Error analysis in biology

Marek Gierliński
Division of Computational Biology

Hand-outs available at http://tiny.cc/statlec

http://www.compbio.dundee.ac.uk/user/mgierlinski/statalk.html

Previously on Errors…

Confidence interval of a correlation

- Fisher's transformation (Gaussian)

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

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

Confidence interval of a proportion

- From binomial distribution we have standard error of a proportion

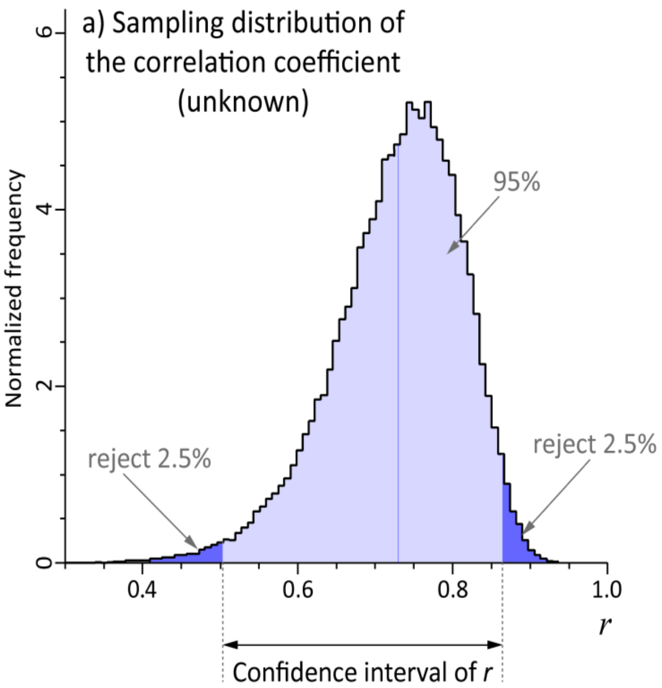
$$SE_{\Phi} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

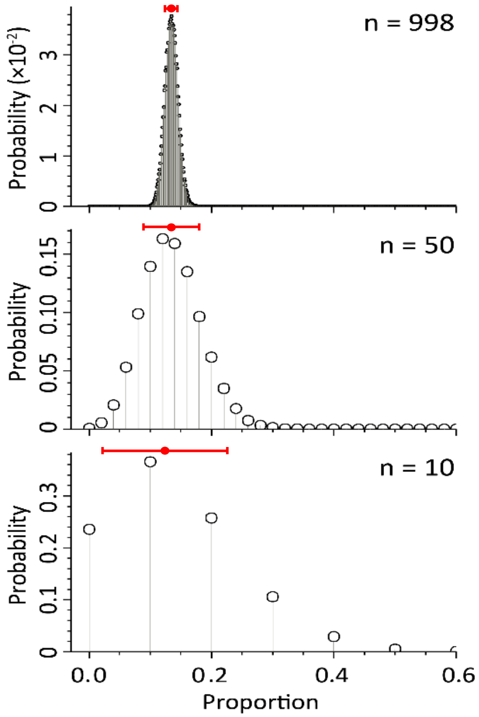
Confidence interval of count data

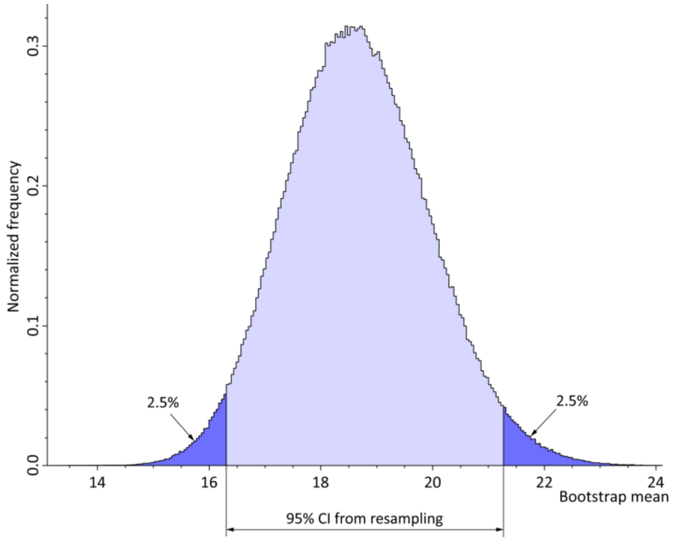
- accurate approximation

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

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







Bootstrapping

- when everything else fails

5. Error bars

“Errors using inadequate data are much less than those using no data at all”

Charles Babbage

A good plot

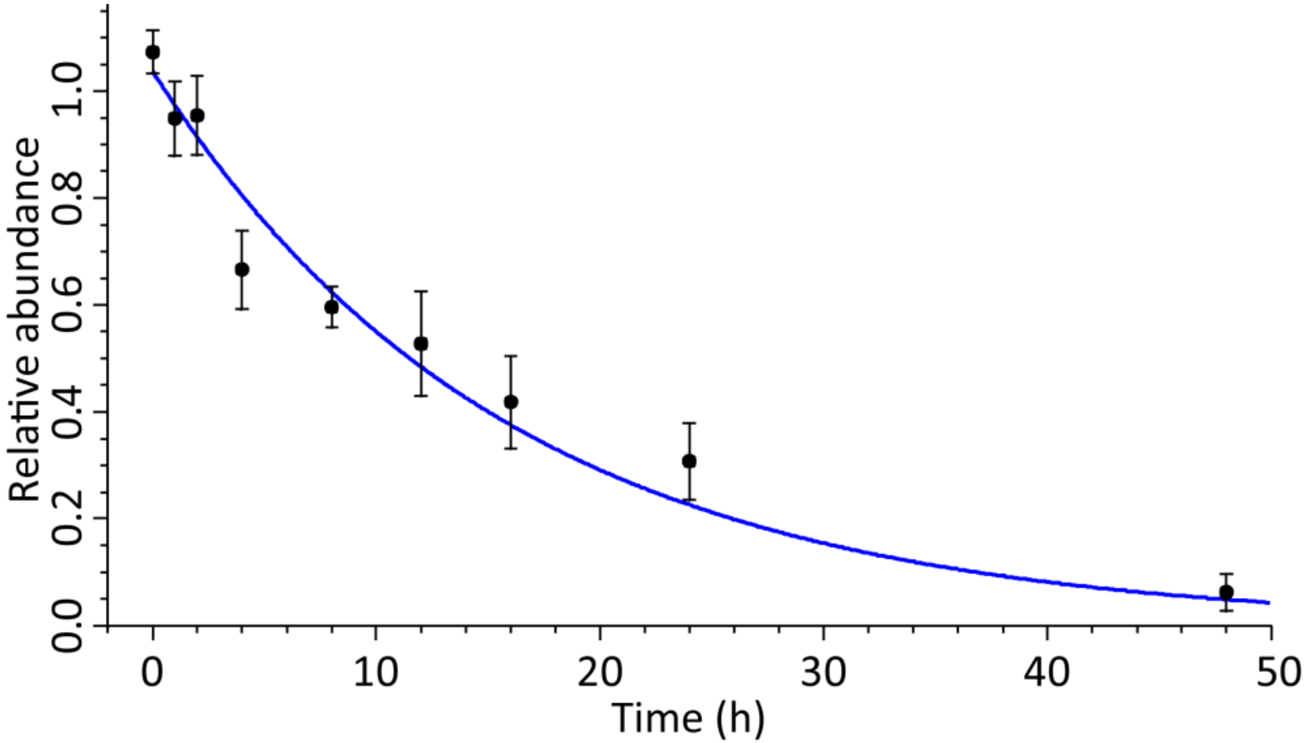


Figure 6-1. Exponential decay of a protein in a simulated experiment. Error bars represent propagated standard errors from individual peptides. The curve shows the best-fitting exponential decay model, $y(t) = Ae^{-t/\tau}$, with $A = 1.04 \pm 0.05$ and $\tau = 16 \pm 3$ h (95% confidence intervals).

