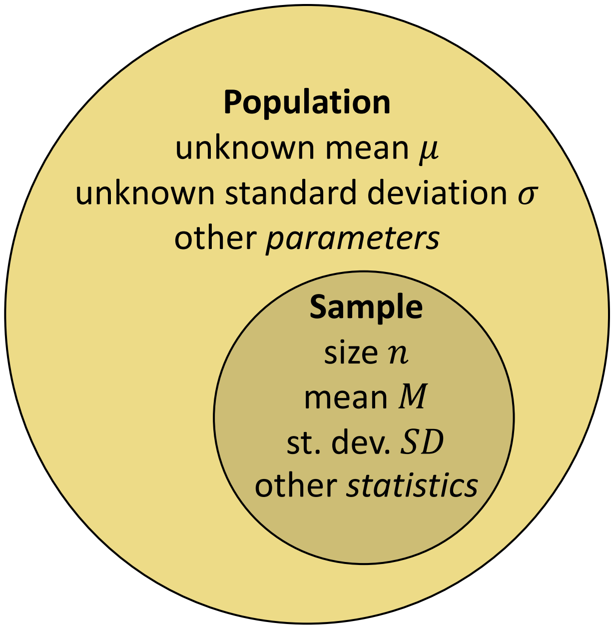
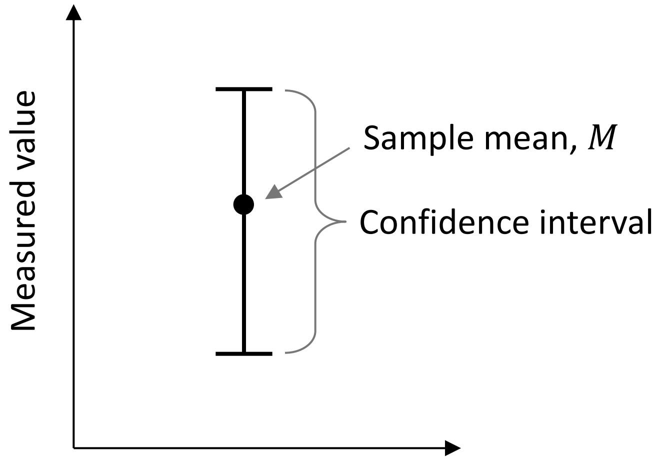
3. Confidence intervals

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

Woody Allen

Confidence intervals





- Sample mean, M, estimates the true mean, μ
- How good is M?

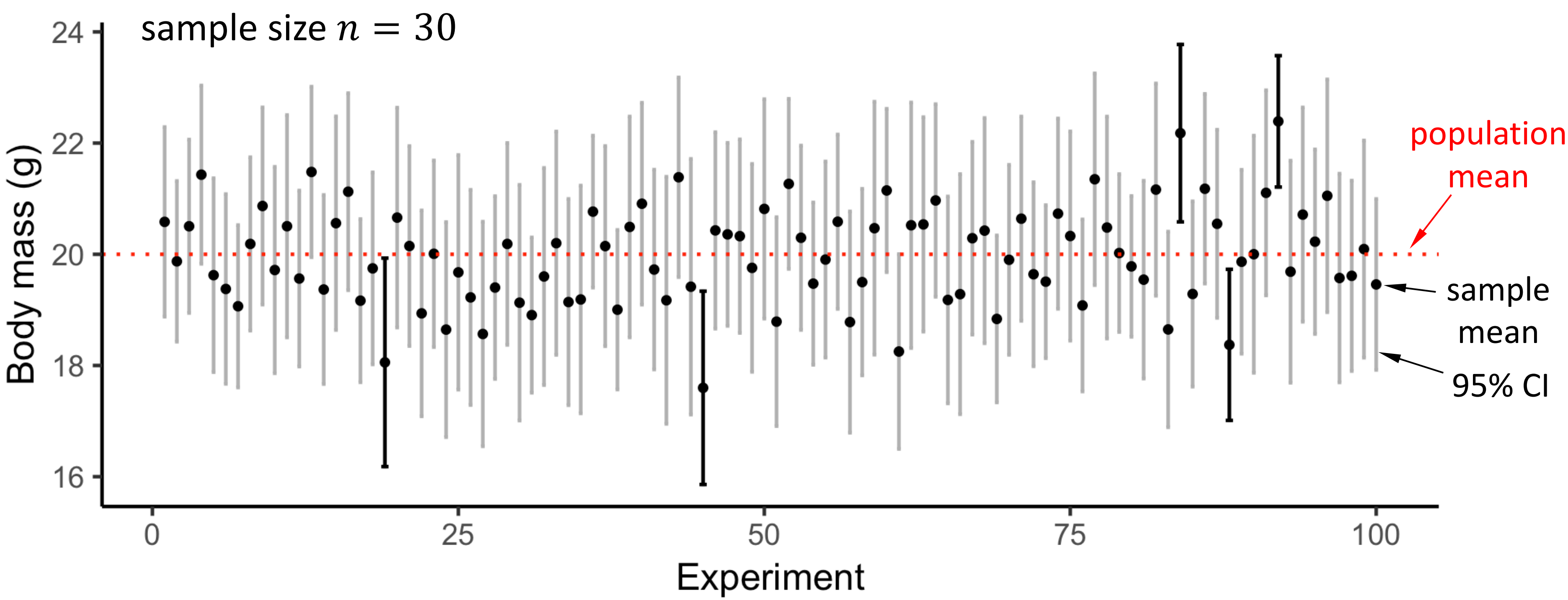
- Confidence interval: a range $[M_L, M_U]$, where we expect the true mean be with a *certain confidence*

- This can be done for any population parameter
  - mean
  - median
  - standard deviation
  - correlation
  - proportion
  - etc.

What is confidence?

- Consider a 95% confidence interval of the mean

- If you were to repeat the entire experiment many times
  - 95% of cases the true mean would be within the calculated interval
  - 5% of cases (1 in 20) it would be outside it (false result)

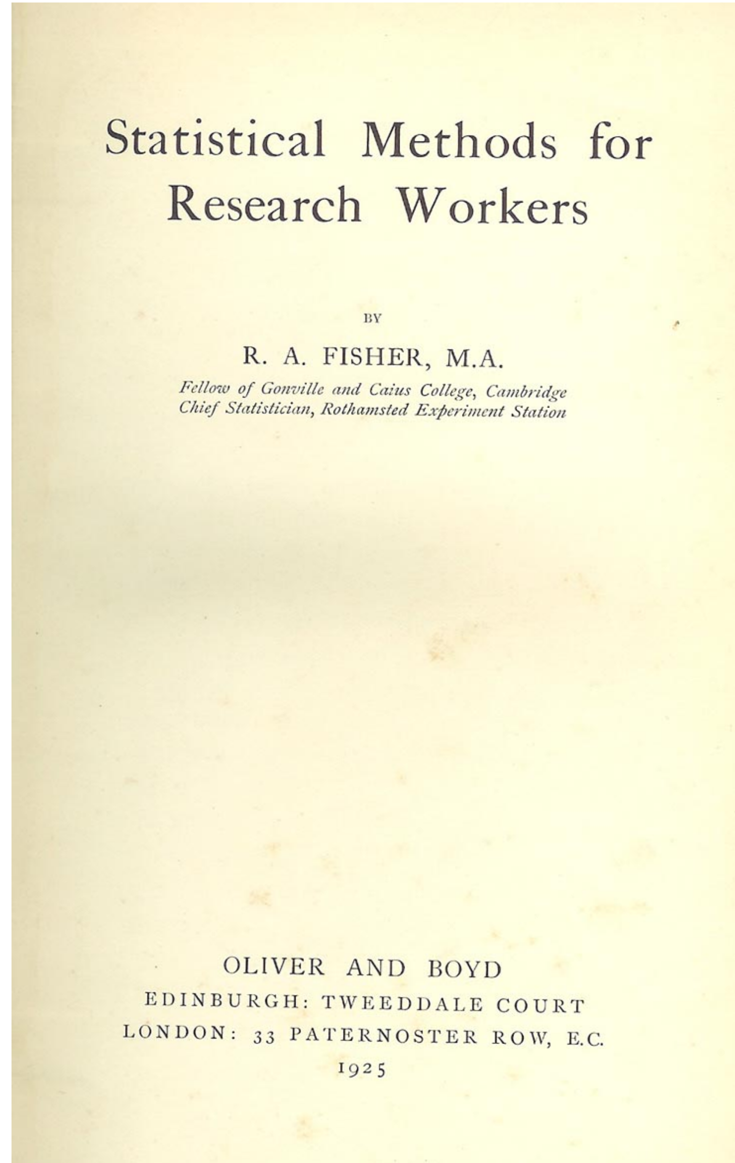


Why 95%?

