# 5. Data presentation

“Above all else show the data”

*Edward R. Tufte*

# How to make a good plot

# A good plot

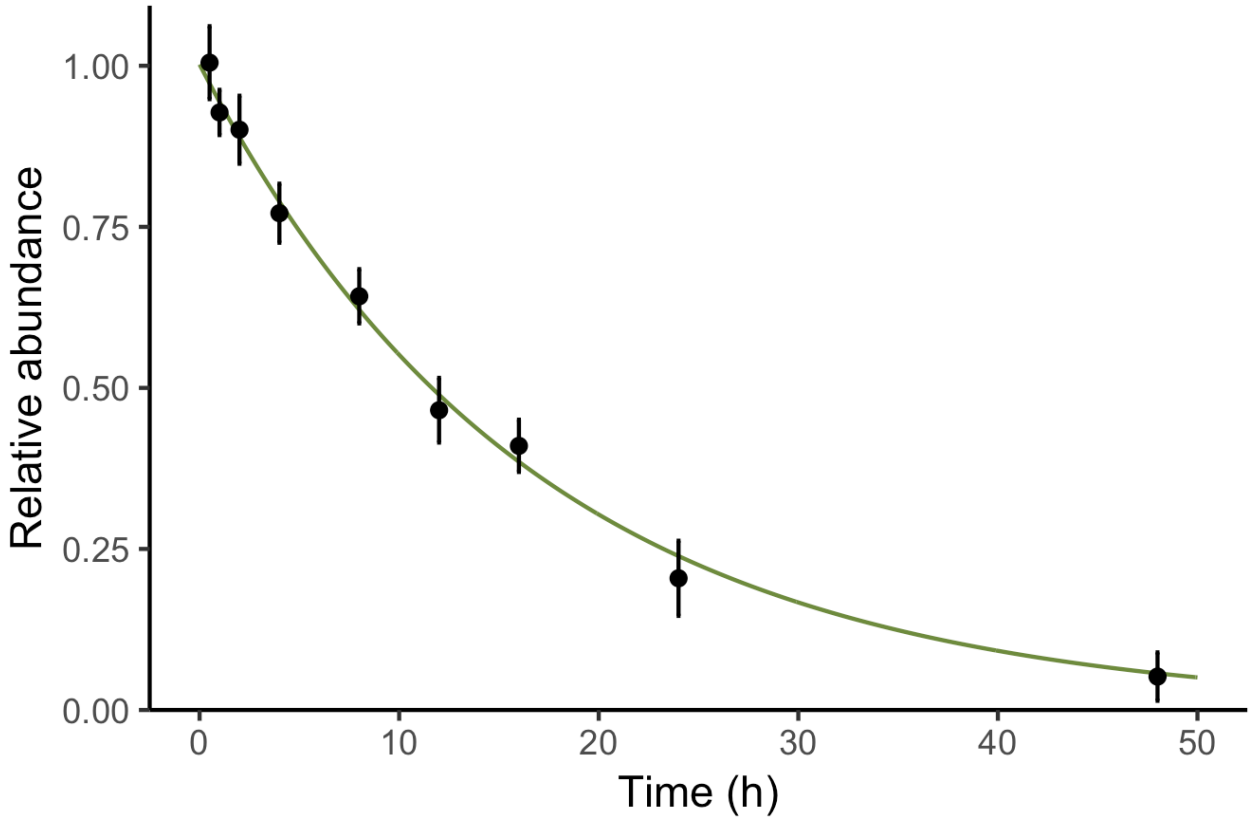



Figure 6-1. Exponential decay of a protein in a simulated experiment. Error bars represent standard errors. The curve shows the best-fitting exponential decay model, $y(t) = Ae^{-t/\tau}$, with $A = 1.00 \pm 0.03$ and $\tau = 17 \pm 1$ h (95% confidence intervals).

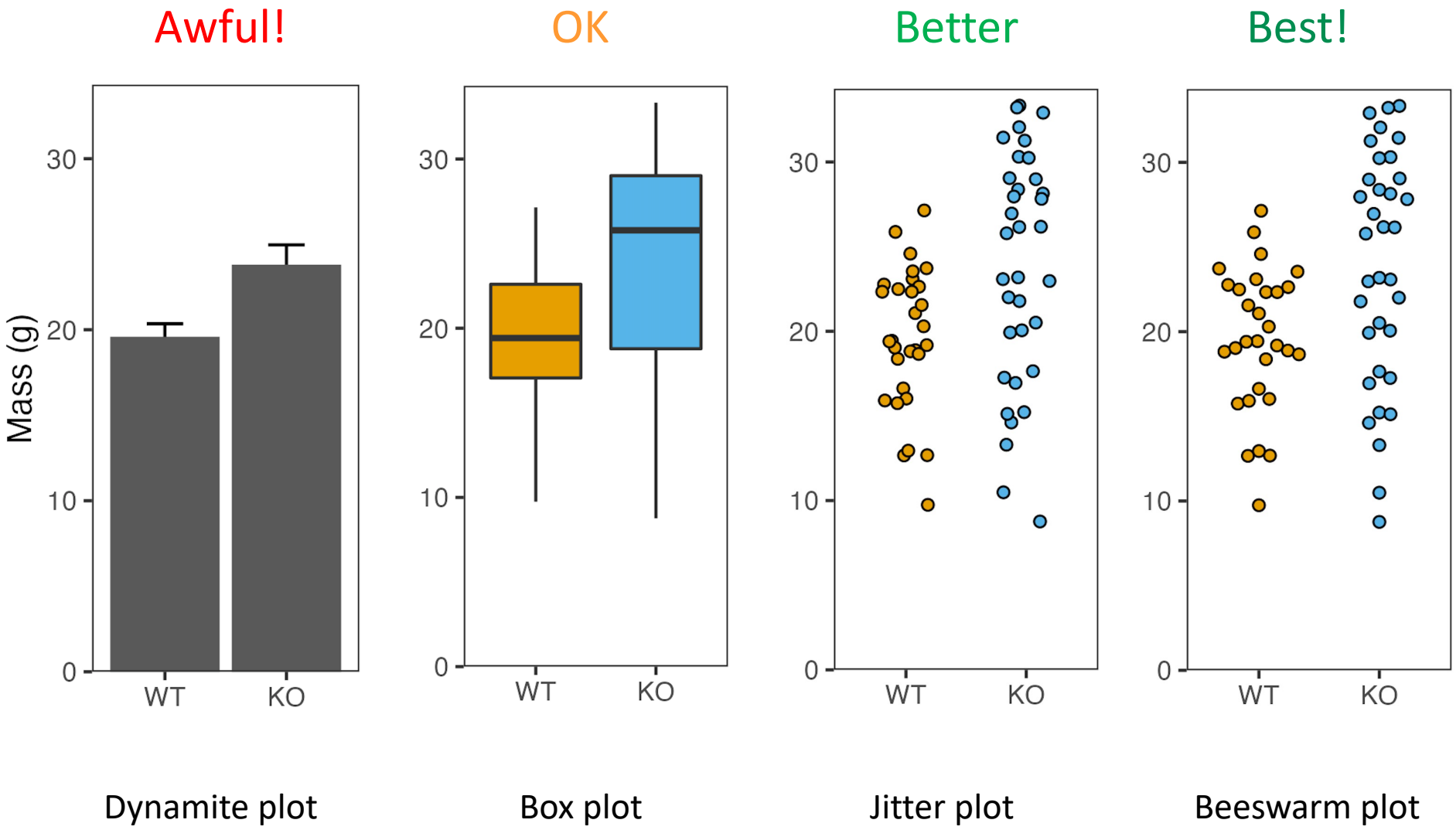
# 3 rules for making good plots

1. Clarity of presentation

2. Clarity of presentation

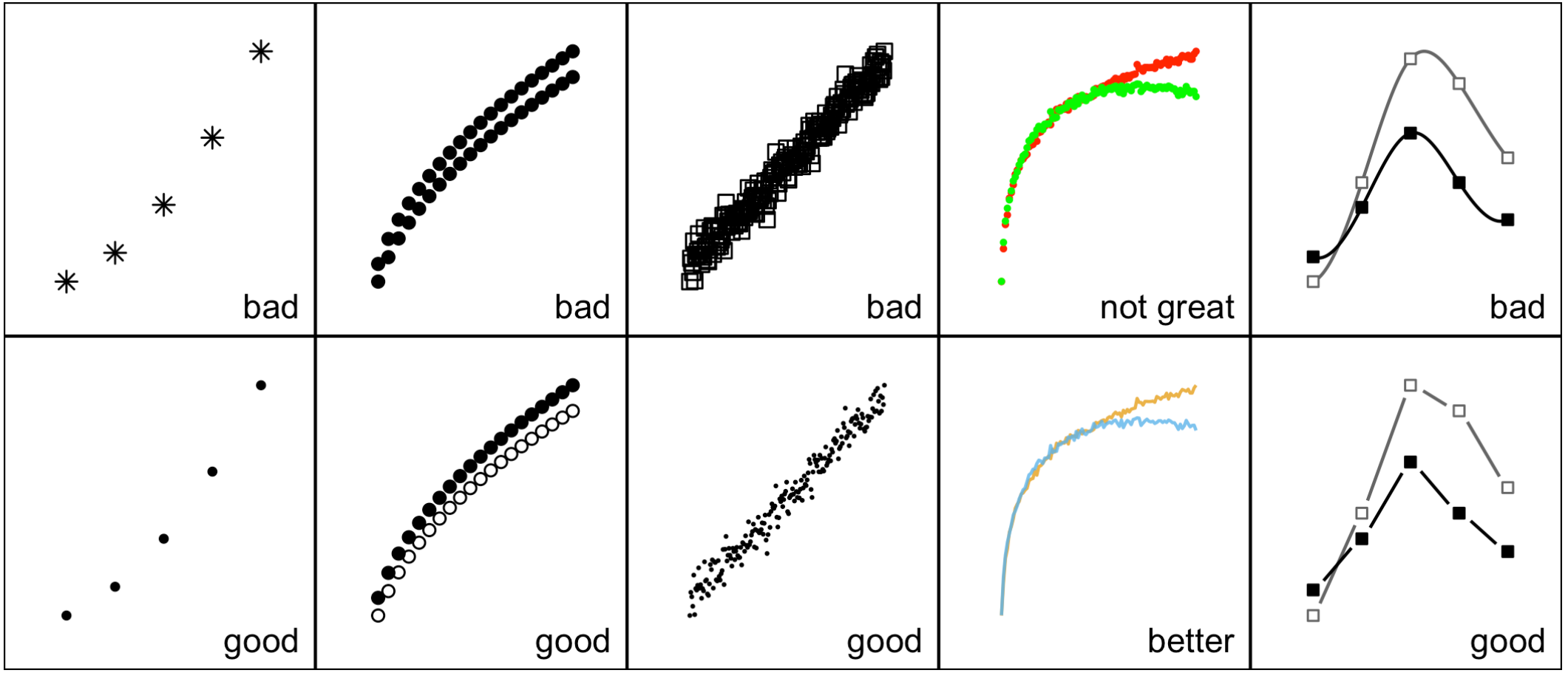
3. Clarity of presentation

# Show your data!



```
library(ggplot2)
library(ggbeeswarm)
library(dplyr)

set.seed(1001)
n1 <- 30
n2 <- 35
d <- data.frame(
  type = c(rep("WT", n1), rep("KO", n2)),
  value = c(rnorm(n1, 20, 5), rnorm(n2, 25, 7))
)
d$type <- relevel(d$type, ref="WT")

dm <- d %>% group_by(type) %>% summarise(M = mean(value), SE = sd(value) / sqrt(n()))

g0 <- ggplot() +
  theme_classic() + theme(legend.position = "none") +
  scale_fill_manual(values=c("#E69F00", "#56B4E9")) + labs(x=NULL, y=NULL)

g0 +
  geom_errorbar(data=dm, aes(x=type, ymin=M-SE, ymax=M+SE), width=0.3, colour="black") +
  geom_col(data=dm, aes(x=type, y=M)) +
  scale_y_continuous(expand=c(0,0), limits=c(0, max(dm$M + dm$SE) * 1.03 )) +
  labs(y="Mass (g)")

g0 +
  geom_boxplot(data=d, aes(x=type, y=value, fill=type))

g0 +
  geom_jitter(data=d, aes(x=type, y=value, fill=type), shape=21, width=0.2, height=0)

g0 +
  geom_beeswarm(data=d, aes(x=type, y=value, fill=type), shape=21, cex=4.5)
```

