Scientific Inquiry
Part II: Experimental Design and Data Analysis

Charlotte Soneson & Michael Stadler

FMI
Friedrich Miescher Institute
for Biomedical Research

SIB

Who we are

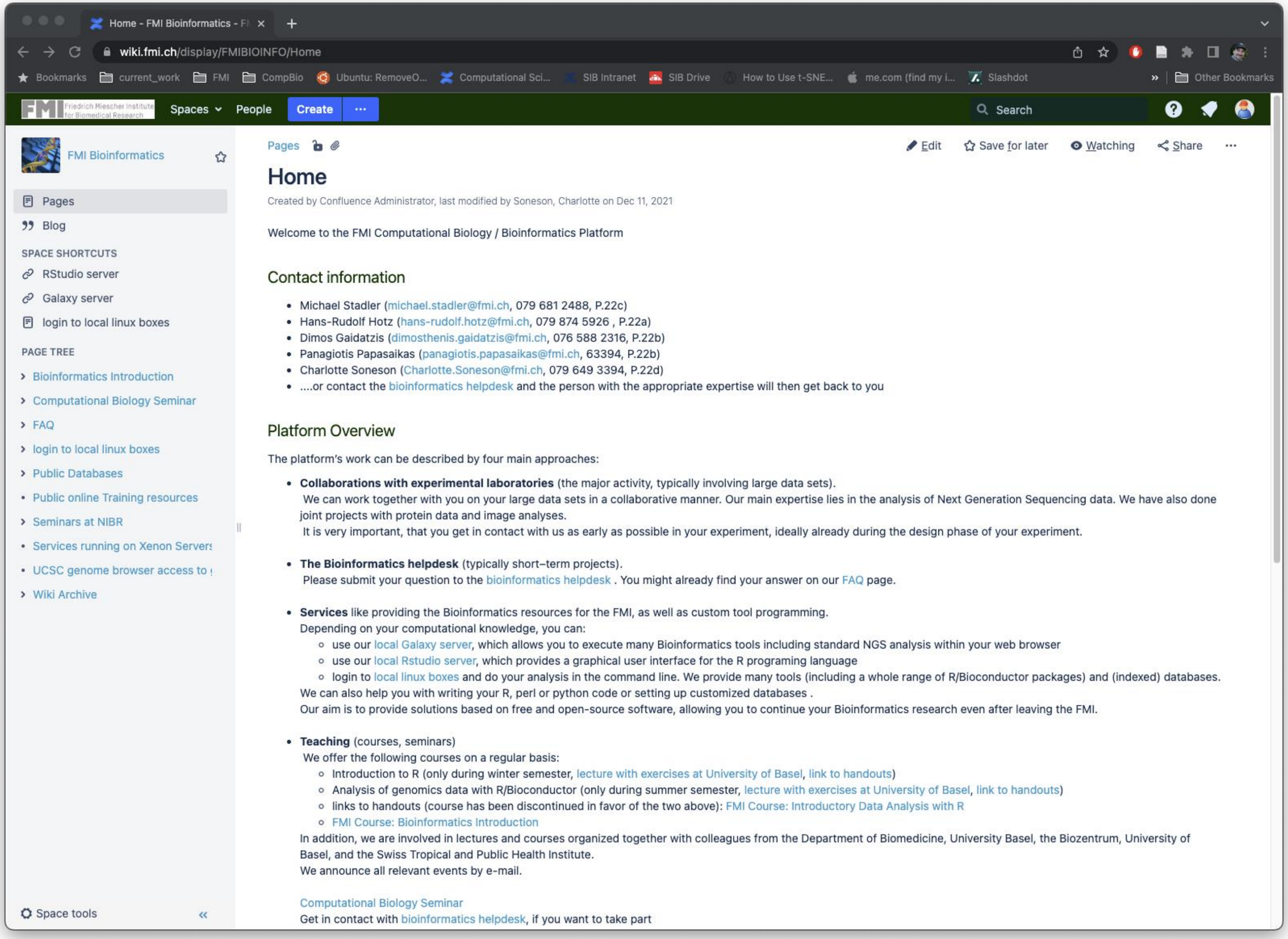
- Charlotte studied mathematics and computational biology (University of Lund, University of Zürich)
- Michael studied “wet” biology (University of Bern) and “dry” biology (University of Geneva, EPFL, MIT)

