Error analysis in biology

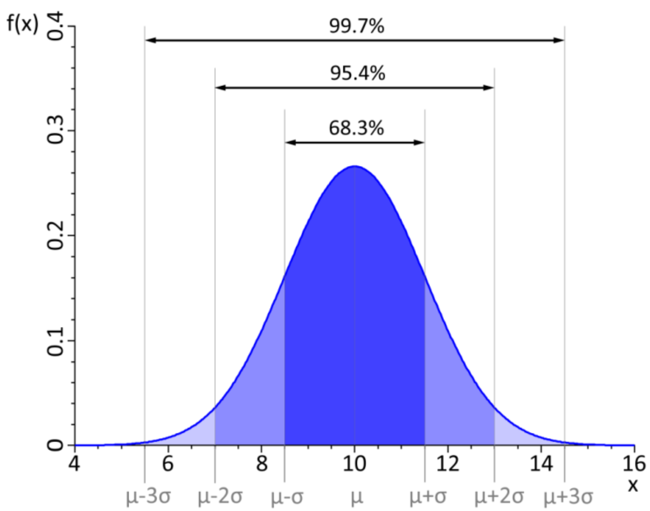
Marek Gierliński
Division of Computational Biology

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

Errors, like straws, upon the surface flow;
He who would search for pearls must dive below
John Dryden (1631-1700)

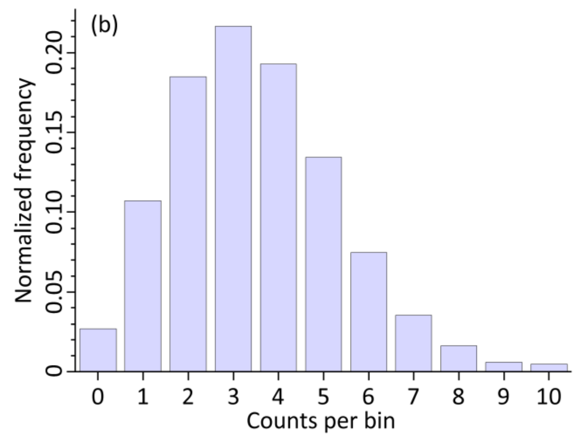
Previously on Errors…

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



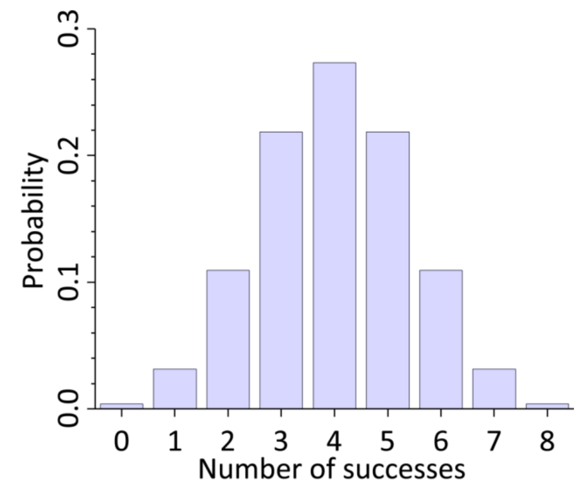
Gaussian (normal) distribution

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



Poisson (count) distribution

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



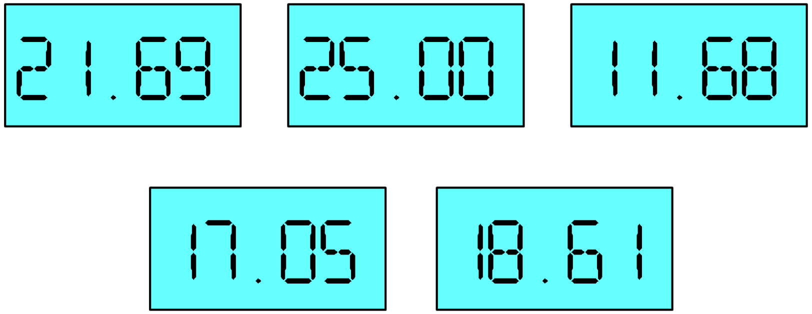
Binomial distribution

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

Example

- Take one mouse and weight it
- Result: 18.21 g
- *Reading error*

- Take five mice and find mean weight
- Results 18.81 g
- *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

Different types of errors

Systematic errors

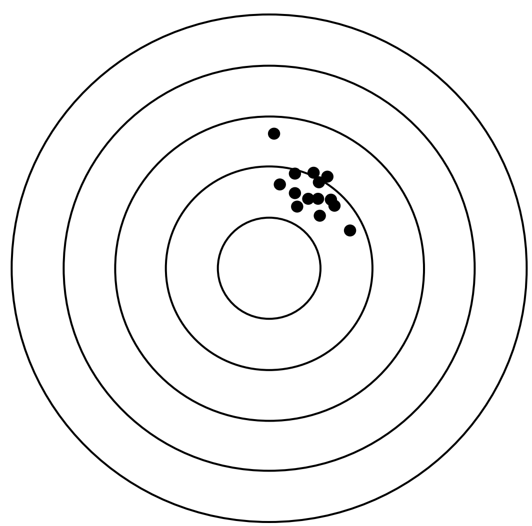
- Incorrect instrument calibration
- Model uncertainties
- Change in experimental conditions
- Mistakes!

Systematic errors can be eliminated in good experiments

Random errors

- Reading errors
- Sampling errors
- Counting errors
- Background noise
- Intrinsic variability
- Sensitivity limits

You can't eliminate random errors, you have to live with them. You can estimate (and reduce) random error by taking multiple measurements