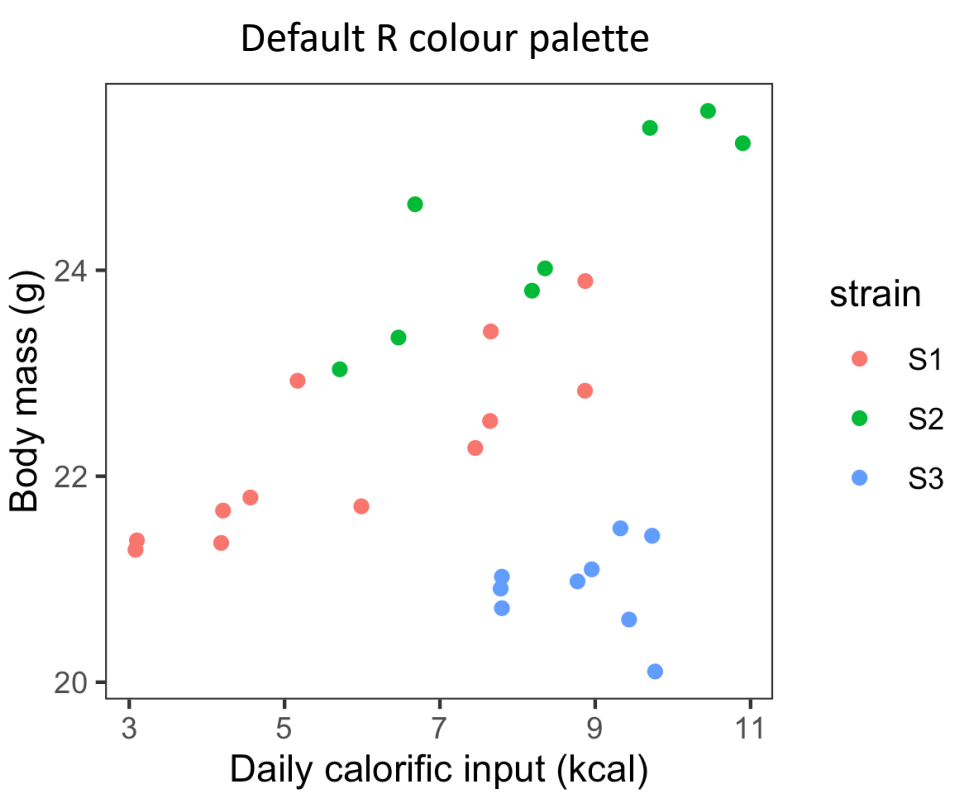
# Lines and symbols

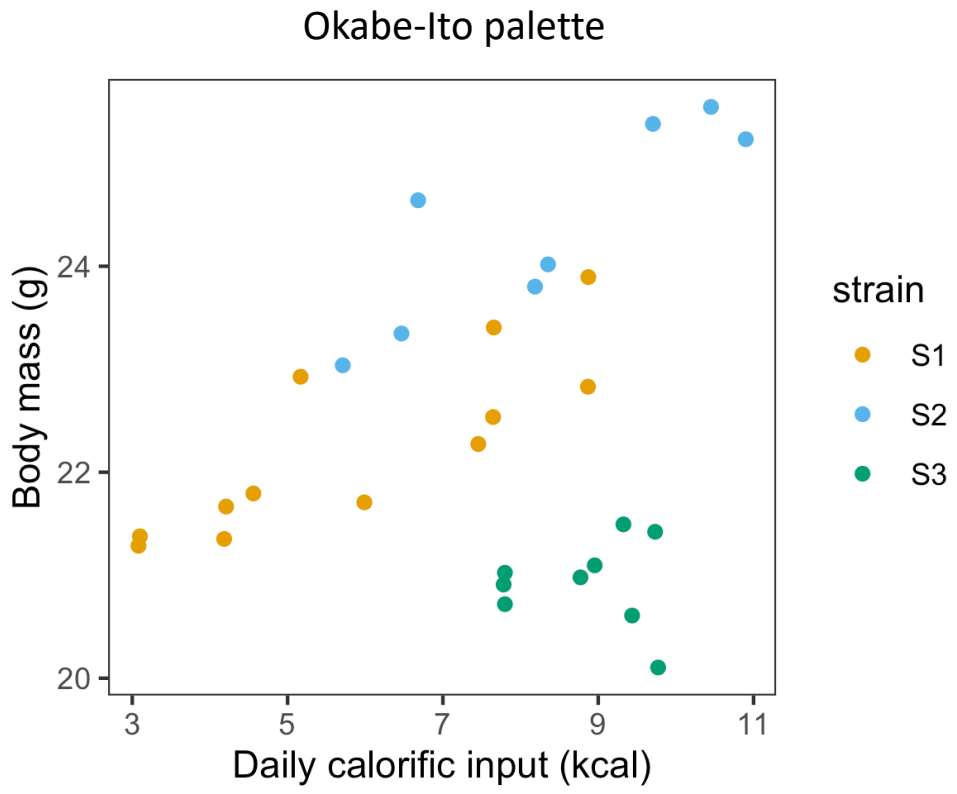



- Clarity!
- Symbols shall be easy to distinguish
- It is OK to join data points with lines for guidance

# Make your plots colour-blind friendly

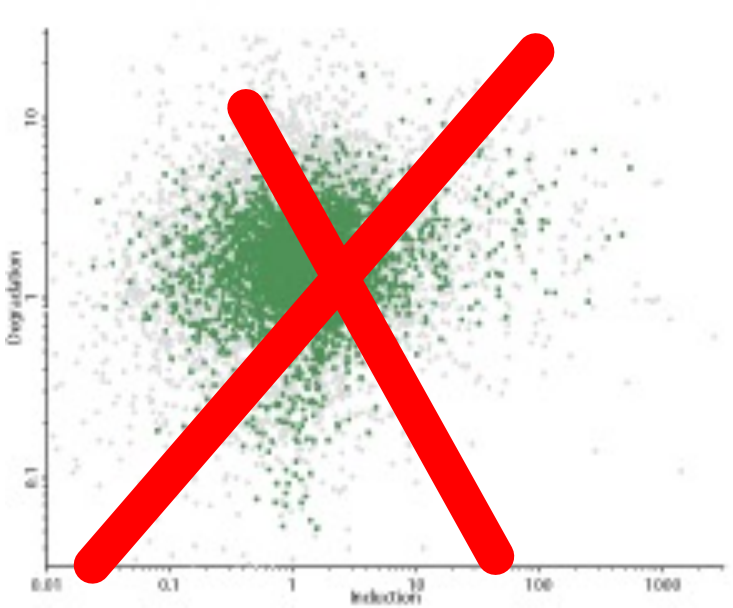
- Colour-blindness affects about 8% men and 0.5% women
- Depending on the type of colour blindness people have difficulties distinguishing some colours, for example red and green

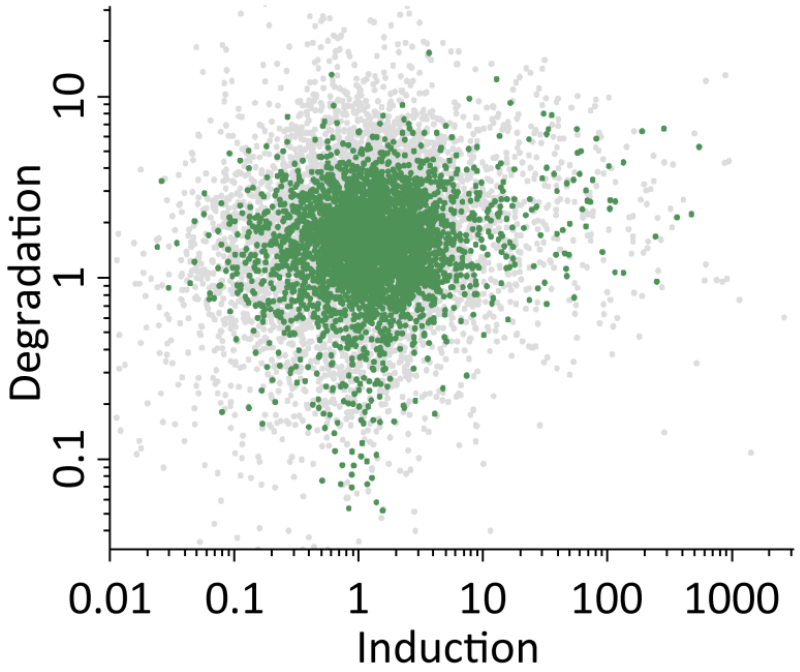


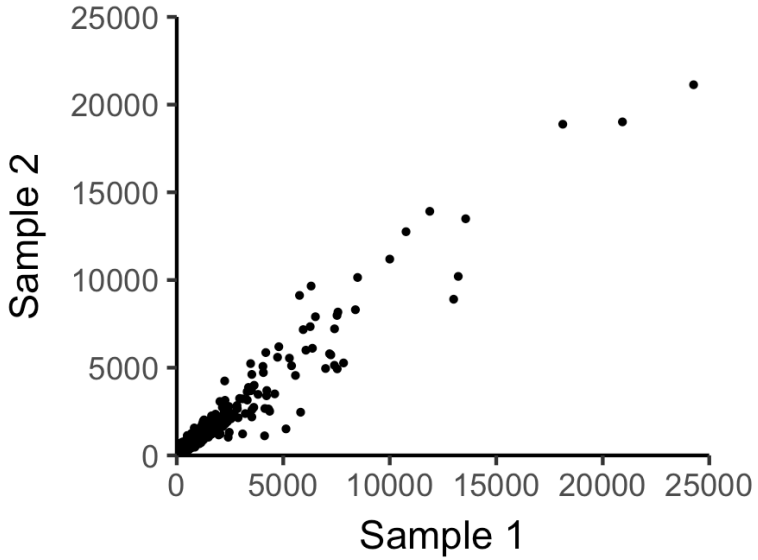

```
okabe_ito_palette <- c("#E69F00", "#56B4E9", "#009E73", "#F0E442", "#0072B2", "#D55E00", "#CC79A7")
```

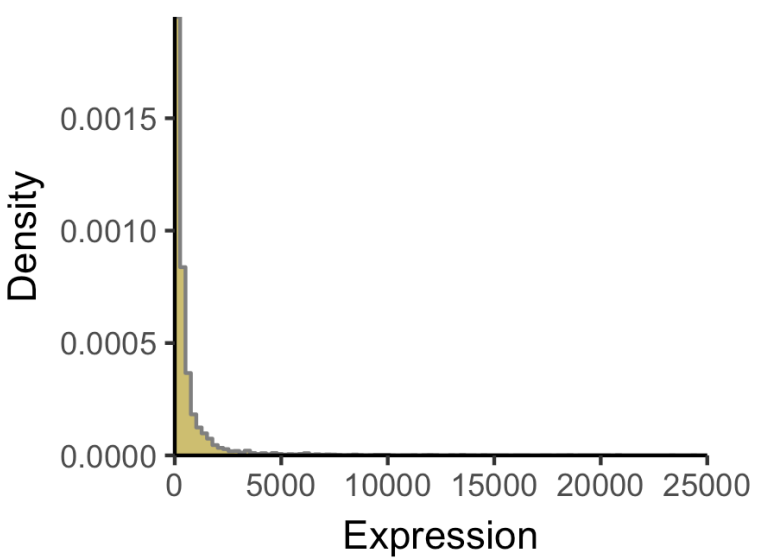
# Labels!

