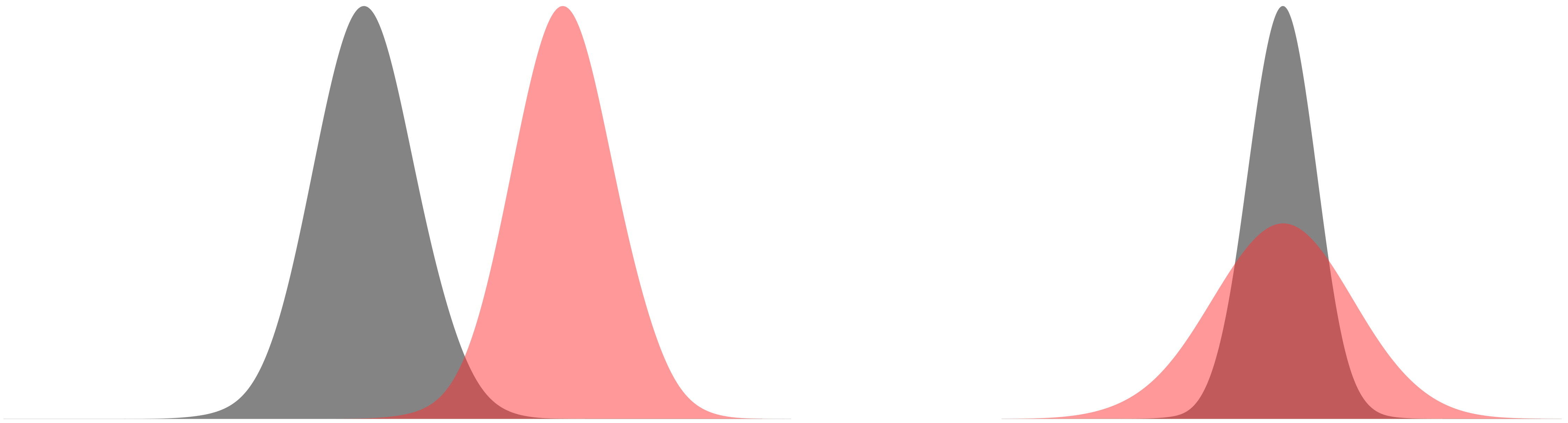
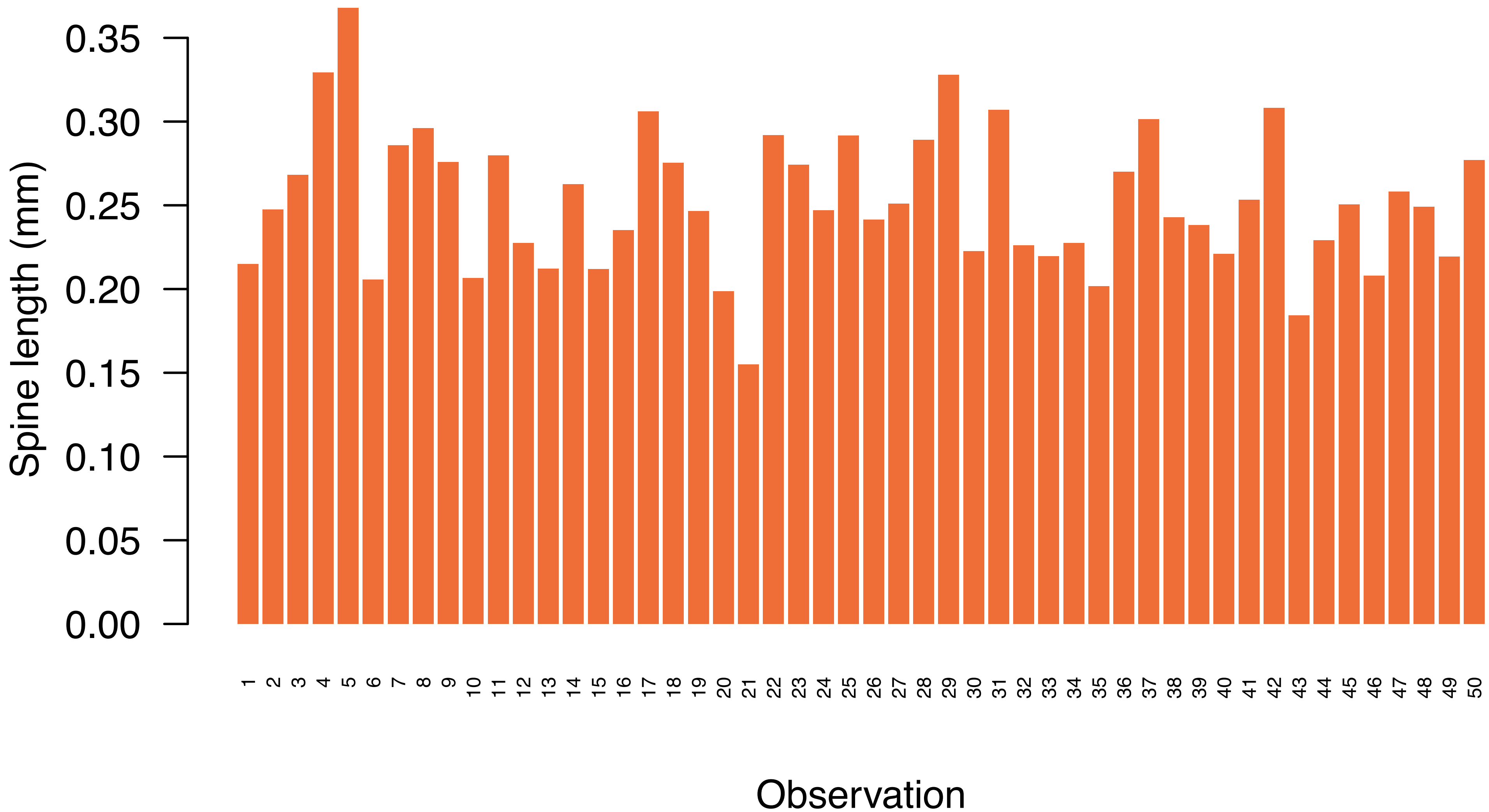
Summary Statistics

Owen Jones
jones@biology.sdu.dk



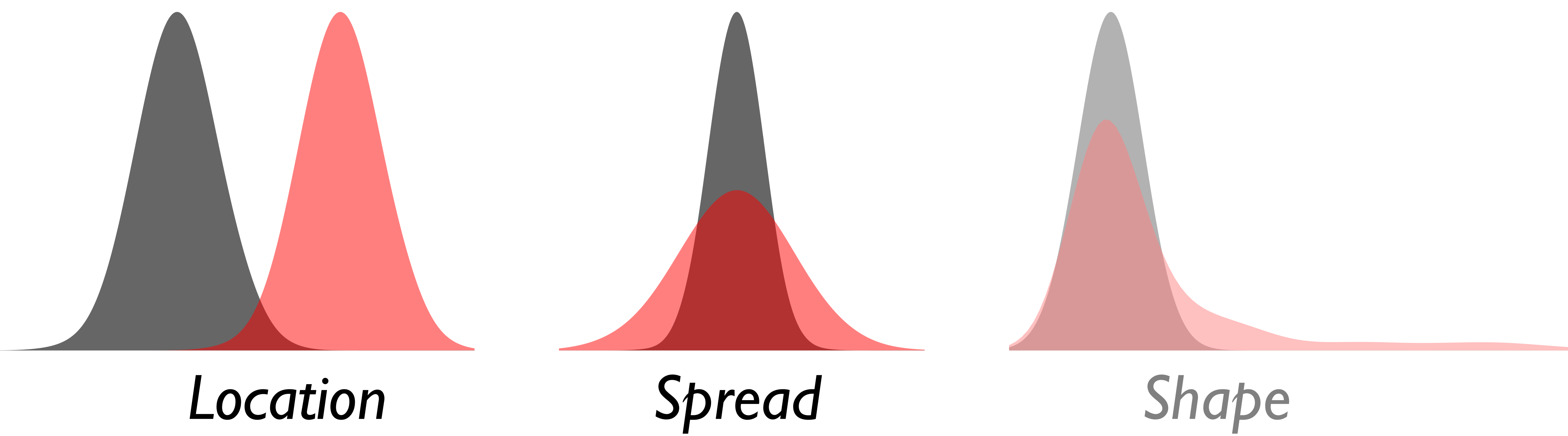
Summary statistics

- Doing science involves collecting data.

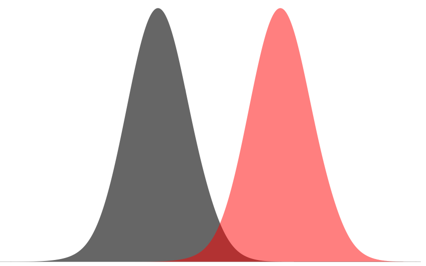


Summary statistics

- Doing science involves collecting data.
- Describing data requires ways to summarise without showing all the data.



Location



The harmonic mean is always the smallest, the arithmetic mean is always the largest. They are all equal if the values are all the same.

Location/central tendency

Average

- *Mean (arithmetic, geometric, harmonic)*
- *Median*
- *Mode*

Arithmetic mean

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

Data: 7, 9, 12, 10, 3, 8, 4

Sum: 7 + 9 + 12 + 10 + 3 + 8 + 4 = 53

n: 7

Arithmetic mean: 53/7 = 7.571

Arithmetic mean

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

Arithmetic mean of sample is an ***unbiased estimator*** *of population mean if:*

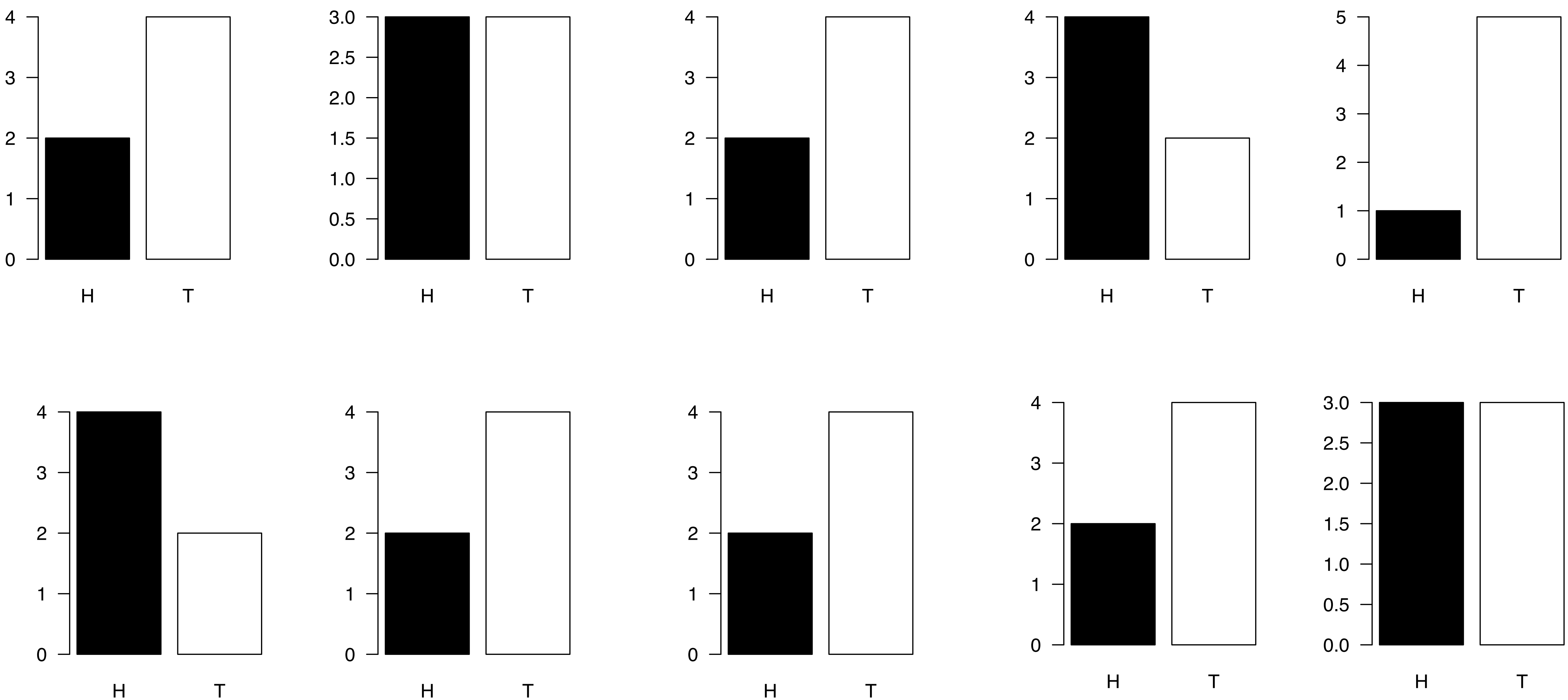
1. Observations are made on randomly selected individuals.

