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)

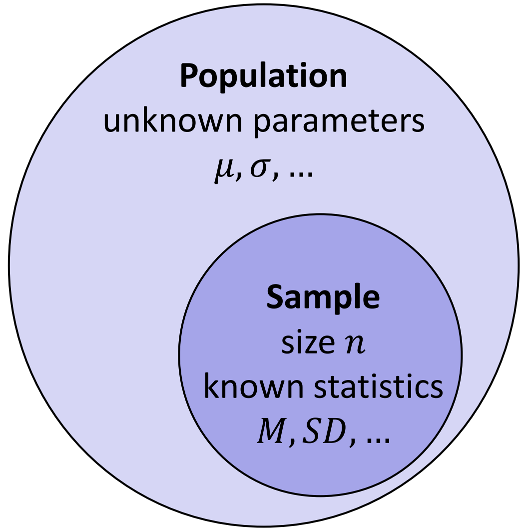
Previously on Errors…

Random errors

- measurement error
- reading error
- counting error
- **sampling error**

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

- mean, median, mode
- variance, standard deviation
- correlation

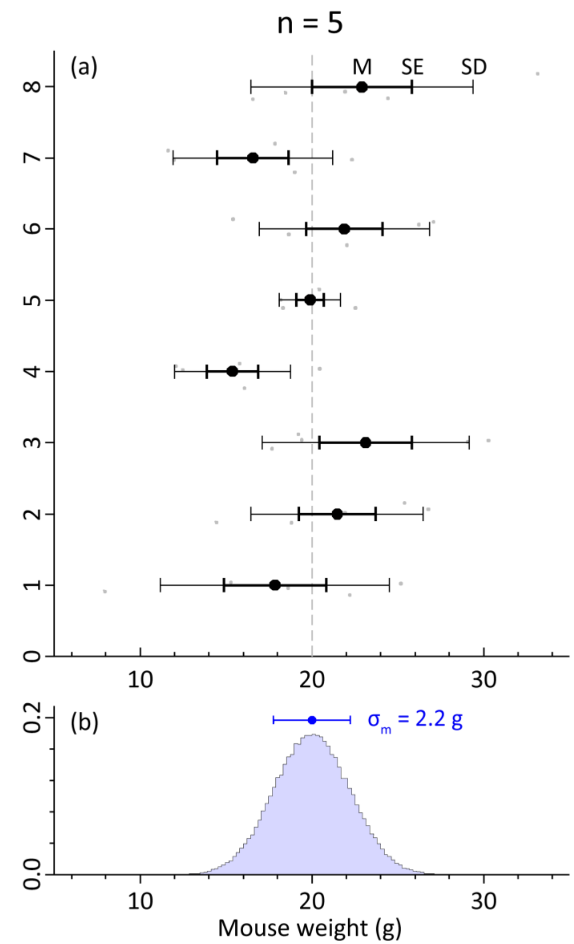


Standard deviation

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

Standard error

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



Sampling distribution of the mean – distribution of sample means from repeated experiments

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

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

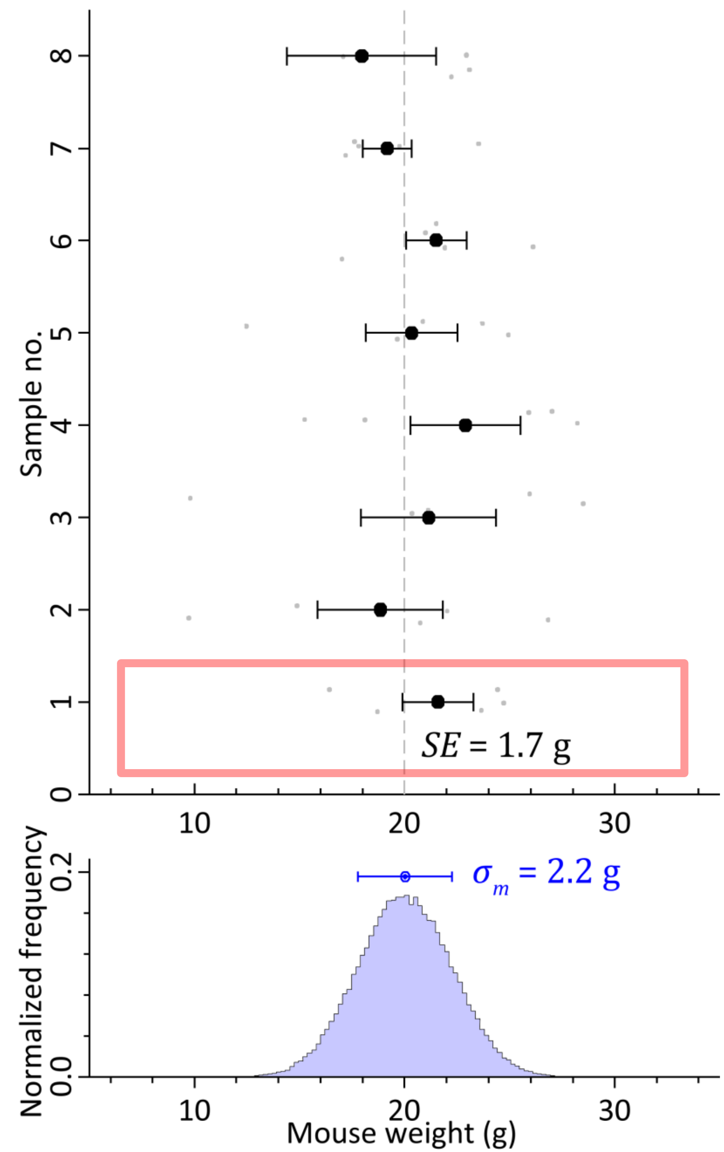
Real experiment

- 5 mice
- Measure body mass:

  16.4, 18.7, 23.7, 24.4, 24.7 g
- Find standard error

$$SE = \frac{SD}{\sqrt{n}} = 1.7 \text{ g}$$

SE is an approximation of σ_m

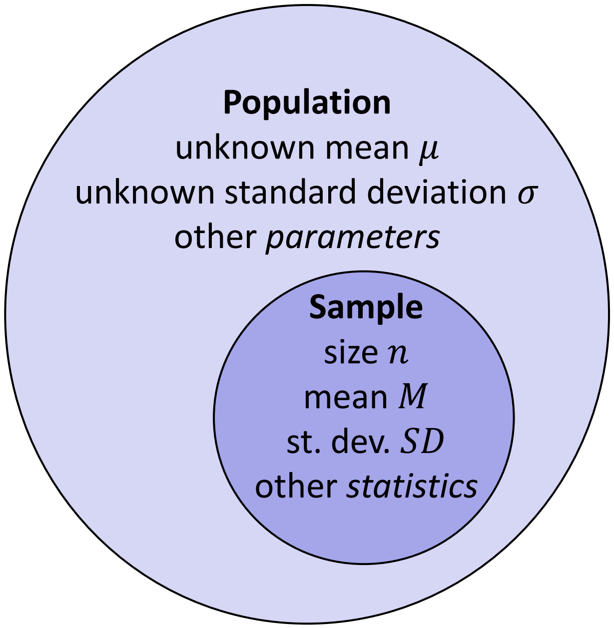


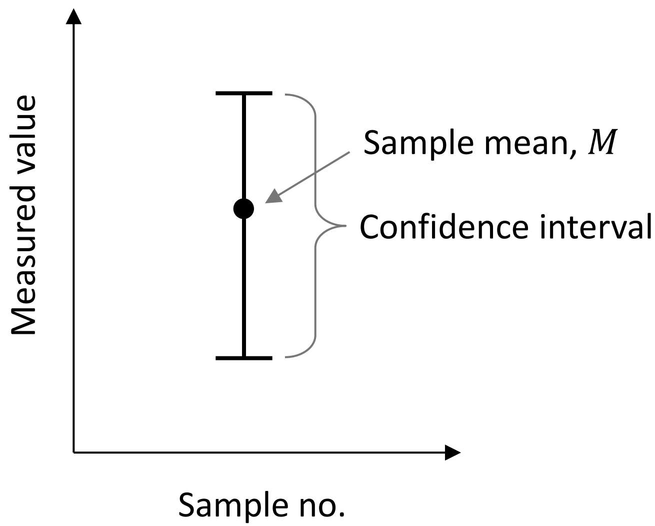
4. Confidence intervals I

“Confidence is what you have before you understand the problem”