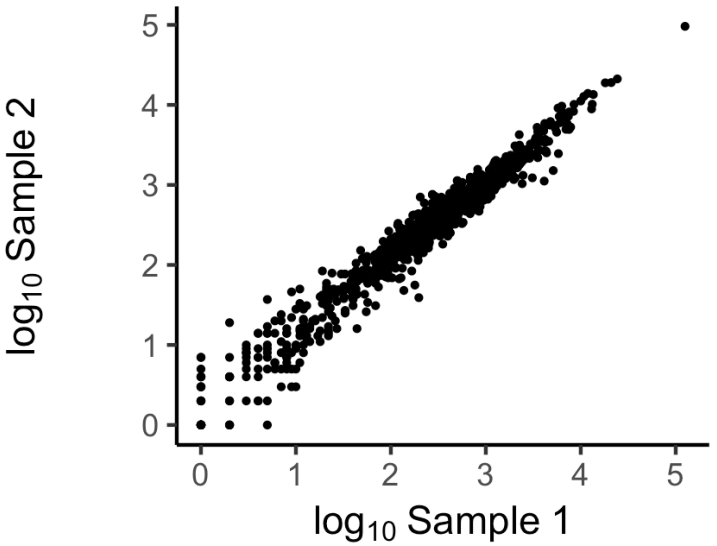


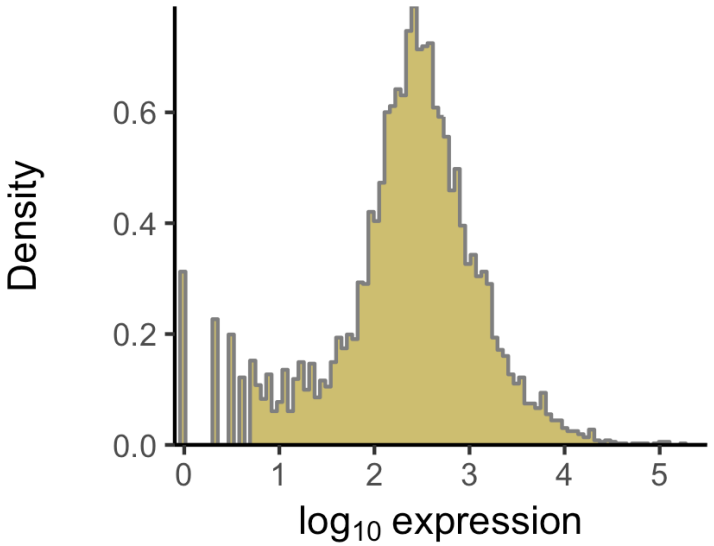


# Logarithmic plots

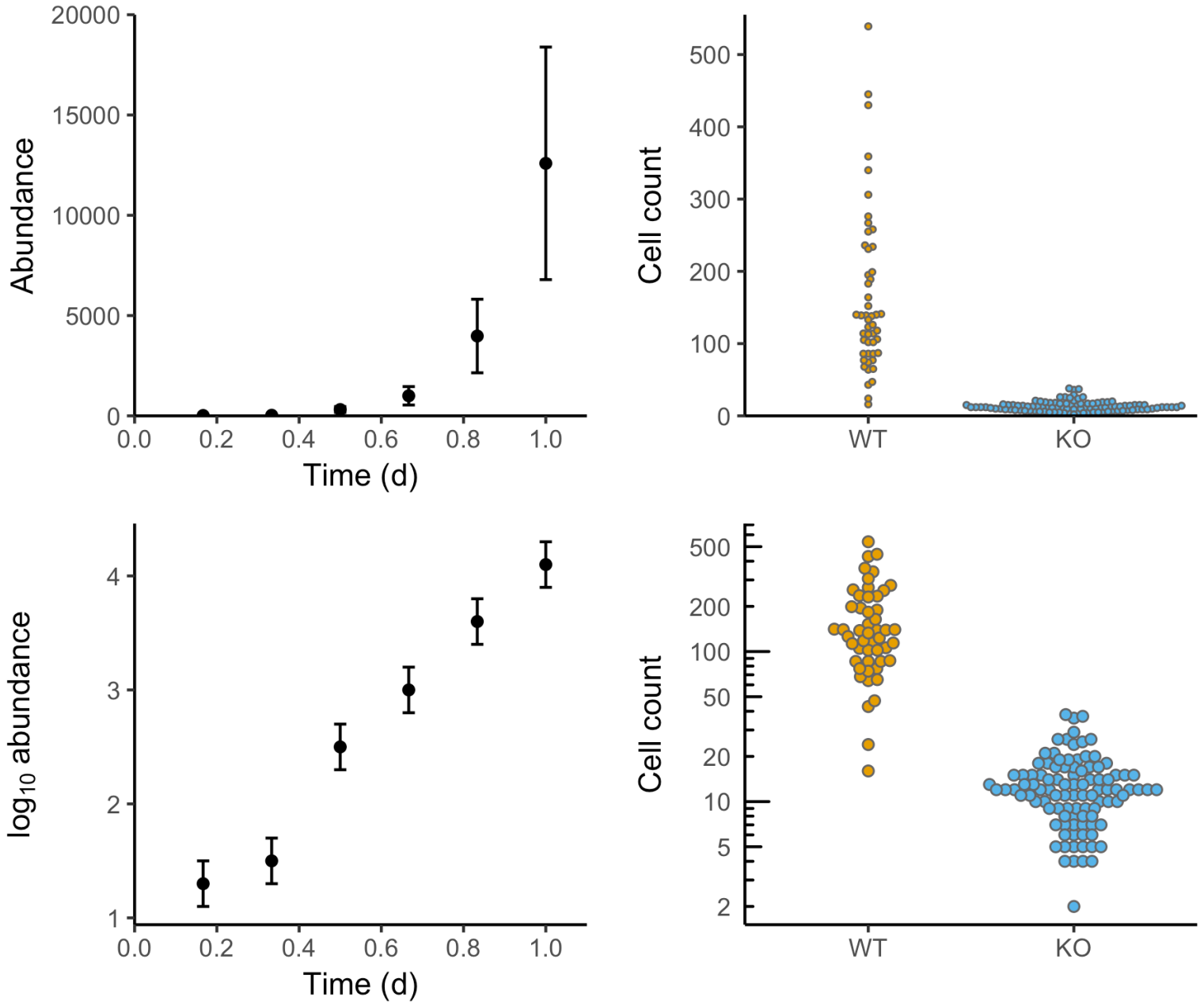


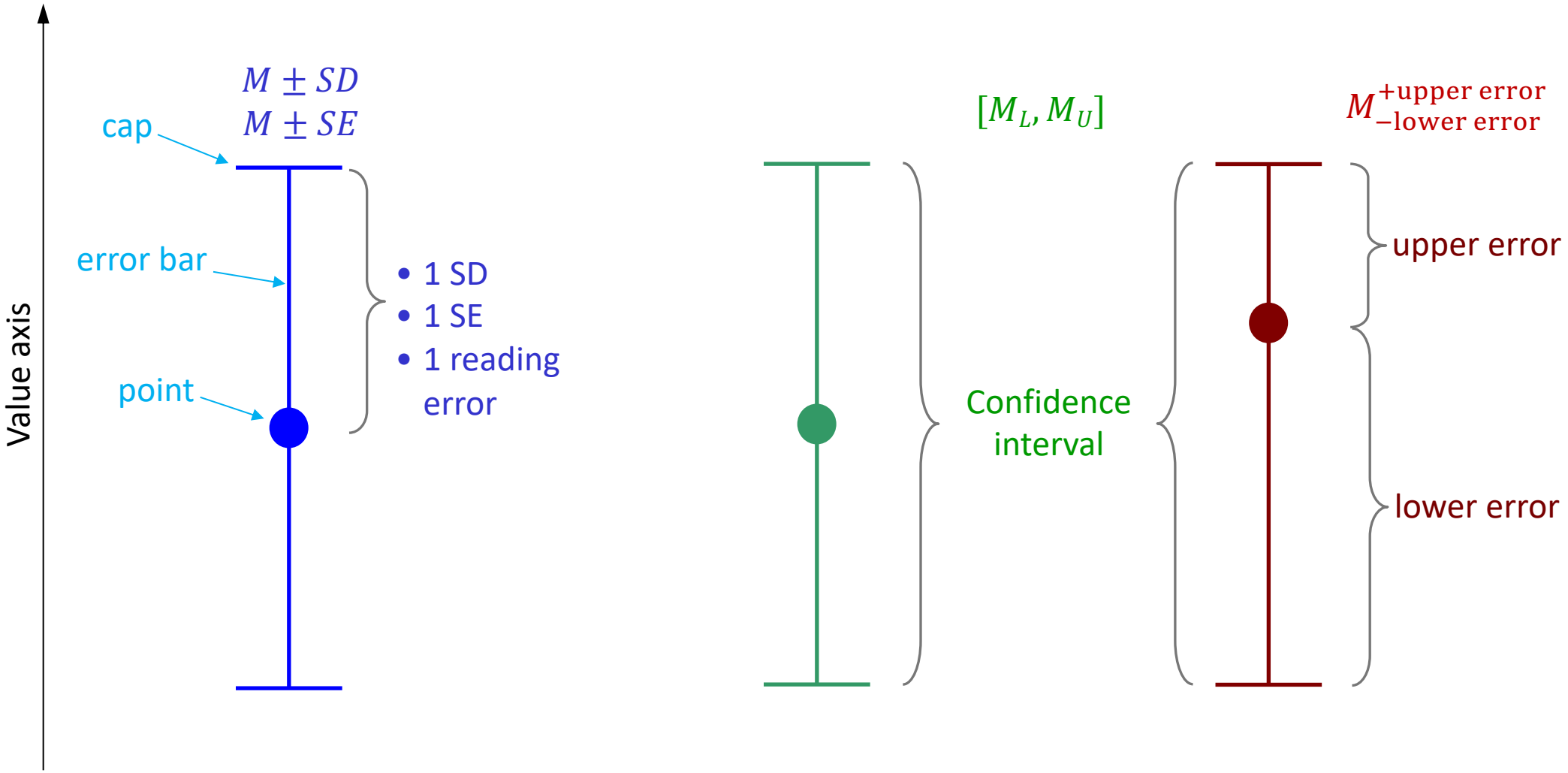



# Logarithmic plots

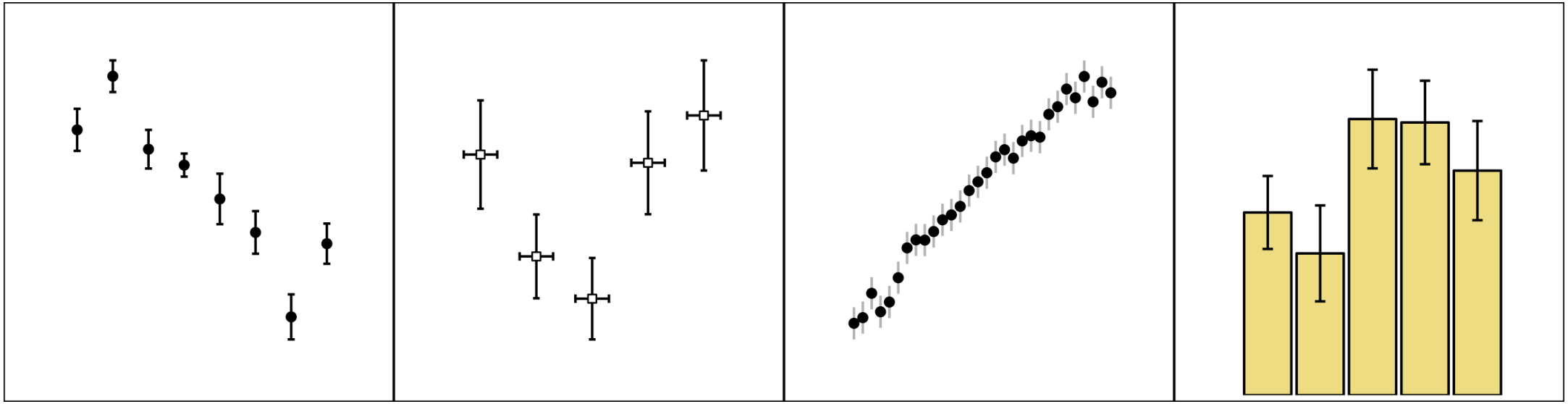


# Error bars

# How to plot error bars



# How to plot error bars



- Clarity!
- Make sure error bars are visible

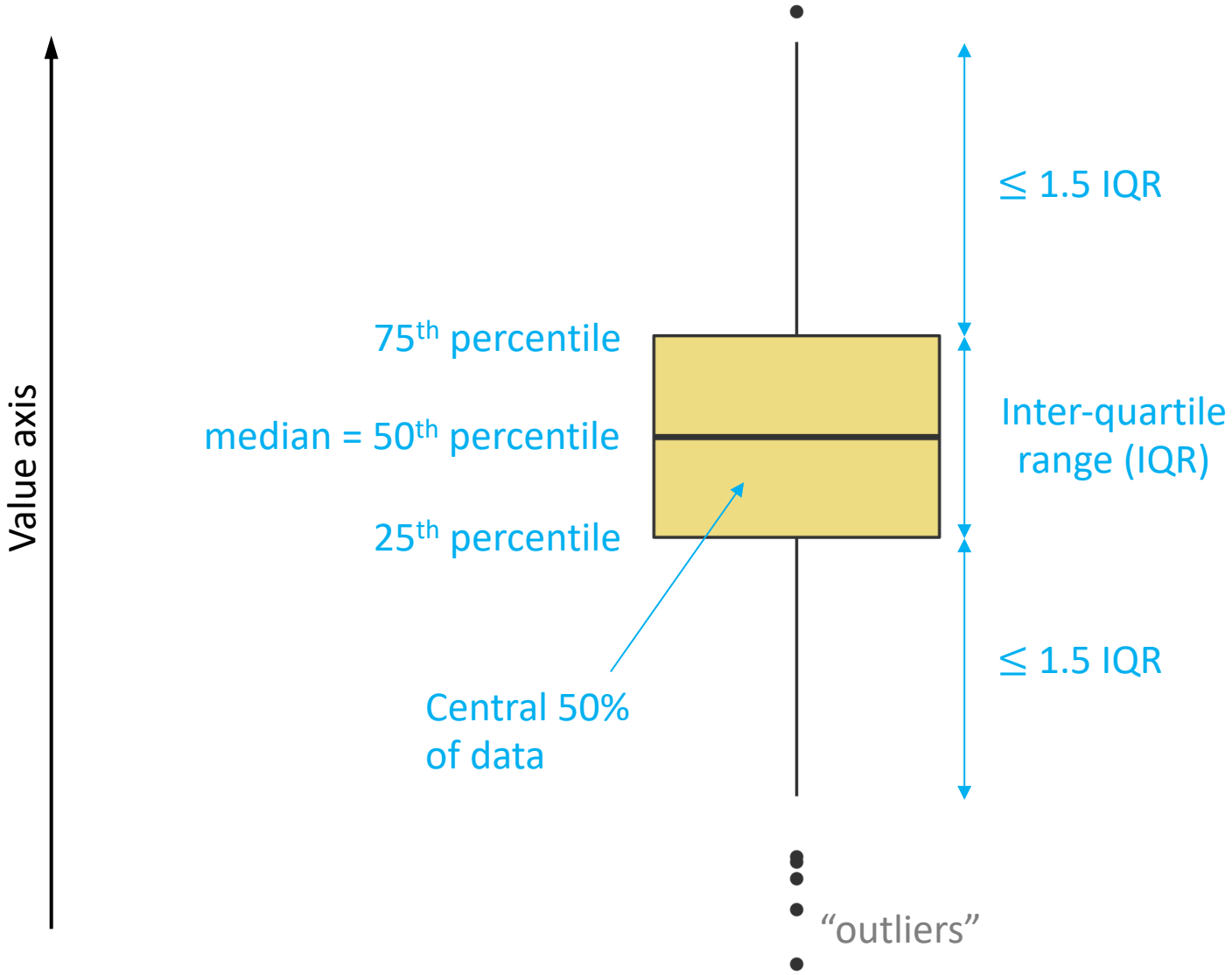
# Types of errors

| 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 |

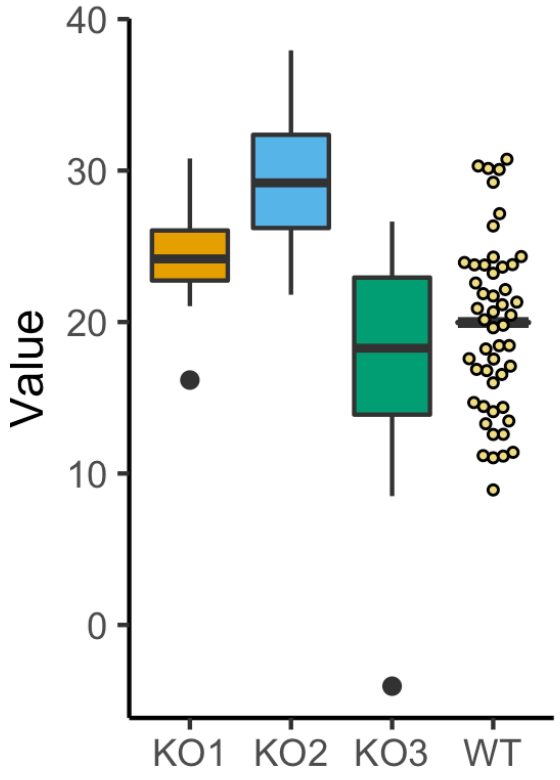
- Always state what type of uncertainty is represented by your error bars