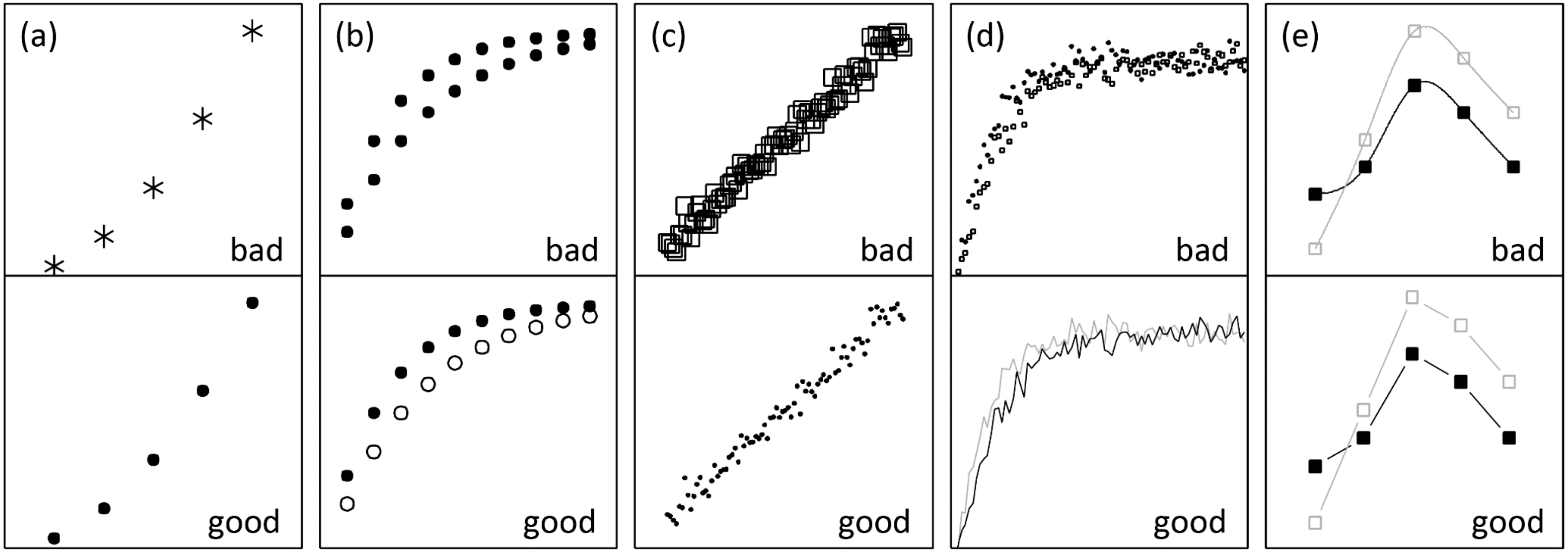
3 rules for making good plots

1. Clarity of presentation

2. Clarity of presentation

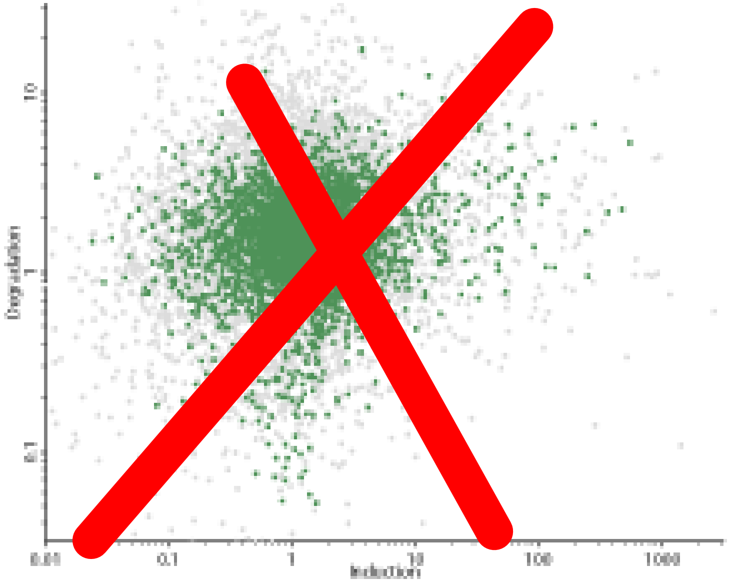
3. Clarity of presentation

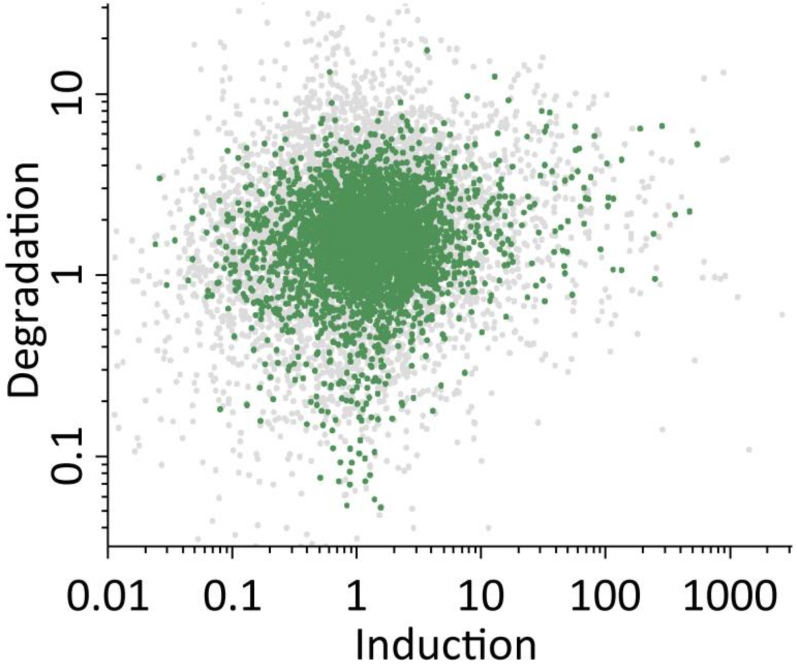
Lines and symbols



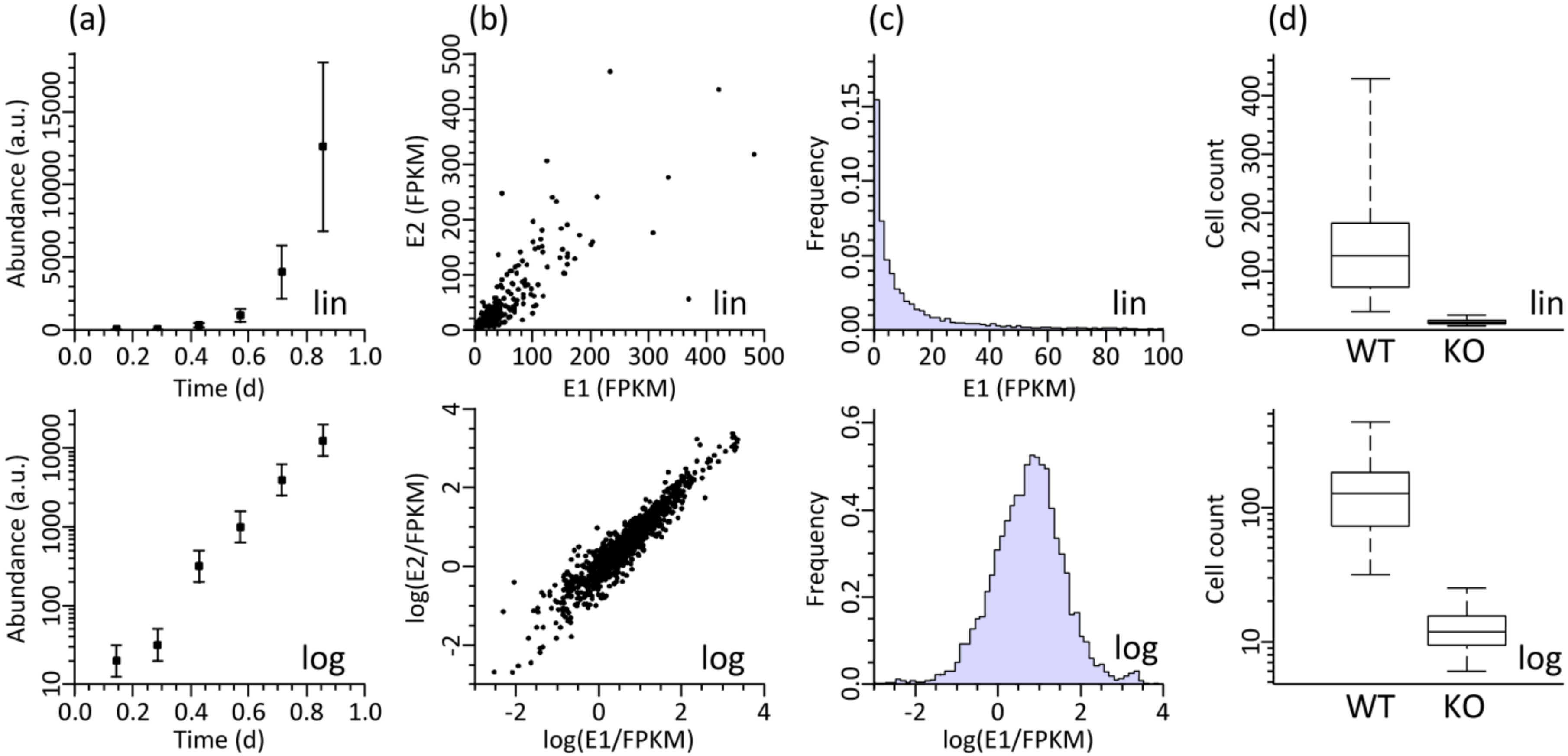
- Clarity!
- Symbols shall be easy to distinguish
- It is OK to join data points with lines for guidance

Labels!