Woody Allen

Confidence intervals





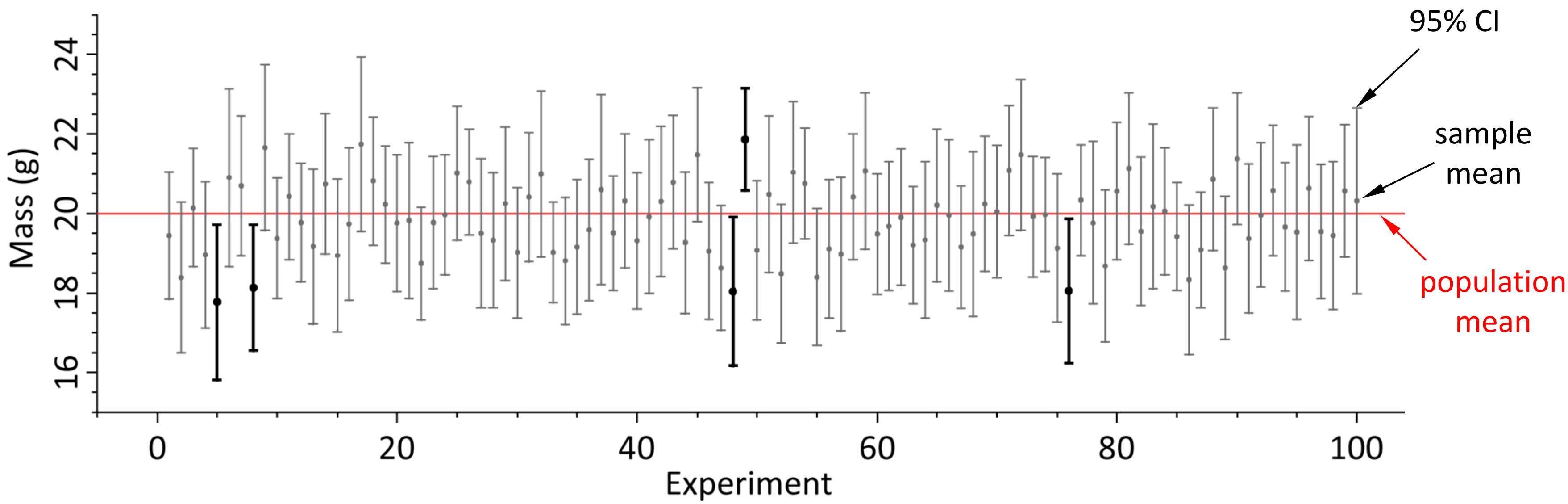
- Sample mean, M, is a statistical estimator of the true mean, μ
- How good is M?
- Confidence interval: a range $[M_L, M_U]$ around M, where we think the true mean is, with a *certain confidence*
- This can be done for any population parameter
  - mean
  - median
  - standard deviation
  - correlation
  - proportion
  - etc.

What is confidence?

- Consider a 95% confidence interval for the mean $[M_L, M_U]$

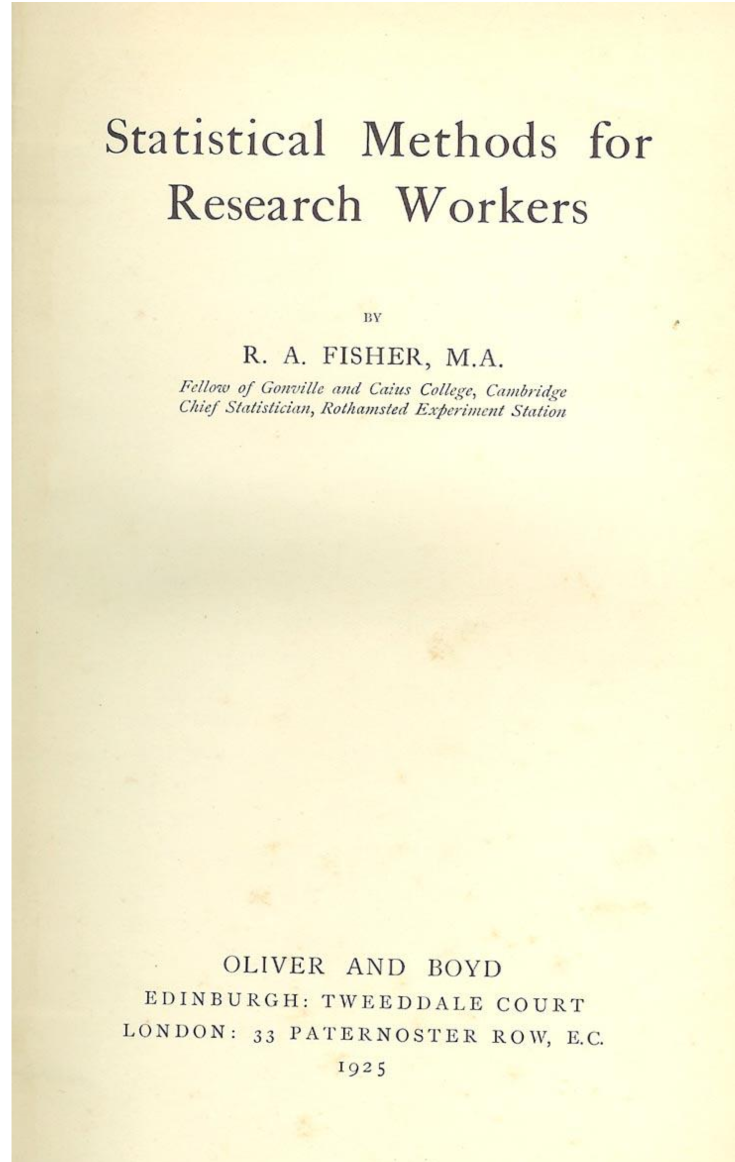
- This **does not mean** there is a 95% probability of finding the true mean in the calculated interval
- The true population mean is a constant number, not a random variable!

- Repeated experiments: in 95% of cases the true mean would be within the calculated interval, and in 5% of cases outside it



Why 95%?

- Textbook by Ronald Fisher (1925)
- He though 95% confidence interval was "convenient" as it resulted in 1 false indication in 20 trials
- He published tables for a few probabilities, including $p = 5\%$
- The book had become one of the most influential textbooks in 20th century statistics
- However, there is nothing special about 95% confidence interval or *p*-value of 5%



Statistical Methods for Research Workers

BY

R. A. FISHER, M.A.

Fellow of Gonville and Caius College, Cambridge
Chief Statistician, Rothamsted Experiment Station

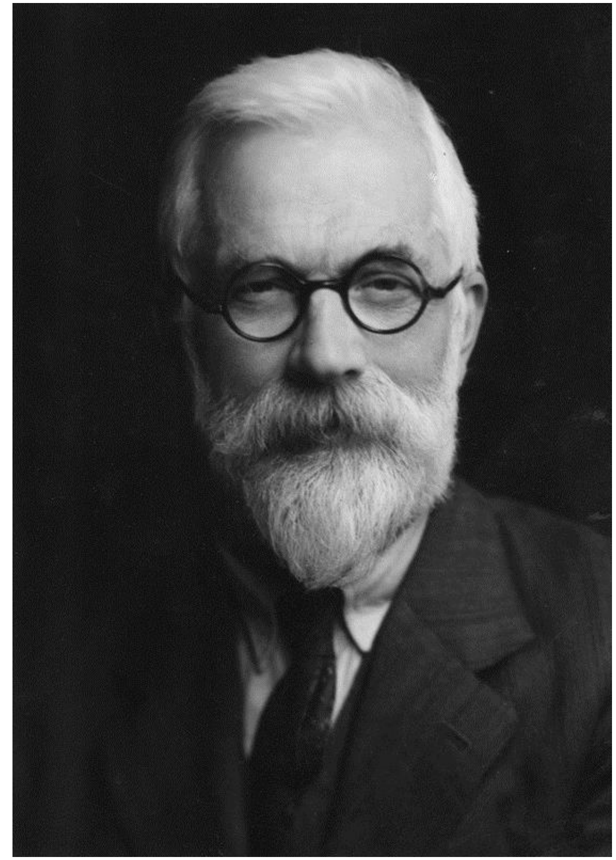
OLIVER AND BOYD
EDINBURGH: TWEEDDALE COURT
LONDON: 33 PATERNOSTER ROW, E.C.
1925