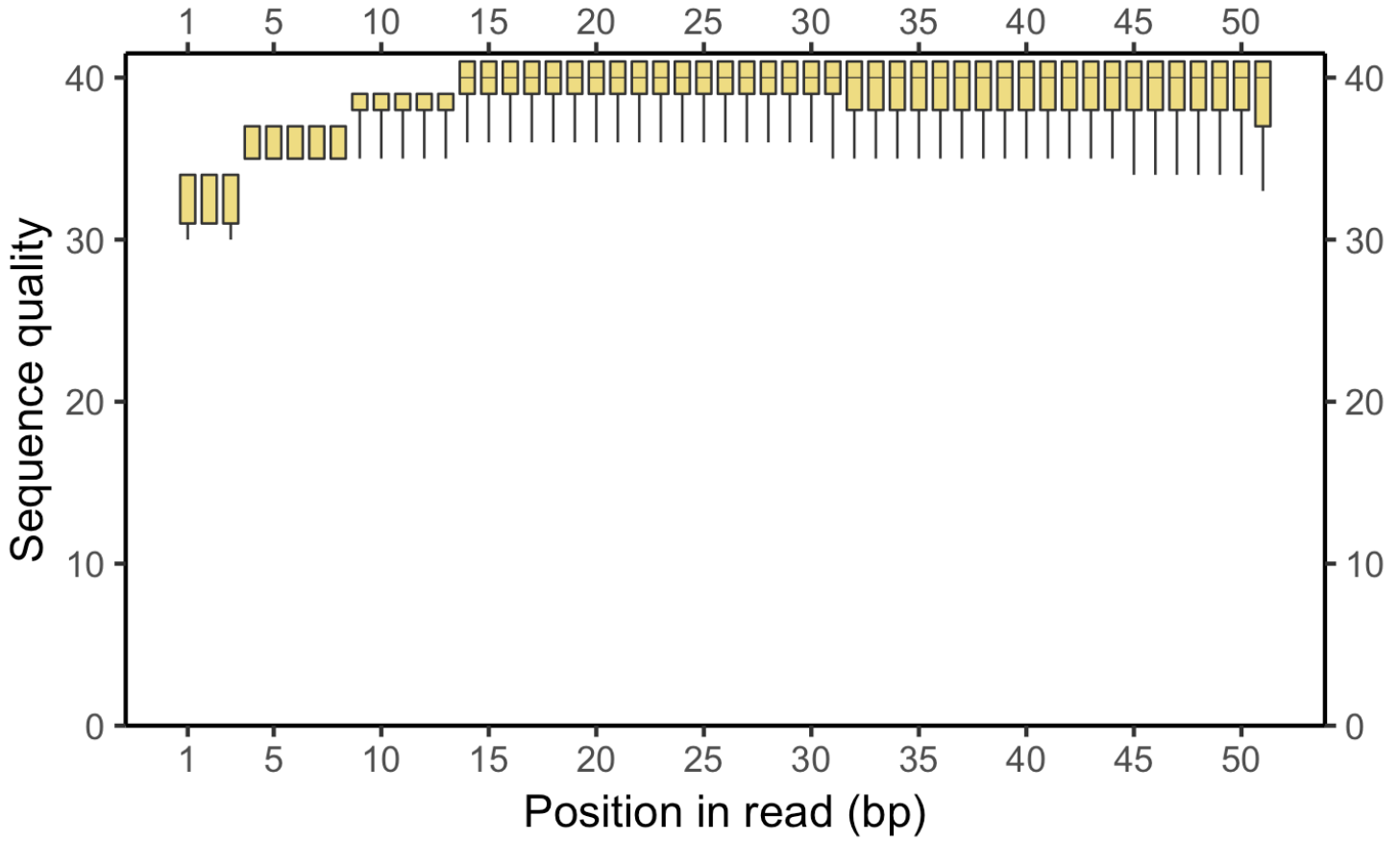
# Box plots

# Box plots


Value axis
≤ 1.5 IQR
75th percentile
median = 50th percentile
25th percentile
Inter-quartile range (IQR)
≤ 1.5 IQR
Central 50% of data
“outliers”

# Box plots

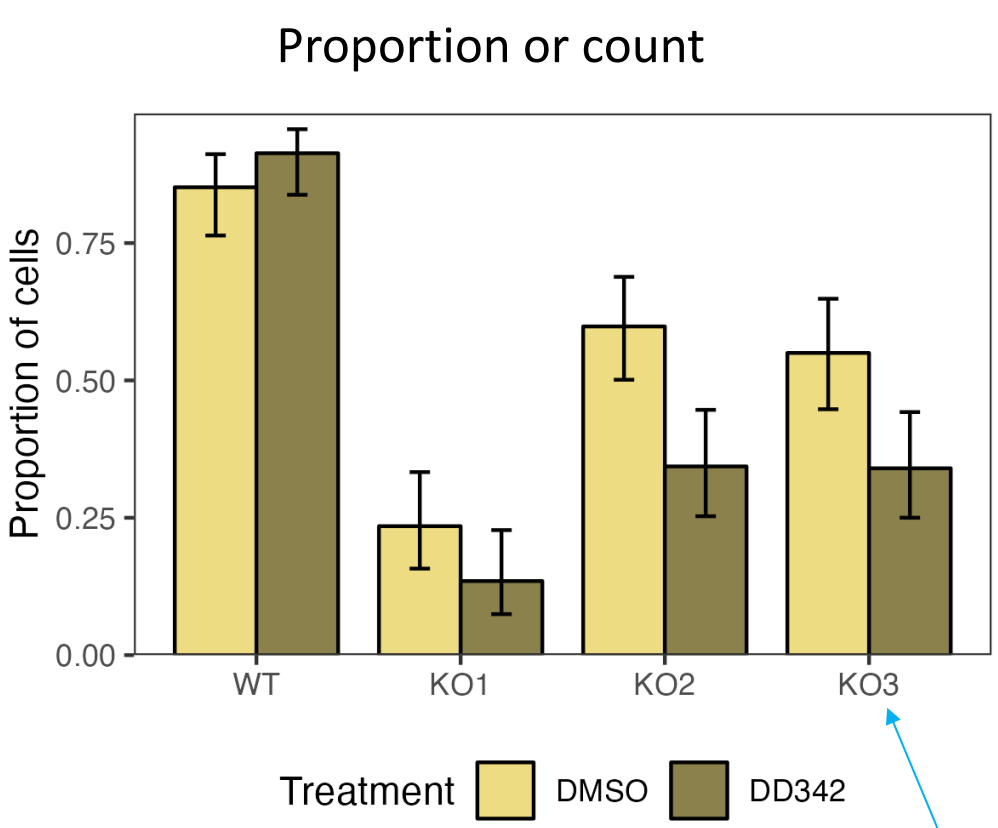


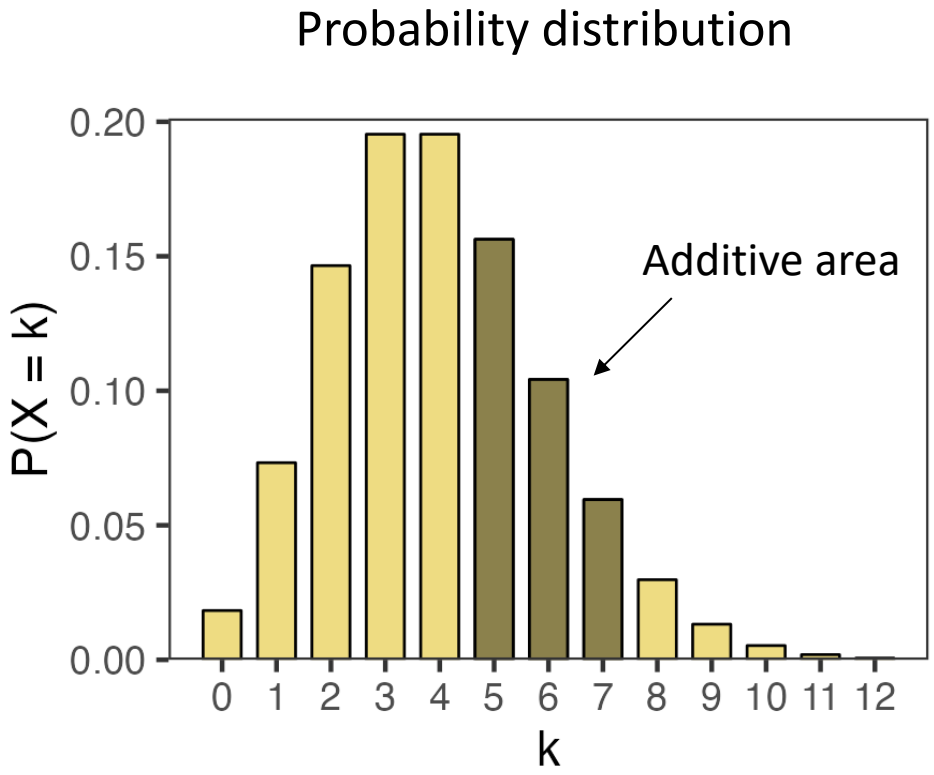




- Box plots are a good alternative to standard deviation error bars
- They are non-parametric and show pure data
- Useful for large number of data points