2. Observations are independent of eachother.

3. Observations are drawn from a large population that can be described by a normal random variable.

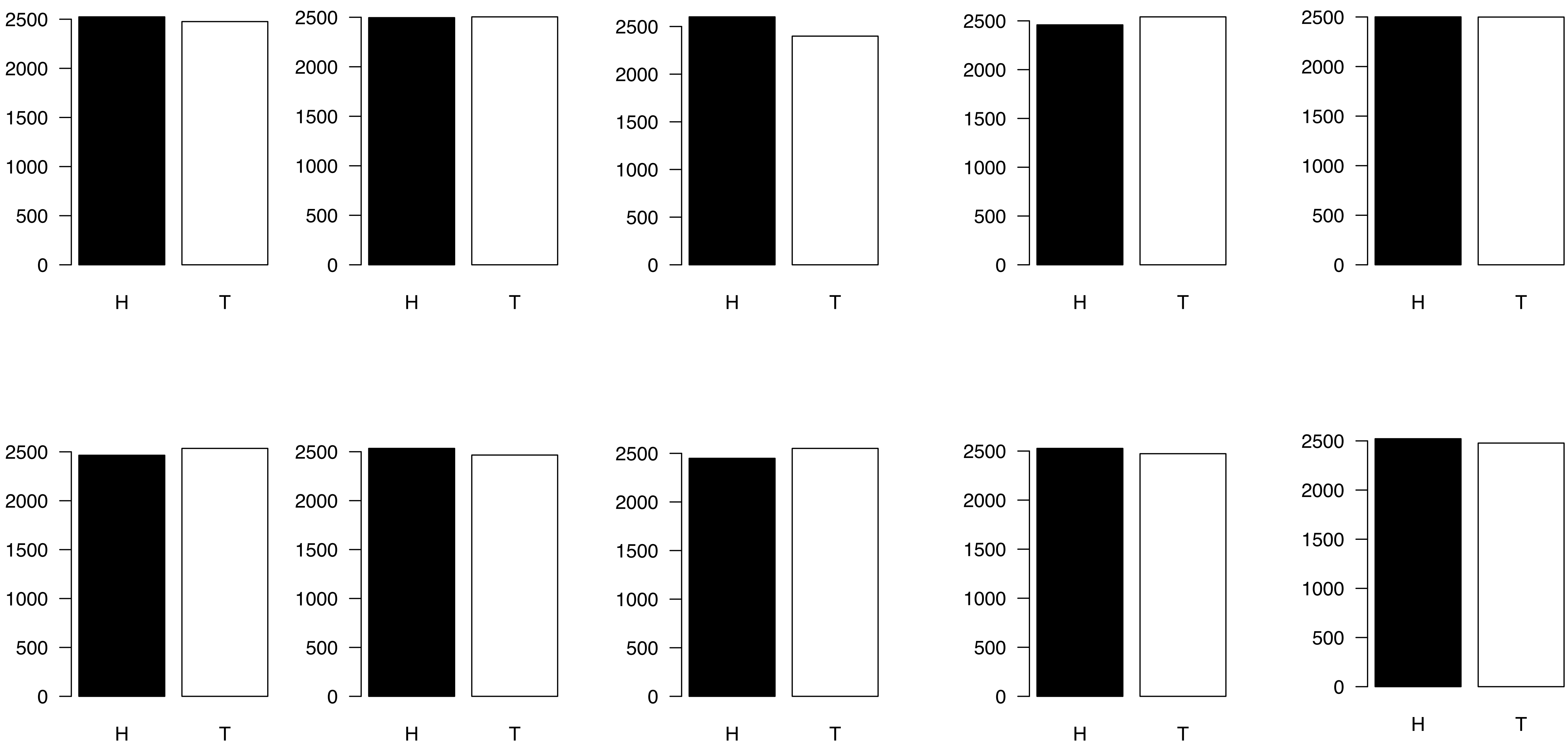
Law of large numbers

"As sample grows large, the sample mean converges to the population mean."



Law of large numbers

"As sample grows large, the sample mean converges to the population mean."



This is why large sample sizes are good!

Geometric and harmonic?

Geometric mean

$$\bar{Z} = \frac{\sum_{i=1}^{n} Z_i}{n}$$

where $Z = ln(Y)$

thus $Y = e^Z$

and, to back-transform to the units of Y:

$$GM_Y = e^{\left[\frac{\sum_{i=1}^{n} Z_i}{n}\right]}$$ or... $$GM_Y = e^{\left[\frac{\sum_{i=1}^{n} ln(Y_i)}{n}\right]}$$

```
exp(mean(log(x)))
```

Geometric and harmonic?

Geometric mean

$$\bar{Z} = \frac{\sum_{i=1}^{n} Z_i}{n}$$

where $Z = ln(Y)$

thus $Y = e^Z$

and, to back-transform
to the units of Y:

$$GM_Y = e^{[\frac{\sum_{i=1}^{n} Z_i}{n}]}$$

or...

$$GM_Y = e^{[\frac{\sum_{i=1}^{n} ln(Y_i)}{n}]}$$

```
exp(mean(log(x)))
```

Harmonic mean

$$H_Y = \frac{1}{[\frac{\sum_{i=1}^{n} \frac{1}{Y_i}}{n}]}$$

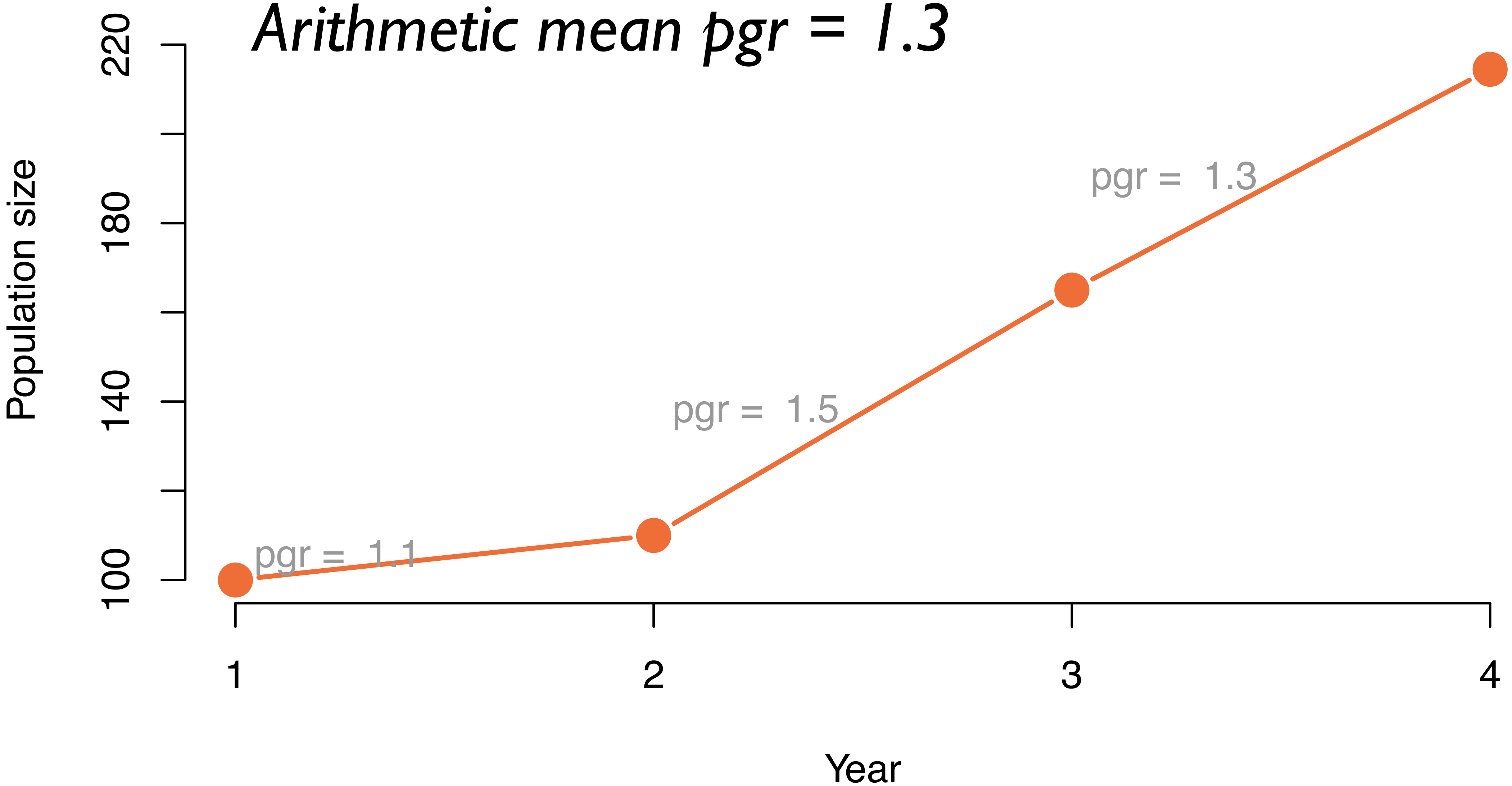
```
1/(mean(1/x))
```

Geometric mean

Population growth: a multiplicative process

What average population growth rate would lead to the same population size over the 3 years?

Arithmetic mean pgr = 1.3

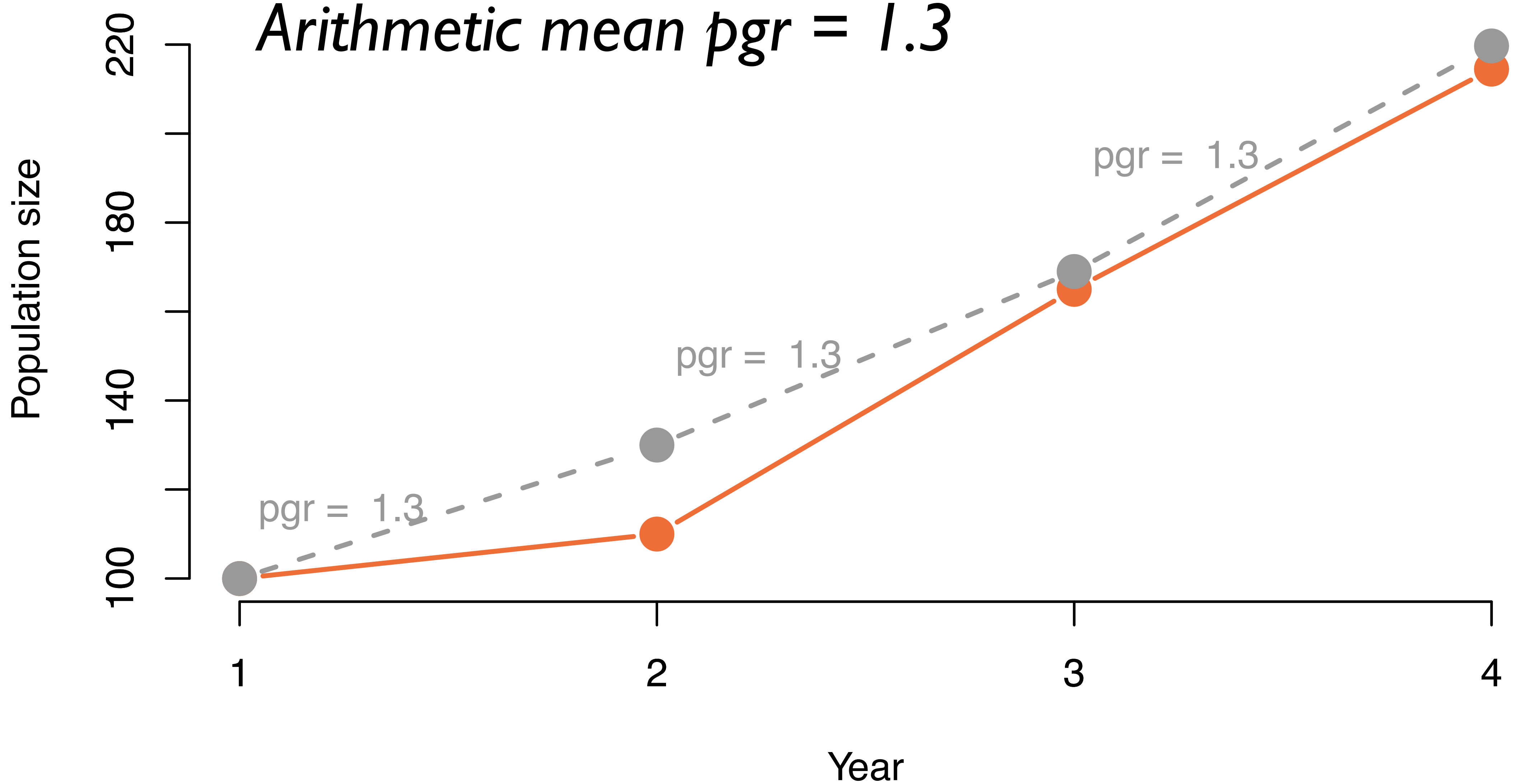


Geometric mean

Population growth: a multiplicative process

What average population growth rate would lead to the same population size over the 3 years?

