Error analysis in biology

Marek Gierliński
Division of Computational Biology

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

Previously on Errors…

Confidence intervals (CI)

- probabilistic measure of uncertainty
- in 95% of repeated experiments the true parameter is within 95% CI
- better than standard error

Sampling distribution

- distribution of a sample statistic
- idea: central 95% of samples gives us a confidence interval

CI of the mean

- a statistic

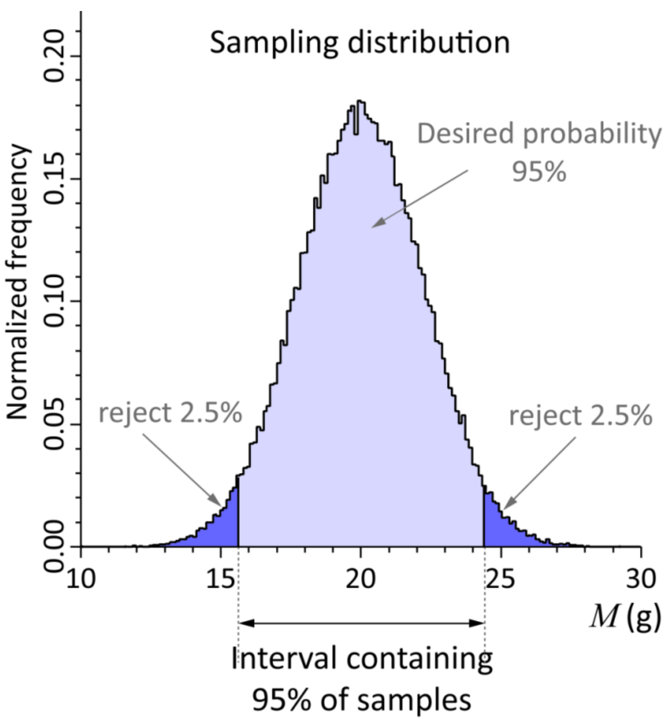
$$t = \frac{M - \mu}{SE}$$

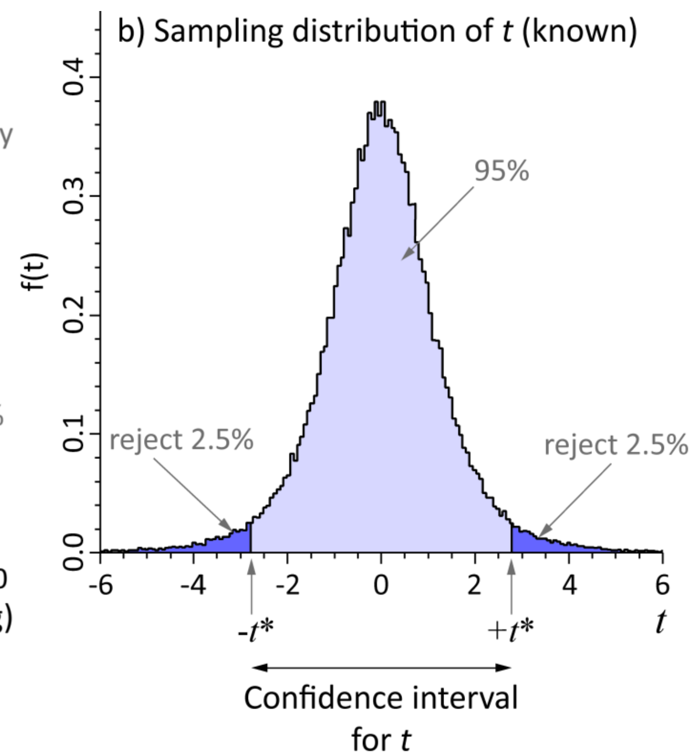
- has known sampling distribution
- Student's t-distribution
- CI of the mean:

$$CI = t^{*} SE$$

CI of the median

- calculated from the binomial distribution
- a simple approximation given





4. Confidence intervals II

"Confidence is what you have before you understand the problem"

Woody Allen

Confidence interval for count data

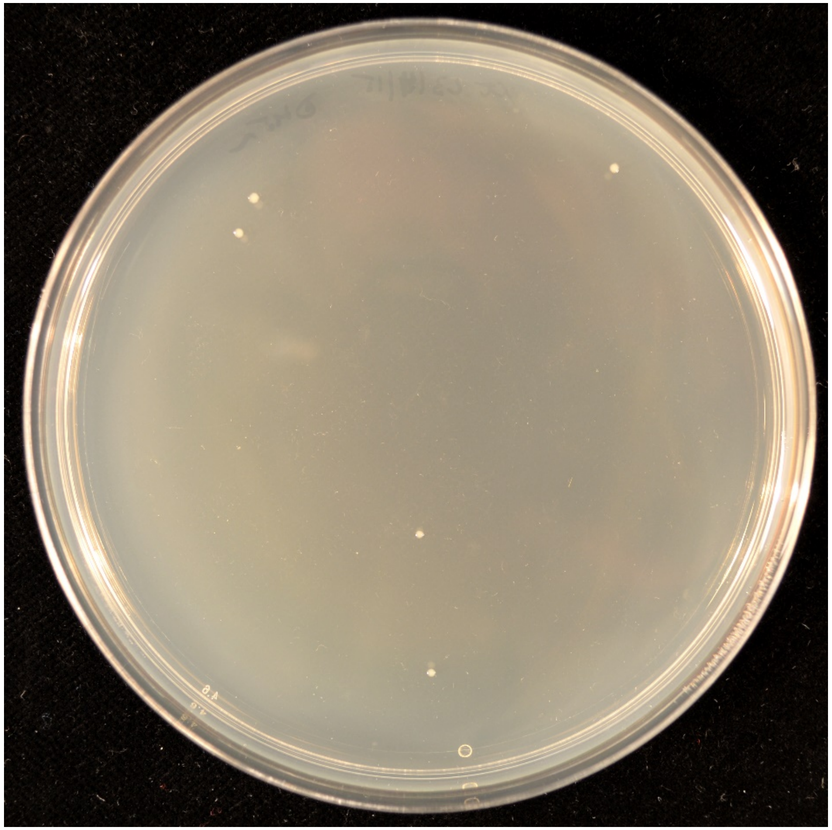
- Standard error of a count, C, is

$$SE = \sqrt{C}$$

- For example 5 ± 2 (after rounding up)

- How to find a confidence interval on μ?
- Exact method: a bit complicated

- We have a good approximation!



$C = 5 \pm 2$ (SE)

Gehrels, N. 1986. Confidence-Limits for Small Numbers of Events in Astrophysical Data. *Astrophysical Journal,* 303**,** 336-346

Confidence interval for count data: approximation

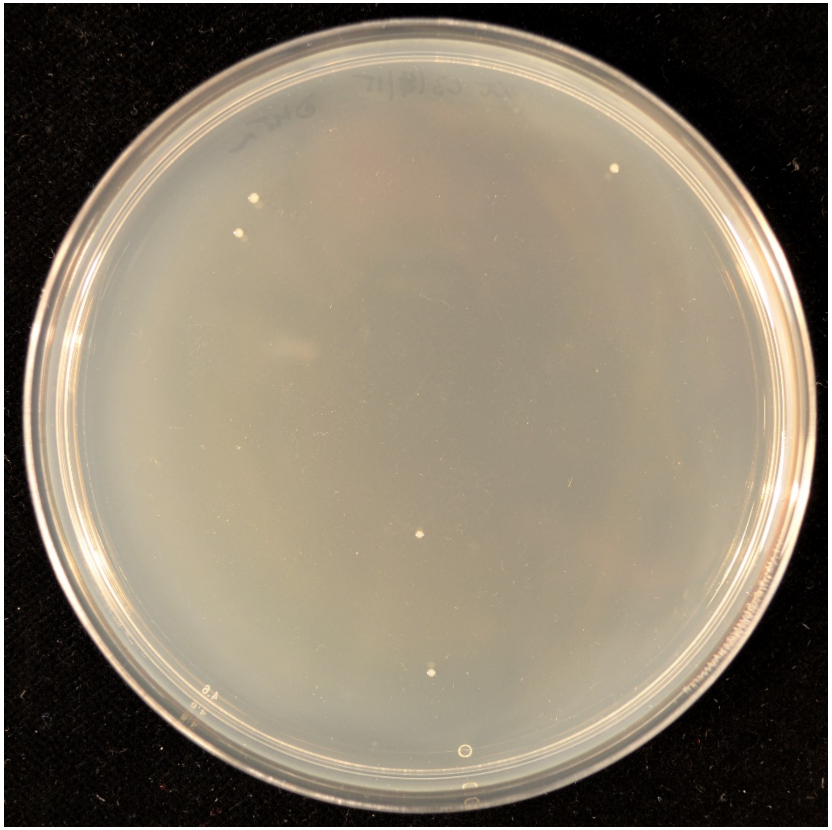
- For the given confidence level find a Gaussian critical value Z
  - for example $Z = 1.96$ for 95% CI