- Textbook by Ronald Fisher (1925)
- He thought 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 20^{th} century statistics

- However, there is nothing special about 95% confidence interval or *p*-value of 5%

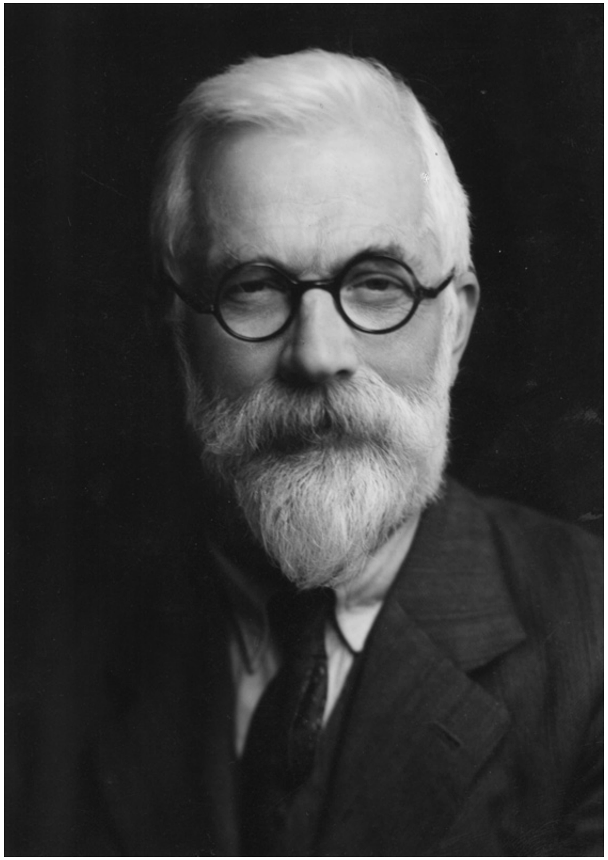


Ronald Fisher

- 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

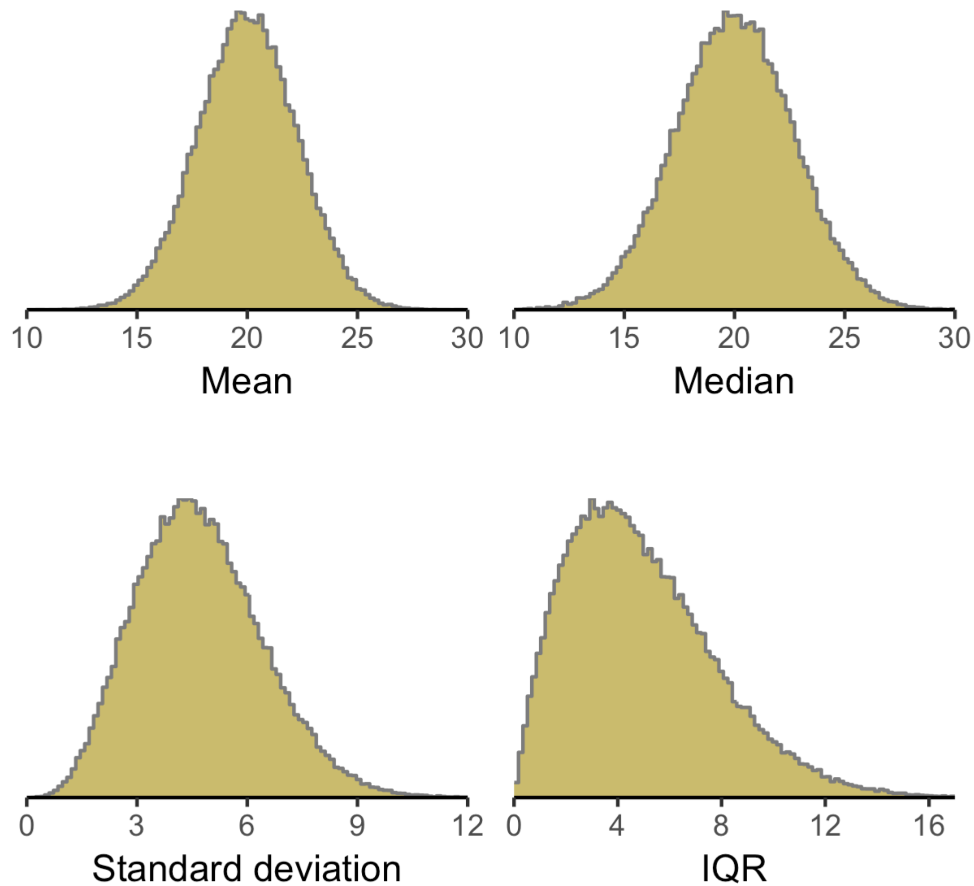
- Didn't like administration and admin people: "an administrator, not the highest form of human life"



Ronald Fisher (1890-1962)

Sampling distribution

- *Gedankenexperiment*
- Consider an unknown population
- Draw lots of samples of size n
- Calculate an estimator from each sample
- Build a frequency distribution of the estimator
- This is a *sampling distribution*
- Width of the sampling distribution is a standard error



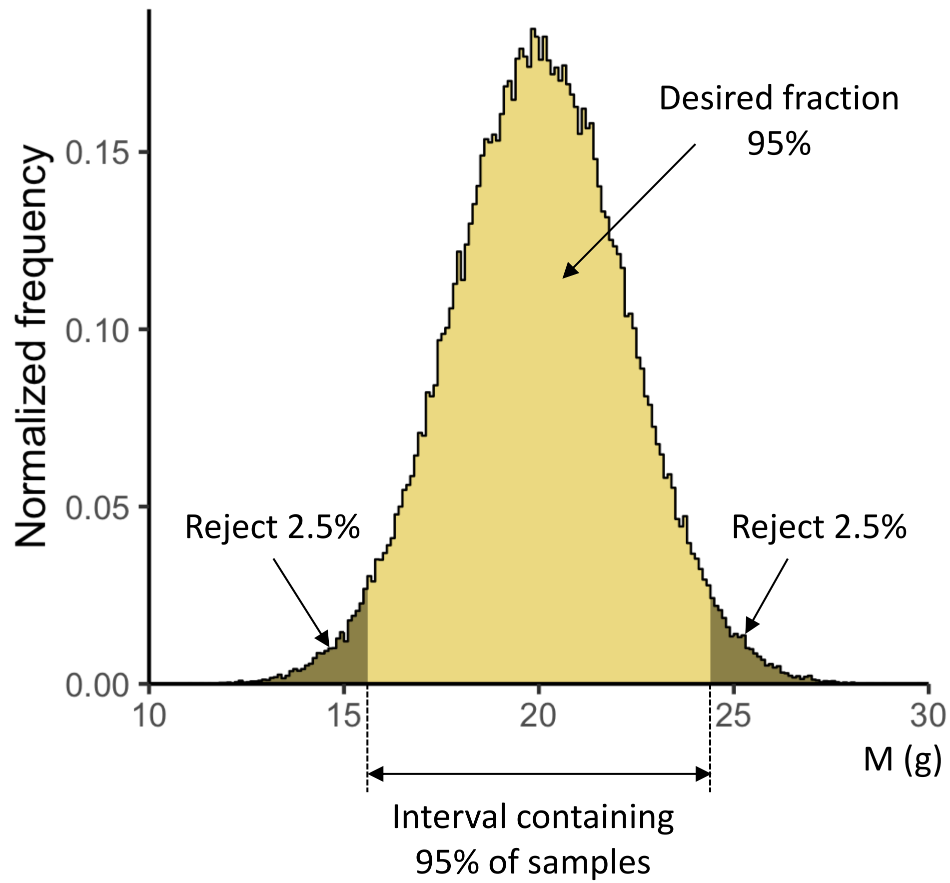
Examples of sampling distribution

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

Confidence interval of the mean

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 a confidence interval of the mean



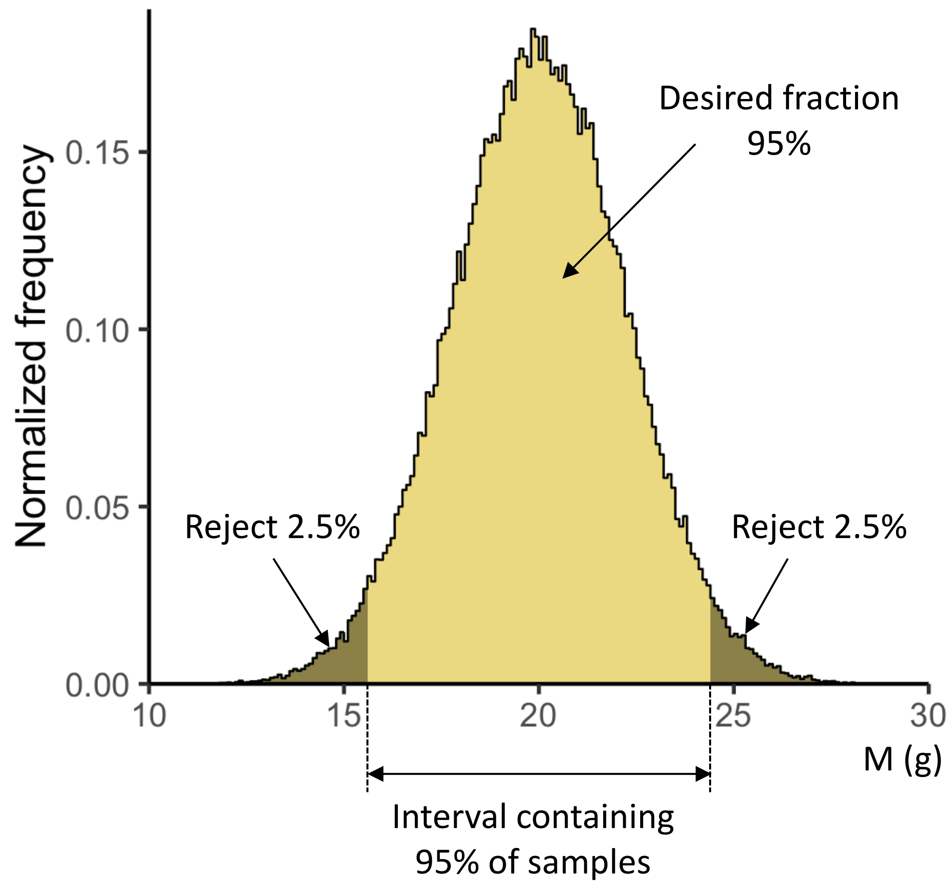
100,000 samples of 5 mice from normal population with $\mu = 20$ g and $\sigma = 5$ g

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 a confidence interval of the mean
- 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