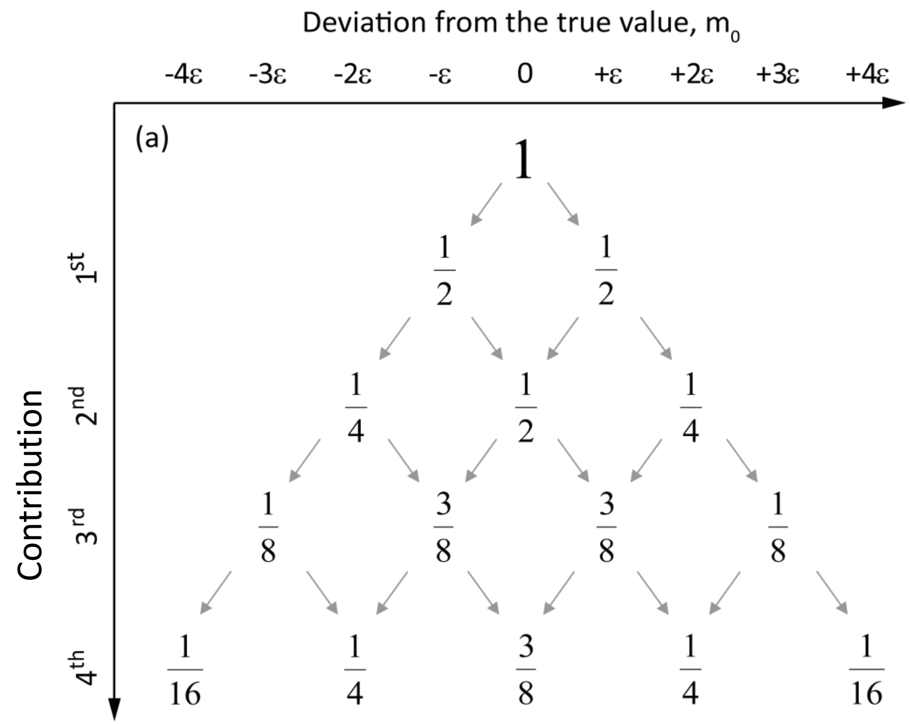


Random measurement error

- Determine the strength of oxalic acid in a sample
- Method: find the volume of NaOH solution required to neutralize a given volume of the acid by observing a phenolphthalein indicator
- 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

A model of random measurement error

- Laplace 1783
- Consider a measurement of a certain quantity
- Its unknown true value is m_0
- Measurement is perturbed by small uncertainties
- Each of them contributes a small **random deviation**, $\pm\varepsilon$, from the measured value



A model of random measurement error

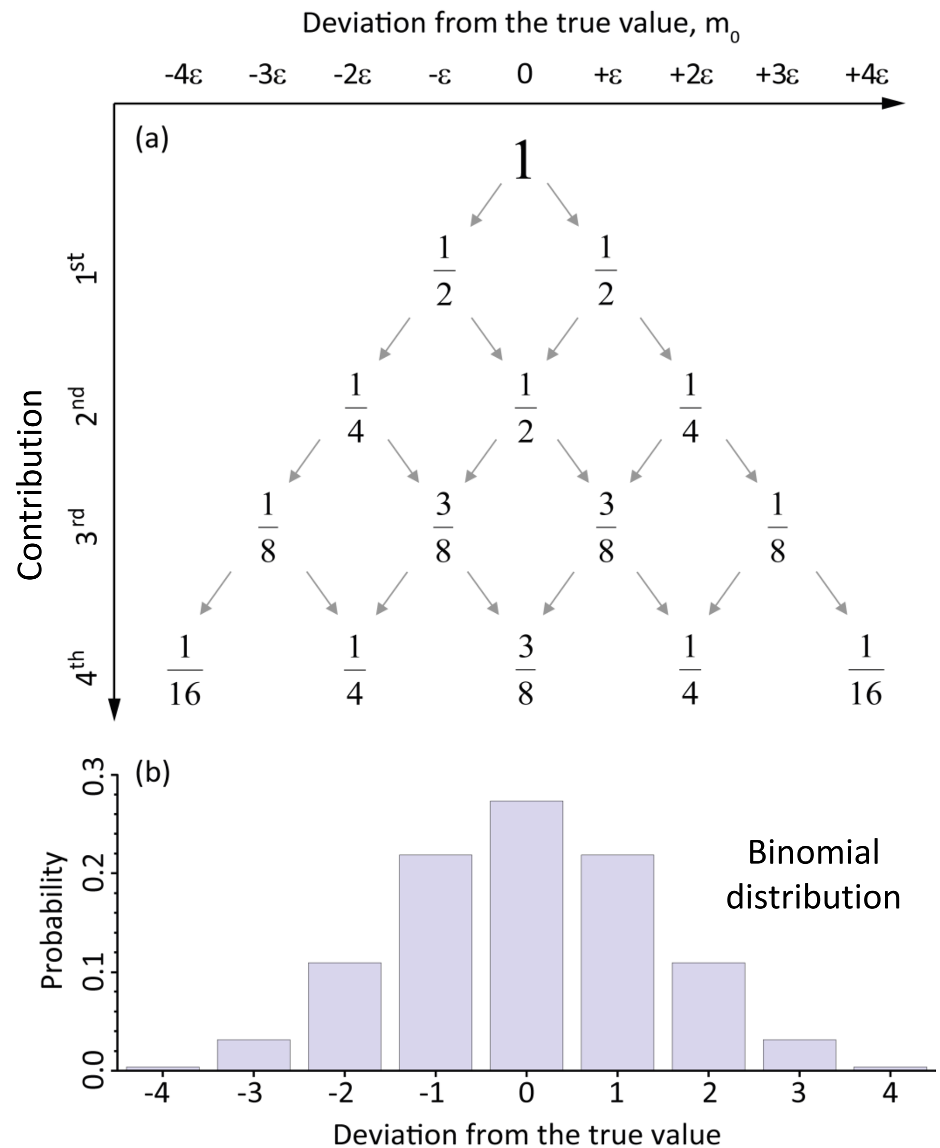
- Laplace 1783

- Consider a measurement of a certain quantity
- Its unknown true value is m_0

- Measurement is perturbed by small uncertainties
- Each of them contributes a small **random deviation**, $\pm\varepsilon$, from the measured value

- This creates binomial distribution
- For large n it approximates Gaussian

- **We expect random measurement errors to be normally distributed**



Test Dynamics

GALTON-BOARD

Biological and technical variability

Biological variability

- Molecular level
- Phenotype variability
- From subject to subject
- Variability in time
- Life is stochastic!

Technical variability

- Random measurement errors
- Accumulation of errors

- In most experiments biological variability dominates
- It is hard to disentangle the two types of variability

Sampling error

- Repeated measurements give us
  - mean value
  - variability scale

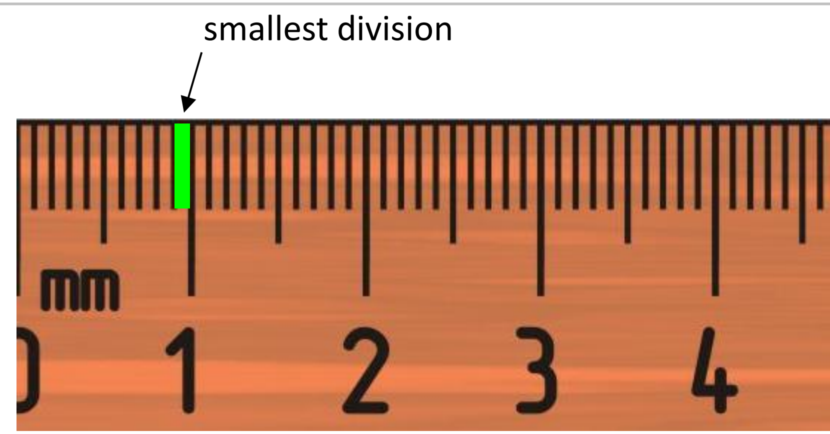
- Sampling from a population
  - Measure the body weight of a mouse
  - *Sample*: 5 mice
  - *Population*: all mice on the planet