10
1
0.1
Degradation
0.01
0.1
1
10
100
1000
Induction

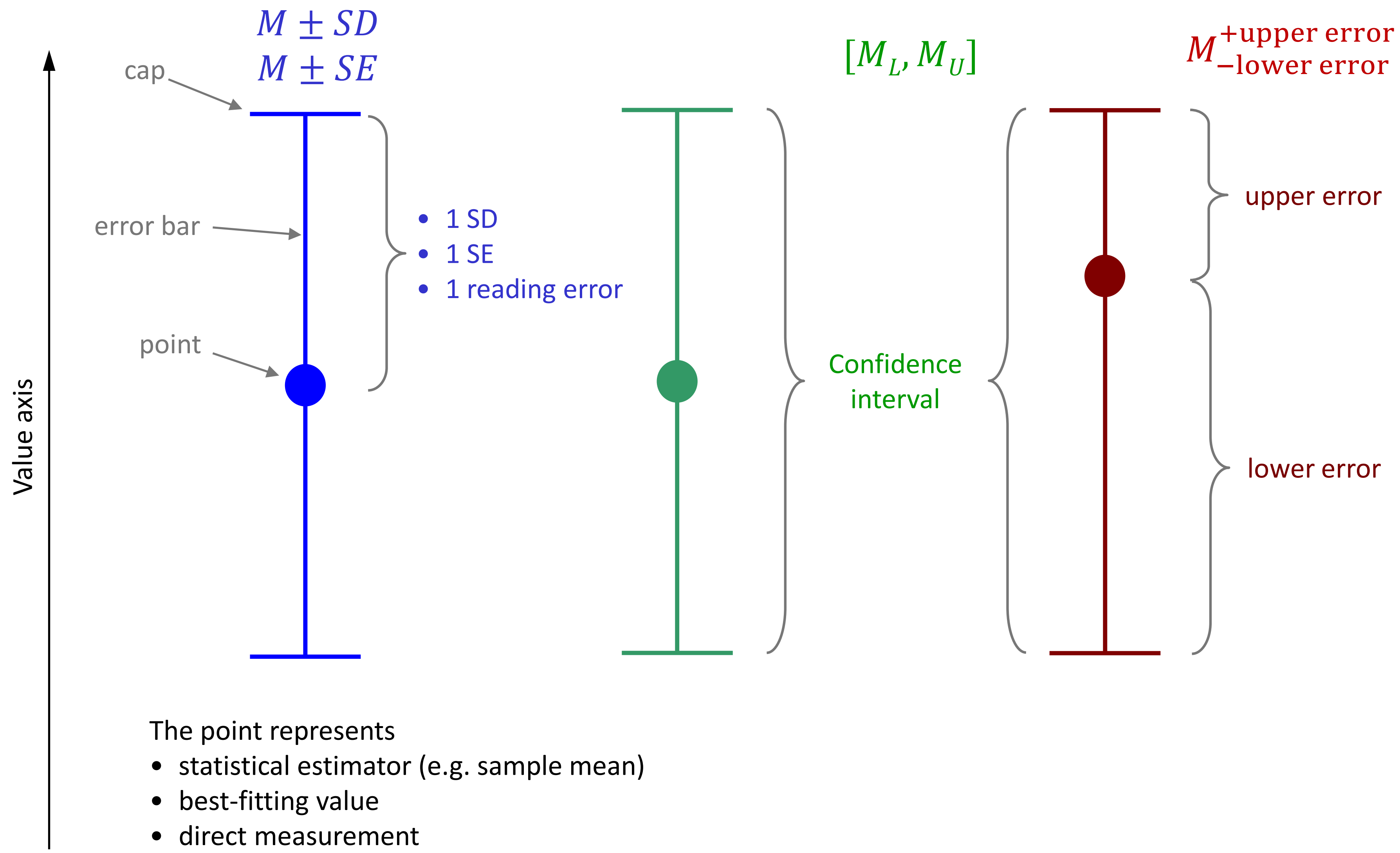

10
1
0.1
Degradation
0.01
0.1
1
10
100
1000
Induction

Logarithmic plots



- Clarity!
- Use logarithmic axes to show data spanning many orders of magnitude

How to plot error bars



Value axis

$M \pm SD$

$M \pm SE$

cap

error bar

point

- 1 SD
- 1 SE
- 1 reading error

$[M_L, M_U]$

Confidence interval

$M^{+\text{upper error}}_{-\text{lower error}}$

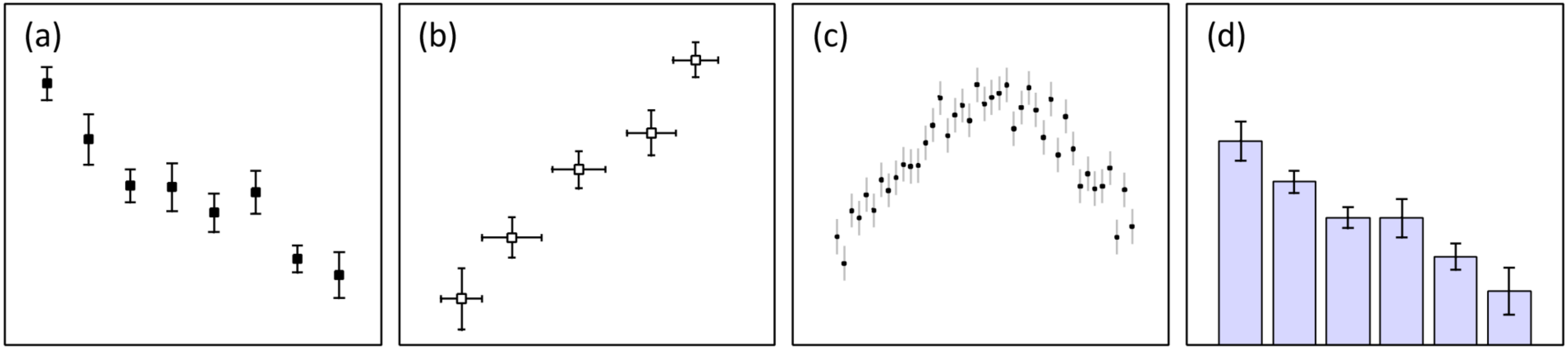
upper error

lower error

The point represents

- statistical estimator (e.g. sample mean)
- best-fitting value
- direct measurement

How to plot error bars



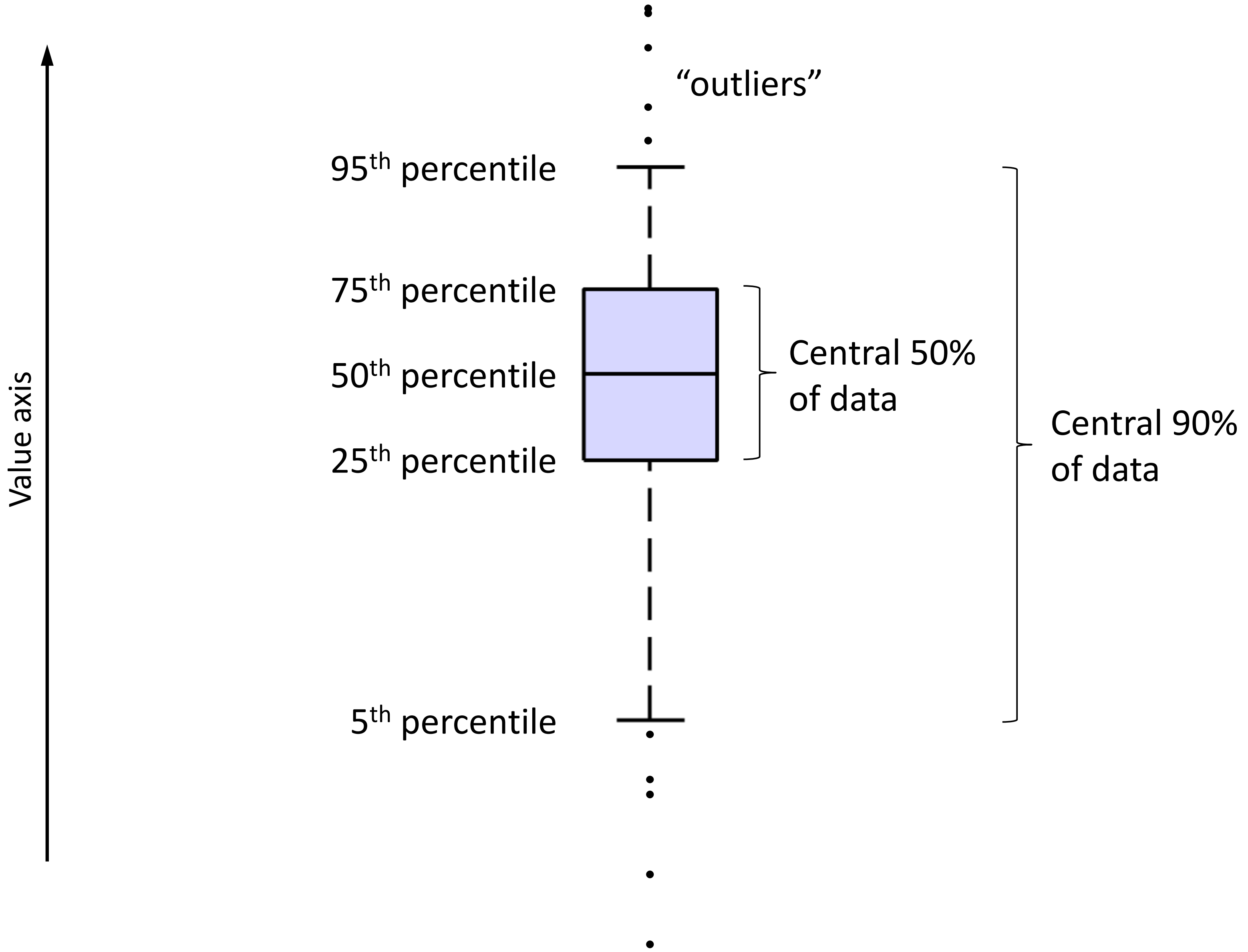
- Clarity!
- Make sure error bars are visible

Types of errors

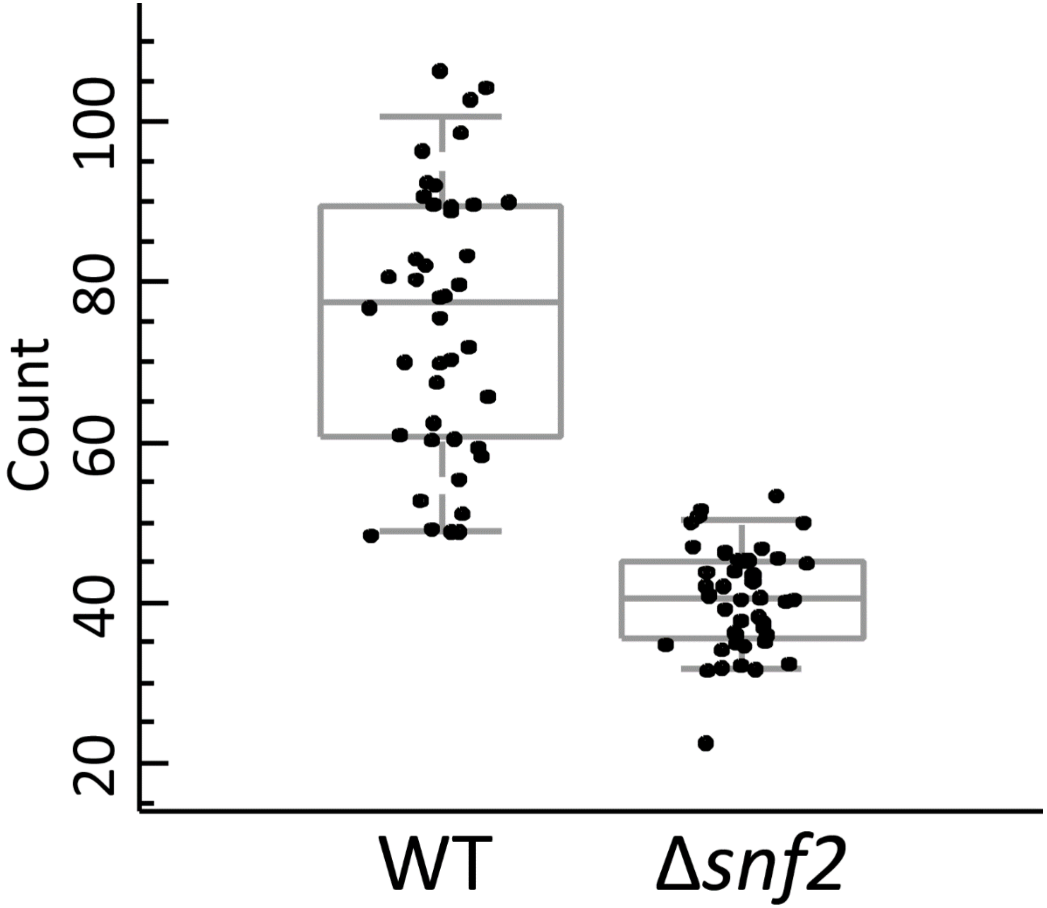
Error bar	What it represents	When to use
Standard deviation	Scatter in the sample	Comparing two or more samples, though box plots (with data points) make a good alternative
Standard error	Error of the mean	Most commonly used error bar, though confidence intervals have better statistical intuition
Confidence interval	Confidence in the result	The best representation of uncertainty; can be used in almost any case

- Always state what type of uncertainty is represented by your error bars

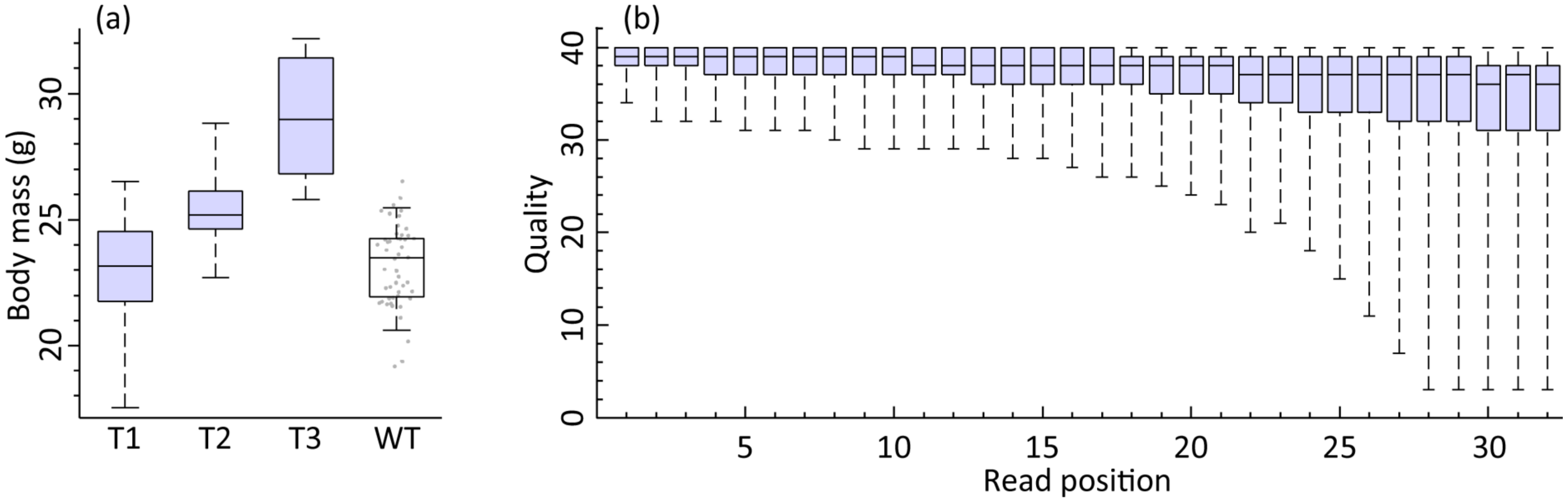
Box plots


“outliers”
95th percentile
75th percentile
50th percentile
25th percentile
5th percentile
Central 50% of data
Central 90% of data
Value axis

