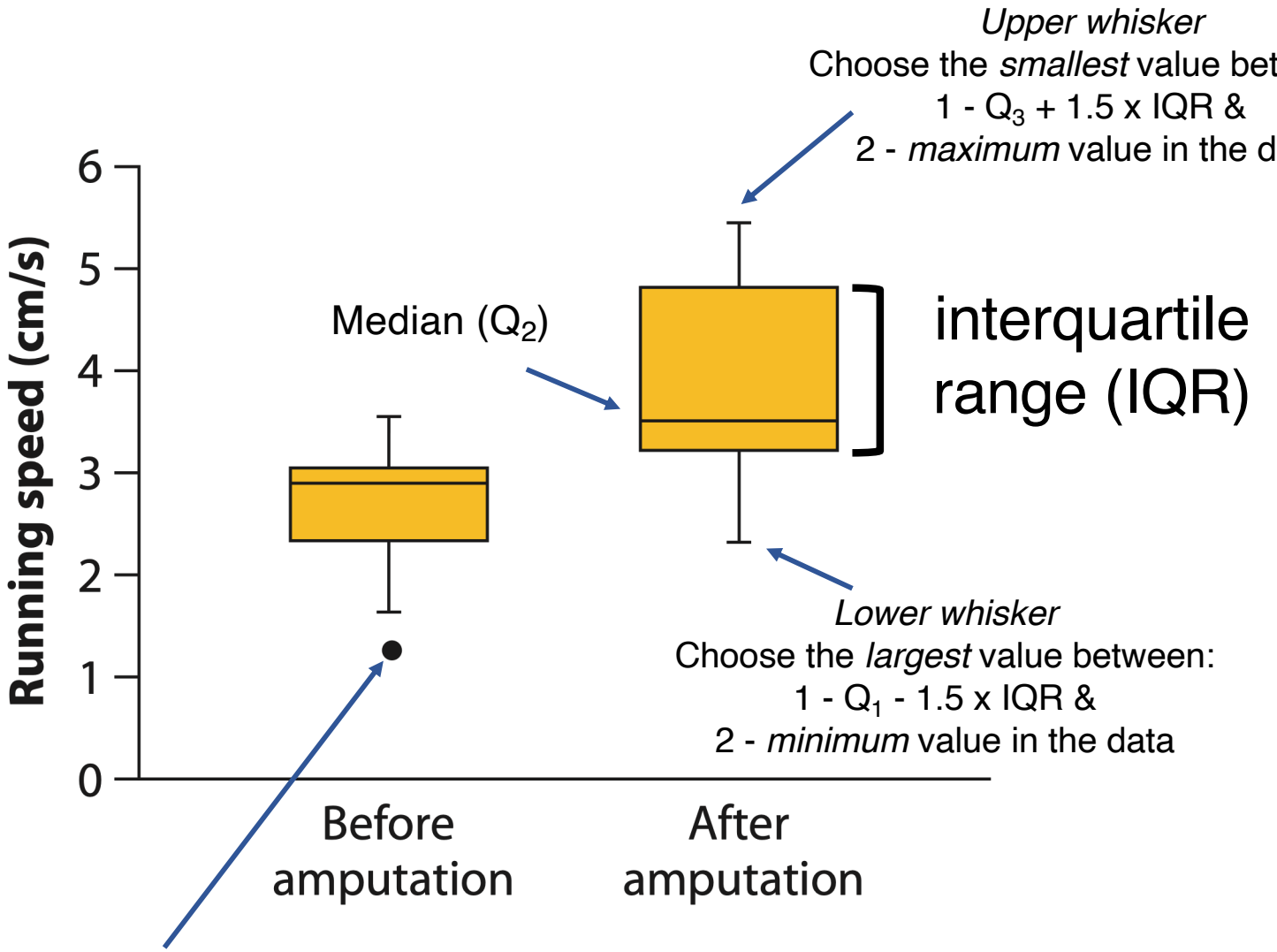
John Tukey
(box-and-whisker 1977)

Representing data distributions by their quartiles: Boxplot (box-and-whisker plot)



There are many ways to calculate whiskers; this is one common way and one of the methods in R

Representing data distributions by their quartiles: Boxplot (box-and-whisker plot)