Arithmetic mean pgr = 1.3

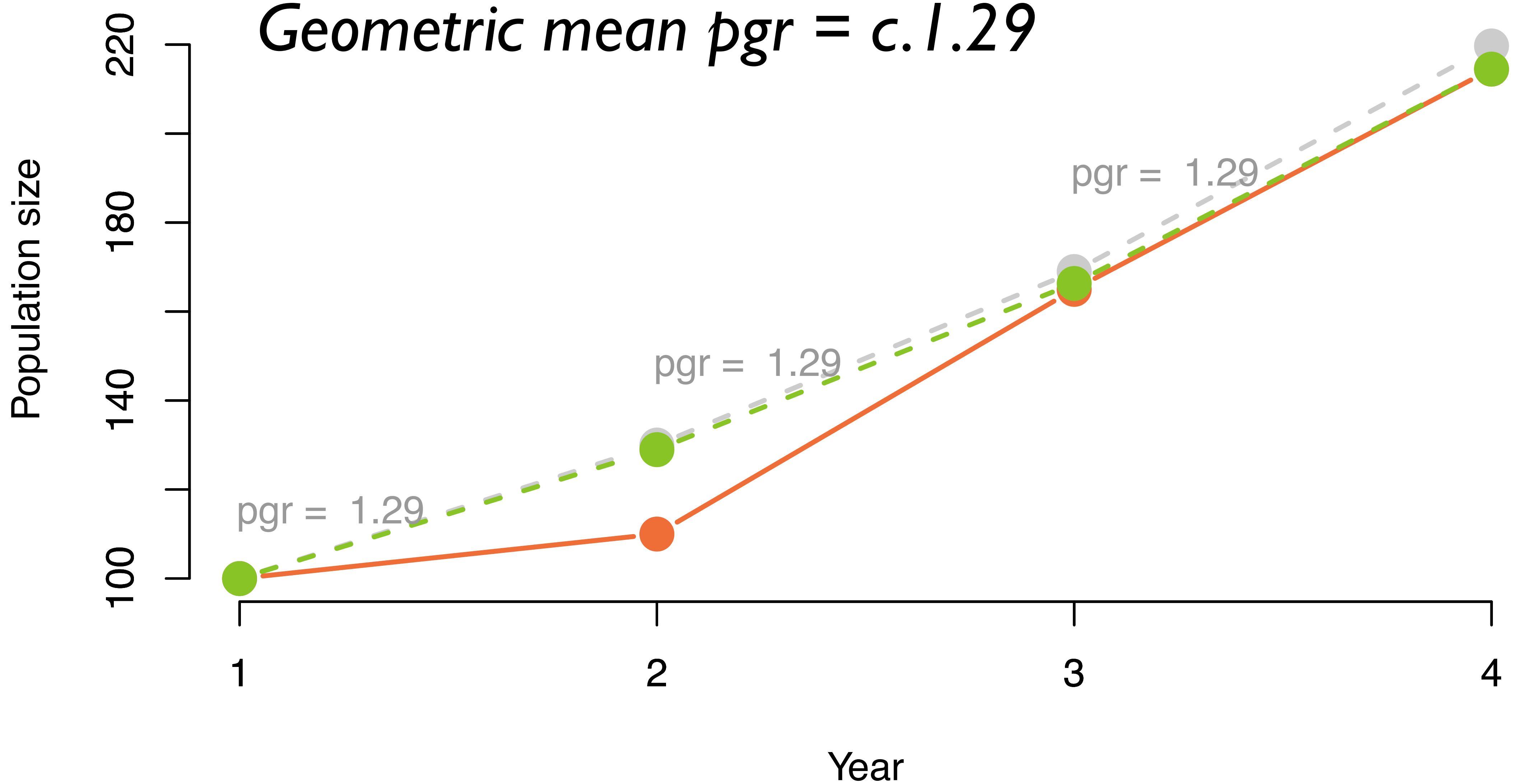


Geometric mean

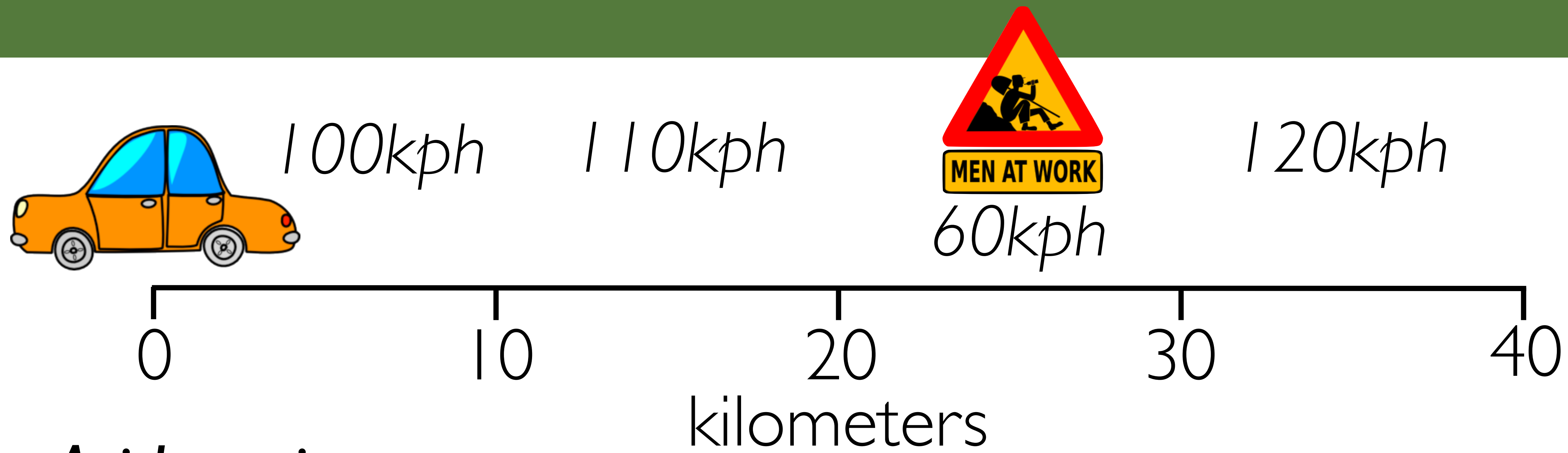
Population growth: a multiplicative process

What average population growth rate would lead to the same population size over the 3 years?

Geometric mean pgr = c.1.29



Harmonic mean



Arithmetic mean

(100+110+60+120)/4 = 97.5 kph

Harmonic mean

100kph 110kph

60kph

120kph

0 10 20 30 40

kilometers

Arithmetic mean

(100+110+60+120)/4 = 97.5 kph

Distance and time

	Distance	Speed	Time
	10	100	0.1
	10	110	0.091
	10	60	0.166
	10	120	0.083
Total	40	-	0.441

Speed = Distance/Time

40/0.441 =90.72

Harmonic mean

100kph 110kph

60kph

120kph

0 10 20 30 40

kilometers

Arithmetic mean

(100+110+60+120)/4 = 97.5 kph

Harmonic mean

$$H_Y = \frac{1}{\left[\frac{\sum_{i=1}^{n} \frac{1}{Y_i}}{n}\right]} \qquad \frac{1}{\left(\frac{\frac{1}{100}+\frac{1}{110}+\frac{1}{60}+\frac{1}{120}}{4}\right)} = 90.72$$

Geometric and harmonic?

Use arithmetic mean:
when the situation is additive.
"if all the quantities had the same value, what would that value have to be in order to achieve the same total?" e.g. weight, length, volume.

Use geometric mean:
when the situation is multiplicative.
"if all the quantities had the same value, what would that value have to be in order to achieve the same product?" e.g. growth rates, investment returns.

Use harmonic mean:
when members are defined in relation to a fixed unit. e.g. fractions, rates.

Median

Median: the middle number in a sorted vector.

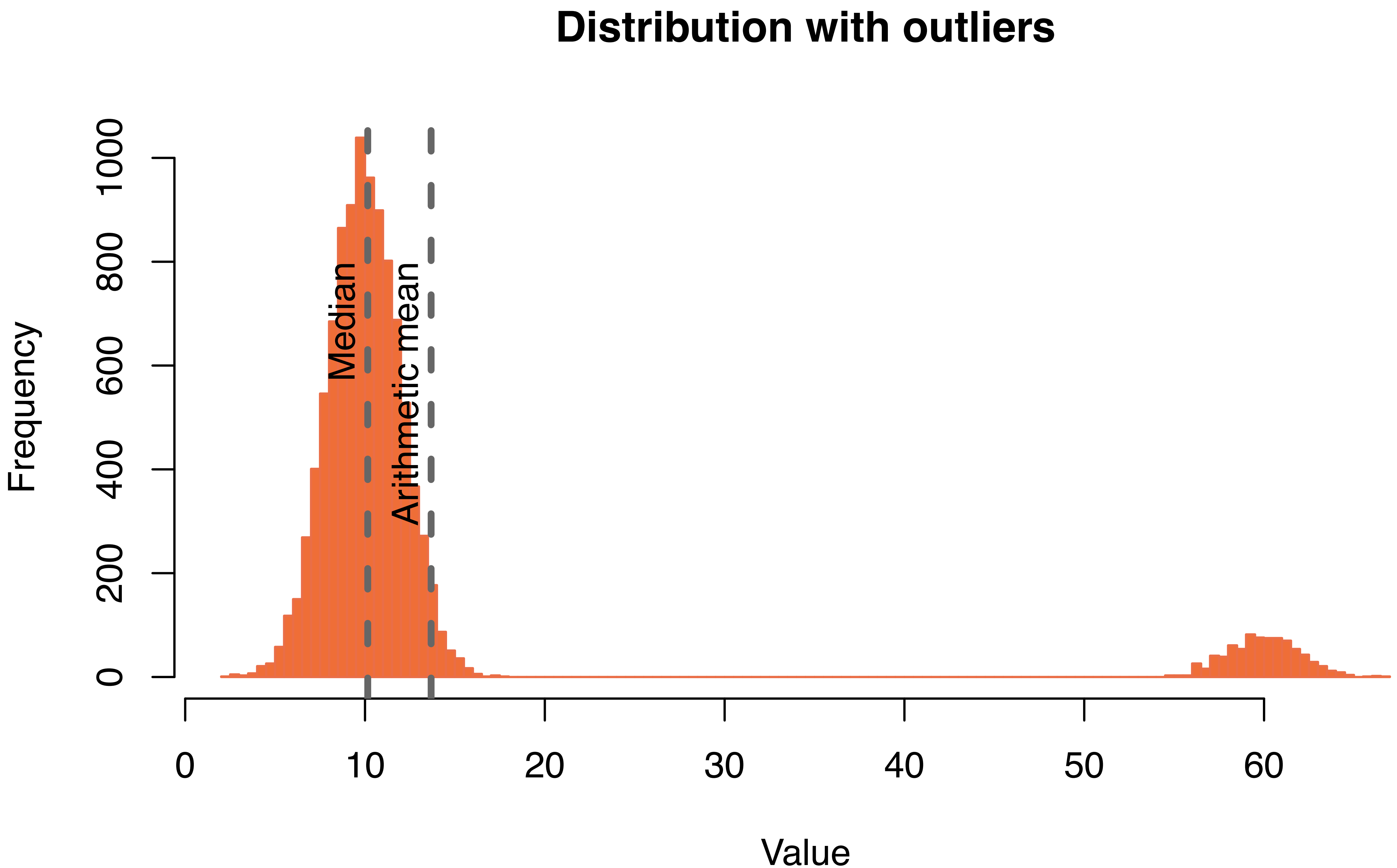
Less affected by skewed data and outliers than means.