Box plots



Box plots



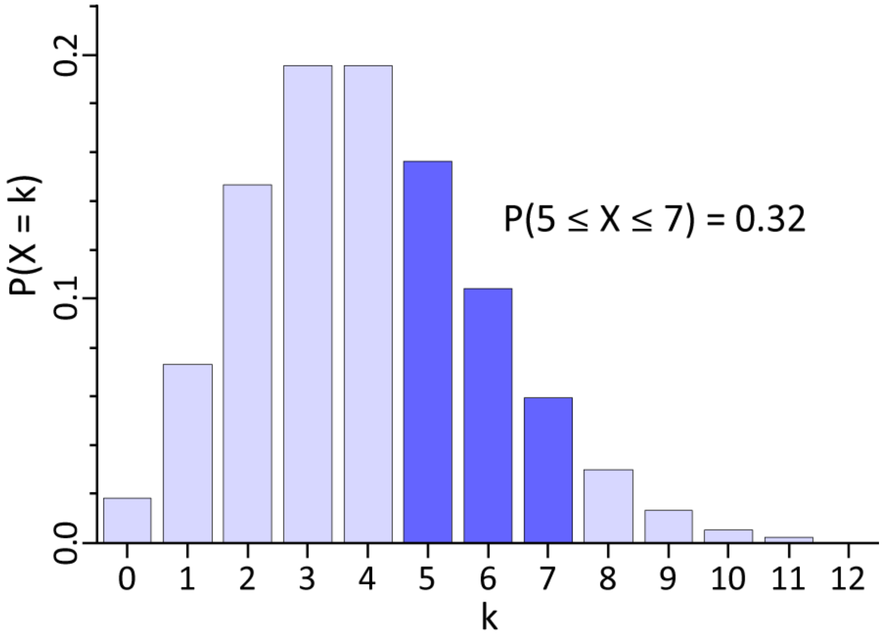
- Box plots are a good alternative to standard deviation error bars
- They are non-parametric and show pure data

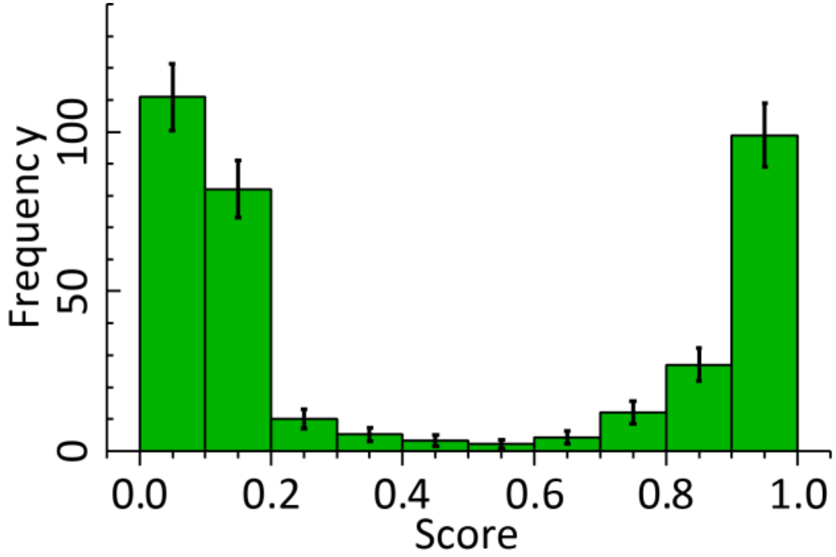
Bar plots

- Area of a bar is proportional to the value presented
- Summed area of several bars represents the total value

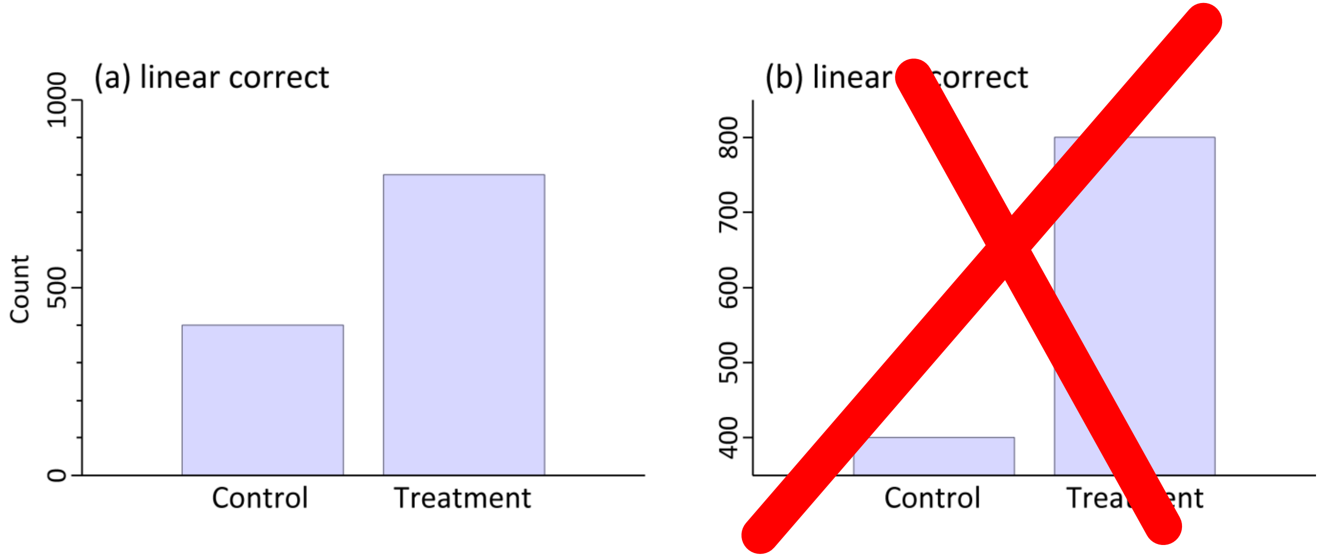
- Bar plots should **only** be used to present additive quantities: counts, fractions, proportions, probabilities, etc.

- Against a continuous variable data are integrated over the bar width
- Each bar is two-dimensional
- Bar width matters!

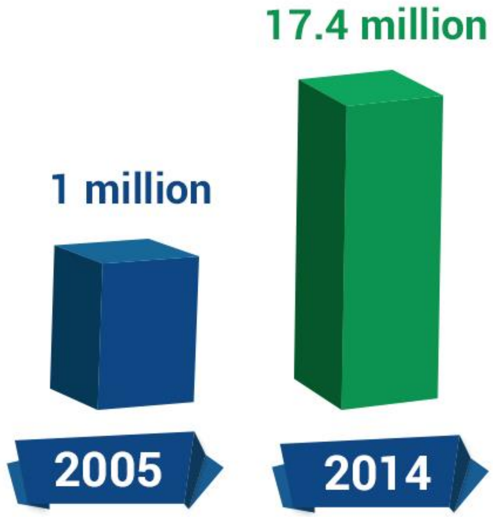




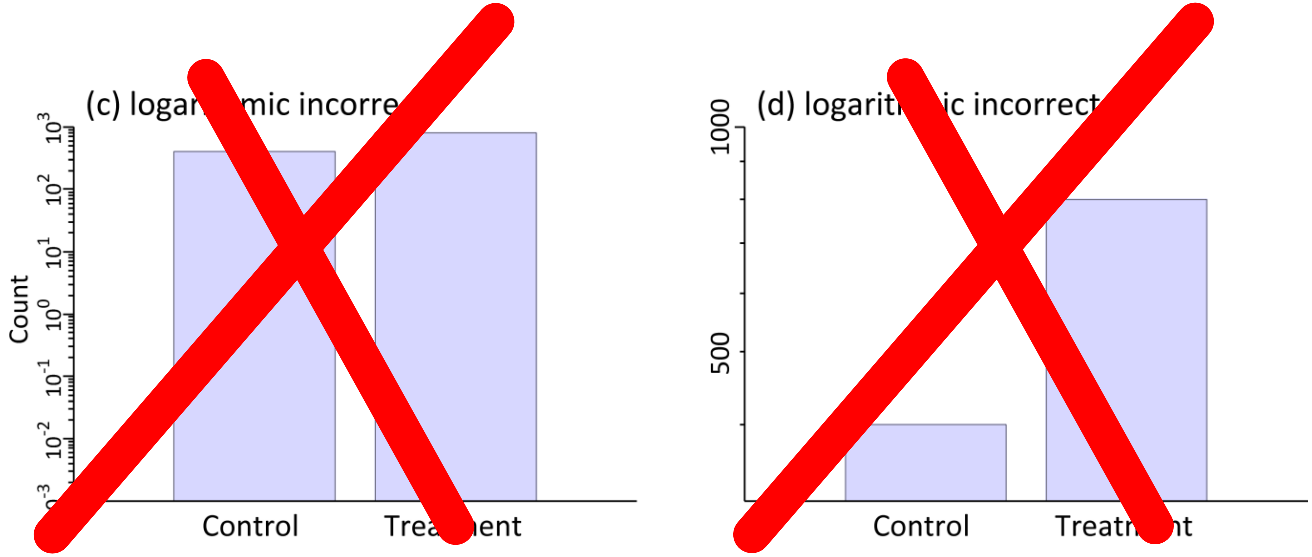
Bar plots start at zero



- Bar area represents its value
- Hence, baseline must be at zero
- If not, the plot is very misleading
- Don’t do it!