# Bar plots

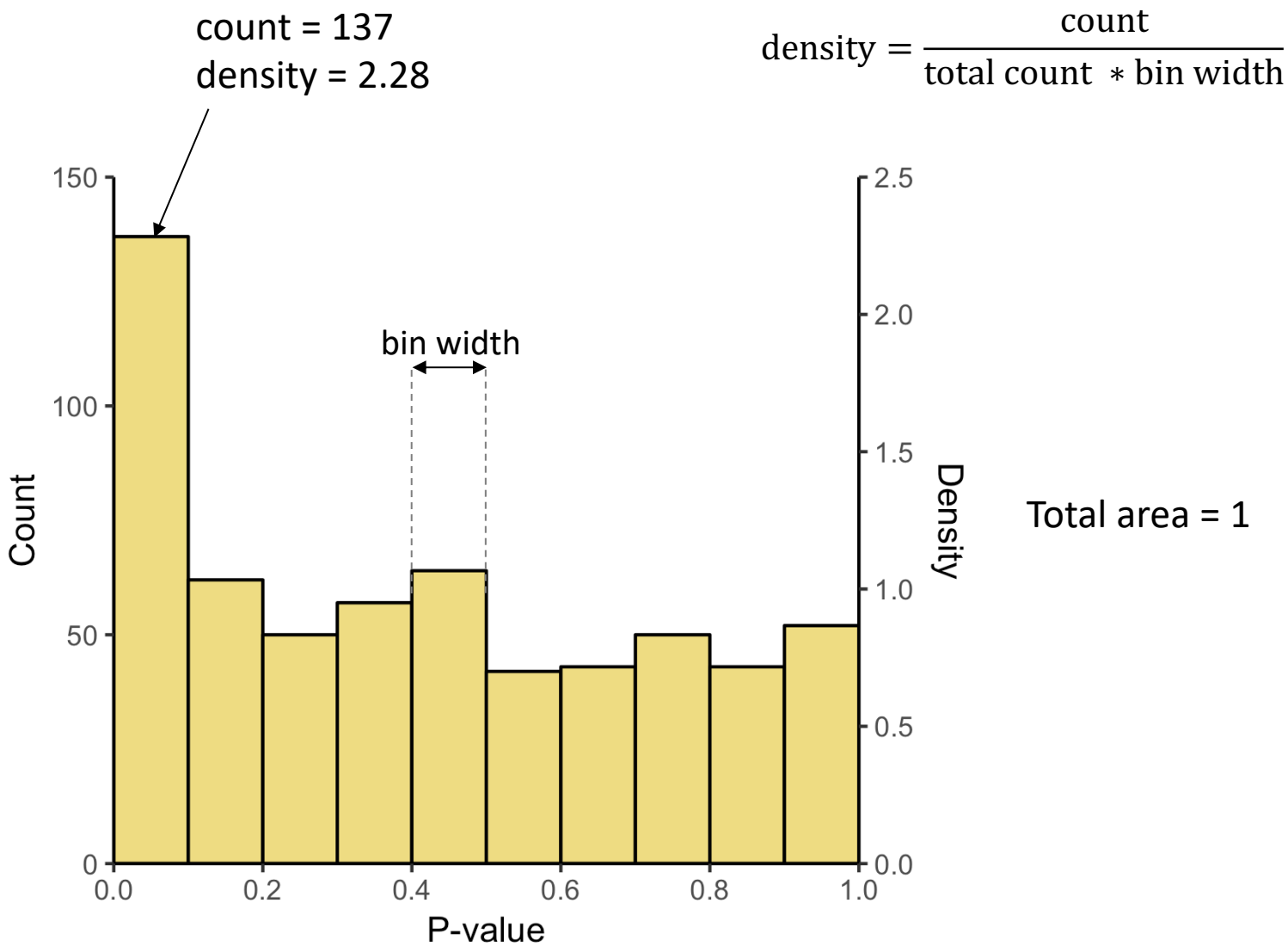
# Bar plots: categorical variable



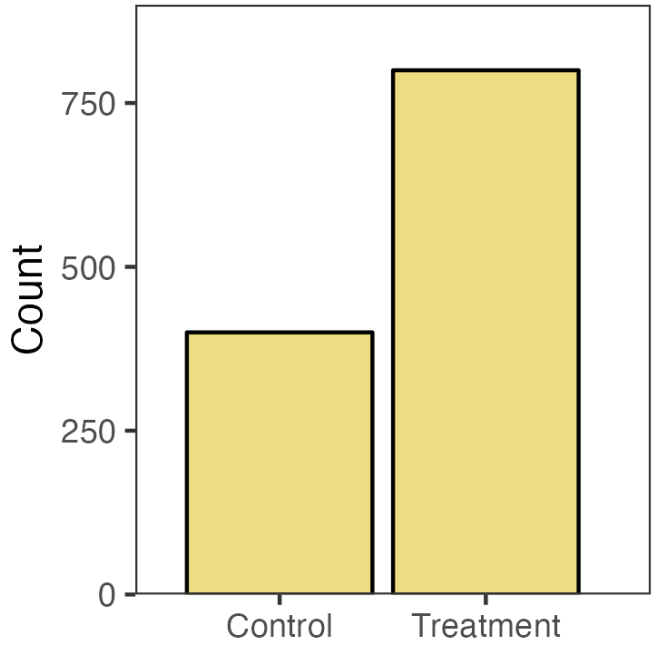



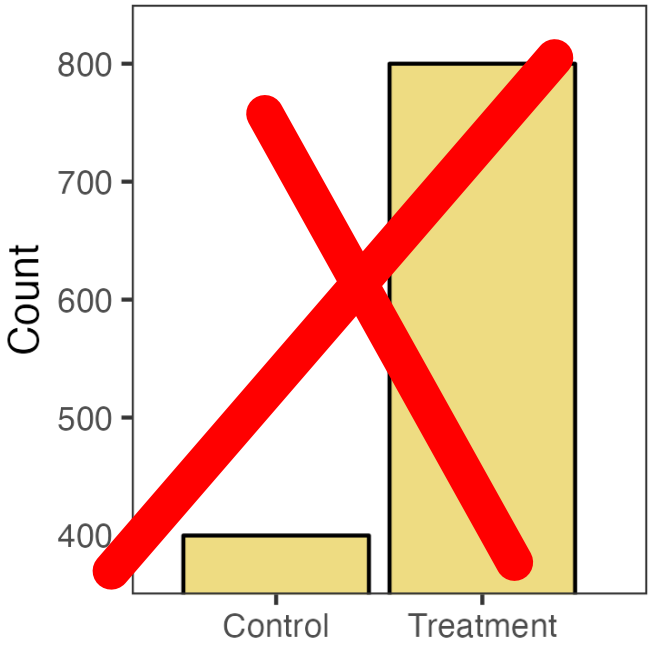

Category

# Bar plots: continuous variable (histogram)

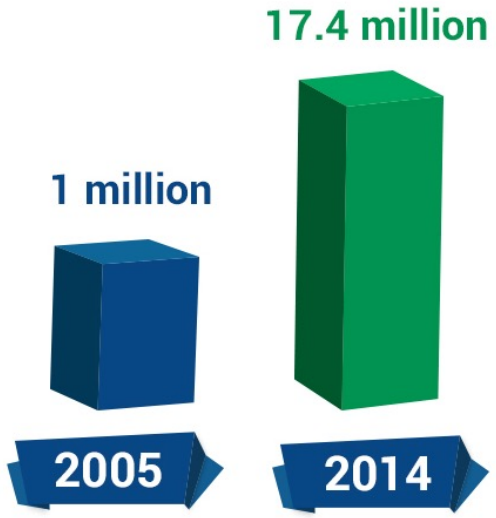


# Bar plots start at zero

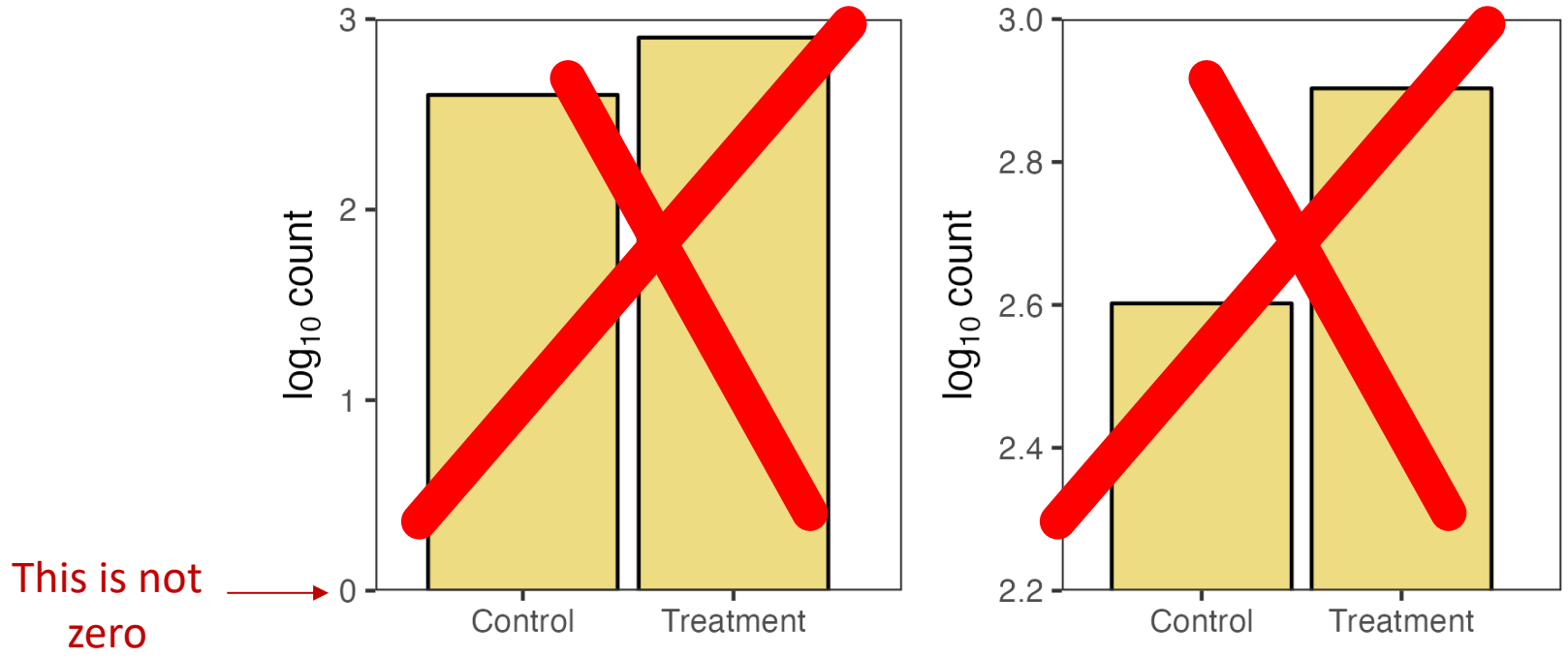



- Bar area represents its value
- Hence, baseline must be at zero!



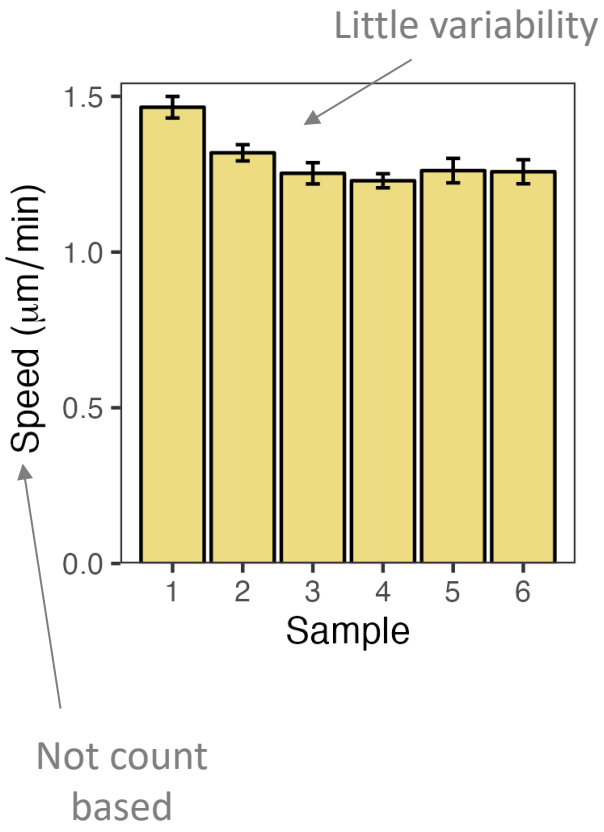
# Bar plots in logarithmic scale



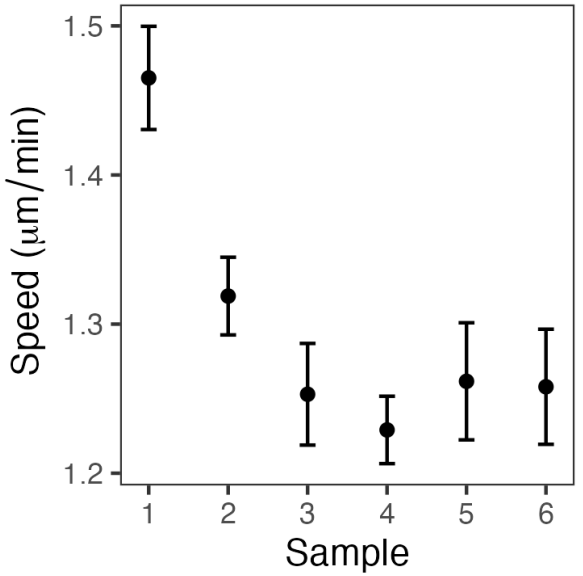

- There is no zero in a logarithmic scale!
- Bar size depends on an arbitrary lower limit of the vertical axis
- Don’t do it!