Data: 7, 9, 12, 10, 3, 8, 4

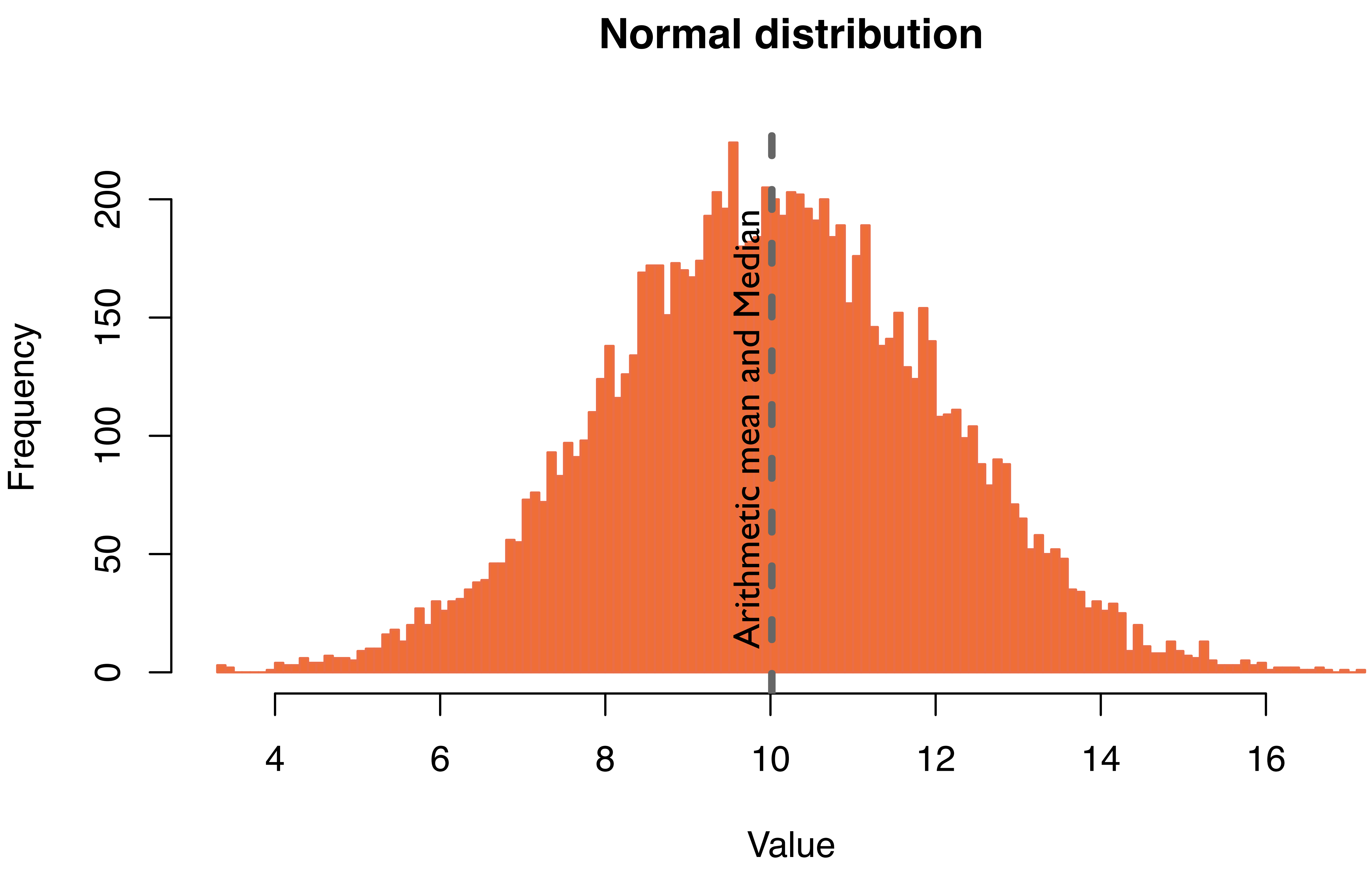
Sorted: 3, 4, 7, 8, 9, 10, 12

```
median(x)
```

Median

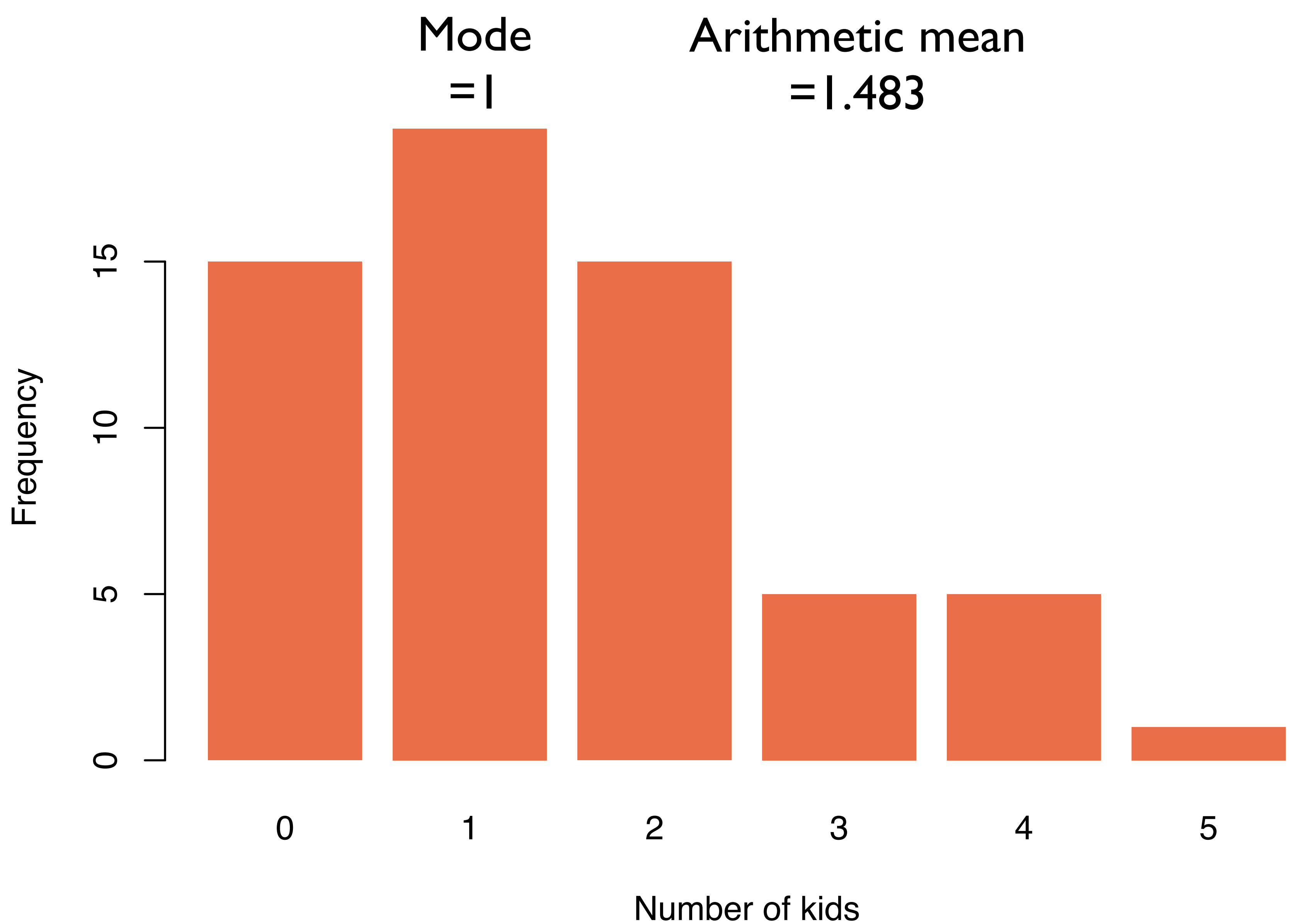

Distribution with outliers
Median
Arithmetic mean
Frequency
0
200
400
600
800
1000
0
10
20
30
40
50
60
Value

Median

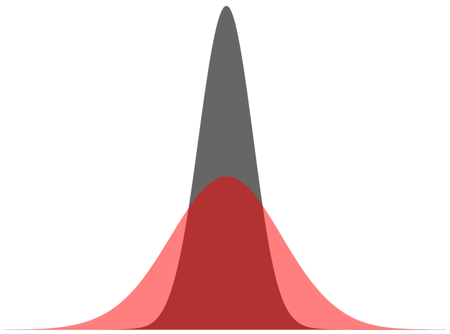

Normal distribution
Frequency
0
50
100
150
200
Arithmetic mean and Median
4
6
8
10
12
14
16
Value

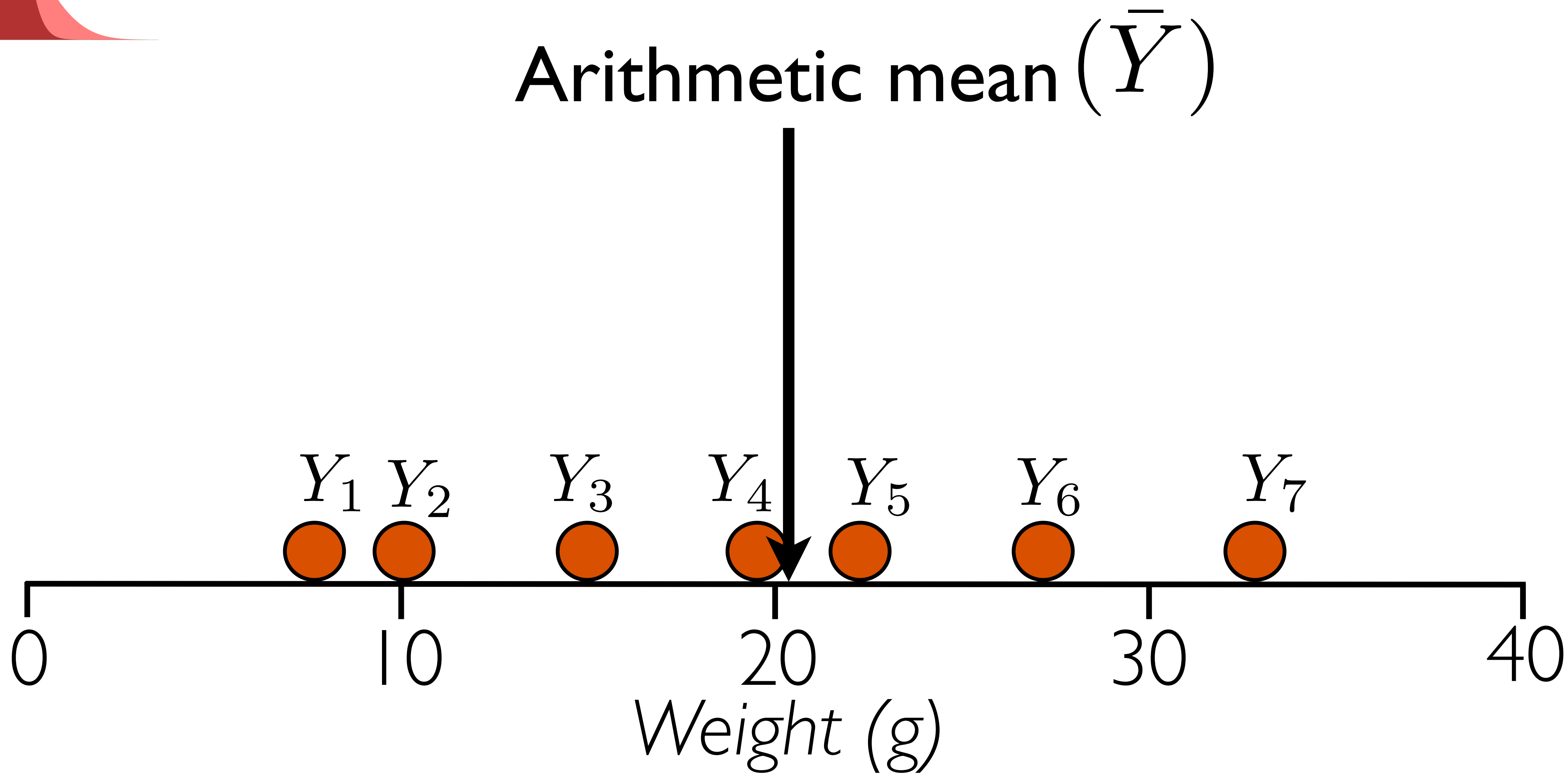
Mode

Mode: the most frequent number
(not useful for continuous variables).