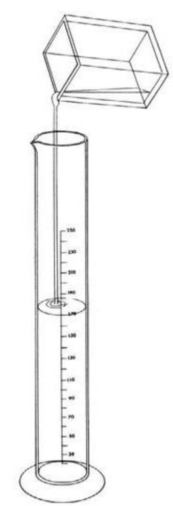
- Small sample size introduces uncertainty

Body weight of 5 mice (g)					**Mean (g)**
20.38	20.73	23.24	15.39	12.58	**18.5**
27.48	12.52	21.95	12.54	21.19	**19.1**
14.73	16.37	28.21	21.18	13.48	**18.9**

Reading error

- When you do one simple measurement using
  - ruler
  - micrometer
  - voltmeter
  - thermometer
  - measuring cylinder
  - stopwatch
- The reading error is ±half of the smallest division
- A ruler with 1-mm scale can give a reading 23±0.5 mm
- 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**





Counting error

- Dilution plating of bacteria

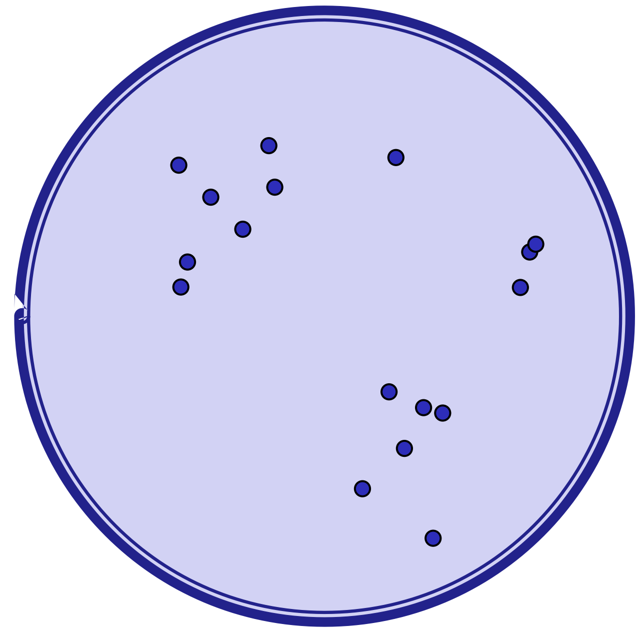
- Counted $C = 17$ colonies on a plate at the 10^{-5} dilution
- Counting statistics: Poisson distribution

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

- Use standard deviation as error estimate

$$S = \sqrt{C} = \sqrt{17} \approx 4$$

$$C = 17 \pm 4$$



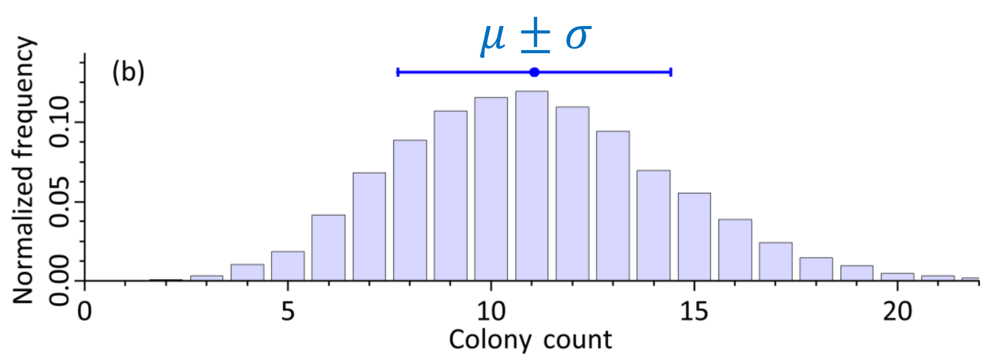
Counting error

- *Gedankenexperiment*
- True mean count, $\mu = 11$

- Measure counts on 10,000 plates (!)
- Plot counts, C_i, and their errors, $S_i = \sqrt{C_i}$
- Plot distribution of counts from 10,000 plates and its mean, μ, and standard deviation, σ

- Counting errors, $S_i = \sqrt{C_i}$ 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