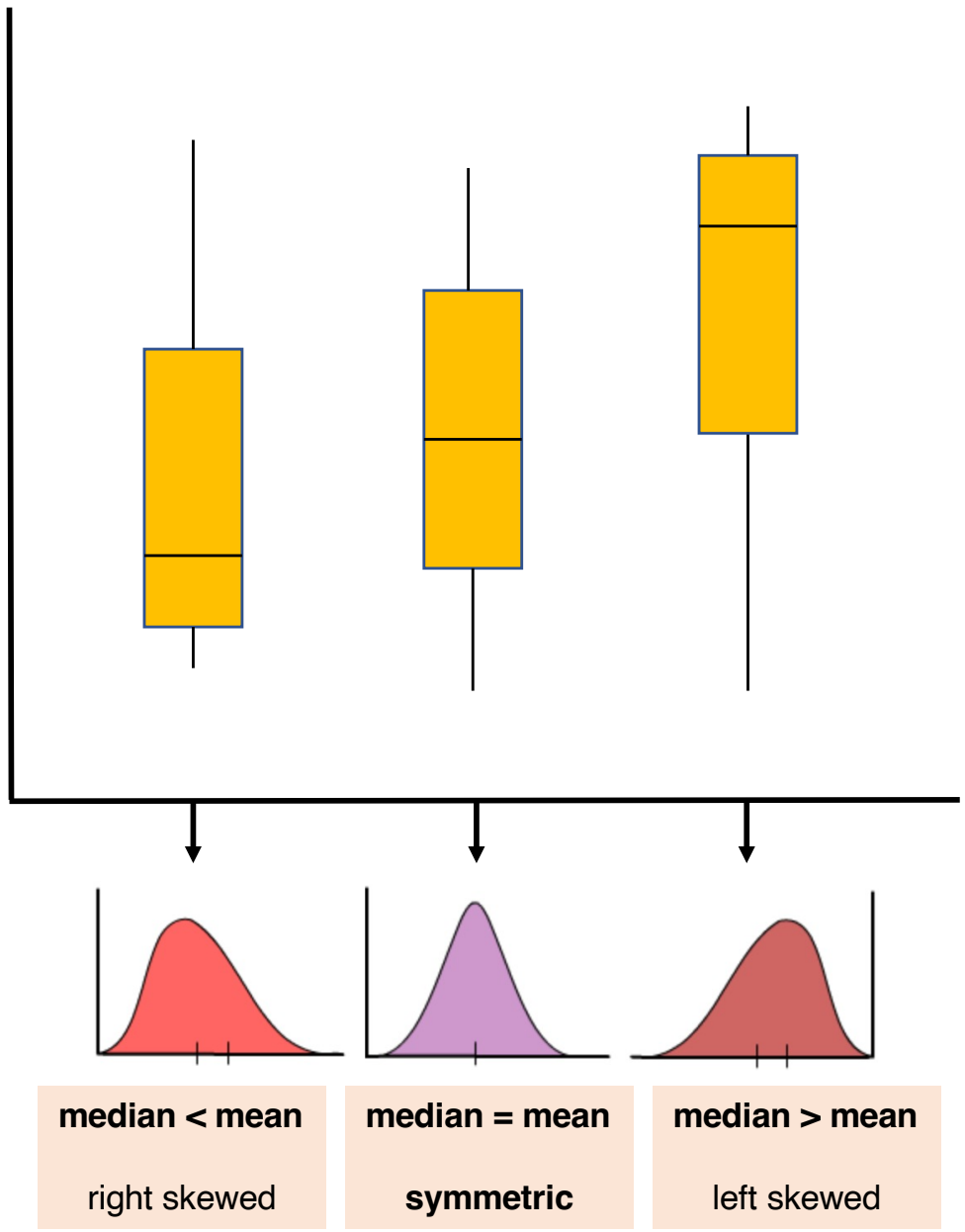
There are many ways to calculate whiskers; this is one common way and one of the methods in R

Representing data distributions by their quartiles: Boxplot (box-and-whisker plot)

What are the advantages of a box plot?

A boxplot quickly shows where most values lie (location) and how spread out the data are.

