Error analysis in biology

Marek Gierliński
Division of Computational Biology

Hand-outs available at http://is.gd/statlec

Oxford Latin dictionary

error ~ōris, *m.* [ERRO[1] + -OR]

1 The act or fact of travelling on an uncertain or devious course, wandering about, roaming, etc. **b** (of things); (esp. of unsteady movements of the head or eyes). **c** the devious and perplexing course of a labyrinth or sim.

2 Uncertainty of mind, doubt, perplexity.

3 A deviation from one's path, going astray.

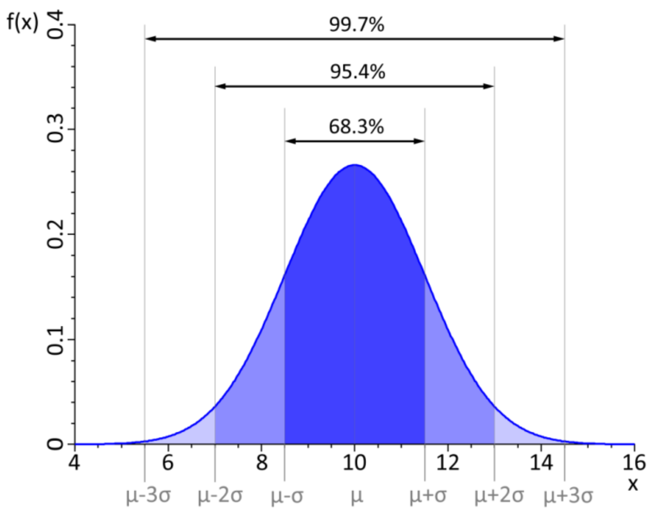
4 A derangement of the mind.

5 A mistake or mistaken condition, error (in thought or action).

6 A departure from right principles, moral lapse or sim. (usu. by implication venial).

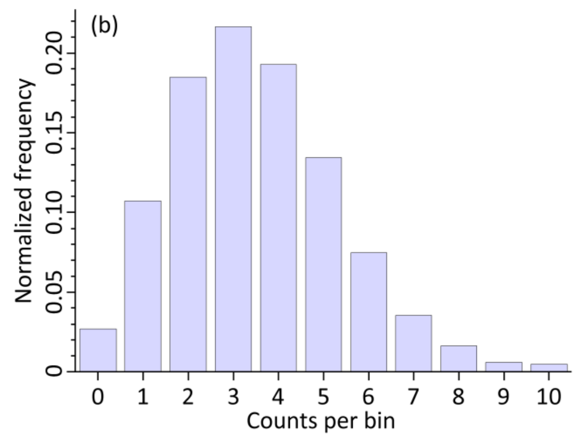
Previously on Errors…

- Random variable: numerical outcome of an experiment
- Probability distribution: how random values are distributed
- Discrete and continuous probability distributions



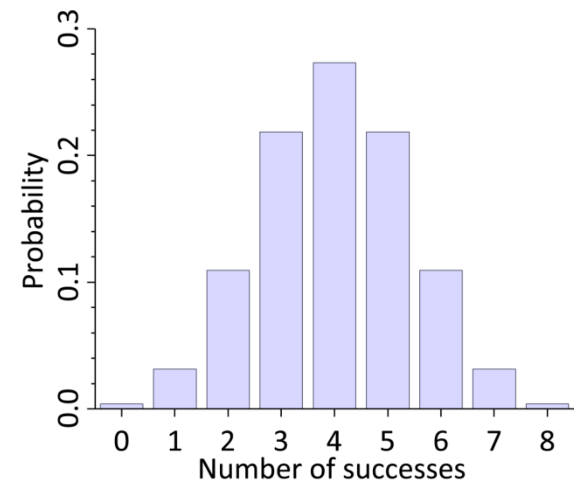
Gaussian (normal) distribution

- very common
- 95% probability within $\mu \pm 1.96\sigma$



Poisson (count) distribution

- random and independent events
- mean = variance
- approximates Gaussian for large n



Binomial distribution

- probability of k successes out of n trials
- toss a coin
- approximates Gaussian for large n

Example

- Take one cuvette with bacterial culture
- Measure optical density (OD600)
- Result: 0.37
- *Reading error*

- Take five cuvettes and find mean OD600
- Results 0.42
- *Sampling error*

- These are examples of **measurement errors**

2. Measurement errors

"If your experiment needs statistics, you ought to have done a better experiment"

Ernest Rutherford

Systematic and random errors

Systematic errors

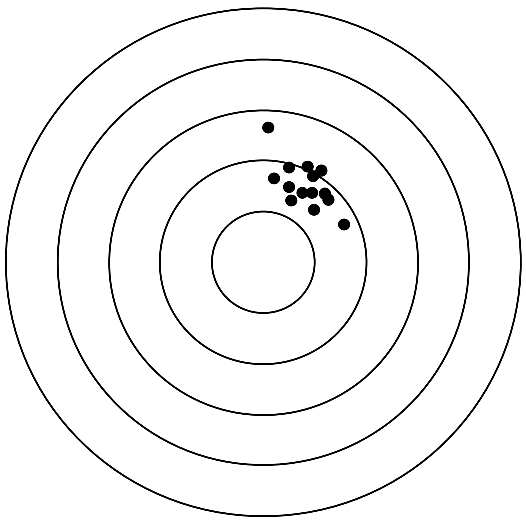
your mistakes

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

Random errors

statistics sucks

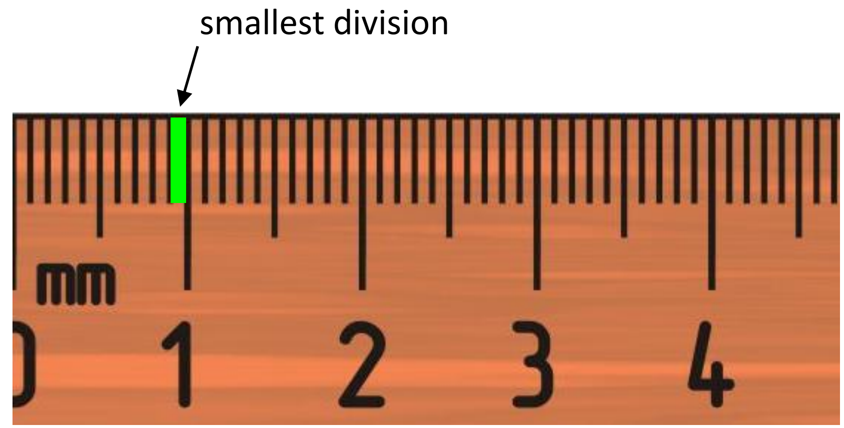
- Reading errors
- Sampling errors
- Counting errors
- Intrinsic variability



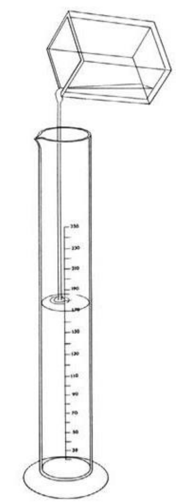
YOU NEED REPLICATES

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!

