

Crash Course in Statistics

ZNZ 2026

III

Christoph Luchsinger and Zofia Baranczuk

Based on Script by Daniel J. Stekhoven

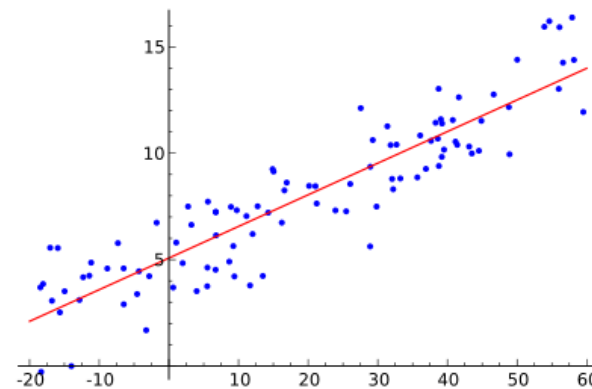
Descriptive statistics and visualizations

- Describing and looking at data (Chapter 2 was models)
- Median versus mean, IQR versus variance, ...
- Don't use bar plots!
- Normal distribution extended
- Transformation of data
- Detection of influential observations

- Much more with Zofia/R

Describe data (goal: use very few parameters to explain a lot)

- Where is the data? (location)
 - mean or median
- How is the data shaped? (spread)
 - standard deviation
 - inter-quartile range
- Is the data linearly distributed (2 dim)?
 - correlation



The average - median versus mean



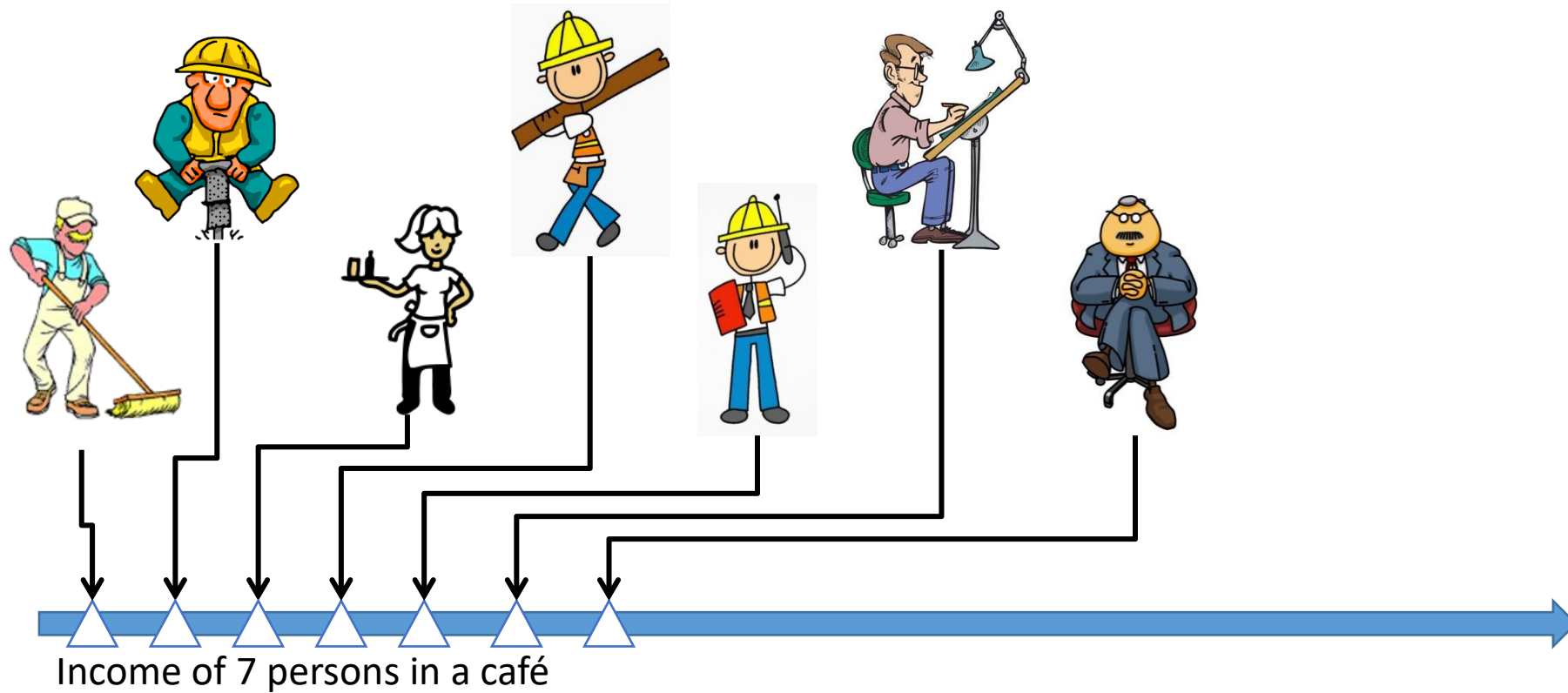
- The arithmetic mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

- The median (*non-parametric*)

$$\tilde{x} = q_{0.5}(x_1, \dots, x_n)$$

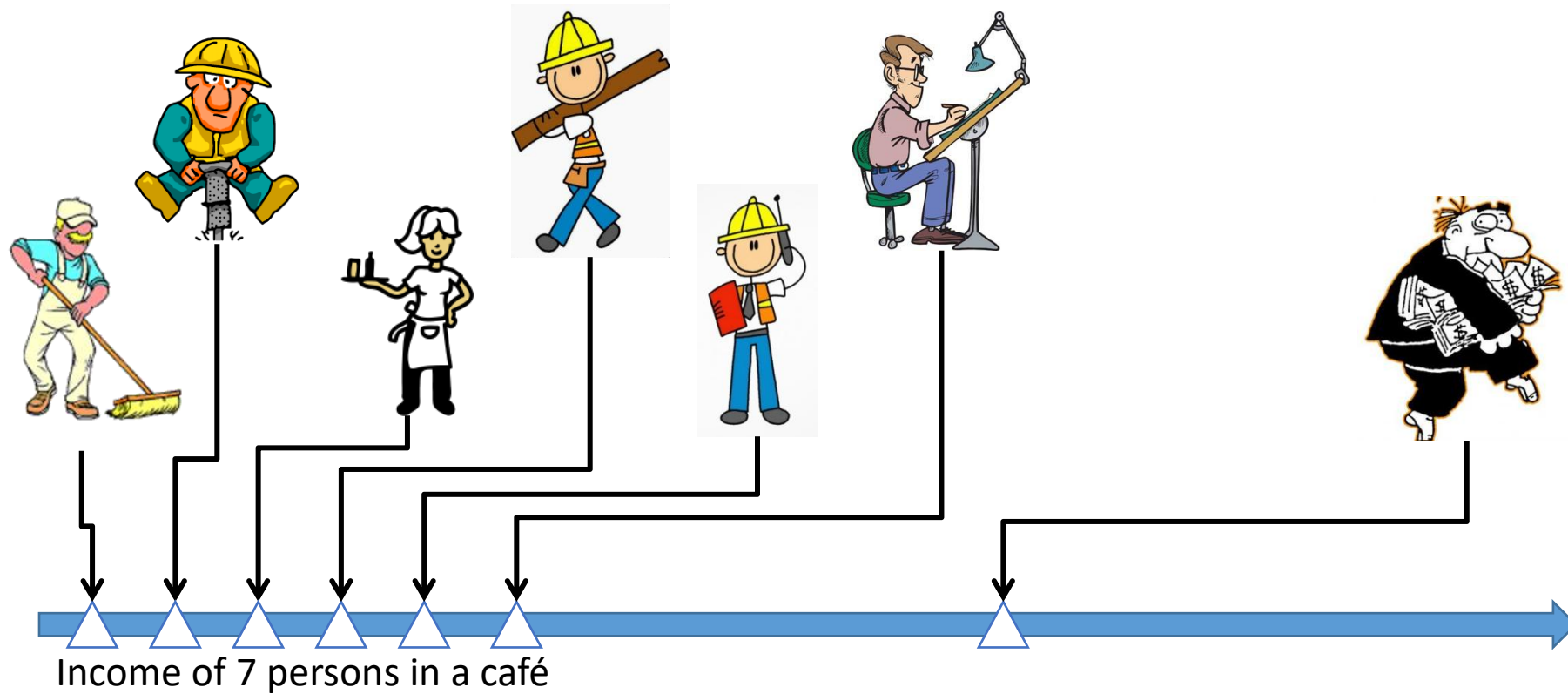
The average - median versus mean



★ mean

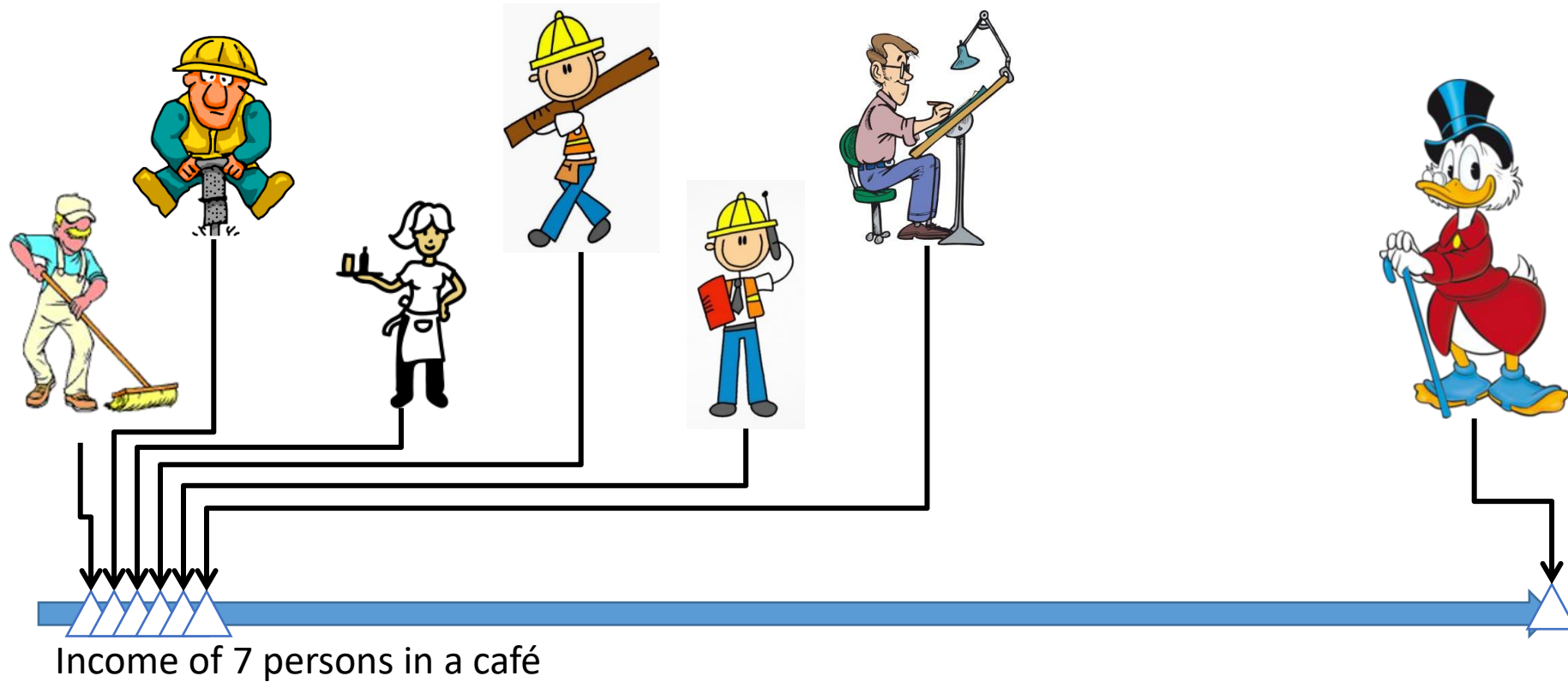
★ median

The average - median versus mean



★ mean
★ median

The average - median versus mean



★ median

★ mean

Robust for outliers!

The spread



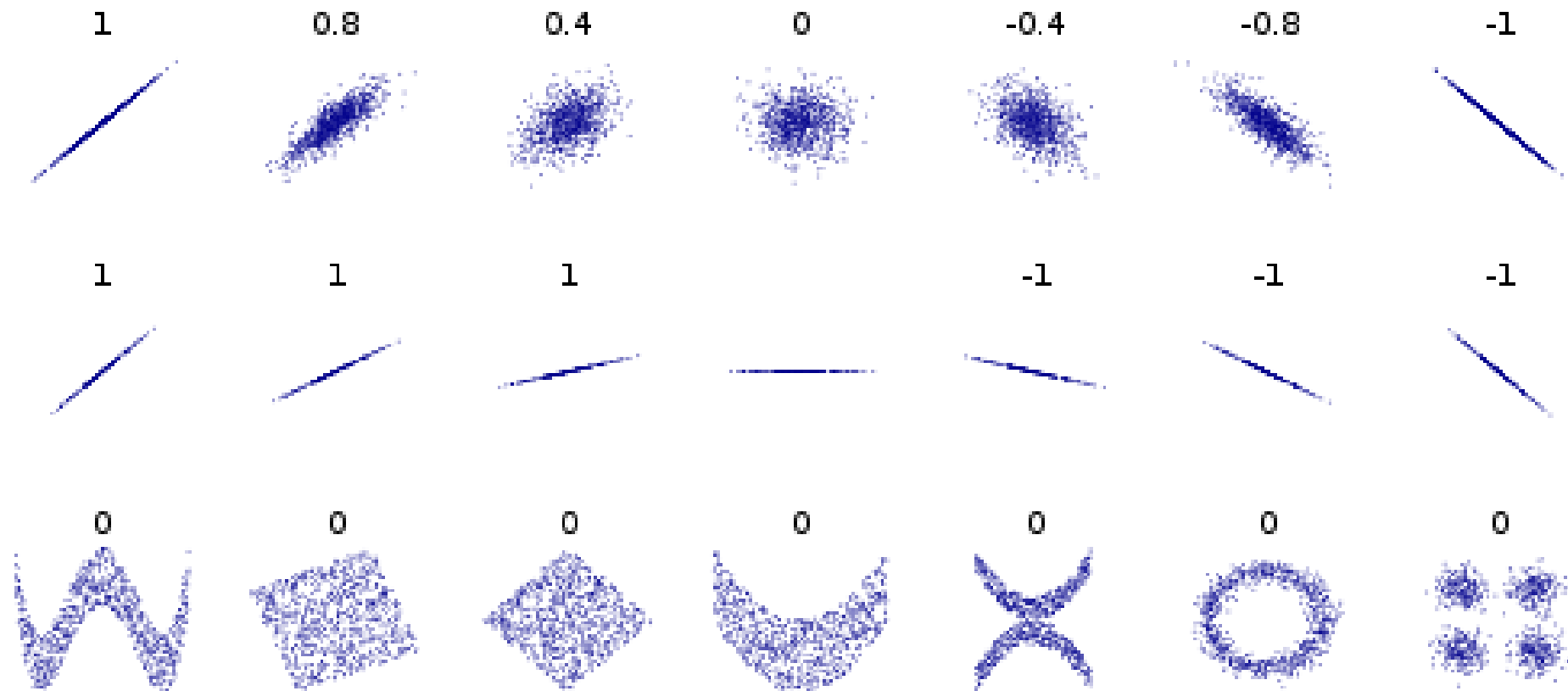
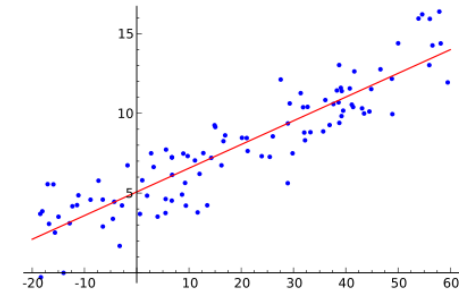
- The standard deviation

$$s_x = \sqrt{\text{Var}(\mathbf{x})} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

- The inter-quartile range (IQR; *non-parametric*)

$$IQR(\mathbf{x}) = q_{0.75}(x_1, \dots, x_n) - q_{0.25}(x_1, \dots, x_n)$$

Correlation

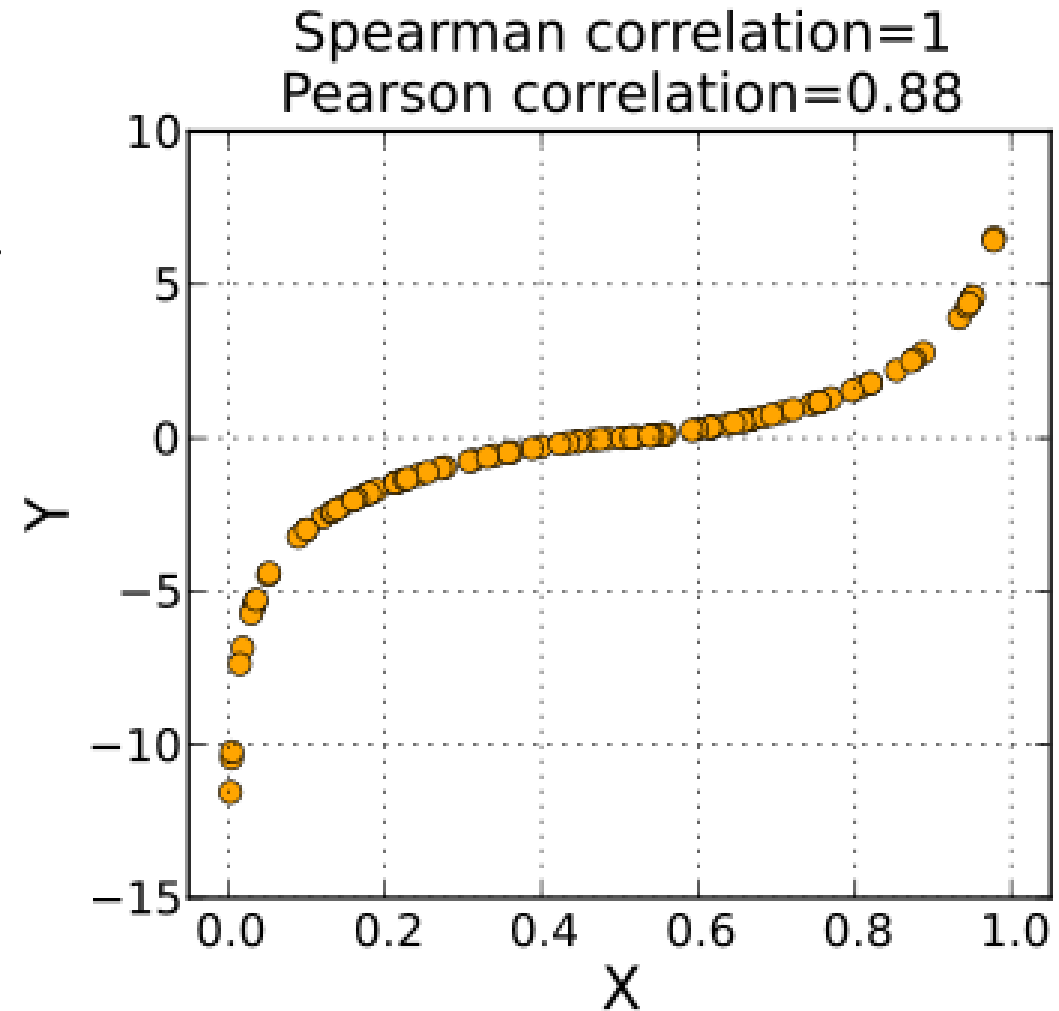


The coefficient of correlation

Yes, it can be understood:

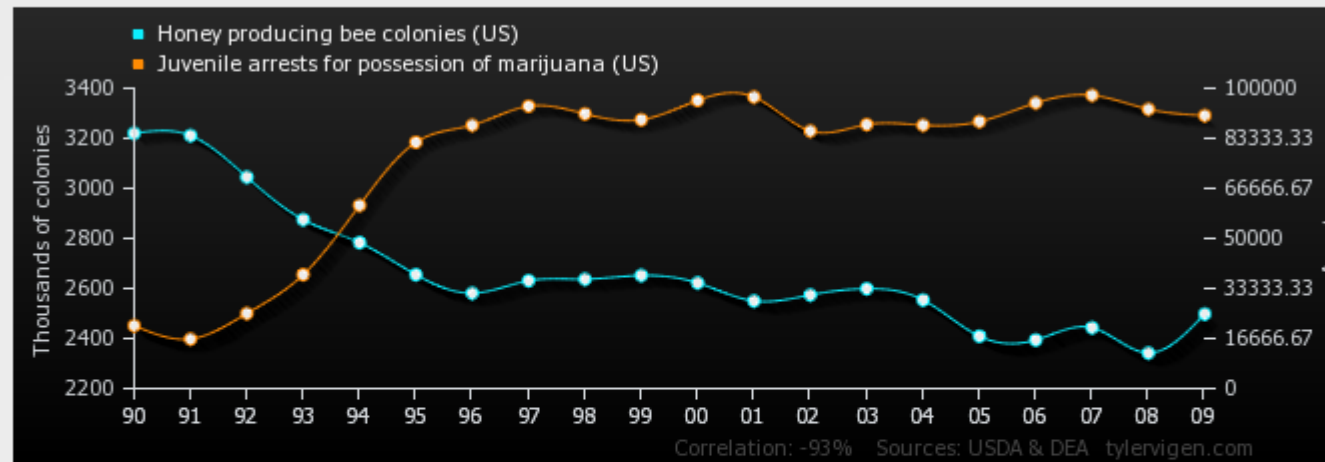
Correlation

- Pearson Correlation
 - Typically used correlation coefficient
 - Measures linear dependence (dt: Gleichläufigkeit)
 - Uses standard deviations
- Spearman's rank correlation
 - Non-parametric alternative to Pearson
 - Method of choice if the data is not *nice*



Correlation does not imply causation

Honey producing bee colonies (US) inversely correlates with Juvenile arrests for possession of marijuana (US)



Honey producing bee colonies (US)
Thousands of colonies (USDA)

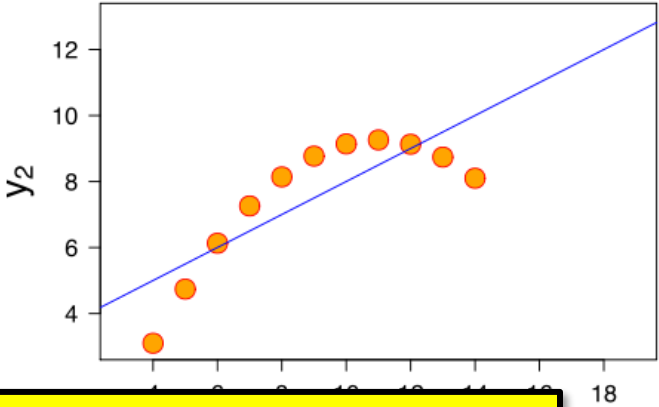
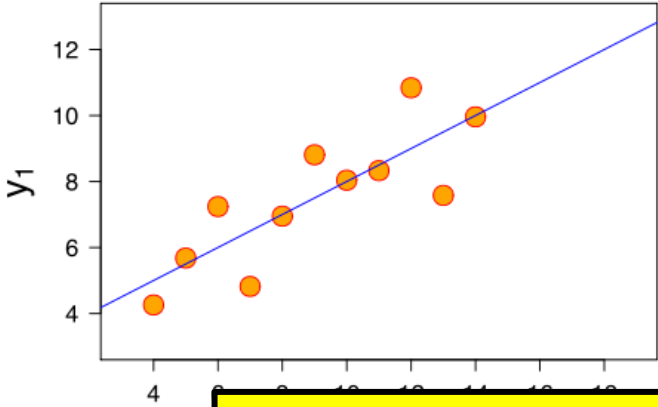
'90: 3,220; '91: 3,211; '92: 3,045; '93: 2,875; '94: 2,783; '95: 2,655; '96: 2,581; '97: 2,631; '98: 2,637; '99: 2,652; '00: 2,622; '01: 2,550; '02: 2,574; '03: 2,599; '04: 2,554; '05: 2,409; '06: 2,394; '07: 2,443; '08: 2,342; '09: 2,498

Juvenile arrests for possession of marijuana (US)
Arrests (DEA)

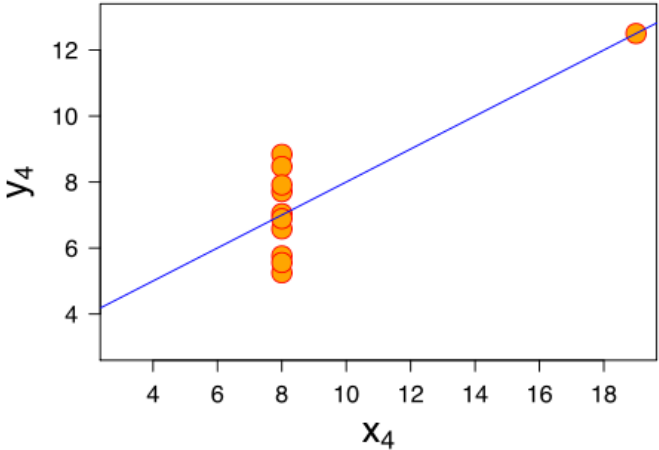
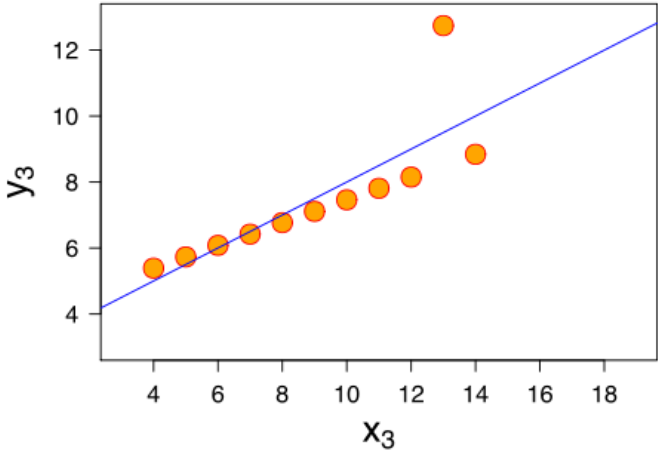
'90: 20,940; '91: 16,490; '92: 25,004; '93: 37,915; '94: 61,003; '95: 82,015; '96: 87,712; '97: 94,046; '98: 91,467; '99: 89,523; '00: 95,962; '01: 97,088; '02: 85,769; '03: 87,909; '04: 87,717; '05: 88,909; '06: 95,120; '07: 97,671; '08: 93,042; '09: 90,927

Correlation: -0.933389

Pearson correlation and linearity



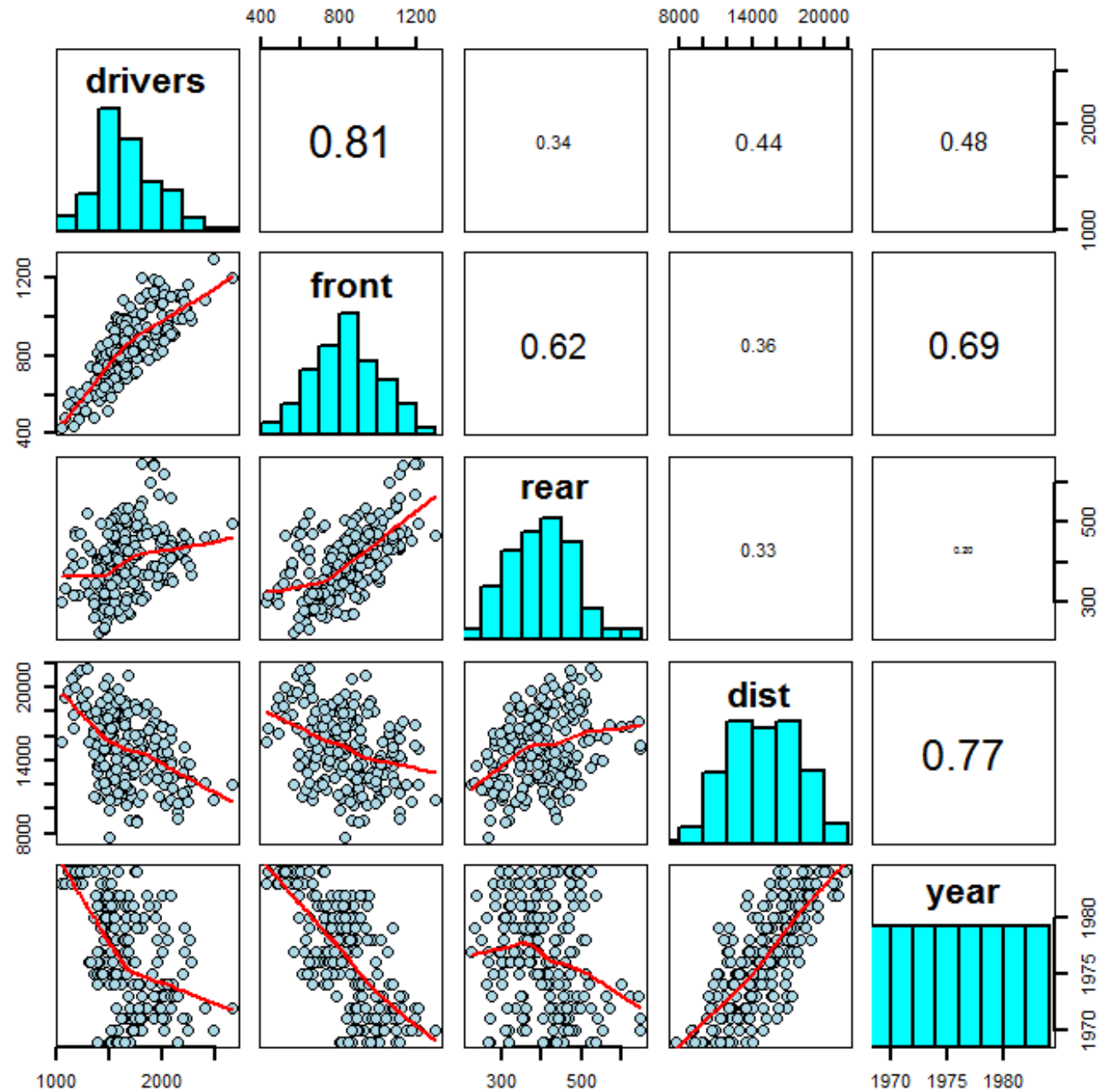
Always take a look at the data



	drivers	front	rear	dist	law	year	month
1	1687	867	269	9059	0	1969	1
2	1508	825	265	7685	0	1969	2
3	1507	806	319	9963	0	1969	3
4	1385	814	407	10955	0	1969	4
5	1632	991	454	11823	0	1969	5
6	1511	945	427	12391	0	1969	6
7	1559	1004	522	13460	0	1969	7
8	1630	1091	536	14055	0	1969	8
9	1579	958	405	12106	0	1969	9
10	1653	850	437	11372	0	1969	10
11	2152	1109	434	9834	0	1969	11
12	2148	1113	437	9267	0	1969	12
13	1752	925	316	9130	0	1970	1
14	1765	903	311	8933	0	1970	2
15	1717	1006	351	11000	0	1970	3
16	1558	892	362	10733	0	1970	4
17	1575	990	486	12912	0	1970	5
18	1520	866	429	12926	0	1970	6
19	1805	1095	551	13990	0	1970	7
20	1800	1204	646	14926	0	1970	8
21	1719	1029	456	12900	0	1970	9
22	2008	1147	475	12034	0	1970	10
23	2242	1171	456	10643	0	1970	11
24	2478	1299	468	10742	0	1970	12
25	2030	944	356	10266	0	1971	1
26	1655	874	271	10281	0	1971	2

Pattern?

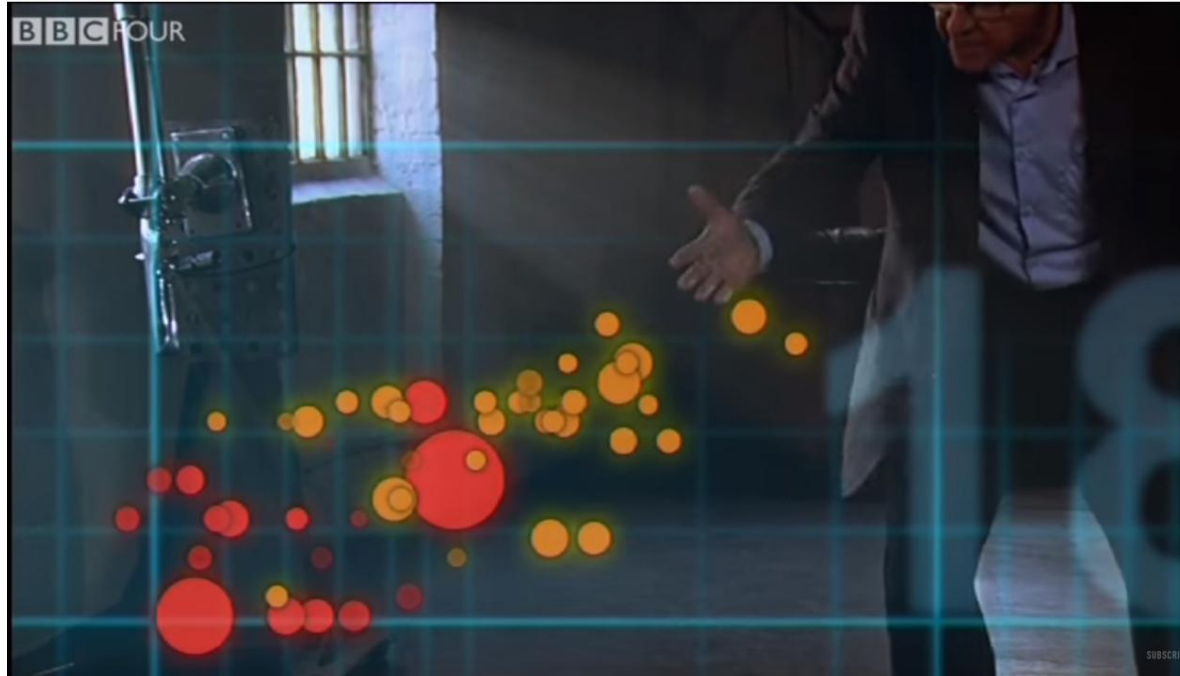
Relation?



Pattern!
Relation!

GapMinder.org

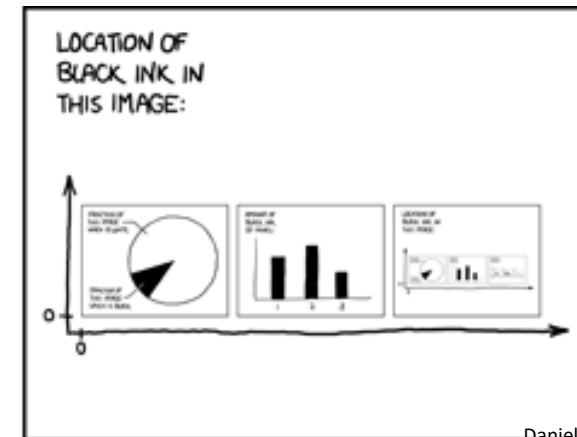
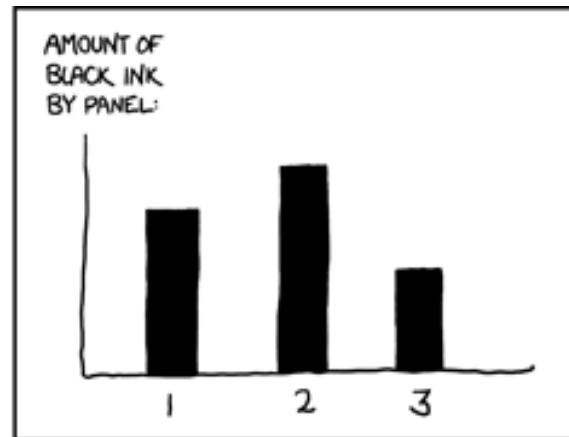
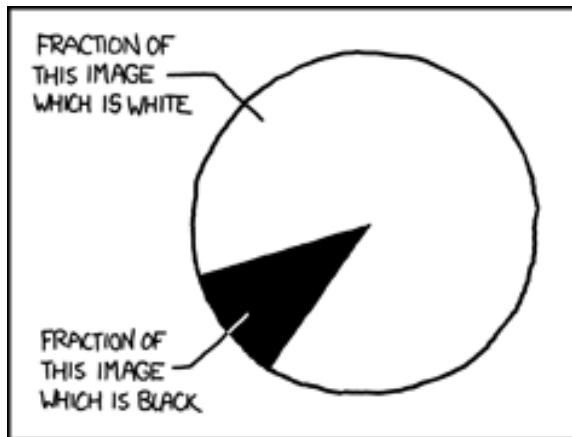
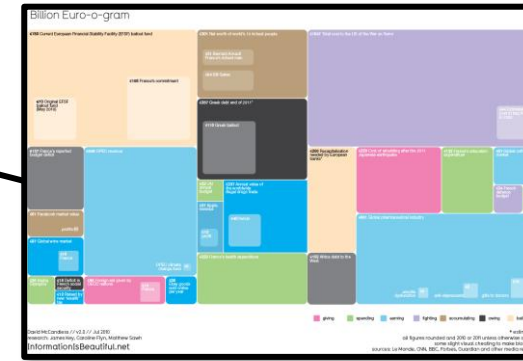
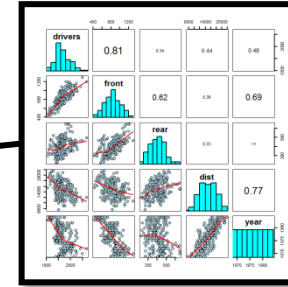
- How to win actual insights from massive amounts of data



- Look at it: <https://www.youtube.com/watch?v=jbkSRLYSojo>
- ...or better try it yourself: <http://www.gapminder.org/tools/bubbles>

Types of Visualization

- Three types...
 - Exploring data
 - Explaining data
 - Visual Art



Common Visualizations

Barplot

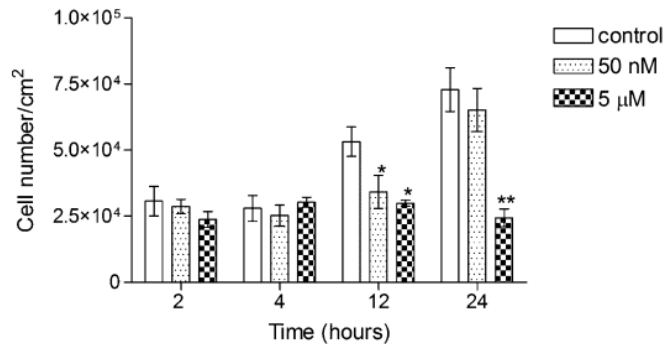
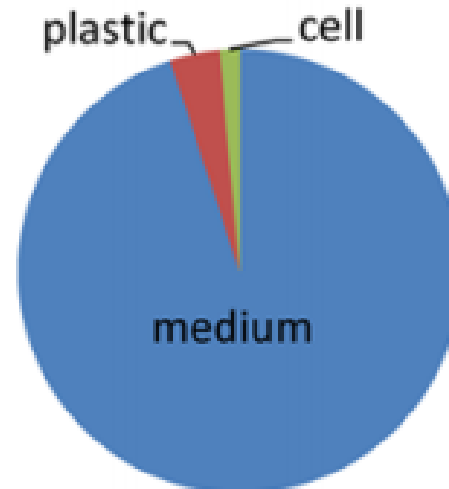


Figure 2. Impact of BaP on Hepa1c1c7 cell proliferation. Cell number was determined over time after exposure to 50 nM BaP, 5 μM BaP, or DMSO as solvent control. All experiments were performed in biological triplicates, and average and standard deviations are shown. Asterisks indicate significant difference as compared to the control column of each time point as determined by one-way ANOVA followed by Dunnett's multiple comparison test (**p*-value <0.05; ***p*-value <0.01).