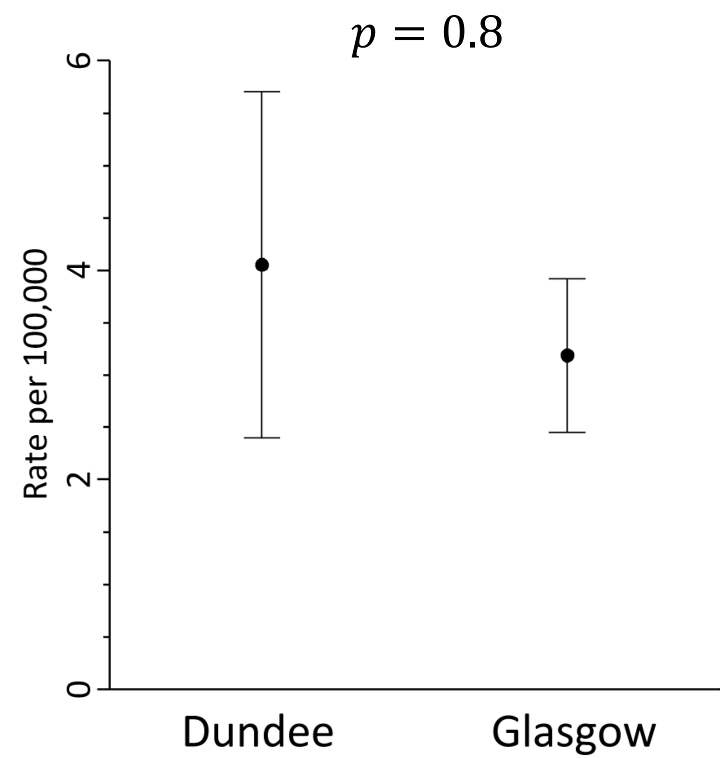
$$\frac{\Delta C_D}{C_D} = 0.41$$

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

- and apply them to murder rate

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

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

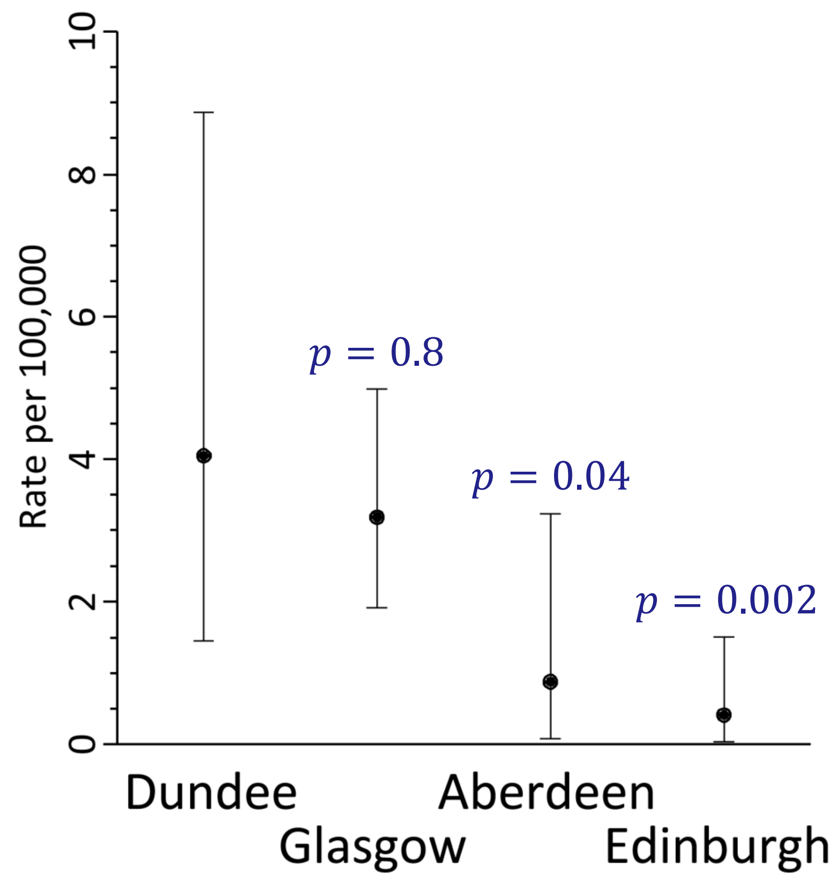


Exercise: is Dundee a murder capital of Scotland?

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



Measurement errors: summary

- Experimental random 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

- Body mass of 5 mice
- This is a **sample**
- We can find
  - mean = 18.8 g
  - median = 18.6 g
  - standard deviation = 5.0 g
  - standard error = 2.2 g

- These are examples of **statistical estimators**

21.69 25.00 11.68

17.05 18.61

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
- Experiment in 5 biological replicates requires 5 biological samples
- After quantification (e.g. protein abundance) we get a set of 5 numbers
- **Statistics**: sample is (usually) a set of numbers (measurements)
- In these talks: $x_1, x_2, \ldots, x_n$

biological samples
(specimens)

quantification

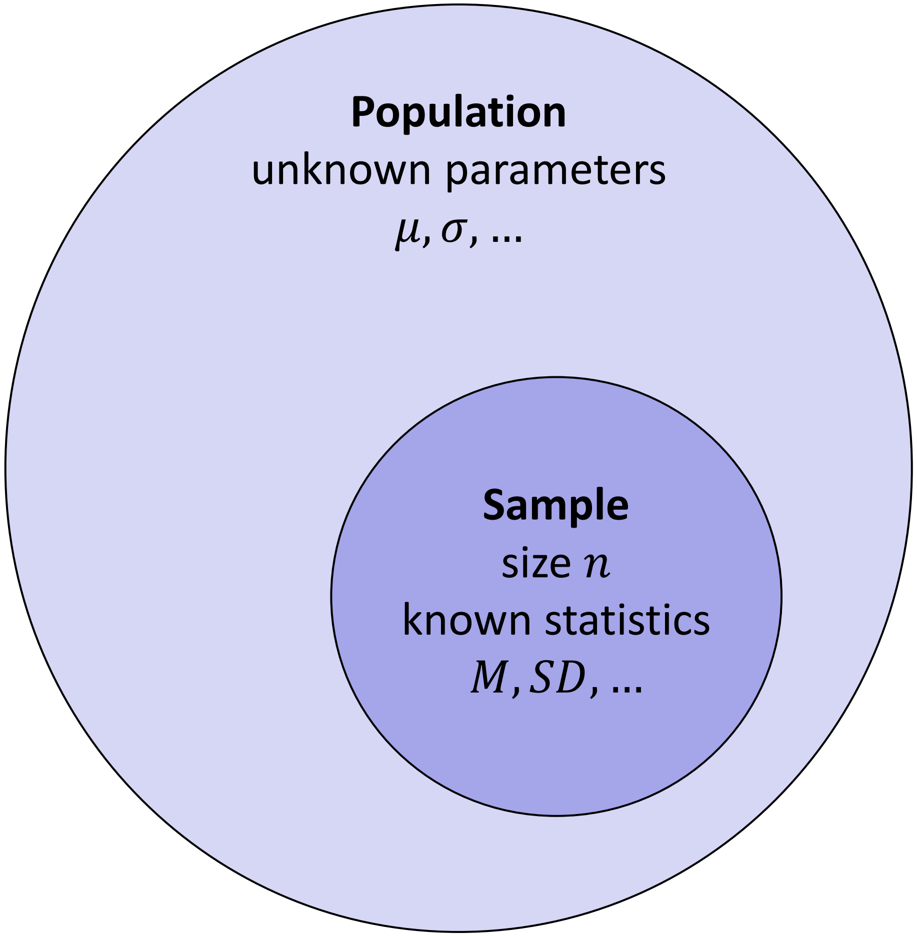
Statistical sample (set of numbers)

1.32 1.12 0.98 0.80 1.07

Population and sample

Population	Sample
Population can be a somewhat abstract concept	Sample is what you get from your experiments
Huge size, impossible to handle	Manageable size, n measurements
▪ all mice on Earth ▪ all people with eczema ▪ all possible measurements of gene expression (infinite population)	▪ 12 mice in a particular experiment ▪ 26 patients with eczema ▪ 5 biological replicates to measure gene expression