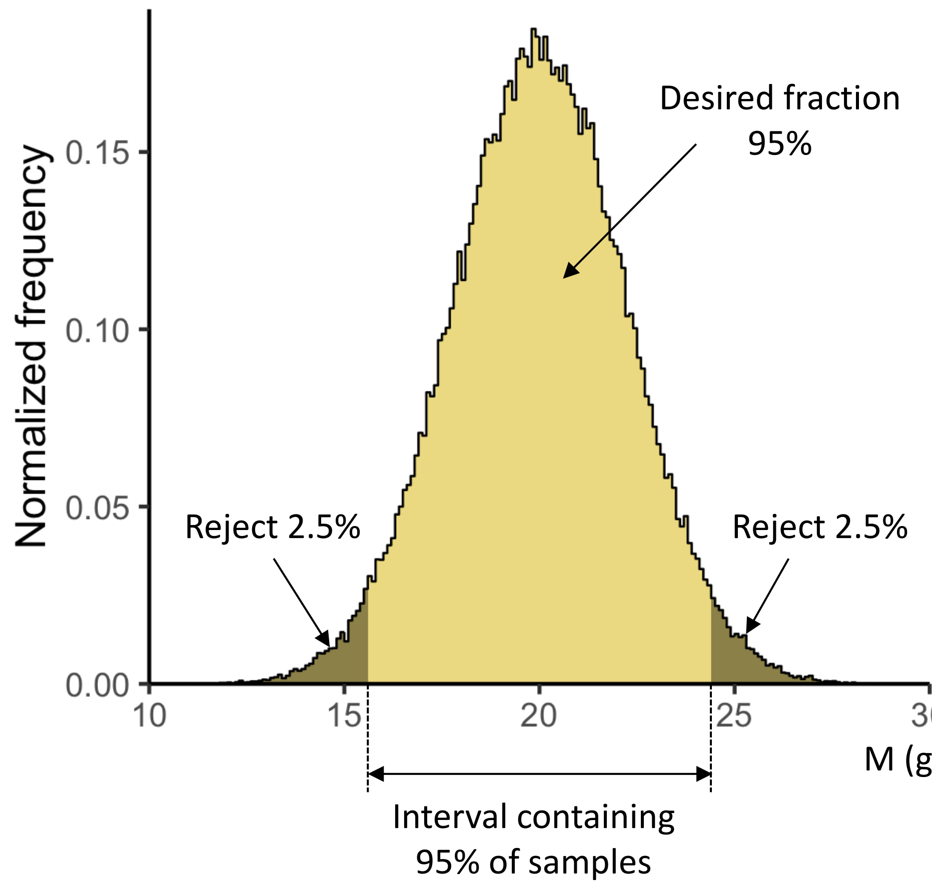
Sampling distribution of the mean

- 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

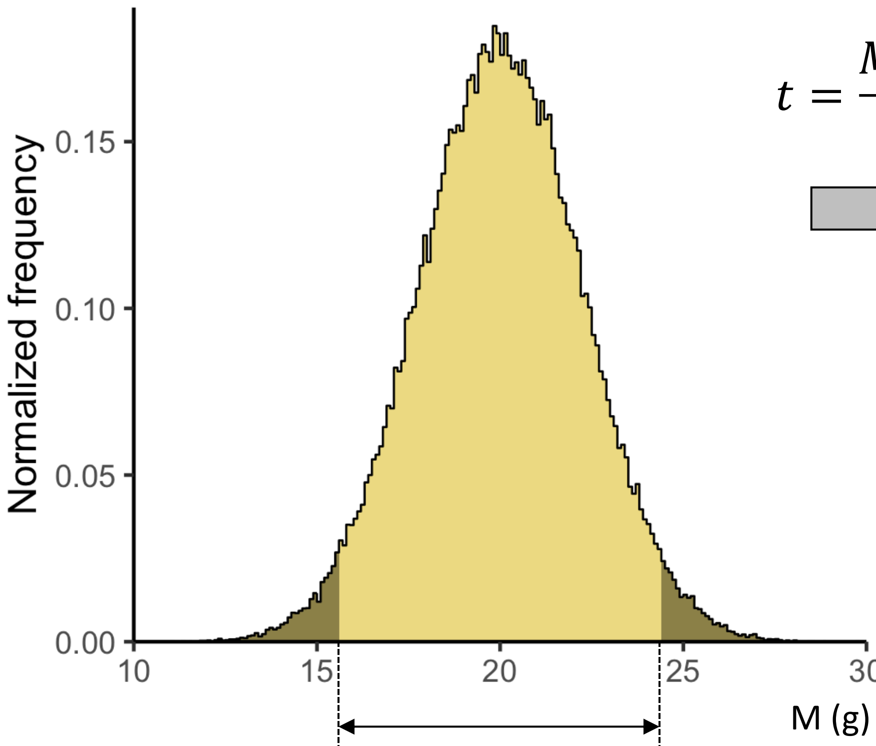


100,000 samples of 5 mice from normal population with $\mu = 20$ g and $\sigma = 5$ g

Mean body weight calculated for each sample

Sampling distribution of t-statistic

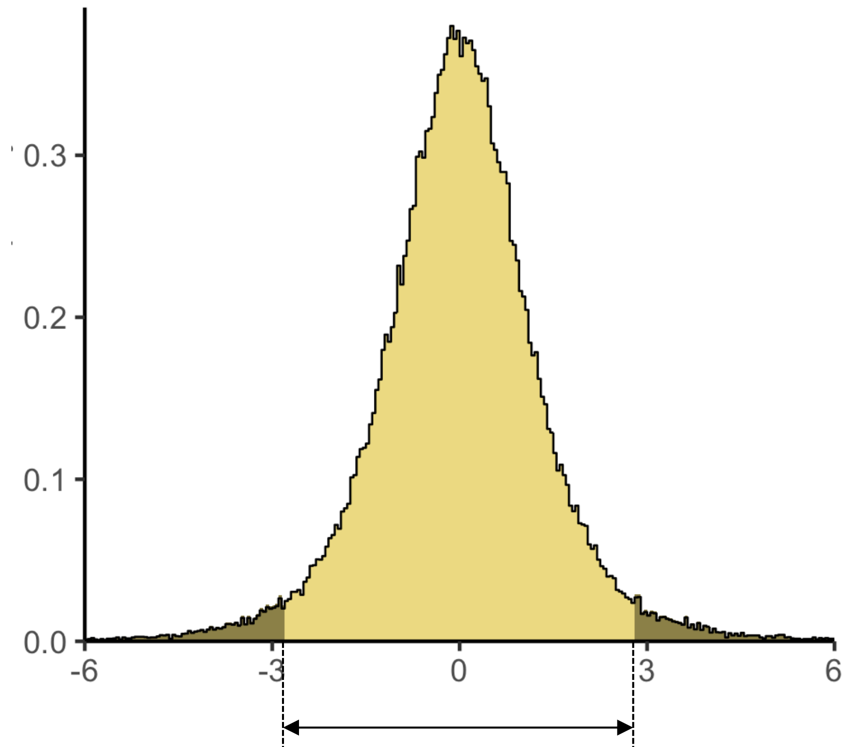
Sampling distribution of M



Confidence interval for μ

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

Sampling distribution of t



Confidence interval for t

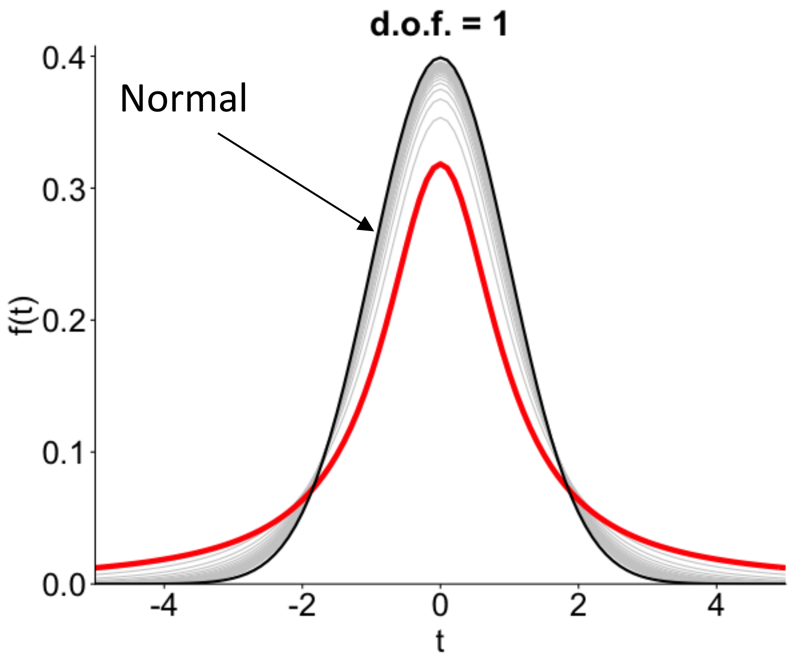
Confidence interval of the mean

- 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!



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?

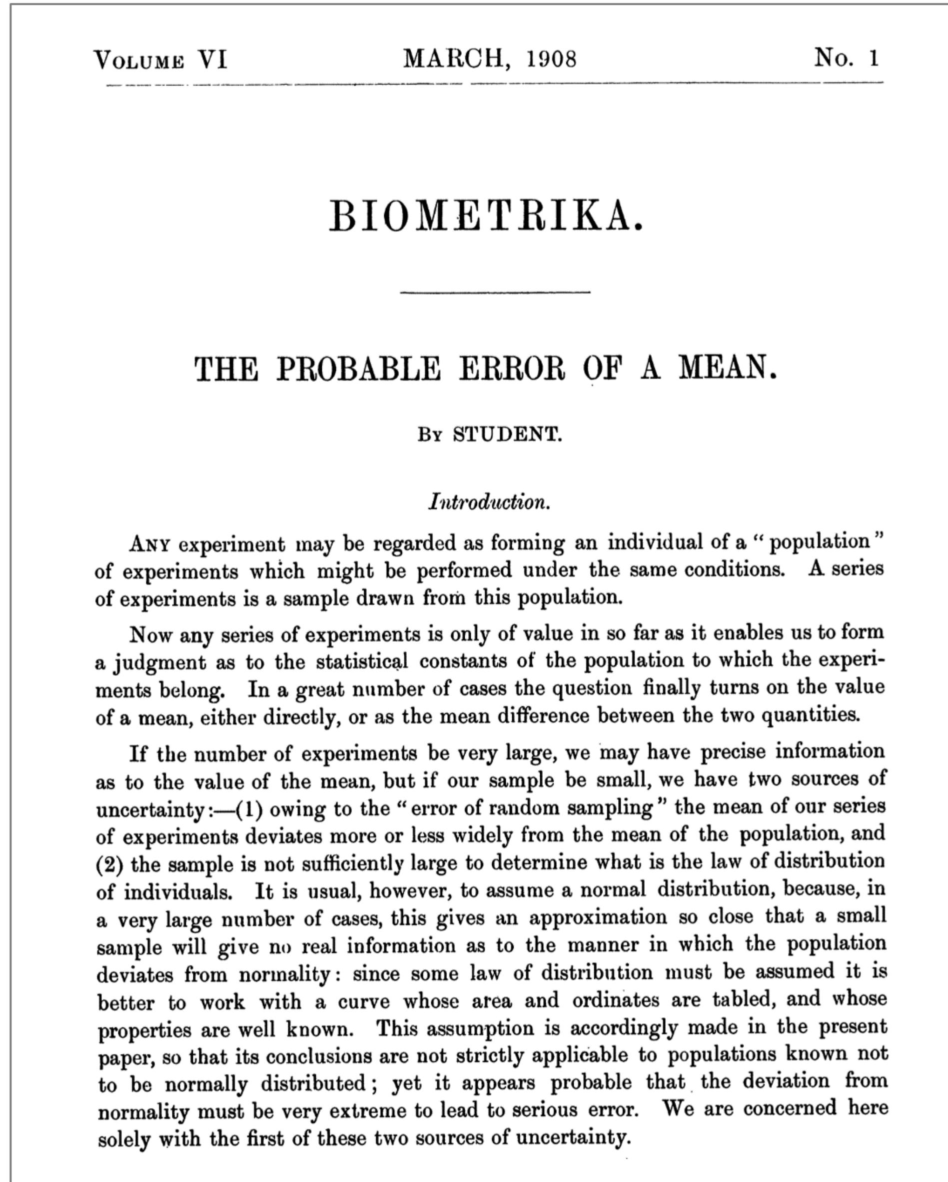
William Sealy Gosset (1876-1937)

William Gosset

- 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?