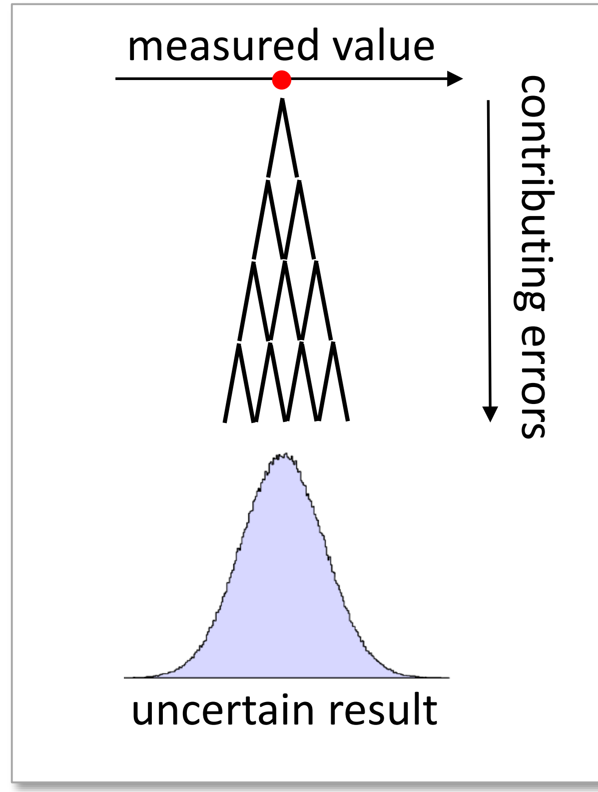


- **Reading error does not take into account biological variability**

Random measurement error

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

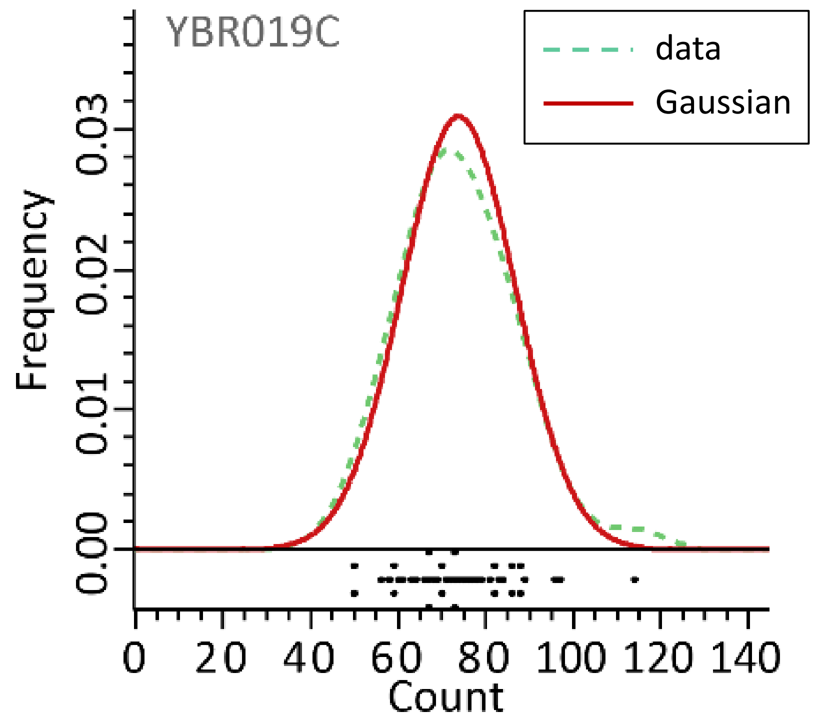
- Uncertainties contributing to the final result
  - volume of the acid sample
  - judgement at which point acid is neutralized
  - volume of NaOH solution used at this point
  - accuracy of NaOH concentration
    - weight of solid NaOH dissolved
    - volume of water added



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

- Measurement errors are normally distributed

Random measurement error



Gene expression from RNA-seq in 42 replicates

Counting error

- Dilution plating of bacteria

- Found $C = 10$ colonies

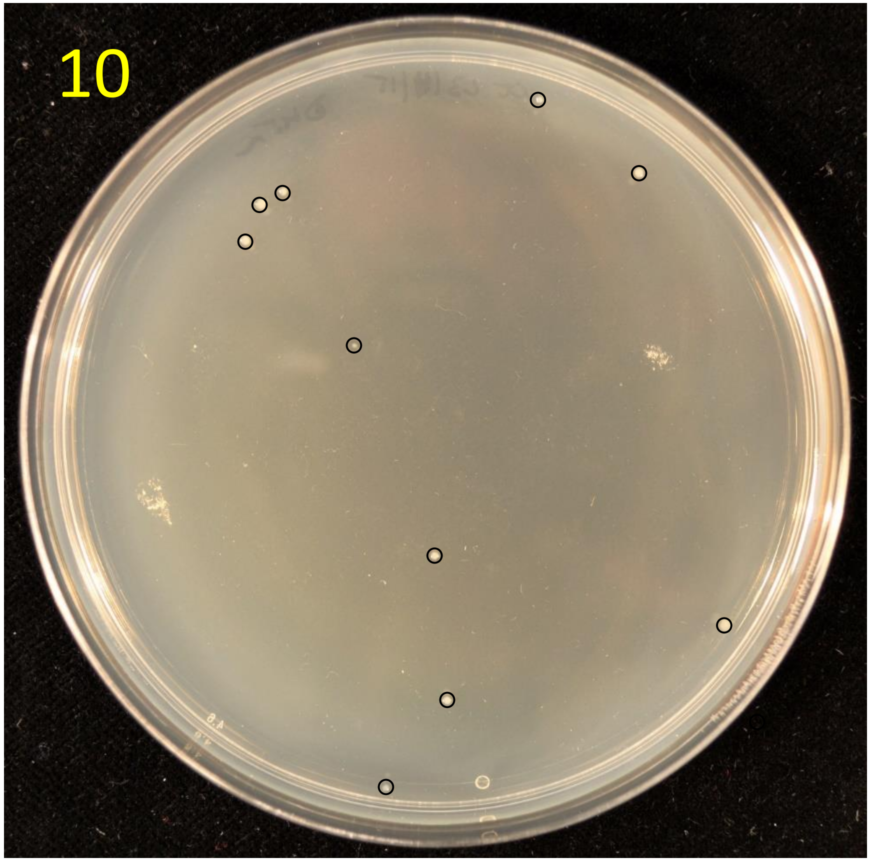
- Counting statistics: Poisson distribution

