2. Measurement errors; statistical estimators

"Errors using inadequate data are much less than those using no data at all"

Charles Babbage

error, n.

The action or state of erring.

Something incorrectly done through ignorance or inadvertence; a mistake, e.g. in calculation, judgement, speech, writing, action, etc.

Mathematics. The quantity by which a result obtained by observation or by approximate calculation differs from an accurate determination.

Oxford English Dictionary

Measurement errors

Systematic and random errors

Systematic errors

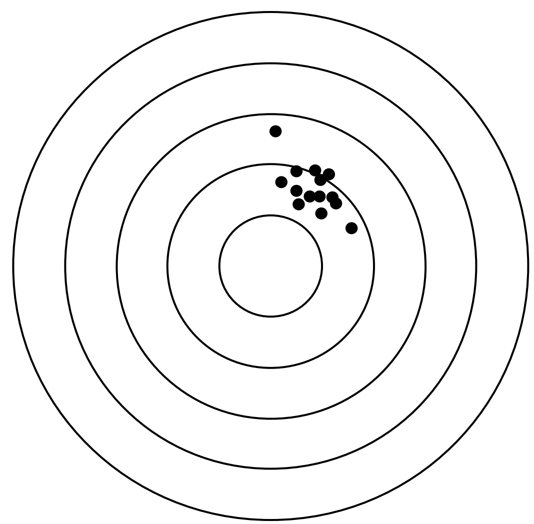
your mistakes

- Incorrect instrument calibration
- Change in experimental conditions
- Pipetting errors

Random errors

statistics sucks

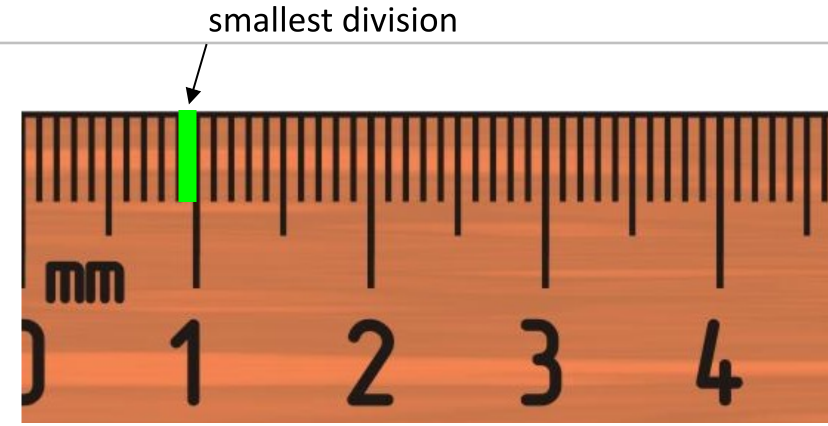
- Reading errors
- Sampling errors
- Intrinsic variability



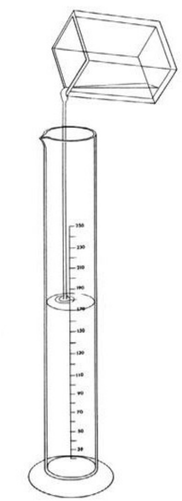
YOU NEED REPLICATES

Reading error

- The reading error is ±half of the smallest division
- Example: 23±0.5 mm from a ruler



- Beware of digital instruments that sometimes give readings much better than their real accuracy
- Read the instruction manual!

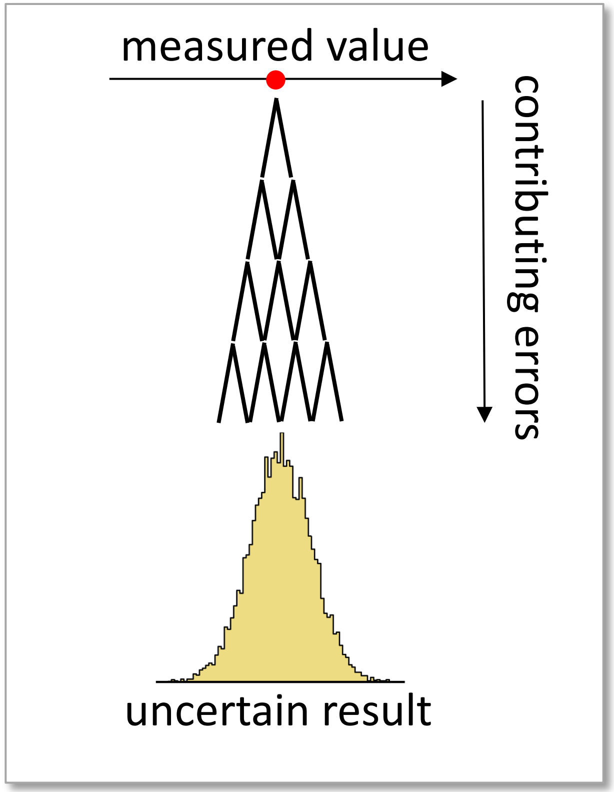


- **Reading error does not take into account biological variability**

Random measurement error

- Determine the strength of oxalic acid in a sample
- Method: sodium hydroxide titration

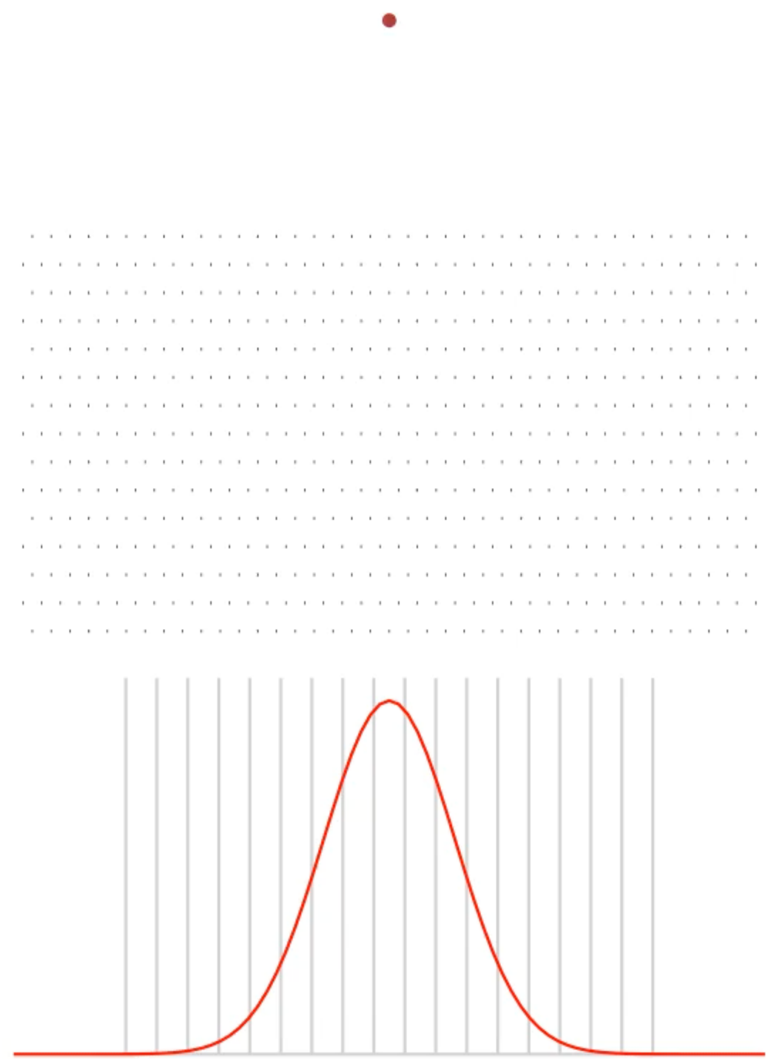
- Uncertainties contributing to the final result
  - volume of the acid sample
  - judgement at which point acid is neutralized
  - volume of NaOH solution used at this point
  - accuracy of NaOH concentration
    - weight of solid NaOH dissolved
    - volume of water added