Ronald Fisher

- Probably the most influential statistician of the 20th century
- Also evolutionary biologists
- Went to Harrow School and then Cambridge
- Arthur Vassal, Harrow's schoolmaster:

  I would divide all those I had taught into two groups: one containing a single outstanding boy, Ronald Fisher; the other all the rest

- Didn't like administration and admin people: "an administrator, not the highest form of human life"



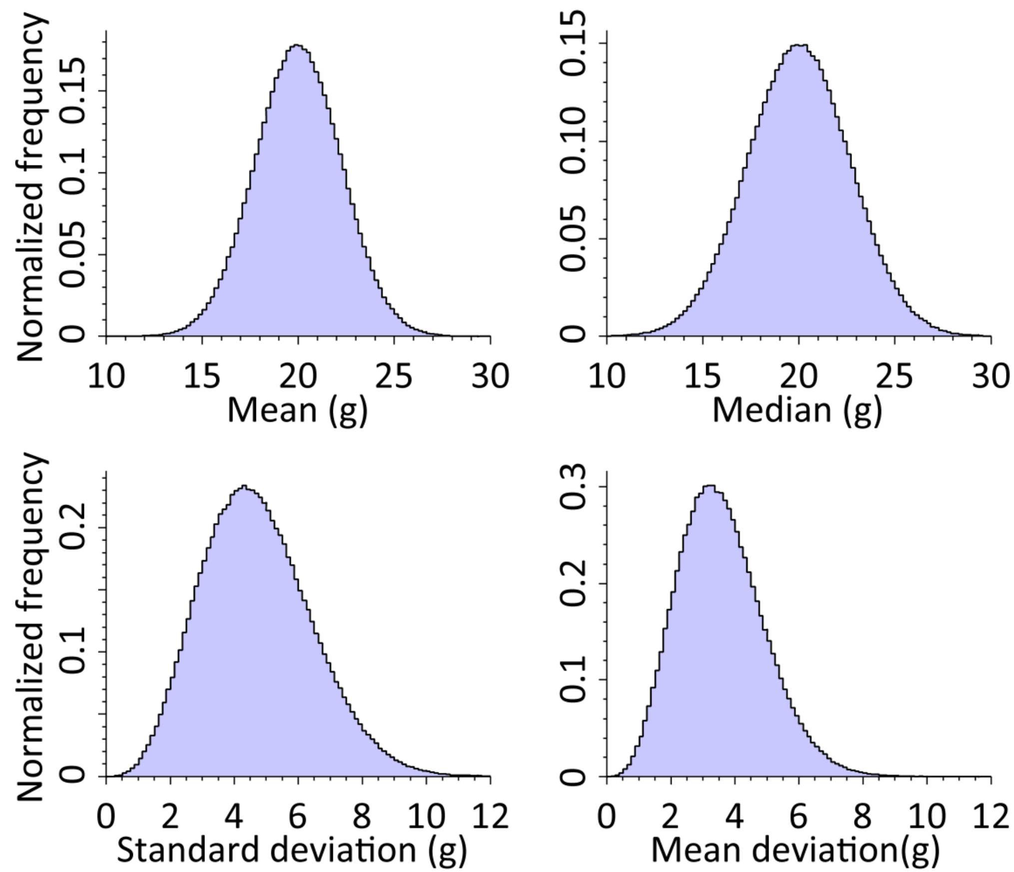
Ronald Fisher (1890-1962)

Sampling distribution

- *Gedankenexperiment*
- Consider an unknown population you wish to characterize
- Draw lots of samples of size n from this population
- Calculate an estimator from each sample, for example

- Build a distribution of the estimator
- This is a *sampling distribution*

- Width of the sampling distribution is a standard error

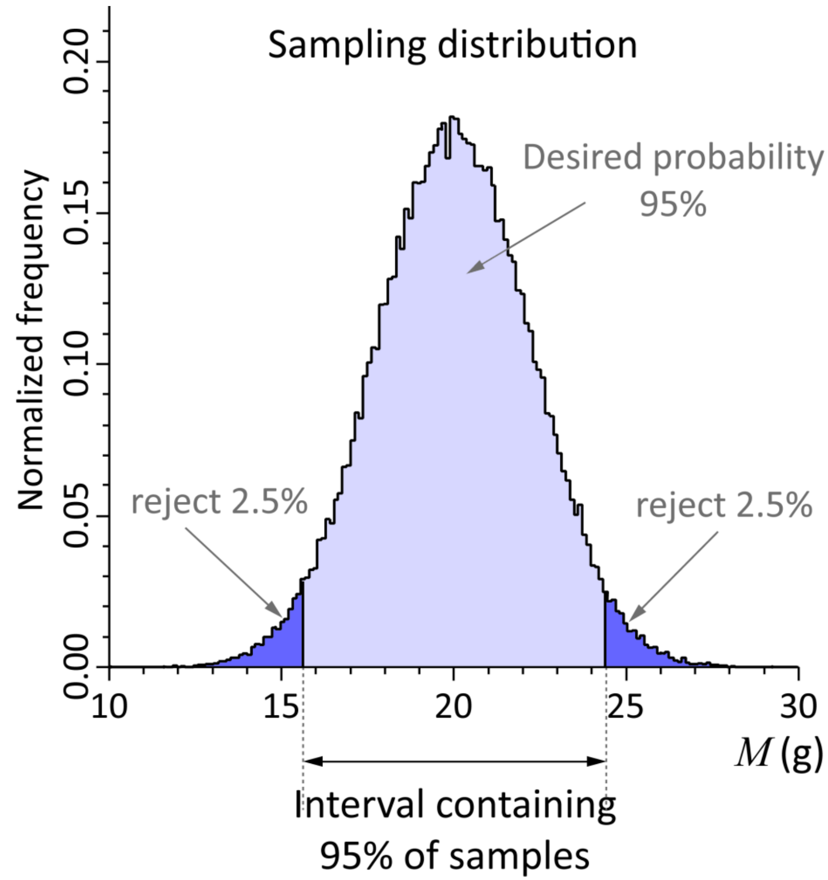


Examples of sampling distribution

10^6 samples of $n = 5$ from $\mathcal{N}(20, 5)$

Sampling distribution of the mean

- The distribution curve represents all samples
- Keep the region corresponding to the required confidence, e.g. 95%
- Reject 2.5% on each side
- This gives confidence interval for our estimator