Population and sample



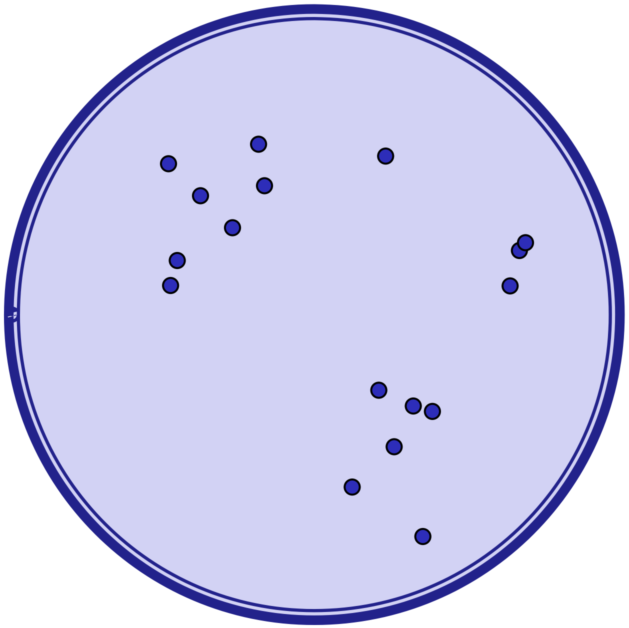
A **parameter** describes a population

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

A statistical estimator approximates the corresponding parameter

Sample size

Dilution plating experiment



17 colonies

What is the sample size?

$$n = 1$$

This sample consists of one measurement: $x_1 = 17$

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 $x_1, x_2, \ldots, x_n$ we can find