  $\sigma = \sqrt{\mu}$

- Use standard deviation as error estimate

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

  $C = 10 \pm 3$

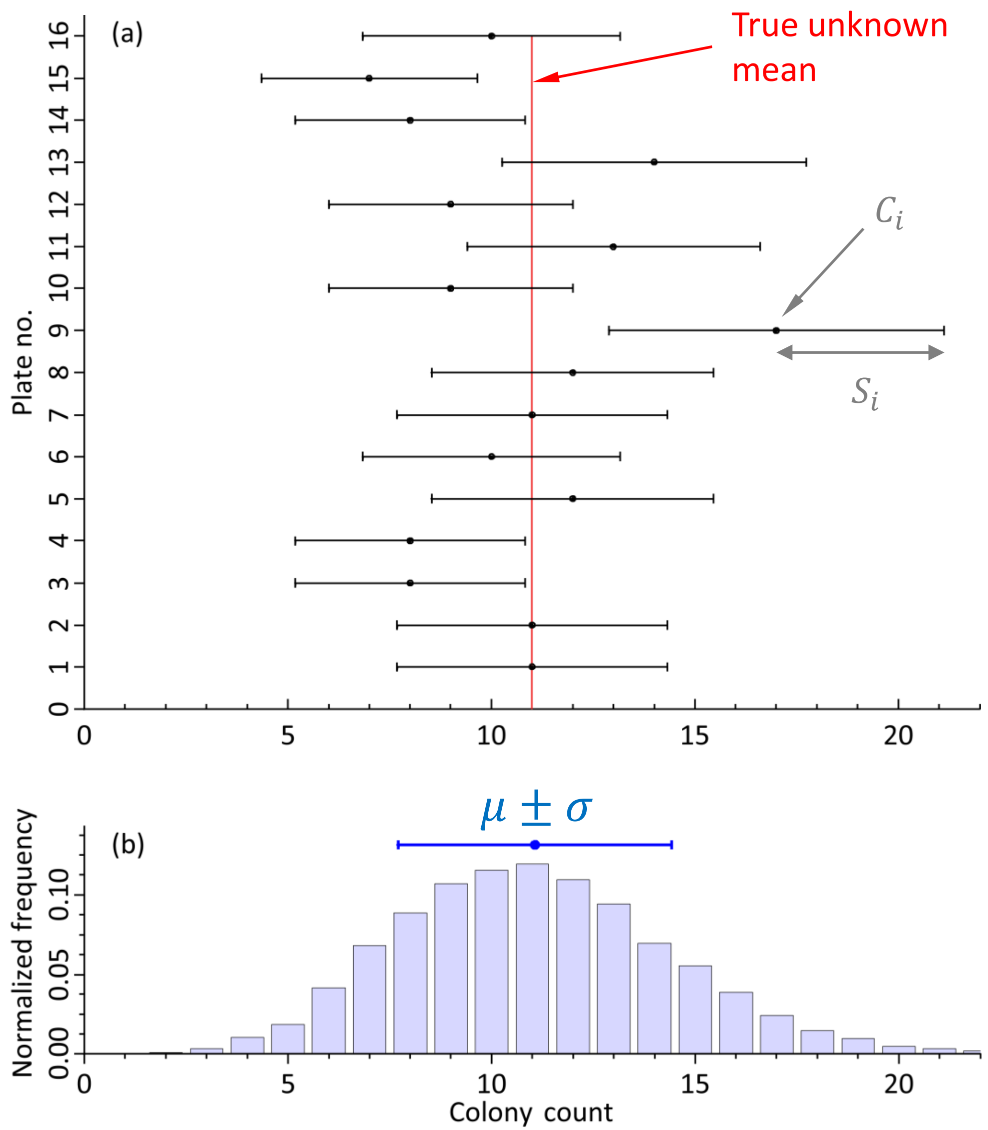


Counting error

- *Gedankenexperiment*
- Measure counts on 10,000 plates

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

- Counting errors, S_i, are similar, but not identical, to σ
- C_i is an estimator of μ
- S_i is an estimator of σ



Exercise: is Dundee a murder capital of Scotland?

- On 2 October 2013 *The Courier* published an article “Dundee is murder capital of Scotland”
- Data in the article (2012/2013):

City	Murders	Per 100,000
Dundee	6	4.1
Glasgow	19	3.2
Aberdeen	2	0.88
Edinburgh	2	0.41

- Compare Dundee and Glasgow
- Find errors on murder rates
- Hint: find errors on murder count first

Exercise: is Dundee a murder capital of Scotland?

City	Murders	Per 100,000
Dundee	6	4.1
Glasgow	19	3.2

$$\Delta C_D = \sqrt{6} \approx 2.4$$
$$\Delta C_G = \sqrt{19} \approx 4.4$$

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

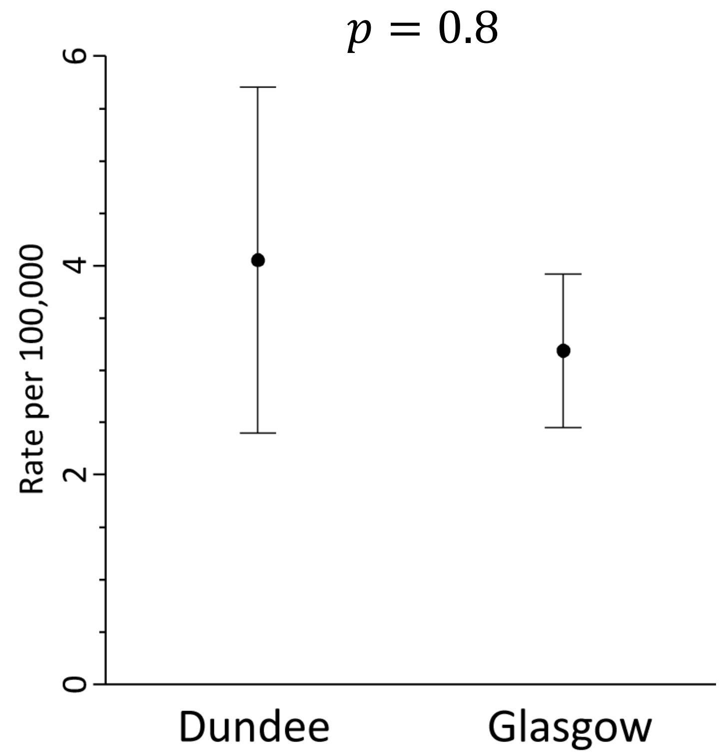
$$\frac{\Delta C_D}{C_D} = 0.41$$

$$\frac{\Delta C_G}{D_G} = 0.23$$

- and apply them to murder rate

$$\Delta R_D = 4.1 \times 0.41 = 1.7$$

$$\Delta R_G = 3.2 \times 0.23 = 0.74$$

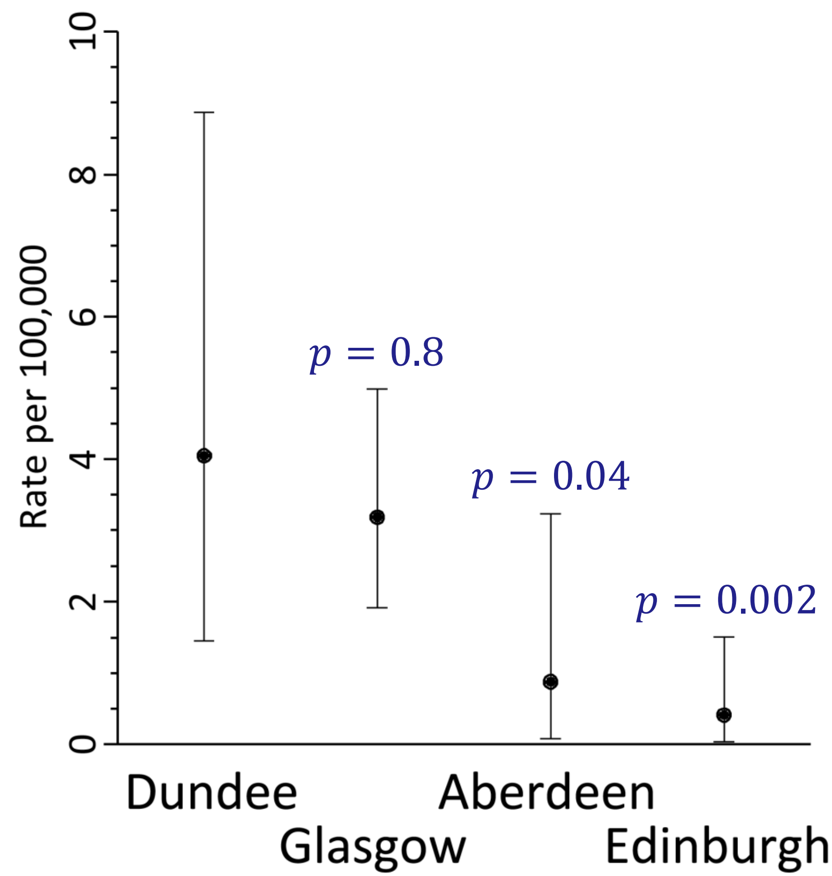


Exercise: is Dundee a murder capital of Scotland?