Madureira et al., 2014,
Chem. Res. Toxicol., 27, 443–453

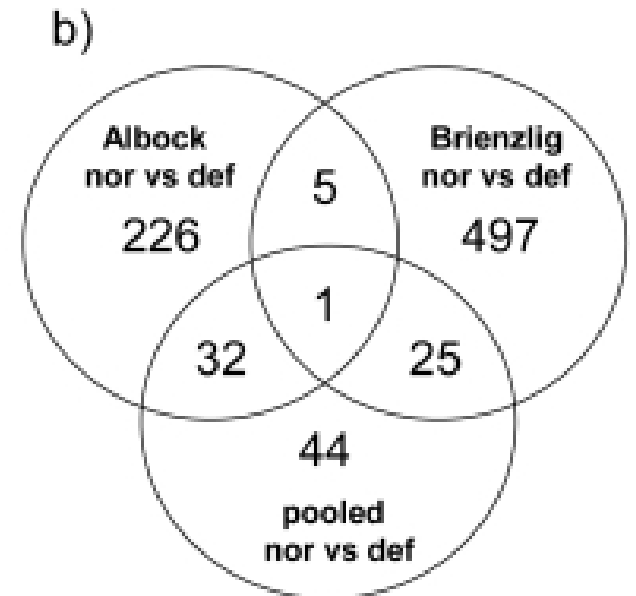
Pie Chart



Example: 50 nM BaP, 4 h

Madureira et al., 2014,
Chem. Res. Toxicol., 27, 443–453

Venn Diagram

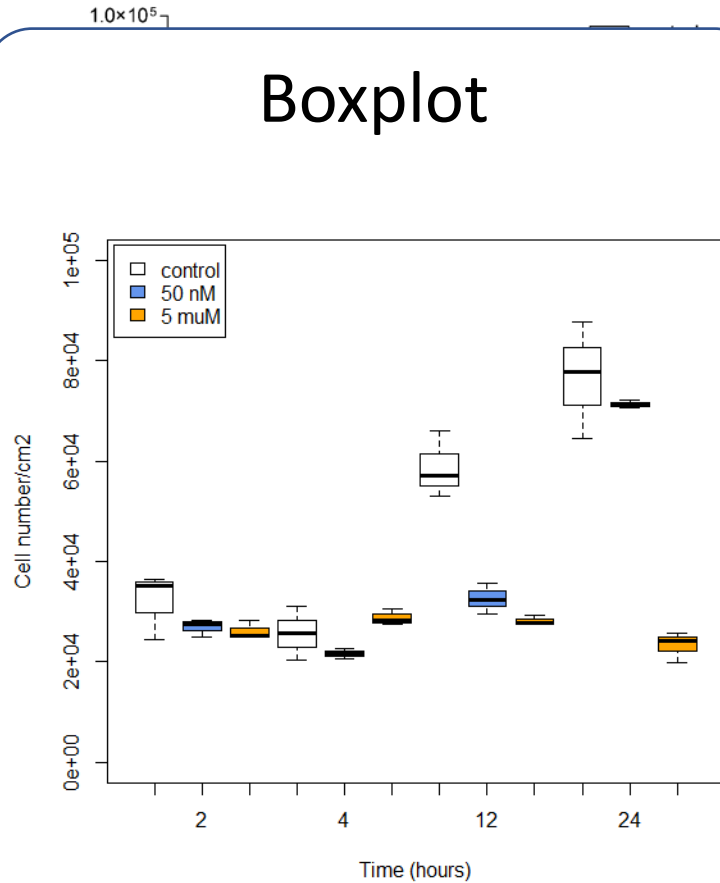


Bittner et al., 2011, Int J
Environ
Res Public Health

These are all **suboptimal** ways to represent the data

...better alternatives!

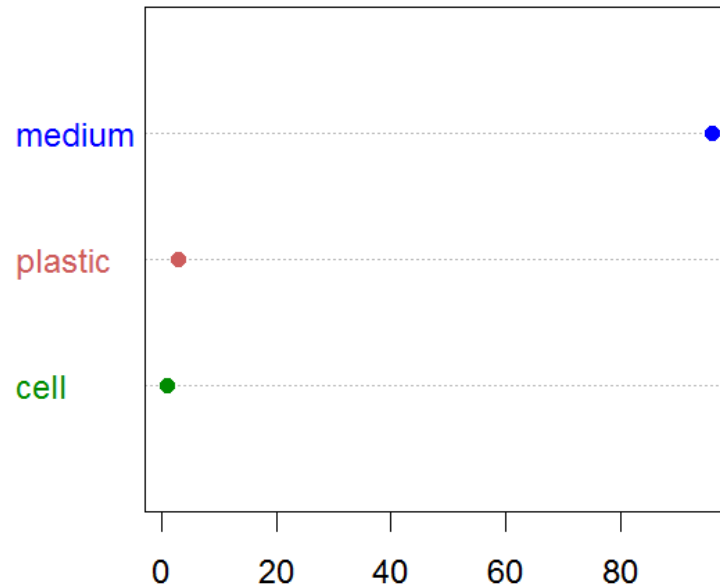
Barplot



Pie Chart

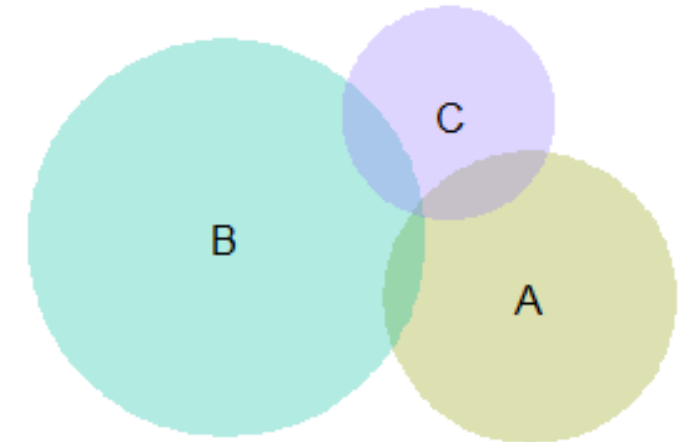
Dot Chart

Example: 50nM BaP, 4h



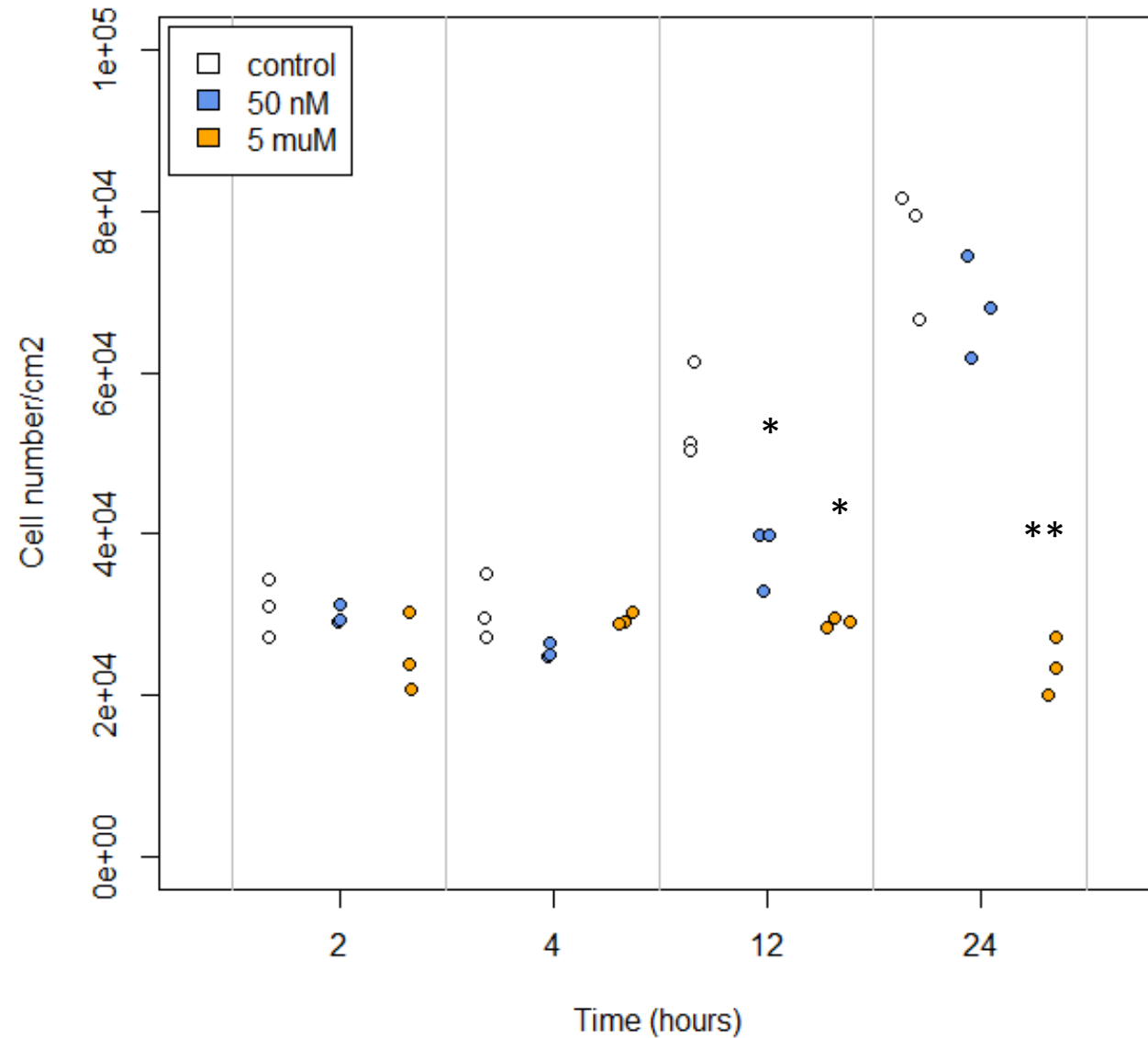
Venn Diagram

Euler Diagram



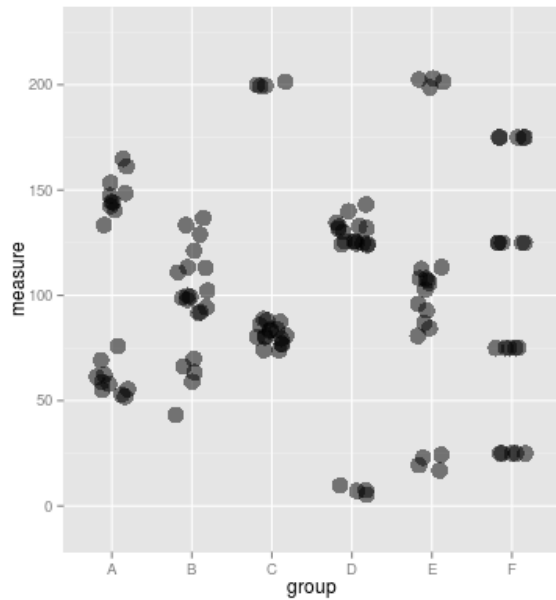
A=Albock, B=Brienzig, C=pooled

...or even simpler

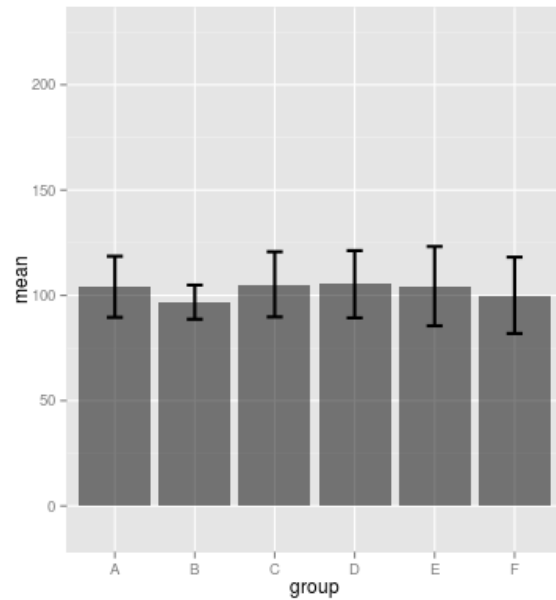


Barplots are usually nonsense...