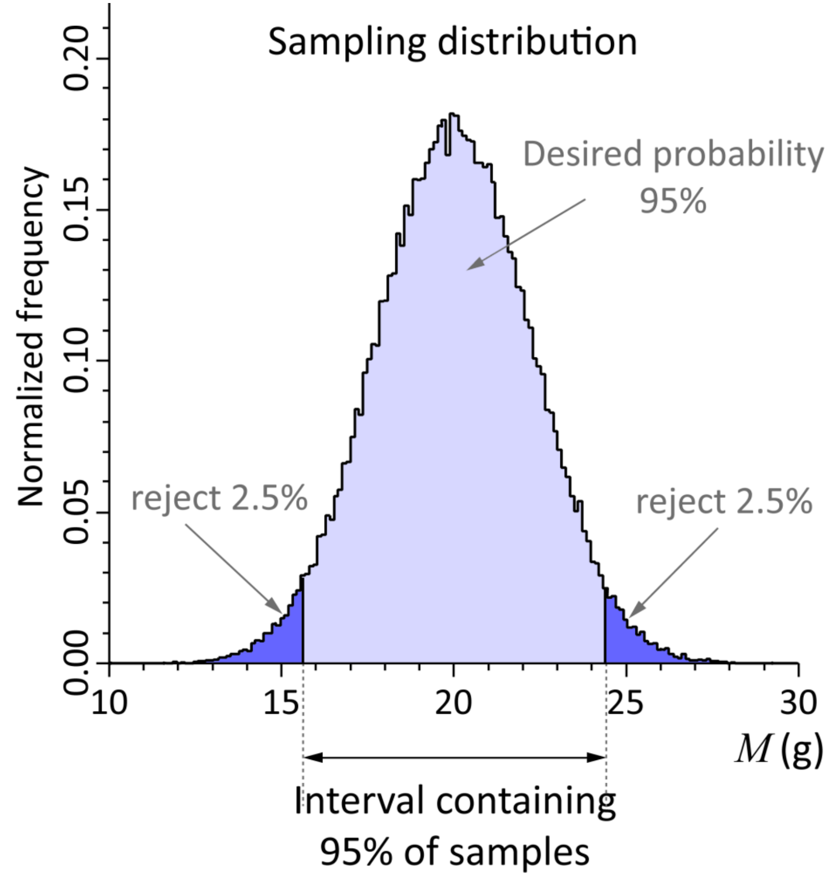


100,000 samples of 5 mice from normal population with $\mu = 20$ g and $\sigma = 5$ g

Mean body weight calculated for each sample

Sampling distribution of the mean

- The distribution curve represents all samples
- Keep the region corresponding to the required confidence, e.g. 95%
- Reject 2.5% on each side
- This gives confidence interval for our estimator
- In real life you can't draw thousands of samples!
- Instead you can use a *known probability distribution* to calculate probabilities



100,000 samples of 5 mice from normal population with $\mu = 20$ g and $\sigma = 5$ g

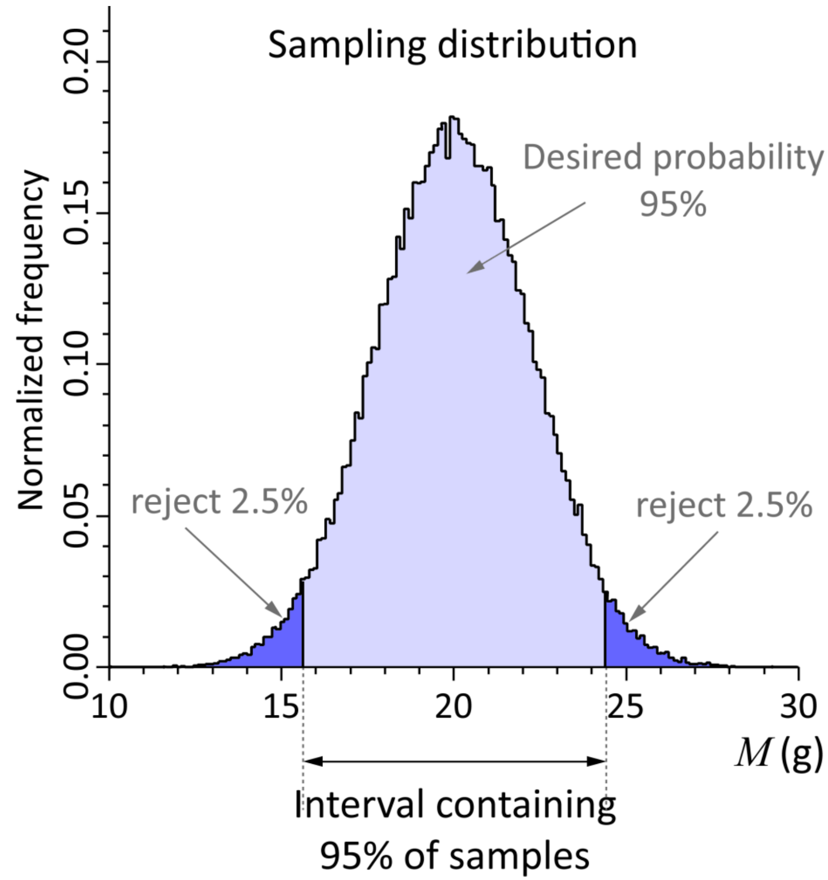
Mean body weight calculated for each sample

Sampling distribution of the mean

- Sampling distribution of the mean is not known
- For the given sample find M, SD and n
- Let us define a statistic

$$t = \frac{M - \mu}{SE}$$

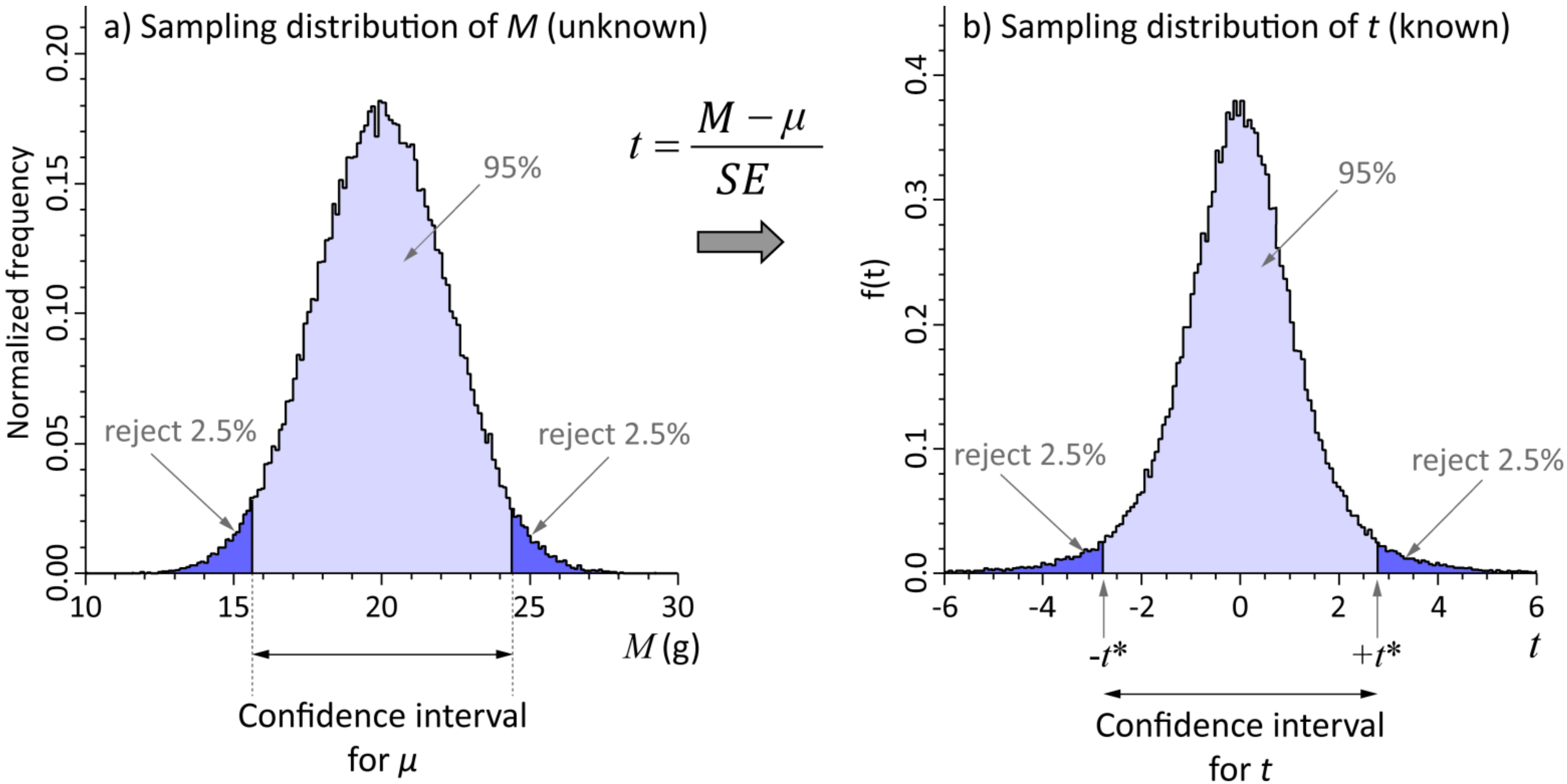
- Mathematical trick – we cannot calculate t
- *Gedankenexperiment*: create a sampling distribution of t