City	Murders	Per 100,000
Dundee	6	4.1
Glasgow	19	3.2
Aberdeen	2	0.88
Edinburgh	2	0.41

95% confidence intervals (Lecture 4)

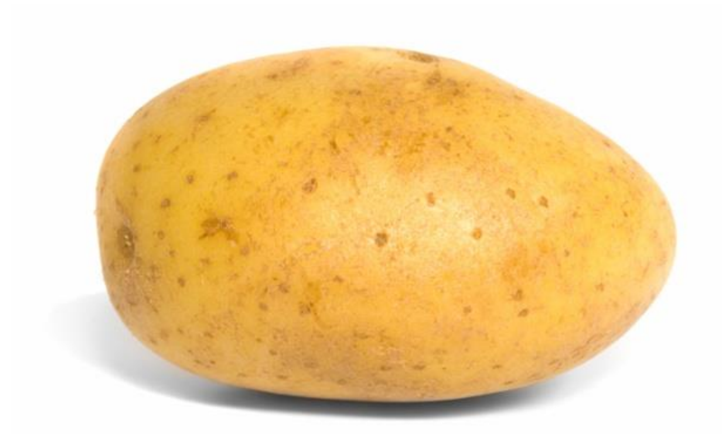
p-values from chi-square test vs Dundee



What’s in the box?

Sampling error

- Repeated measurements give us
  - mean value
  - variability scale
- Sampling from a population
  - Measure the weight of a potato
  - *Sample*: 5 potatoes
  - *Population*: all potatoes
- Small sample size introduces uncertainty



Body weight of 5 potatoes (g)					**Mean (g)**
115	174	178	149	137	**151**
175	162	119	134	66	**131**
194	245	62	177	112	**158**

Measurement errors: summary

- Random measurement errors are expected to be normally distributed

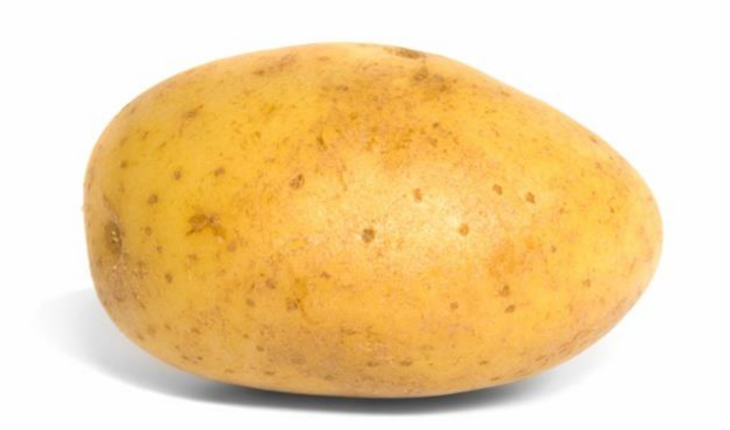
- Some errors can be estimated directly
  - reading (scale, gauge, digital read-out)
  - counting

- Other uncertainties require replicates (a sample)
  - this introduces sampling error

Example

- Weight of 5 potatoes
- This is a **sample**
- We can find
  - mean = 150 g
  - median = 150 g
  - standard deviation = 26 g
  - standard error = 12 g

- These are examples of **statistical estimators**



No.	Weight (g)
1	115
2	174
3	178
4	149
5	137

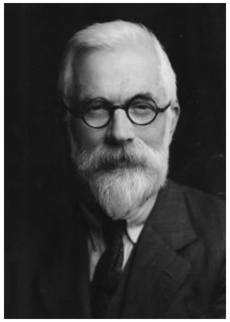
3. Statistical estimators

"The average human has one breast and one testicle"

Des MacHale

Population and sample

Sample selection

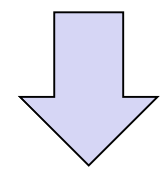


- Terms nicked from social sciences
- Most biological experiments involve sample selection
- Terms “population” and “sample” are not always literal

What is a sample?

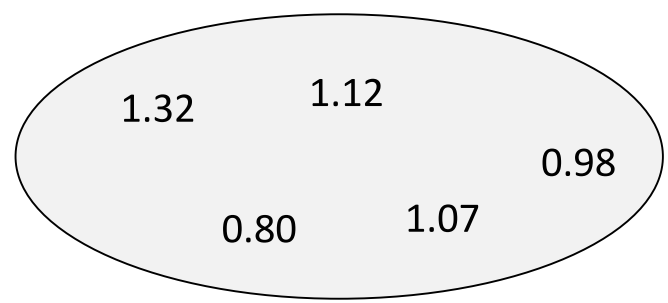
- The term “sample” has different meanings in biology and statistics
- **Biology**: sample is a specimen, e.g., a cell culture you want to analyse
- **Statistics**: sample is (usually) a set of numbers (measurements)
- In these talks: $x_1, x_2, \dots, x_n$

biological samples
(specimens)

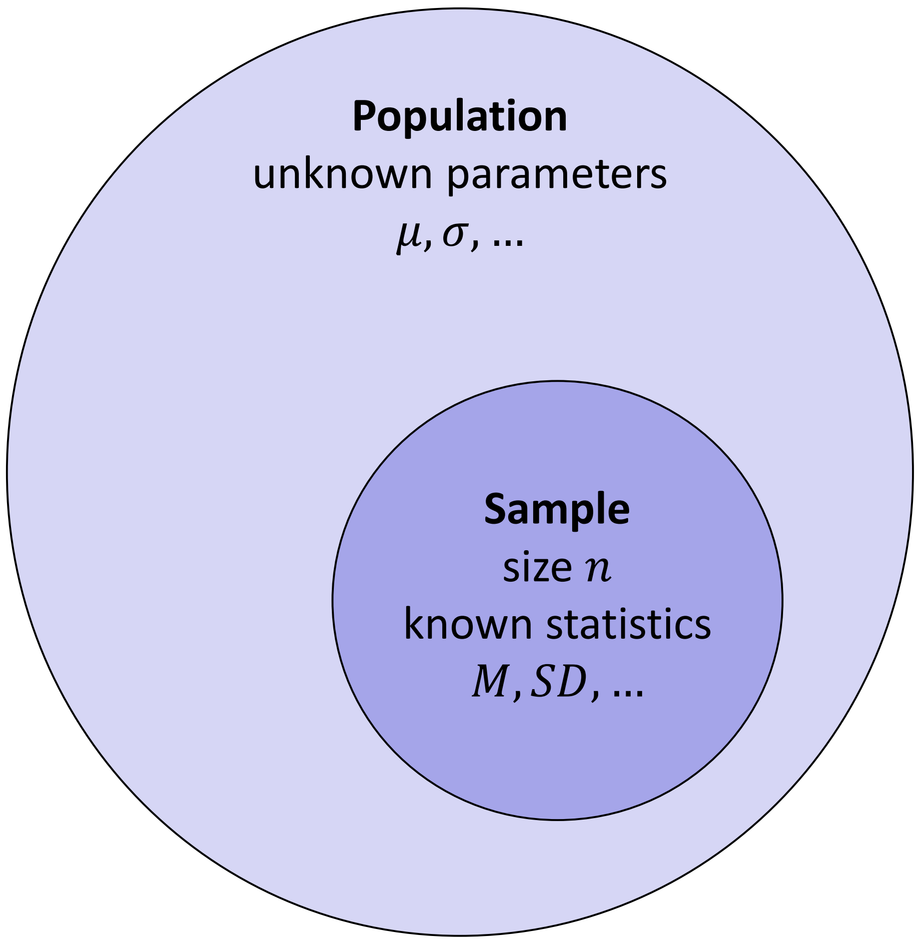


quantification

Statistical sample (set of numbers)