- For the given count number, C, calculate lower and upper limits:

$$C_L = C - Z\sqrt{C} + \frac{Z^2 - 1}{3}$$

$$C_U = C + Z\sqrt{C+1} + \frac{Z^2 + 2}{3}$$

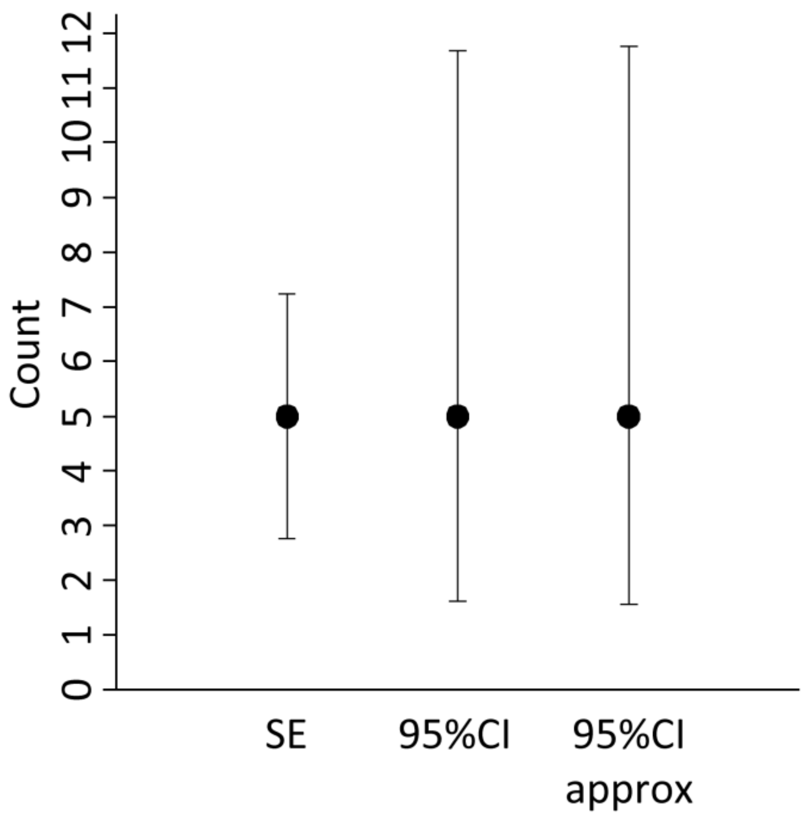
- Example:
  - $C = 5$
  - $Z = 1.96$
  - $C_L = 1.6, C_U = 11.8$
  - $C = 5^{+7}_{-3}$
  - It is asymmetric!

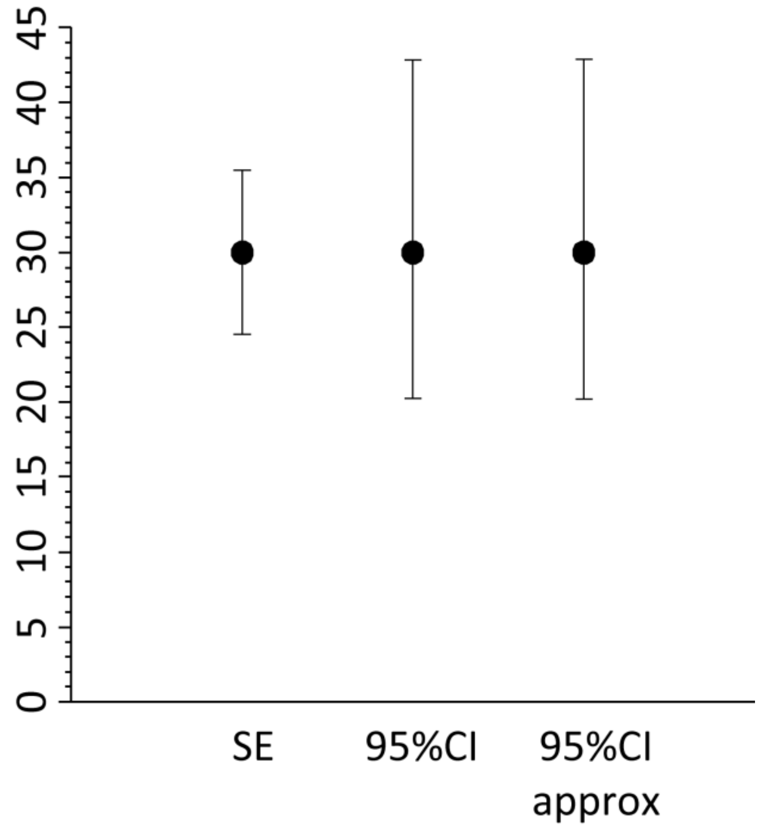


$C = 5 \pm 2$ (SE)

$C = 5^{+7}_{-3}$ (95% CI)

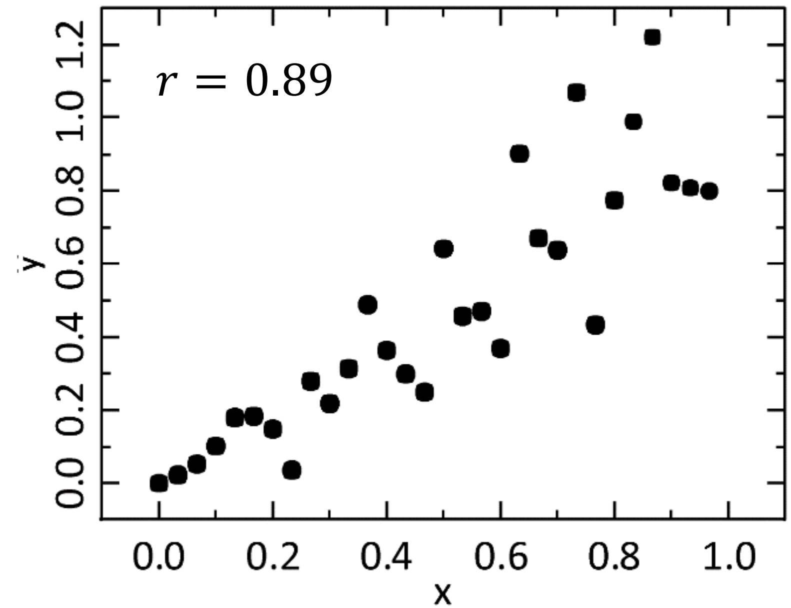
Count errors: example





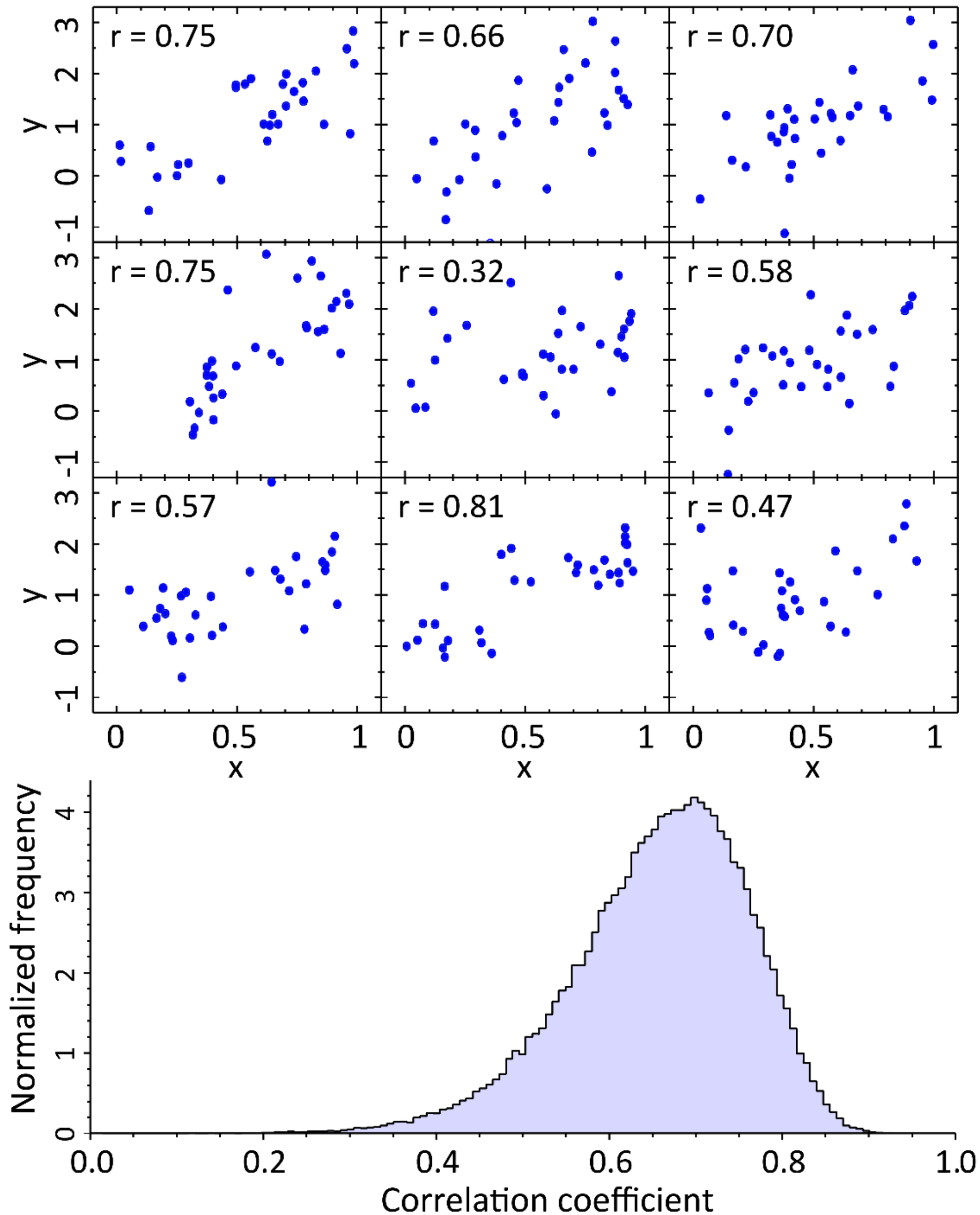
Confidence interval of the correlation coefficient