- Each of these uncertainties adds a random error to the final result

- Measurement errors are normally distributed

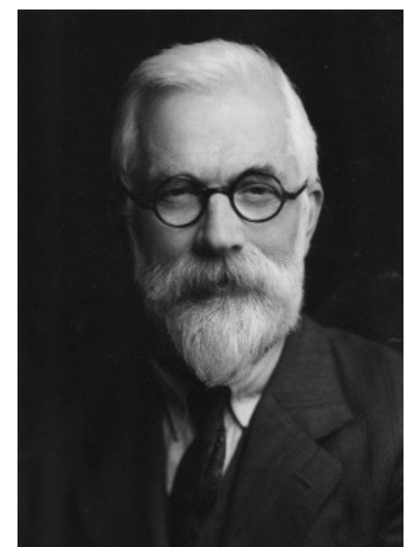
Galton board



Population and sample

Population and sample

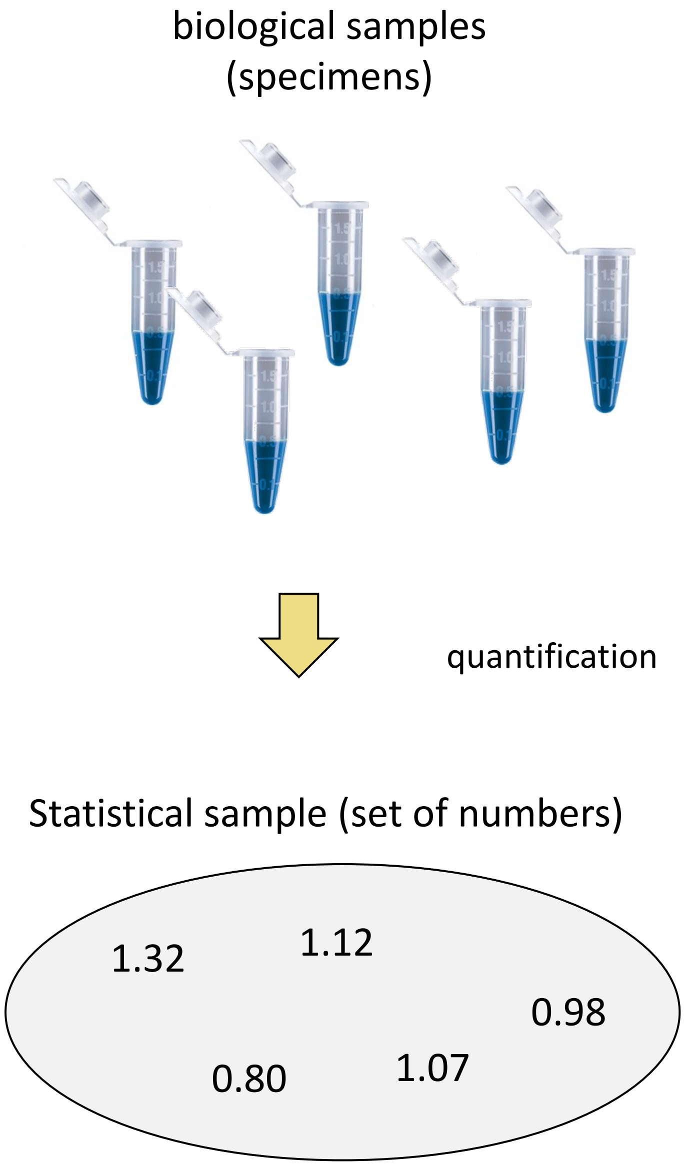
Sample selection



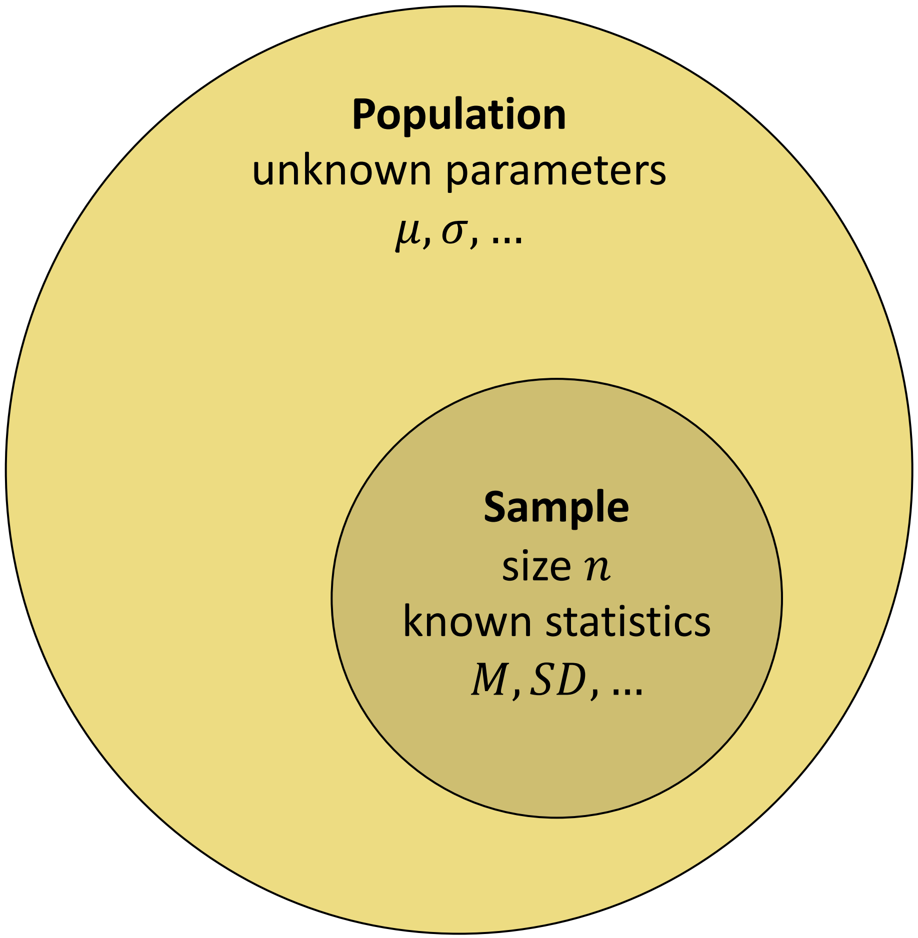
- Terms nicked from social sciences
- Most biological experiments involve sample selection
- Terms “population” and “sample” are not always literal

What is a sample?

- The term “sample” has different meanings in biology and statistics
- **Biology**: sample is a specimen, e.g., a cell culture you want to analyse
- **Statistics**: sample is (usually) a set of numbers (measurements)
- In these talks: $x_1, x_2, \dots, x_n$



Population and sample



A **parameter** describes a population

A **statistical estimator** describes a sample

A statistical estimator (statistic) approximates the corresponding parameter

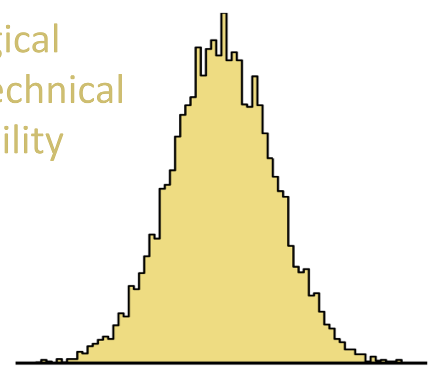
Sampling from a population = experiment

Population
all possible measurements

Experiment

Sample
5 biological replicates

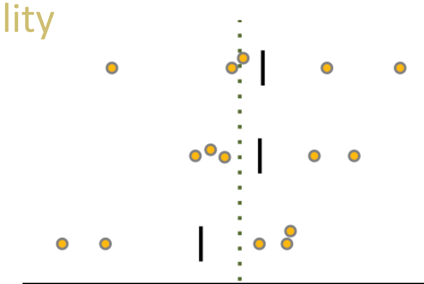
Biological and technical variability



Measured quantity

Sampling

Sampling variability



Measured quantity

Statistical estimators

"The average human has one breast and one testicle"

Des MacHale

What is a statistical estimator?

"Right and lawful rood*" from *Geometrei*, by Jacob Köbel (Frankfurt 1575)

Stand at the door of a church on a Sunday and bid 16 men to stop, tall ones and small ones, as they happen to pass out when the service is finished; then make them put their left feet one behind the other, and the length thus obtained shall be a right and lawful rood to measure and survey the land with, and the 16th part of it shall be the right and lawful foot.

Over 400 years ago Köbel:

- introduced random sampling from a population
- required a representative sample
- defined standardized units of measure
- used 16 replicates to minimize random error
- calculated an estimator: the sample mean

*rood – a unit of measure equal to 16 feet

Example

- Weight of 7 mice
- This is a **sample**
- We can find
  - mean = 19.2 g
  - median = 18.7 g
  - standard deviation = 4.4 g
  - standard error = 1.7 g
  - interquartile range = 6.0 g
- These are examples of **statistical estimators**