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 of experiments deviates more or less widely from the mean of the population, and (2) the sample is not sufficiently large to determine what is the law of distribution of individuals. It is usual, however, to assume a normal distribution, because, in a very large number of cases, this gives an approximation so close that a small sample will give no real information as to the manner in which the population deviates from normality: since some law of distribution must be assumed it is better to work with a curve whose area and ordinates are tabled, and whose properties are well known. This assumption is accordingly made in the present paper, so that its conclusions are not strictly applicable to populations known not to be normally distributed; yet it appears probable that the deviation from normality must be very extreme to lead to serious error. We are concerned here solely with the first of these two sources of uncertainty.

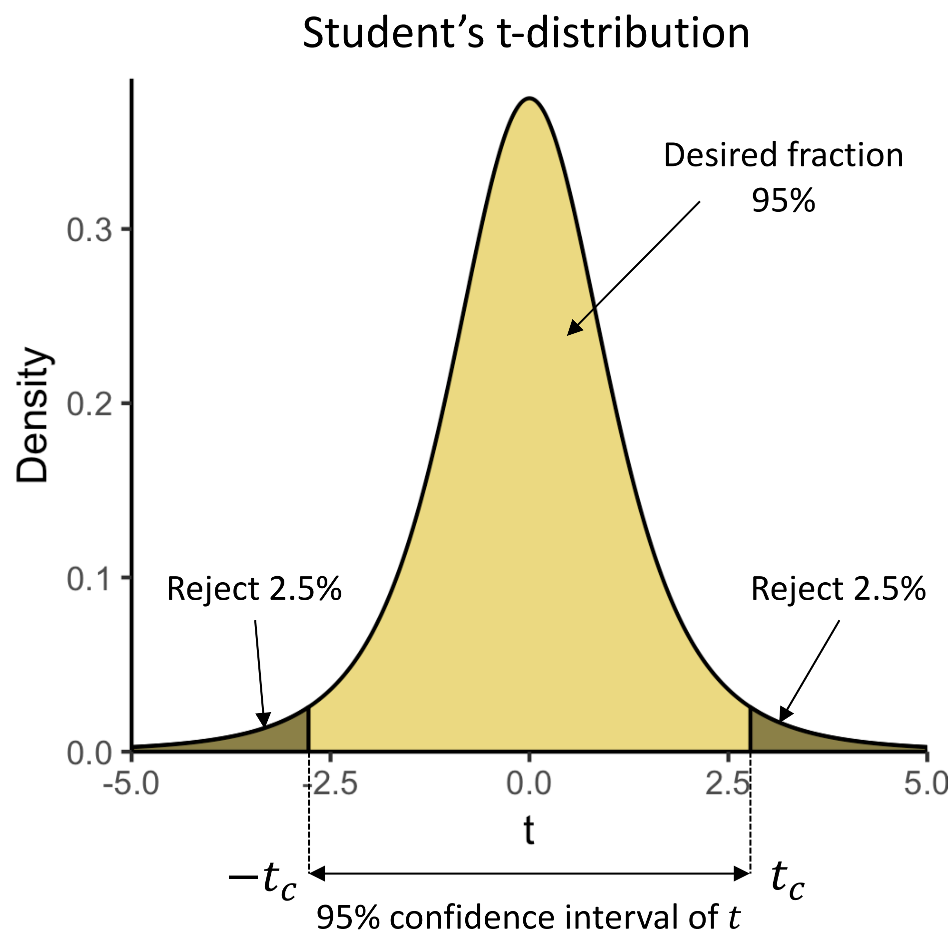
Confidence interval of the mean

- 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_c to cut off required confidence interval
- R function qt
- Confidence interval on t is $[-t_c, +t_c]$



Confidence interval of the mean

- We used transformation

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

- Confidence interval on t is $[-t_c, +t_c]$

- Find μ from the equation above

$$\mu = M + tSE$$

- From limits on t we find limits on μ:

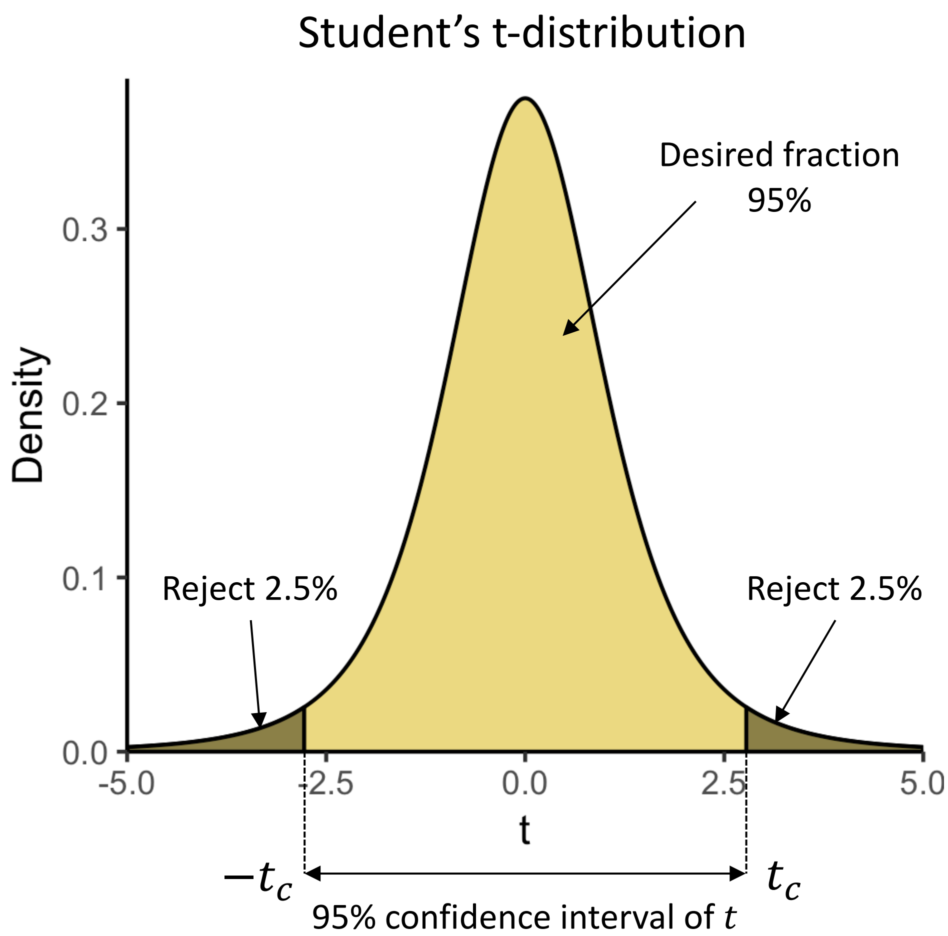
$$M_L = M - t_c SE$$
$$M_U = M + t_c SE$$

- Or

$$\mu = M \pm CI$$

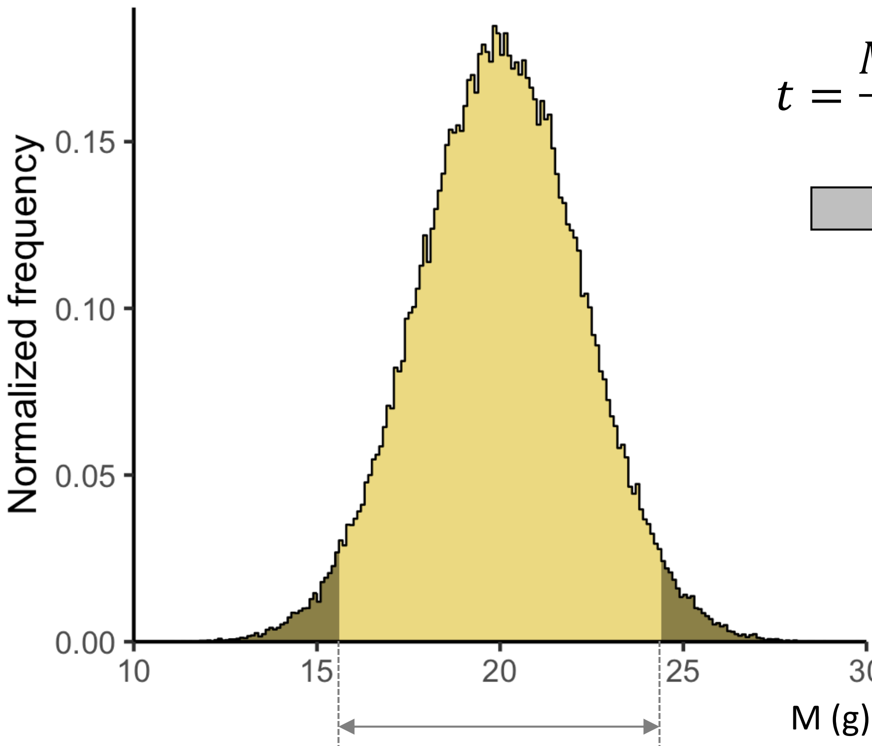
where confidence interval is a scaled standard error