- Pearson's correlation coefficient r for a sample of pairs (x_i, y_i)
- It is a number between -1 and 1
- It is not enough to say "we find $r = 0.89$, therefore our samples are correlated"
- Confidence limits on r **or** significance of correlation



Sampling distribution of the correlation coefficient

- *Gedankenexperiment*
- Consider a population of pairs of numbers (x_i, y_i)
- The (unknown) population correlation coefficient, $\rho = 0.66$
- Draw lots of samples of pairs, size n
- Calculate the correlation coefficient for each sample
- Build a sampling distribution of the correlation coefficient



Sampling distribution of the correlation coefficient

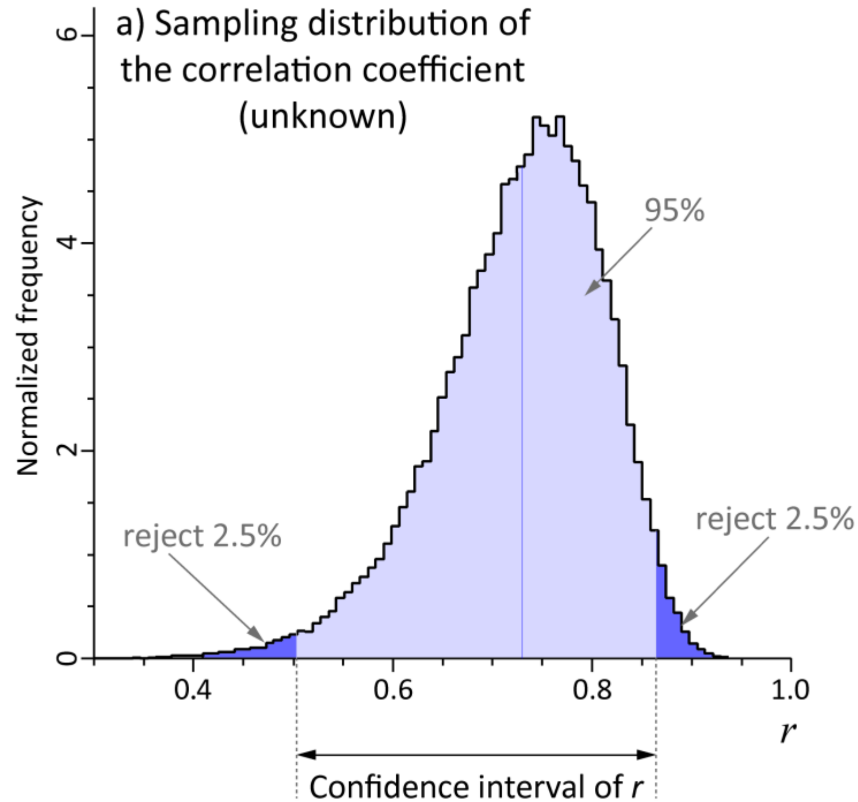
- Sampling distribution of r
- Unknown in analytical form

- Let us transform it into a known distribution

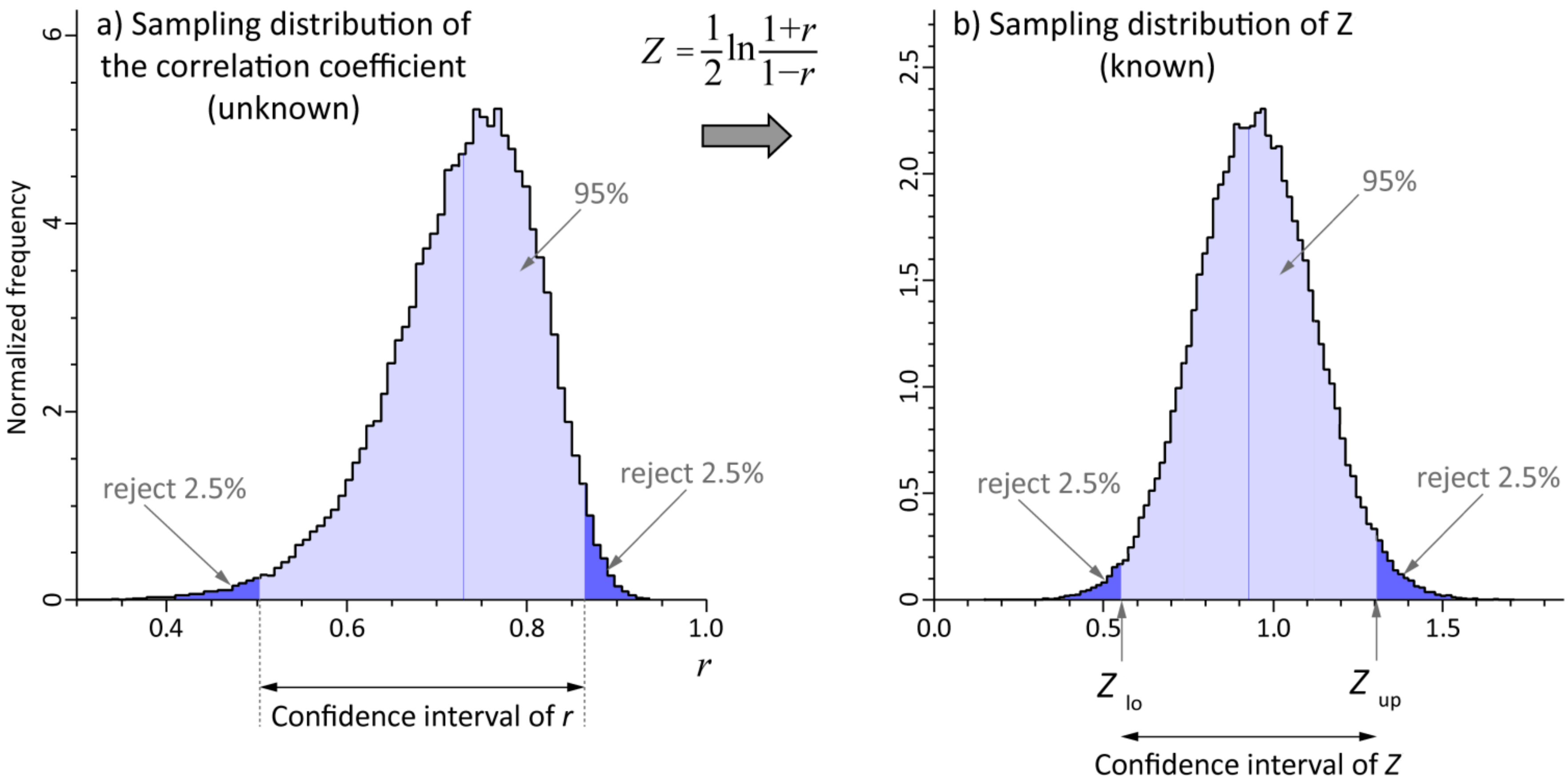
- Fisher's transformation:

$$Z = \frac{1}{2}\ln\frac{1+r}{1-r}$$

- Build a sampling distribution of Z



Confidence interval of the correlation coefficient



Gaussian with standard deviation

$$\sigma = \frac{1}{\sqrt{n-3}}$$

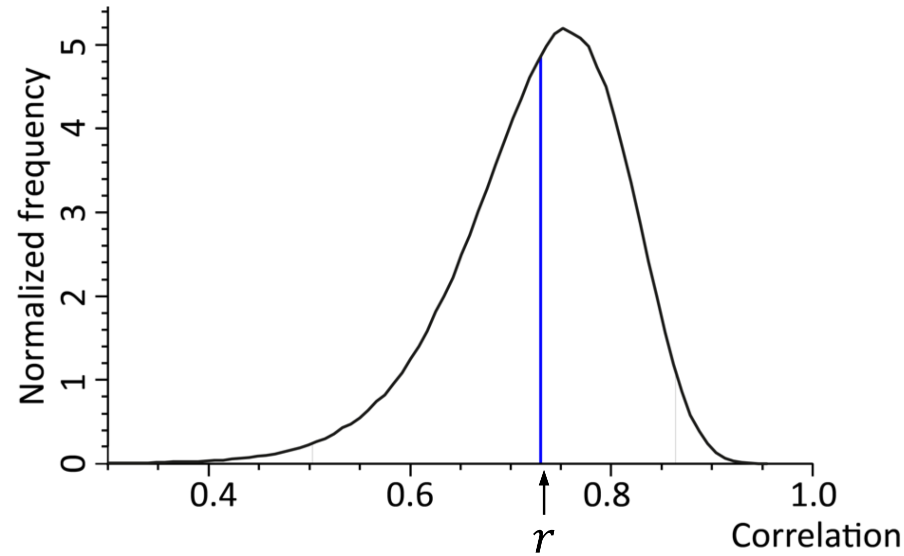
Example: 95% confidence limits on r

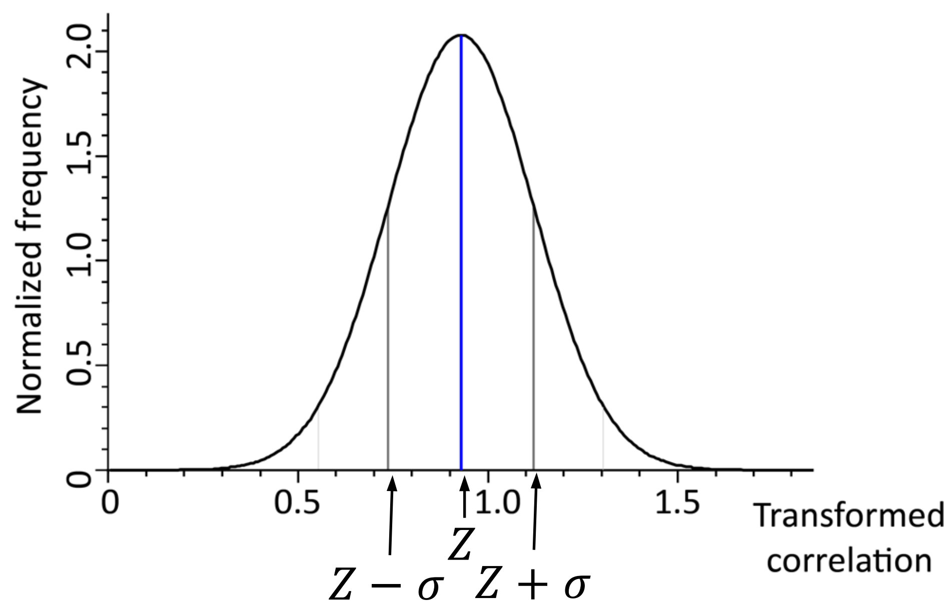
- A sample of $n = 30$ pairs of numbers, correlation coefficient $r = 0.73$
- First, find

$$Z = \frac{1}{2}\ln\frac{1+r}{1-r} = 0.929$$

$$\sigma = \frac{1}{\sqrt{n-3}} = 0.192$$

- Z is normally distributed





Example: 95% confidence limits on r

- A sample of $n = 30$ pairs of numbers, correlation coefficient $r = 0.73$
- First, find

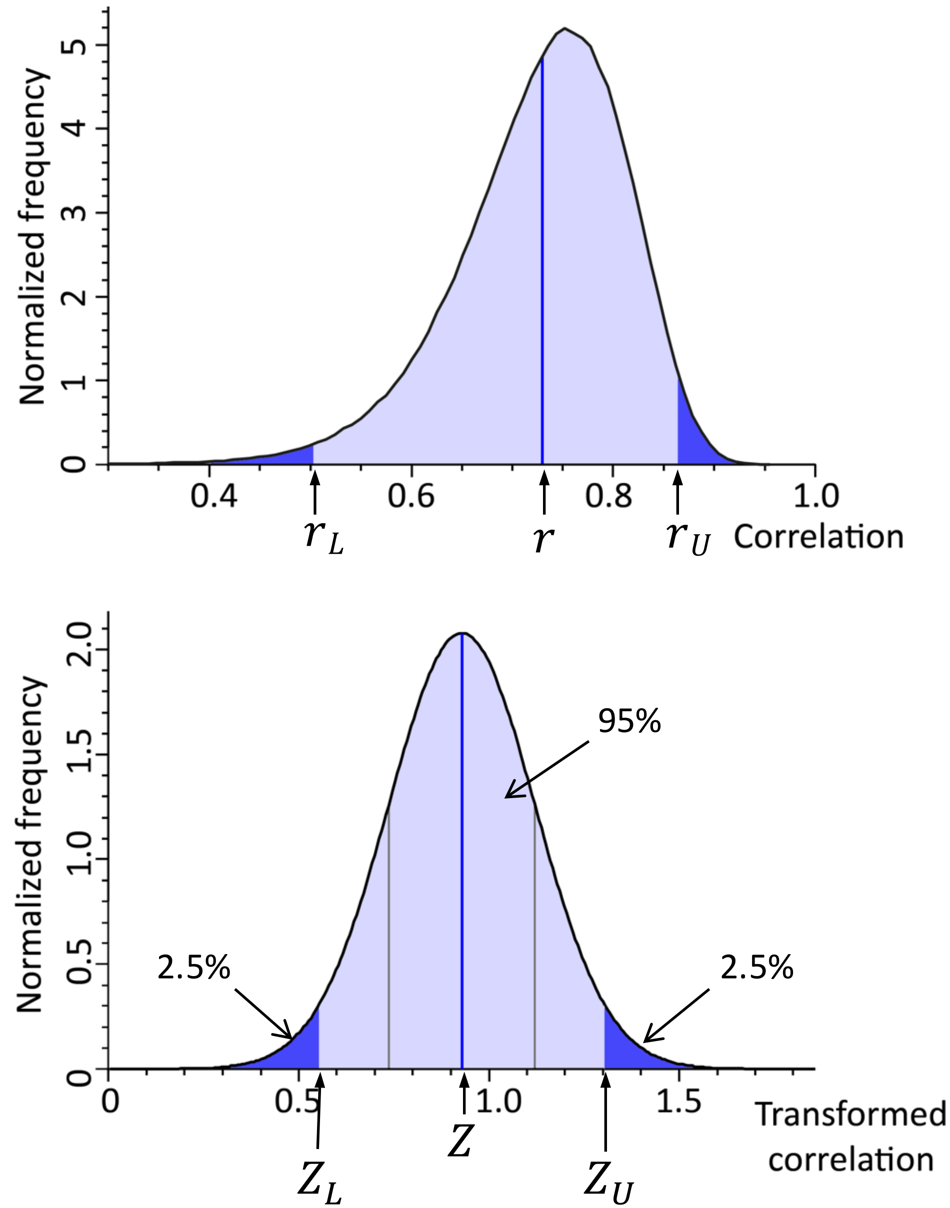
$$Z = \frac{1}{2}\ln\frac{1+r}{1-r} = 0.929$$

$$\sigma = \frac{1}{\sqrt{n-3}} = 0.192$$

- 95% CI corresponds to $Z \pm 1.96\sigma$:
  - $Z_L = Z - 1.96\sigma = 0.553$
  - $Z_U = Z + 1.96\sigma = 1.31$
- Now we find the corresponding limits on r

$$r = \frac{e^{2Z} - 1}{e^{2Z} + 1}$$

  - $r_L = \ 0.503$
  - $r_U = \ 0.864$
- Hence, with 95% confidence, $r = 0.73^{+0.13}_{-0.23}$