Population and sample



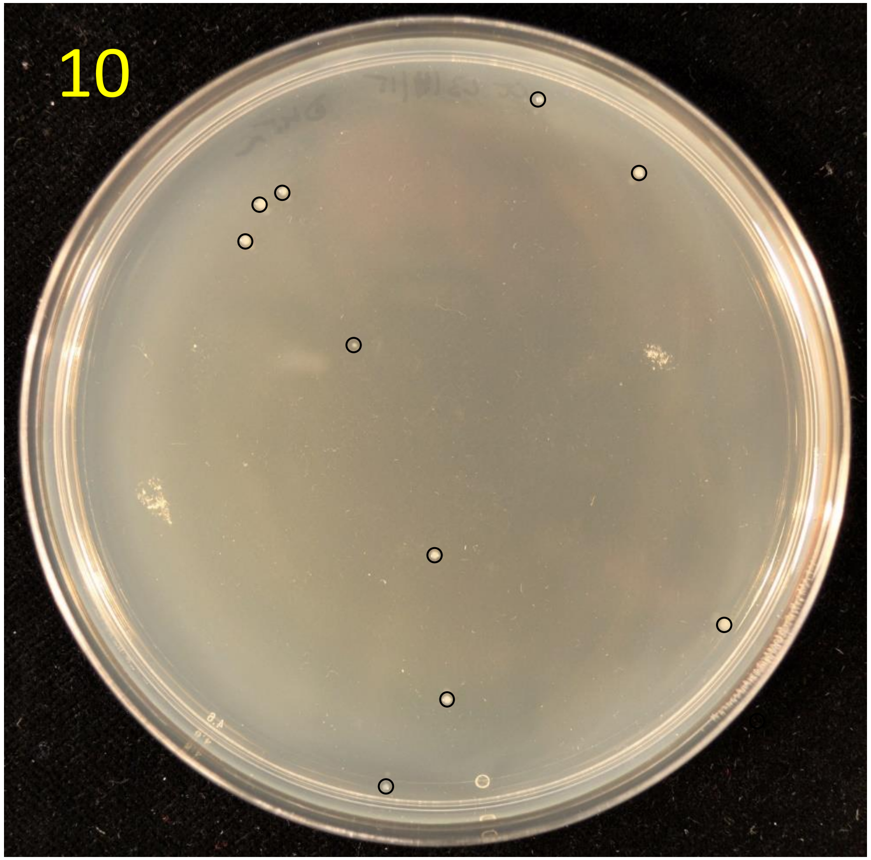
A **parameter** describes a population

A **statistical estimator** (statistic) describes a sample

A statistical estimator approximates the corresponding parameter

Sample size

Dilution plating experiment



10 colonies

What is the sample size?

$$n = 1$$

This sample consists of one measurement: $x_1 = 10$

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

Statistical estimators

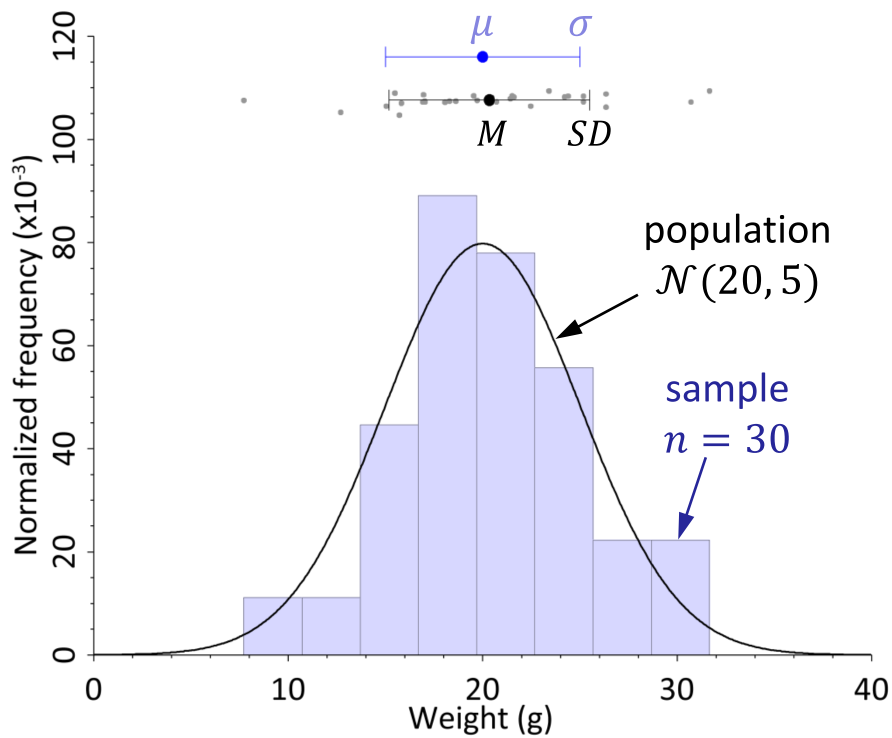
- Statistical estimator is a sample attribute used to estimate a population parameter

- From a sample $x_1, x_2, \dots, x_n$ we can find

$$M = \frac{1}{n}\sum_{i=1}^{n} x_i$$ mean

$$SD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - M)^2}$$ standard deviation