$$CI = t_c SE$$



How to use t-distribution to get 95% CI of the mean

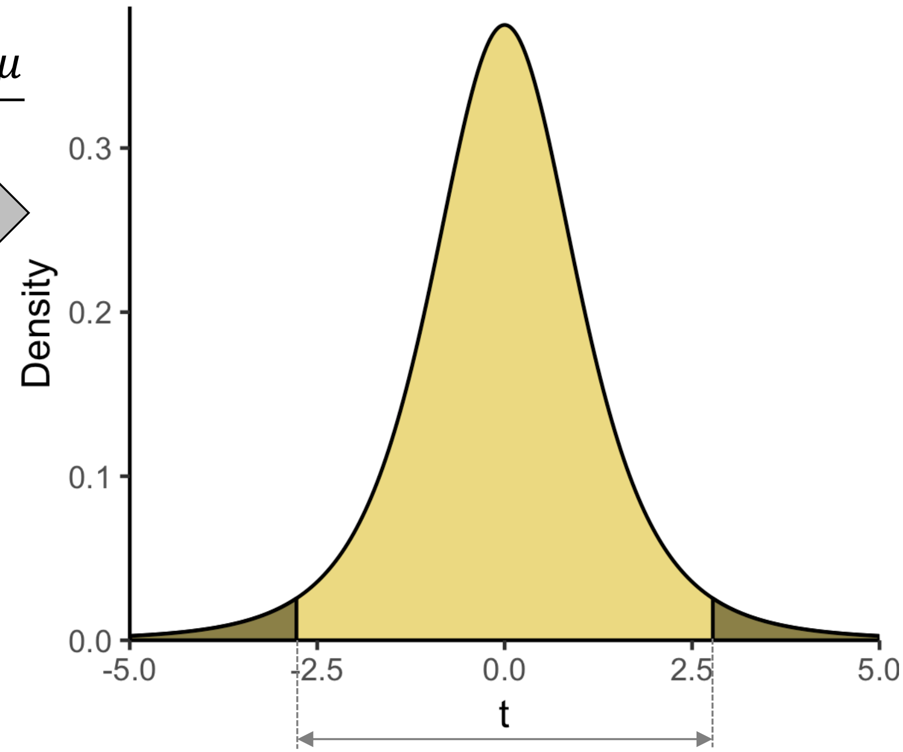
Sampling distribution of M



M_L Confidence interval for μ M_U

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

Student's t-distribution
$n - 1$ degrees of freedom



$-t_c$ Confidence interval for t t_c

$$M_L = M - t_c SE$$
$$M_U = M + t_c SE$$

Exercise: 95% confidence interval for the mean

- We have 5 mice with measured body weights 16.8, 21.8, 29.2, 23.3 and 26.3 g
- Estimators from the sample

$$
\begin{aligned}
M &= 23.48\text{ g} \\
SD &= 4.69\text{ g} \\
SE &= 2.10\text{ g}
\end{aligned}
$$

- Critical value from t-distribution for two-tail probability and 4 degrees of freedom

$$t_c = 2.776$$

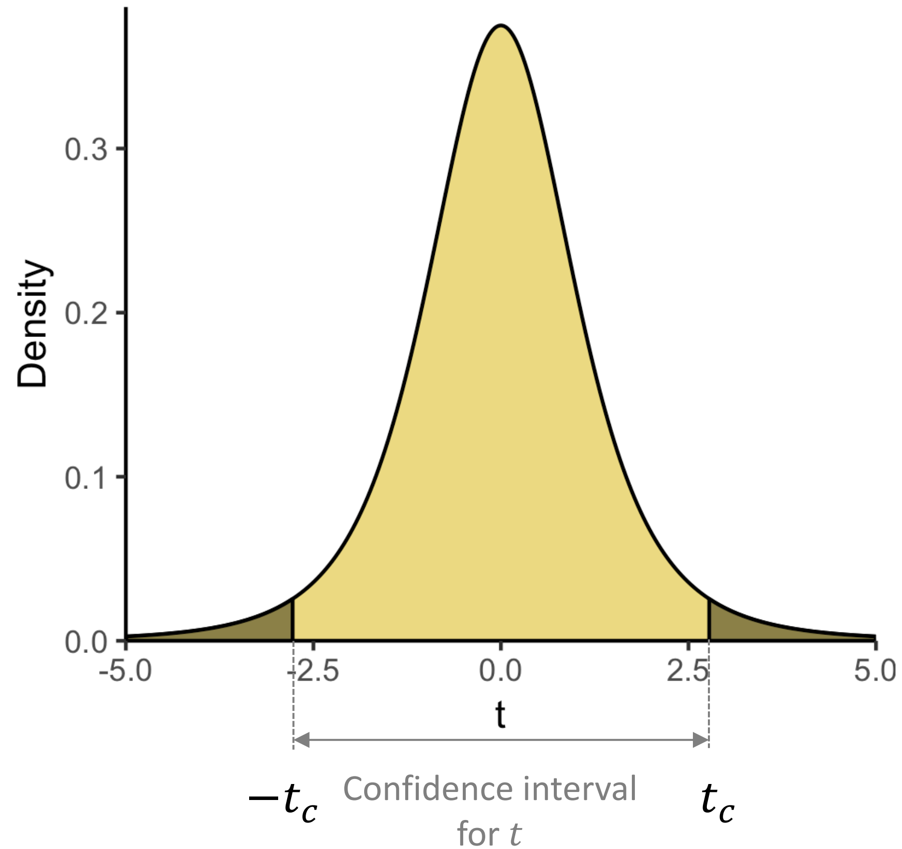
- Confidence limits are

$$
\begin{aligned}
M_L &= M - t_c SE = 17.65\text{ g} \\
M_U &= M + t_c SE = 29.31\text{ g}
\end{aligned}
$$

- Estimate of the mean with 95% confidence is

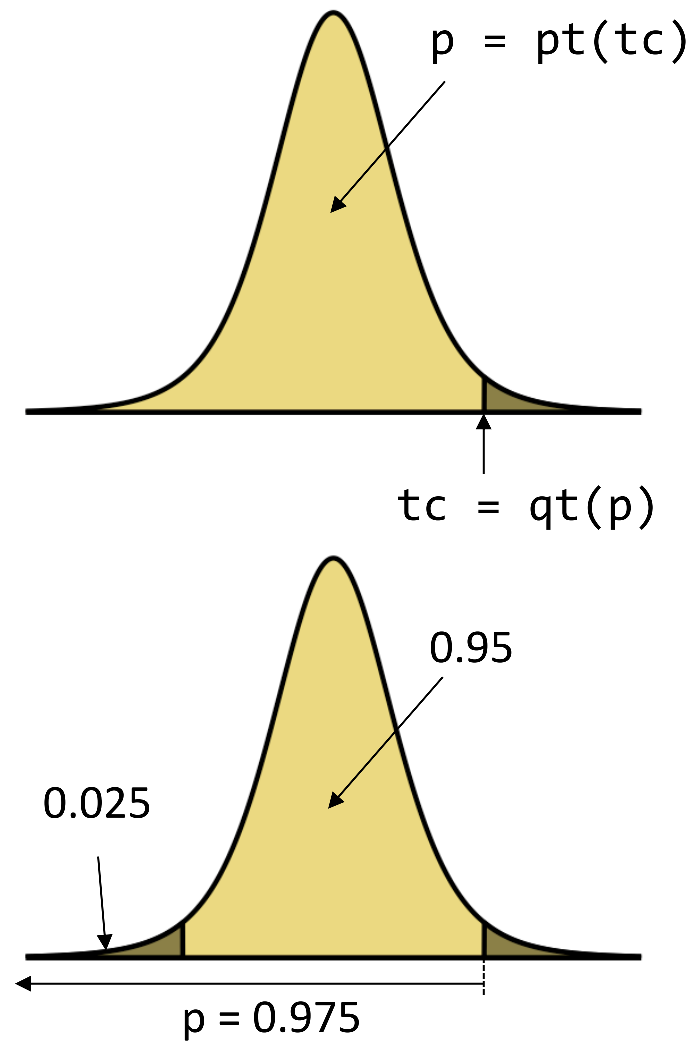
$$\mu = 23 \pm 6\text{ g}$$

$$\cancel{23.48 \pm 5.83\text{ g}}$$



Confidence interval in R

```
> d <- c(16.8, 21.8, 29.2, 23.3, 26.3)
> n <- length(d)
> M <- mean(d)
> SE <- sd(d) / sqrt(n)
# critical t
> tc <- qt(0.975, df = n - 1)
# lower confidence limit
> M - tc * SE
[1] 17.65118
# upper confidence limit
> M + tc * SE
[1] 29.30882
```




```
tc = qt(0.975, df = 4)
[1] 2.776445
```

Confidence interval in R: the simple way

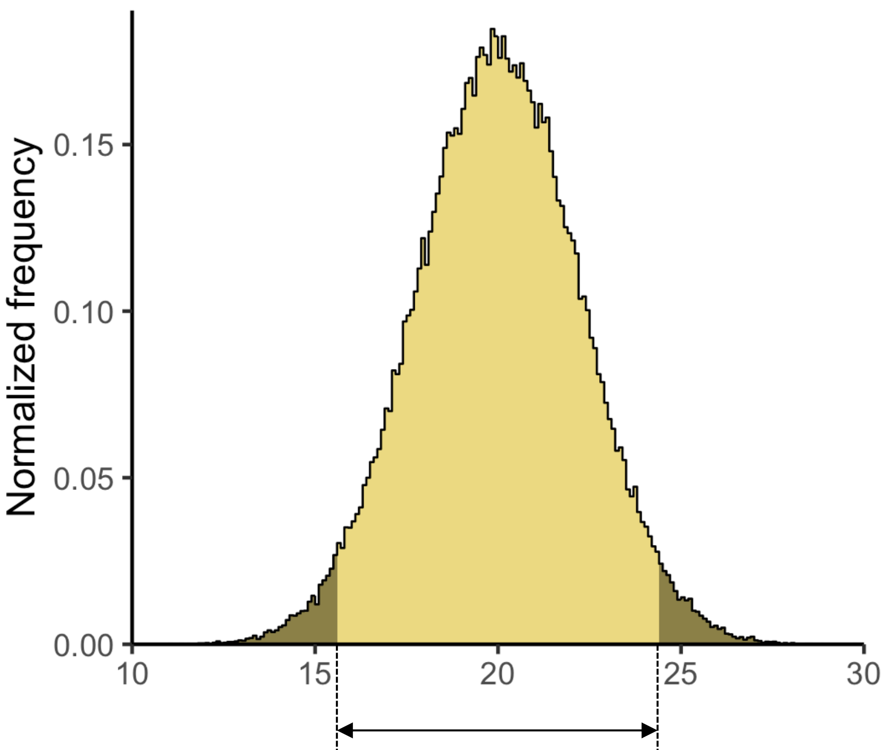
```
> d <- c(16.8, 21.8, 29.2, 23.3, 26.3)
> t.test(d)

        One Sample t-test

data:  d
t = 11.184, df = 4, p-value = 0.0003639
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 17.65118 29.30882
sample estimates:
mean of x
    23.48
```