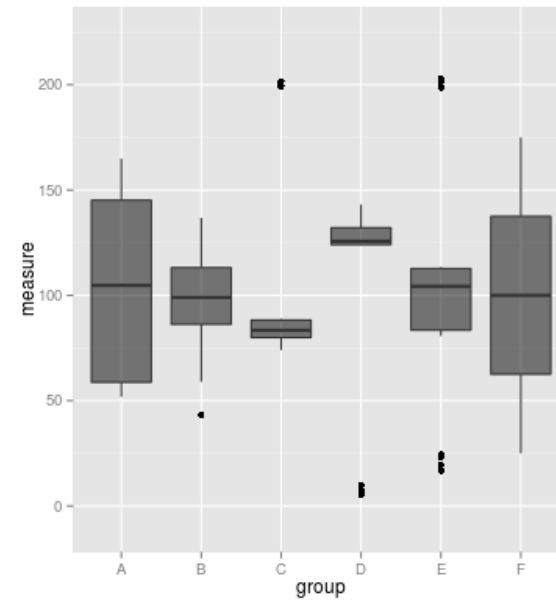
Scatter Plot



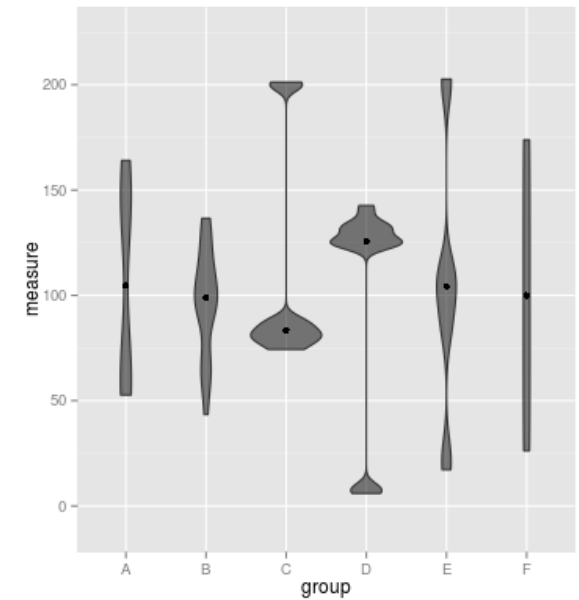
Bar Plot



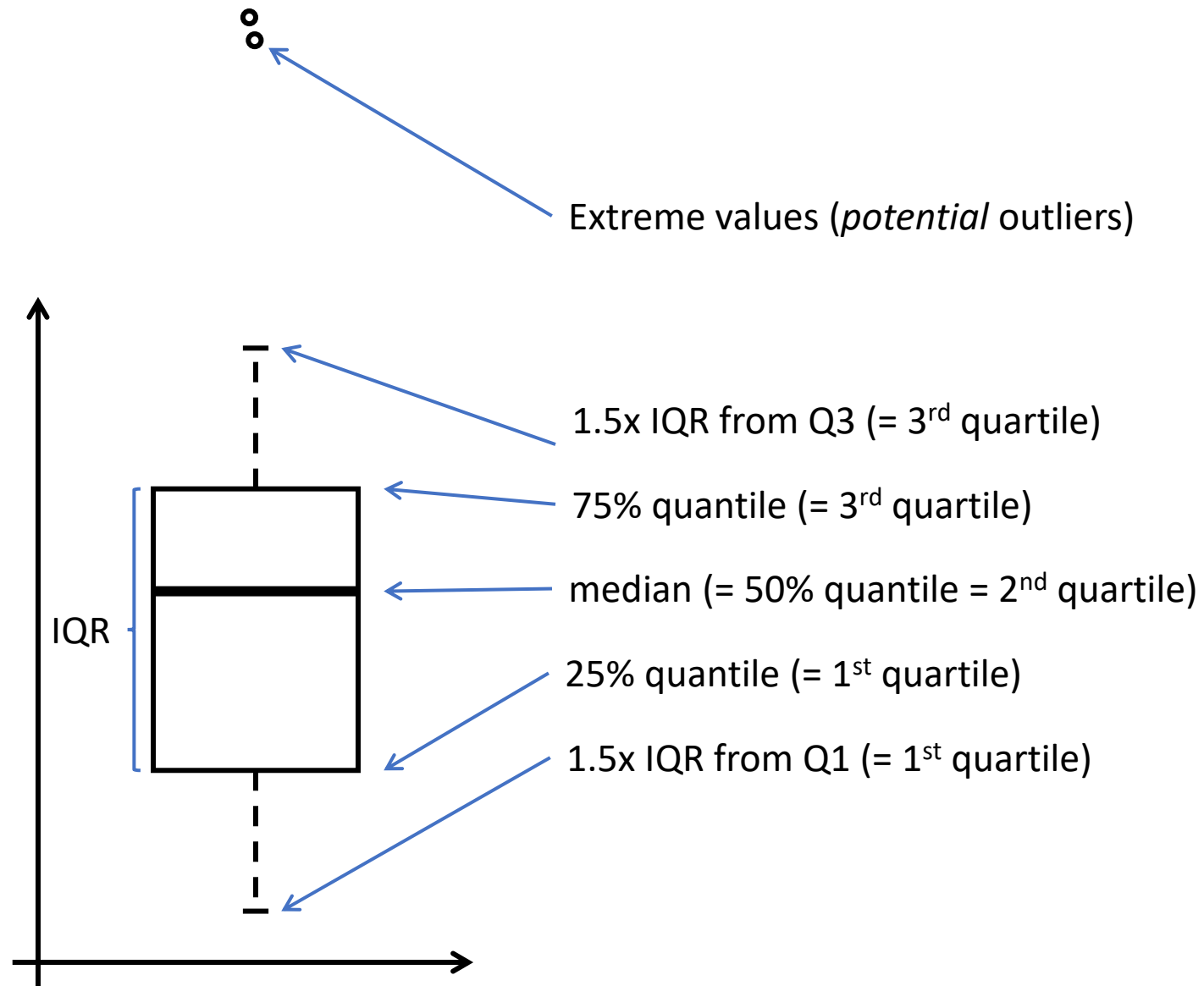
Box Plot



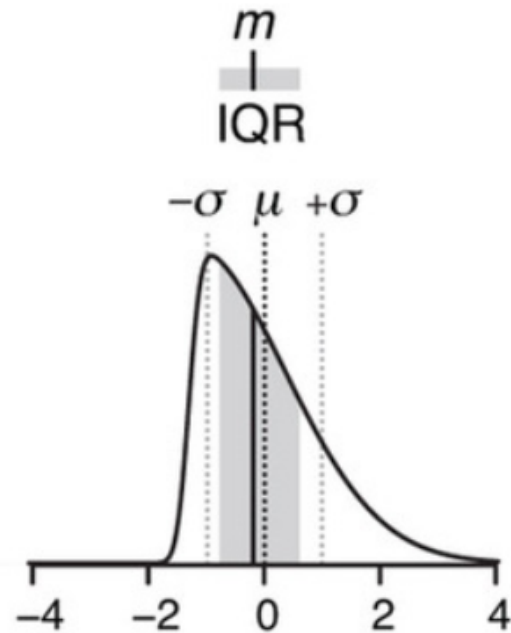
Violin Plot



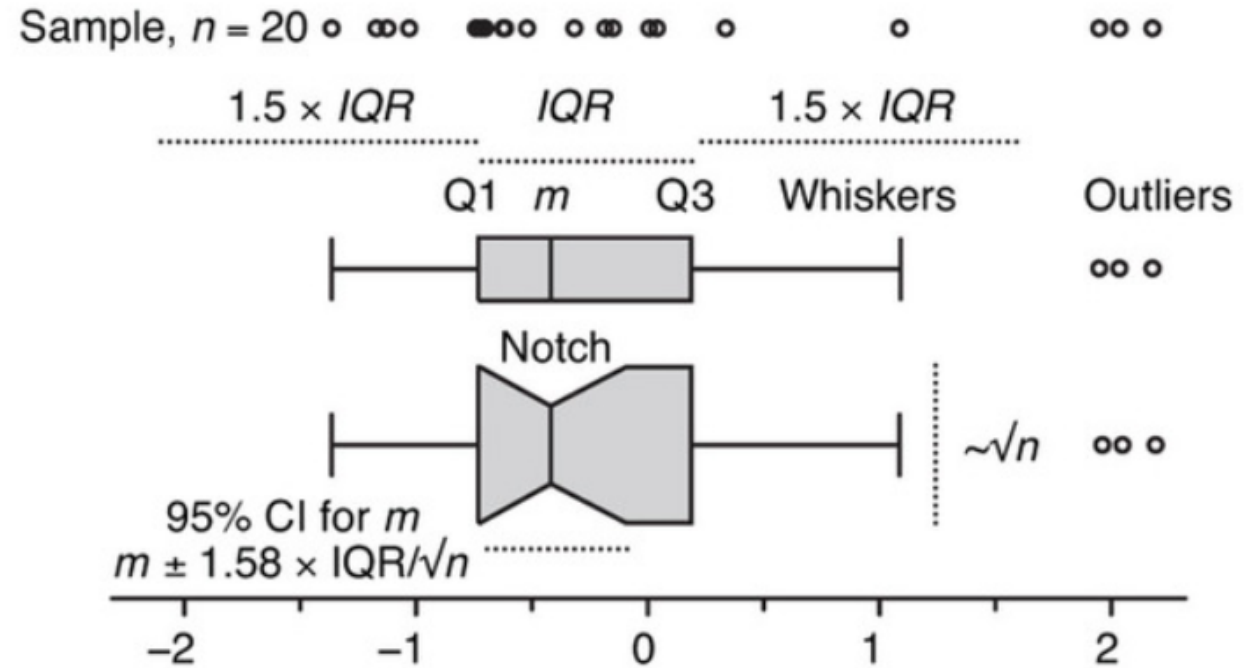
Boxplot



a Population distribution

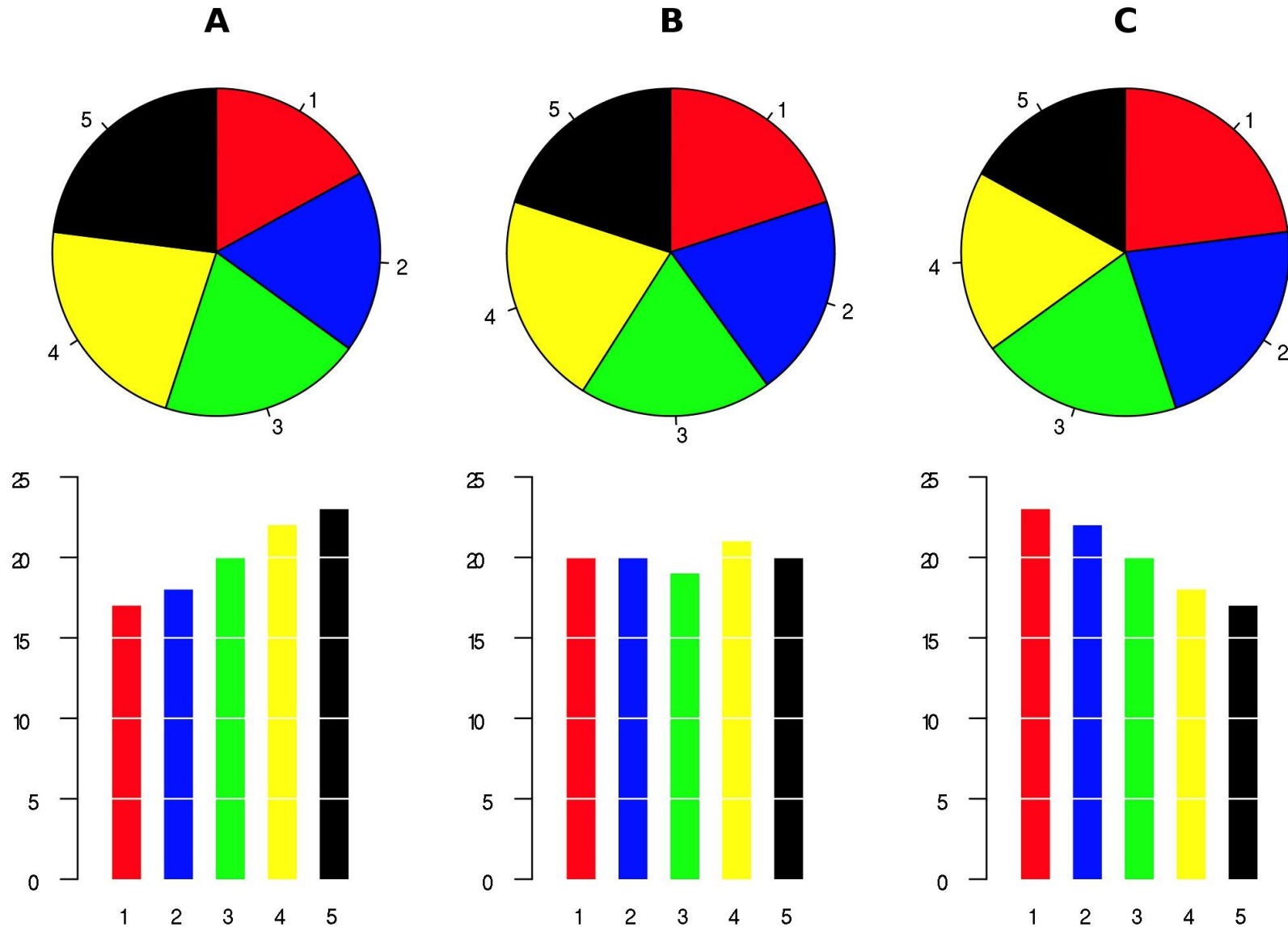


b Construction of a box plot



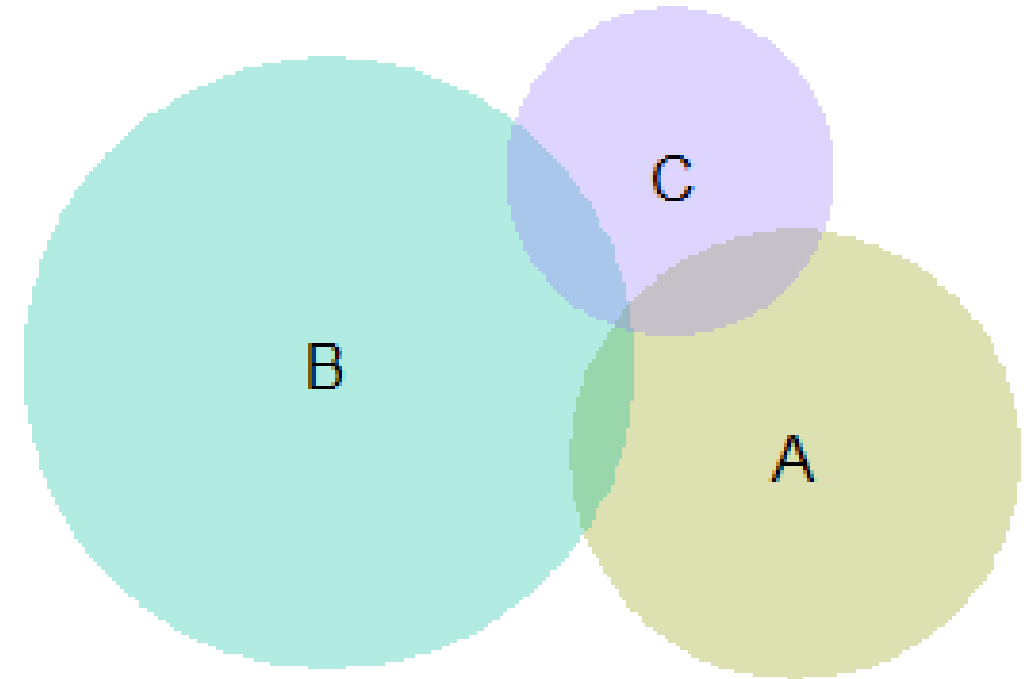
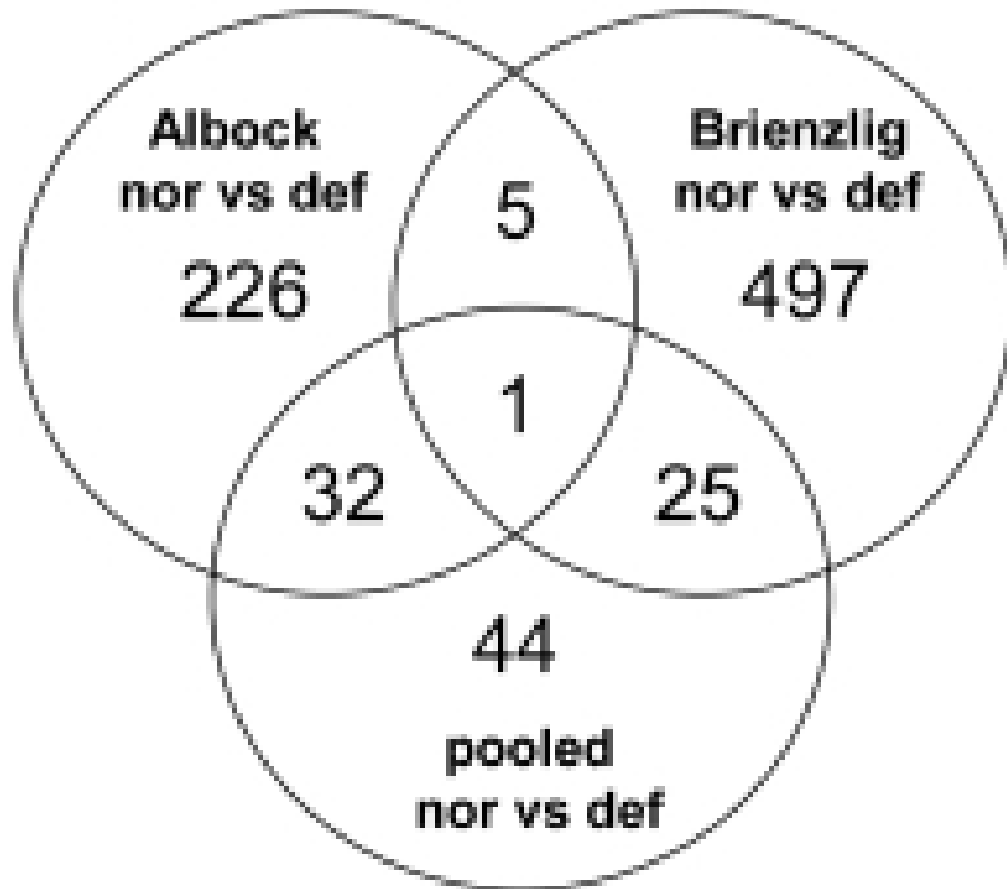
(a) The median ($m = -0.19$, solid vertical line) and interquartile range ($IQR = 1.38$, gray shading) are ideal for characterizing asymmetric or irregularly shaped distributions. A skewed normal distribution is shown with mean $\mu = 0$ (dark dotted line) and s.d. $\sigma = 1$ (light dotted lines). (b) Box plots for an $n = 20$ sample from a. The box bounds the IQR divided by the median, and Tukey-style whiskers extend to a maximum of $1.5 \times IQR$ beyond the box. The box width may be scaled by \sqrt{n} , and a notch may be added approximating a 95% confidence interval (CI) for the median. Open circles are sample data points. Dotted lines indicate the lengths or widths of annotated features.

Use dot chart (or even bar charts) instead of pie chart



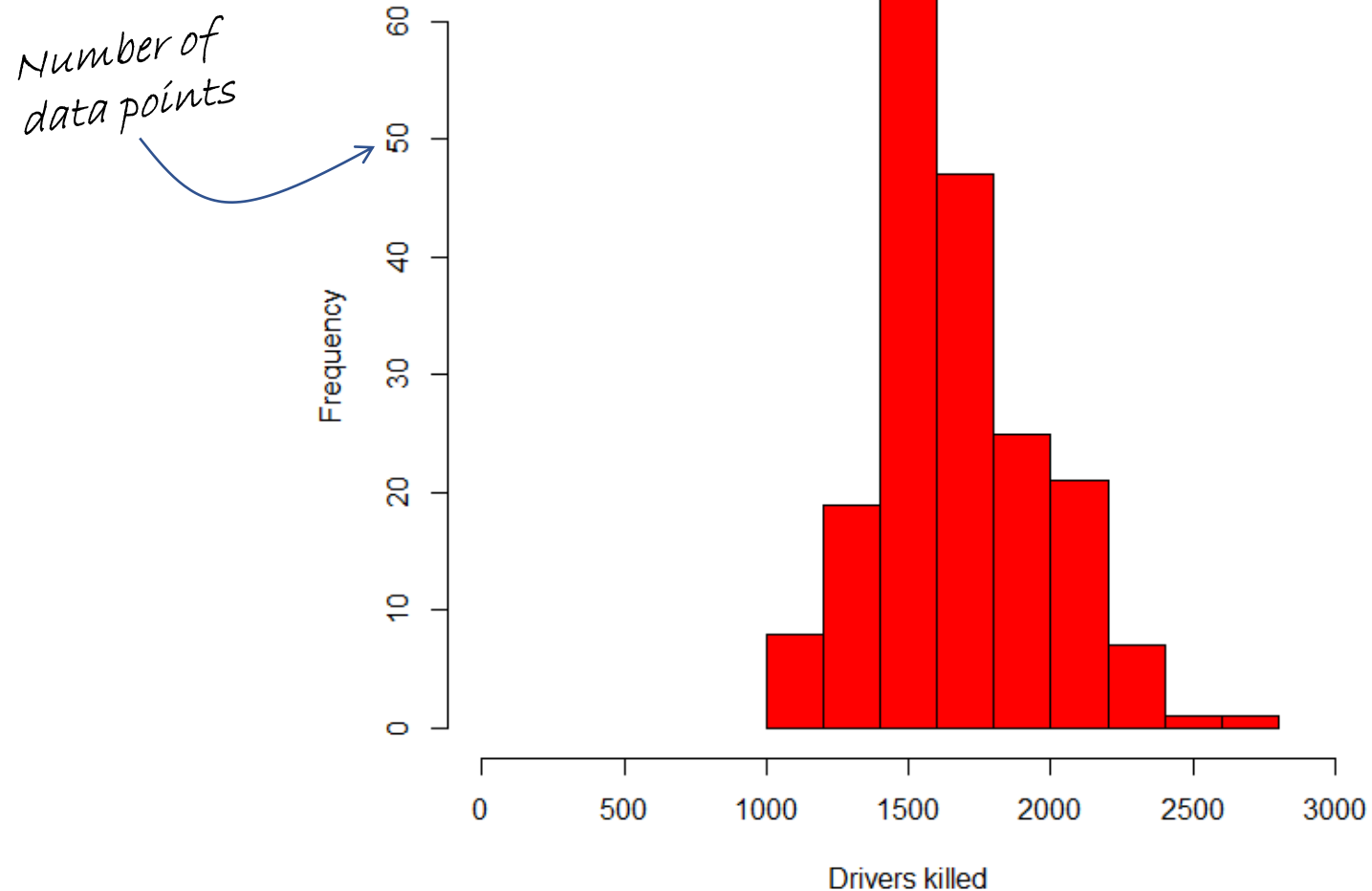
Venn versus Euler diagrams – what is the difference?

b)