# Bar plot problems

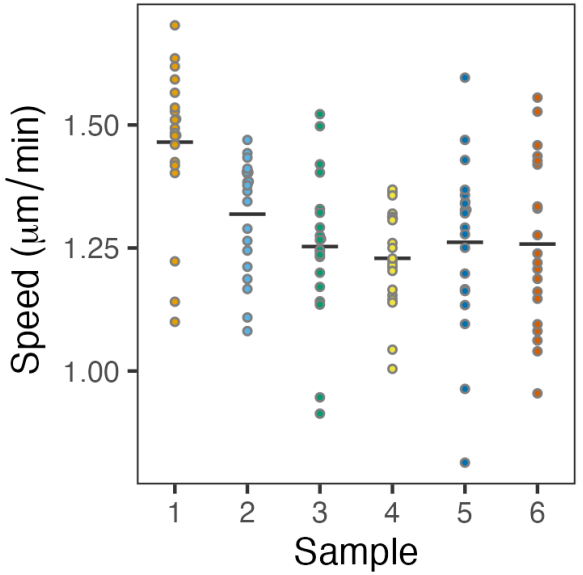
Bad



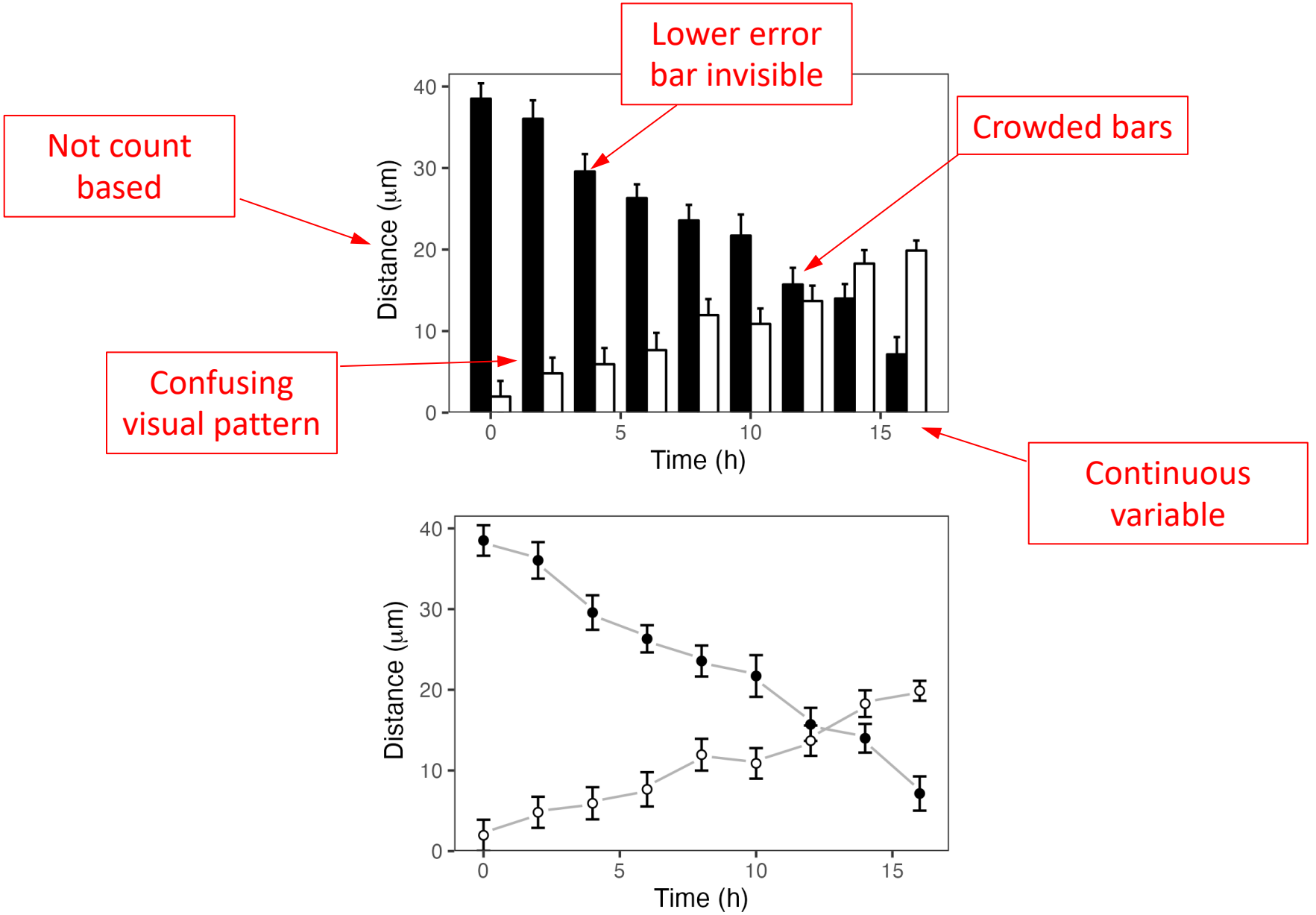

Better



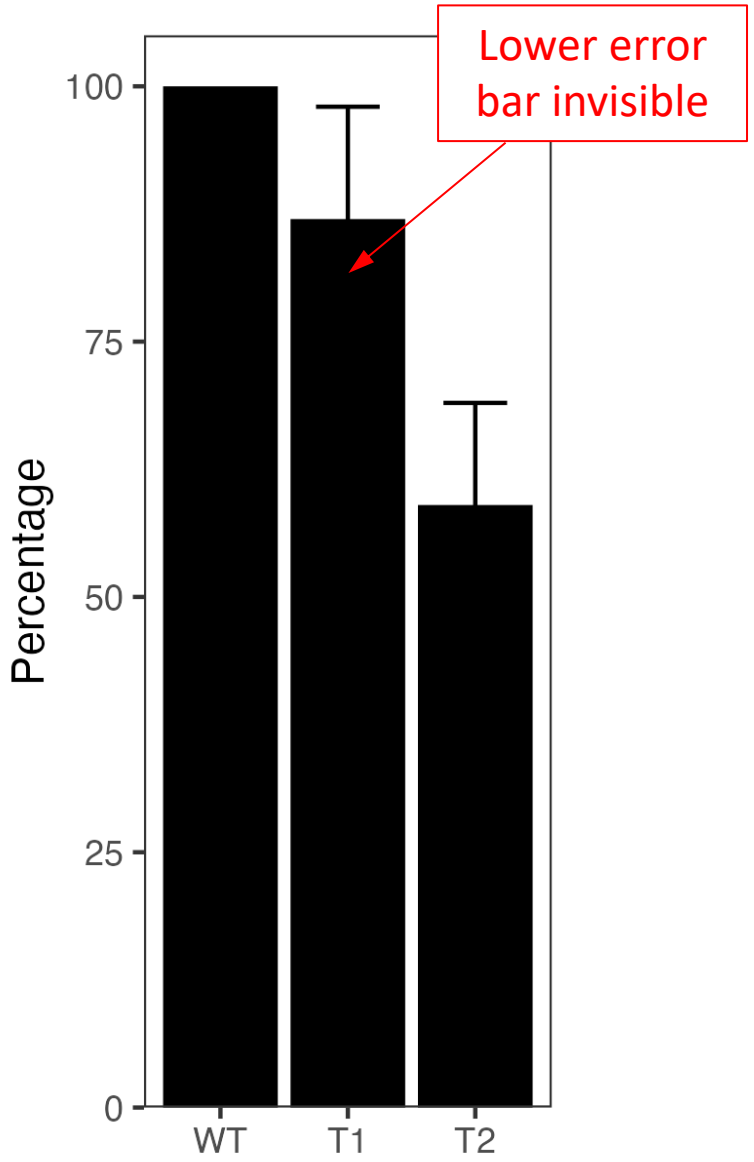

Best



# Bar plot problems



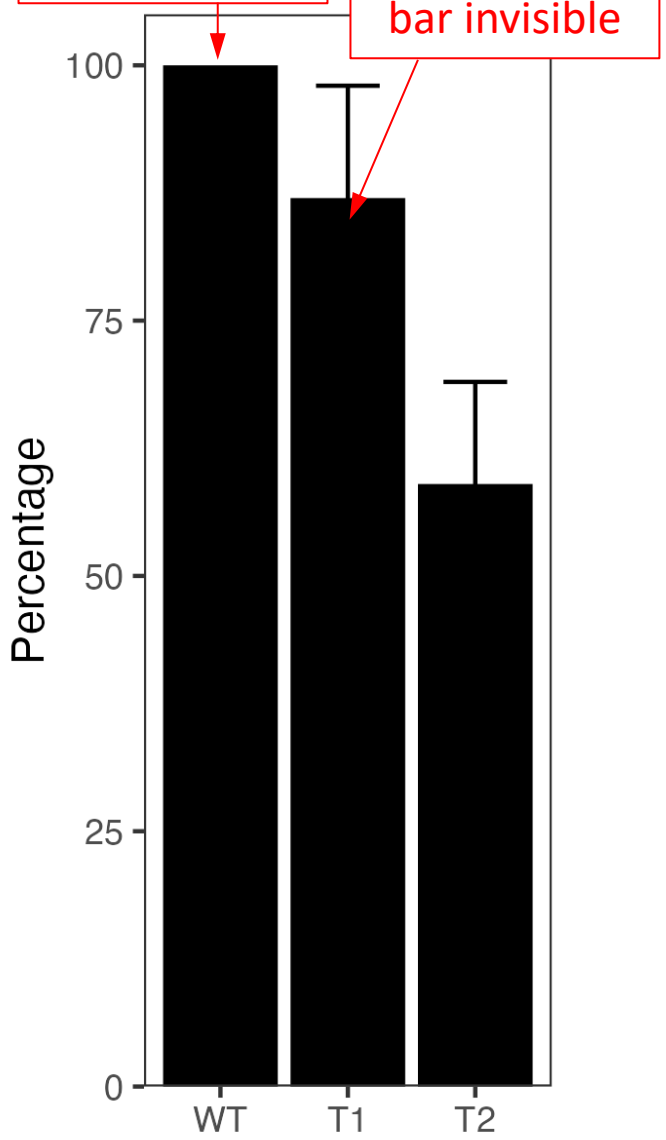
# Bar plots with error bars



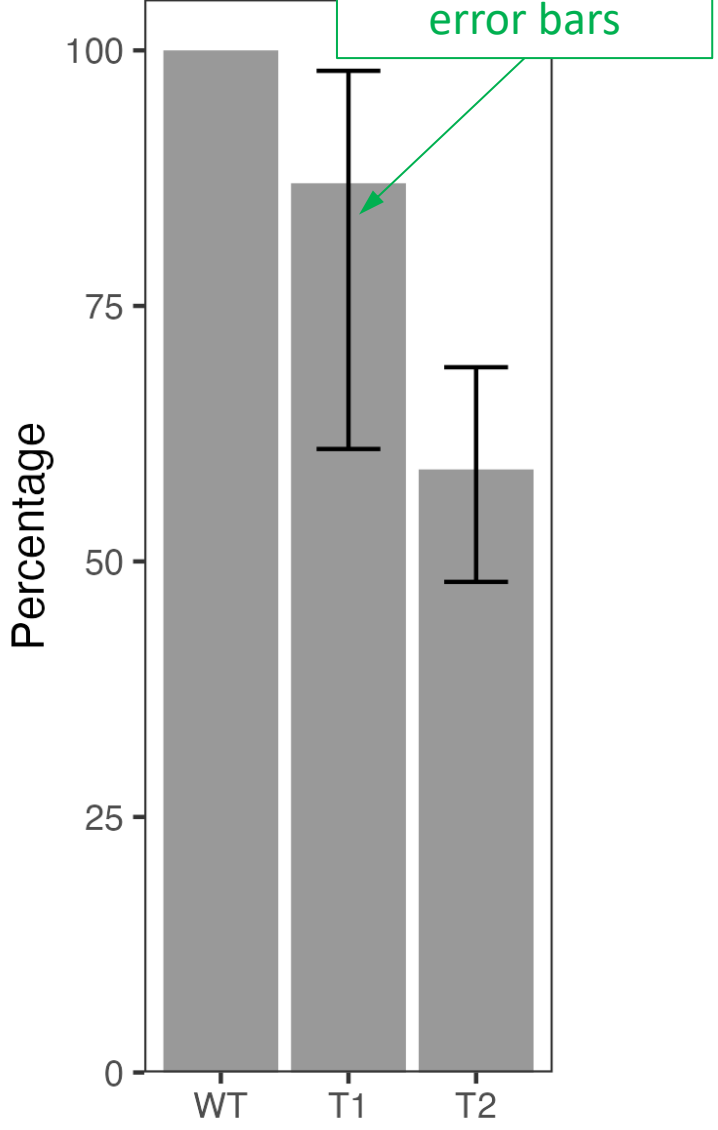

# Bar plots with error bars

Redundant information

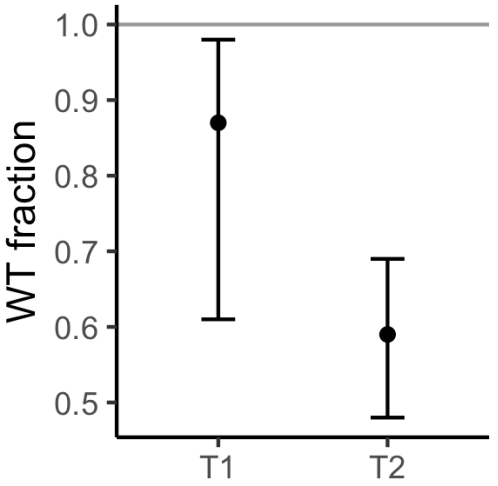
Lower error bar invisible




Always show upper and lower error bars




Other types of plots might be better

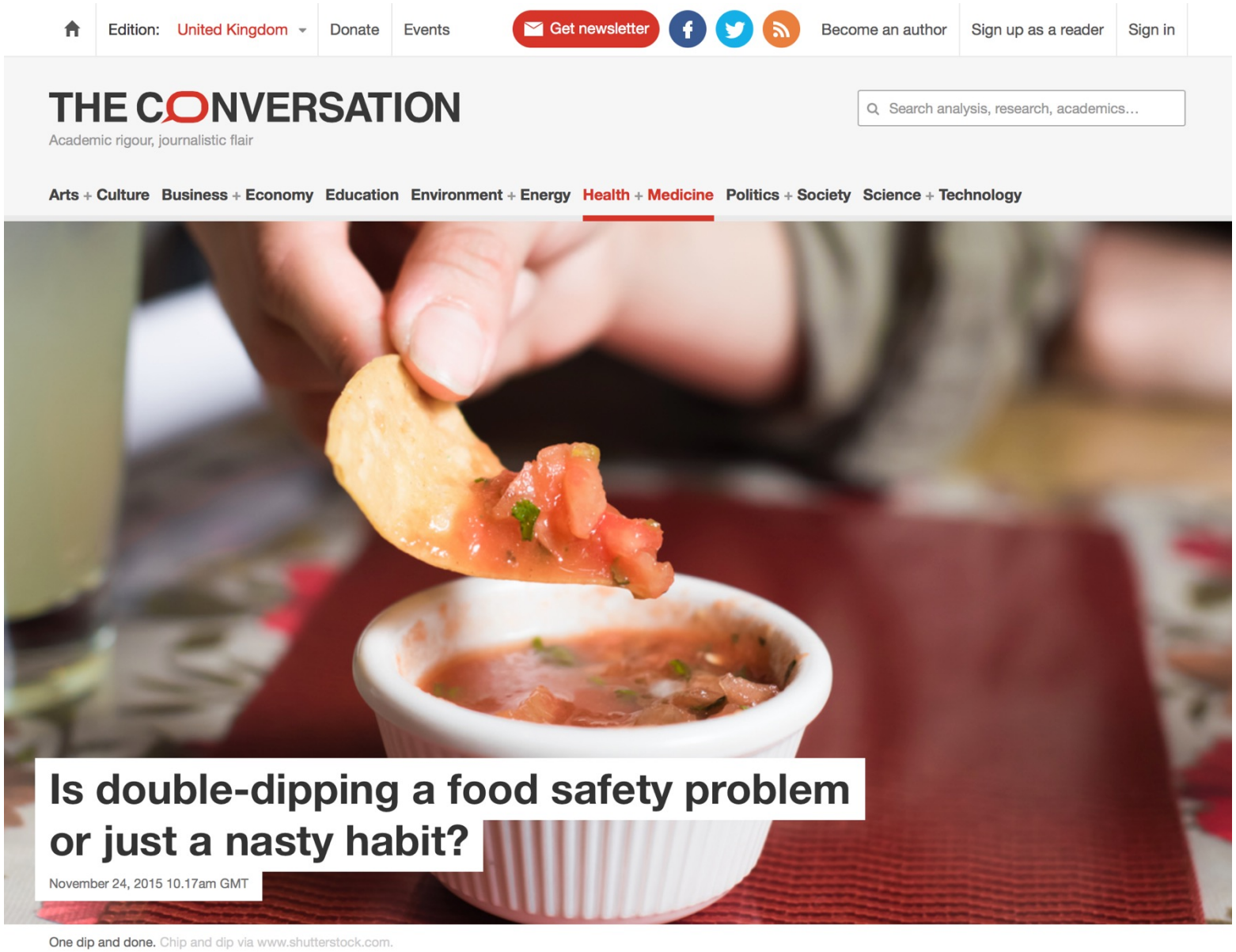



$T1 = 0.87^{+0.11}_{-0.26}$

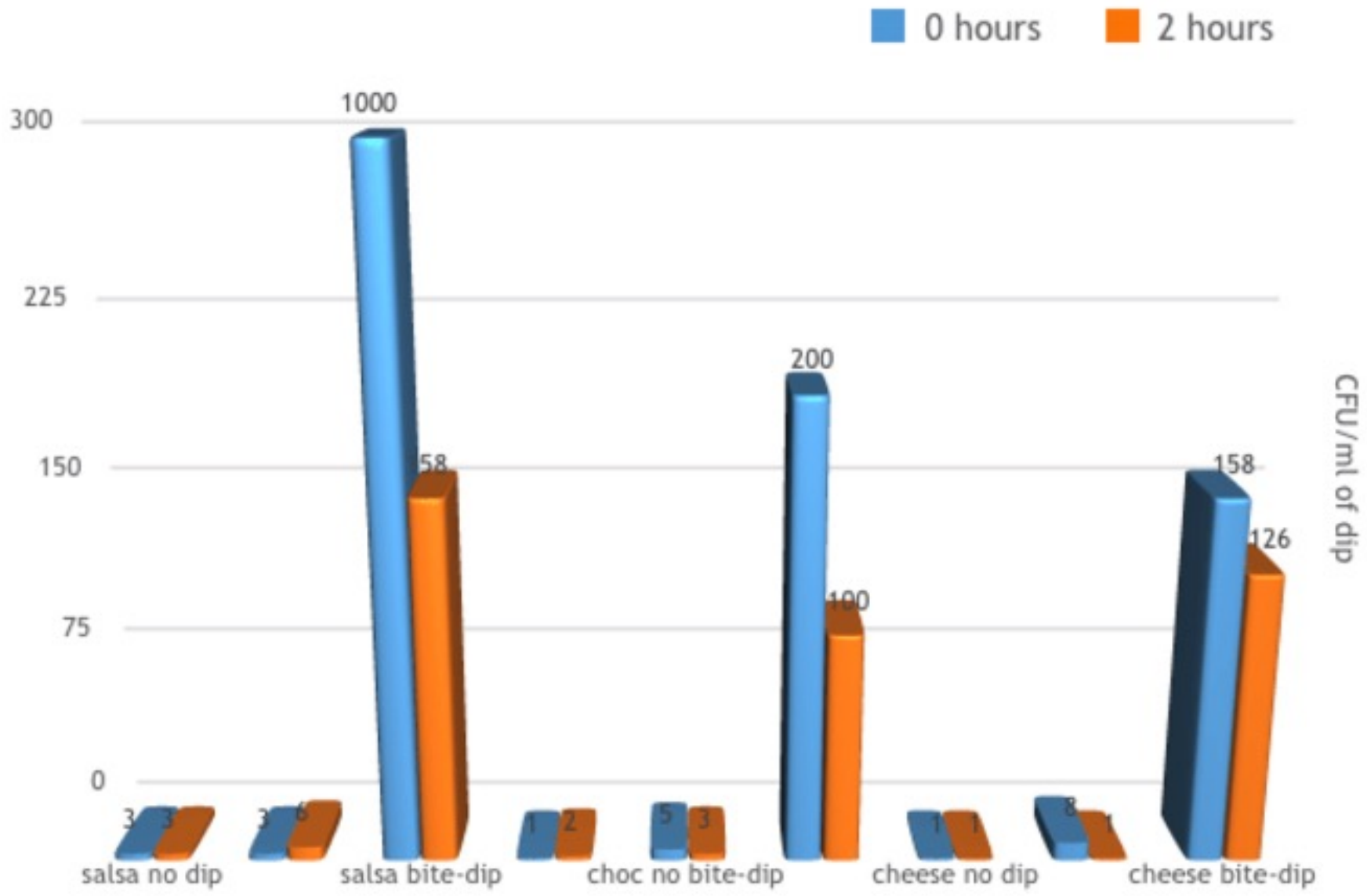
$T2 = 0.59^{+0.10}_{-0.11}$

Do you really need a plot?

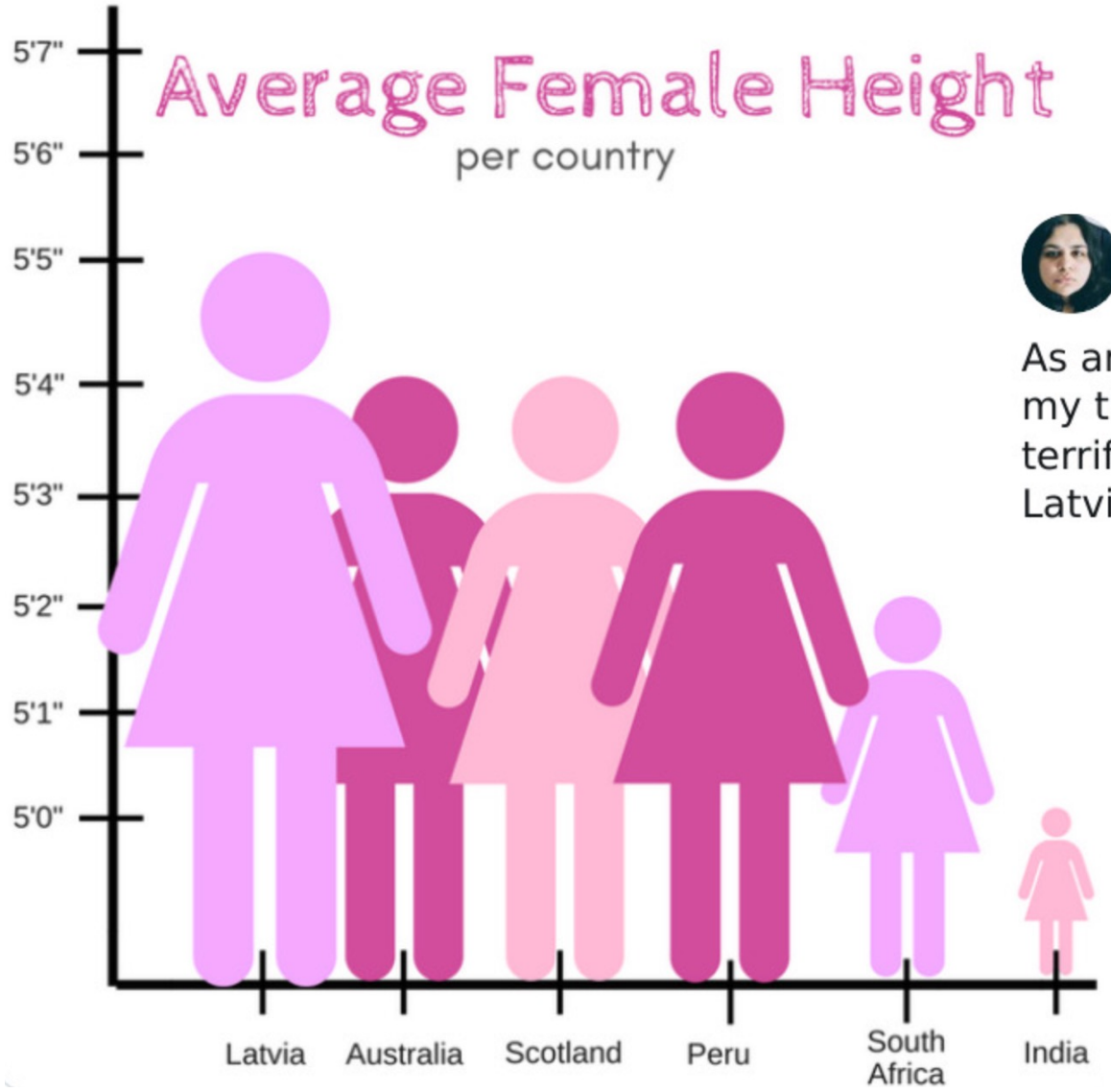
# The worst bar plot in the world?



# The worst bar plot in the world?



# Average female height




**Sabah Ibrahim**
@reina_sabah

As an Indian woman, I can confirm that too much of my time is spent hiding behind a rock praying the terrifying gang of international giant ladies and their Latvian general don't find me

# Pie charts

**NO**

# Summary of plots

# Ten commandments of making good plots

1. Thou shalt always prioritize clarity in thy presentation.
2. Thou shalt adorn thy axes with both scales and labels.
3. Thou shalt employ logarithmic scales when data spans vast orders of magnitude.