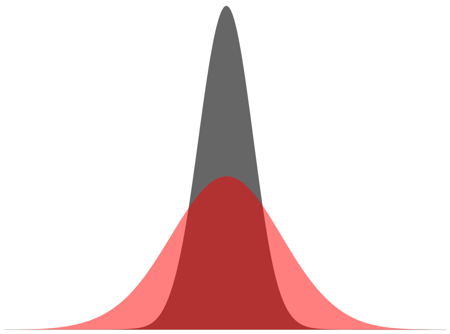


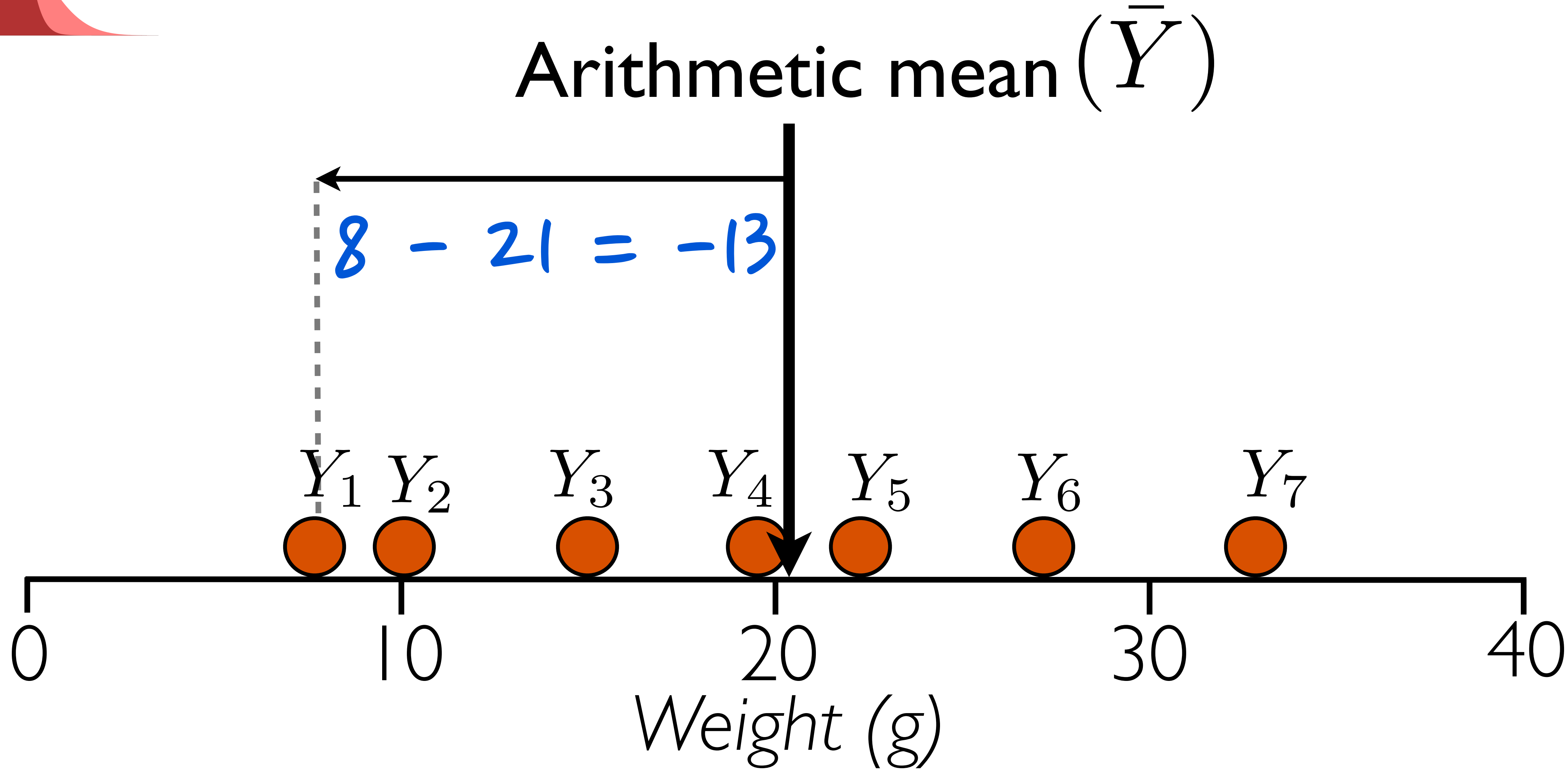
Spread





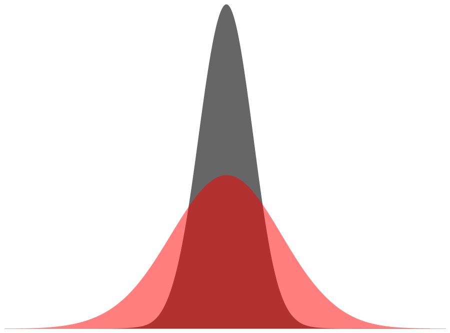
Spread

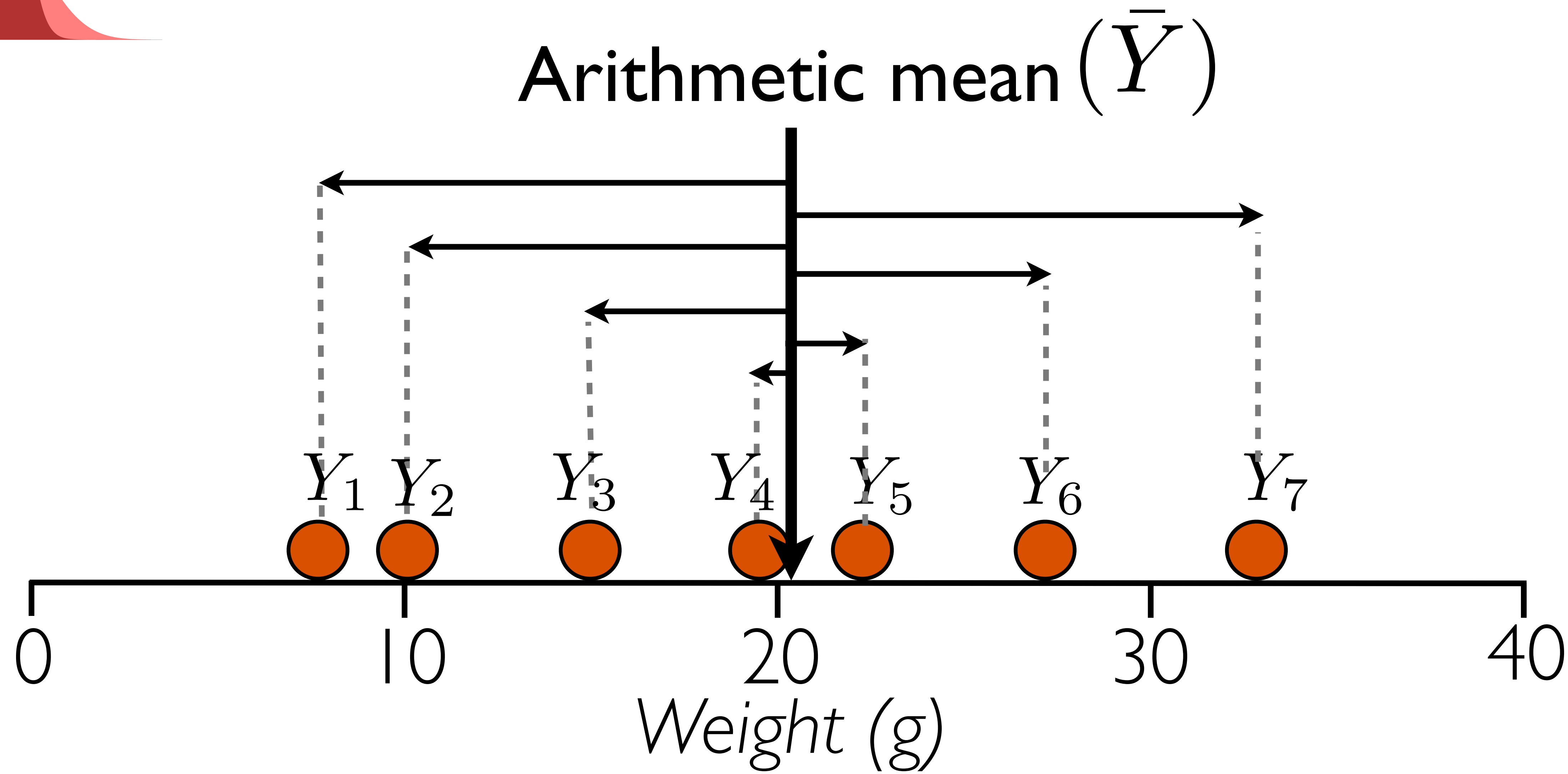




$$(Y_i - \bar{Y})$$

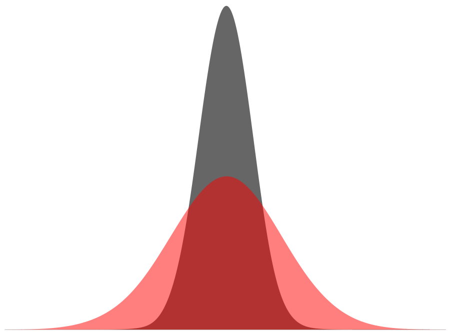
Spread

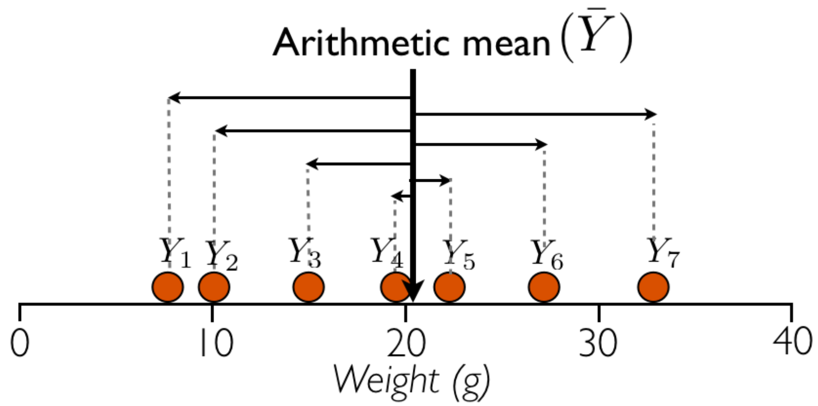




$$(Y_i - \bar{Y})$$

Spread





$$(Y_i - \bar{Y})$$ *A measure of distance from mean*

$$(Y_i - \bar{Y})^2$$ *Squared - to keep positive values*

$$\sum (Y_i - \bar{Y})^2$$ *Summed - to get a single measure*

$$\boxed{s^2 = \frac{\sum (Y_i - \bar{Y})^2}{n-1}}$$

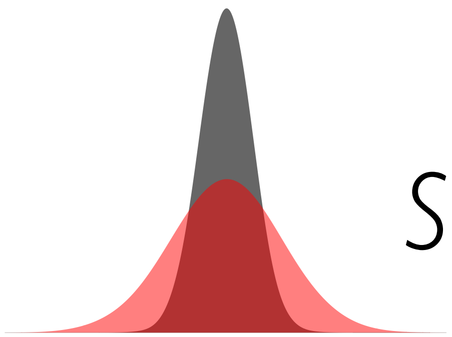
Sample variance

Divided by n-1 - to control for size of sample

Why $n-1$ **?**

See what happens with sample size of one.