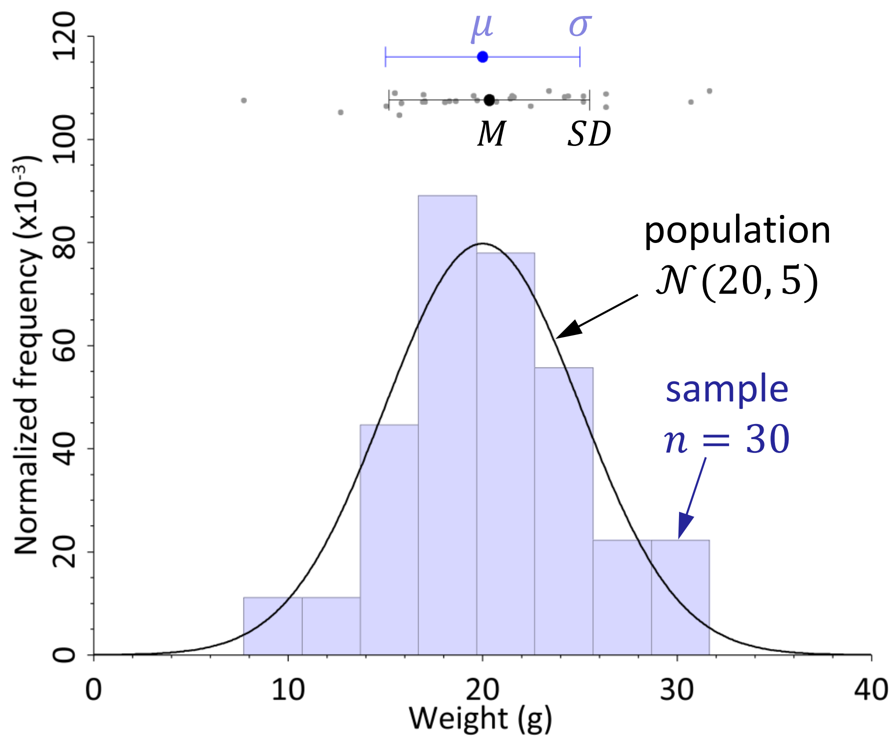
$$M = \frac{1}{n}\sum_{i=1}^{n} x_i$$

mean

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

standard deviation

median, proportion, correlation, …

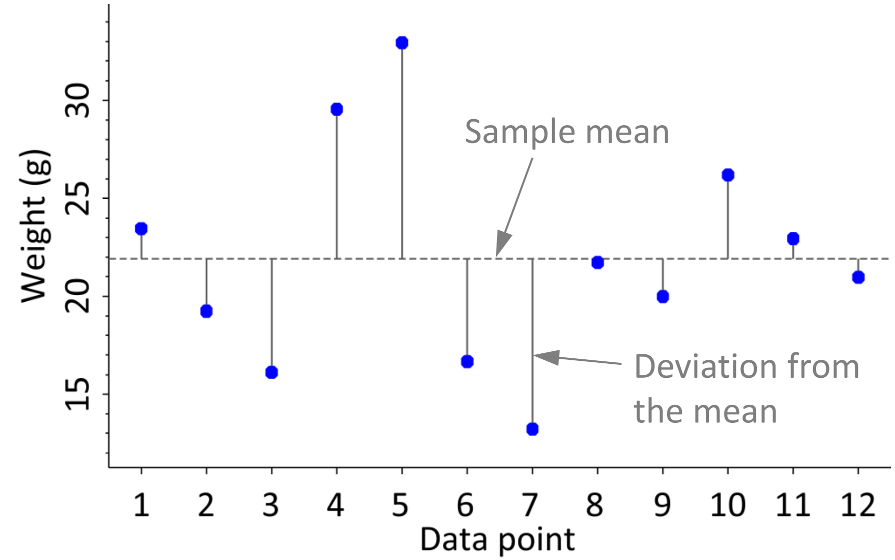


- $n = 30$
- $M = 20.3$ g
- $SD = 5.2$ g
- $SE = 0.94$ g

$$M = (20.3 \pm 0.9)\text{ g}$$

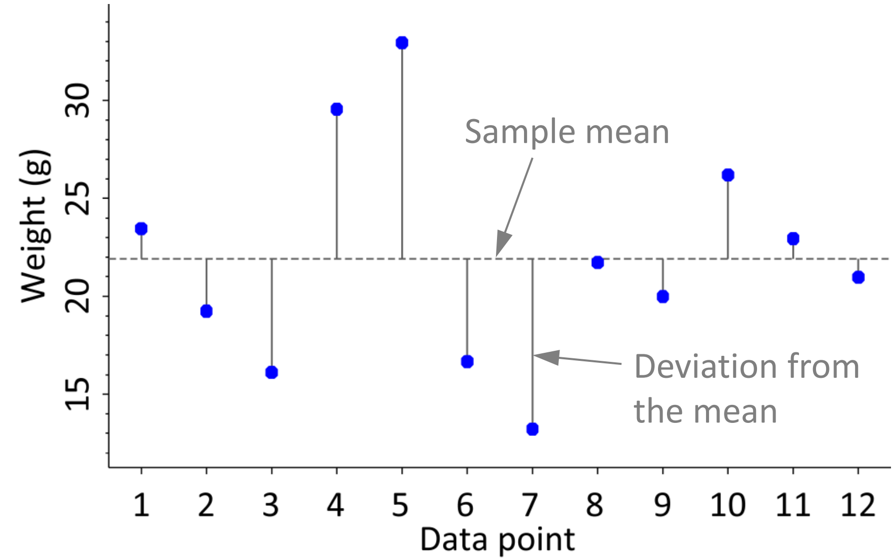
Standard deviation

- 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 $x_1, x_2, \ldots, x_n$

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

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

SD_{n-1}^2 is unbiased estimator of variance

- Mean deviation

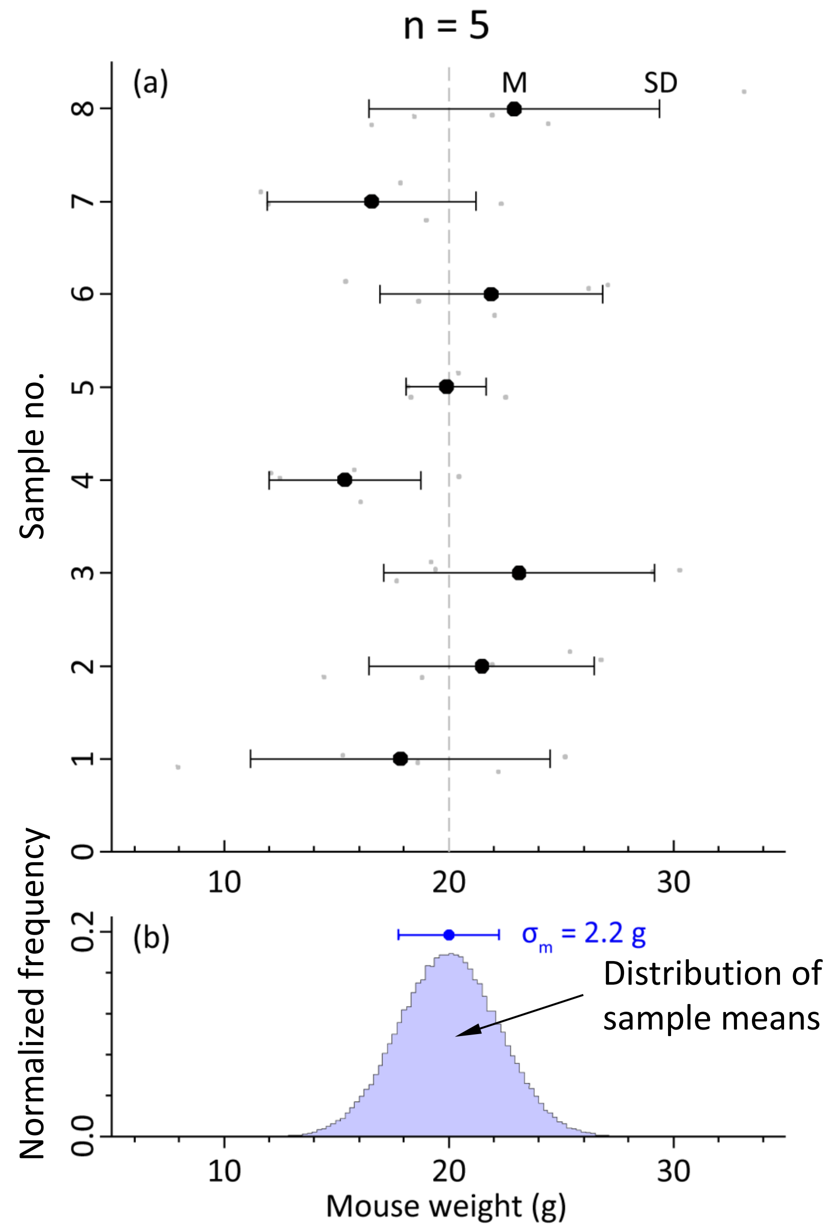
$$MD = \frac{1}{n}\sum_i |x_i - M|$$

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

Standard error of the mean

- *Gedankenexperiment*
- Consider a population of mice with normally distributed body weight with $\mu = 20$ g and $\sigma = 5$ g

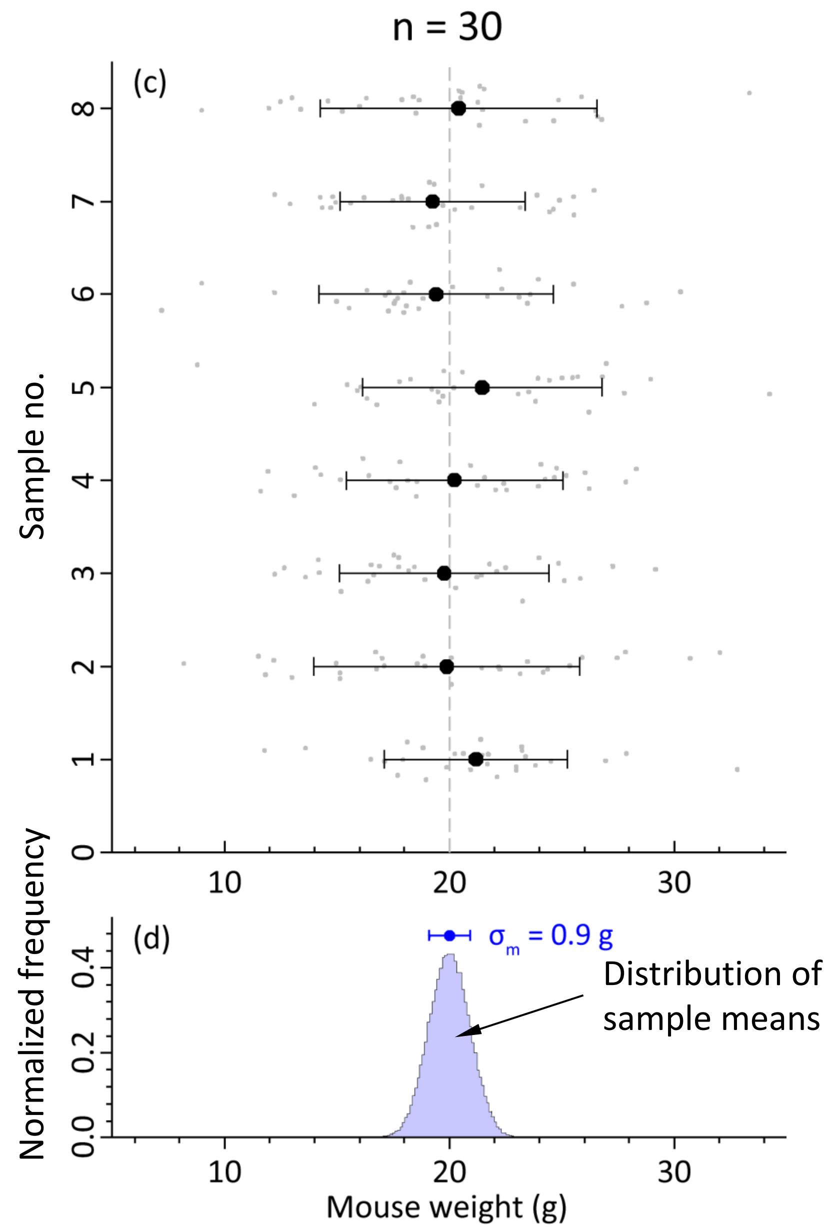
- Take a sample of 5 mice
- Calculate sample mean, M
- Repeat many times
- Plot distributions of sample means