4. Thou shalt ensure that all labels and numbers are legible to all.
5. Thou shalt choose symbols that stand distinct from one another.
6. Thou shalt select colours discernible even to those with colour-blindness.
7. Thou shalt bestow error bars upon thy plots where possible.
8. Thou shalt always proclaim the nature of uncertainty depicted by thy error bars.
9. Thou shalt incorporate model lines when deemed fitting.
10. Thou shalt connect data points with guiding lines if it brings clarity.
11. Thou shalt refrain from using bar plots, unless truly warranted.

# Bar plots: recommendations

1. Bar plots should only be used to present count-based quantities: counts, proportions and probabilities
2. Often it is better 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. Thou shalt not make dynamite plots. Ever

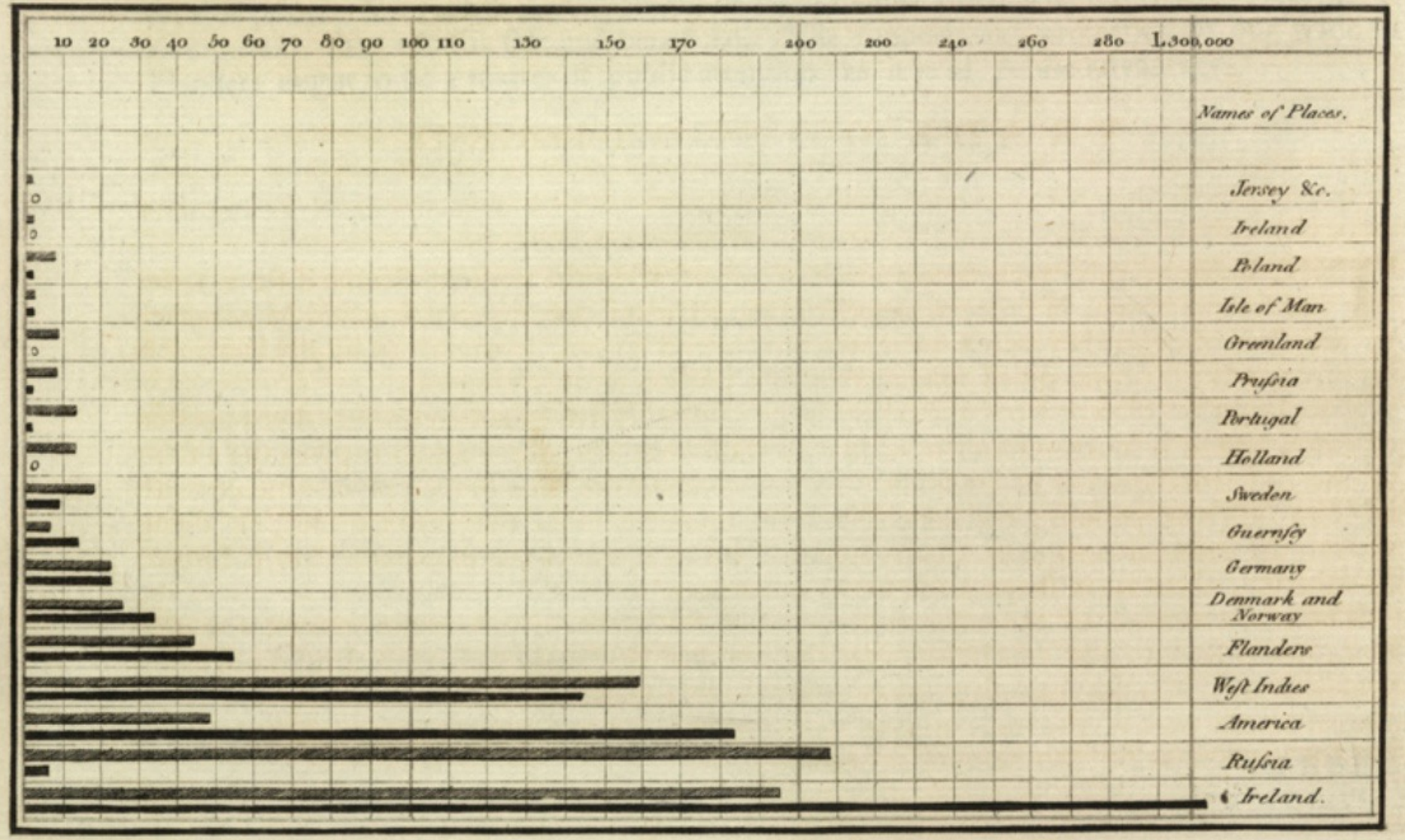
# William Playfair

- Born in Liff near Dundee
- Man of many careers (millwright, engineer, draftsman, accountant, inventor, silversmith, merchant, secret agent, 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)

Exports and Imports of SCOTLAND to and from different parts for one Year from Chriſtmas 1780 to Chriſtmas 1781.




The Upright diviſions are Ten Thouſand Pounds each. The Black Lines are Exports the Ribbed lines Imports.