Example: 95% CI for correlation with $n = 6$ and $n = 30$

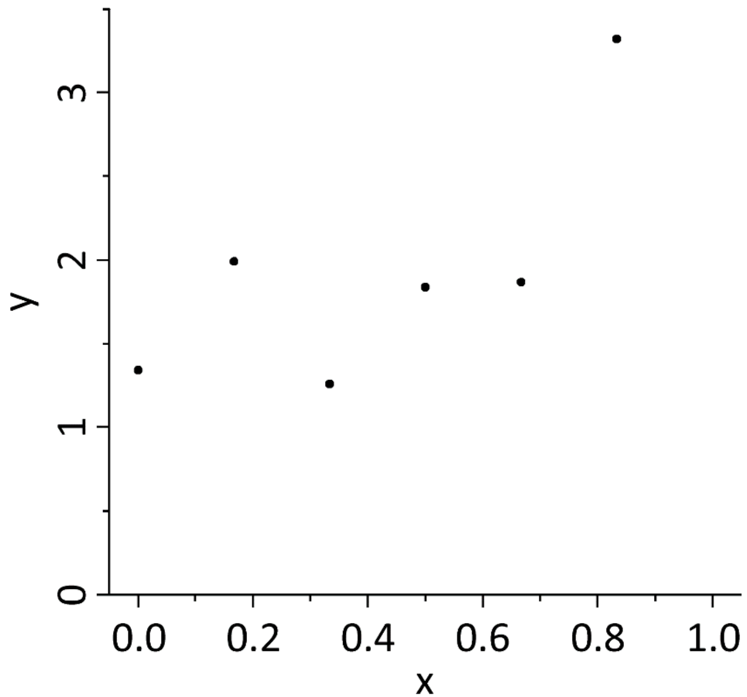
$$r = 0.73$$

	$n = 6$	$n = 30$
$Z = \frac{1}{2}\ln\frac{1+r}{1-r}$	0.929	0.929
$\sigma = \frac{1}{\sqrt{n-3}}$	0.577	0.192
$Z_L = Z - 1.96\sigma$	-0.20	0.553
$Z_U = Z + 1.96\sigma$	2.06	1.31
$r_L = \frac{e^{2Z_L} - 1}{e^{2Z_L} + 1}$	-0.20	0.503
$r_U = \frac{e^{2Z_U} - 1}{e^{2Z_U} + 1}$	0.97	0.864
	$r = 0.7^{+0.3}_{-0.9}$	$r = 0.73^{+0.13}_{-0.23}$

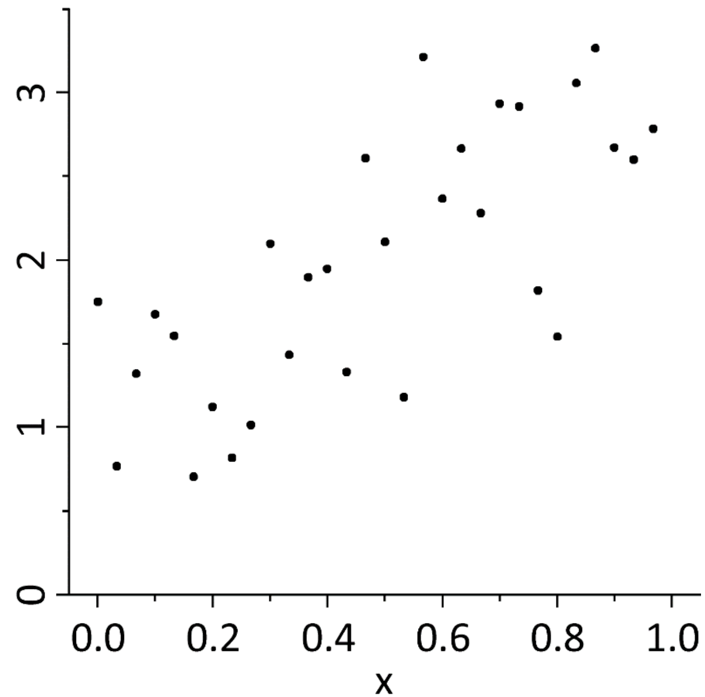
Significance of correlation

- H_0: the sample is drawn from a population with no correlation ($\rho = 0$)
- Calculate $t = r\sqrt{\frac{n-2}{1-r^2}}$
- It follows a Student's t-distribution with $n - 2$ degrees of freedom
- Calculate p-value: probability of getting the observed correlation by chance

$n = 6, r = 0.73\ [-0.20, 0.97], p = 0.05$



$n = 30, r = 0.73\ [0.50, 0.86], p = 2\times10^{-6}$



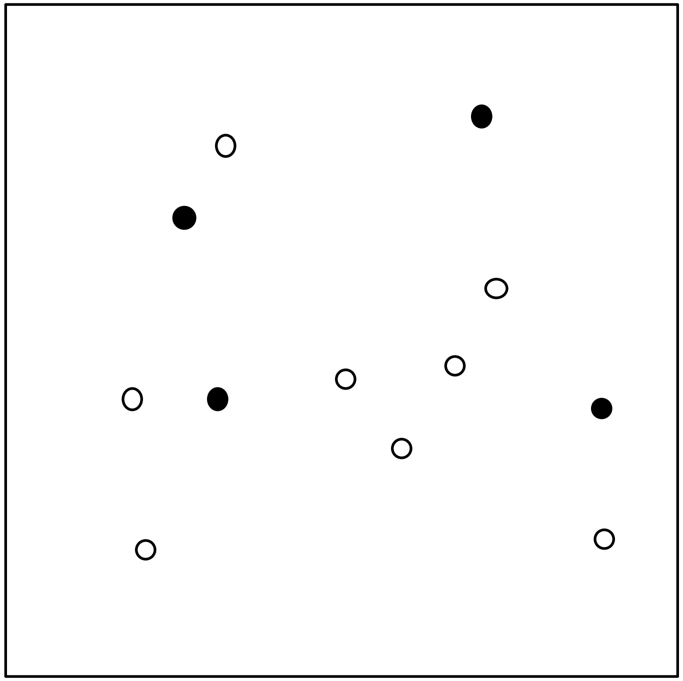
Confidence interval of a proportion

- Proportion:

$$\hat{p} = \frac{\hat{S}}{n} = \frac{\text{number of successes}}{\text{sample size}}$$

- Examples:
  - poll results
  - survival experiments
  - counting cells with a property

- Sample proportion, $\hat{p}$, is an estimator of the (unknown) population proportion, p



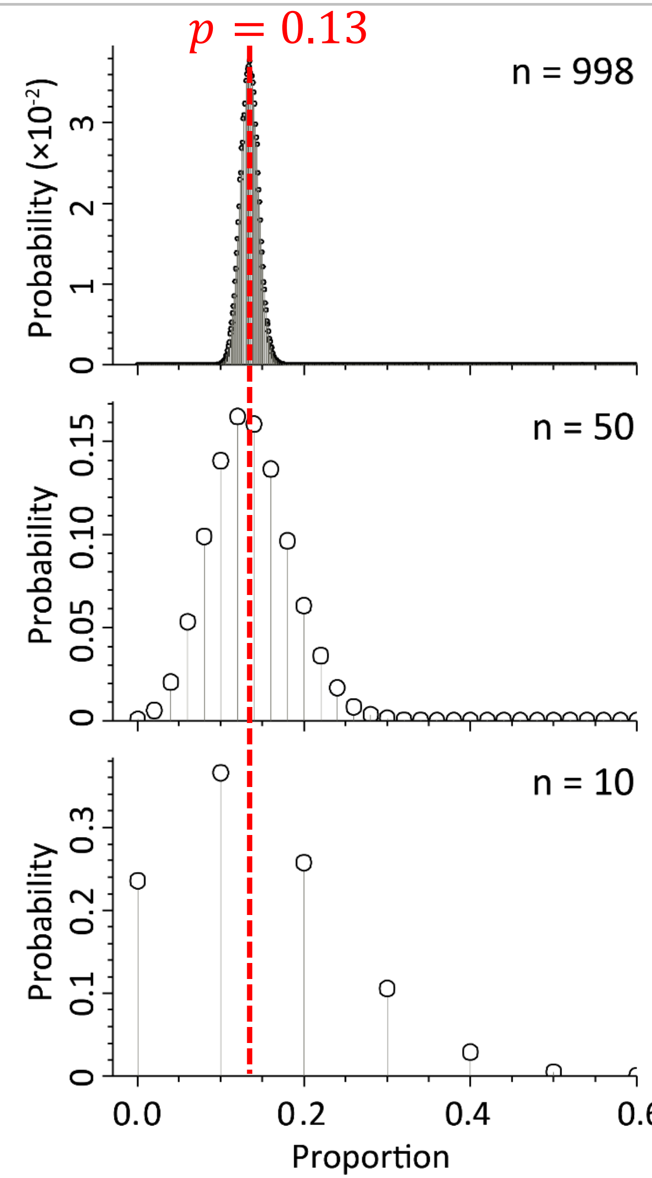
● $\hat{S} = 4$

○+● $n = 12$

$$\hat{p} = \frac{4}{12} = 0.33$$

Sampling distribution of a proportion

- *Gedankenexperiment*
- Consider a population of mice where $p = 13\%$ are immune to a certain disease
- Draw a random sample of size n and find the proportion of immune mice, $\hat{p}$, in the sample
- Repeat 100,000 times and plot the distribution of $\hat{p}$
- What kind of distribution is it?
- Hint: every time you select a mouse, it can be either immune or not, with probability p or $1 - p$
- Binomial distribution
  - immune = "success", probability p
  - not immune = "failure", probability $1 - p$
- Good! Sampling distribution is known