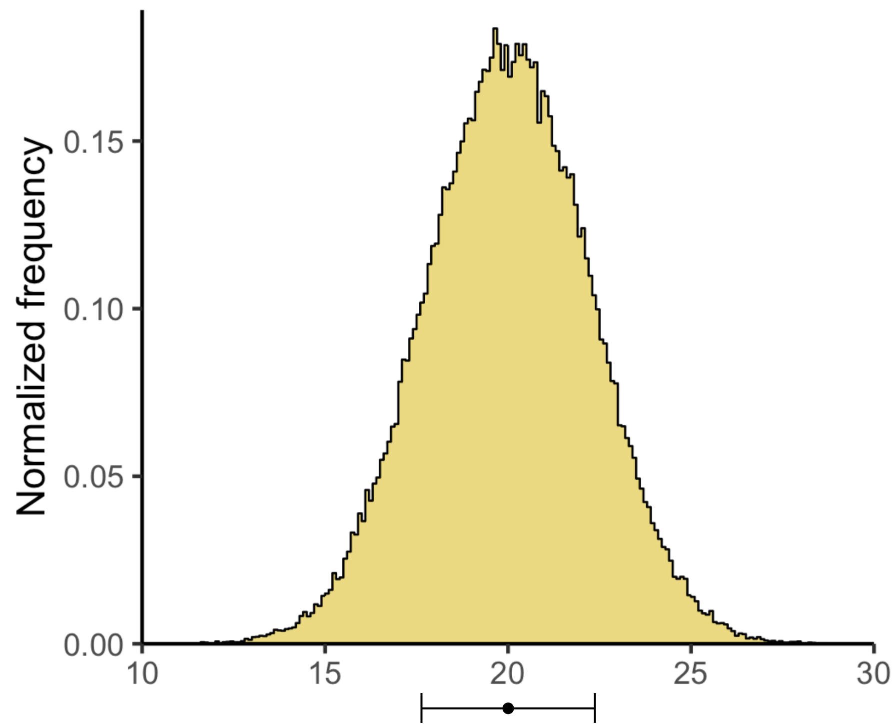
Confidence interval vs. standard error

Confidence interval



Interval containing
95% of samples

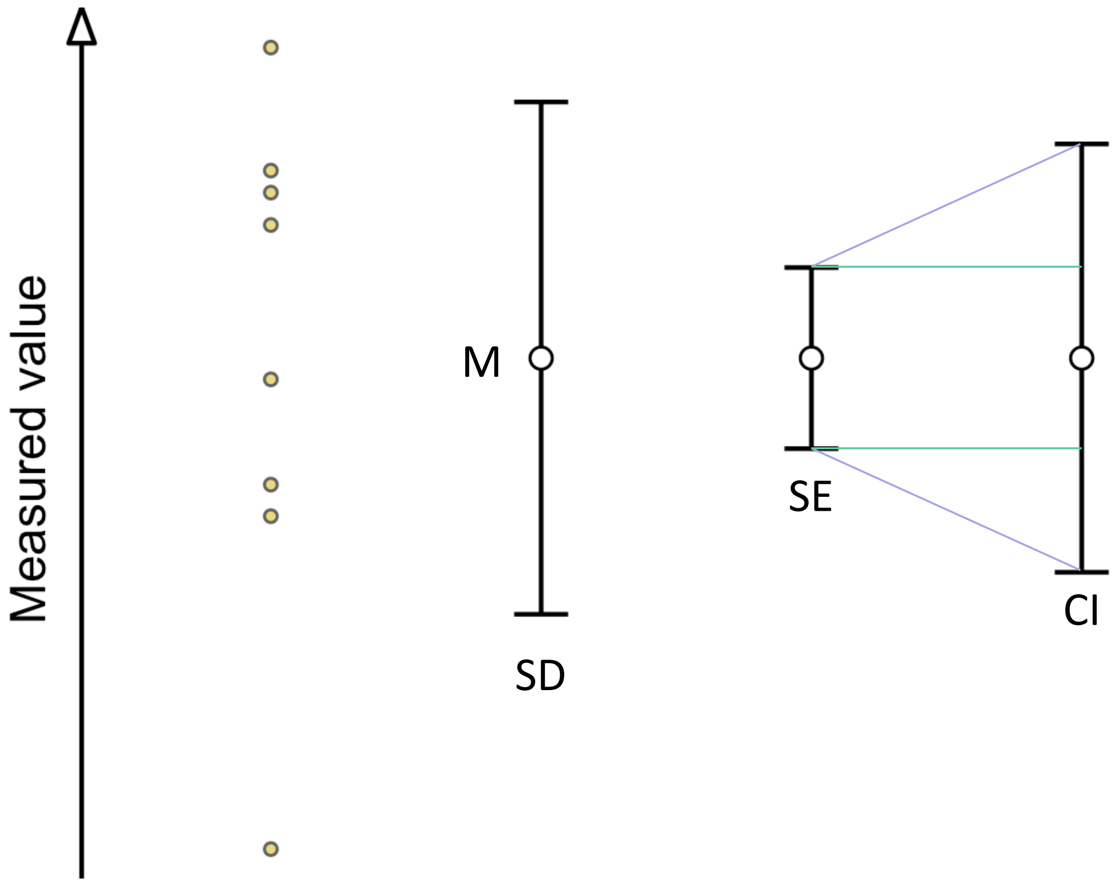
Standard error



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

Width of the
sampling distribution

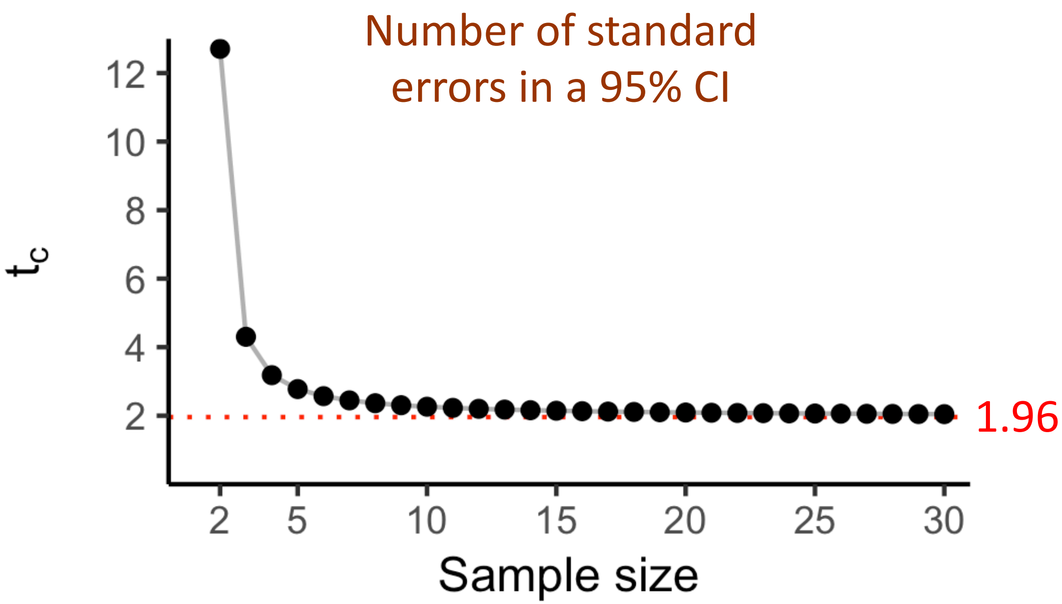
Confidence interval vs standard error



How many standard errors are in a confidence interval?

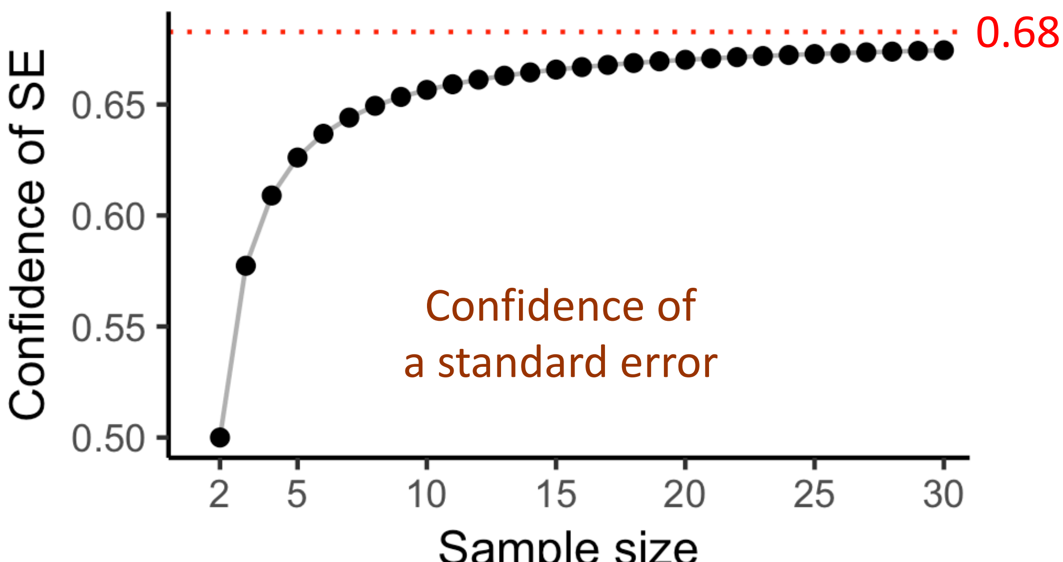
What is the confidence of the standard error?