100,000 samples of 5 mice from normal population with $\mu = 20$ g and $\sigma = 5$ g

Mean body weight calculated for each sample

Sampling distribution of t-statistic



Confidence interval of the mean

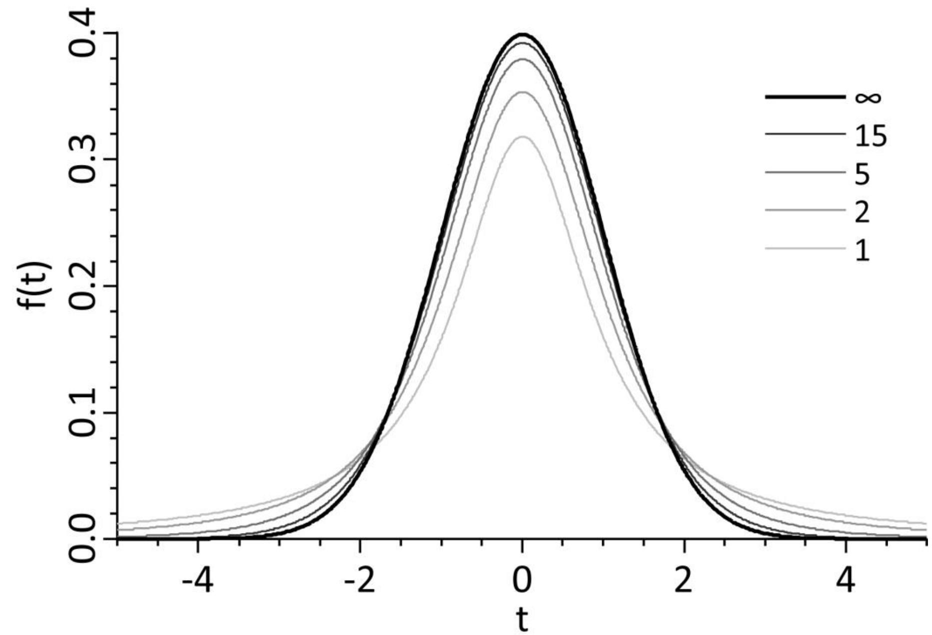
- Statistic

$$t = \frac{M - \mu}{SE}$$

has a *known* sampling distribution: Student's t-distribution with $n - 1$ degrees of freedom

- We can calculate probabilities!

Student's t-distribution



This is the t-distribution for 1, 2, 5, 15 and ∞ degrees of freedom. For large number of d.o.f. this turns into a Gaussian distribution.

William Gosset

- Brewer and statistician
- Developed Student’s *t*-distribution

- Worked for Guinness, who prohibited employees from publishing any papers
- Published as “Student”

- Worked with Fisher and developed the *t*-statistic in its current form
- Always worked with experimental data
- Progenitor bioinformatician?

William Sealy Gosset (1876-1937)

William Gosset

VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Introduction.

ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

If the number of experiments be very large, we may have precise information as to the value of the mean, but if our sample be small, we have two sources of uncertainty:—(1) owing to the "error of random sampling" the mean of our series

William Gosset’s calculator

Confidence interval of the mean

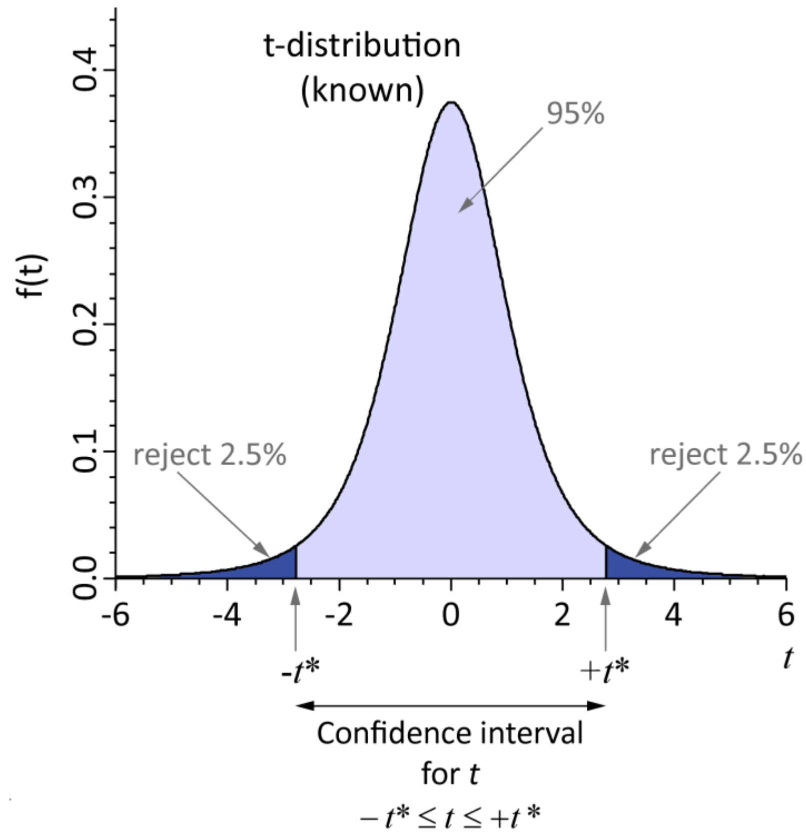
- Statistic

$$t = \frac{M - \mu}{SE}$$

has a *known* sampling distribution: Student's t-distribution with $n - 1$ degrees of freedom

- We can find a critical value of t^* to cut off required confidence interval
- Use tables of t-distribution or any statistical package

- Confidence interval on t is $[-t^*, +t^*]$



Tables of t-distribution usually give critical values of t^* for the tail probability.

Confidence interval of the mean

- We used transformation

$$t = \frac{M - \mu}{SE}$$

- Confidence interval on t is $[-t^*, +t^*]$

- Find μ from the equation above

$$\mu = M + tSE$$

- From limits on t we find limits on μ:

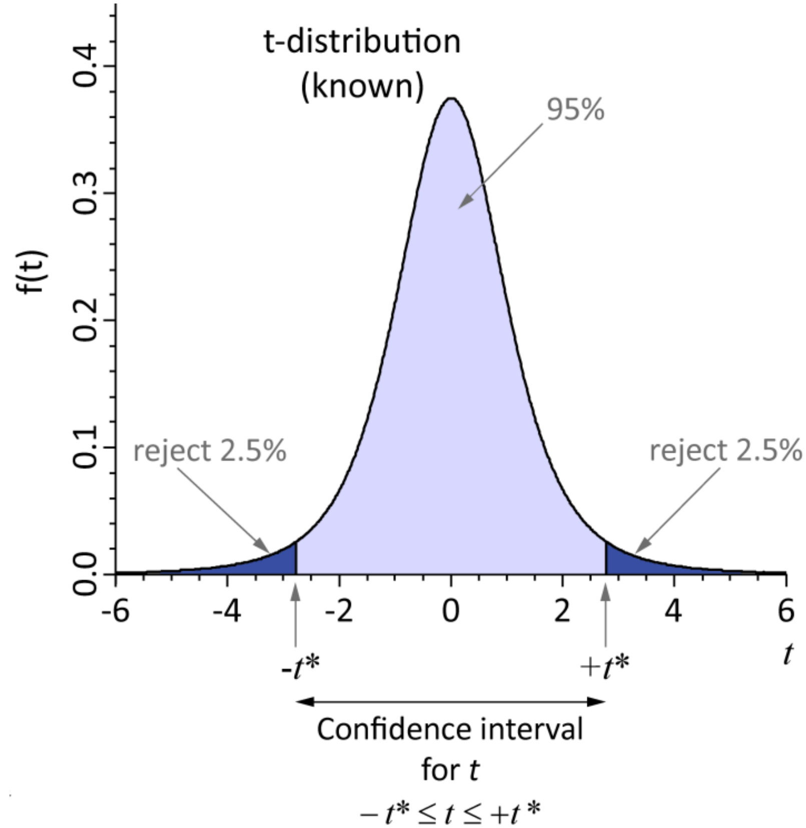
$$M_L = M - t^* SE$$
$$M_U = M + t^* SE$$

- Or

$$\mu = M \pm t^* SE$$

- Confidence interval is a scaled standard error

$$CI = t^* SE$$



Tables of t-distribution usually give critical values of t^* for the tail probability.