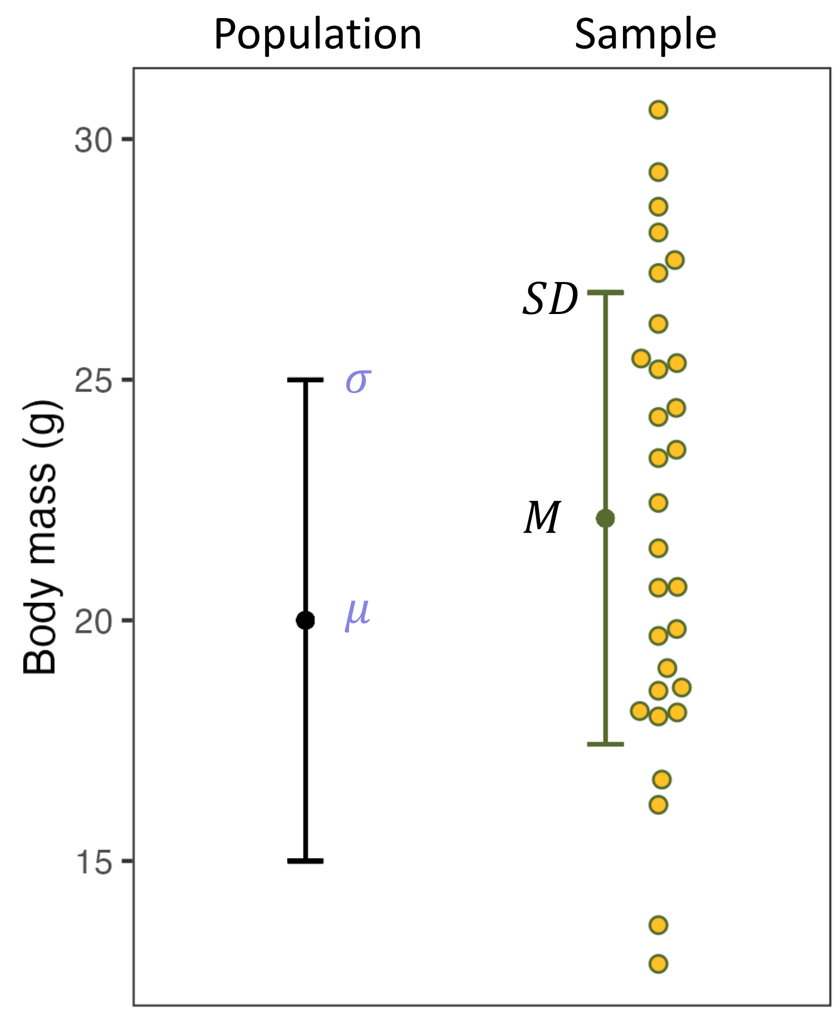
No.	Weight (g)
1	13.6
2	16.1
3	25.1
4	24.8
5	16.6
6	19.8
7	18.7

Statistical estimators

- Statistical estimator is a sample attribute used to estimate a population parameter

- Population parameters
  - mean, μ
  - standard deviation, σ

- Sample $(x_1, x_2, \ldots, x_n)$ estimates
  - mean, M
  - standard deviation, SD



Mean vs median

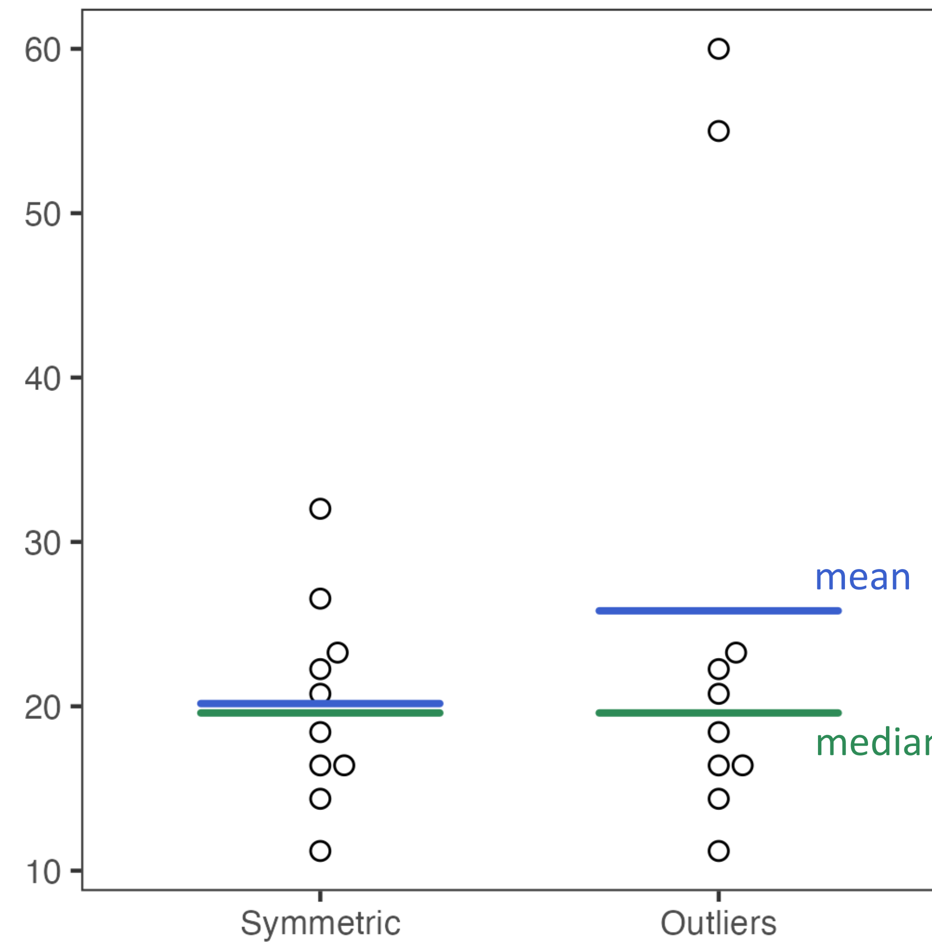
Median

- More appropriate for skewed distributions
- Not sensitive to outliers

Mean

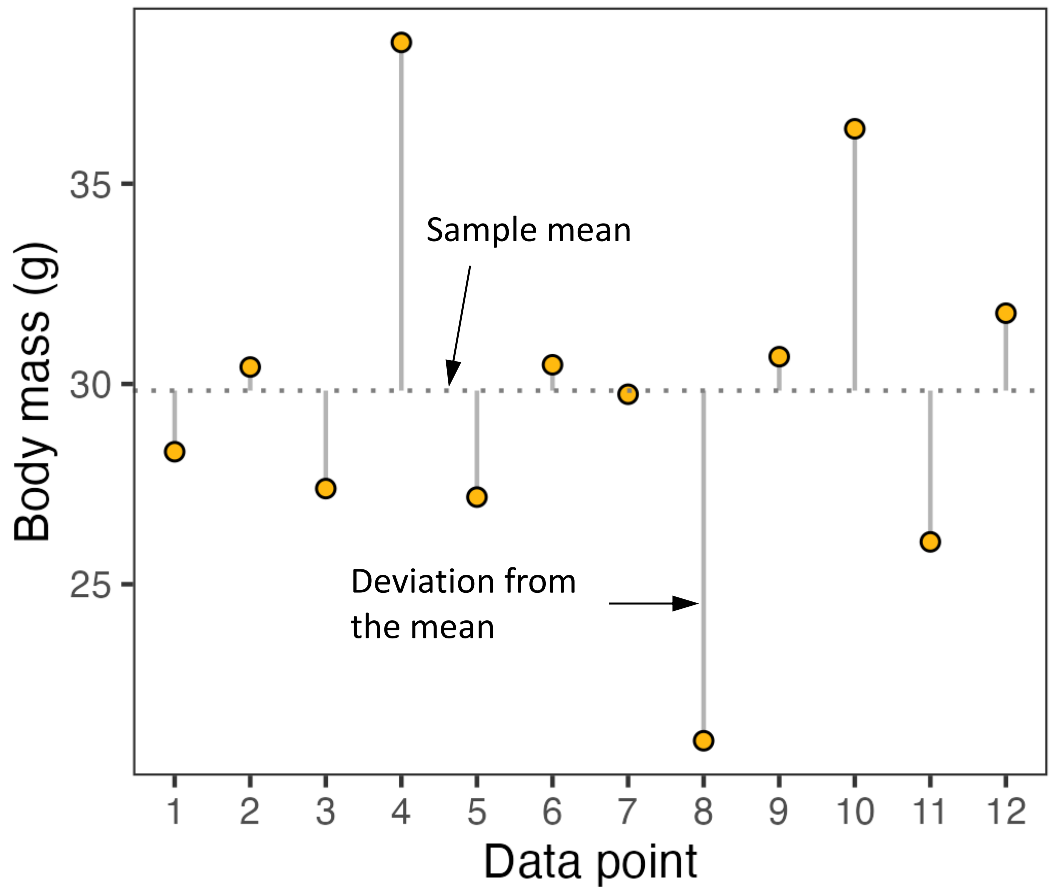
- Better estimate of the central value
- Statistical tests on the mean (e.g. t-test) are more power full than non-parametric tests