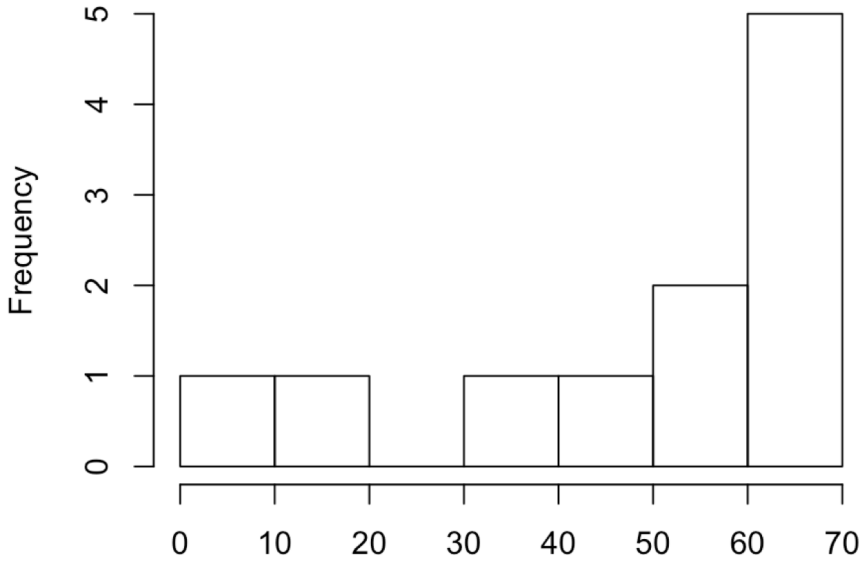
It can often provide an indication of the data's symmetry and skewness, though this is not always the case.

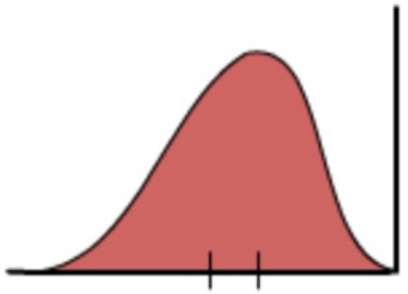


Representing data distributions by their quartiles: Boxplot (box-and-whisker plot)

Three fictional data sets to illustrate the calculation and properties of distributions using their boxplot.

Left-skewed distribution: 9,11,31,44,52,58,61,61,63,64,66

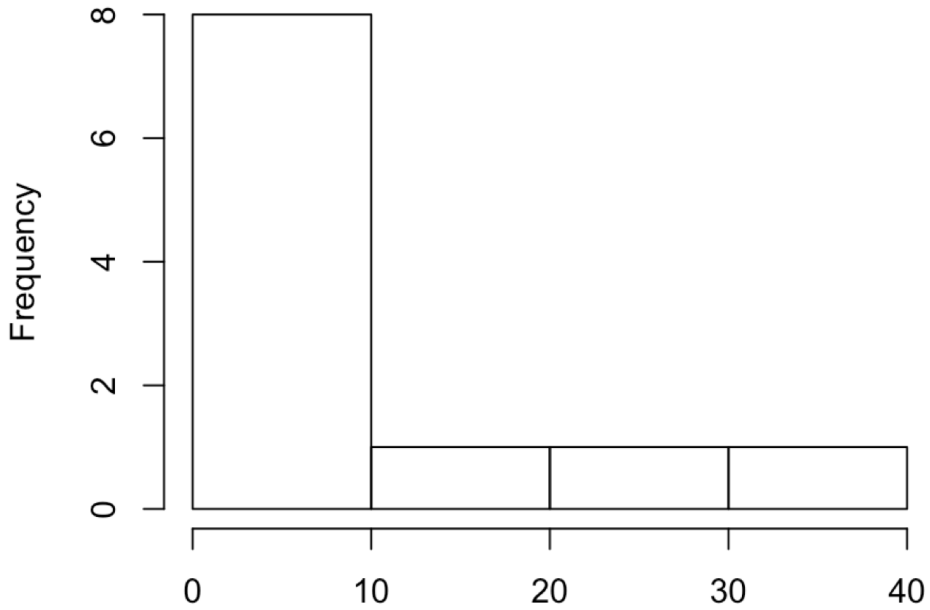


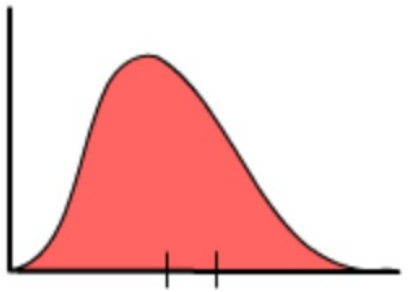


Representing data distributions by their quartiles: Boxplot (box-and-whisker plot)

Three fictional data sets to illustrate the calculation and properties of distributions using their boxplots.