Exercise: 95% confidence interval for the mean

- We have 7 mice with measured body weights 16.8, 21.8, 29.2, 23.3, 19.5, 18.2 and 26.3 g
- Estimators from the sample

$$M = 22.16 \text{ g}$$
$$SD = 4.46 \text{ g}$$
$$SE = 1.69 \text{ g}$$

- Find the 95% confidence interval for the mean

Tail Probabilities						
One Tail		0.10	0.05	0.025	0.01	0.005
Two Tails		0.20	0.10	0.05	0.02	0.01
D	1	3.078	6.314	12.71	31.82	63.66
E	2	1.886	2.920	4.303	6.965	9.925
G	3	1.638	2.353	3.182	4.541	5.841
R	4	1.533	2.132	2.776	3.747	4.604
E	5	1.476	2.015	2.571	3.365	4.032
E	6	1.440	1.943	2.447	3.143	3.707
S	7	1.415	1.895	2.365	2.998	3.499
	8	1.397	1.860	2.306	2.896	3.355
O	9	1.383	1.833	2.262	2.821	3.250
F	10	1.372	1.812	2.228	2.764	3.169
	11	1.363	1.796	2.201	2.718	3.106
F	12	1.356	1.782	2.179	2.681	3.055
R	13	1.350	1.771	2.160	2.650	3.012
E	14	1.345	1.761	2.145	2.624	2.977
E	15	1.341	1.753	2.131	2.602	2.947
D	16	1.337	1.746	2.120	2.583	2.921
O	17	1.333	1.740	2.110	2.567	2.898
M	18	1.330	1.734	2.101	2.552	2.878
	19	1.328	1.729	2.093	2.539	2.861
	20	1.325	1.725	2.086	2.528	2.845
	[illegible]	[illegible]	1.721	[illegible]	[illegible]	[illegible]
	[illegible]	[illegible]	1.717	[illegible]	[illegible]	[illegible]
	60	[illegible]	[illegible]	[illegible]	2.390	[illegible]
	65	1.295	[illegible]	1.997	2.385	2.654
	70	1.294	1.667	1.994	2.381	2.648
	80	1.292	1.664	1.990	2.374	2.639
	100	1.290	1.660	1.984	2.364	2.626
	150	1.287	1.655	1.976	2.351	2.609
	200	1.286	1.653	1.972	2.345	2.601
Two Tails		0.20	0.10	0.05	0.02	0.01
One Tail		0.10	0.05	0.025	0.01	0.005
Tail Probabilities						

Exercise: 95% confidence interval for the mean

- We have 7 mice with measured body weights 16.8, 21.8, 29.2, 23.3, 19.5, 18.2 and 26.3 g
- Estimators from the sample

$$M = 22.16\text{ g}$$
$$SD = 4.46\text{ g}$$
$$SE = 1.69\text{ g}$$

- Critical value from t-distribution for one-tail probability 0.025 and 6 degrees of freedom

$$t^* = 2.447$$

- Half of the confidence interval is

$$CI = t^* SE = 4.14\text{ g}$$

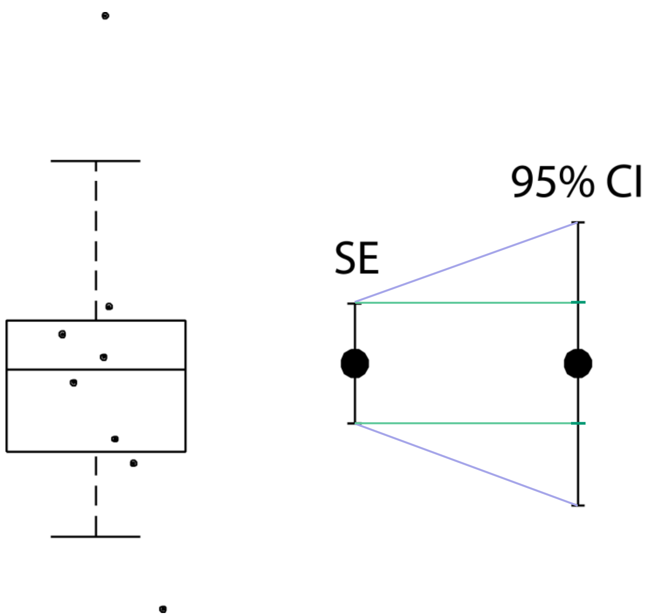
- Estimate of the mean with 95% confidence is

$$\mu = 22 \pm 4\text{ g}$$

~~22.16 ± 4.14 g~~

Tail Probabilities						
One Tail		0.10	0.05	0.025	0.01	0.005
Two Tails		0.20	0.10	0.05	0.02	0.01
D	1	3.078	6.314	12.71	31.82	63.66
E	2	1.886	2.920	4.303	6.965	9.925
G	3	1.638	2.353	3.182	4.541	5.841
R	4	1.533	2.132	2.776	3.747	4.604
E	5	1.476	2.015	2.571	3.365	4.032
E	6	1.440	1.943	2.447	3.143	3.707
S	7	1.415	1.895	2.365	2.998	3.499
	8	1.397	1.860	2.306	2.896	3.355
O	9	1.383	1.833	2.262	2.821	3.250
F	10	1.372	1.812	2.228	2.764	3.169
	11	1.363	1.796	2.201	2.718	3.106
F	12	1.356	1.782	2.179	2.681	3.055
R	13	1.350	1.771	2.160	2.650	3.012
E	14	1.345	1.761	2.145	2.624	2.977
E	15	1.341	1.753	2.131	2.602	2.947
D	16	1.337	1.746	2.120	2.583	2.921
O	17	1.333	1.740	2.110	2.567	2.898
M	18	1.330	1.734	2.101	2.552	2.878
	19	1.328	1.729	2.093	2.539	2.861
	20	1.325	1.725	2.086	2.528	2.845
	[illegible]	[illegible]	1.721	[illegible]		
	[illegible]		1.717			
	60	[illegible]		[illegible]	2.390	[illegible]
	65	1.295	[illegible]	1.997	2.385	2.654
	70	1.294	1.667	1.994	2.381	2.648
	80	1.292	1.664	1.990	2.374	2.639
	100	1.290	1.660	1.984	2.364	2.626
	150	1.287	1.655	1.976	2.351	2.609
	200	1.286	1.653	1.972	2.345	2.601
Two Tails		0.20	0.10	0.05	0.02	0.01
One Tail		0.10	0.05	0.025	0.01	0.005
Tail Probabilities						

Confidence interval vs standard error



How many standard errors are in a confidence interval?

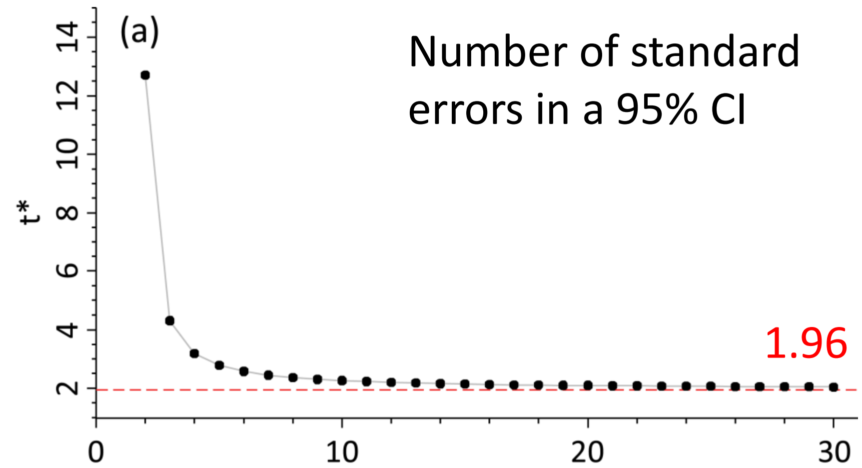
What is the confidence of the standard error?