Spread



Sample variance

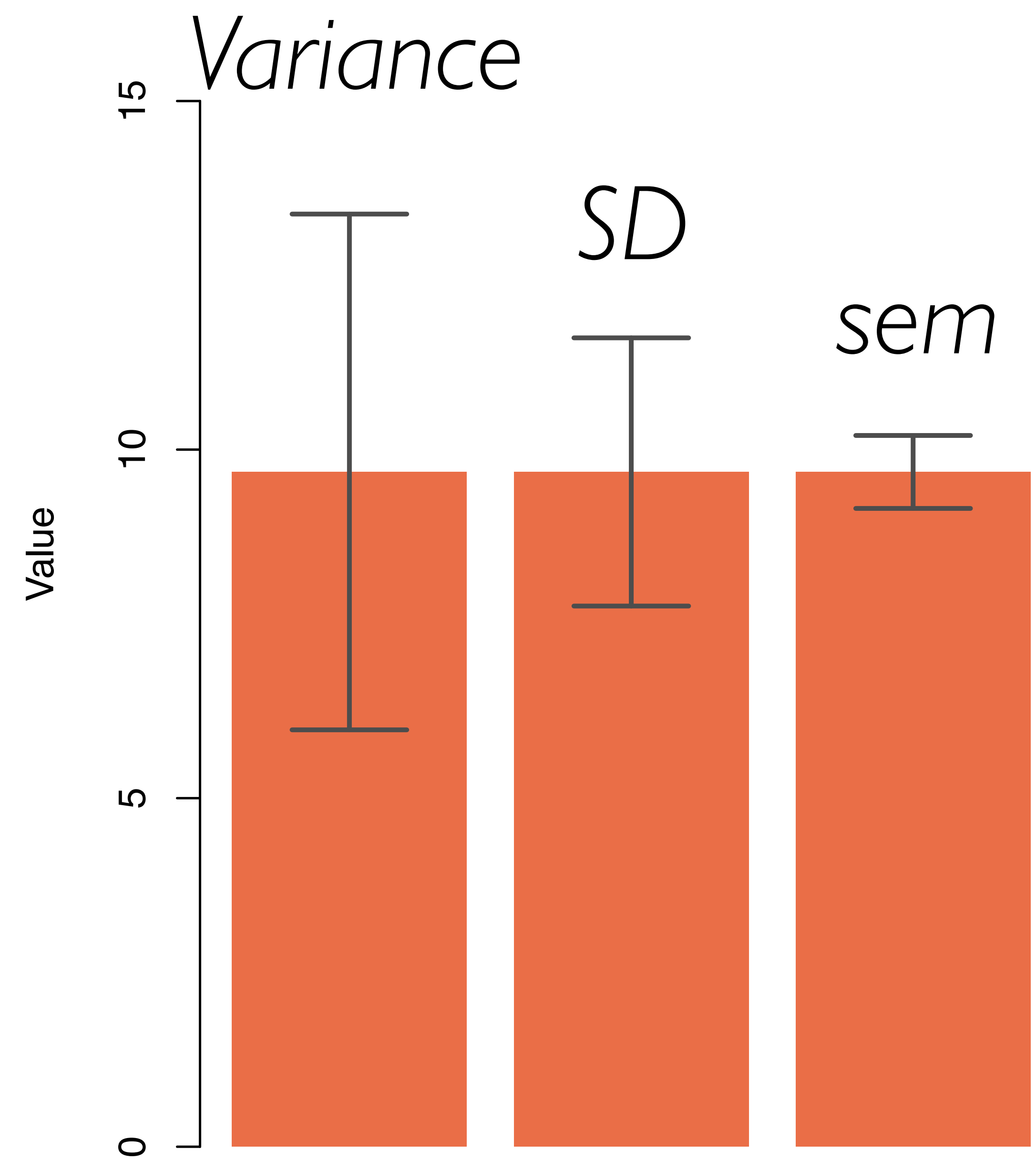
$$s^2 = \frac{\sum (Y_i - \bar{Y})^2}{n - 1}$$