- We work in the FMI Computational Biology platform
  - Bioinformatics infrastructure (e.g. galaxy.fmi.ch, rstudio.fmi.ch, xenon7/8, software)
  - Teaching (R/Bioconductor, visual data exploration, experiment design)
  - Project collaborations (often NGS)

What we will not do today

- Learn in general how to design your experiments.
  → specific (FMI) courses, or come talk to us
- Learn how to analyze your data,
  discover patterns and test hypotheses.
  → specific UniBas lectures (spring/fall semester, 2 hours/week)
- Rather, we will:
  - Introduce relevant aspects of experimental design and data analysis
  - Aim: Raise your awareness, enable you to decide when to reach out

https://wiki.fmi.ch/display/FMIBIOINFO



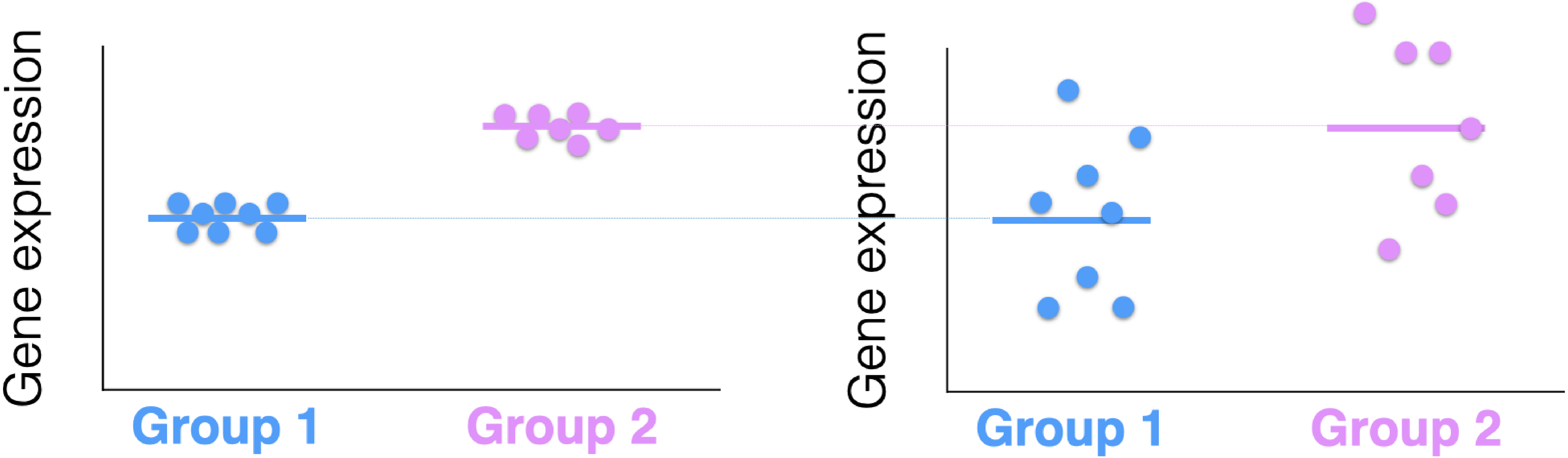
Episodes

1. Statistical significance
2. Multiple testing correction
3. Correlation and causation (or “Eat more chocolate!”)
4. Good practices for data visualization

Episodes

1. Statistical significance
2. Multiple testing correction
3. Correlation and causation (or “Eat more chocolate!”)
4. Good practices for data visualization

1. Statistical significance

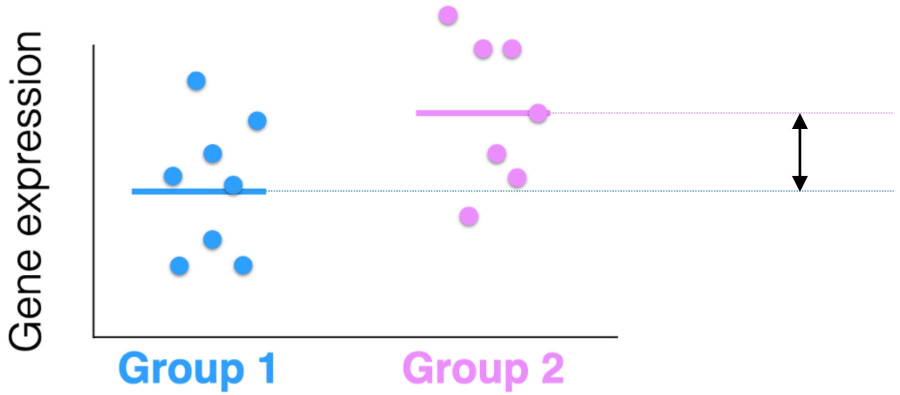


Which difference of means between Group 1 and Group 2
do you trust more (left or right)?