Confidence interval vs standard error

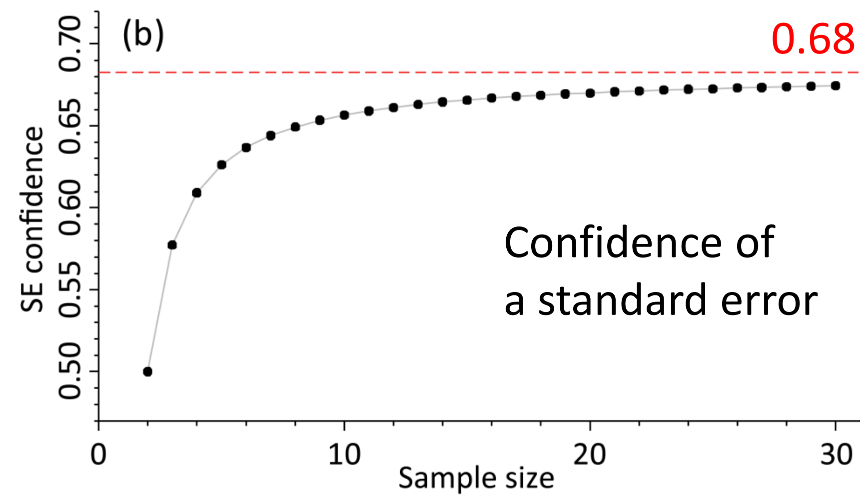
- Confidence interval is a scaled standard error

$$CI = t^*_{n-1}(\gamma) \times SE$$

- Scaling factor, t^*, depends on
  - confidence, γ
  - sample size, n



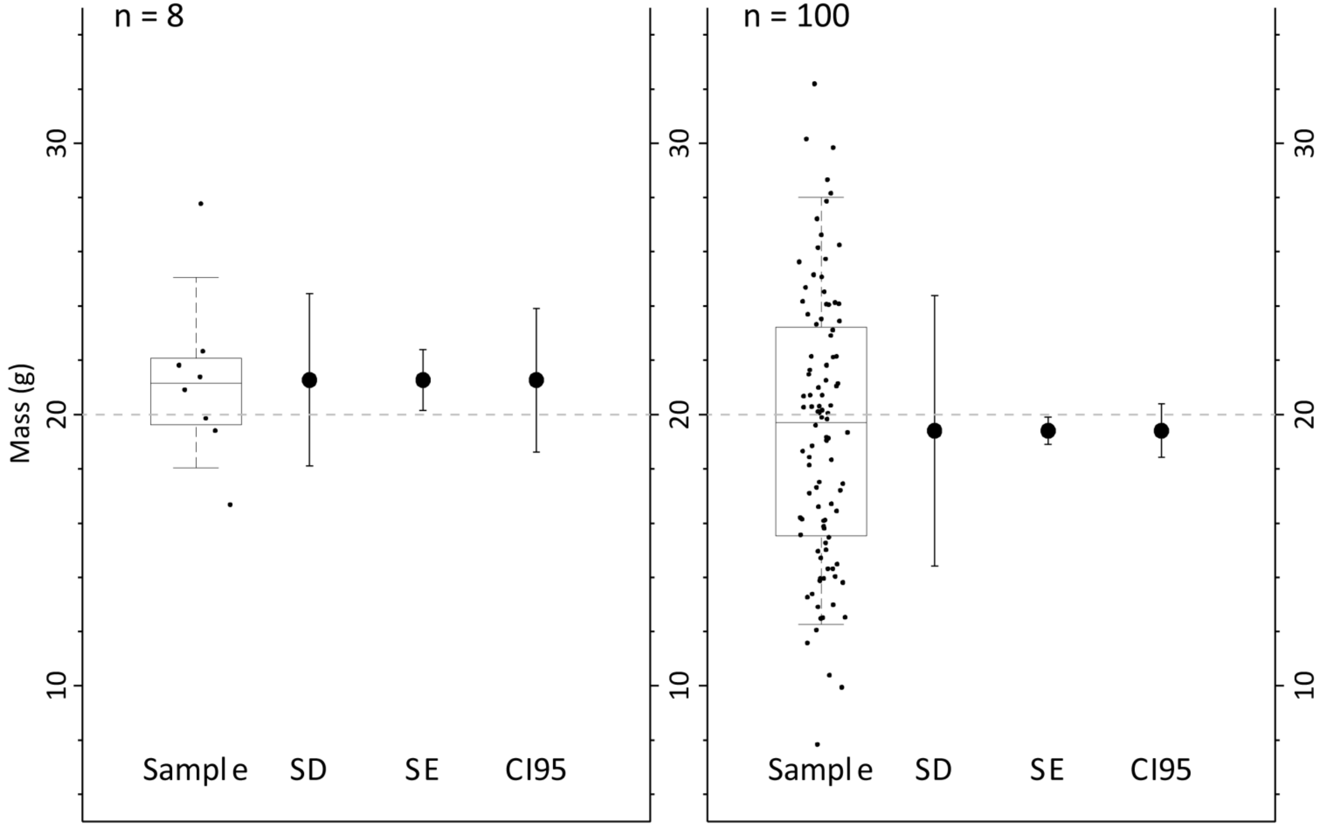
- We can find how t^* depends on the sample size
- We can find confidence corresponding to $t^* = 1$



- 95% CI is huge for $n = 2$
- SE has no simple probabilistic interpretation

- Use confidence intervals!

SD, SE and 95% CI



- Normal population of $\mu = 20$ g and $\sigma = 5$ g
- Sample of $n = 8$ and $n = 100$
- Whiskers in the box plot encompass 90% of data – nothing to do with 90% confidence interval

Exercise: confidence intervals

- Experiment where a reporter measures transcriptional activity of a gene
- Day 1: 3 biological replicates, day 2: 5 biological replicates (the same experiment)
- Normalized data:

Day 1	0.89	0.92	0.89		
Day 2	0.55	0.76	0.61	0.83	0.75

- 95% confidence intervals for the mean:

  Day 1: [0.86, 0.94]

  Day 2: [0.56, 0.84]

- What can you say about these results? What else can you do with these data?

Exercise: confidence intervals

- Experiment where a reporter measures transcriptional activity of a gene
- Day 1: 3 biological replicates, day 2: 5 biological replicates (the same experiment)
- Normalized data:

Day 1	0.89	0.92	0.89		
Day 2	0.55	0.76	0.61	0.83	0.75

- 95% confidence intervals for the mean:

  Day 1: [0.86, 0.94]

  Day 2: [0.56, 0.84]

- What can you say about these results? What else can you do with these data?
- You could pool data together to get [0.66, 0.89]
- But t-test between two days gives $p = 0.03$
- Perhaps something had changed between day 1 and day 2?