Sample standard deviation

$$s = \sqrt{\frac{\sum (Y_i - \bar{Y})^2}{n - 1}}$$

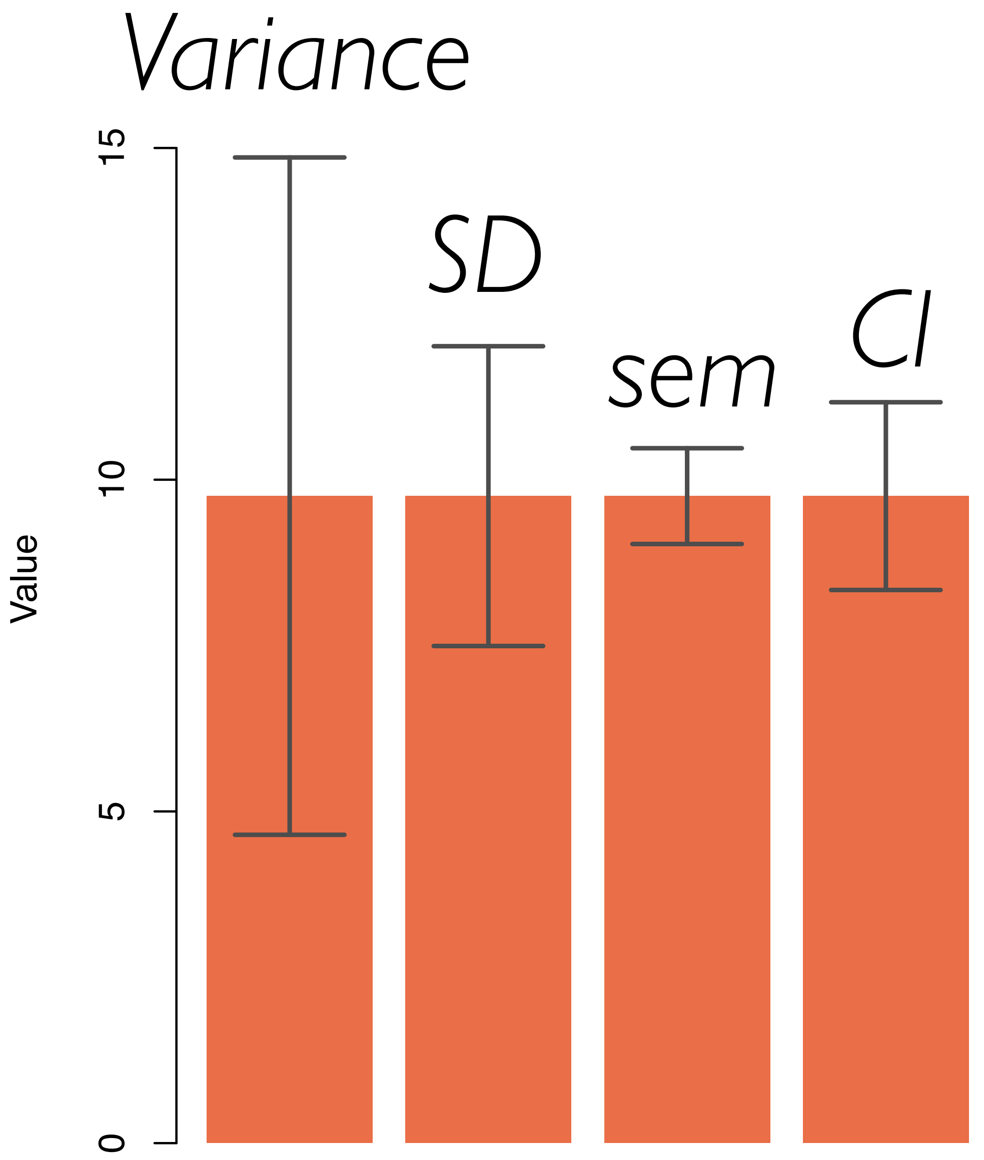
Standard error of the mean

$$s_{\bar{Y}} = \frac{s}{\sqrt{n}}$$



Confidence intervals

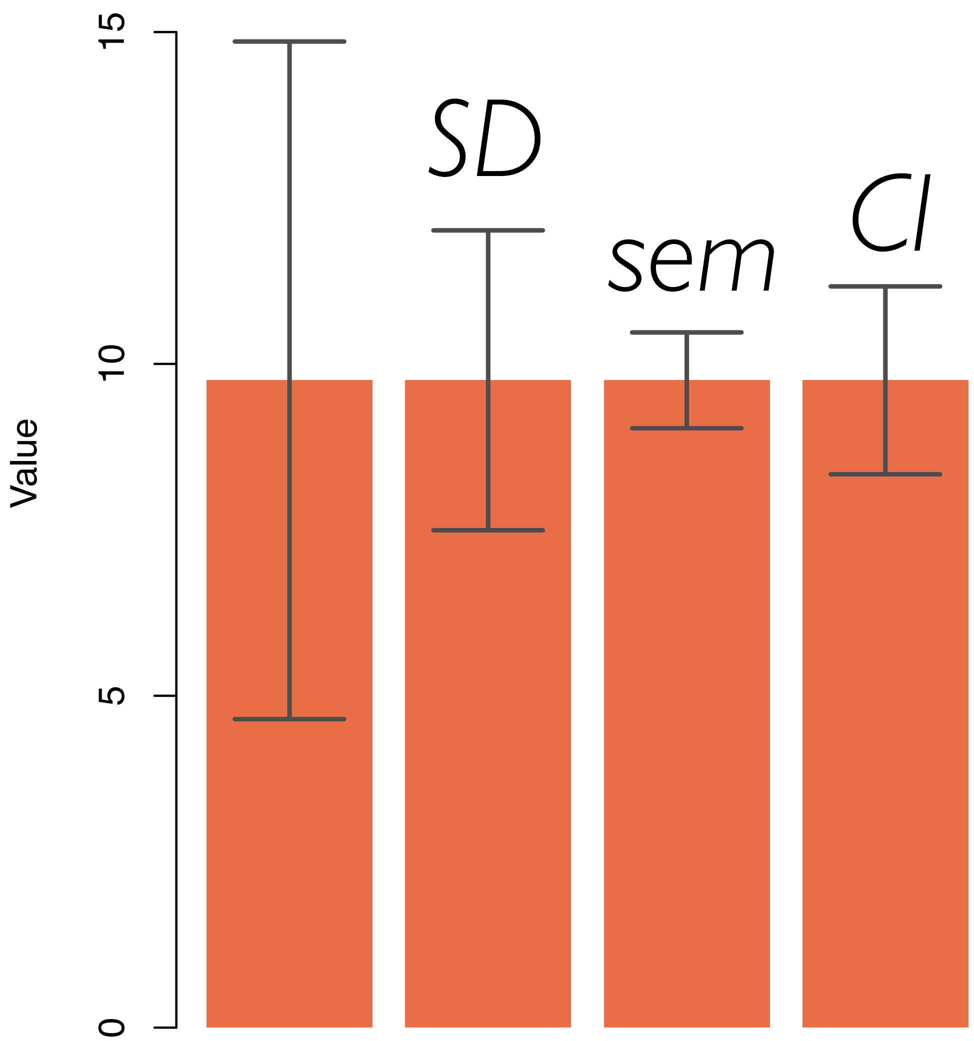
*Confidence interval = sem *1.96*



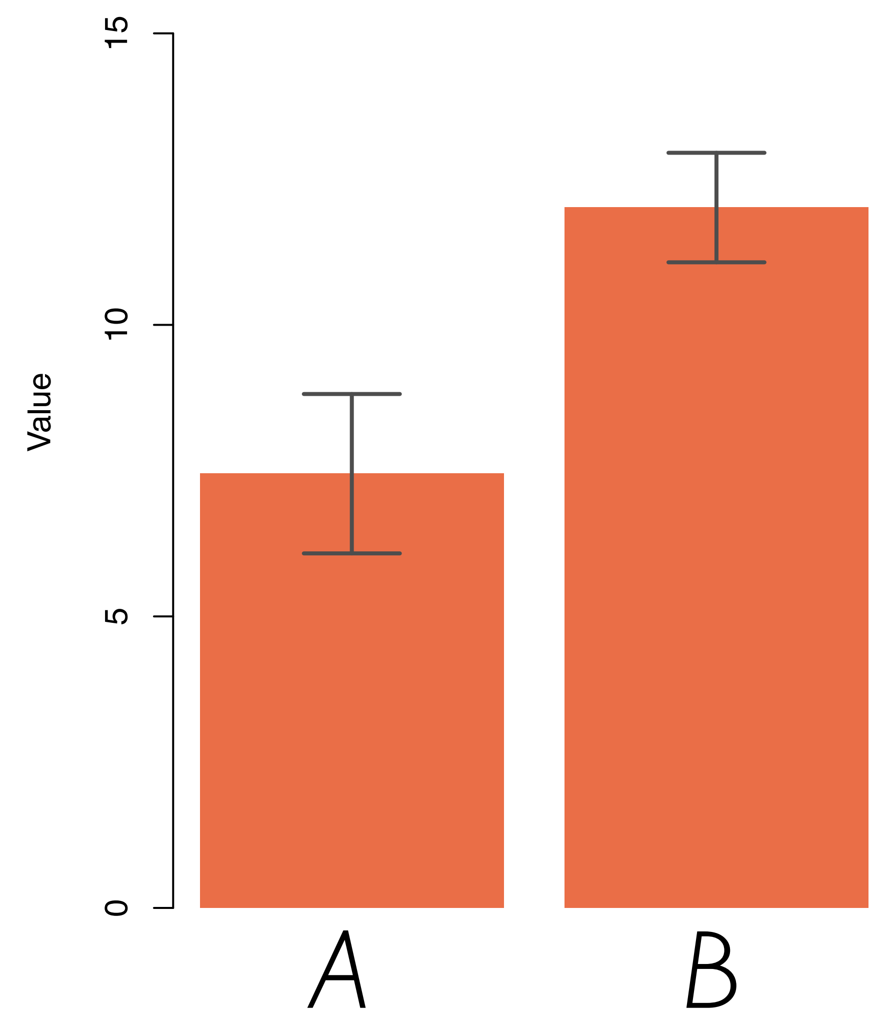
Confidence intervals

*Confidence interval = sem *1.96*

Variance

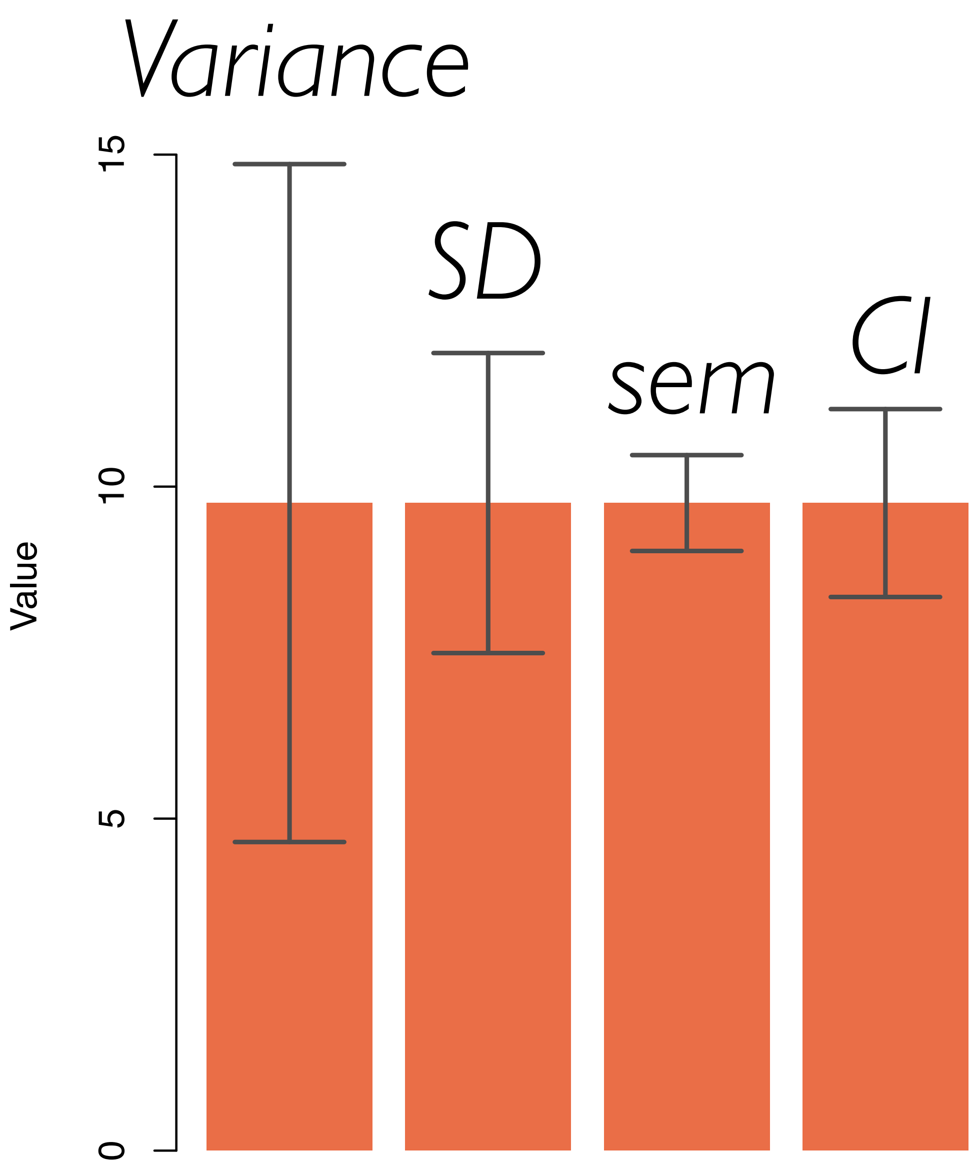


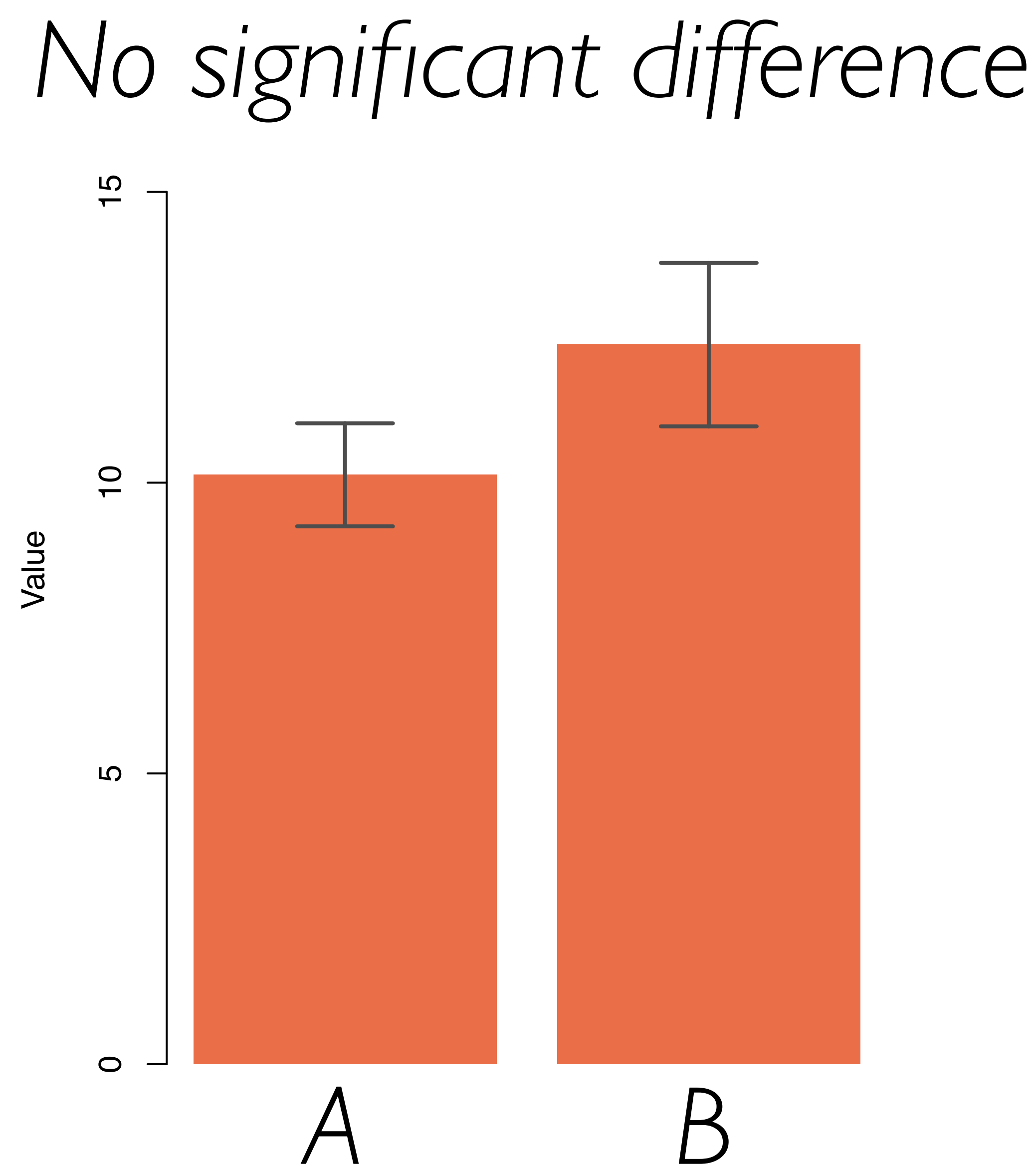
Significantly different



Confidence intervals

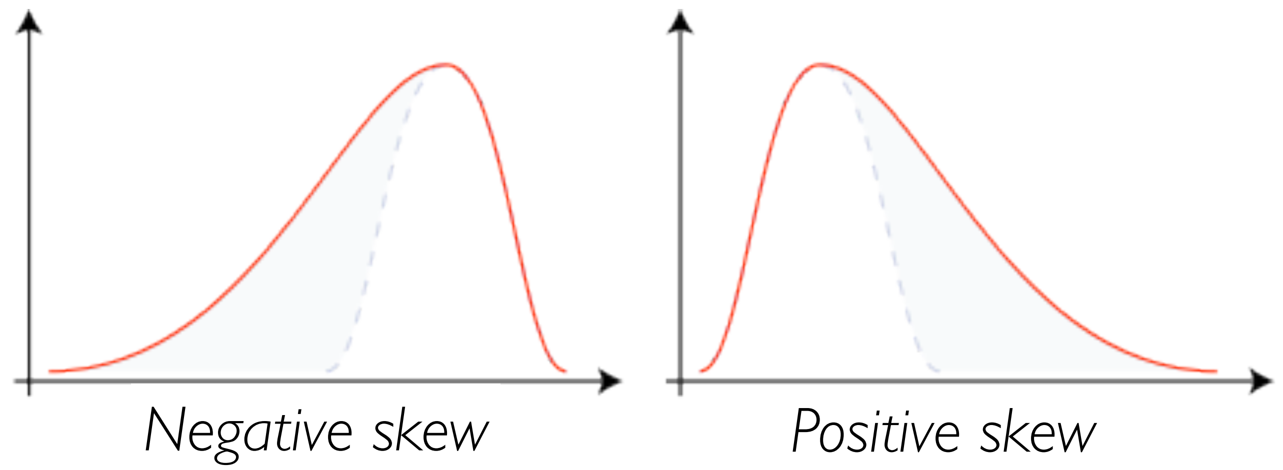