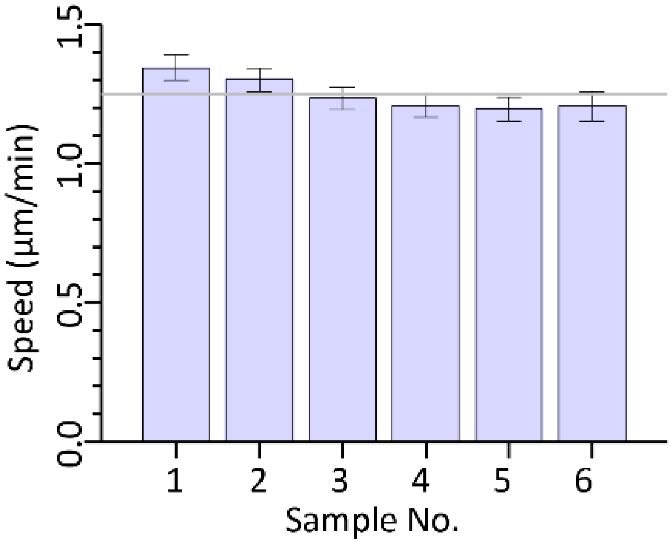
Bar plots in logarithmic scale



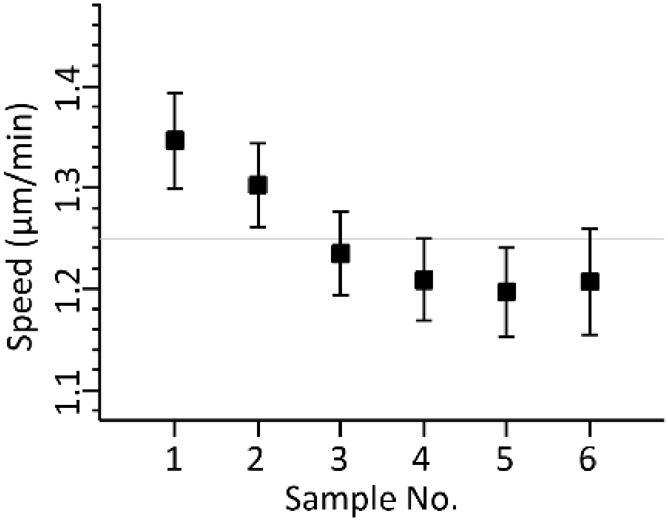
- There is no zero in a logarithmic scale!
- Bar size depends on an arbitrary lower limit of the vertical axis
- Don’t do it!