Sampling distribution of a proportion: scaled binomial

Absolute numbers

- S – binomial random variable
- Mean and standard deviation

$$\mu = np$$

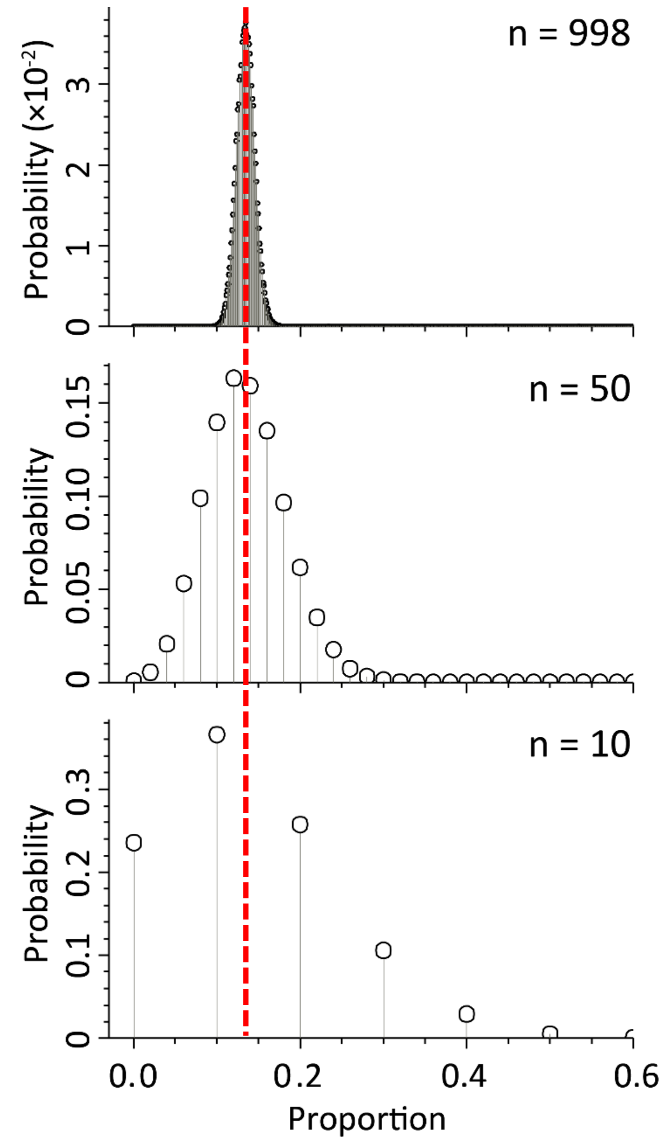
$$\sigma = \sqrt{np(1-p)}$$

Proportion

- $R = S/n$ – scaled binomial random variable
- Mean and standard deviation scaled by n:

$$\mu_R = p$$

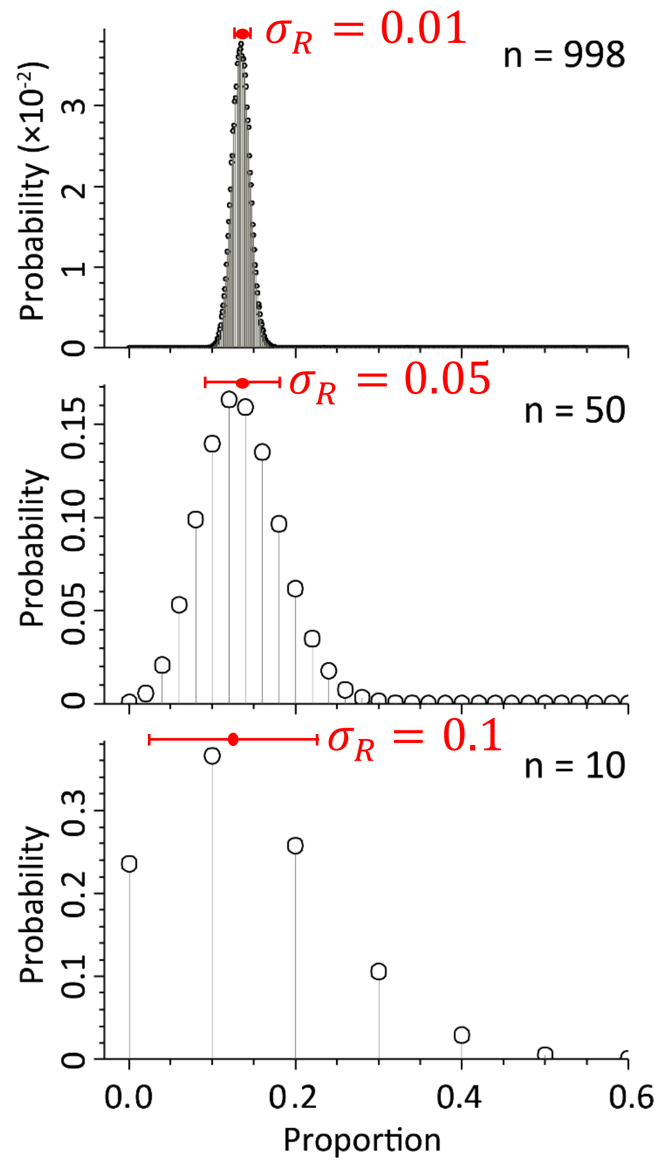
$$\sigma_R = \sqrt{\frac{p(1-p)}{n}}$$



Sampling distribution of a proportion

- Width of the sampling distribution of a proportion

$$\sigma_R = \sqrt{\frac{p(1-p)}{n}}$$



Reminder from lecture 2

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

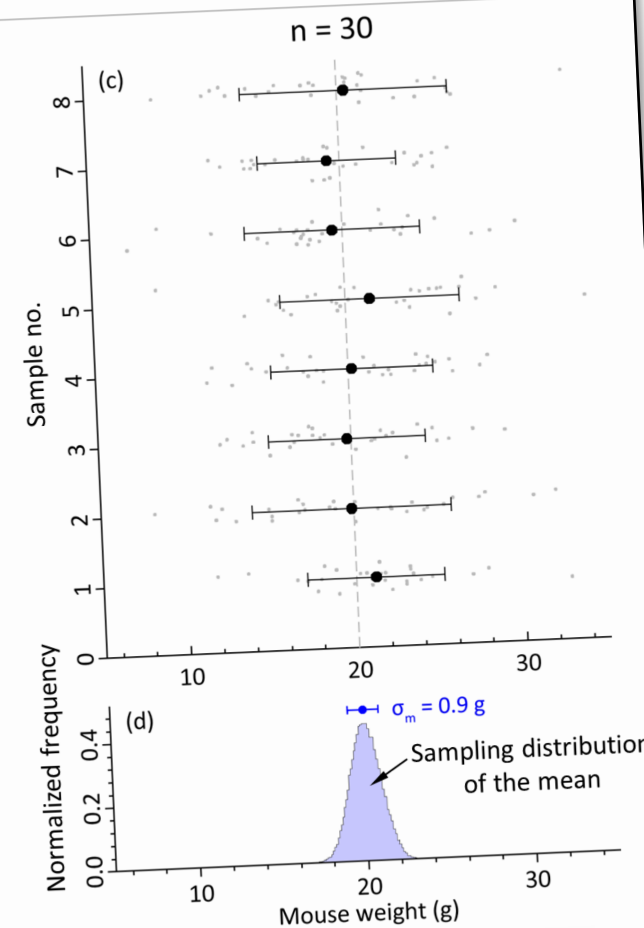
- Sampling distribution is Gaussian, with standard deviation