Standard error of the mean

- *Gedankenexperiment*
- Consider a population of mice with normally distributed body weight with $\mu = 20$ g and $\sigma = 5$ g

- Take a sample of 30 mice
- Calculate sample mean, M
- Repeat many times
- Plot distributions of sample means



Standard error of the mean

- Distribution of sample means is called *sampling distribution of the mean*
- The larger the sample, the narrower the sampling distribution

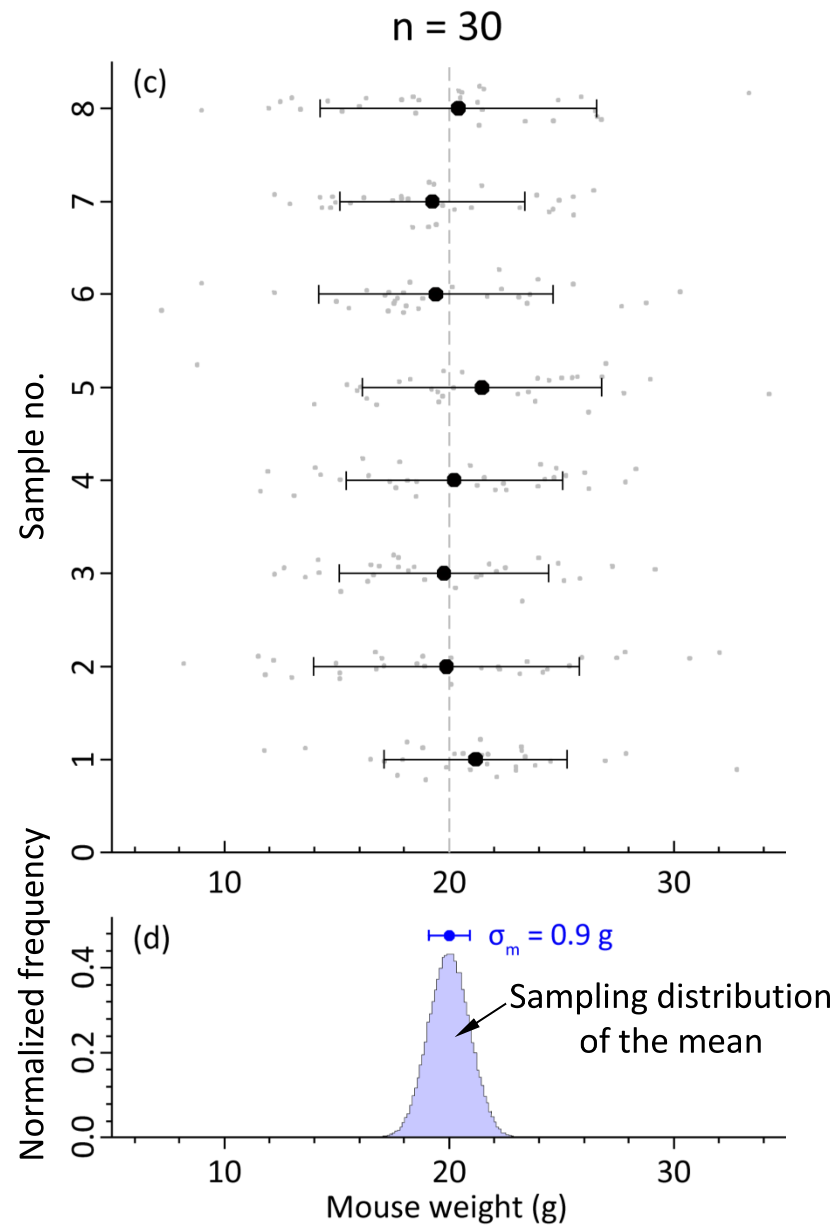
- Sampling distribution is Gaussian, with standard deviation

$$\sigma_m = \frac{\sigma}{\sqrt{n}}$$

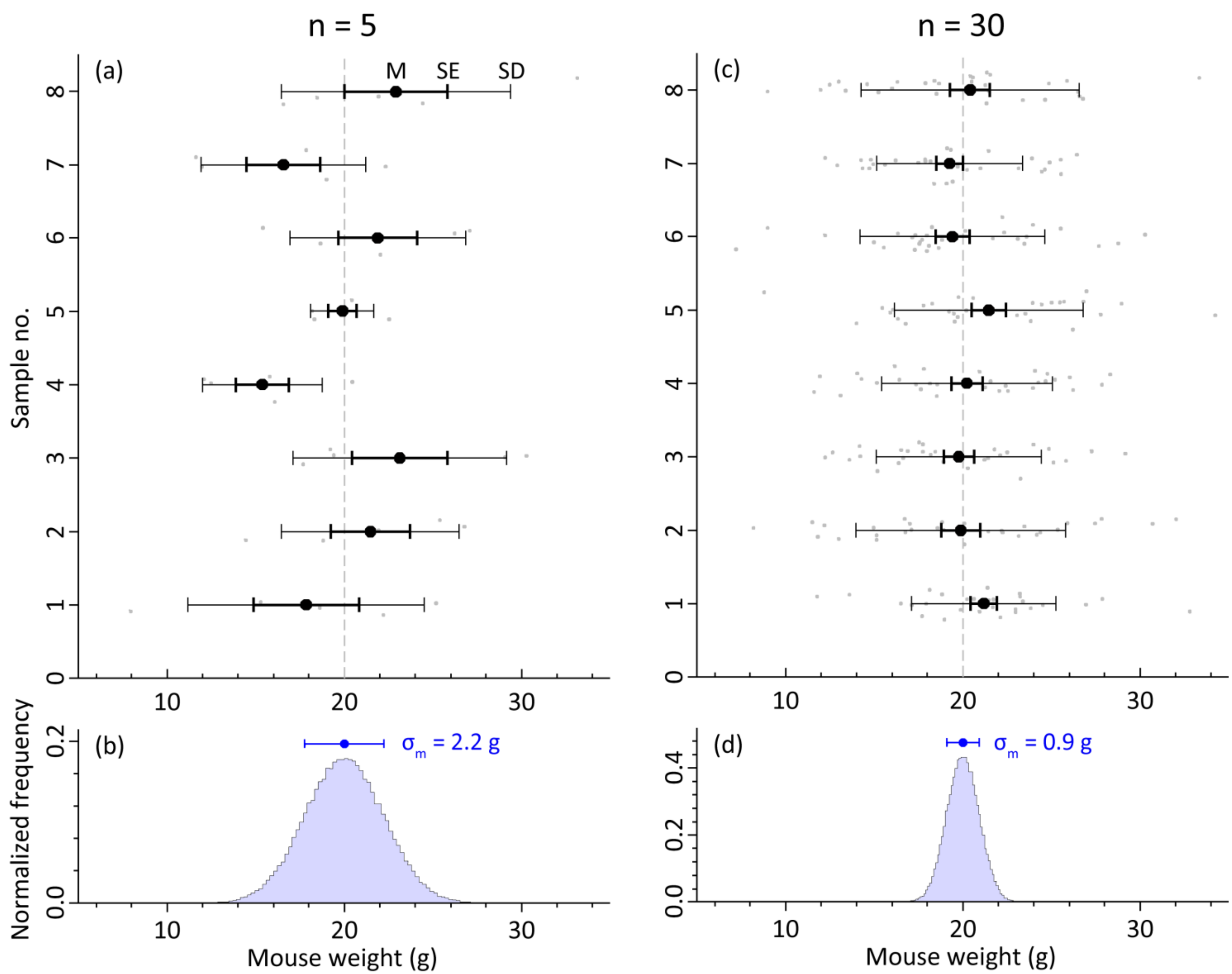
- Hence, **uncertainty of the mean** can be estimated by