# 6. Quoting numbers and errors

“23.230814584530889987945556640625%”

*Anonymous*

# What is used to quantify errors

- In a publication you typically quote:

$$x = x_{\text{best}} \pm \Delta x$$

best estimate — error

- 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

# 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?

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

- 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 $\text{min}^{-1}$

# How to communicate significant figures

- 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 |
| **4**000 | 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 |

# How to remove non-significant figures

- 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 $\rightarrow$ 1.3)
  - 5-9: round up (**1.3**56 $\rightarrow$ 1.4)
- So, how many figures are significant?

Suppose we have 2 s.f. in each number

| Raw number | Quote |
|---|---|
| **12**34 | 1200 |
| **12**87 | 1300 |
| **1.4**91123 | 1.5 |
| **1.4**49999 | 1.4 |

# How to find which figures are significant

- Look at the error of the number, e.g., $SE$

- Error in the error is

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

- This formula can be applied to $SD$ and $CI$

**Example**

$n = 12$

$SE = 23.17345$

$$\Delta SE = \frac{23.17345}{\sqrt{2 \times 11}} \approx \frac{23.17}{4.69} \approx 5$$

$SE = 23 \pm 5$

- We can trust only one figure in the error
- Round $SE$ to one s.f.:

$SE = 20$

# Error in the error

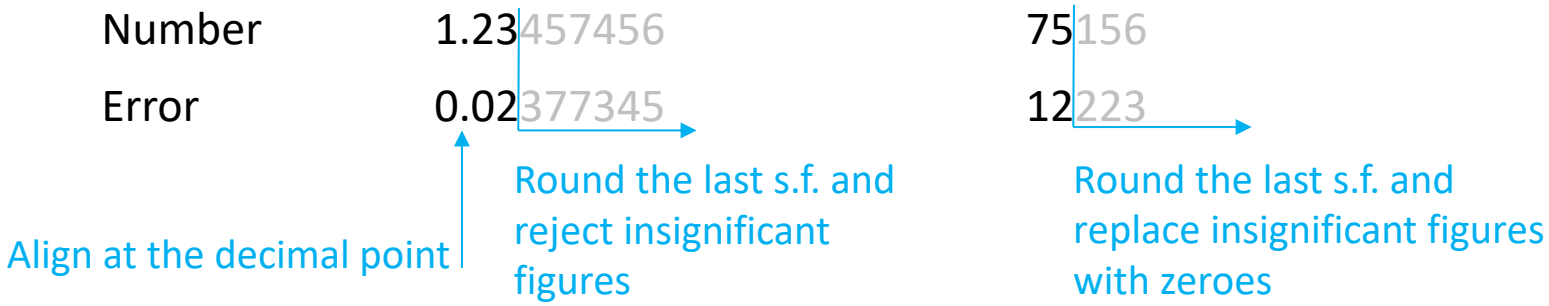
| $n$ | $\frac{\Delta SE}{SE}$ | s.f. to quote |
|---|---|---|
| 10 | 0.24 | 1 |
| 100 | 0.07 | 2 |
| 1,000 | 0.02 | 2 |
| 10,000 | 0.007 | 3 |
| 100,000 | 0.002 | 3 |

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

# How to quote number and error

# 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




| Correct | Incorrect |
|---|---|
| $1.23 \pm 0.02$ | $1.2 \pm 0.02$ |
| $1.2 \pm 0.5$ | $1.23423 \pm 0.5$ |
| $6.0 \pm 3.0$ | $6 \pm 3.0$ |
| $75000 \pm 12000$ | $75156 \pm 12223$ |
| $(3.5 \pm 0.3)\times10^{-5}$ | $3.5 \pm 0.3\times10^{-5}$ |

# Number 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

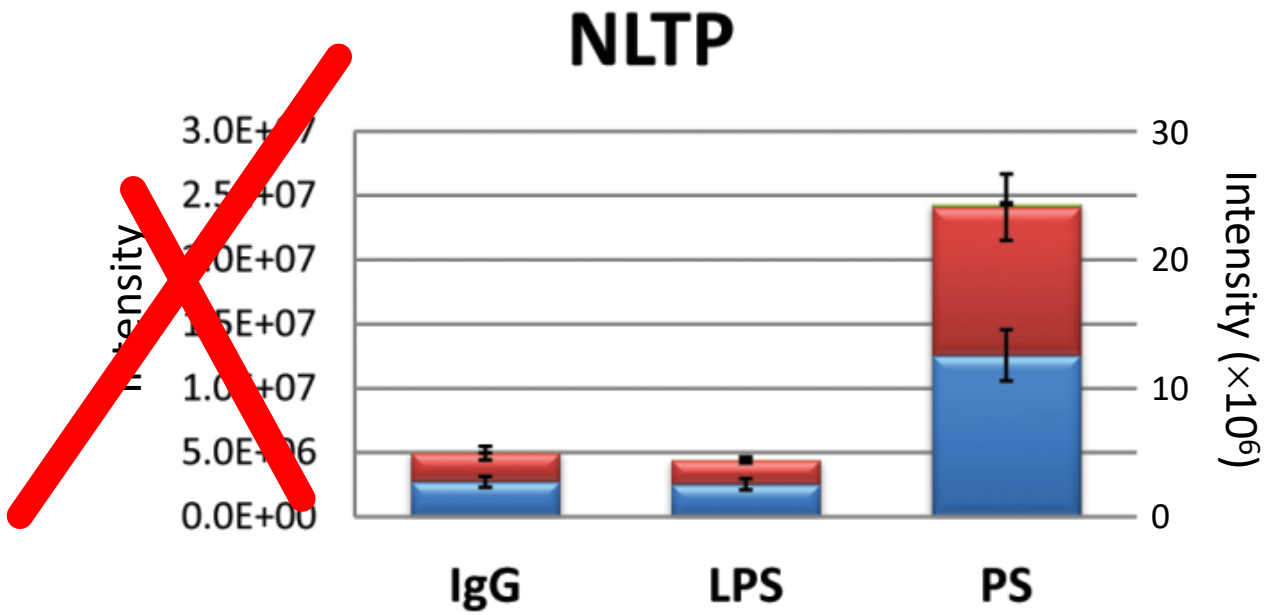
# Avoid computer notation

- Example from a random paper off my shelf

  “p-value = 5.51E-14”

- I’d rather put it down as

  p-value = $6\times10^{-14}$



# 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 | $1.5\times10^4$ |
| 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

# P-values

- It is not easy to estimate uncertainty of a p-value
- Typically, p-values are uncertain

- Quote 1 or 2 digits
  - 0.01
  - 0.5
  - 0.043
  - $2 \times 10^{-8}$

# 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**
  - $1.23 \pm 0.02$
  - $1.23423 \pm 0.00005$ (rather unlikely in biological experiments)
  - $6 \pm 3$
  - $75 \pm 12$
  - $(3.2 \pm 0.3) \times 10^{-5}$

## WHEN YOU DON'T KNOW ERROR

- You still need to guesstimate your error!
- Quote only figures that are significant
- P-values have 1 or 2 s.f.: $p = 0.03$, 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 $\rightarrow$ 6.6)
5-9: up (6.65 $\rightarrow$ 6.7)

# Don’t use Excel

CSV file

```
gene_symbol,sample_1,sample_2,sample_3
CDK1,1203,234,2134
MARCH1,2987,4234,12334
SEPT1,0,0,2
```

Open in Excel

| | A | B | C | D |
|---|---|---|---|---|
| 1 | gene_symbol | sample_1 | sample_2 | sample_3 |
| 2 | CDK1 | 1203 | 234 | 2134 |
| 3 | Mar-01 | 2987 | 4234 | 12334 |
| 4 | Sep-01 | 0 | 0 | 2 |
| 5 | | | | |

~30% of published papers contain mangled gene names in supplementary data

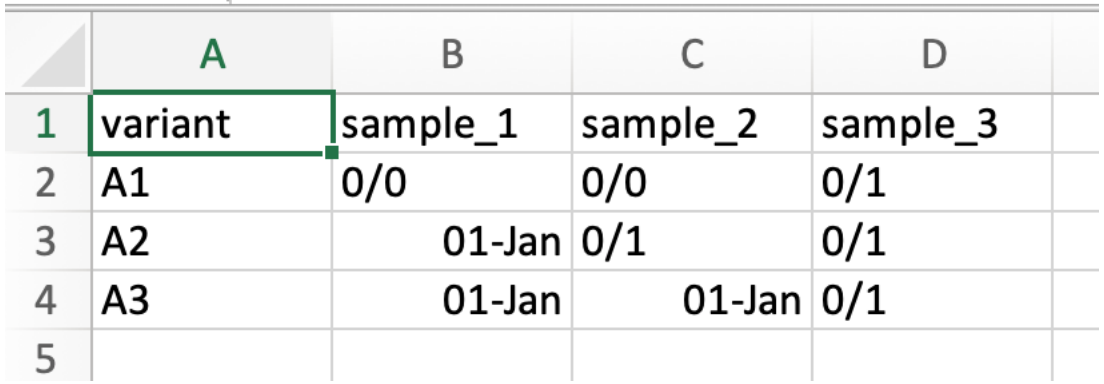
HUGO Gene Nomenclature Committee changed 27 gene symbols, e.g.
MARCH1 ⟶ MARCHF1
SEPT1 ⟶ SEPTIN1

# Don’t use Excel

CSV file

```
variant,sample_1,sample_2,sample_3
A1,0/0,0/0,0/1
A2,1/1,0/1,0/1
A3,1/1,1/1,0/1
```

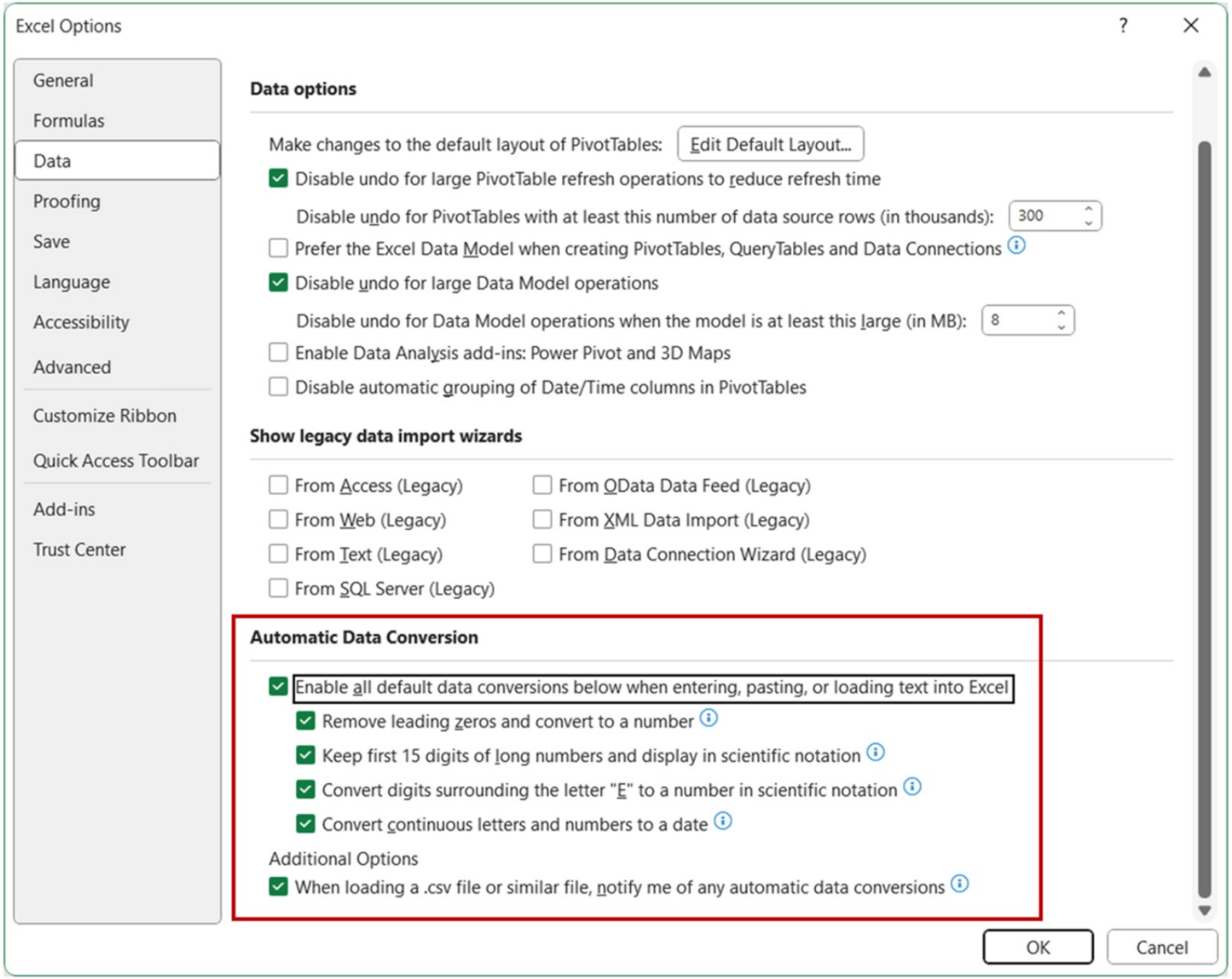
Open in Excel



| | A | B | C | D |
|---|---|---|---|---|
| 1 | variant | sample_1 | sample_2 | sample_3 |
| 2 | A1 | 0/0 | 0/0 | 0/1 |
| 3 | A2 | 01-Jan | 0/1 | 0/1 |
| 4 | A3 | 01-Jan | 01-Jan | 0/1 |
| 5 | | | | |

“Excel is where data go to die”

19 October 2023



Slides available at
https://dag.compbio.dundee.ac.uk/training/Statistics_lectures.html