- If your data are symmetric, use mean



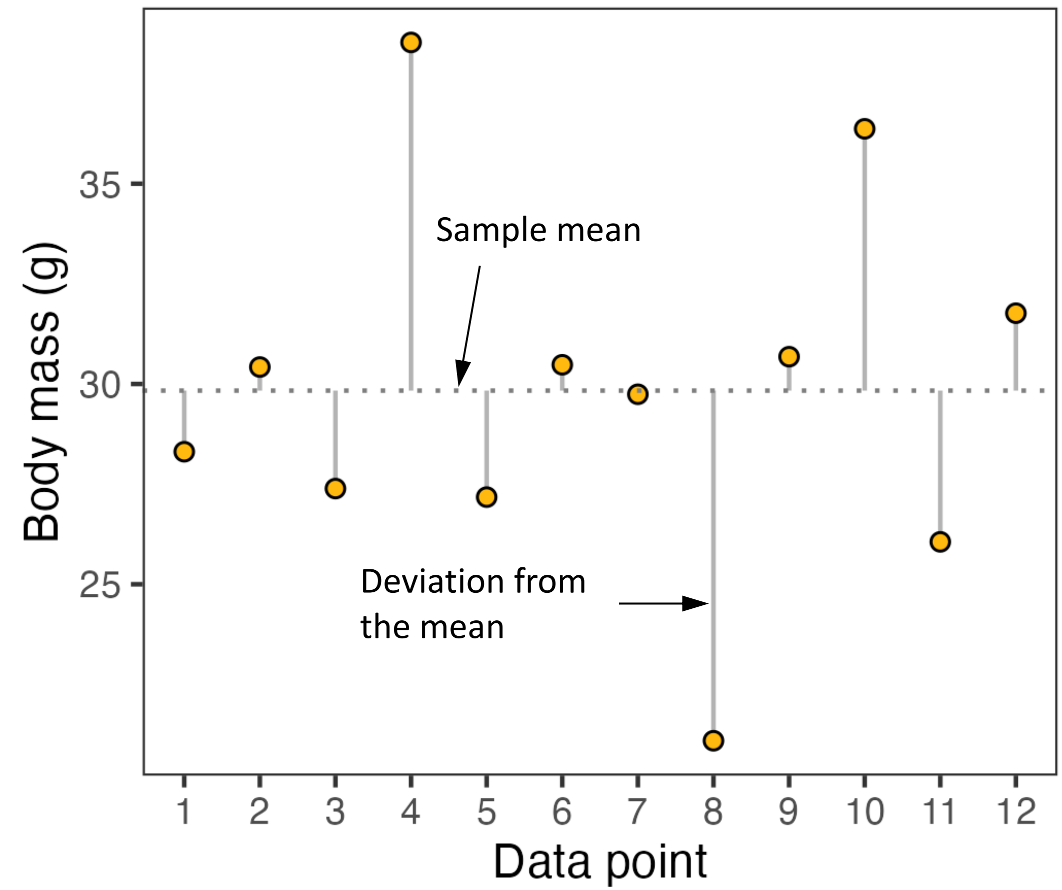
Standard deviation

- Standard deviation is a measure of spread of data points
- Idea:
  - calculate the mean
  - find deviations from the mean
  - get rid of negative signs
  - combine them together



Standard deviation

- Standard deviation is a measure of spread of data points
- Idea:
  - calculate the mean
  - find deviations from the mean
  - get rid of negative signs
  - combine them together



- **Standard deviation of $x_1, x_2, \dots, x_n$**

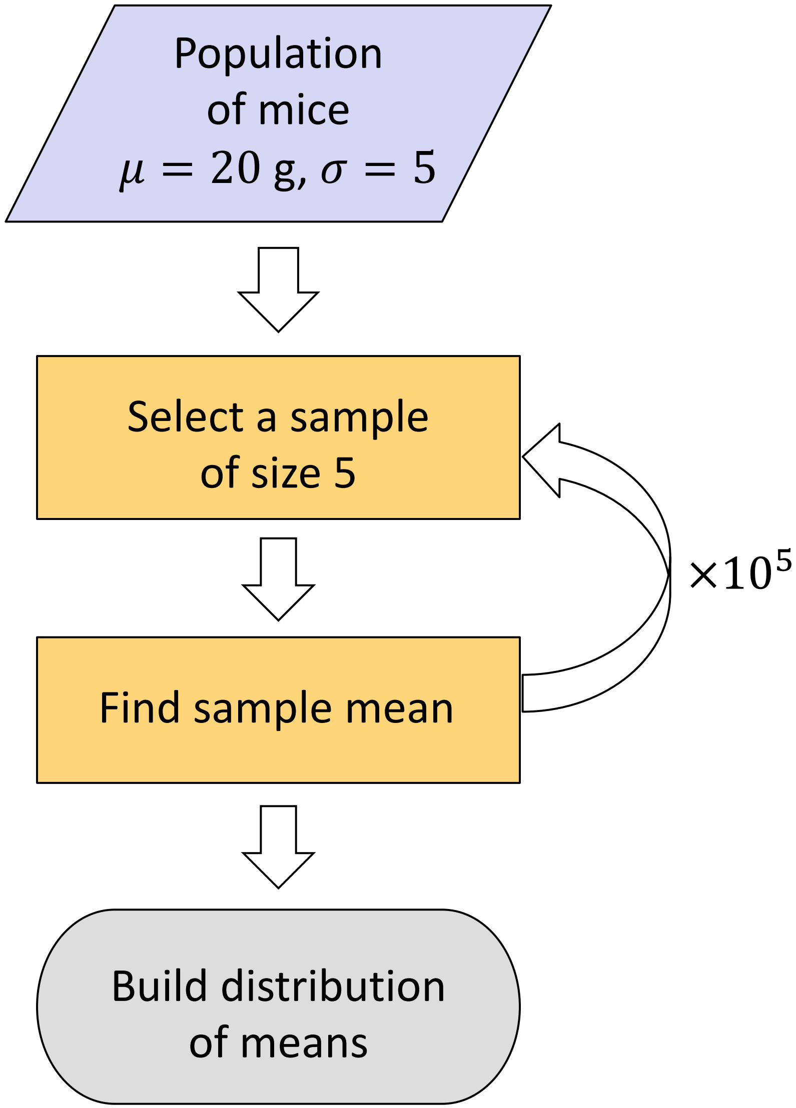
$$SD_n = \sqrt{\frac{1}{n}\sum_i (x_i - M)^2}$$

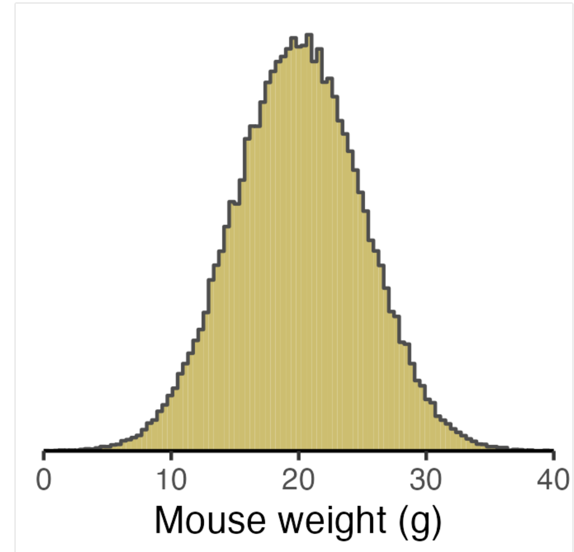
$$SD_{n-1} = \sqrt{\frac{1}{n-1}\sum_i (x_i - M)^2}$$

SD_{n-1}^2 estimates true variance better than SD_n^2

Standard error of the mean

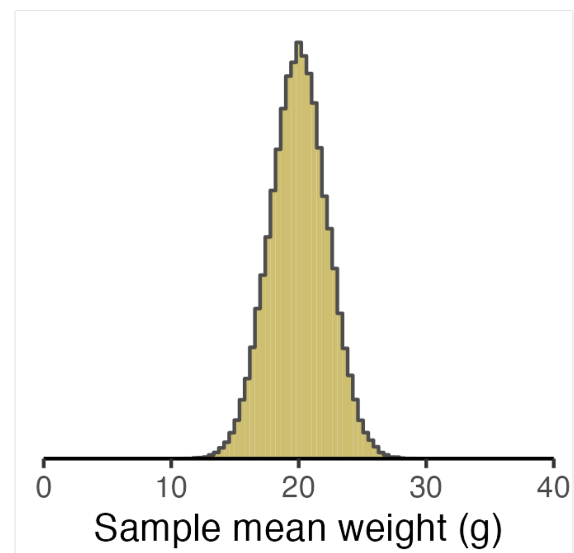
Sampling distribution of the mean





Population

Mean	Standard deviation
μ	σ



Sample means

Mean	Standard deviation
μ	$\frac{\sigma}{\sqrt{n}}$

Standard error of the mean

Hypothetical experiment

- 100,000 samples of 5 mice
- Build a distribution of sample means
- Width of this distribution is the true uncertainty of the mean