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

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

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

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



32

Sampling distribution of a proportion

- Width of the sampling distribution of a proportion

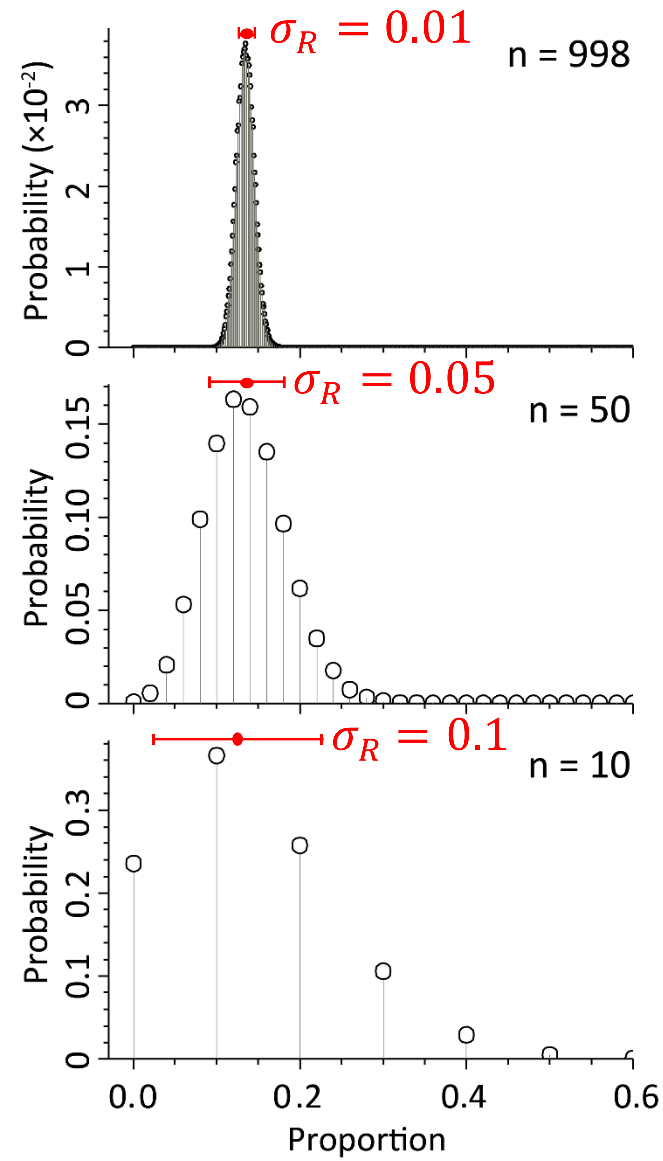
$$\sigma_R = \sqrt{\frac{p(1-p)}{n}}$$

- Replace an unknown population parameter, p, with the observed estimator, $\hat{p}$

$$SE_R = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- Standard error of a proportion
- SE_R **estimates** the width of the sampling distribution

- However, this doesn't work for small n, or when proportion is close to 0 or 1



Wald method

- Sample: size n with $\hat{S}$ successes
- Select Gaussian Z for given confidence (e.g. $Z = 1.96$ for 95%)
- Calculate *corrected* quantities

$$S' = \hat{S} + \frac{Z^2}{2} \qquad n' = n + Z^2$$

- and then:

$$p' = \frac{S'}{n'} \qquad SE'_R = \sqrt{\frac{p'(1-p')}{n'}}$$

- Margin of error:

$$W = Z \times SE'_R$$

- Confidence interval is $p' \pm W$:

$$[p' - Z \times SE'_R, p' + Z \times SE'_R,]$$

Example

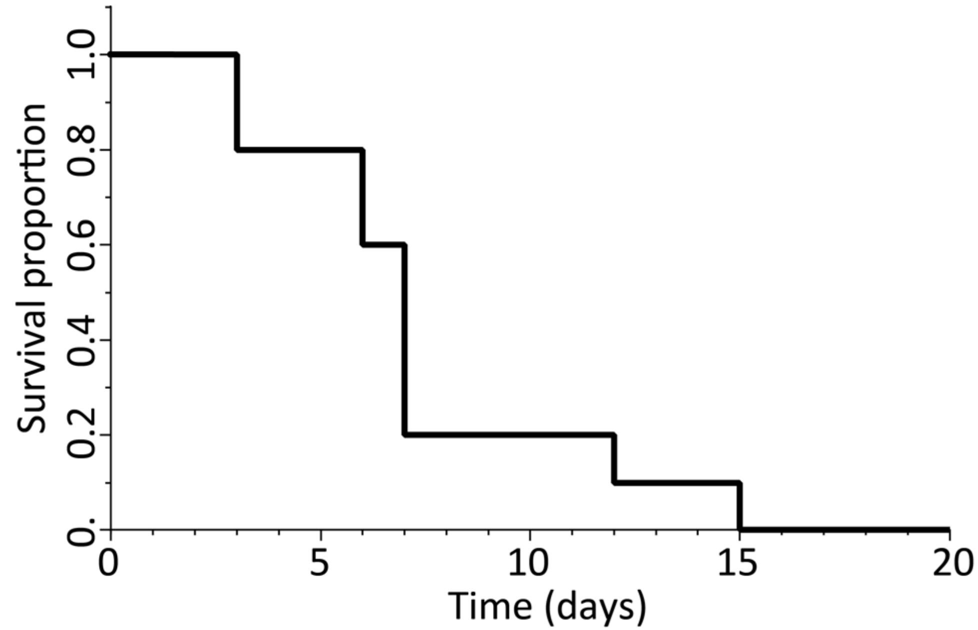
$n = 10$
$\hat{S} = 1$
$\hat{p} = 0.1$

- Uncorrected standard error
  $SE = 0.1$
- Corrected values
  $S' = 1 + 1.92 = 2.92$
  $n' = 10 + 3.84 = 13.84$
- Corrected proportion and error
  $p' = 0.21$
  $SE'_R = 0.11$
- Margin of error
  $W = Z \times SE'_R = 0.21$
- 95% confidence interval is [0, 0.43]

Confidence intervals of a proportion

- Consider survival experiment
  - take 10 mice
  - infect with something nasty
  - apply treatment
  - count survival proportion over time

- We need errors of proportion!



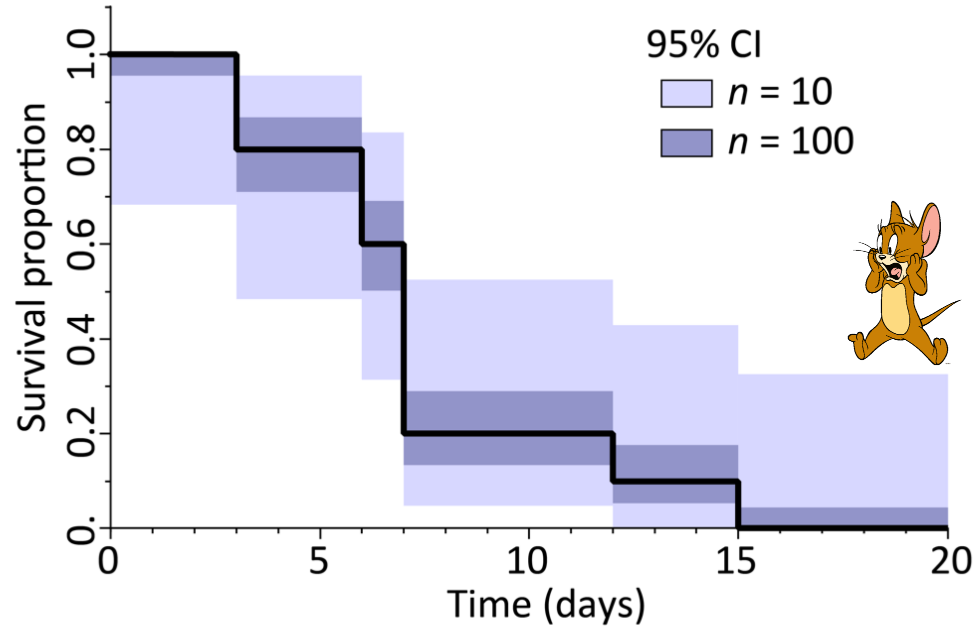
Confidence intervals of a proportion

- Consider survival experiment
  - take 10 mice
  - infect with something nasty
  - apply treatment
  - count survival proportion over time

- 95% CIs using Wald method
- The bigger sample, the smaller error

- Even when $\hat{p} = 0$, error allows for non-zero proportion

- We have zombie mice!



What’s in the box?

Exercise: error of proportion

- What is the proportion of black balls in the box?

Sample size	12	Z	1.96
Black	2	p'	0.25
Proportion	17%	W	0.2132
Error	21%		

95% confidence interval

4%	46%

$$p' = \frac{\hat{S} + Z}{n + Z^2}$$

Modified proportion, where Z is a z-score corresponding to needed confidence (e.g. $Z = 1.96$ for 95%)

$$W = Z\sqrt{\frac{p'(1-p')}{n + Z^2}}$$

Margin of error

$$[p' - W, p' + W]$$

Confidence interval for proportion

Beware of small samples!