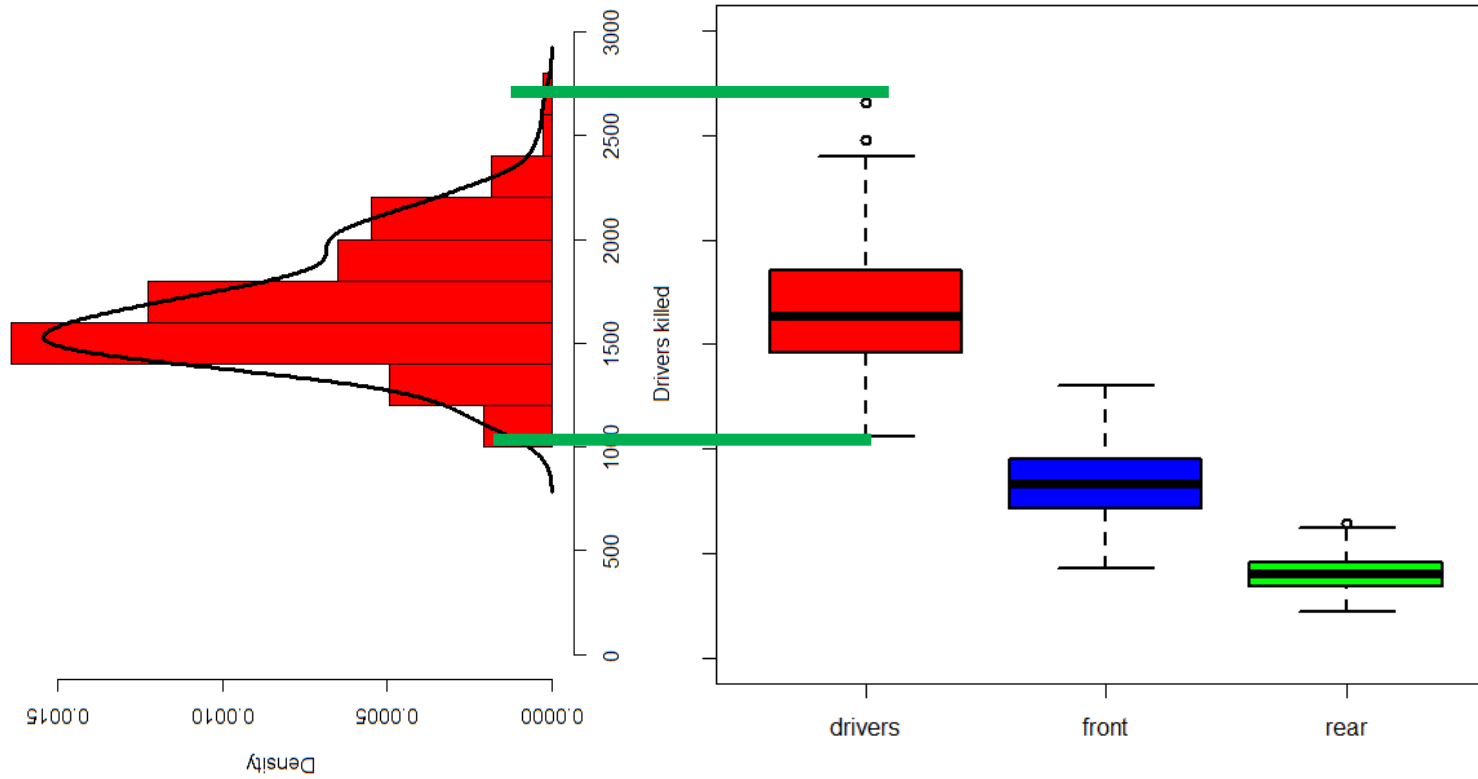


A=Albock, B=Brienzlig, C=pooled

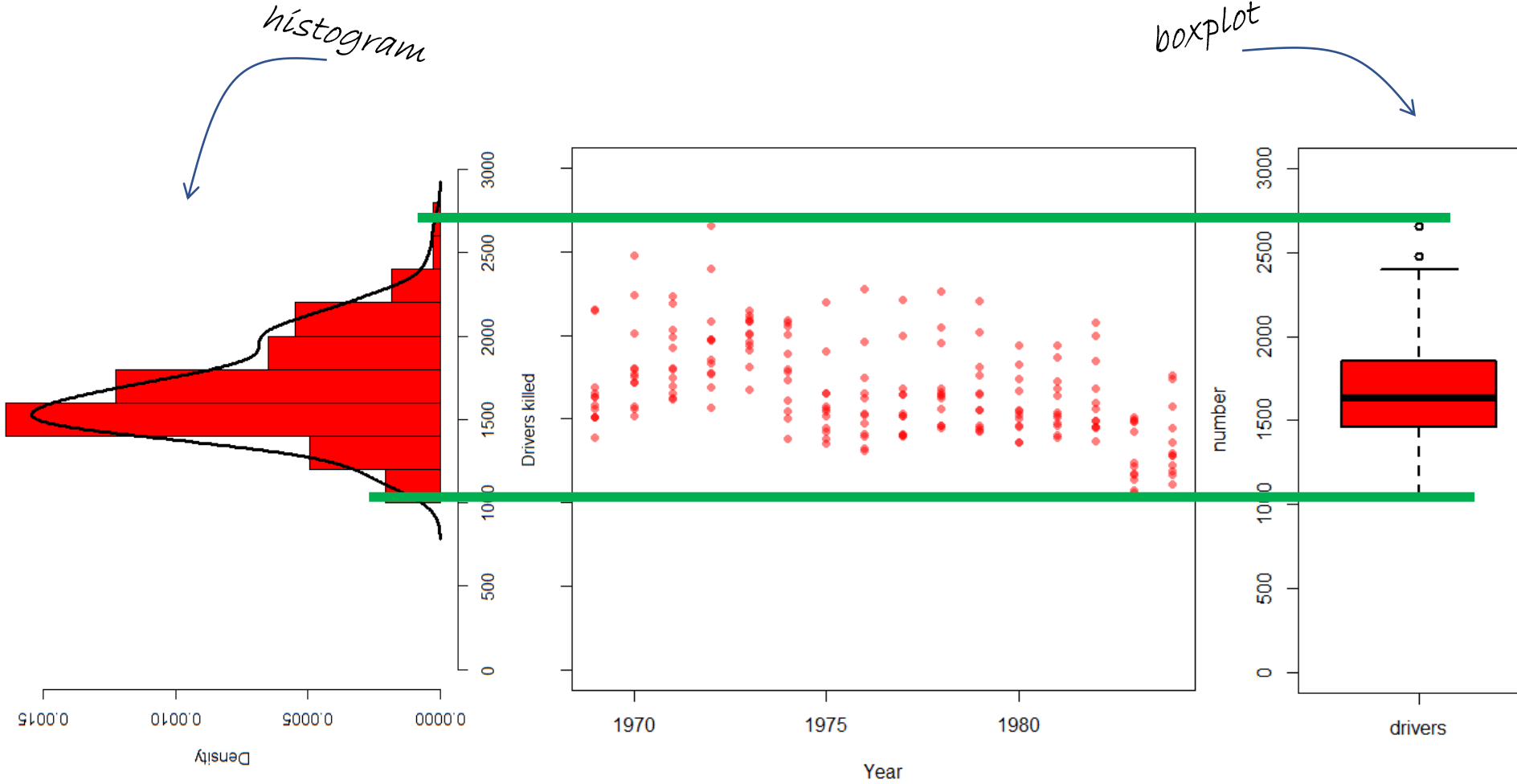
Histogram (a.k.a. frequency distribution)



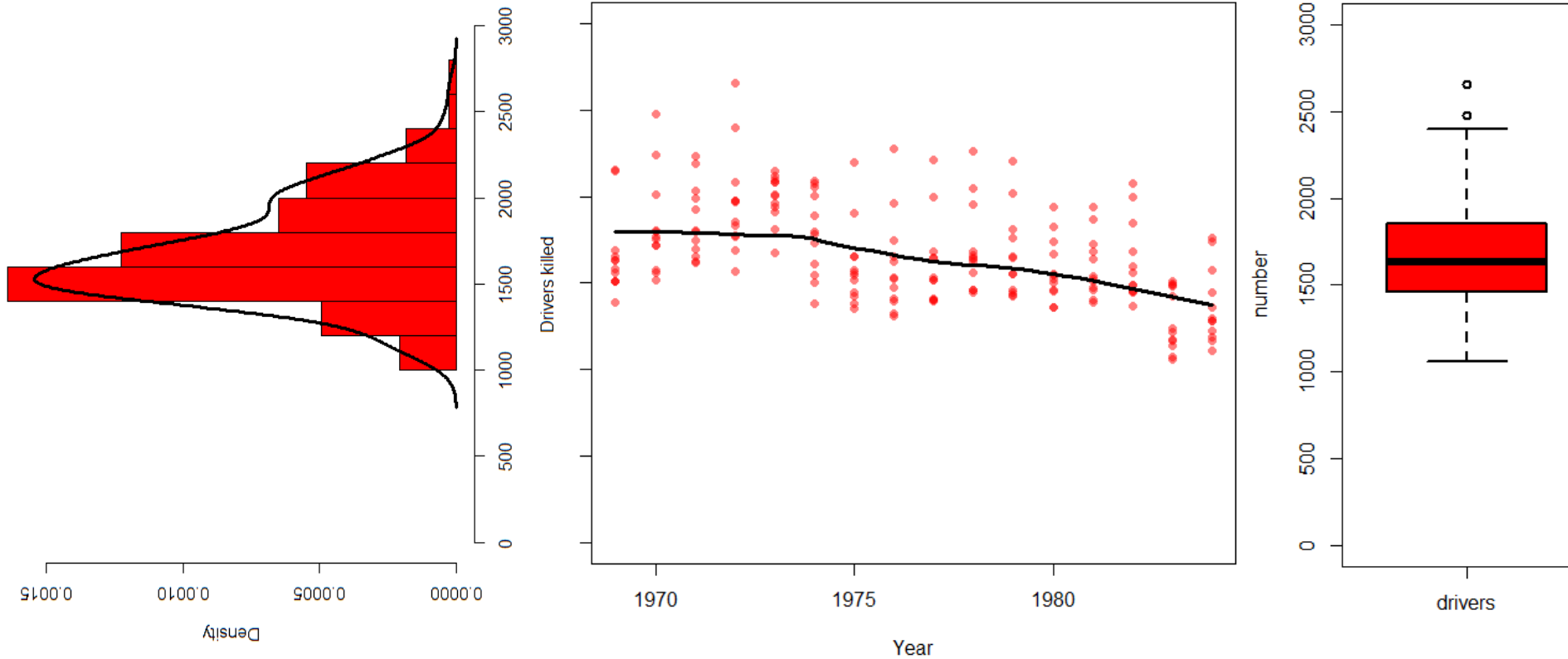
Histogram and box plots



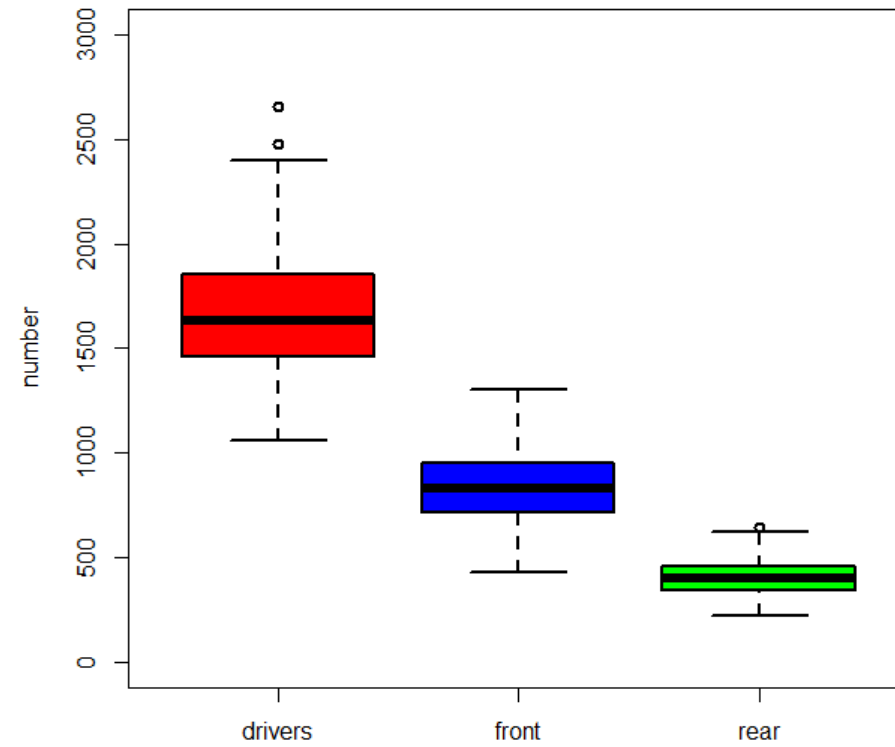
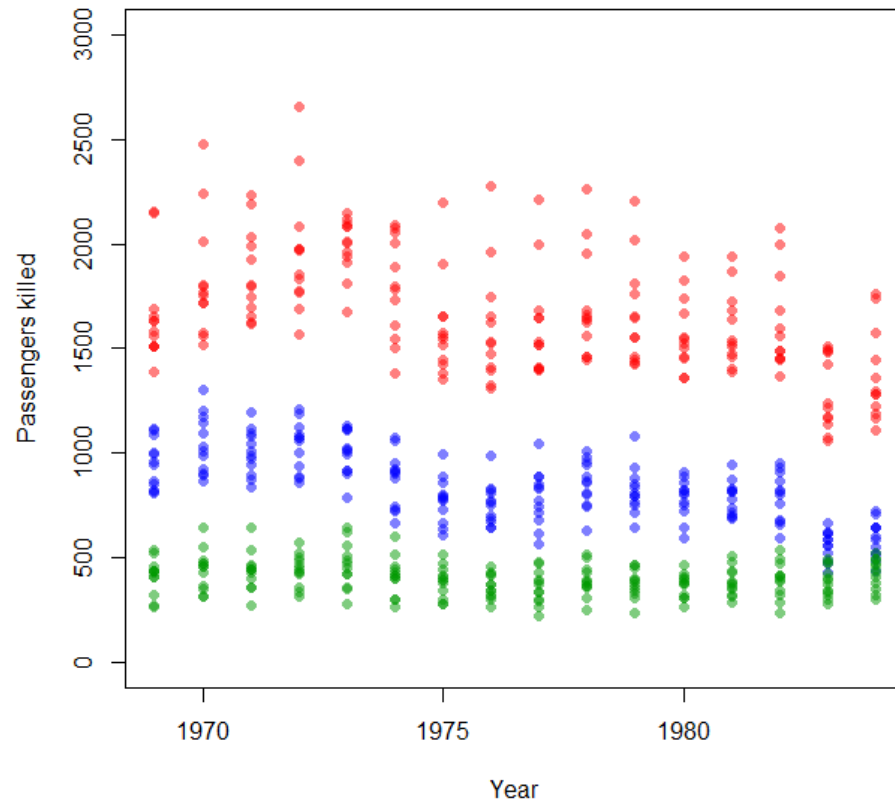
Scatterplot



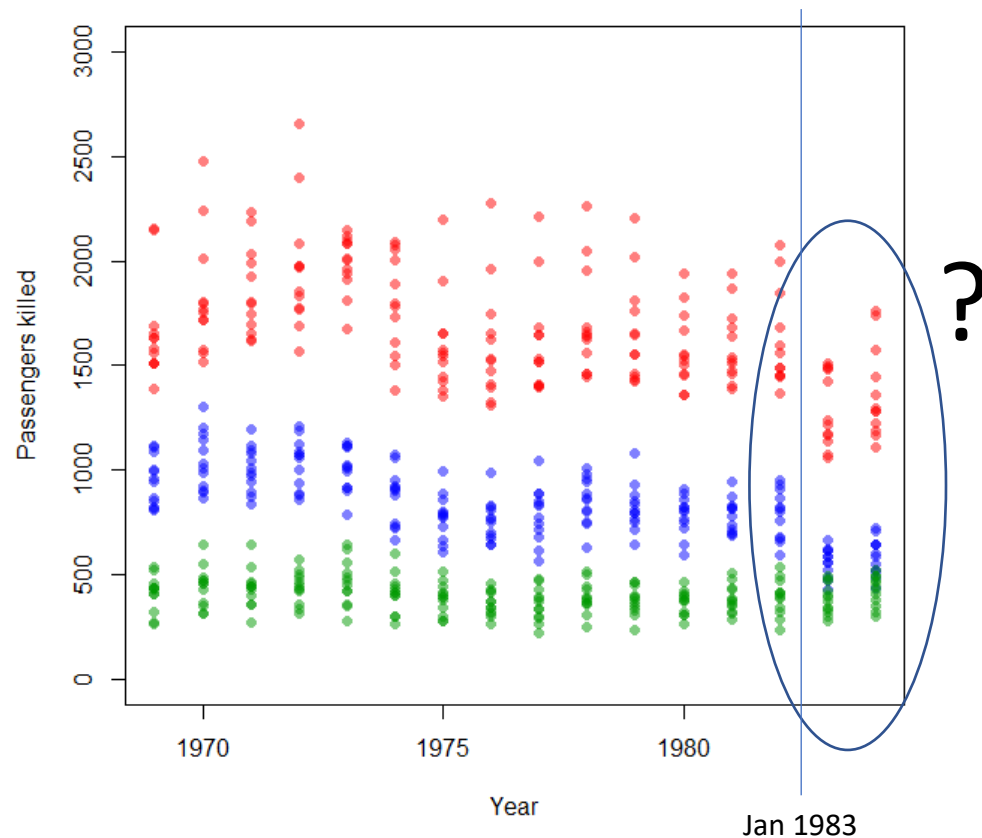
Scatterplot (with some stats added)



Scatterplot versus boxplot



...and now some statistical thinking

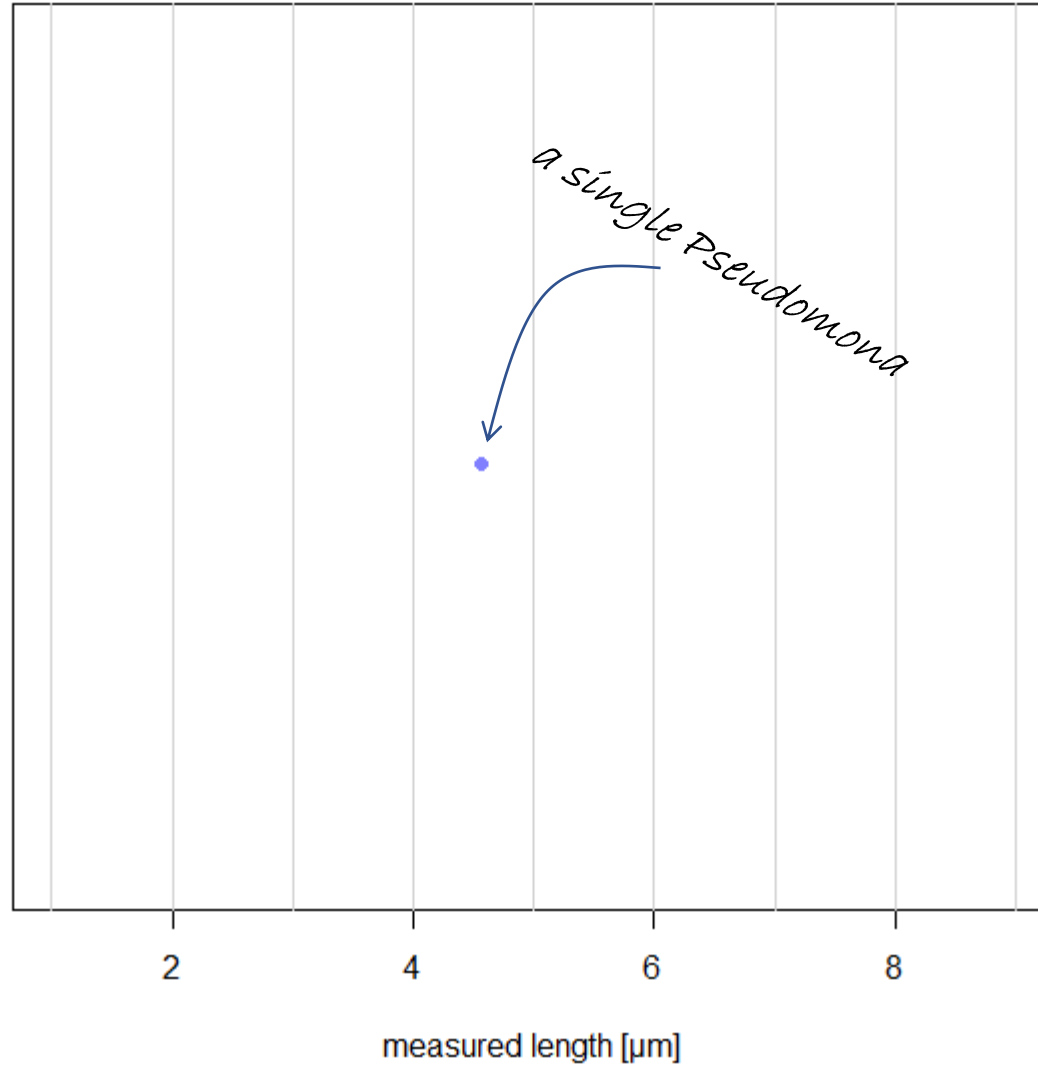


In January 1983 a law was passed in the UK which made wearing seat belts mandatory