$$\sigma_m = \frac{\sigma}{\sqrt{n}} = 2.2 \text{ g}$$

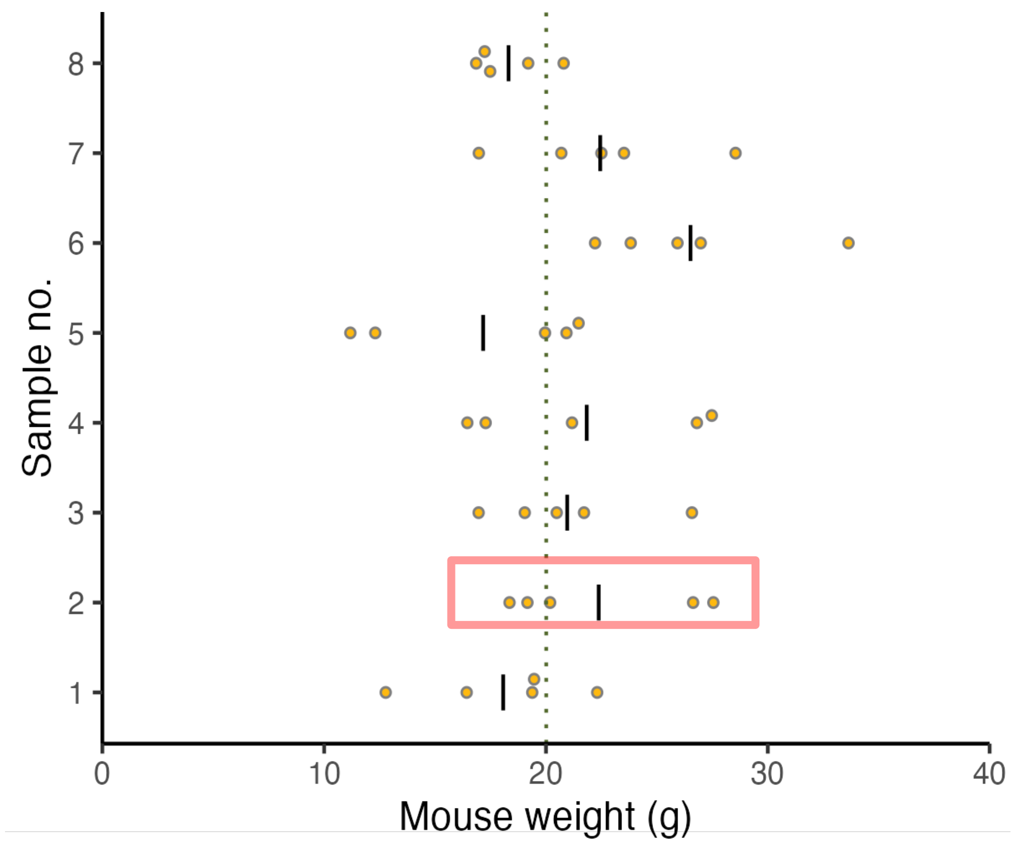
Real experiment

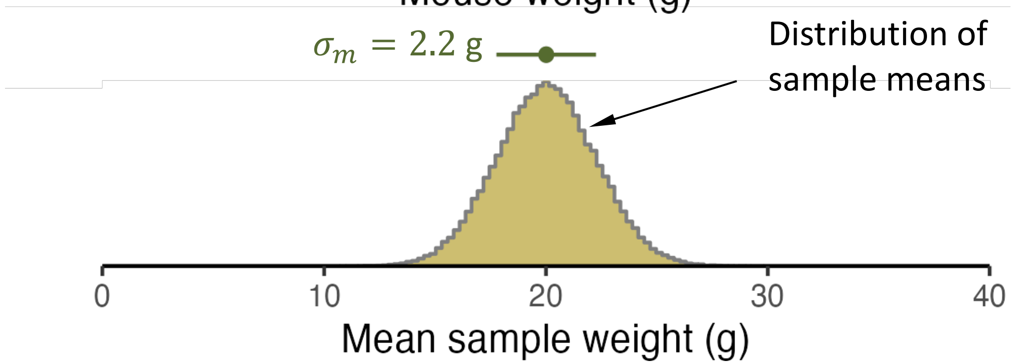
- 5 mice
- Measure body mass:

  18.4, 19.2, 20.2, 26.6, 27.5 g
- Find standard error

$$SE = \frac{SD}{\sqrt{n}} = 2.0 \text{ g}$$

***SE* is an approximation of σ_m**





Standard error of the mean

Hypothetical experiment

- 100,000 samples of 30 mice
- Build a distribution of sample means
- Width of this distribution is the true uncertainty of the mean

$$\sigma_m = \frac{\sigma}{\sqrt{n}} = 0.9 \text{ g}$$

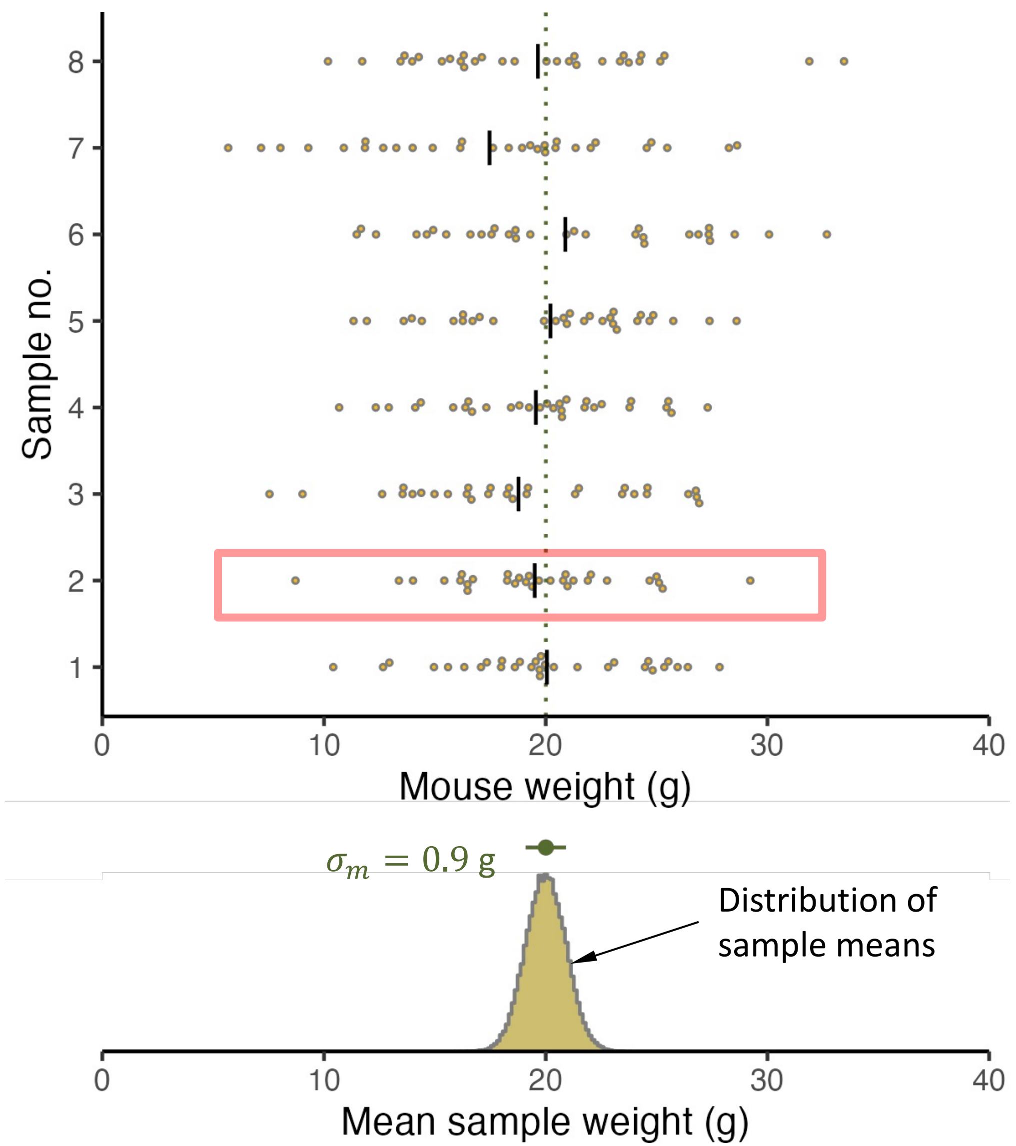
Real experiment

- 30 mice
- Measure body mass:

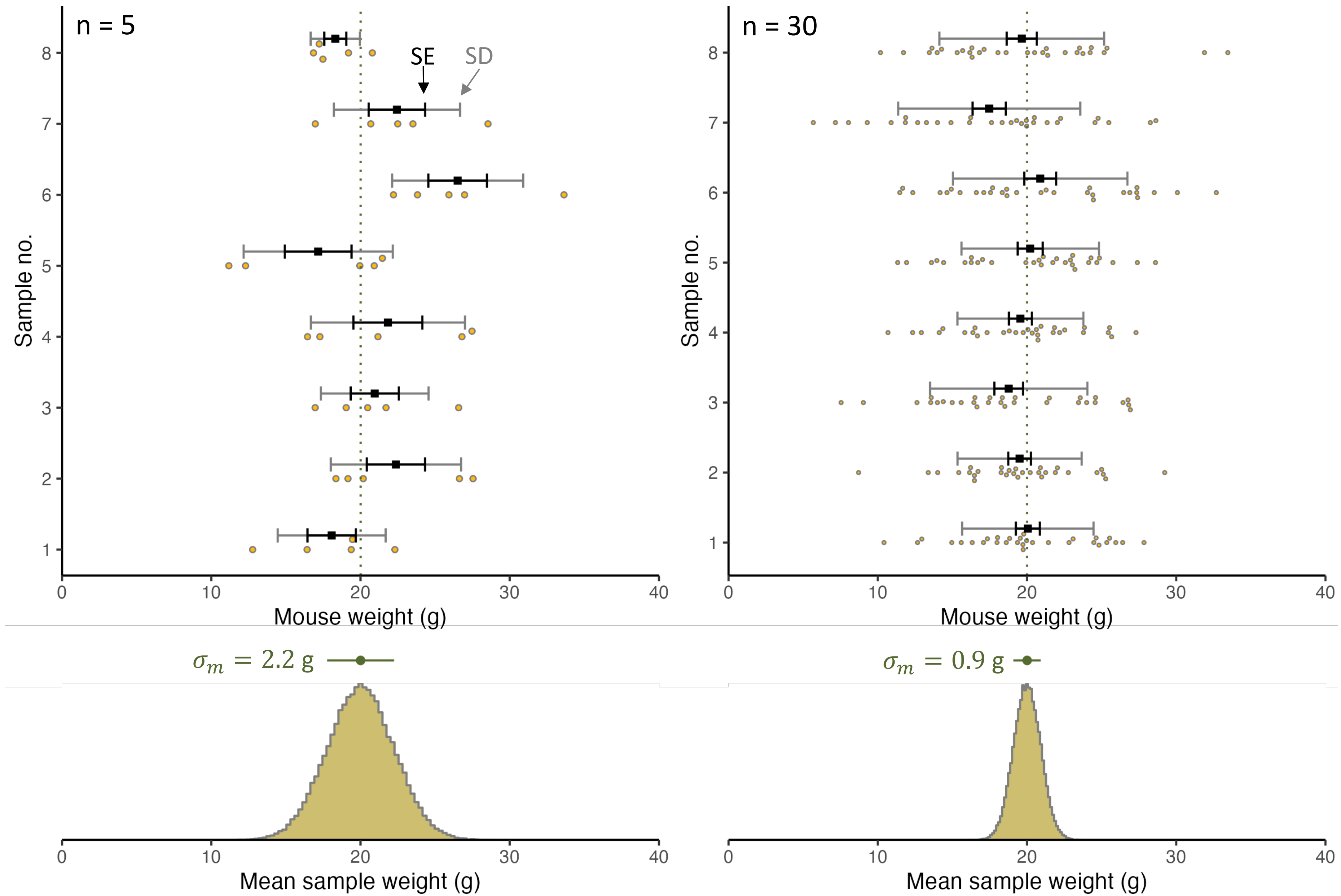
  8.7, 13.4, ..., 29.2 g
- Find standard error

$$SE = \frac{SD}{\sqrt{n}} = 0.76 \text{ g}$$

***SE* is an approximation of σ_m**



Standard error of the mean



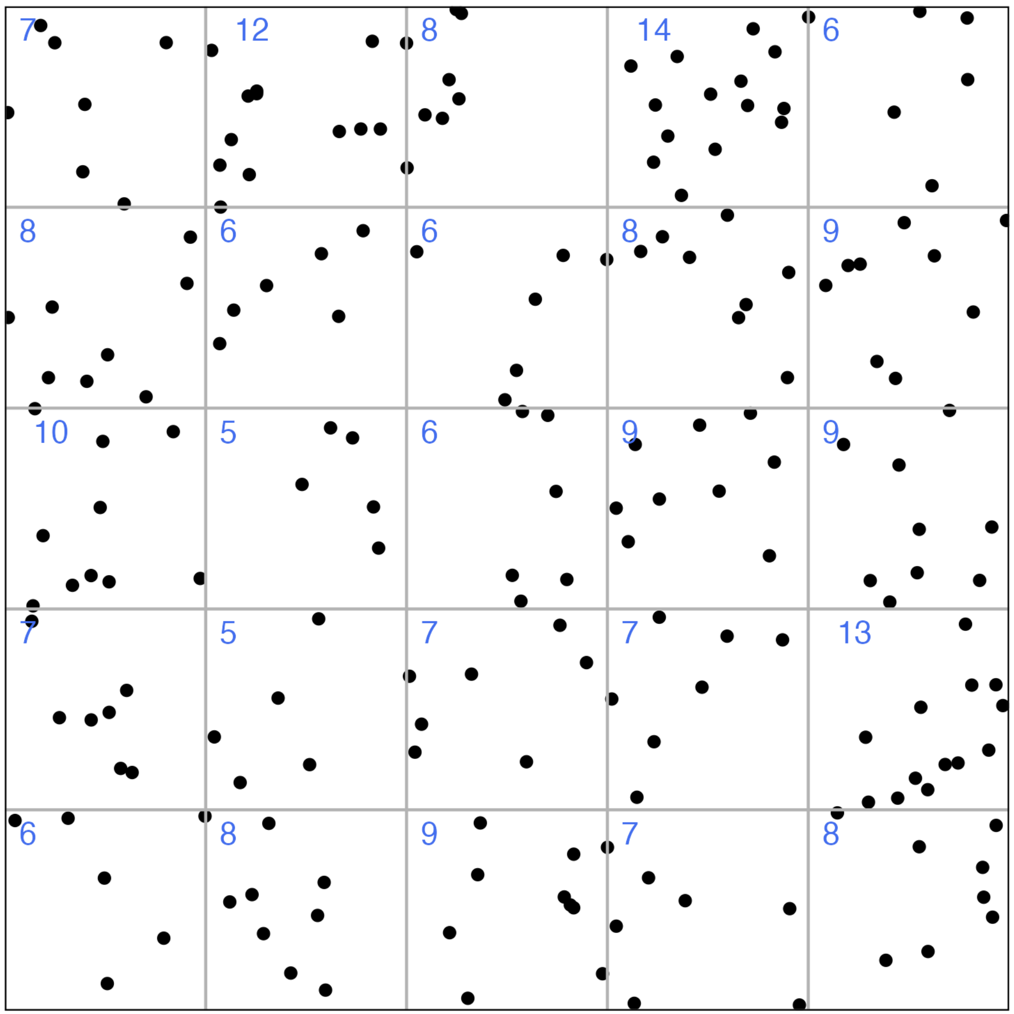
Standard deviation and standard error

Standard deviation	**Standard error**
$SD = \sqrt{\frac{1}{n-1}\sum_i (x_i - M)^2}$	$SE = \frac{SD}{\sqrt{n}}$
Measure of dispersion in the sample	Error of the mean
Estimates the true standard deviation in the population, σ	Estimates the width (standard deviation) of the distribution of the sample means
Does not depend on sample size	Gets smaller with increasing sample size

Counting error

(standard error of the count)

Counting bacteria



- Bacteria uniformly distributed
- One box = one aliquot
- Random count, Poisson law

Counting error

- Dilution plating of bacteria

- Found $C = 10$ colonies

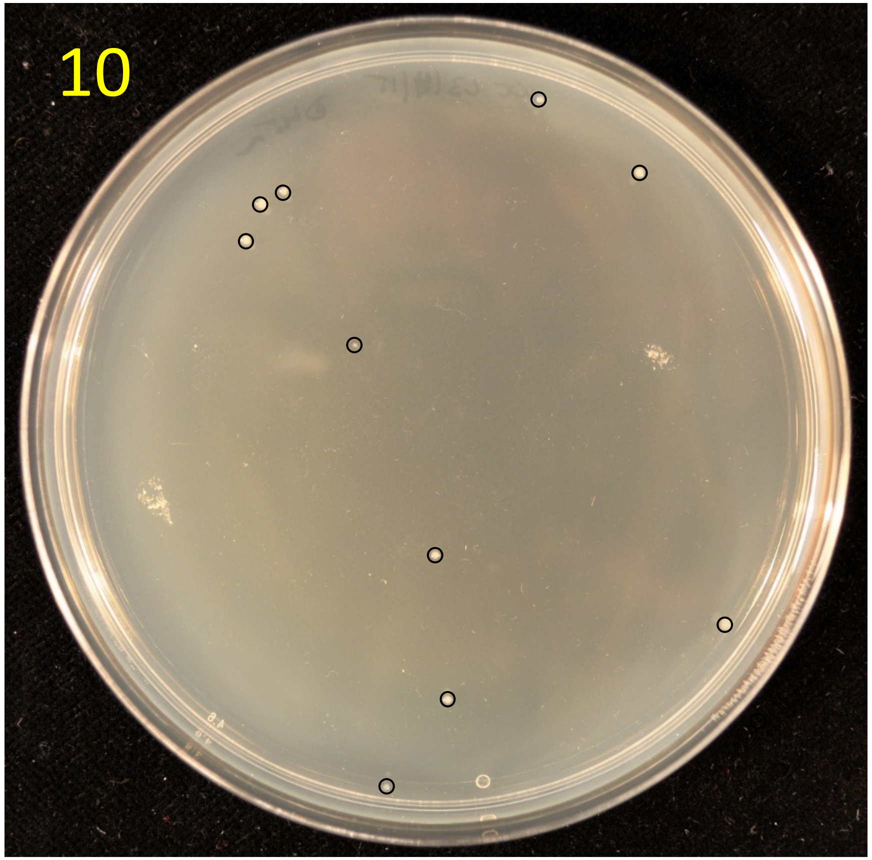
- Counting statistics: Poisson distribution

  $$\sigma = \sqrt{\mu}$$

- Use standard deviation as error estimate to obtain the *standard error of the count*