- When you count things, small samples are not very good
- Consider a sample size of $n = 10$

	Counts	Correlation $r = 0.73$	Proportion $\hat{S} = 3$
95% CI	$[4.8, 18.4]$	$[0.19, 0.93]$	$[0.10, 0.61]$
Fractional half error	68%	51%	71%

- When you do counts, you need $n > 30$

Bootstrapping

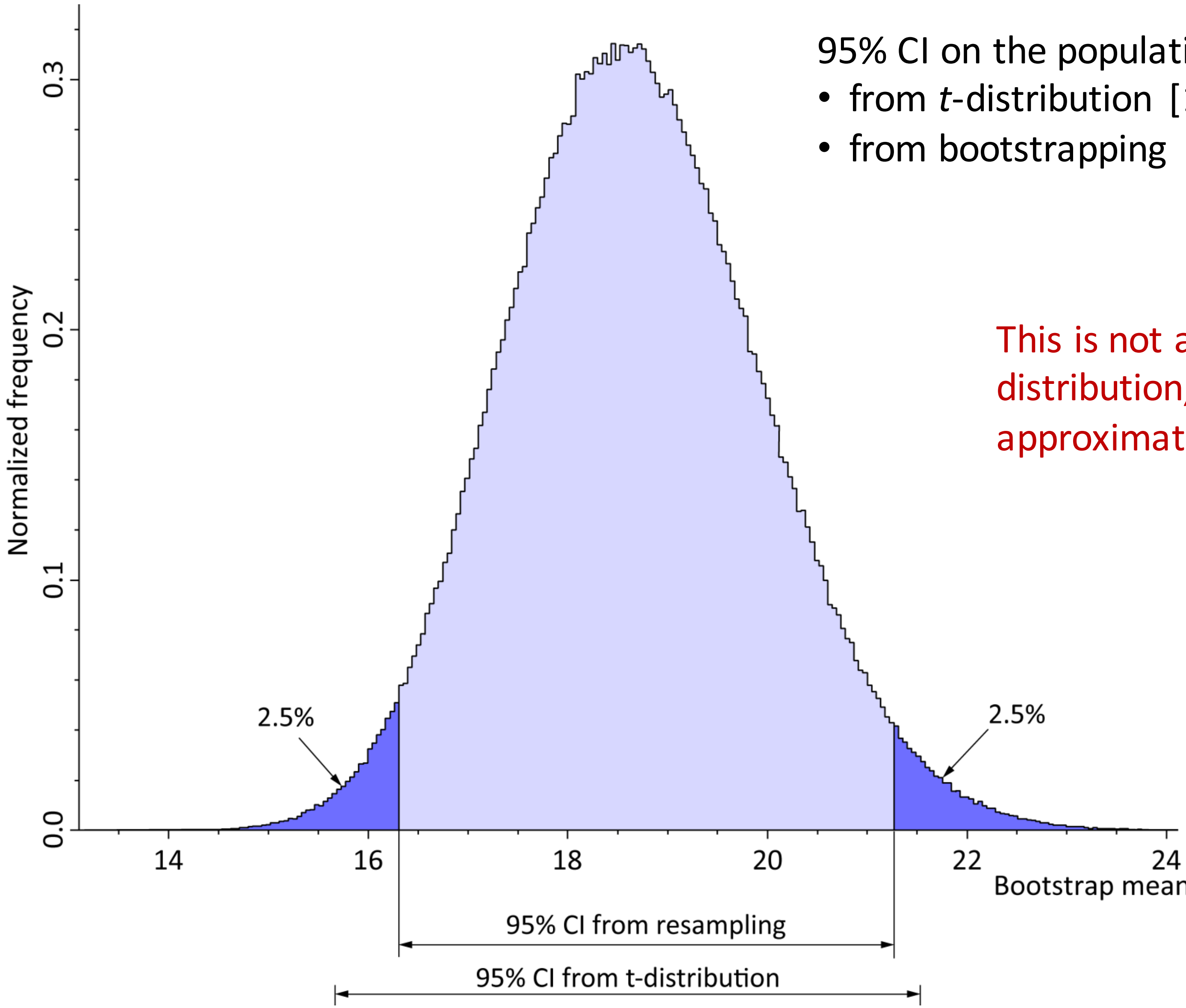
- Versatile technique used when
  - distribution of the estimator is complicated or unknown
  - for power calculations
- Approximate sampling distribution from one sample only
- Use random resampling *with replacement*

19.4	**18.2**	**11.5**	**17.2**	**25.7**	**19.2**	**21.5**	**16.7**	**15.6**	**27.7**	**14.3**	**16.3**	$\boldsymbol{M = 18.6}$	original sample
27.7	18.2	18.2	25.7	11.5	17.2	17.2	25.7	21.5	11.5	14.3	17.2	$M = 18.8$	
19.2	14.3	19.2	15.6	14.3	14.3	17.2	16.3	19.2	19.2	16.3	21.5	$M = 17.2$	
14.3	17.2	18.2	18.2	18.2	11.5	14.3	18.2	17.2	19.4	11.5	16.3	$M = 16.2$	resamples
25.7	18.2	15.6	15.6	19.4	19.2	18.2	19.4	21.5	16.7	14.3	18.2	$M = 18.5$	
19.2	21.5	16.7	17.2	21.5	18.2	21.5	17.2	21.5	15.6	21.5	21.5	$M = 19.4$	

…

- Repeat this many times (e.g. 10^6) and collect all means
- Build the bootstrap distribution of the mean

Bootstrapping



95% CI on the population mean

- from *t*-distribution [15.7, 21.5]
- from bootstrapping [16.3, 21.3]

This is not a sampling distribution, it only approximates it

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

YOU NEED REPLICATES

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

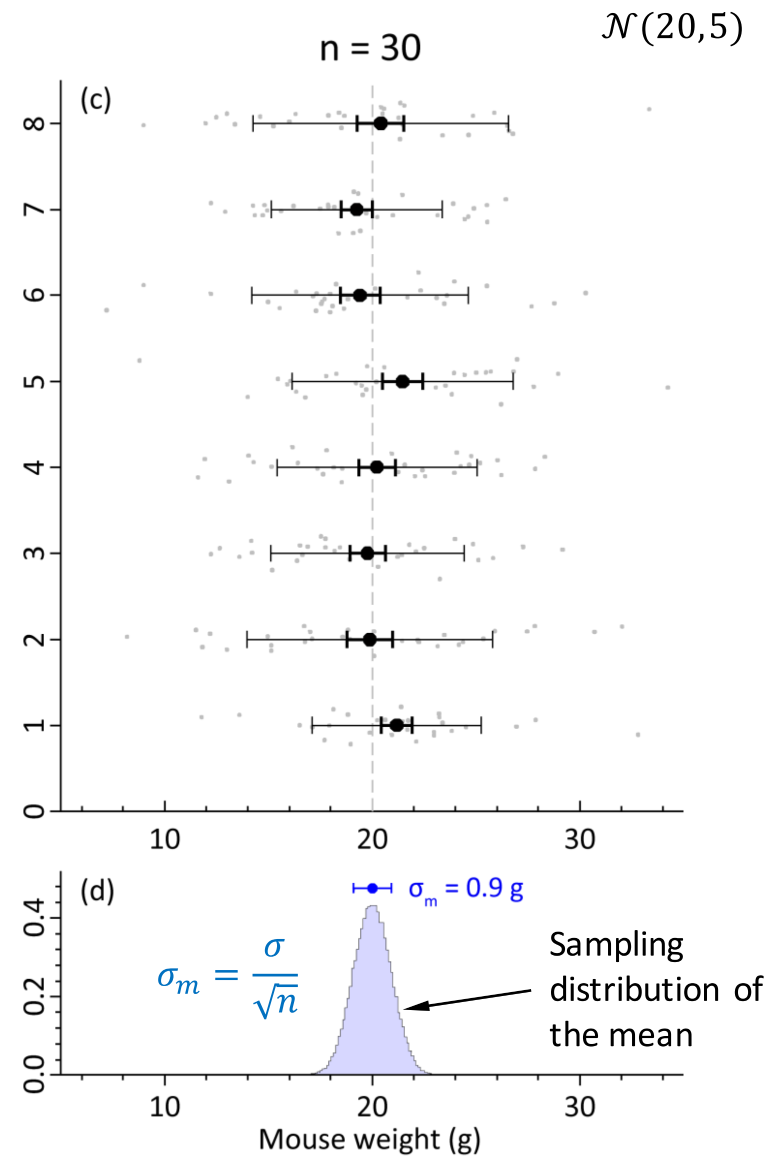
$$\epsilon \approx 2\sigma_m = \frac{2\sigma}{\sqrt{n}}$$

- 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

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



Homework

In a study of a new antibiotic a sample of bacterial cells was treated with the drug, while the control sample remained untreated. Both treatment and control were done in three replicates. An aliquot of each replicate was placed in a counting chamber an living cells counted.

Replicate	1	2	3
Control	111	104	123
Treatment	16	18	14

1. Pool the counts together and find the proportion of cells surviving the treatment and its 95% confidence interval.
2. Do the same using replicated data. Compare the results.

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