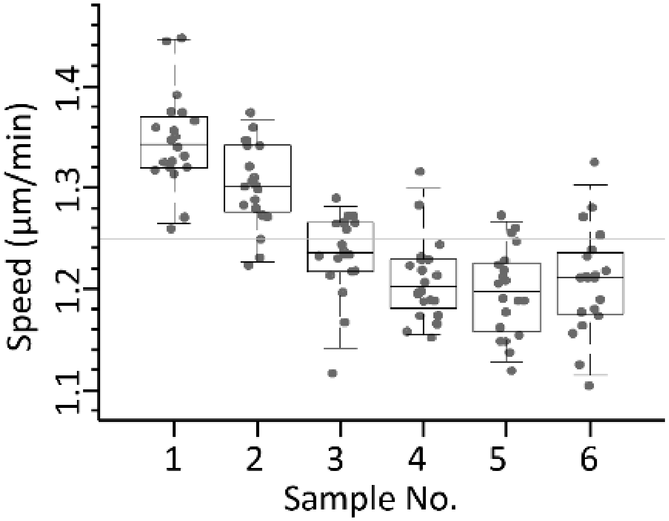
Bar plot problems

Non-additive quantity

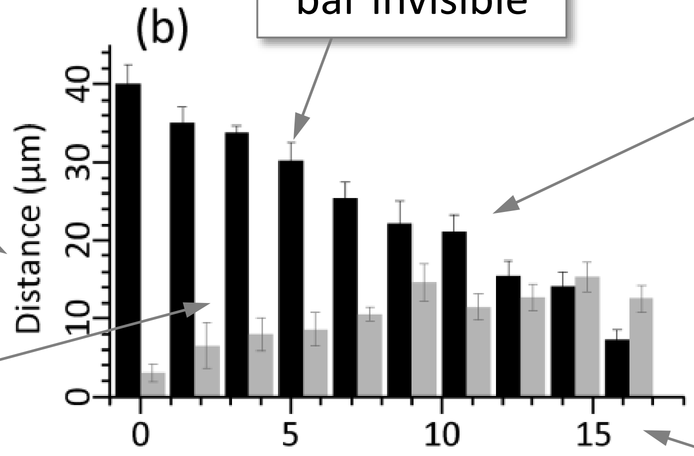


Little variability





Bar plot problems



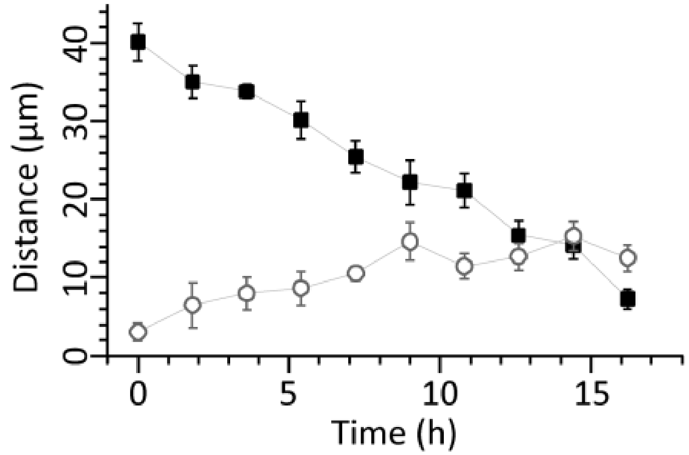
Non-additive quantity

Lower error bar invisible

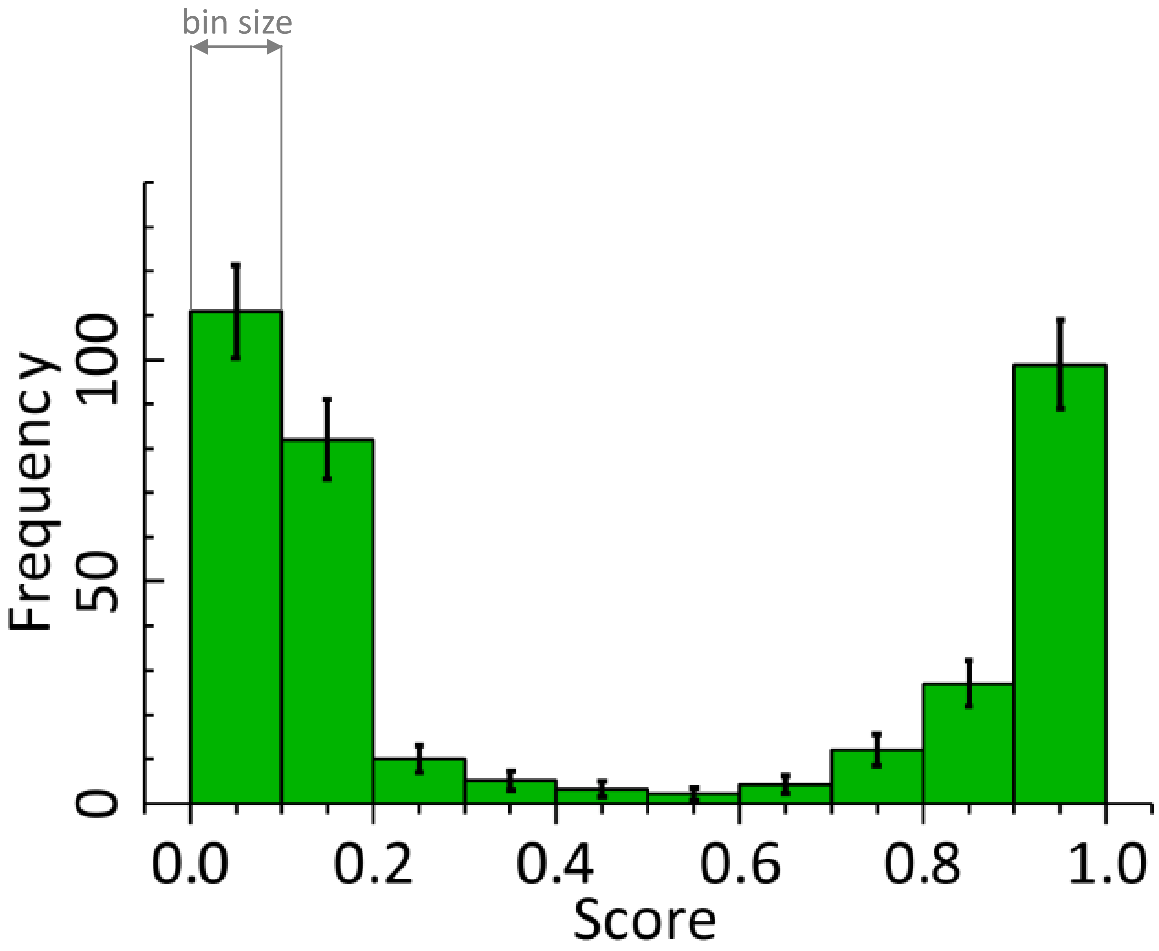
Crowded bars

Confusing visual pattern

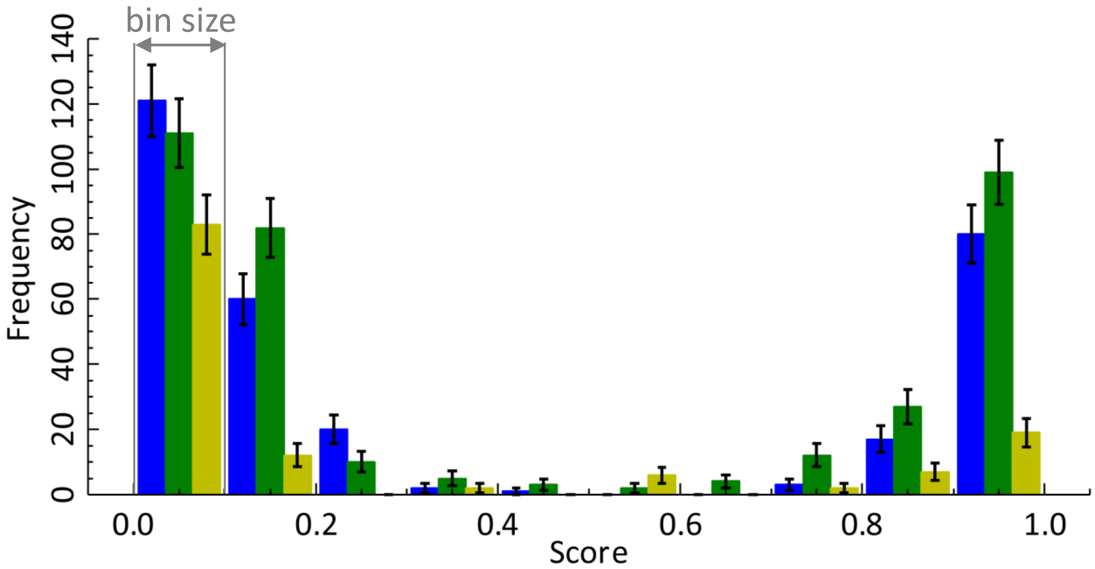
Continuous variable

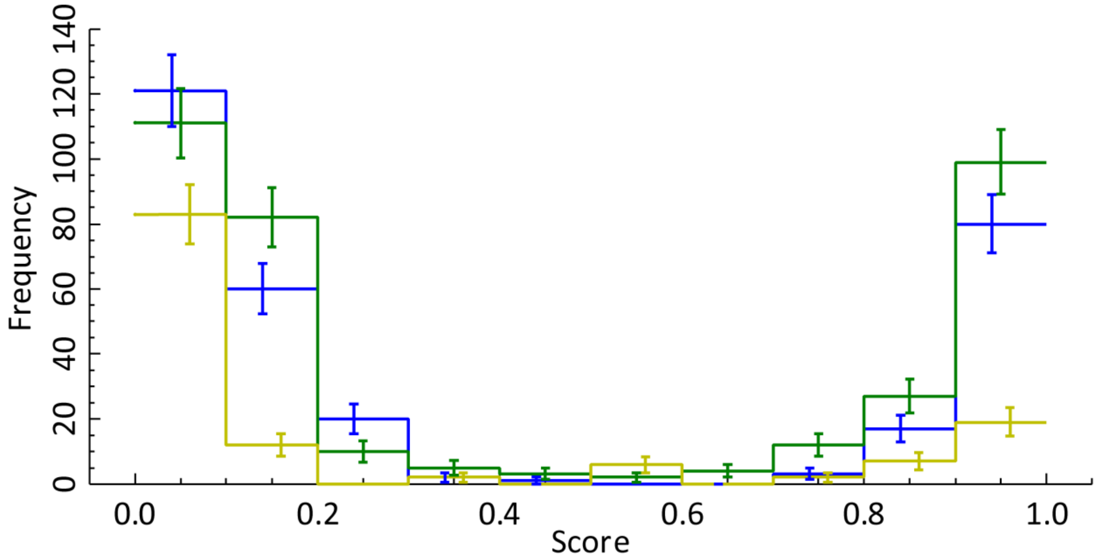


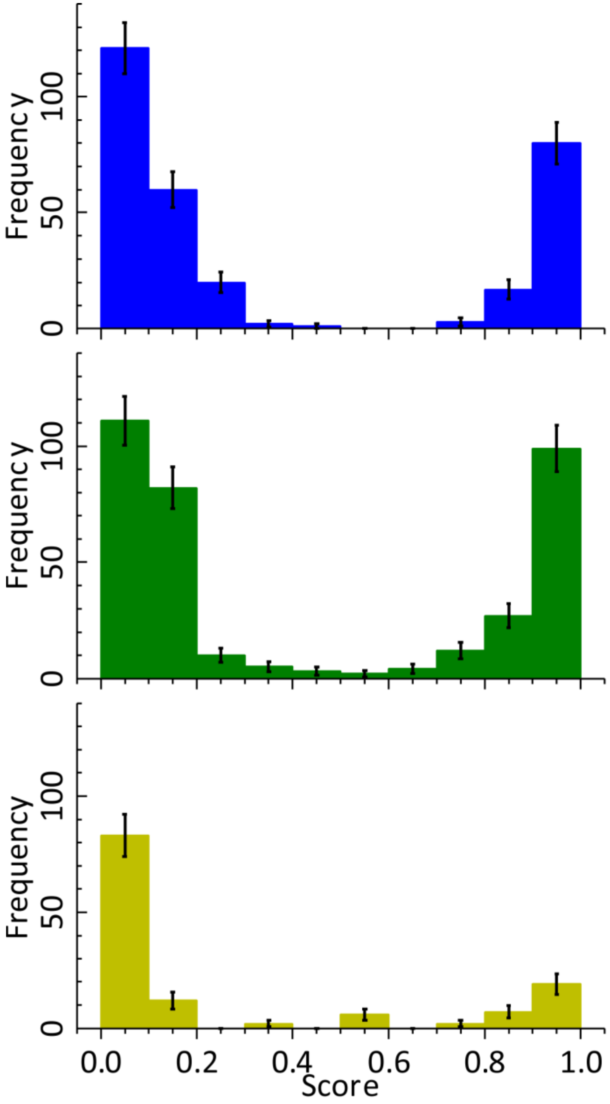
Multiple bar plots and a continuous variable



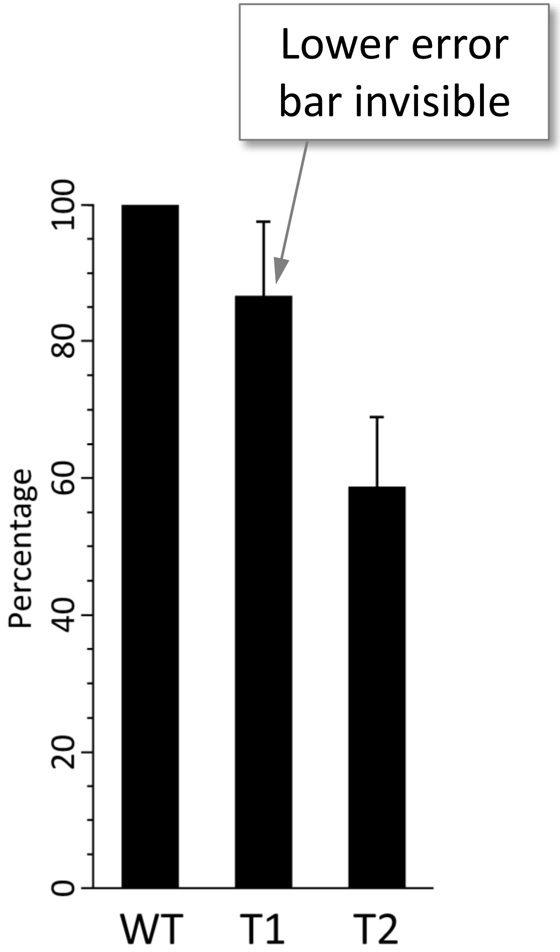
Multiple bar plots and a continuous variable