  $$S = \sqrt{C} = \sqrt{10} \approx 3$$

  $$C = 10 \pm 3$$



Counting error

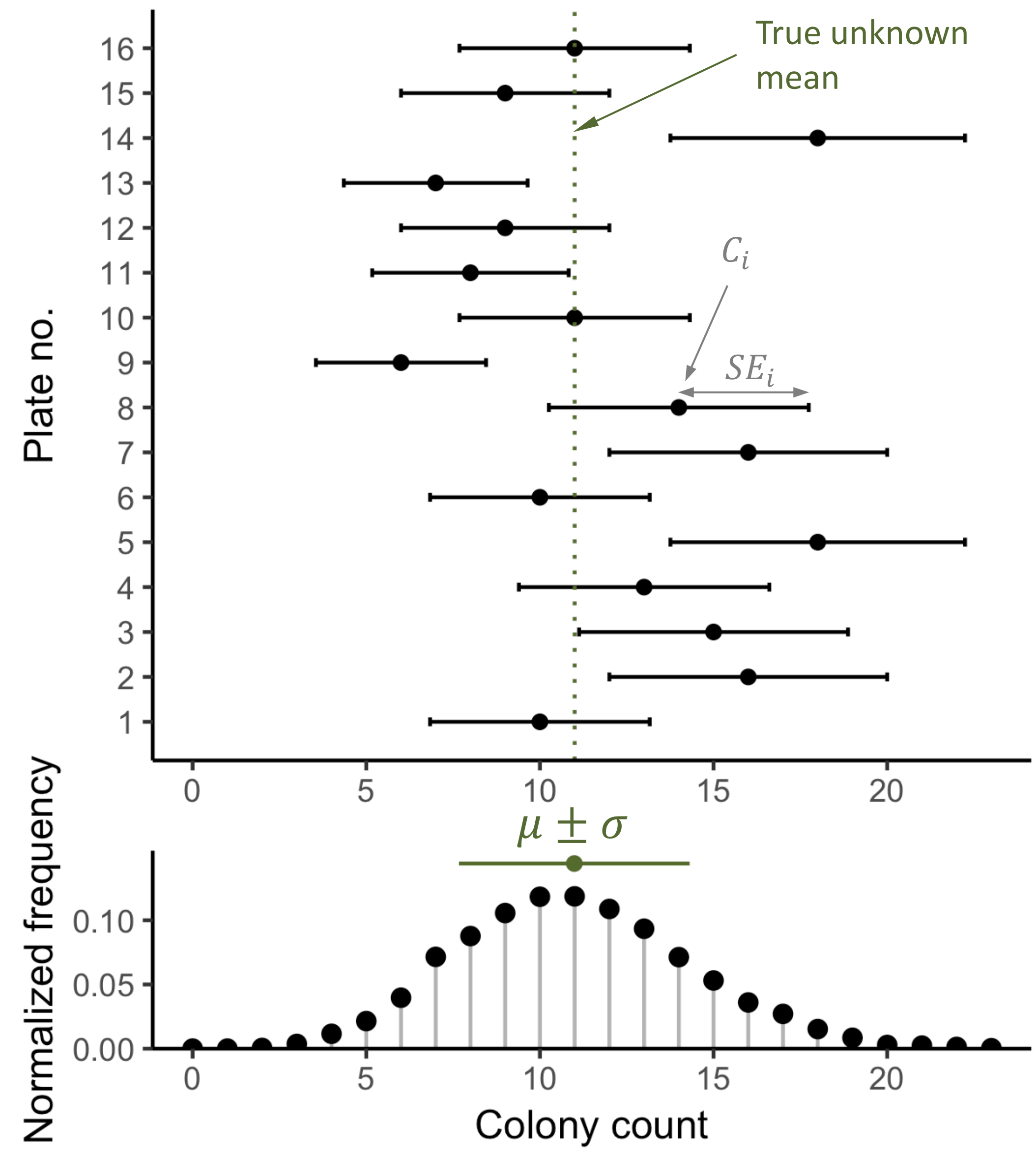
- *Gedankenexperiment*

- Measure counts on 10,000 plates

C_i	Count from plate i
$SE_i = \sqrt{C_i}$	Its error
μ	Unknown population mean
$\sigma = \sqrt{\mu}$	Unknown population SD

- Counting errors, SE_i, are similar, but not identical, to σ

- C_i is an estimator of μ
- SE_i is an estimator of σ



Exercise: is Dundee a murder capital of Scotland?

25 October 2022

SCOTTISH DAILY EXPRESS

NEWS POLITICS COMMENT ROYALS BONNIE SCOTLAND MORE

CONTACT US NEWSLETTERS SIGNUP

News › Scottish News › Crime

West Lothian and Dundee are Scotland's murder capitals as country sees one homicide EVERY WEEK

What was the homicide rate in your local area last year? Although the total number of homicides is now the lowest since 1976, the rate of overall violent crime was up last year

City	Murders	Per 100,000
Dundee	4	2.69
West Lothian	5	2.72
Glasgow	10	1.57
Edinburgh	3	0.57
Aberdeen	1	0.44

Exercise: is Dundee a murder capital of Scotland?