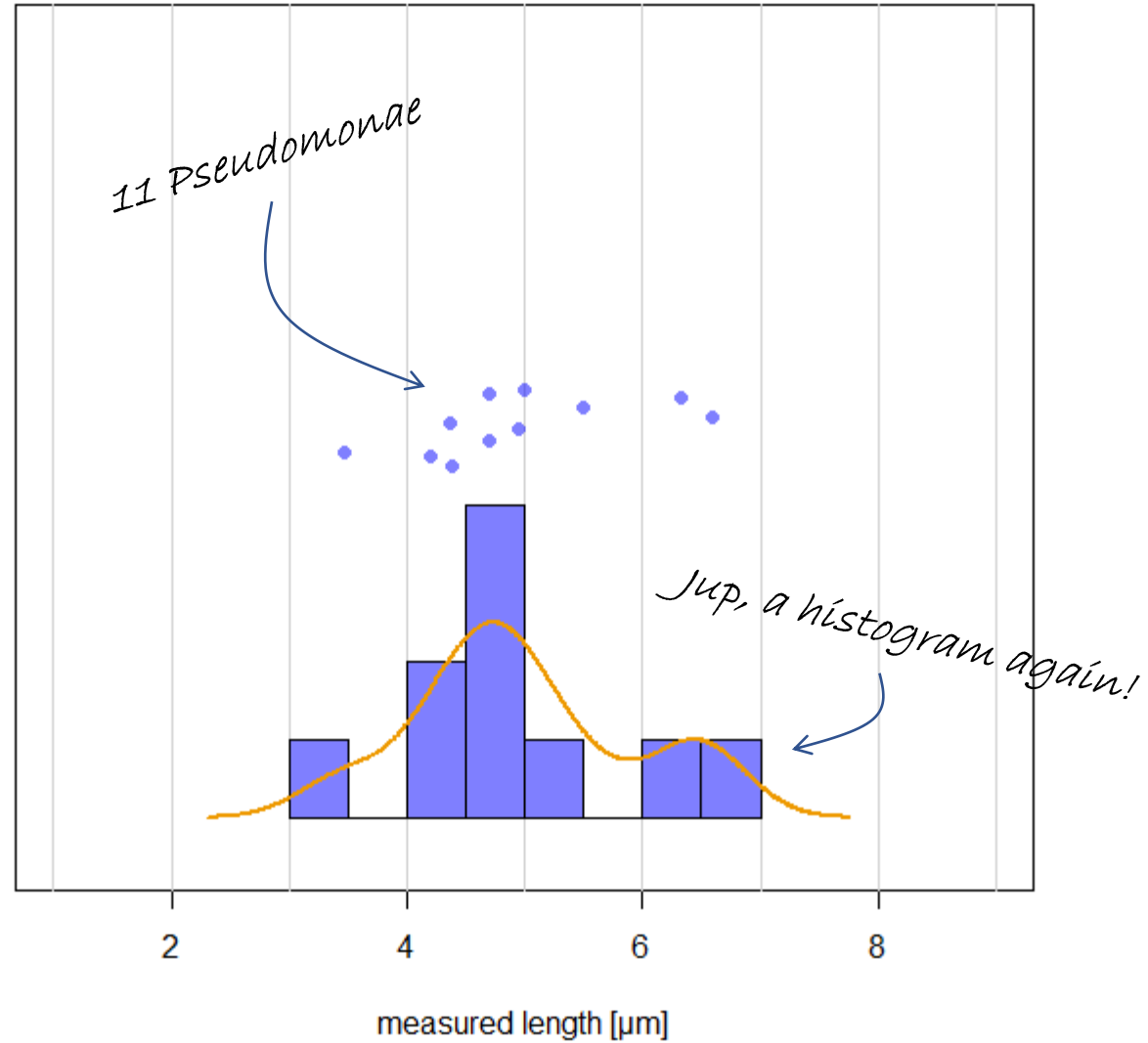
The Normal Distribution



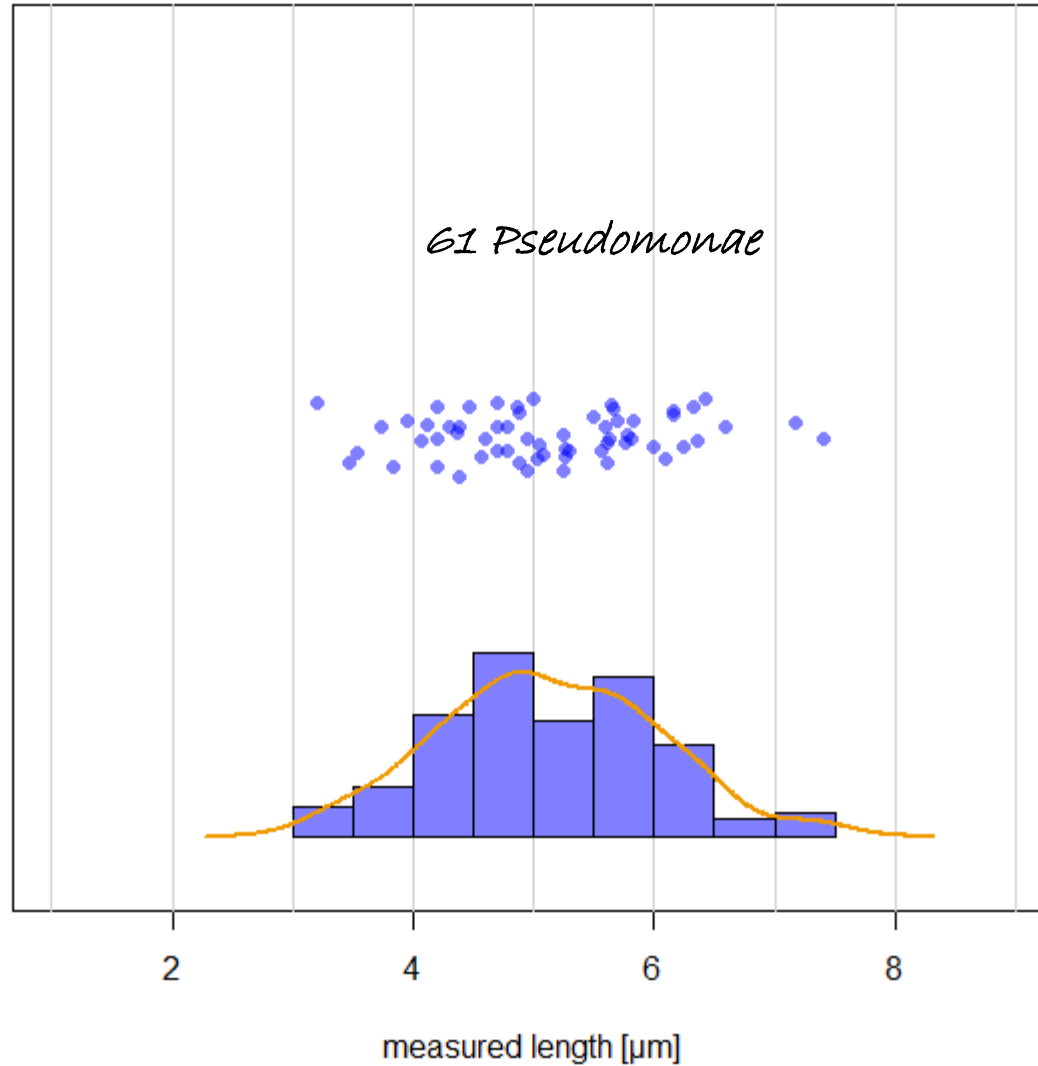
The Normal Distribution



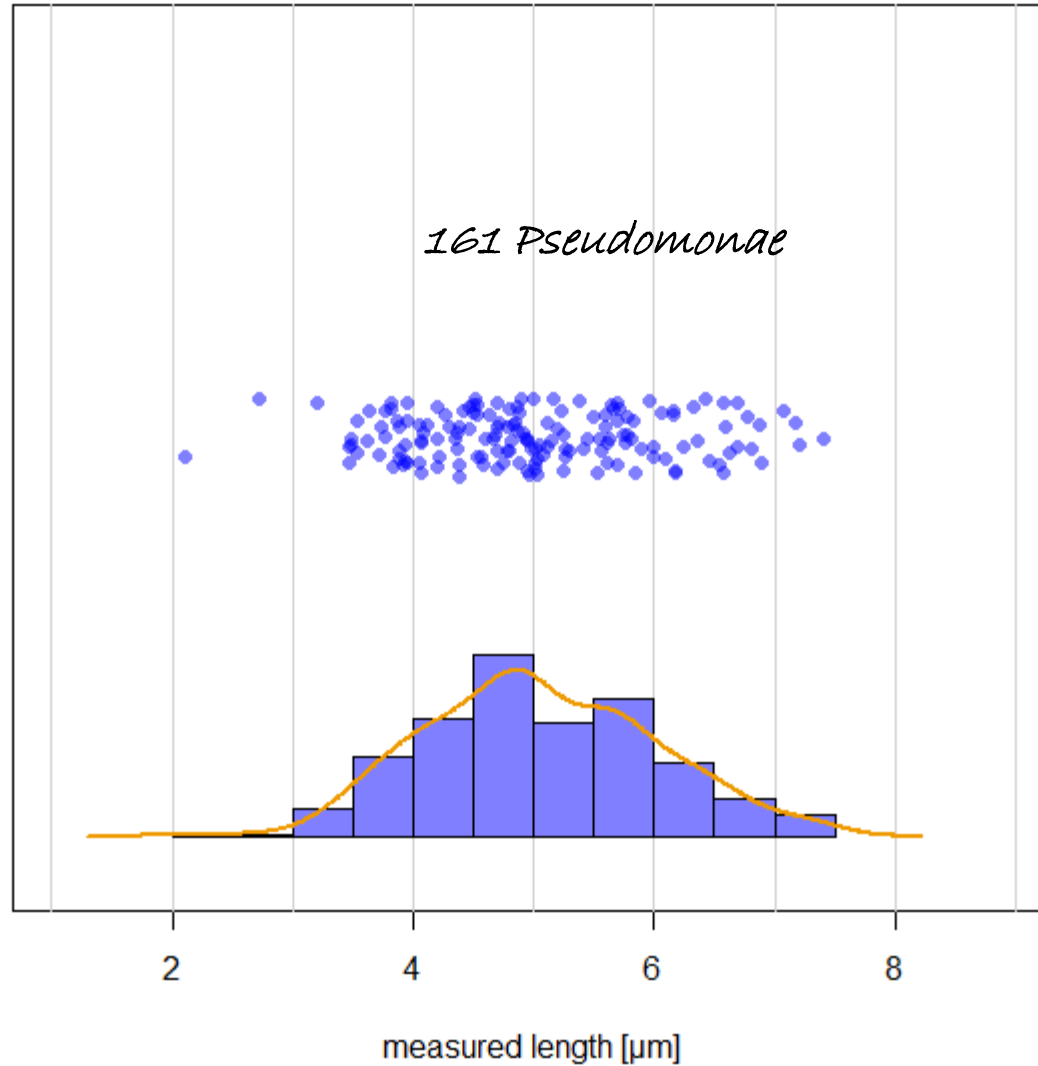
The Normal Distribution



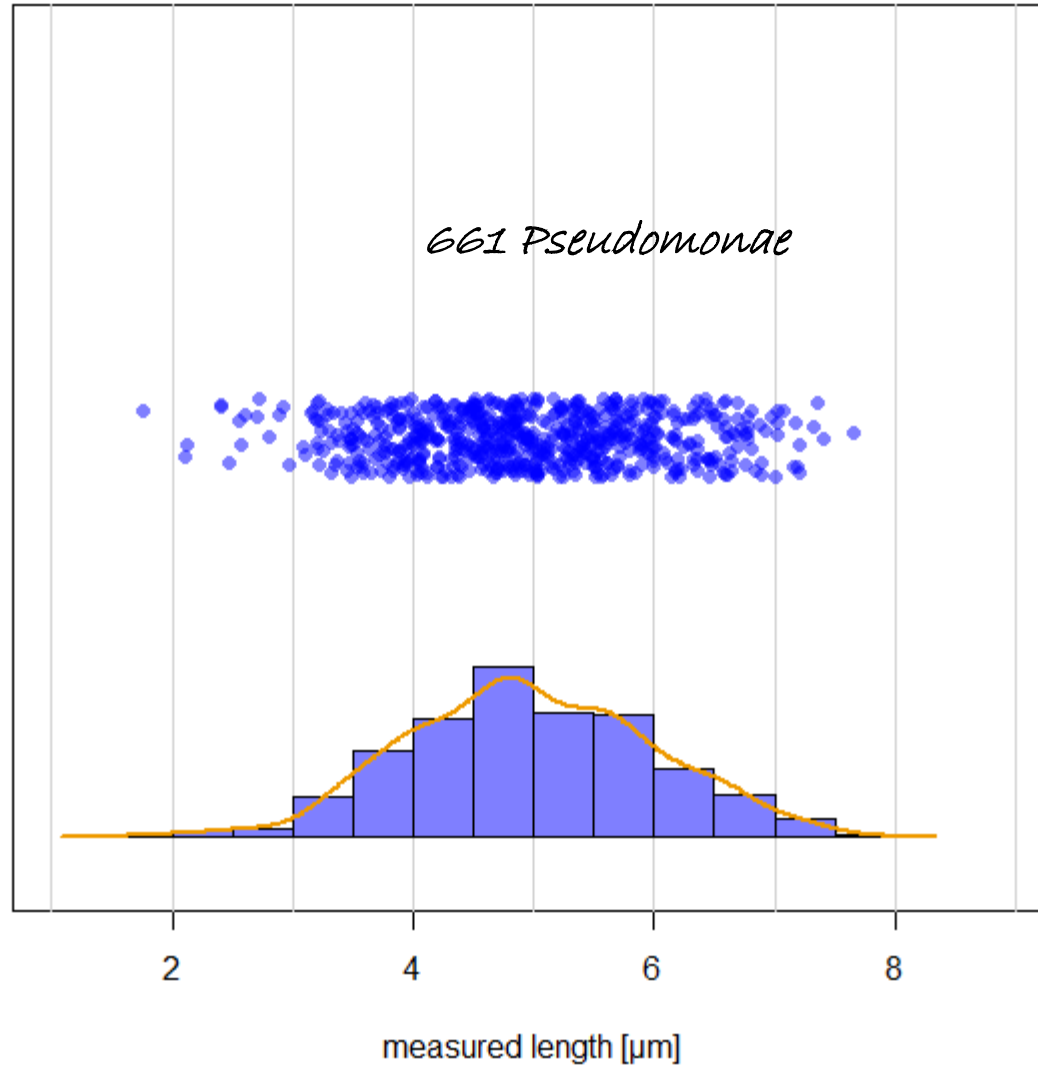
The Normal Distribution



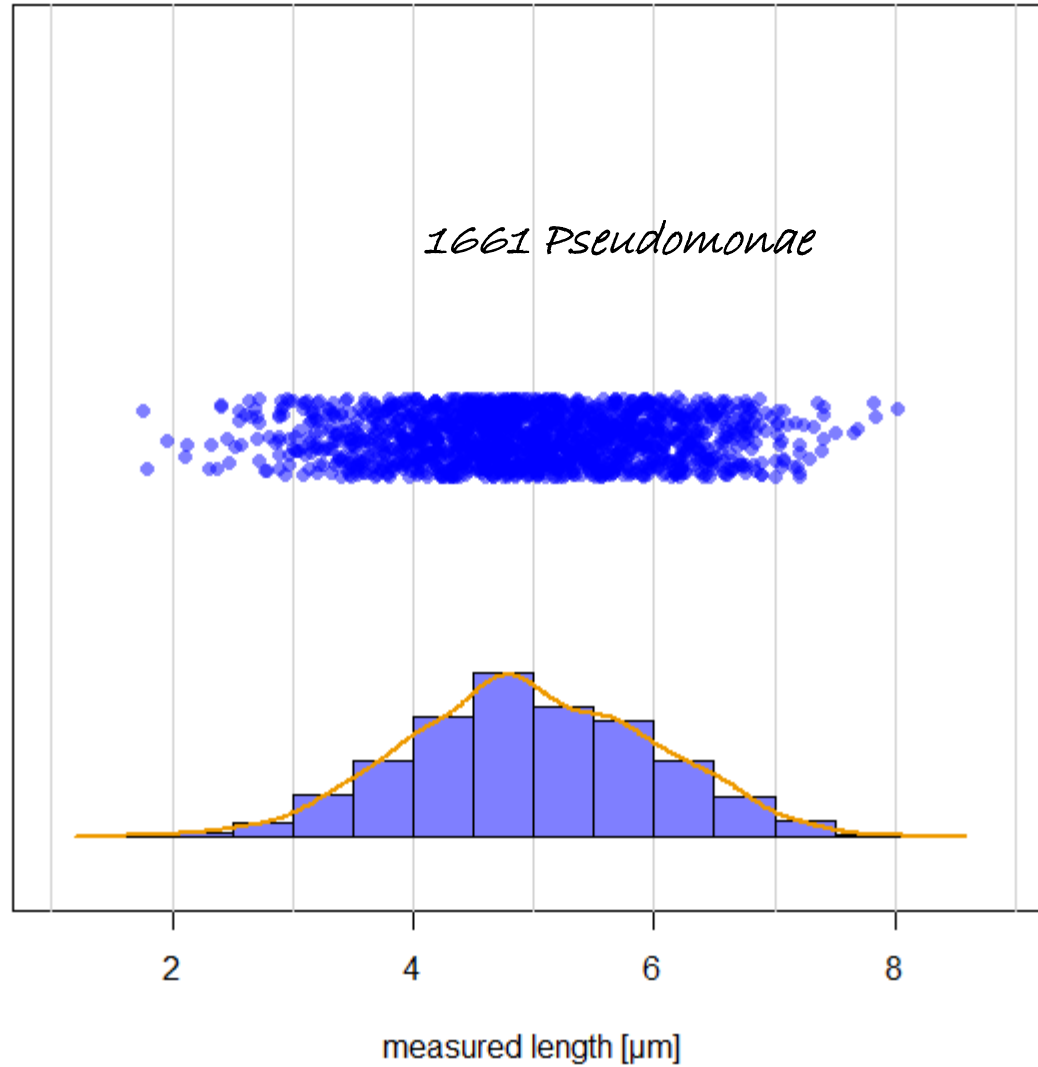
The Normal Distribution



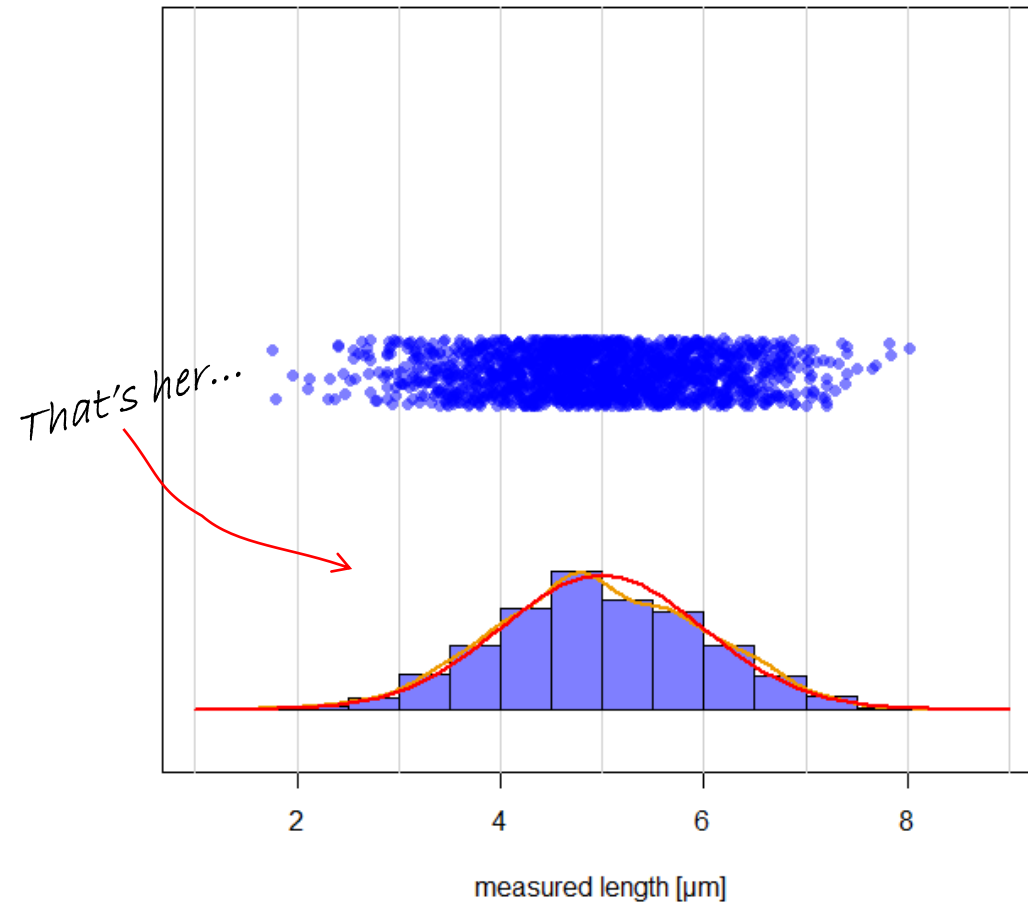
The Normal Distribution



The Normal Distribution



The Normal Distribution, thanks to CLT (Central Limit Theorem; ZGWS=Zentraler GrenzwertSatz)

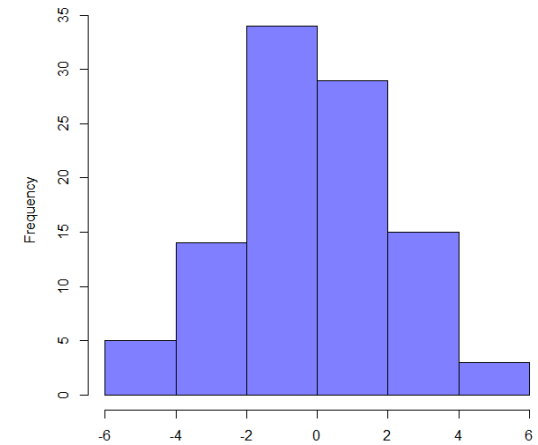
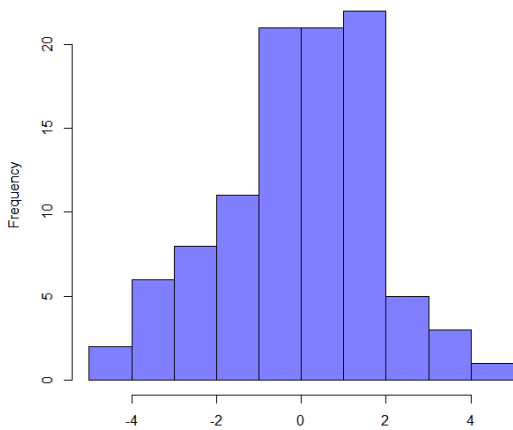
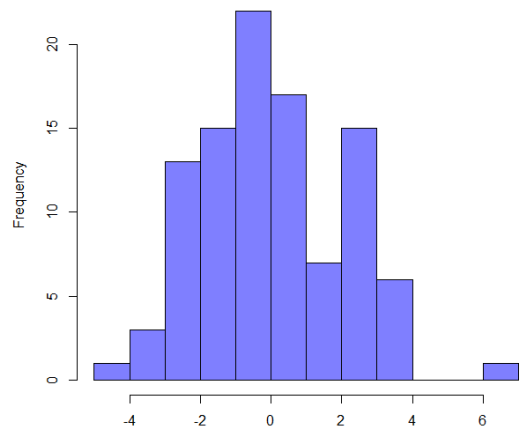
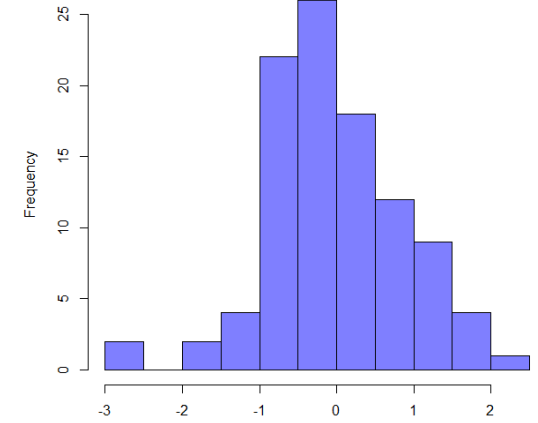
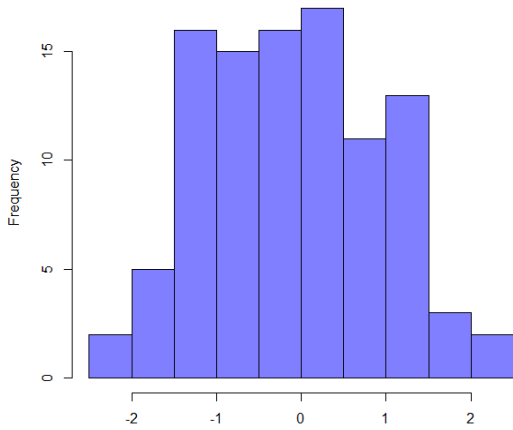
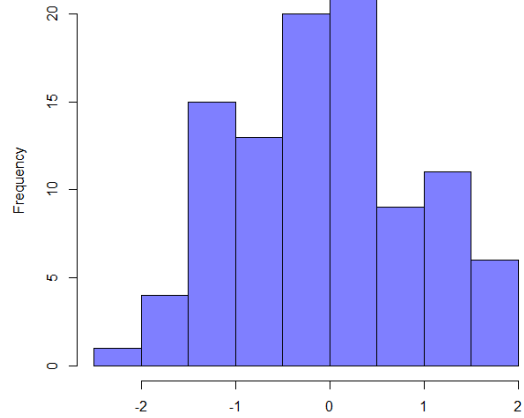


Why are you bugging us with this?