“Statistical relevance—relevant statistics, part I” doi: 10.15252/embj.201592958

1. What is a P value?

- A P value is the **probability** of observing **by chance** a measure **as extreme as the observed** one.



probability of getting such a difference when there is no difference of means

- The P value reported by tests is a probabilistic significance, not a biological one.
- With enough data, arbitrarily small effects become significant

“Significance, P values and t-tests” doi: 10.1038/nmeth.2698
“Power and sample size” doi: 10.1038/nmeth.2738

Episodes

1. Statistical significance
2. Multiple testing correction
3. Correlation and causation (or “Eat more chocolate!”)
4. Good practices for data visualization

2. Multiple testing correction

When we perform many tests, the number of p values that are significant **by chance** increases.

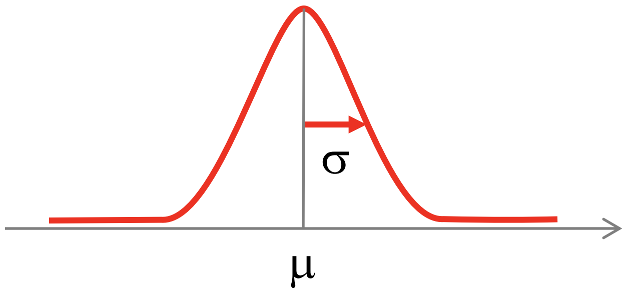
Thought experiment:

given
distribution

draw two
random samples

compare
samples

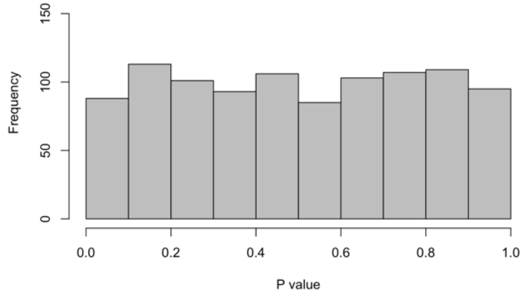
repeat many times;
distribution of p values?



sample 1:
-1.047 0.970 -1.798
-0.653 0.016 0.459

sample 2:
-0.370 -0.017 -0.479
-0.714 0.994 0.559

difference of
means?
(Student’s
t-test)



2. Multiple testing correction simulation in R

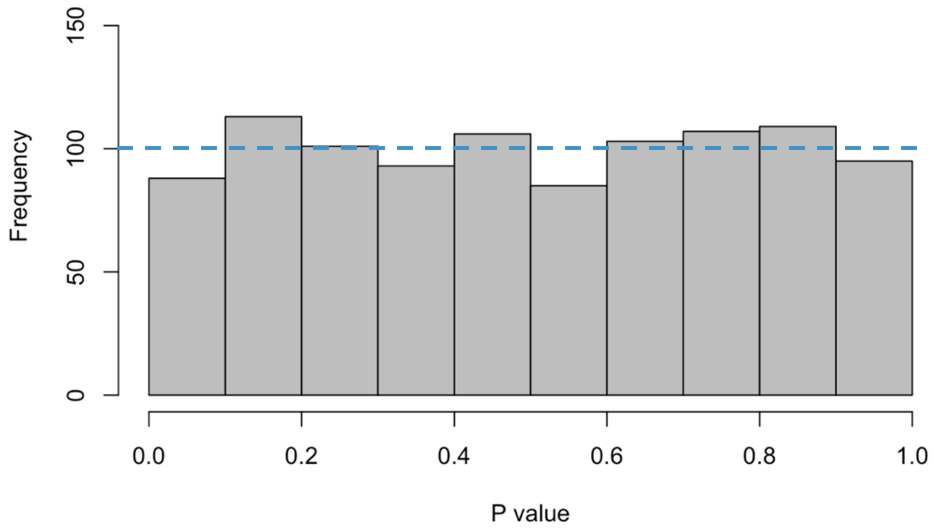
```r
x <- rnorm(100, mean = 0, sd = 1)
y <- rnorm(100, mean = 0, sd = 1)
t.test(x, y)$p.value  # 0.2653
```

```r
for (i in 1:1000) {
    x <- rnorm(100, mean = 0)
    y <- rnorm(100, mean = 0)
    pvals[i] <- t.test(x, y)$p.value
}

mean(pvals <= 0.1)  # 0.088 (~0.1)
padj <- p.adjust(pvals, method = "fdr")
mean(padj <= 0.1)   # 0
```