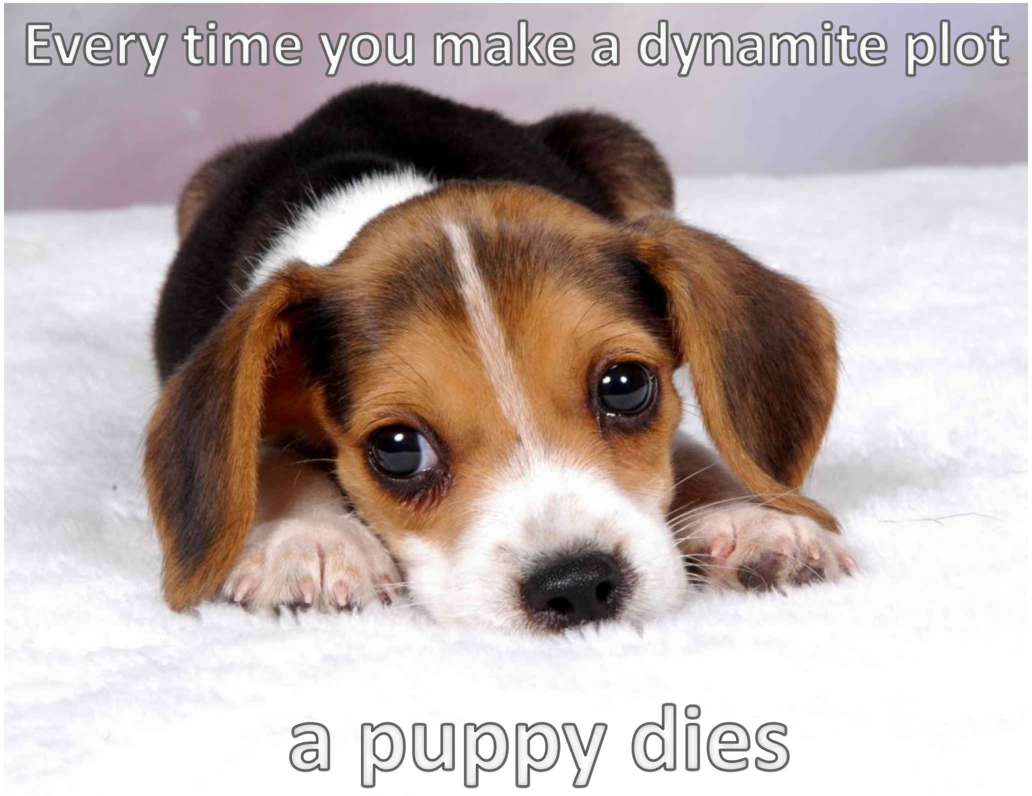






Bar plots with error bars

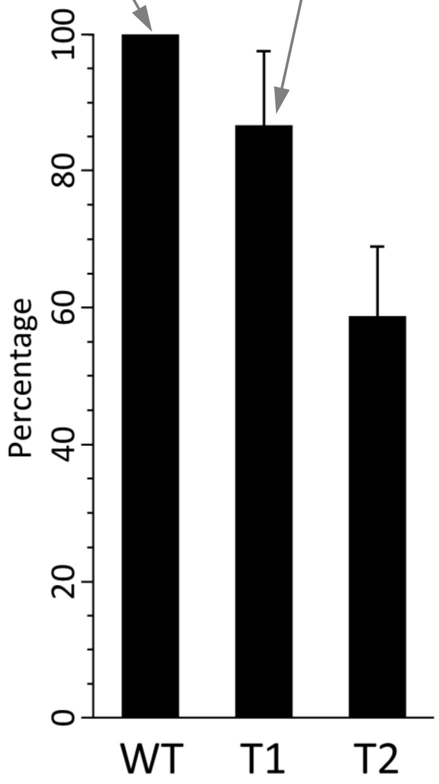

Lower error bar invisible



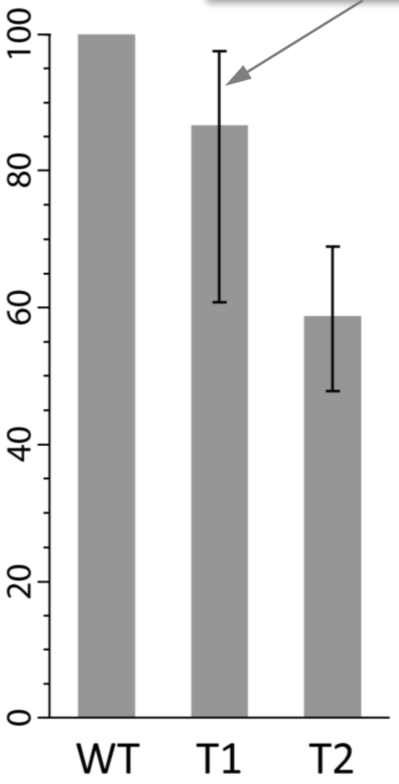
Bar plots with error bars

Useless information

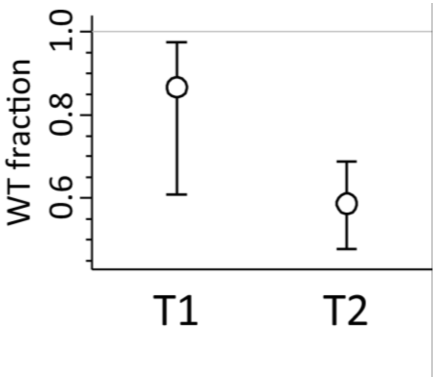
Lower error bar invisible



Always show upper and lower error bars



Other types of plots might be better

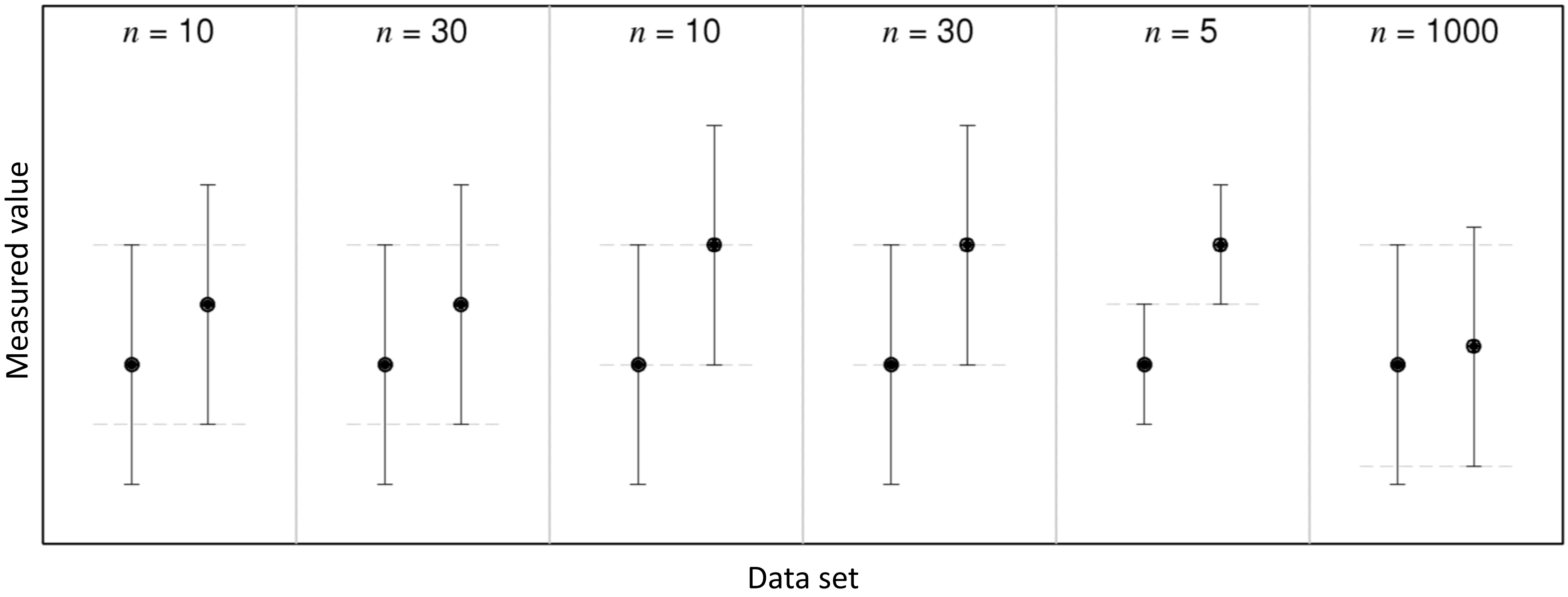


$T1 = 0.87^{+0.11}_{-0.26}$

$T2 = 0.59^{+0.10}_{-0.11}$

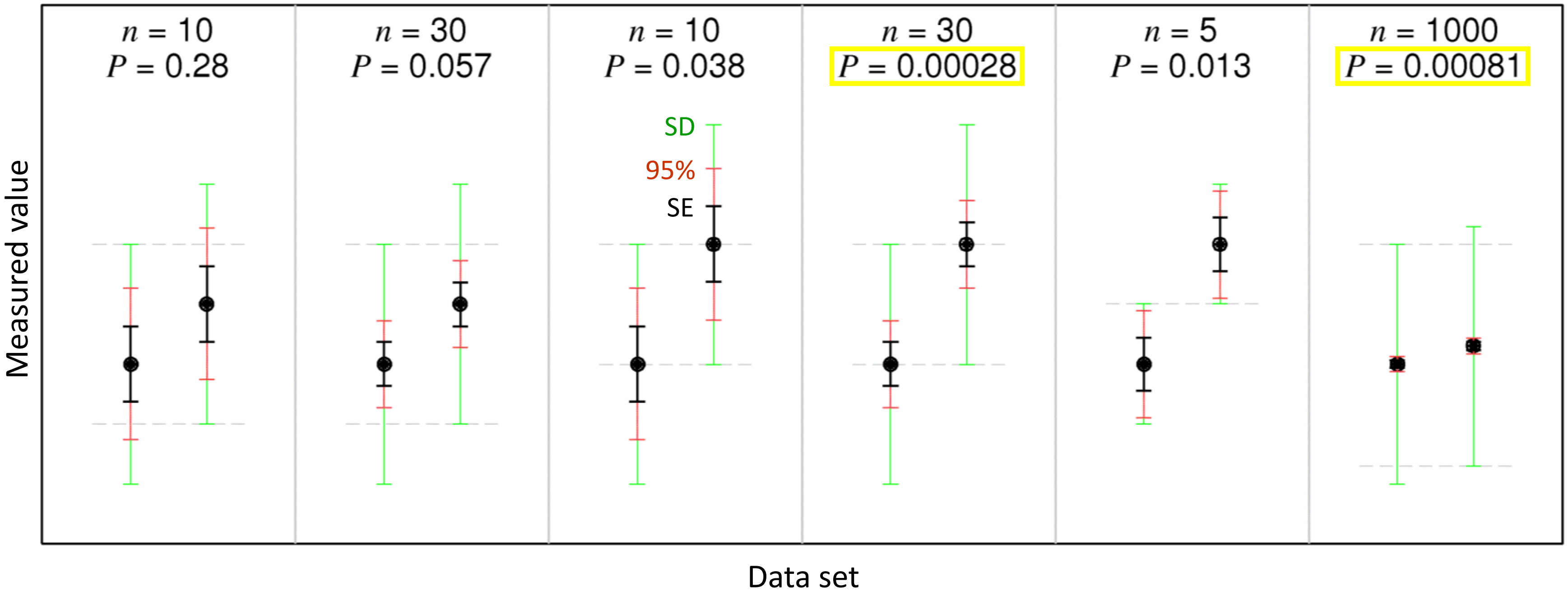
Do you really need a plot?

Exercise: overlapping error bars



- Each panel shows results from a pair of samples of the same size
- **Mean and standard deviation** are shown
- Which of these pairs of means are significantly different?

Exercise: overlapping error bars



- Error bars from inside: standard error of the mean, 95% confidence interval for the mean, standard deviation
- A rule of thumb: “when 95% CI overlap, the difference is not significant”
- Remember: 95% $CI \approx 2SE$
- But only a statistical test, for example a t-test, will tell you real significance!

Rules of making good graphs

1. Always keep **clarity of presentation** in mind
2. You shall use axes with scales and labels
3. Use logarithmic scale to show data spanning over many orders of magnitude
4. All labels and numbers should be easy to read
5. Symbols shall be easy to distinguish
6. Add error bars were possible
7. Always state what type of uncertainty is represented by your error bars
8. Use model lines, where appropriate
9. It is OK to join data points with lines for guidance
10. You shall not use bar plots unless necessary

Bar plots: recommendations

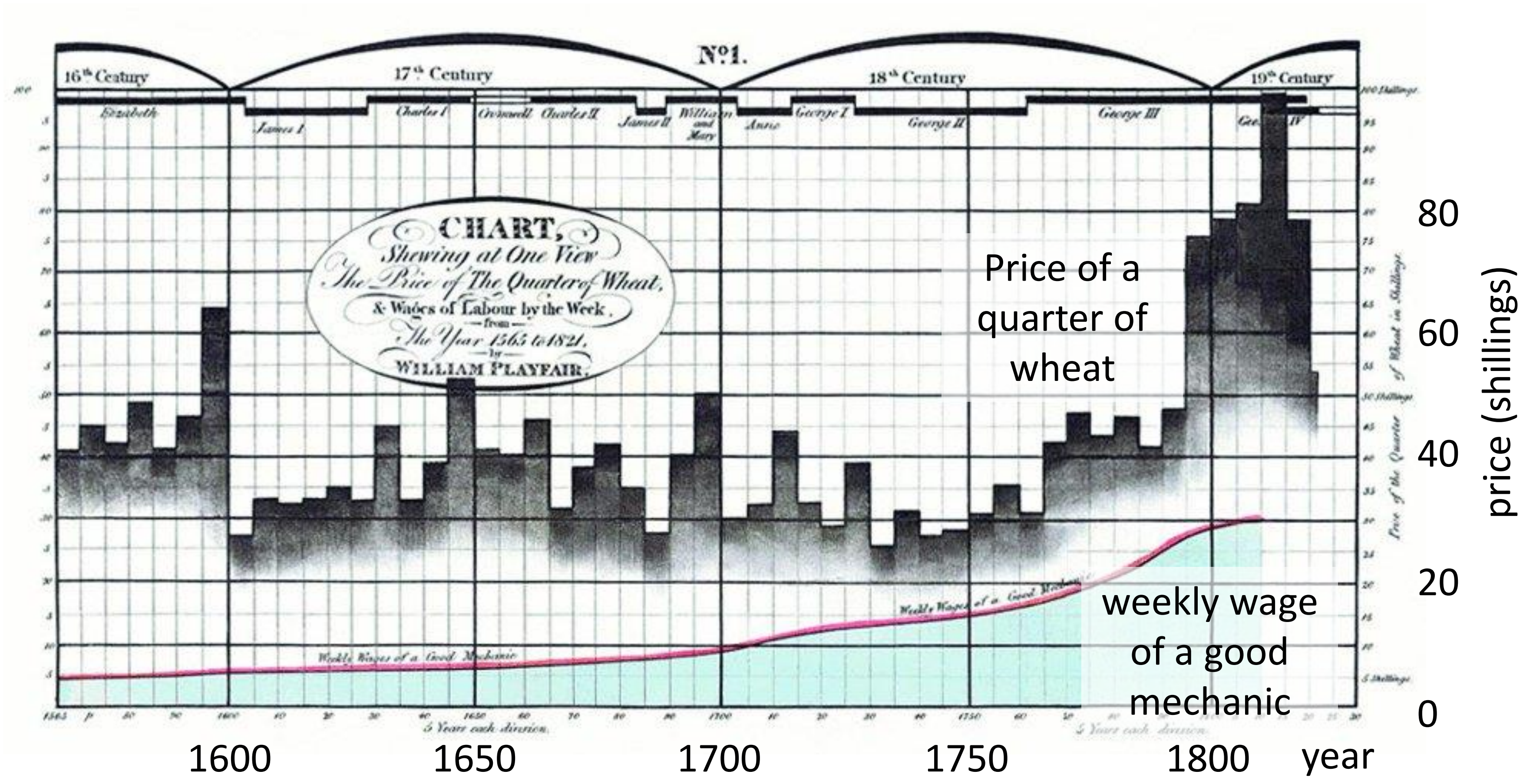
1. Bar plots should only be used to present additive quantities: counts, proportions and probabilities
2. Often it is to show whole data instead, e.g., a box plot or a histogram
3. Each bar has to start at zero
4. Don’t even think of making a bar plot in the logarithmic scale
5. Bar plots are not useful for presenting data with small variability
6. Multiple data bar plots are not suited for plots where the horizontal axis represents a continuous variable
7. Multiple data bar plots can be cluttered and unreadable
8. Make sure both upper and lower errors in a bar plot are clearly visible
9. You shall not make dynamite plots. Ever