- Normality is a typical assumption (a.k.a. requirement) for many tests, models and methods!
 - t-test
 - χ^2 -test (and G-test)
 - Linear regression
 - Correlation

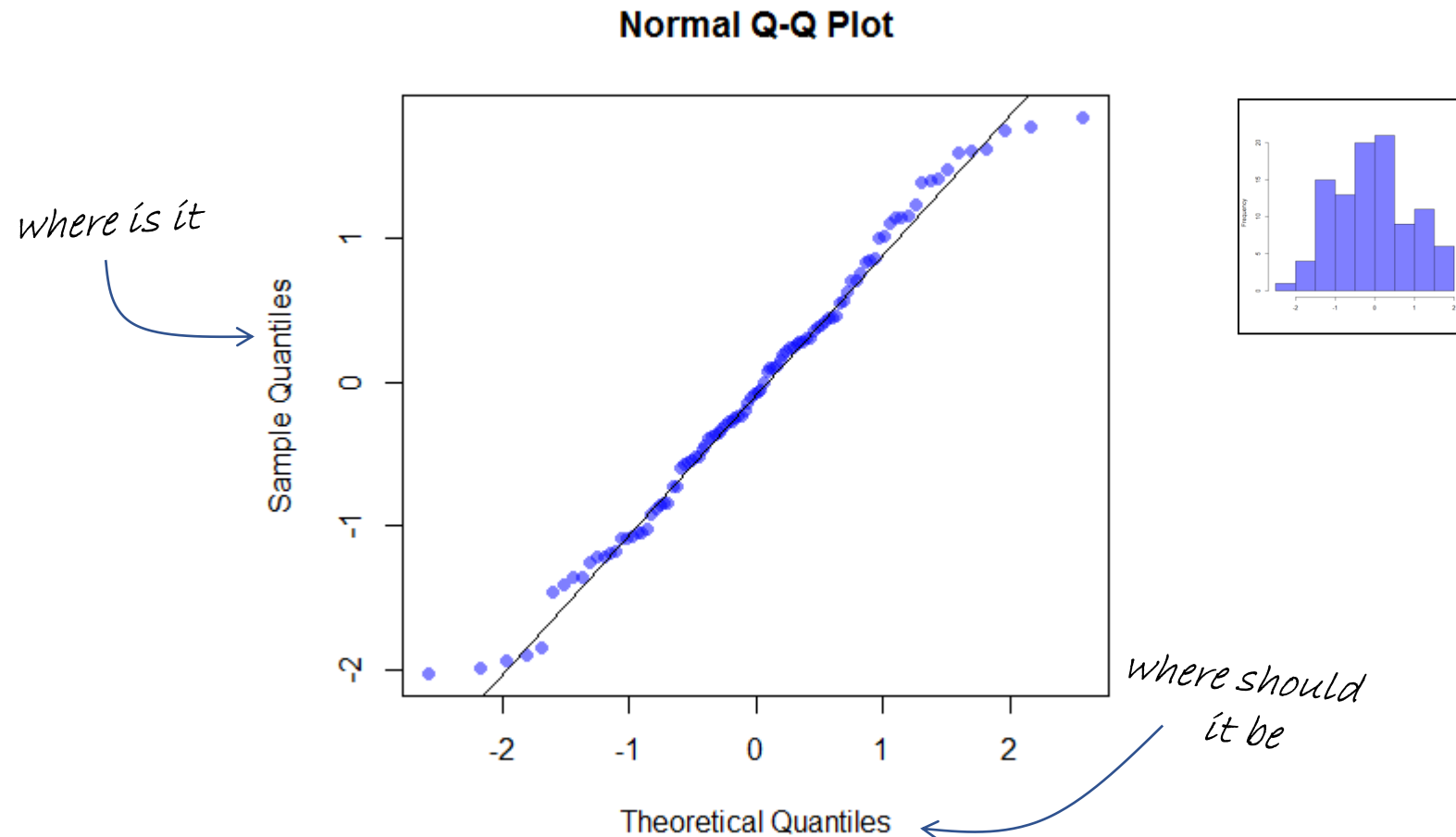
Without Normality these tests and methods are virtually **useless!**

When do we have Normality?

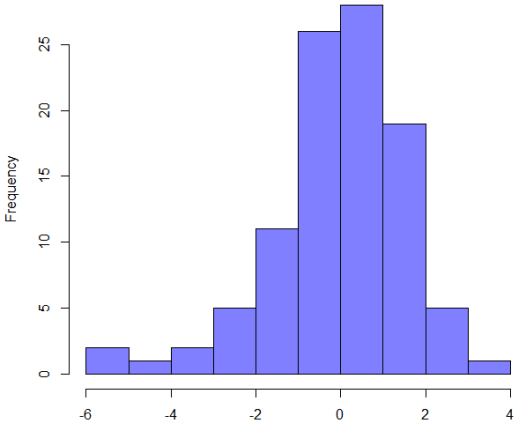


When do we have (**enough**) Normality?

- Take a look at the Quantile-Quantile (QQ) plot

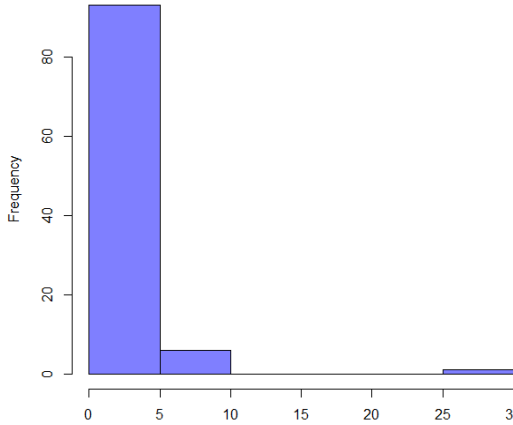
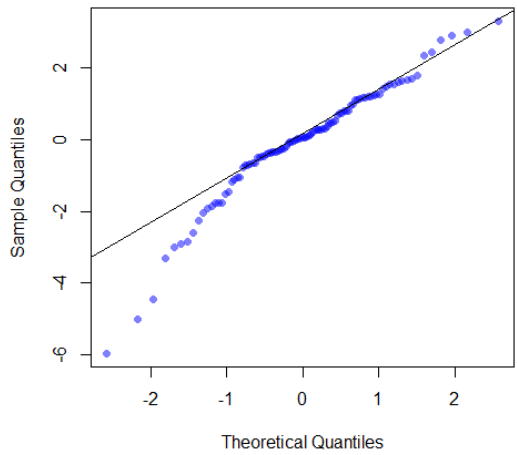


When do we have Normality?



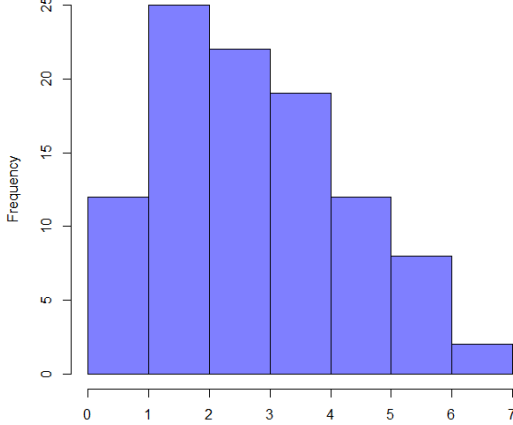
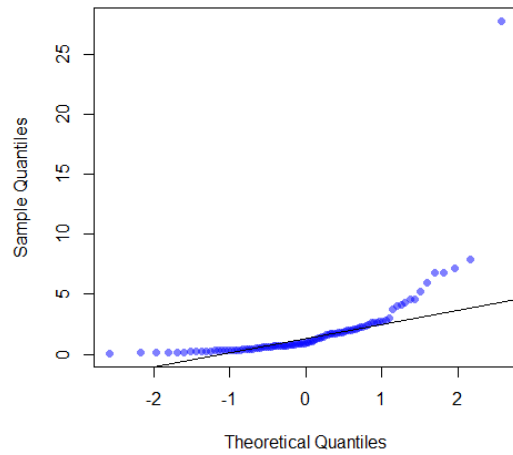
Student's t distribution

Normal Q-Q Plot



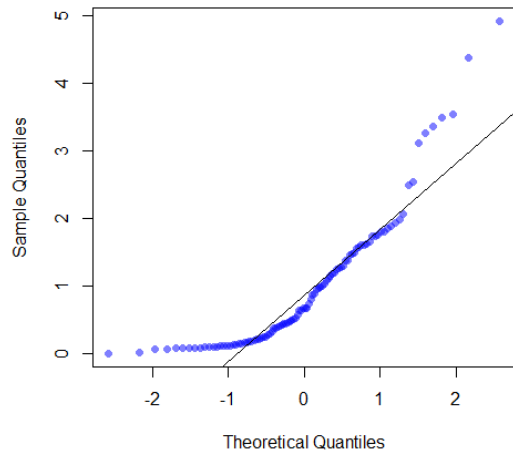
Log-normal distribution

Normal Q-Q Plot

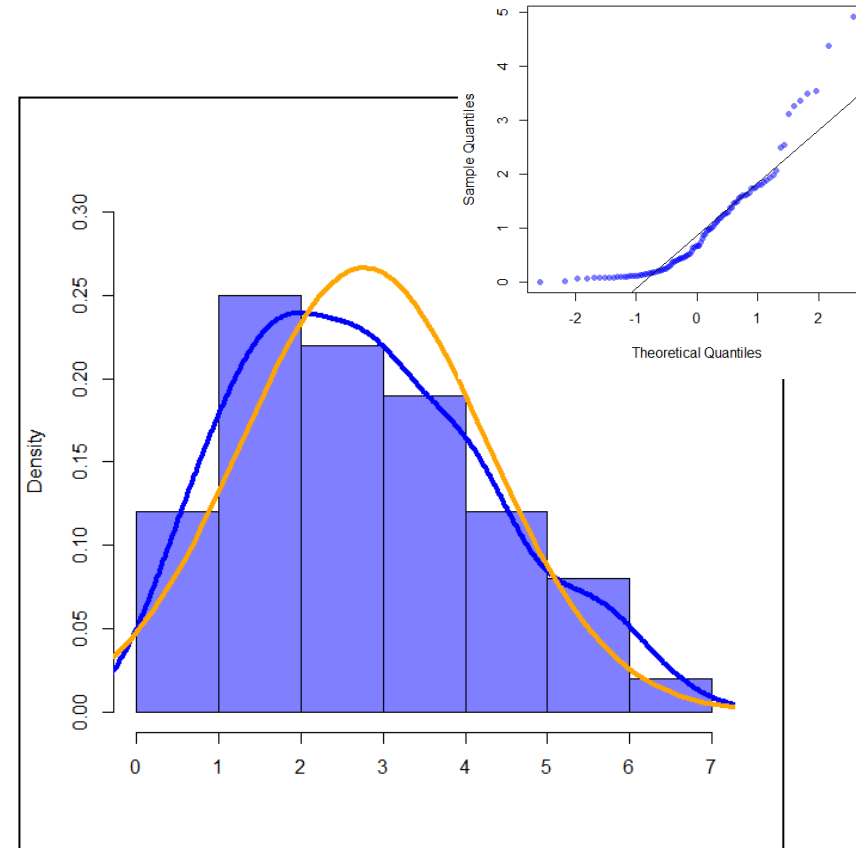
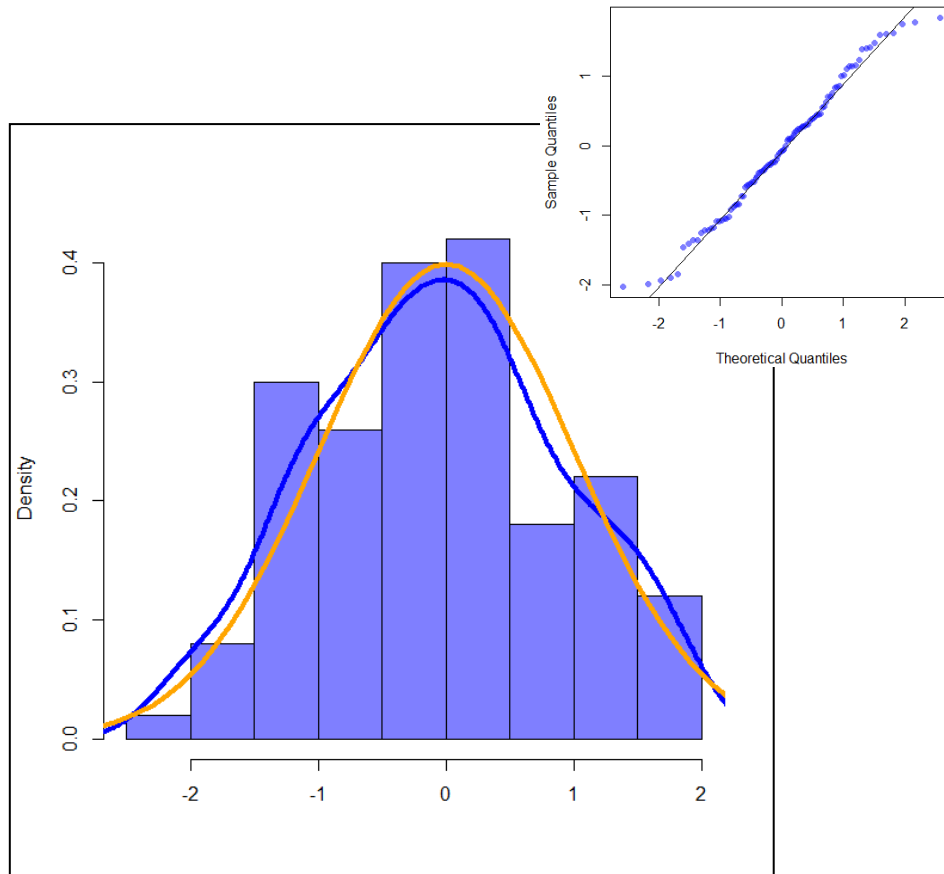