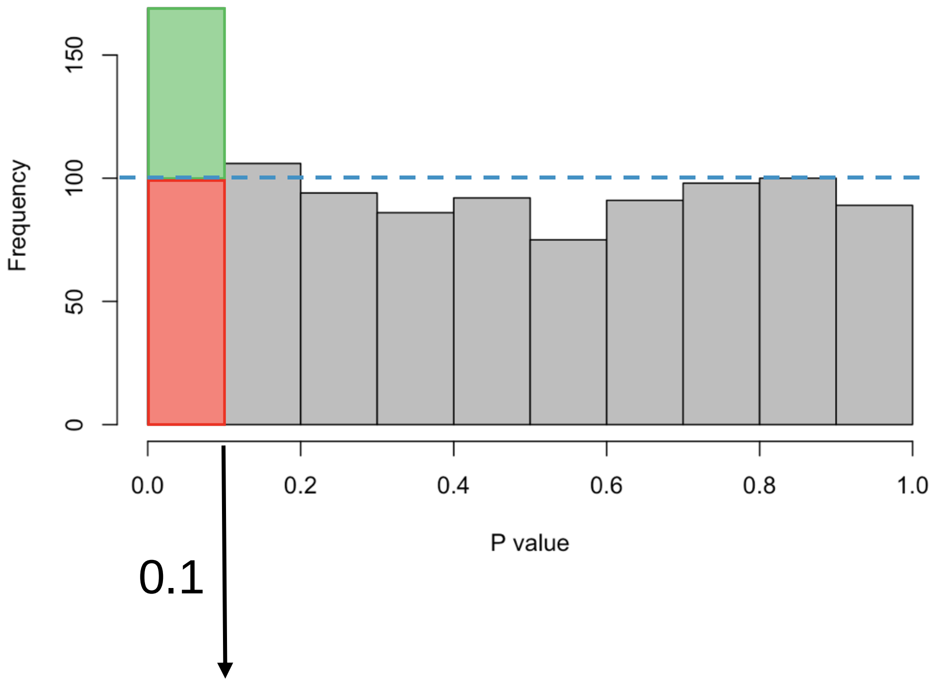
```r
y.mu <- rep(c(0, 0.4), c(900, 100))
for (i in 1:1000) {
    x <- rnorm(100, mean = 0)
    y <- rnorm(100, mean = y.mu[i])
    pvals[i] <- t.test(x, y)$p.value
}

mean(pvals <= 0.1)  # 0.169
padj <- p.adjust(pvals, method = "fdr")
mean(padj <= 0.1)   # 0.067
```

2. Multiple testing correction

histogram of P values (under H_0)



histogram of P values (10% true effect)



$$\text{false discovery rate (FDR)} = \frac{\square}{\square + \square} = \frac{100}{100 + 69} = 0.59$$

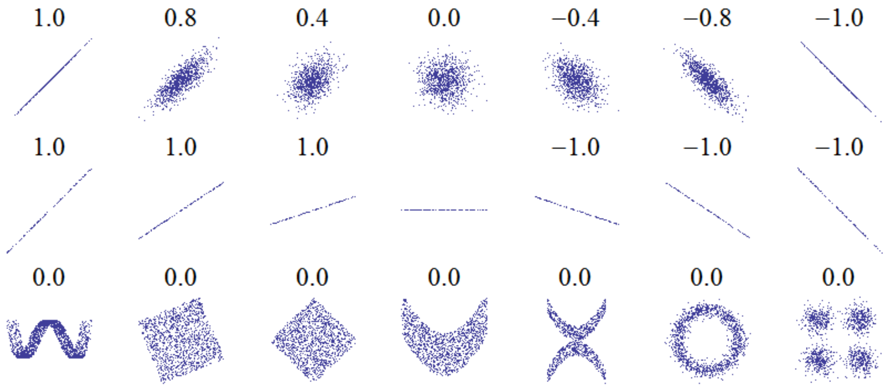
Episodes

1. Statistical significance
2. Multiple testing correction
3. Correlation and causation (or “Eat more chocolate!”)
4. Good practices for data visualization

3. Correlation and causation or “Eat more chocolate!”

“Correlation measures increasing/decreasing trends […].”