William Playfair

- Born in Liff near Dundee
- Man of many careers (millwright, engineer, draftsman, accountant, inventor, silversmith, merchant, investment broker, economist, statistician, pamphleteer, translator, publicist, land speculator, blackmailer, swindler, convict, banker, editor and journalist)
- He invented
  - line graph (1786)
  - bar plot (1786)
  - pie chart (1801)

William Playfair (1759-1823)

William Playfair



"Chart showing at one view the price of the quarter of wheat & wages of labour by the week" (1821)

6. Quoting numbers and errors

"46.345% of all statistics are made up"

Anonymous

What is used to quantify errors

- In a publication you typically quote:

$$x = \underset{\substack{\nearrow \\ \text{best estimate}}}{x_{\mathrm{best}}} \pm \underset{\substack{\nwarrow \\ \text{error}}}{\Delta\mathrm{x}}$$

- Error can be:
  - Standard deviation
  - Standard error of the mean
  - Confidence interval
  - Derived error
- Make sure you tell the reader what type of errors you use

Significant figures (digits)

- Significant figures (or digits) are those that carry meaningful information
- More s.f. – more information
- The rest is meaningless junk!
- Quote only significant digits

Example

- A microtubule has grown 4.1 µm in 2.6 minutes; what is the speed of growth of this microtubule?

$$\frac{4.1\ \mu\mathrm{m}}{2.6\ \mathrm{min}} = 1.576923077\ \mu\mathrm{m\ min}^{-1}$$

- There are only two significant figures (s.f.) in length and speed
- Therefore, only about two figures of the result are meaningful: 1.6 µm min^{-1}

Significant figures in writing

- Non-zero figures are significant
- Leading zeroes are not significant
    - 34, 0.34 and 0.00034 carry the same amount of information
- Watch out for trailing zeroes
    - before the decimal dot: not significant
    - after the decimal dot: significant

Number	Significant figures
365	3
1.**893**	4
4000	1 or 4
4$\times 10^3$	1
4.000$\times 10^3$	4
4000.**00**	6
0.000**34**	2
0.000**3400**	4

Rounding

- Remove non-significant figures by rounding

- Round the last s.f. according to the value of the next digit
  - □ 0-4: round down (**1.3**42 → 1.3)
  - □ 5-9: round up (**1.3**56 → 1.4)

- So, how many figures are significant?

Suppose we have 2 s.f. in each number

Raw number	Quote
1234	1200
1287	1300
1.491123	1.5
1.449999	1.4

Error in the error

- To find how many s.f. are in a number, you need to look at its error

- Use sampling distribution of the standard error

- Error in the error is

$$\Delta SE = \frac{SE}{\sqrt{2(n-1)}}$$

- This formula can be applied to SD and CI

Example

$n = 12$

$SE = 23.17345$

$$\Delta SE = \frac{23.17345}{\sqrt{2 \times 11}} \approx \frac{23.17}{4.69} \approx 4.94$$

$SE = 23.17 \pm 4.94$

- We can trust only one figure in the error

- Round SE to one s.f.:

$SE = 20$

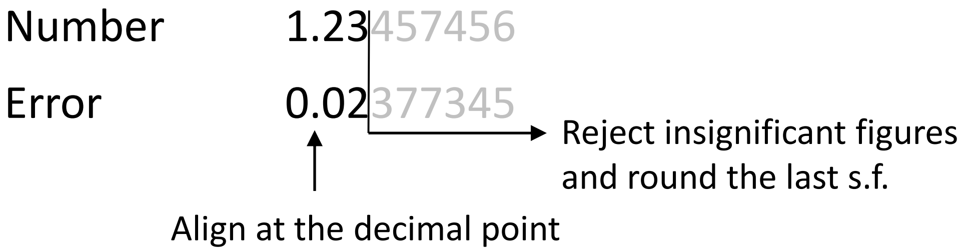
Error in the error

n	$\frac{\Delta SE}{SE}$	s.f. to quote
10	0.24	1
100	0.07	2
1,000	0.02	2
10,000	0.007	3
100,000	0.002	3

- An error quoted with 3 s.f. (2.567±0.165) implicitly states you have 10,000 replicates

Quote number and error

- Get a number and its error
- Find how many significant figures you have in the error (typically 1 or 2)
- Quote the number with the *same decimal precision* as the error



Number 1.23457456

Error 0.02377345

Reject insignificant figures and round the last s.f.

Align at the decimal point

Correct	Incorrect
1.23 ± 0.02	1.2 ± 0.02
1.2 ± 0.5	1.23423 ± 0.5
6.0 ± 3.0	6 ± 3.0
75000 ± 12000	75156 ± 12223
$(3.5 \pm 0.3) \times 10^{-5}$	$3.5 \pm 0.3 \times 10^{-5}$

Error with no error

- Suppose you have a number without error
- (Go back to your lab and do more experiments)

- For example
  - Centromeres are transported by microtubules at an average speed of 1.5 µm/min
  - The new calibration method reduces error rates by ~5%
  - Transcription increases during the first 30 min
  - Cells were incubated at 22°C

- There is an **implicit error** in the last significant figure
- All quoted figures are presumed significant

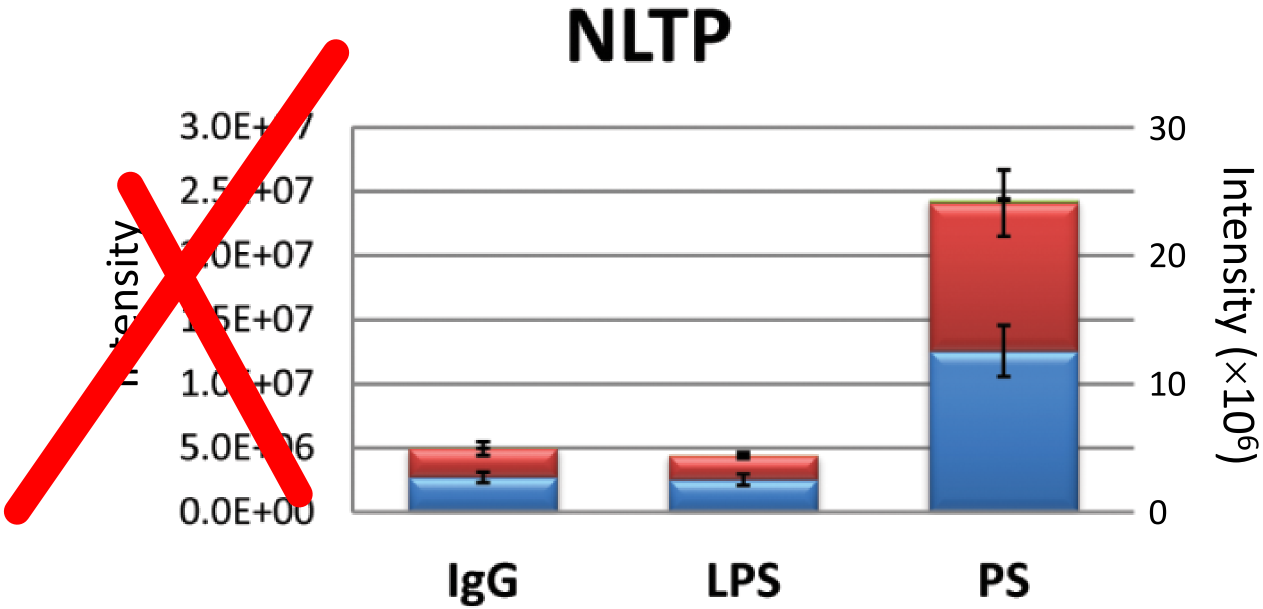
Avoid computer notation

- Example from a random paper off my shelf

  "p-value = 5.51E-14"

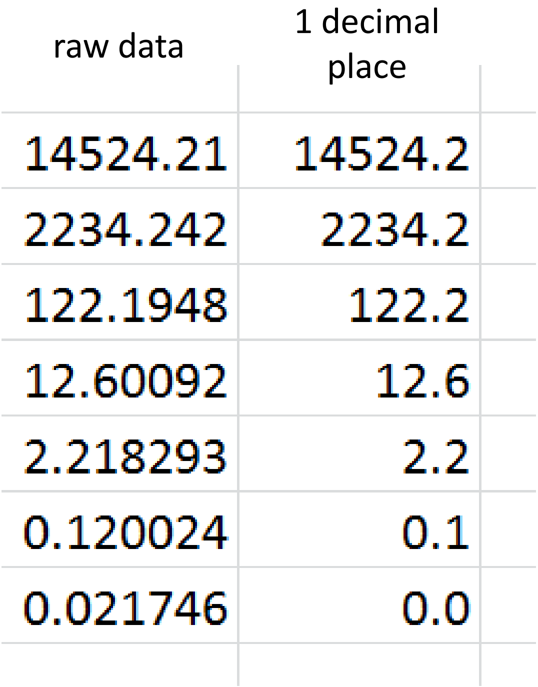
- I'd rather put it down as

  p-value = 6×10^{-14}



Fixed decimal places

- Another example, sometimes seen in papers
- Numbers with fixed decimal places, copied from *Excel*
- Typically fractional errors are similar and we have the same number of s.f.



raw data	1 decimal place
14524.21	14524.2
2234.242	2234.2
122.1948	122.2
12.60092	12.6
2.218293	2.2
0.120024	0.1
0.021746	0.0

Wrong	Right
14524.2	1.5×10^4
2234.2	2200
122.2	120
12.6	13
2.2	2.2
0.1	0.12
0.0	0.022

Assume there are only
2 s.f. in these measurements

How to quote numbers (and errors)

WHEN YOU KNOW ERROR

- First, calculate the error and estimate its uncertainty
- This will tell you how many significant figures of the error to quote
- Typically you quote 1-2 s.f. of the error
- Quote the number with **the same precision as the error**
  - □ 1.23 ± 0.02
  - □ 1.23423 ± 0.00005 (rather unlikely in biological experiments)
  - □ 6 ± 3
  - □ 75 ± 12
  - □ $(3.2 \pm 0.3) \times 10^{-5}$

WHEN YOU DON'T KNOW ERROR

- You still need to guesstimate your error!
- Quote only figures that are significant, e.g. $p = 0.03$, not $p = 0.0327365$
- Use common sense!
- Try estimating order of magnitude of your uncertainty
- Example: measure distance between two spots in a microscope
  - □ Get 416.23 nm from computer software
  - □ Resolution of the microscope is 100 nm
  - □ Quote 400 nm

Rounding numbers
0-4: down (6.64 → 6.6)
5-9: up (6.65 → 6.7)

University of Dundee

THE BARTON GROUP

The Data Analysis Group

Hand-outs available at http://tiny.cc/statlec

Please leave your feedback forms on the table by the door

Centre for Gene Regulation & Expression

wellcome trust