$$SE = \frac{SD}{\sqrt{n}}$$

- Standard error **estimates** the width of the sampling distribution



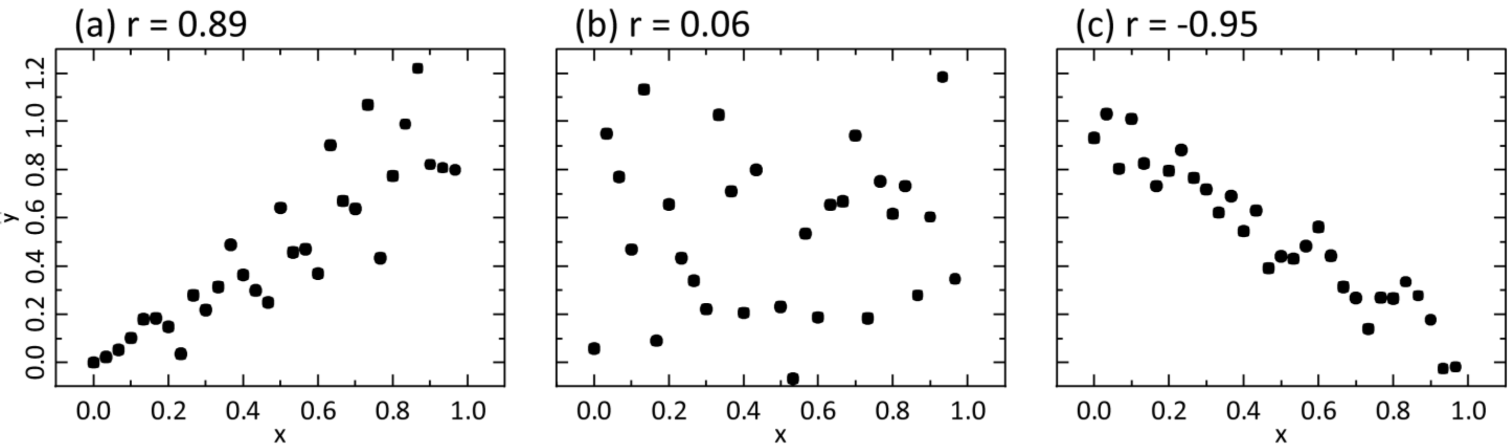
Standard error of the mean



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: $x_1, x_2, \dots, x_n$ and $y_1, y_2, \dots, y_n$

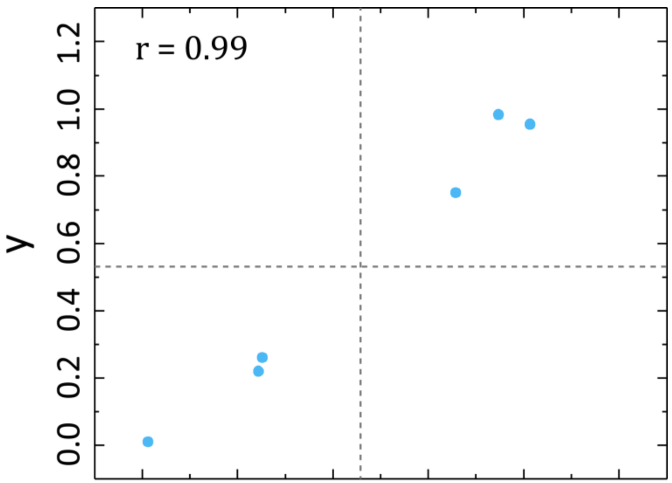
$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - M_x}{SD_x}\right)\left(\frac{y_i - M_y}{SD_y}\right) = \frac{1}{n-1}\sum_{i=1}^{n} Z_{xi} Z_{yi}$$

where Z is a "Z-score"

- Correlation does not mean causation!

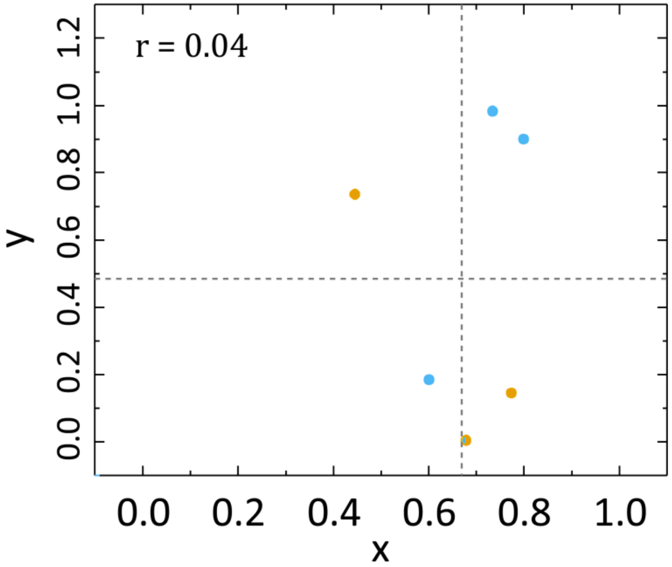
Correlation coefficient: example

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



x	y	Z_x	Z_y	Z_xZ_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

$$\sum Z_xZ_y = 4.96$$



x	y	Z_x	Z_y	Z_xZ_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

$$\sum Z_xZ_y = 0.18$$

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