Confidence interval vs standard error



Large samples:

$95\%\ CI \approx 2\ SE$



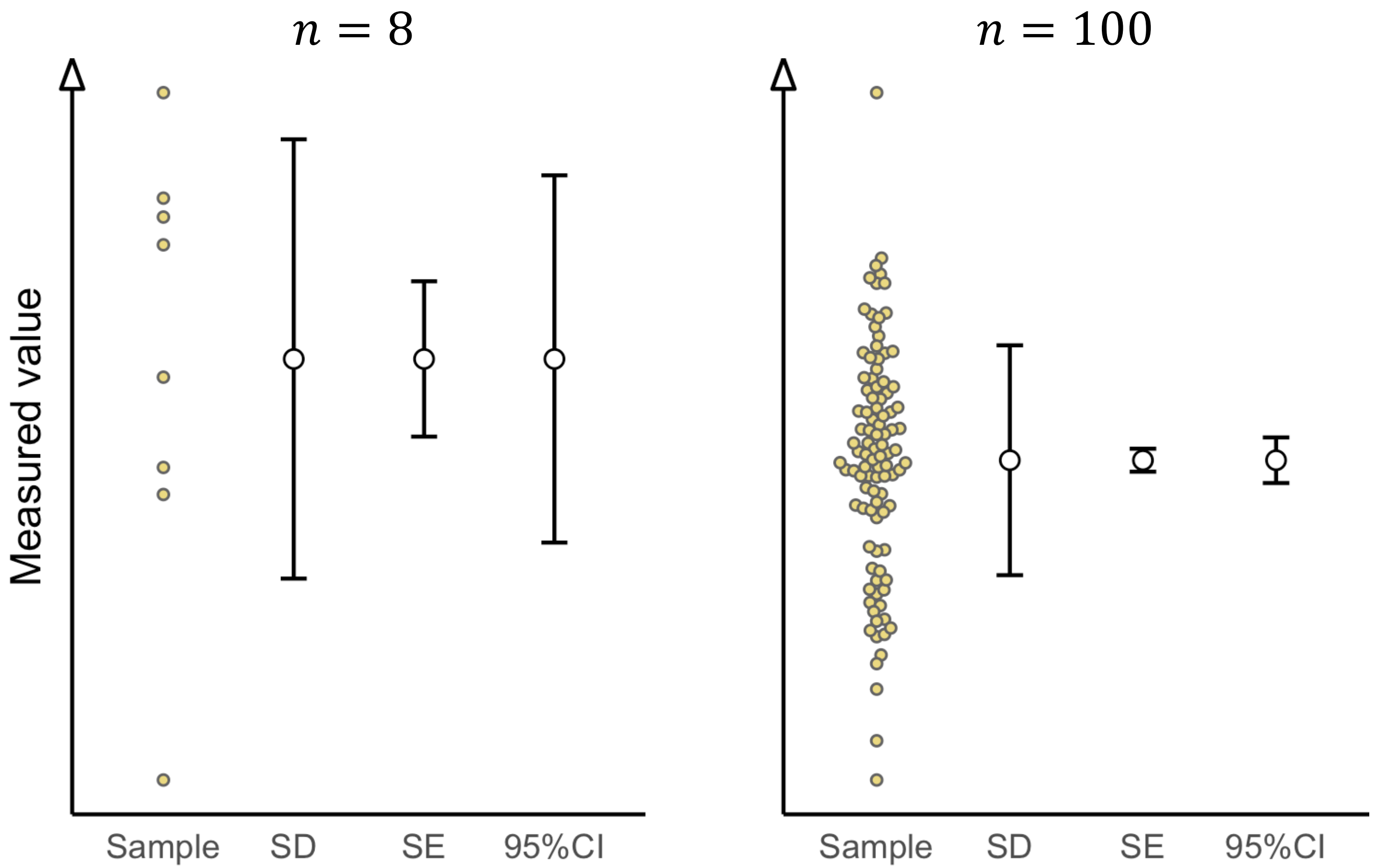
Confidence of SE is ~68%

YOU NEED

more

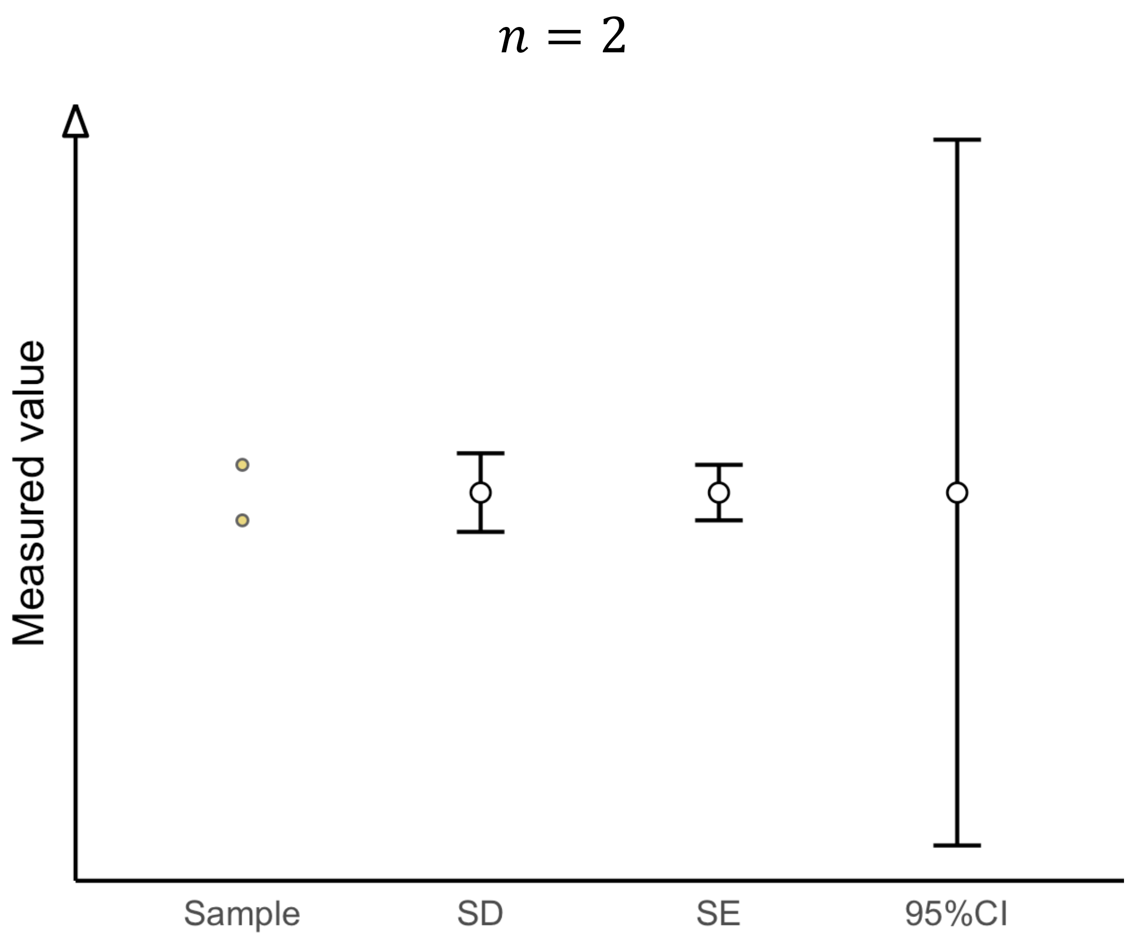
REPLICATES

SD, SE and 95% CI



- Normal population of $\mu = 20$ g and $\sigma = 5$ g
- Sample of $n = 8$ and $n = 100$

2 replicates? NO!



Example: confidence intervals

- Experiment where a reporter measures transcriptional activity of a gene
  - Day 1: 3 biological replicates
  - Day 2: 5 biological replicates
- Normalized data:

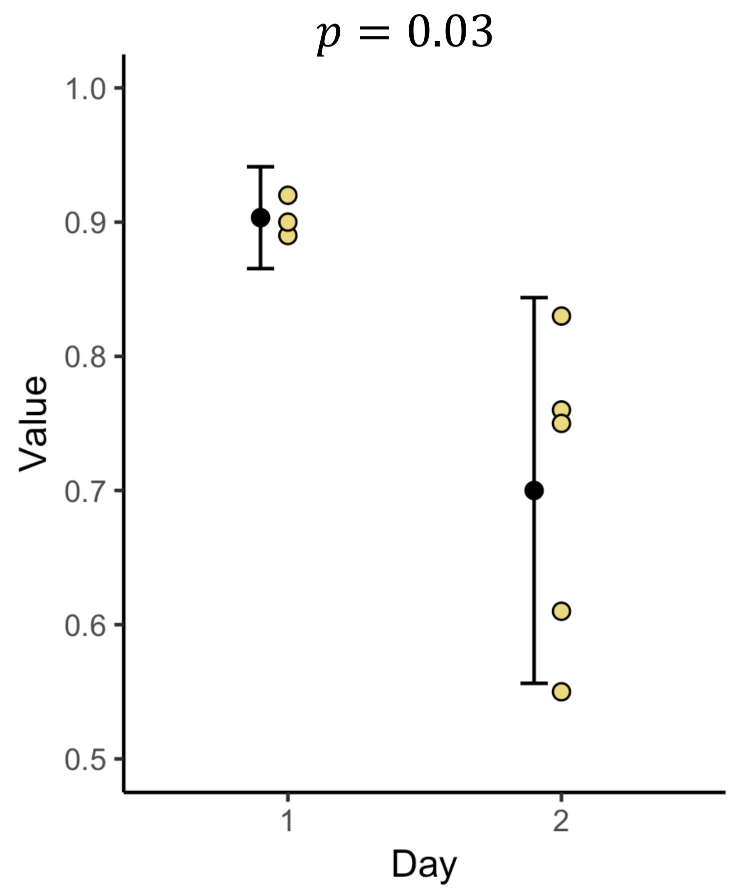
Day 1	0.89	0.92	0.90		
Day 2	0.55	0.76	0.61	0.83	0.75