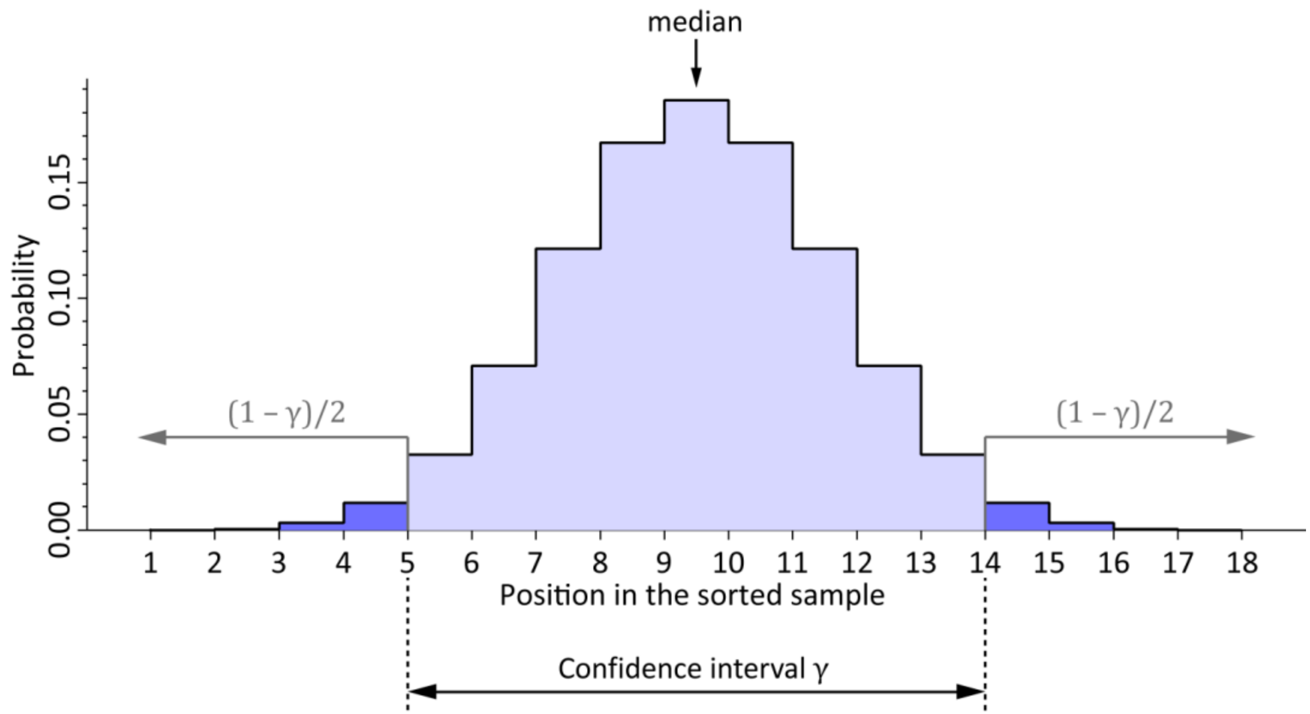
Confidence interval of the median

- Median is quite often used in biology
- Do you know how to find its uncertainty?
- Consider sample $x_1, x_2, \dots, x_n$
- Let (unknown) population median be Θ
- $P(x_i < \Theta) = \frac{1}{2}$ and $P(x_i > \Theta) = \frac{1}{2}$
- It is like tossing a coin

- Sort the sample and find the probability of Θ between k and $n-k$ points:

$$P(x_{(1)} \leq \cdots \leq x_{(k)} \leq \Theta \leq x_{(k+1)} \dots \leq x_{(n)})$$

- Binomial distribution
- You can find precise confidence limits by interpolating in this discrete distribution



Confidence interval of the median: approximation

- Sample $x_1, x_2, \dots, x_n$
- Sorted sample $x_{(1)} \le x_{(2)} \le \dots \le x_{(n)}$
- Find two limiting indices:

$$L = \left\lfloor \frac{n}{2} \right\rfloor - \left\lceil \sqrt{\frac{n}{4}} \right\rceil$$

$\lfloor x \rfloor$ - floor, $\lfloor 3.4 \rfloor = 3$

$\lceil x \rceil$ - ceiling, $\lceil 3.4 \rceil = 4$

$$U = n - L$$

- Standard error of the median

$$\widetilde{SE} = \frac{\mathrm{x}_{(U)} - x_{(L+1)}}{2}$$

- Confidence intervals

$$\widetilde{M}_L = \widetilde{M} - t^* \widetilde{SE}$$

$$\widetilde{M}_U = \widetilde{M} + t^* \widetilde{SE}$$

- Here, t^* is the critical value from t-distribution with $U - L - 1$ degrees of freedom

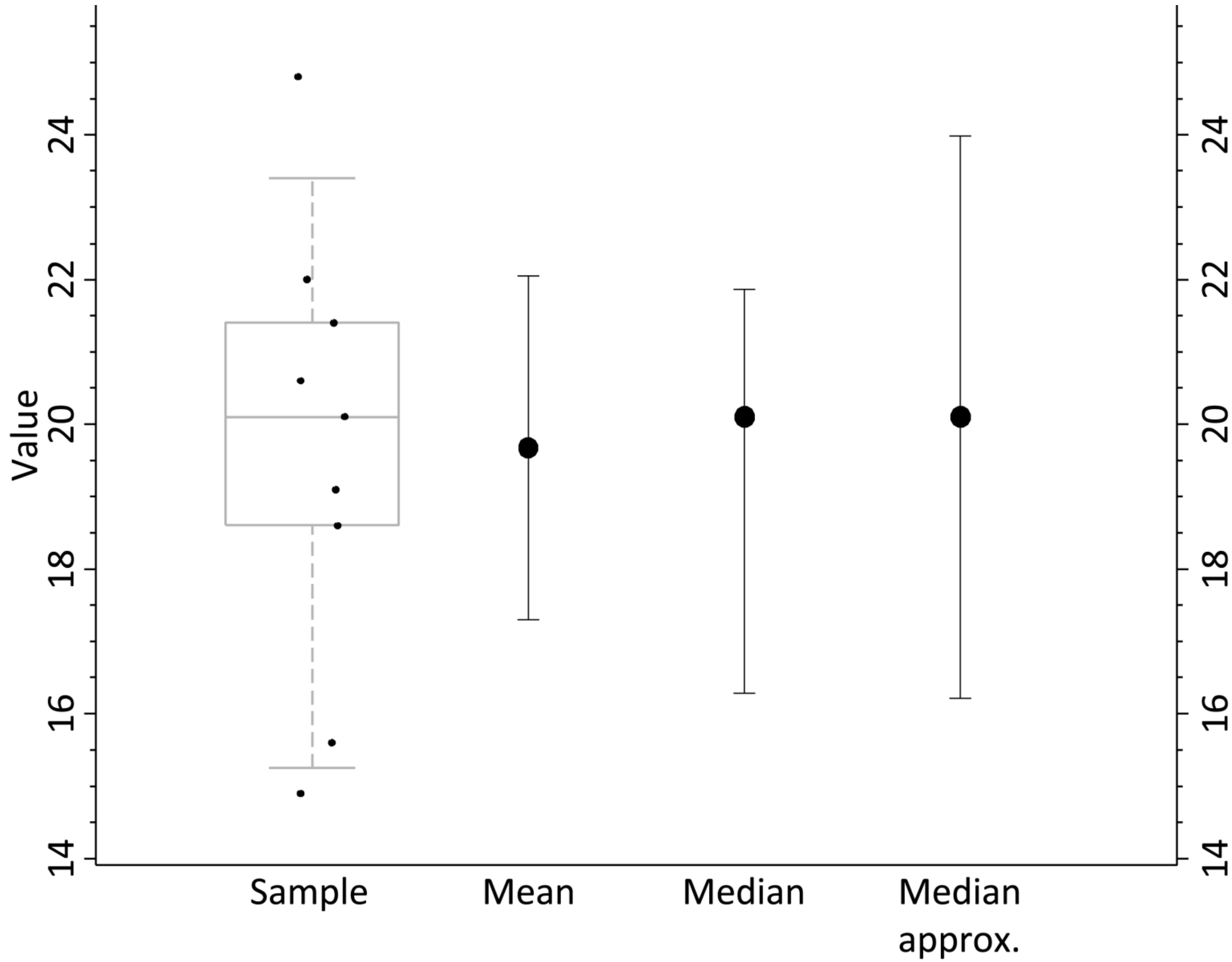
Example

- Weighed 9 mice: 14.9, 22.0, 15.6, 19.1, 21.4, 20.6, 20.1, 24.8, 18.6 g
- Sorted sample:

i	1	2	3	4	5	6	7	8	9
$x_{(i)}$	14.9	15.6	18.6	19.1	20.1	20.6	21.4	22.0	24.8

- Median is $\widetilde{M} = 20.1$ g
- $L = \lfloor 4.5 \rfloor - \lceil 1.5 \rceil = 4 - 2 = 2$
- $U = 9 - 2 = 7$
- $\widetilde{SE} = \frac{21.4 - 18.6}{2} = 1.4$ g
- $t^* = 2.776$ for 4 d.o.f.
- 95% CI is $[16.2, 24.0]$ g

Confidence interval of the median: example



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

Please leave your feedback forms on the table by the door