Right-skewed distribution: 1,2,3,4,5,6,7,10,20,30,40

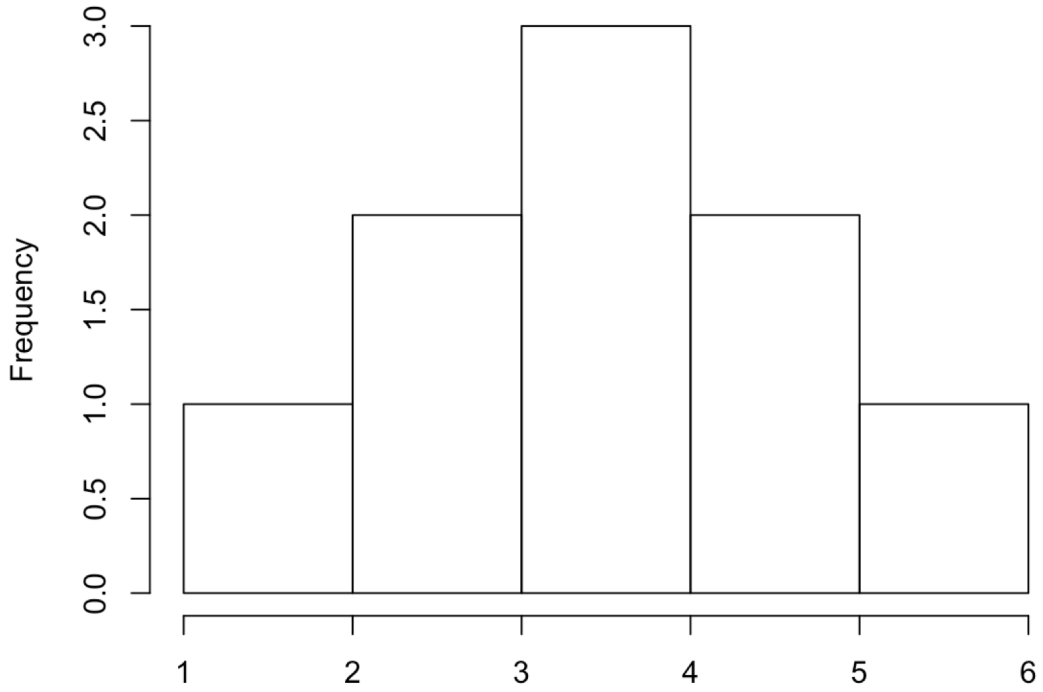




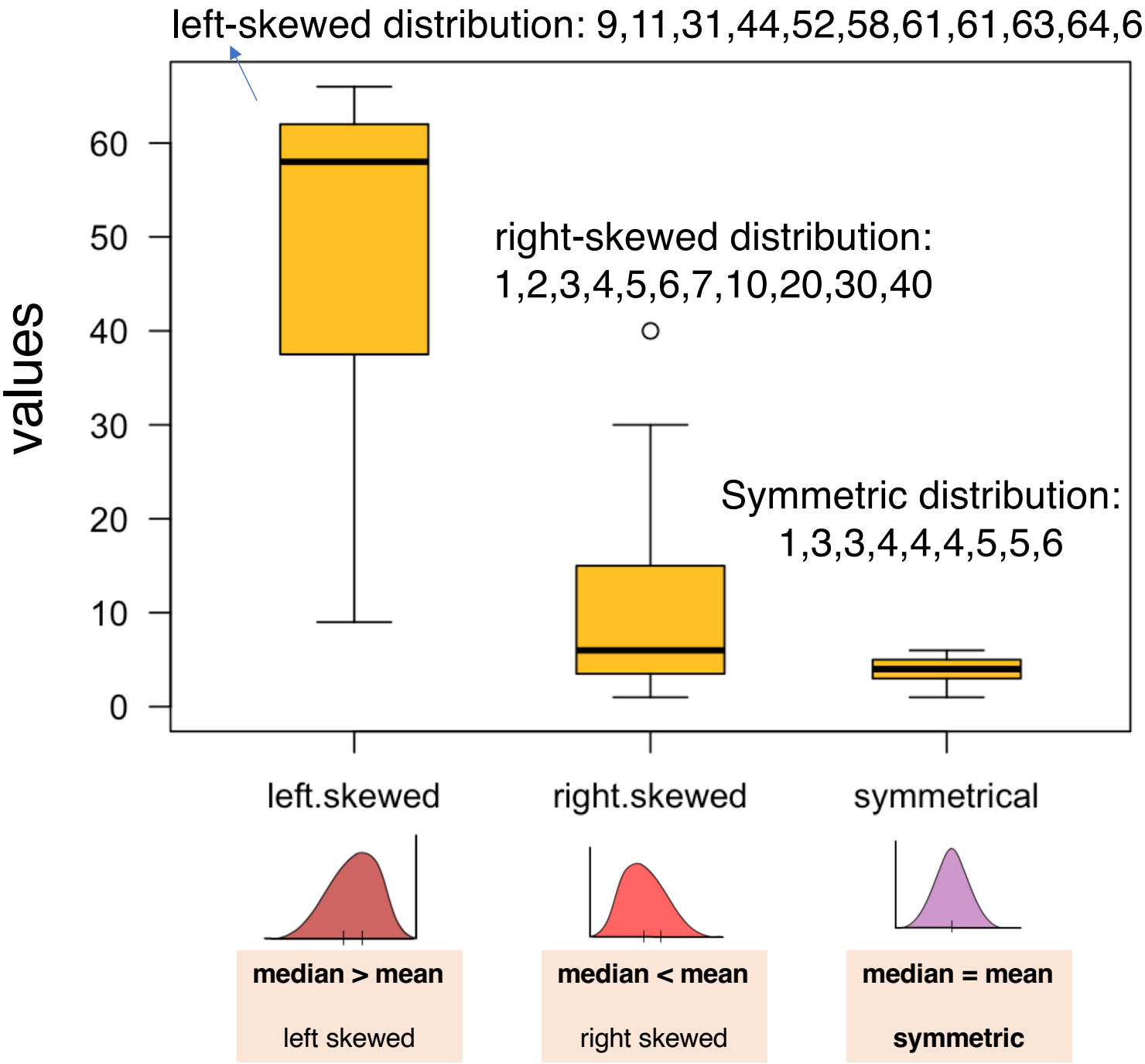
Representing data distributions by their quartiles: Boxplot (box-and-whisker plot)