City	Murders	Per 100,000
Dundee	4	2.69
Glasgow	10	1.57

$$SE_D = \sqrt{4} = 2$$
$$SE_G = \sqrt{10} \approx 3.2$$

- Errors scale with variables, so we can use fractional errors

$$\frac{SE_D}{C_D} = 0.5$$

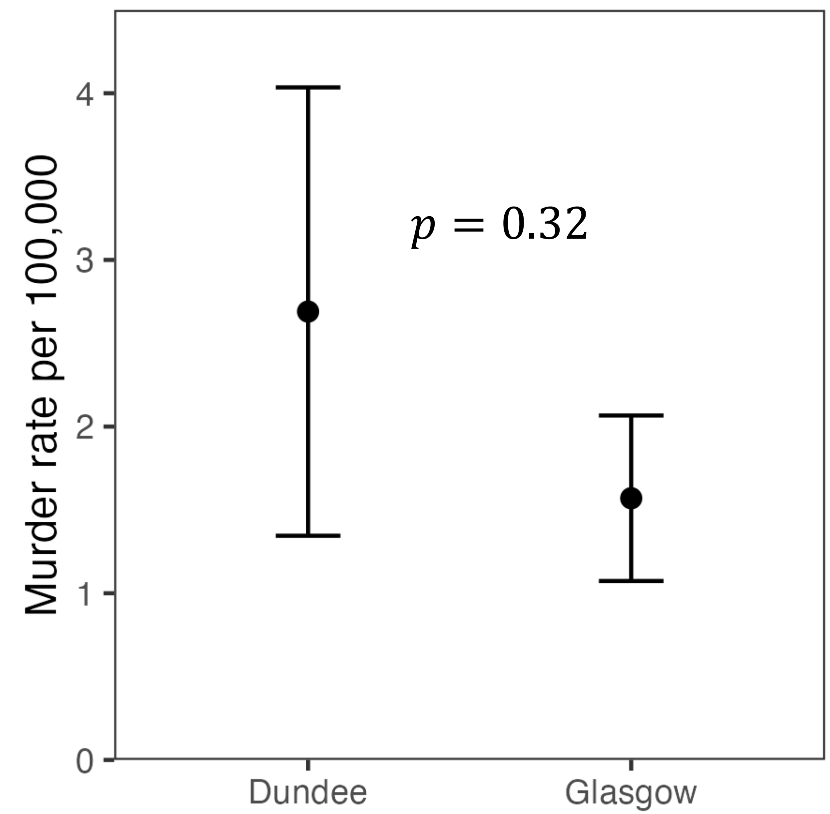
$$\frac{SE_G}{D_G} \approx 0.32$$

- and apply them to murder rate

$$\Delta R_D = 2.69 \times 0.5 = 1.34$$

$$\Delta R_G = 1.57 \times 0.32 = 0.50$$

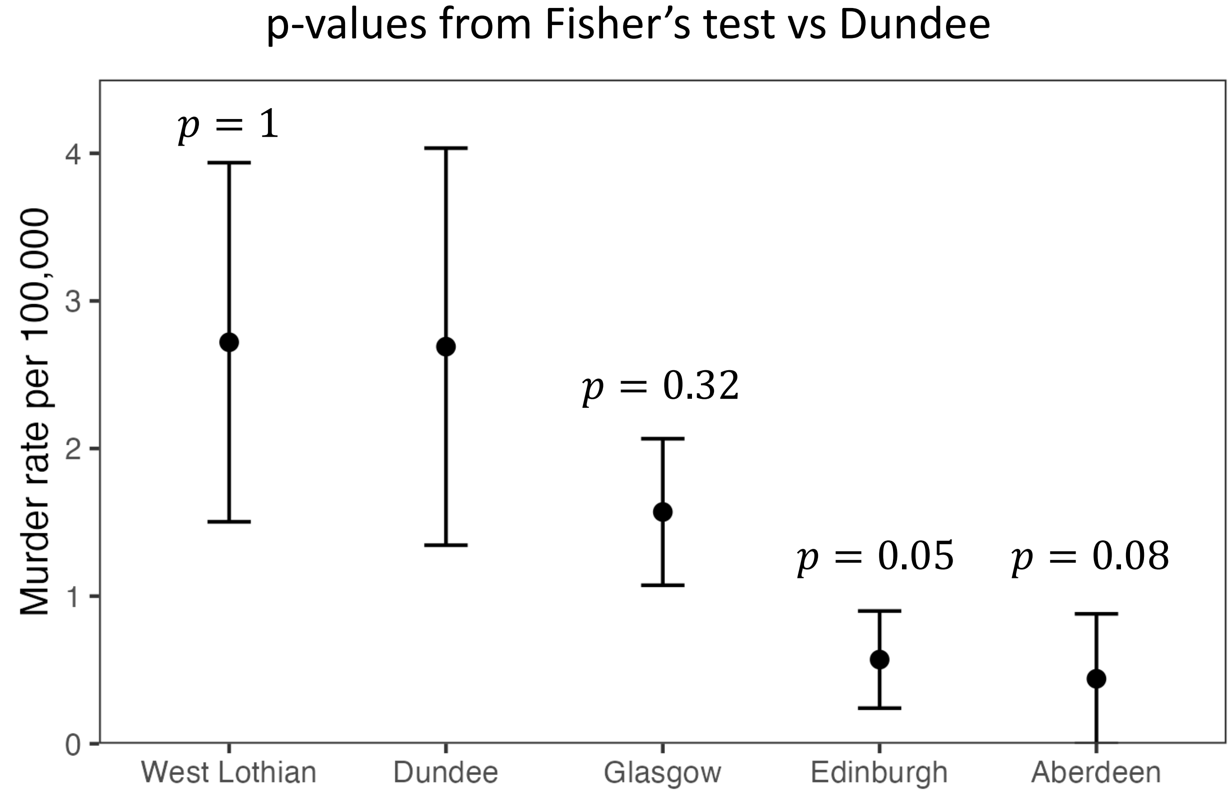
p-value from Fisher's test



Exercise: is Dundee a murder capital of Scotland?

City	Murders	Per 100,000
Dundee	4	2.69
West Lothian	5	2.72
Glasgow	10	1.57
Edinburgh	3	0.57
Aberdeen	1	0.44

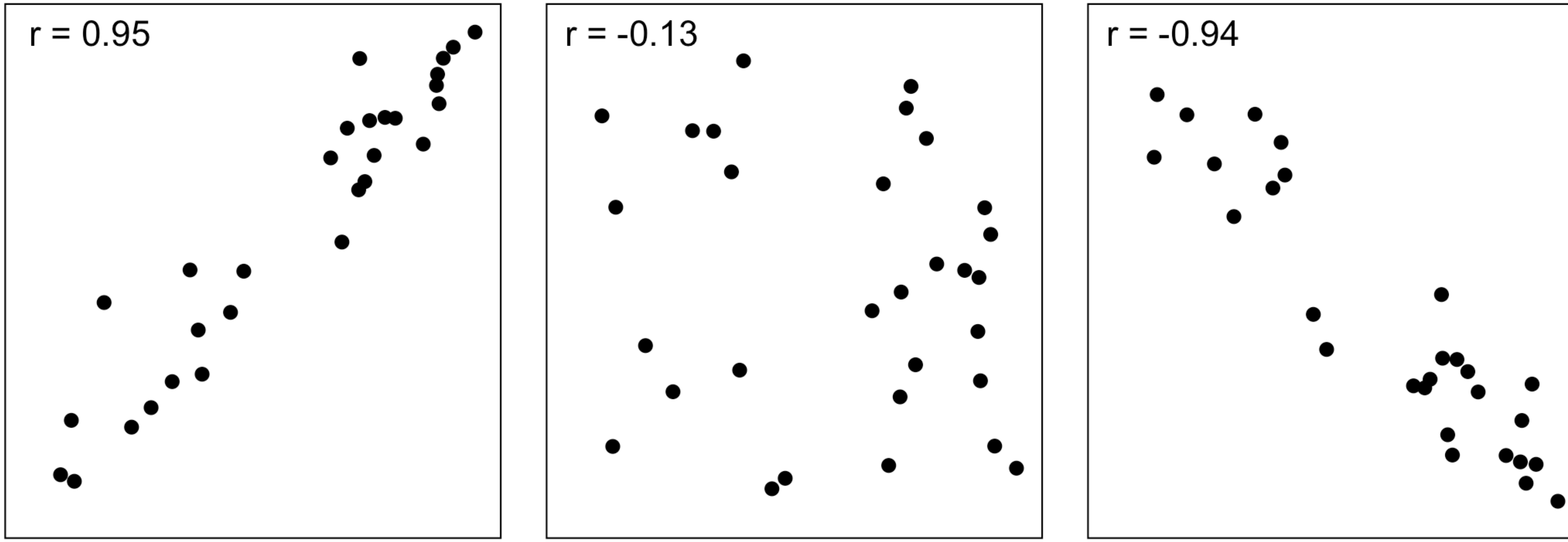
Murder rate elsewhere:
London: 1.4
Cape Town, South Africa: 64
Tijuana, Mexico: 105



> "Misunderstanding of probability may be the greatest of all general impediments to scientific literacy"

Correlation coefficient

Correlation coefficient



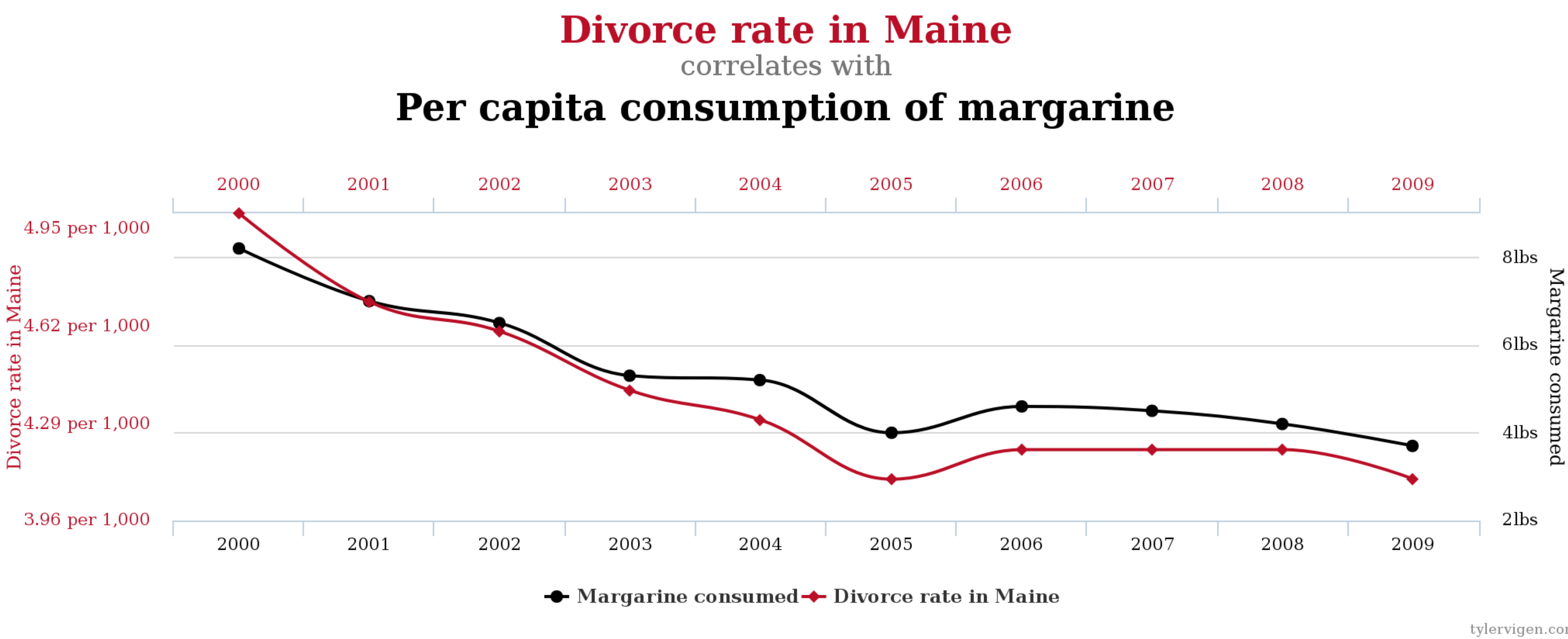
- Two samples: $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - M_x}{SD_x}\right)\left(\frac{y_i - M_y}{SD_y}\right) = \frac{1}{n-1}\sum_{i=1}^{n} Z_{xi}Z_{yi}$$

where Z is a “Z-score”

Correlation doesn't mean causation!

$$r = 0.993$$



tylervigen.com

Statistical estimators

Central point
Mean
Geometric mean
Harmonic mean
Median
Mode
Trimmed mean

Dispersion
Variance
Standard deviation
Standard error
Mean deviation
Range
Interquartile range
Mean difference

Symmetry
Skewness
Kurtosis

Dependence
Pearson's correlation
Rank correlation
Distance

Slides available at
https://dag.compbio.dundee.ac.uk/training/Statistics_lectures.html