*Confidence interval = sem *1.96*





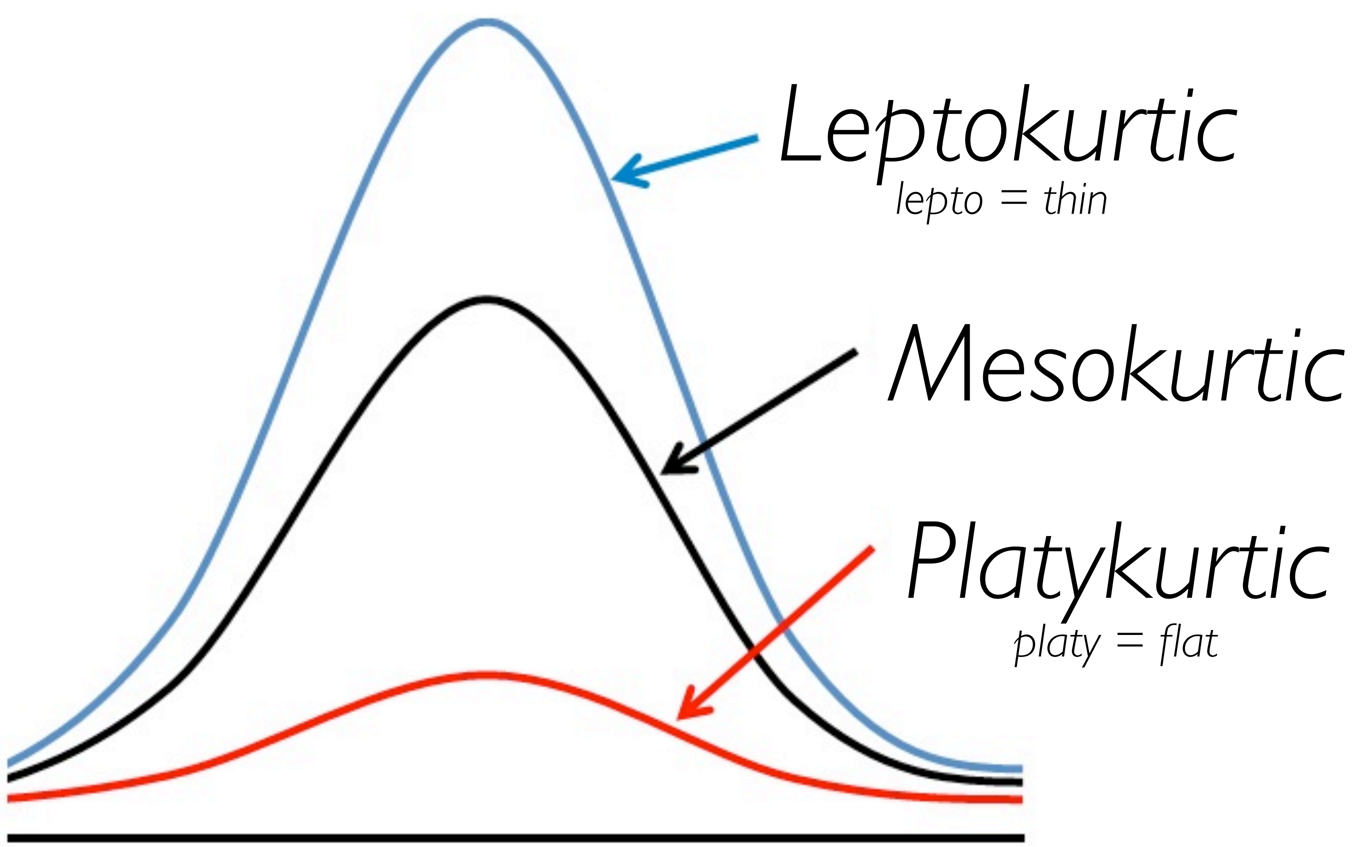
Shape

Skewness

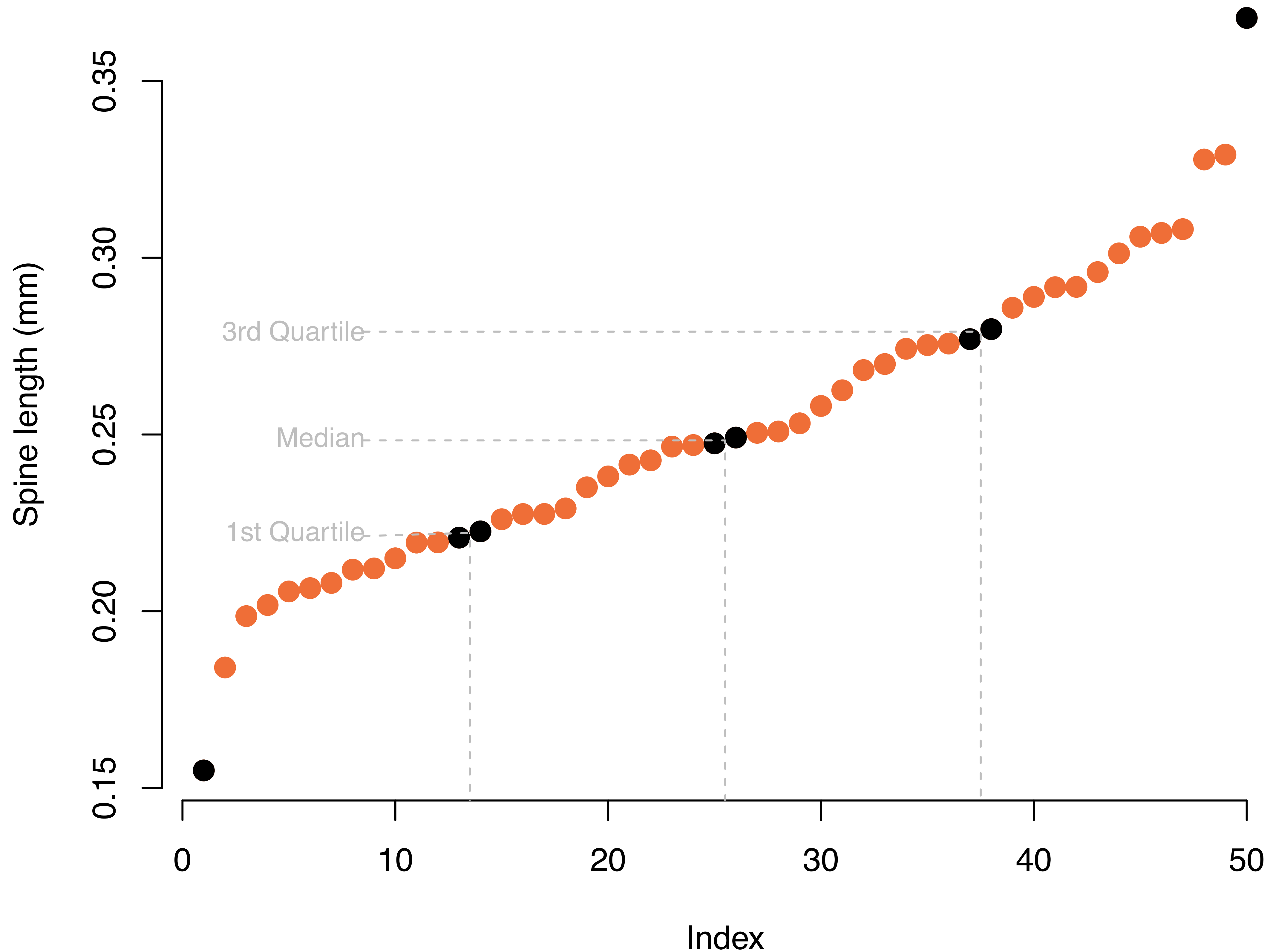


Kurtosis

"squashedness"

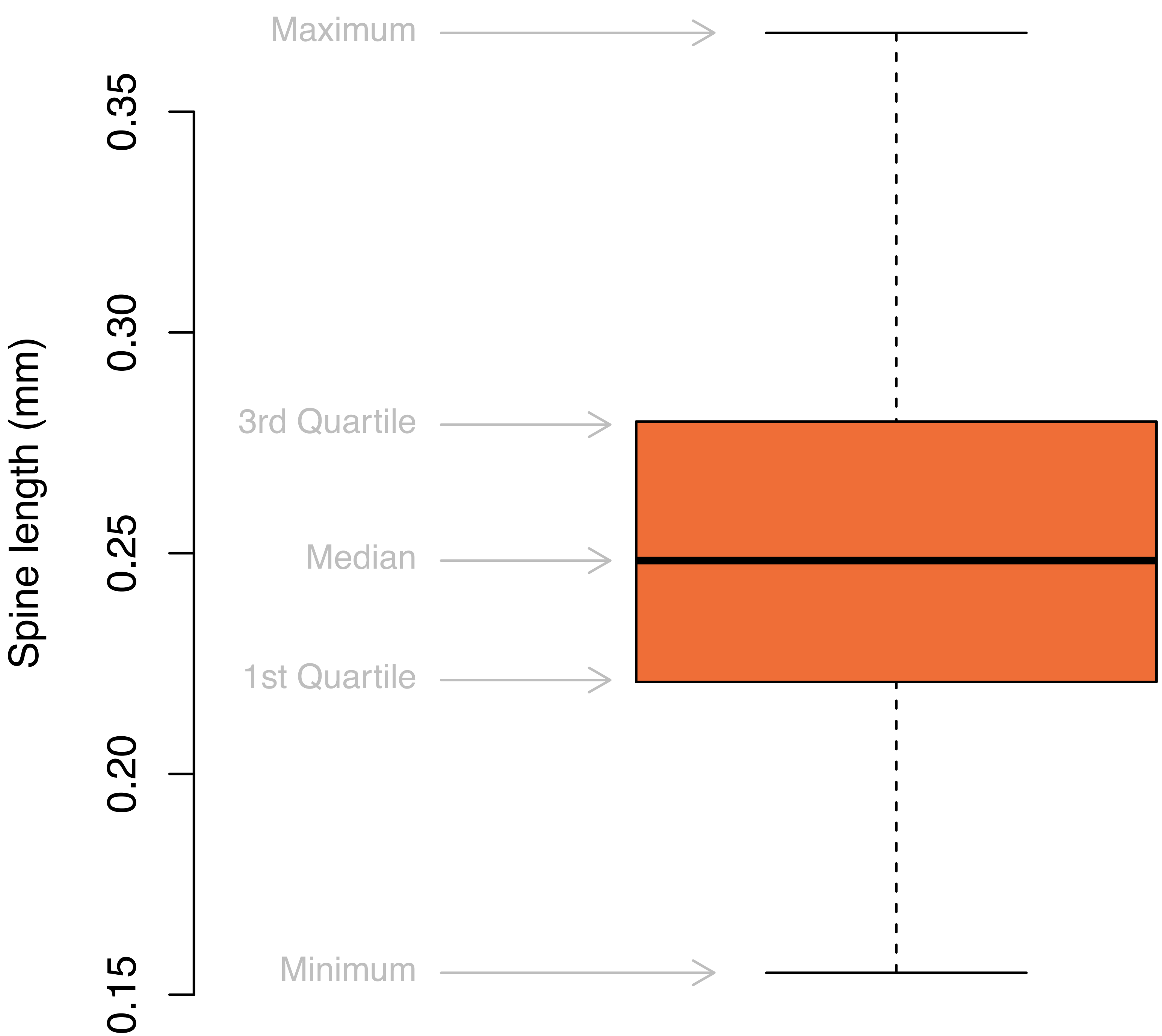


5 number summary



5 number summary

A box plot



Summary

Location: *average (mean, median, mode)*

Means: *arithmetic/geometric/harmonic means*

Law of large numbers: *large samples are good*

Spread: *variance, std. deviation, std. error of the mean*

confidence interval

Shape: *skew/kurtosis*

The 5 number summary: *quartiles, min/max, median*