Three fictional data sets to illustrate the calculation and properties of distributions using their boxplots.

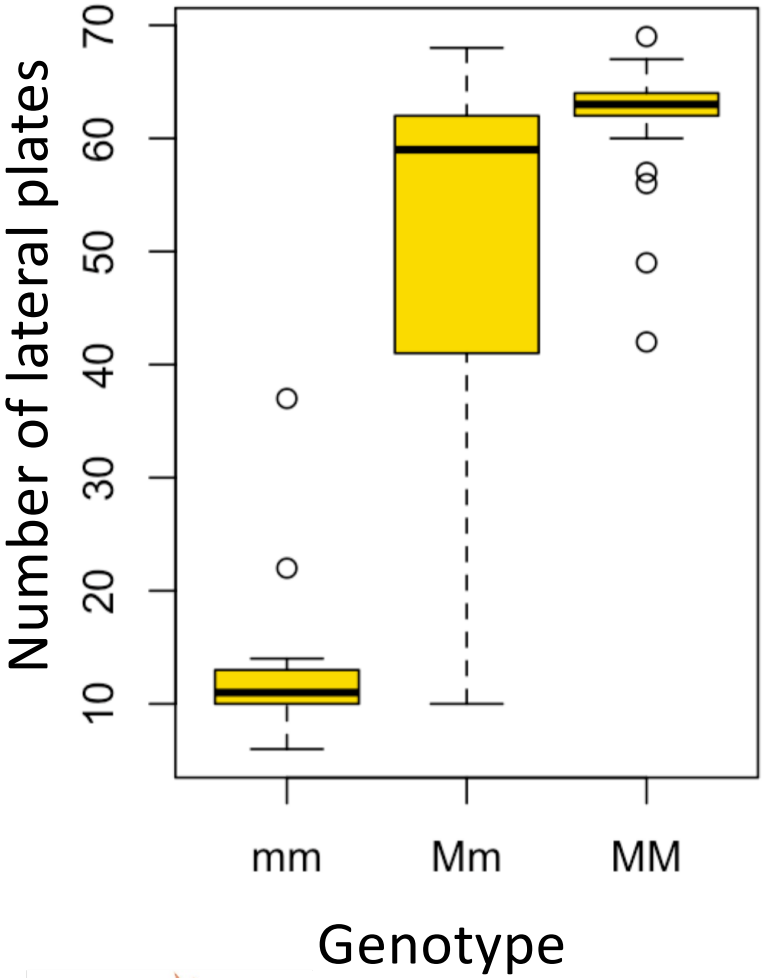
Symmetric distribution: 1,3,3,4,4,4,5,5,6