median, proportion, correlation, …

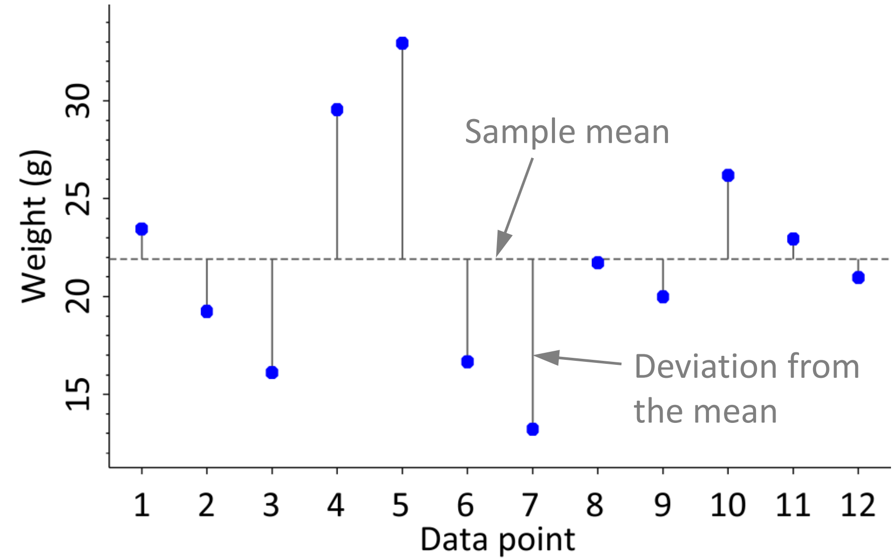


- $n = 30$
- $M = 20.3$ g
- $SD = 5.2$ g
- $SE = 0.94$ g

$M = (20.3 \pm 0.9)$ g

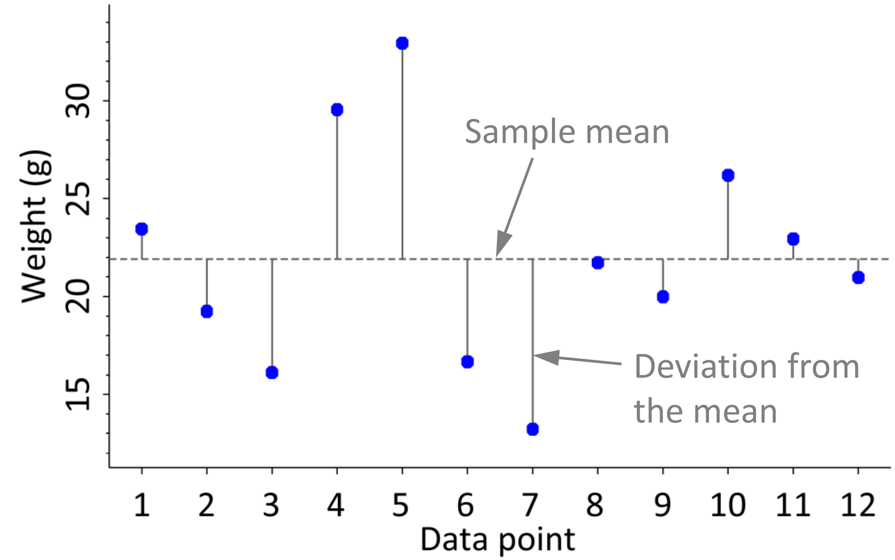
Standard deviation

- Standard deviation is a measure of spread of data points
- Idea:
  - calculate the mean
  - find deviations from the mean of individual points
  - 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 of individual points
  - 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

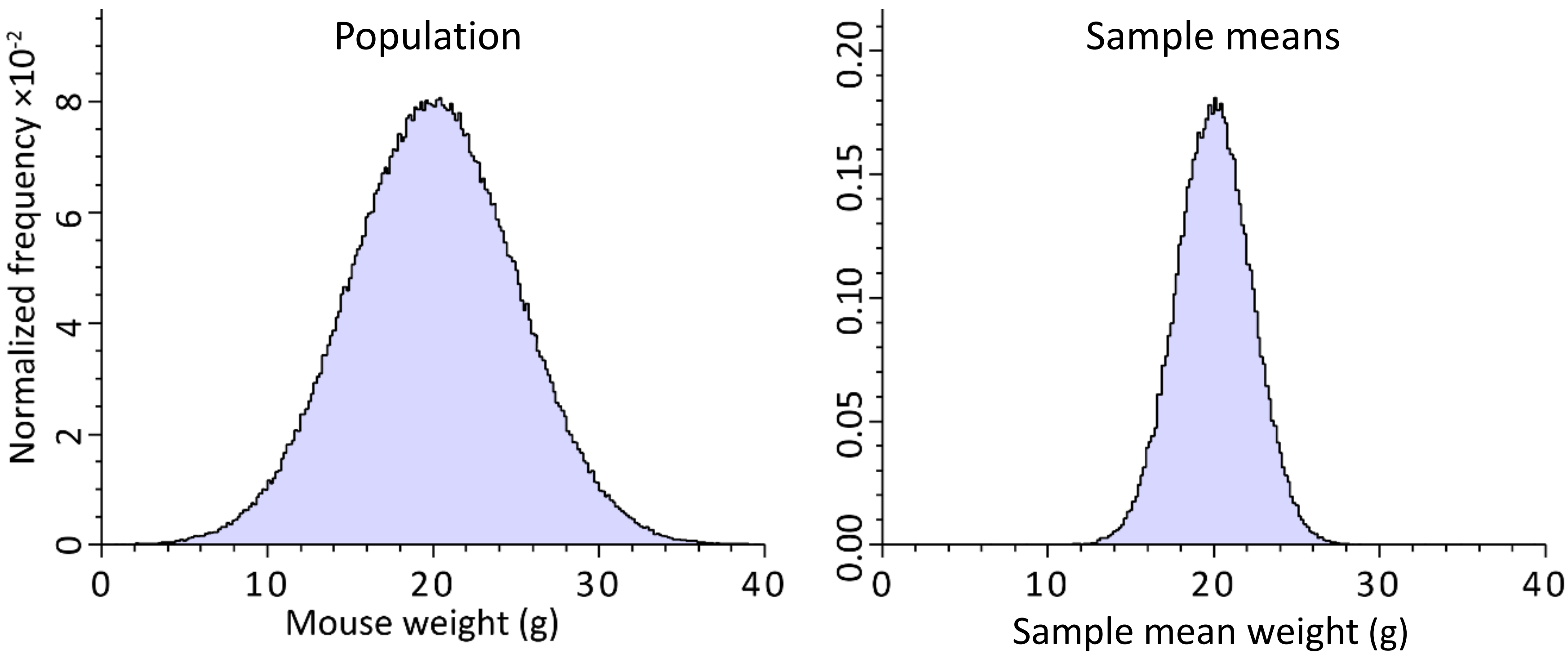
- Mean deviation

$$MD = \frac{1}{n}\sum_i |x_i - M|$$

- doesn't overestimate outliers
- less accurate than SD
- mathematically more complicated
- tradition: use SD

Sampling distribution

Population of mice with Gaussian body weight: $\mu = 20$ g, $\sigma = 5$ g
Draw lots of samples of size $n = 5$



Standard error of the mean

Hypothetical experiment

- 10,000 samples of 5 mice
- Build a distribution of sample means
- Width of this distribution is the true uncertainty of the mean

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

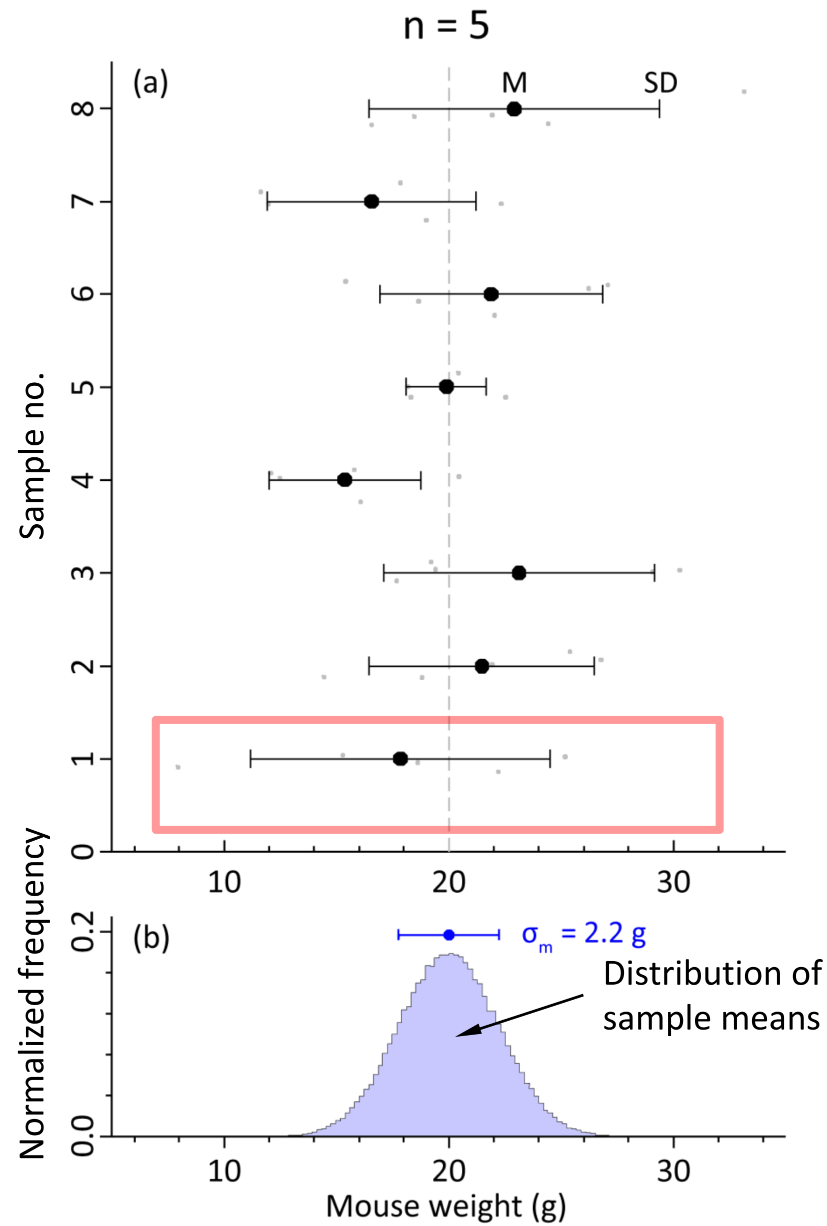
Real experiment

- 5 mice
- Measure body mass:

  7.9, 15.3, 18.5, 22.4, 25.3 g
- Find standard error

$$SE = \frac{SD}{\sqrt{n}} = 3.0\text{ g}$$

SE is an approximation of σ_m



Standard error of the mean

Hypothetical experiment

- 10,000 samples of 30 mice
- Build a distribution of sample means
- Width of this distribution is the true uncertainty of the mean

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

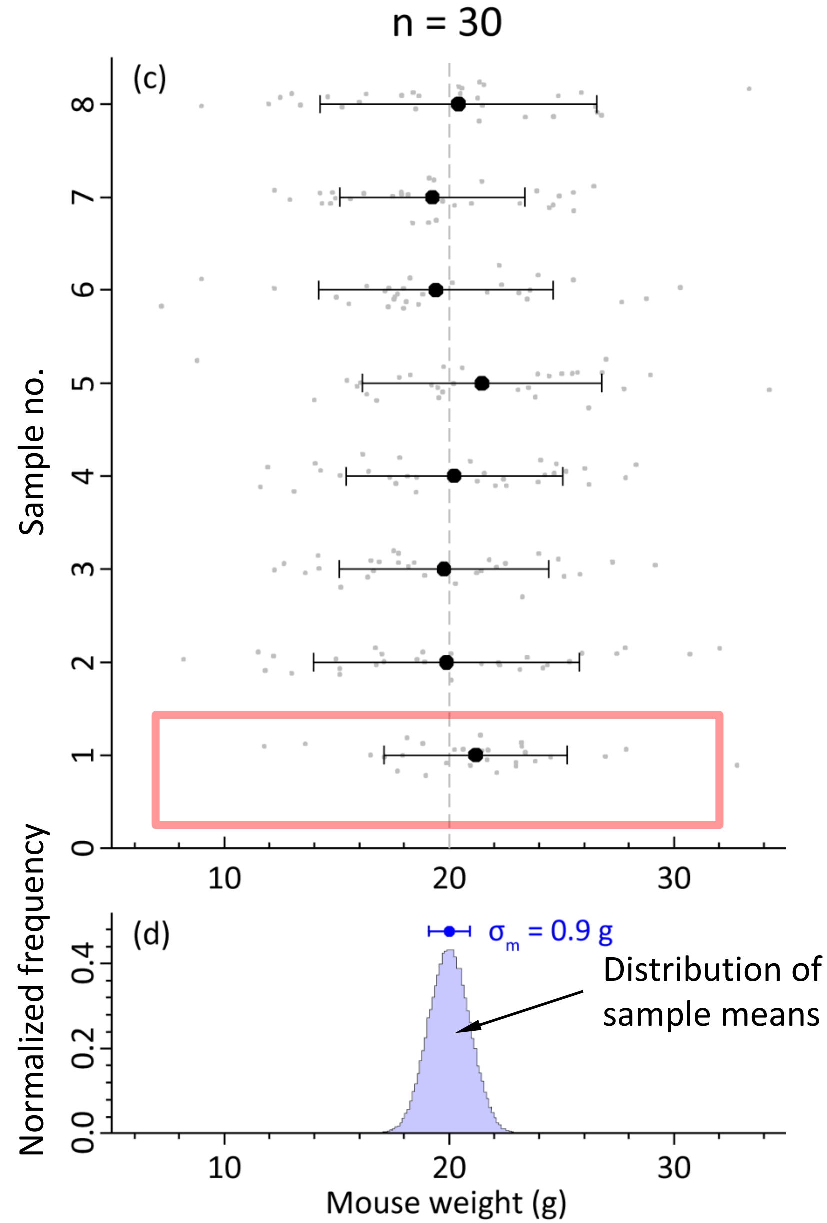
Real experiment

- 30 mice
- Measure body mass:

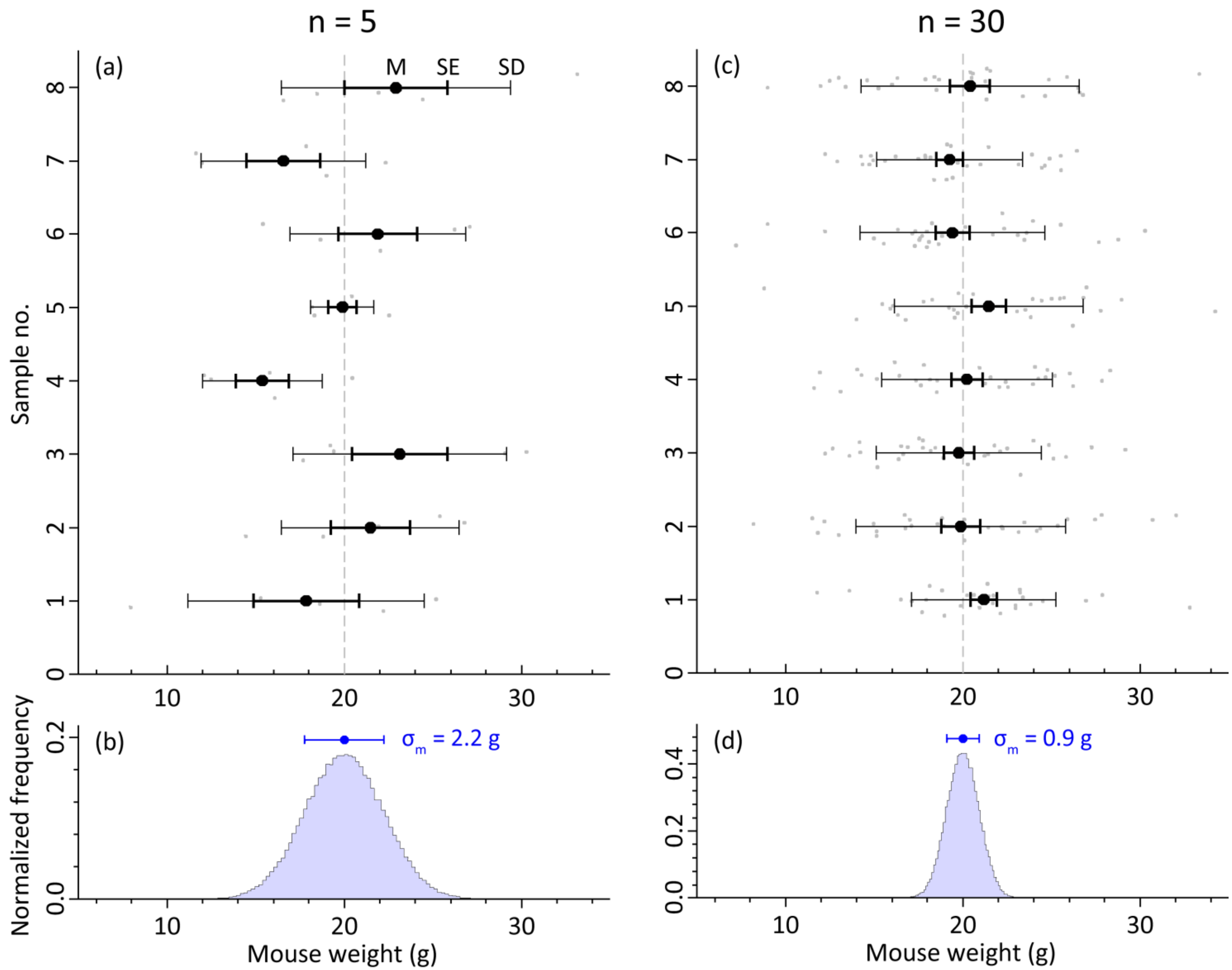
  11.6, 13.7, …, 32.8 g
- Find standard error

