5. Data presentation

“Above all else show the data”

Edward R. Tufte
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 \( \tau = 17 \pm 1 \text{ h} \) (95% confidence intervals).
3 rules for making good plots

1. Clarity of presentation

2. Clarity of presentation

3. Clarity of presentation
Show your data!

- **Awful!**
- **OK**
- **Better**
- **Best!**

- **Dynamite plot**
- **Box plot**
- **Jitter plot**
- **Beeswarm plot**
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)
- Clarity!
- Symbols shall be easy to distinguish
- It is OK to join data points with lines for guidance
Labels!
Logarithmic plots
Logarithmic plots
How to plot error bars

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

$M \pm SD$
$M \pm SE$

$[M_L, M_U]$

$M$ + upper error
$M$ - lower error

upper error
lower error
How to plot error bars

- Clarity!
- Make sure error bars are visible
### Types of errors

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

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

- Central 50% of data
- Inter-quartile range (IQR)
- 25\textsuperscript{th} percentile
- Median = 50\textsuperscript{th} percentile
- 75\textsuperscript{th} percentile
- "Outliers"

Value axis

≤ 1.5 IQR

≤ 1.5 IQR
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: categorical variable

Proportion or count

Probability distribution

Proportion of cells

Additive area

Treatment

DMSO

DD342

Category
Bar plots: continuous variable

count = 137
density = 2.28

density = \frac{\text{count}}{\text{total count} \ast \text{bin width}}

Total area = 1
Bar plots start at zero

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

Little variability

**Better**

**Best**

Not count based
Bar plot problems

- Not count based
- Confusing visual pattern
- Lower error bar invisible
- Crowded bars
- Continuous variable
Bar plots with error bars

Lower error bar invisible

Every time you make a dynamite plot

a puppy dies
Bar plots with error bars

- Always show upper and lower error bars
- Other types of plots might be better

Useless information
- Lower error bar invisible

Do you really need a plot?

T1 = 0.87^{+0.11}_{-0.26}
T2 = 0.59^{+0.10}_{-0.11}
Is double-dipping a food safety problem or just a nasty habit?

November 24, 2015 10.17am GMT

The worst bar plot in the world?
Pie charts

NO
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 where 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 count-based 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, 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)
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:

\[ x = x_{\text{best}} \pm \Delta x \]

- 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?

\[
\frac{4.1 \text{ µm}}{2.6 \text{ min}} = 1.576923077 \text{ µm min}^{-1}
\]

- 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

<table>
<thead>
<tr>
<th>Number</th>
<th>Significant figures</th>
</tr>
</thead>
<tbody>
<tr>
<td>365</td>
<td>3</td>
</tr>
<tr>
<td>1.893</td>
<td>4</td>
</tr>
<tr>
<td>4000</td>
<td>1 or 4</td>
</tr>
<tr>
<td>4\times10^3</td>
<td>1</td>
</tr>
<tr>
<td>4.000\times10^3</td>
<td>4</td>
</tr>
<tr>
<td>4000.00</td>
<td>6</td>
</tr>
<tr>
<td>0.00034</td>
<td>2</td>
</tr>
<tr>
<td>0.0003400</td>
<td>4</td>
</tr>
</tbody>
</table>
Rounding

- Remove non-significant figures by rounding

- Round the last s.f. according to the value of the next digit
  - 0-4: round down (1.342 → 1.3)
  - 5-9: round up (1.356 → 1.4)

- So, how many figures are significant?

Suppose we have 2 s.f. in each number

<table>
<thead>
<tr>
<th>Raw number</th>
<th>Quote</th>
</tr>
</thead>
<tbody>
<tr>
<td>1234</td>
<td>1200</td>
</tr>
<tr>
<td>1287</td>
<td>1300</td>
</tr>
<tr>
<td>1.491123</td>
<td>1.5</td>
</tr>
<tr>
<td>1.449999</td>
<td>1.4</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\frac{\Delta SE}{SE}$</th>
<th>s.f. to quote</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.24</td>
<td>1</td>
</tr>
<tr>
<td>100</td>
<td>0.07</td>
<td>2</td>
</tr>
<tr>
<td>1,000</td>
<td>0.02</td>
<td>2</td>
</tr>
<tr>
<td>10,000</td>
<td>0.007</td>
<td>3</td>
</tr>
<tr>
<td>100,000</td>
<td>0.002</td>
<td>3</td>
</tr>
</tbody>
</table>

- 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

<table>
<thead>
<tr>
<th>Number</th>
<th>Error</th>
<th>Align at the decimal point</th>
<th>Round the last s.f. and reject insignificant figures</th>
<th>Round the last s.f. and replace insignificant figures with zeroes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.23457456</td>
<td>0.02377345</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>75156</td>
<td>12223</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Correct</th>
<th>Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.23 ± 0.02</td>
<td>1.2 ± 0.02</td>
</tr>
<tr>
<td>1.2 ± 0.5</td>
<td>1.23423 ± 0.5</td>
</tr>
<tr>
<td>6.0 ± 3.0</td>
<td>6 ± 3.0</td>
</tr>
<tr>
<td>75000 ± 12000</td>
<td>75156 ± 12223</td>
</tr>
<tr>
<td>(3.5 ± 0.3)×10^{-5}</td>
<td>3.5 ± 0.3×10^{-5}</td>
</tr>
</tbody>
</table>
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
  \[ p\text{-value} = 6 \times 10^{-14} \]
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.

<table>
<thead>
<tr>
<th>raw data</th>
<th>1 decimal place</th>
</tr>
</thead>
<tbody>
<tr>
<td>14524.21</td>
<td>14524.2</td>
</tr>
<tr>
<td>2234.242</td>
<td>2234.2</td>
</tr>
<tr>
<td>122.1948</td>
<td>122.2</td>
</tr>
<tr>
<td>12.60092</td>
<td>12.6</td>
</tr>
<tr>
<td>2.218293</td>
<td>2.2</td>
</tr>
<tr>
<td>0.120024</td>
<td>0.1</td>
</tr>
<tr>
<td>0.021746</td>
<td>0.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Wrong</th>
<th>Right</th>
</tr>
</thead>
<tbody>
<tr>
<td>14524.2</td>
<td>1.5×10^4</td>
</tr>
<tr>
<td>2234.2</td>
<td>2200</td>
</tr>
<tr>
<td>122.2</td>
<td>120</td>
</tr>
<tr>
<td>12.6</td>
<td>13</td>
</tr>
<tr>
<td>2.2</td>
<td>2.2</td>
</tr>
<tr>
<td>0.1</td>
<td>0.12</td>
</tr>
<tr>
<td>0.0</td>
<td>0.022</td>
</tr>
</tbody>
</table>

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
  - $1.23 \pm 0.02$
  - $1.23423 \pm 0.00005$ (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, 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 ($6.64 \rightarrow 6.6$)
- 5-9: up ($6.65 \rightarrow 6.7$)
Hand-outs available at
https://dag.compbio.dundee.ac.uk/training/Statistics_lectures.html