Gamma(3,1) distribution

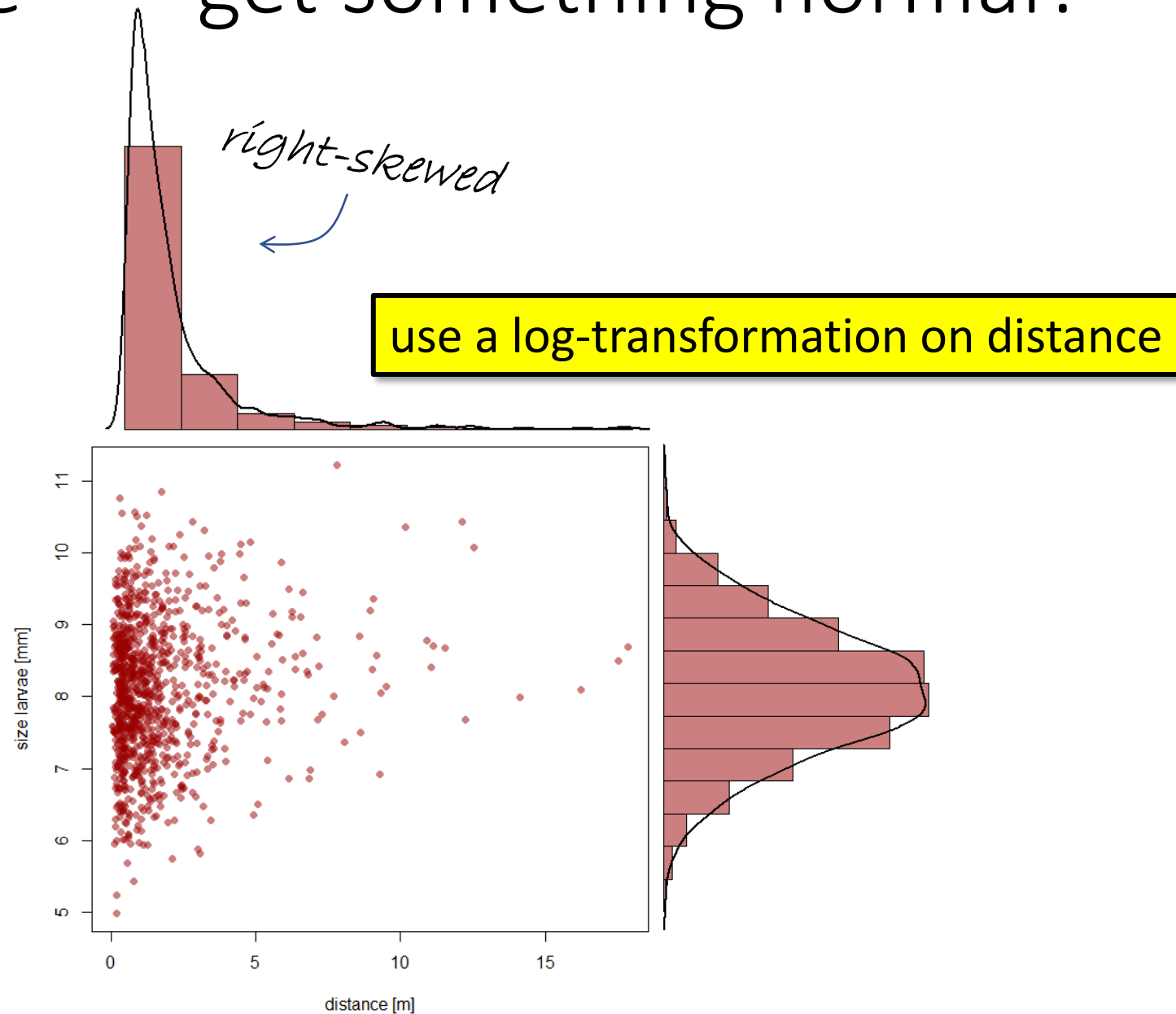
Normal Q-Q Plot



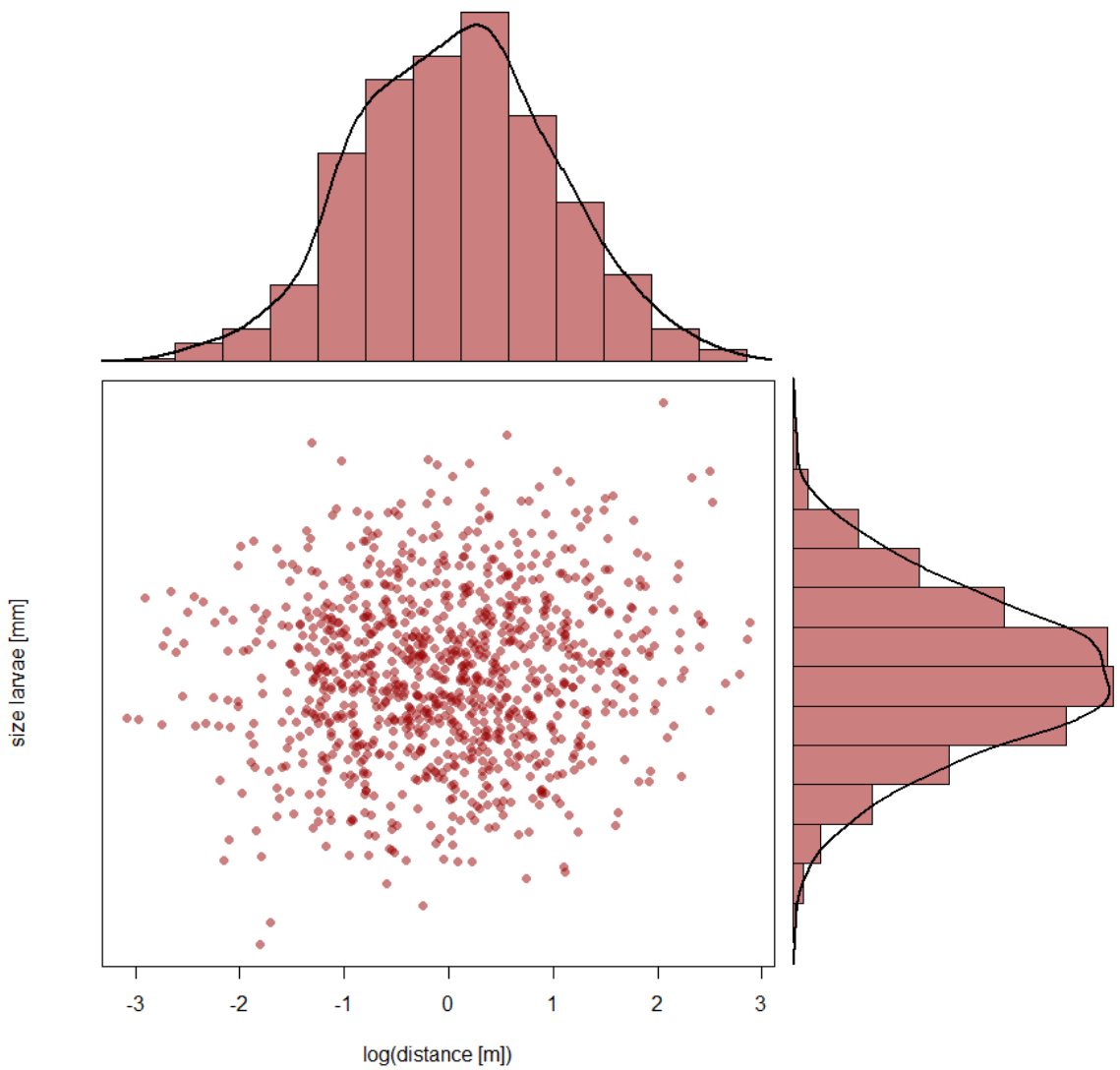
When do we have Normality?



How can we get something normal?



How can we get something normal?



How can we get something normal?

When can I use a log transformation?

Always, when the data...

- Only takes **positive values** (e.g. concentrations)

...but also...

- for **strongly right-skewed** data;
- **count variables**.

Log transforms can be justified in an ideal setting I

- Log

Log transforms can be justified in an ideal setting II

- Log Log

Outliers

- An observation that is distant from other observations

- Indication:

- Measurement error
- Heavy-tails of the distribution

Get rid of it!



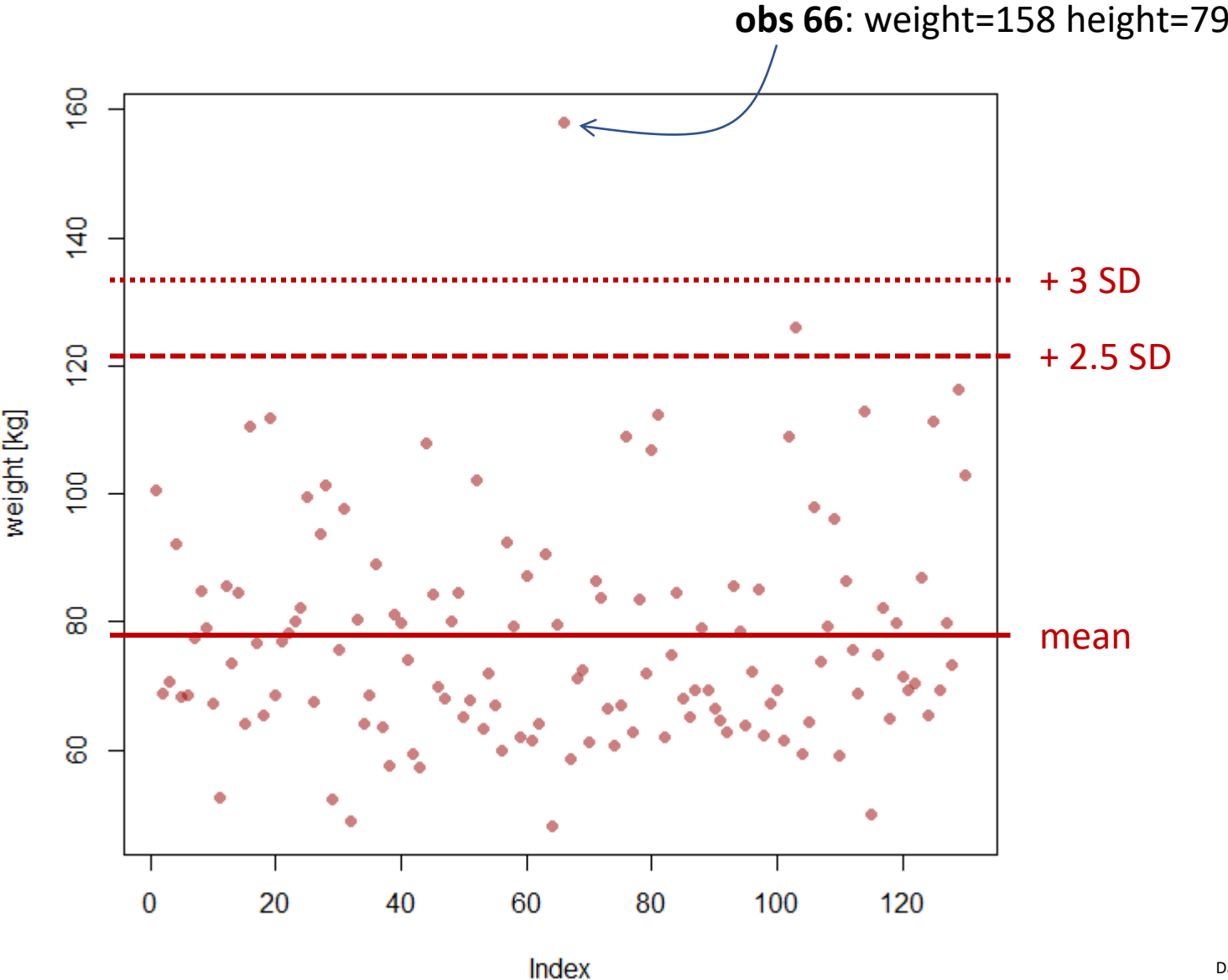
- Best way to find them:

- graphically!

...be very careful!

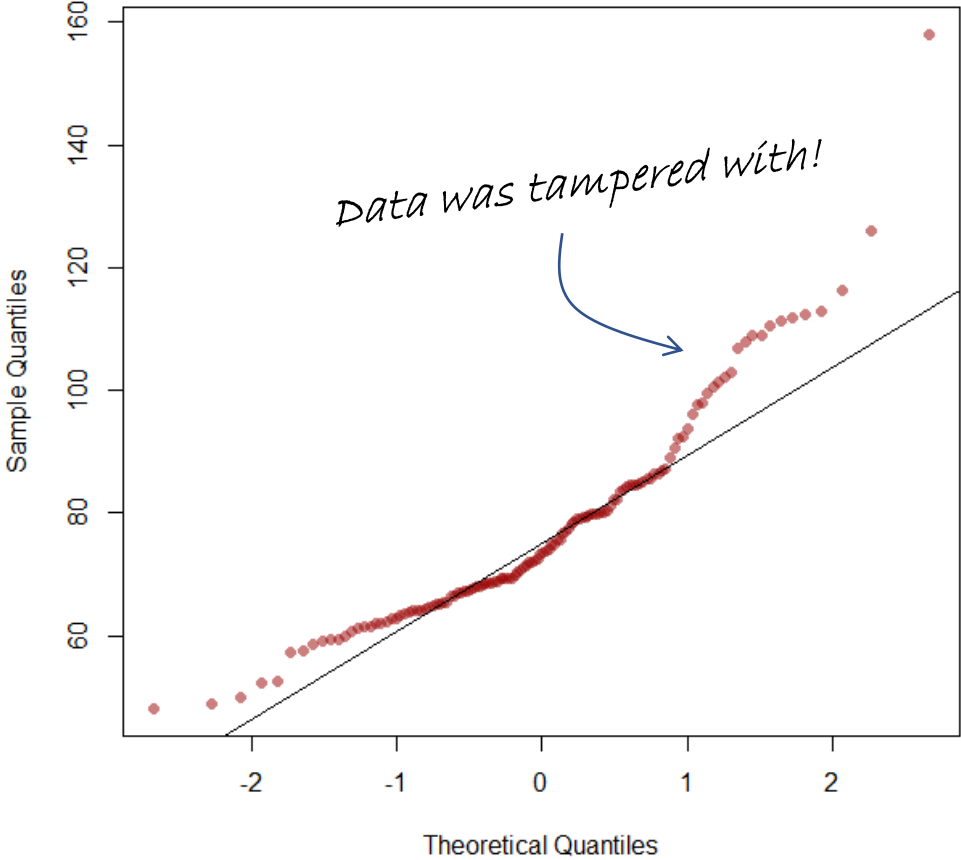


Outliers

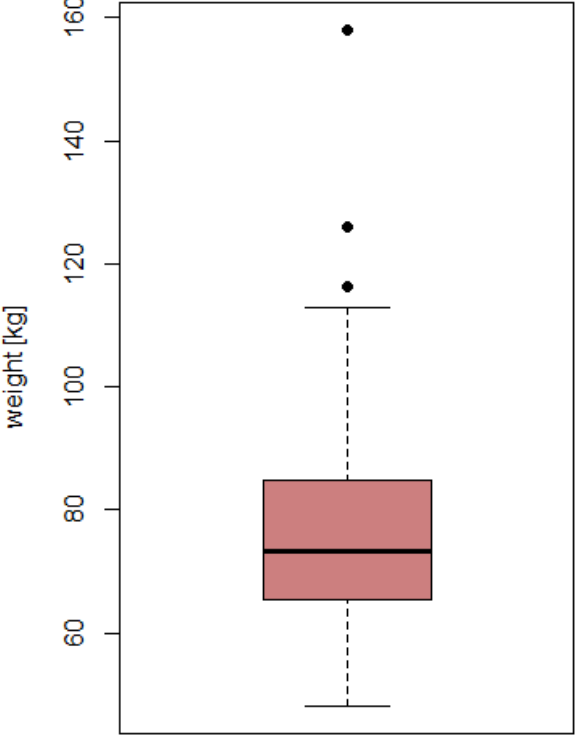


Outliers

QQ-plot



Boxplot



Outliers

Finally, a word of wisdom from a long-ago engineering colleague:
"Whenever I see an outlier, I'm never sure whether to throw it away
or patent it."

-- Berton Gunter (on outlier identification)
R-help (December 2009)

Summary

- Don't use bar plots!
- Always take a look at your data
- The Normal distribution is the single most important distribution of them all
- It's totally OK to transform data
- Outliers are not always evil, sometimes they are just influential observations