$$SE = \frac{SD}{\sqrt{n}} = 0.8 \text{ g}$$

SE is an approximation of σ_m



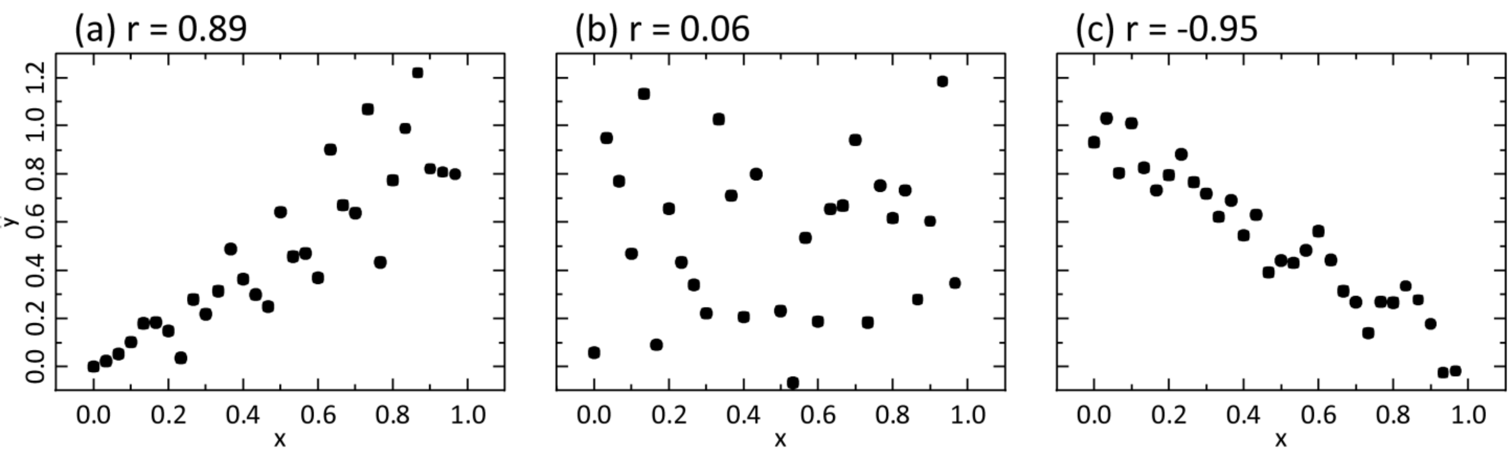
Standard error of the mean



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

Correlation coefficient



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

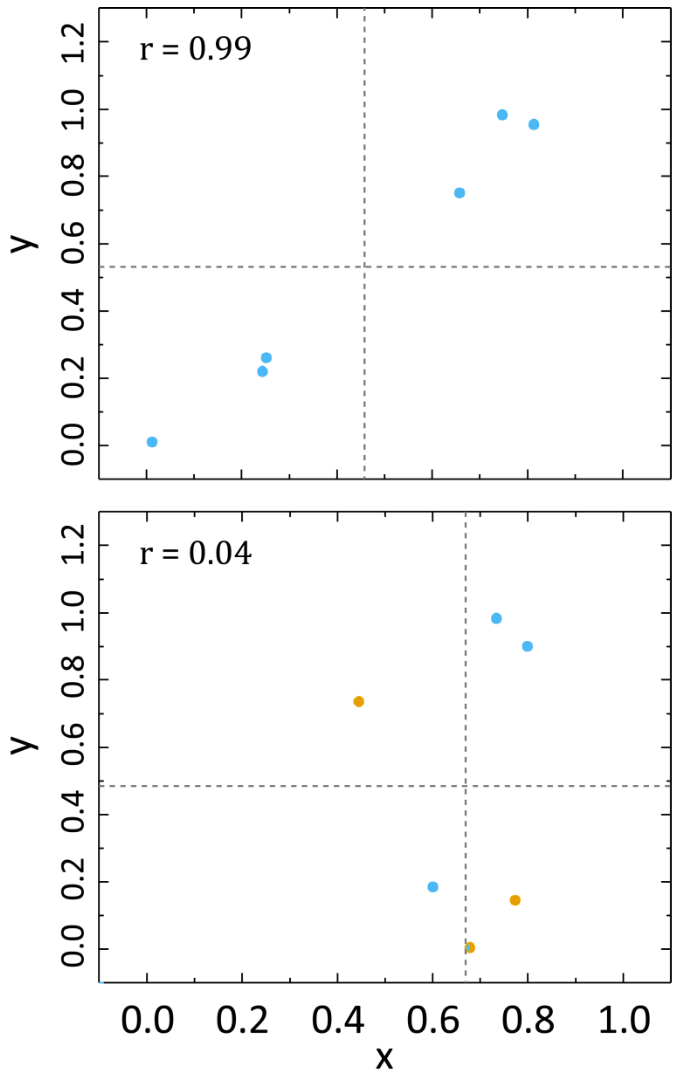
$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - M_x}{SD_x} \right) \left(\frac{y_i - M_y}{SD_y} \right) = \frac{1}{n-1} \sum_{i=1}^{n} Z_{xi} Z_{yi}$$

where Z is a “Z-score”

- Correlation does not mean causation!

Correlation coefficient: example

$$r = \frac{1}{n-1}\sum_{i=1}^{n} Z_{xi} Z_{yi}$$



x	y	Z_x	Z_y	$Z_x Z_y$
0.01	0.01	-1.35	-1.24	1.68
0.24	0.22	-0.64	-0.74	0.48
0.25	0.26	-0.62	-0.64	0.40
0.66	0.75	0.62	0.53	0.33
0.75	0.98	0.89	1.09	0.97
0.81	0.95	1.10	1.02	1.11

$$\sum Z_x Z_y = 4.96$$

x	y	Z_x	Z_y	$Z_x Z_y$
0.45	0.74	-1.72	0.57	-0.98
0.60	0.19	-0.54	-0.72	0.39
0.68	0.00	0.05	-1.14	-0.06
0.73	0.98	0.47	1.14	0.54
0.77	0.15	0.77	-0.81	-0.63
0.80	0.90	0.96	0.95	0.92

$$\sum Z_x Z_y = 0.18$$

Statistical estimators

Central point
Mean
Geometric mean
Harmonic mean
Median
Mode
Trimmed mean

Dispersion
Variance
Standard deviation
Mean deviation
Range
Interquartile range
Mean difference

Symmetry
Skewness
Kurtosis

Dependence
Pearson's correlation
Rank correlation
Distance

Hand-outs available at http://is.gd/statlec

Please leave your feedback forms on the table by the door