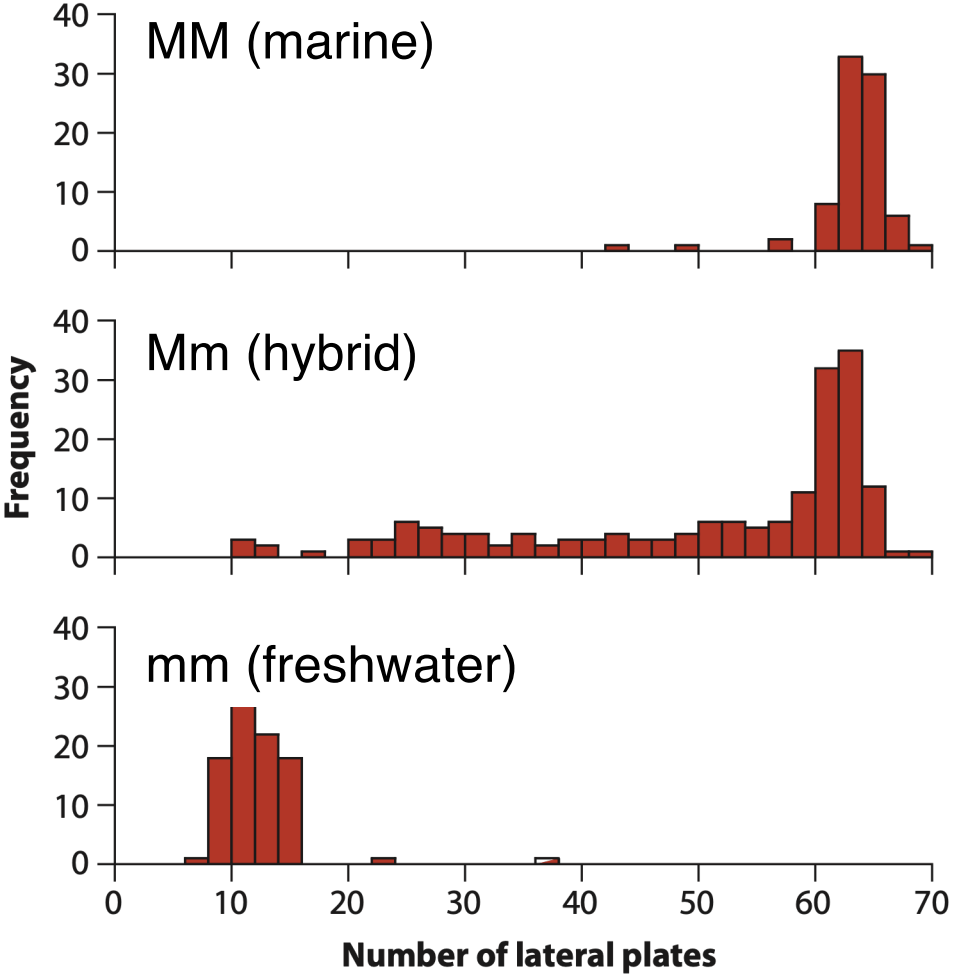


Boxplot (box-and-whisker plot): contrasting distributions



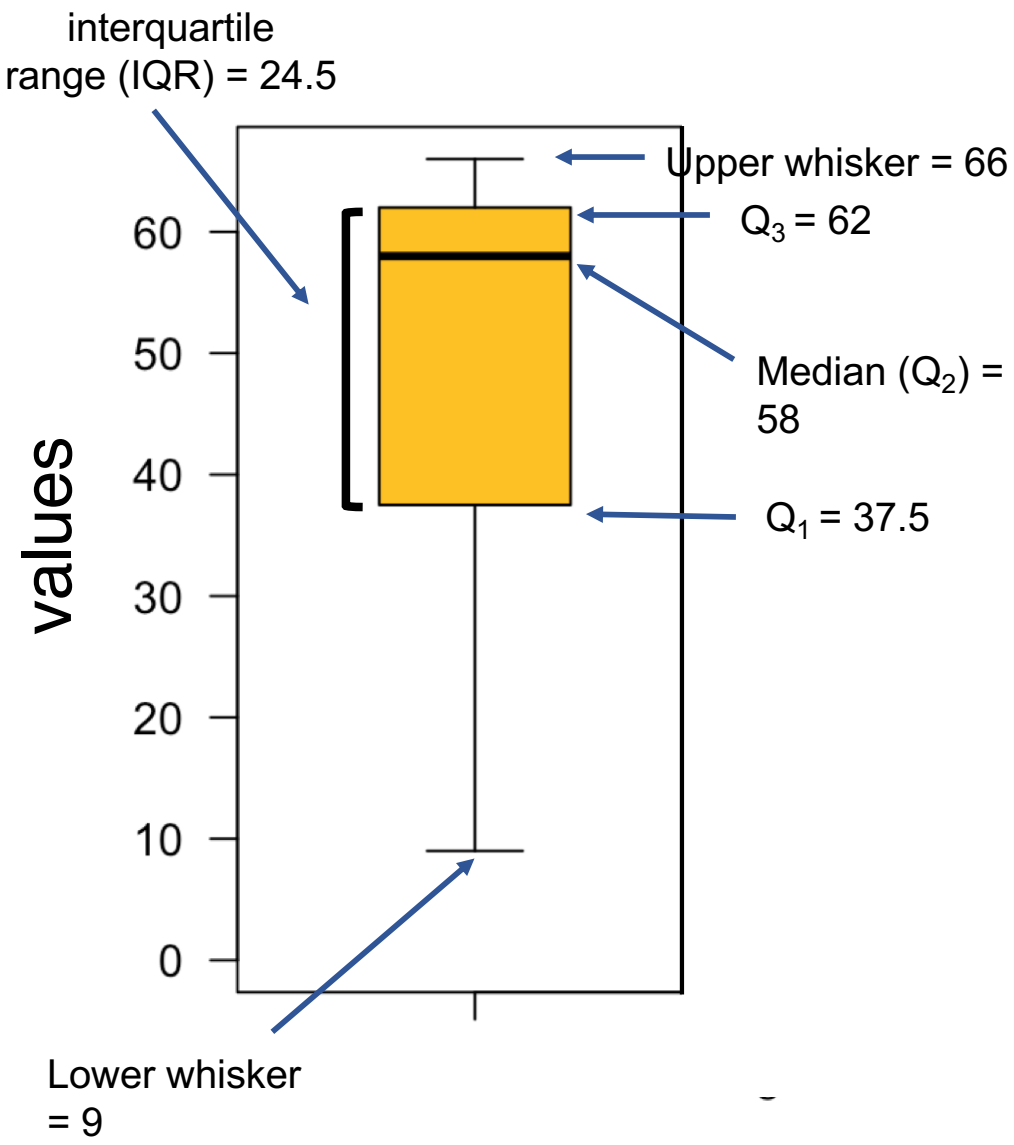
Boxplot (box-and-whisker plot): contrasting distributions





For the mathematically brave (optional): These calculations aren't required, but they answer the questions students usually have.

9,11,31,44,52,**58**,61,61,63,64,66



Q_1 = (31+44) / 2 = 37.5
Median (Q_2) = 58
Q_3 = (63+64) / 2 = 62.0

IQR (Interquartile range) = 62.0 - 37.5 = 24.5

Lower whisker = 9; calculation:
Choose the *largest* value between:
1 - Q_1 - 1.5 x IQR = 37.5 – 1.5 x 24.5 = 0.75
2 - *minimum* value in the data = **9**

Upper whisker = 66; calculation:
Choose the *smallest* value between:
1 - Q_3 + 1.5 x IQR = 62 + 1.5 x 24.5 = 98.8
2 - *maximum* value in the data =

Statistics is based on samples!

The main purpose of statistics is to learn about an entire population by studying a sample.

Statistics helps us make decisions when we don't have complete information, using samples to represent populations whose true features we don't fully know.

Since quantities calculated from samples (like the mean or standard deviation) change from sample to sample, uncertainty is always part of the process.

Next lecture - Estimating with uncertainty