- 95% confidence intervals for the mean:

  Day 1: [0.87, 0.94]

  Day 2: [0.56, 0.84]

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



Confidence interval of the median

Confidence interval of the median

- We do *not* build a sampling distribution
- Draw one random sample of n points, one by one: $x_1, x_2, \dots, x_n$
- Population median θ property: $P(x_i < \theta) = \frac{1}{2}$ and $P(x_i > \theta) = \frac{1}{2}$
- For each data point we have fifty-fifty chance

1. Let true median $\theta = 20$

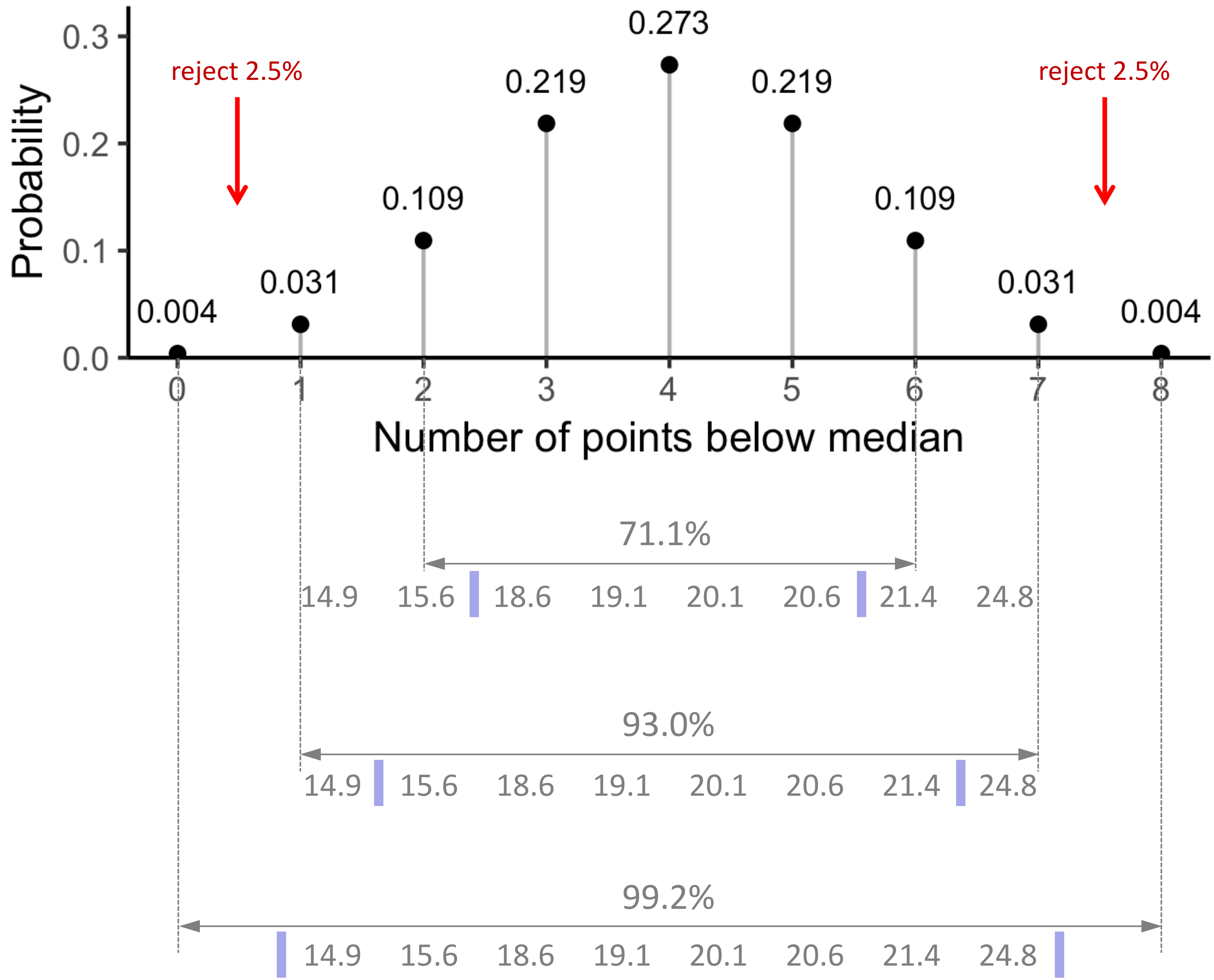
14.9 **15.6** **18.6** **19.1** **20.1** **20.6** **21.4** **24.8**

$P = 0.273$

2. Let true median $\theta = 15$

$P = 0.031$

Limited confidence intervals of the median



Confidence interval of the median - interpolation

- Approach based on all pairs of data points and interpolation
- Hodges-Lehmann estimator

```
> x <- c(14.9, 15.6, 18.6, 19.1, 20.1, 20.6, 21.4, 24.8)
> wilcox.test(x, conf.int = TRUE)

        Wilcoxon signed rank test

data:  x
V = 36, p-value = 0.007813
alternative hypothesis: true location is not equal to 0
95 percent confidence interval:
 16.75 22.45
sample estimates:
(pseudo)median
          19.6
```

Replicates

- Replication is the repetition of an experiment under the same conditions
- Typically, the only way of estimating measurement errors is to do the experiment in replicates
- You need replicates, but how many?

- Statistical power
- Roughly speaking, there are two cases
  - to get an estimate with a required precision
  - to get enough sensitivity for differential analysis

Number of replicates to find the mean

- Sampling distribution of the mean has $\sigma_m = \sigma/\sqrt{n}$
- Interval $\sim 2\sigma_m$ around the true mean contains 95% of all samples
- Let's call it precision of the mean:

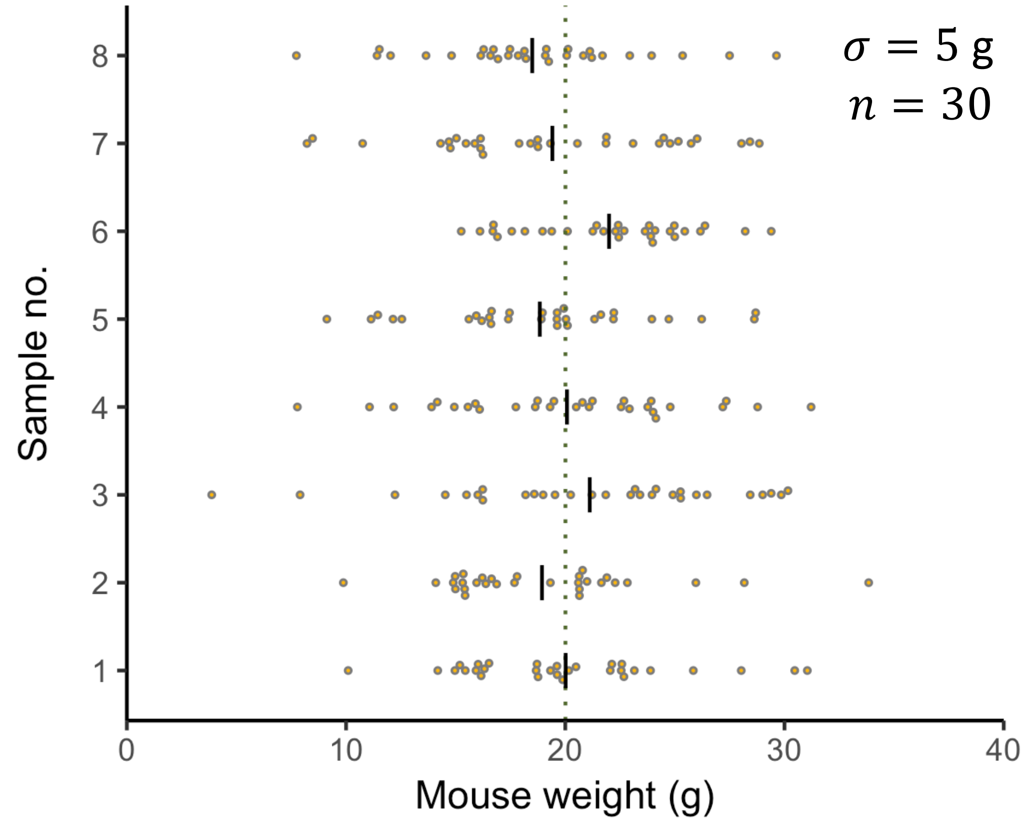
$$\epsilon \approx 2\sigma_m = \frac{2\sigma}{\sqrt{n}}$$

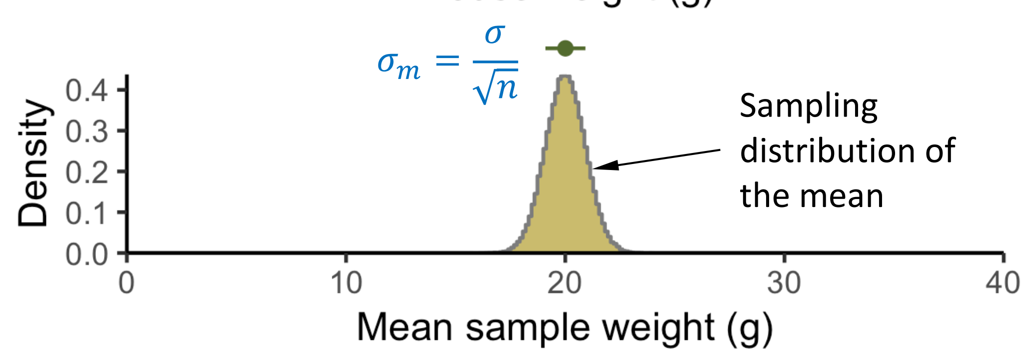
- Sample size to get the required precision:

$$n = \frac{4\sigma^2}{\epsilon^2}$$

- This requires a priori knowledge of σ (do a pilot experiment to estimate)
- Example: $\sigma = 5$ g, required precision of ± 2 g

$$n = 4 \times \frac{5^2}{2^2} = 25$$





Hand-outs available at
https://dag.compbio.dundee.ac.uk/training/Statistics_lectures.html