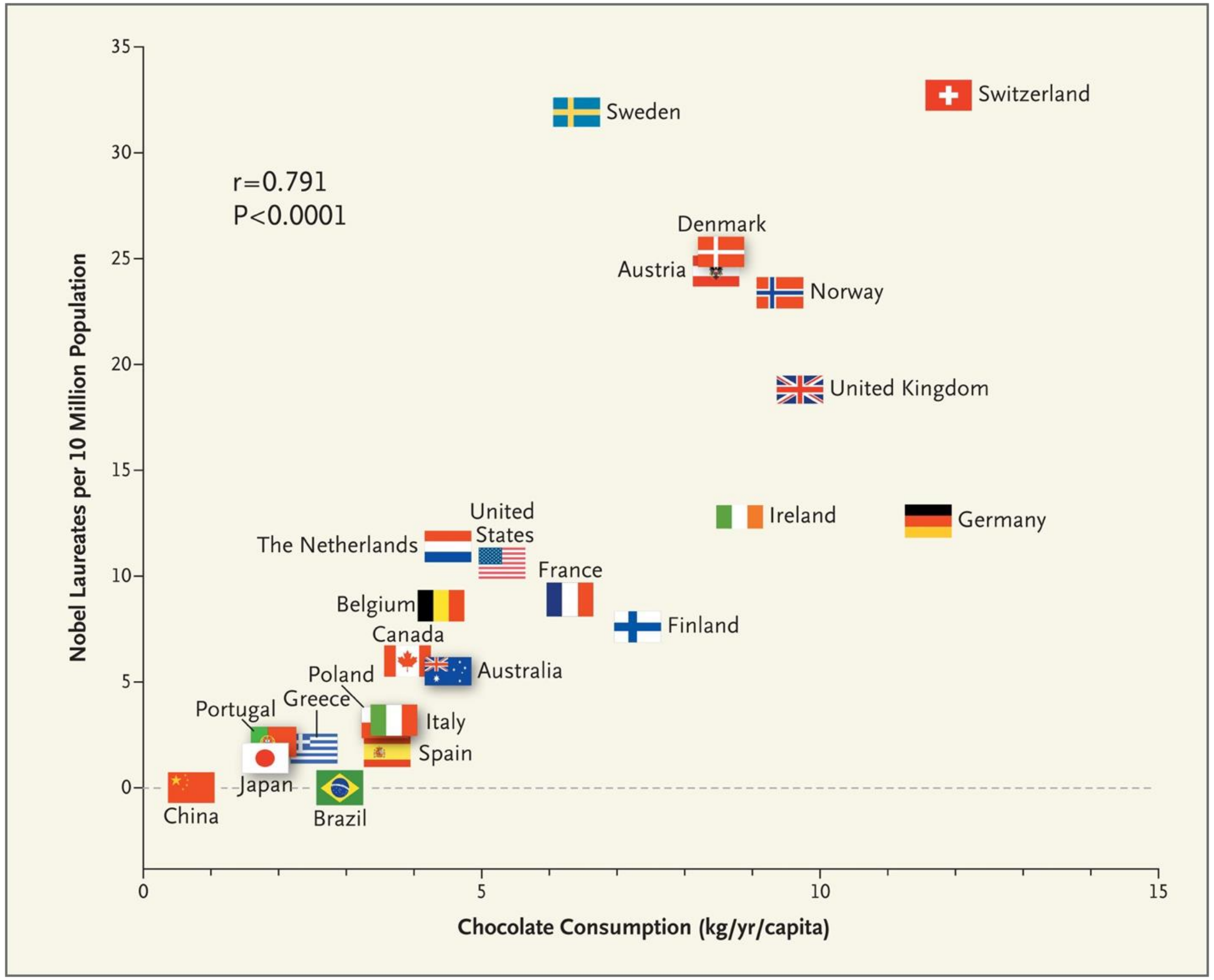
Pearson’s correlation coefficient r [-1, +1]
x, y: independent variables ⇒ $r_{xy} = 0$ (but not the inverse)



“Correlation implies association, but not causation.”

“Association, correlation and causation” doi: 10.1038/nmeth.3587
en.wikipedia.org/wiki/Correlation_and_dependence

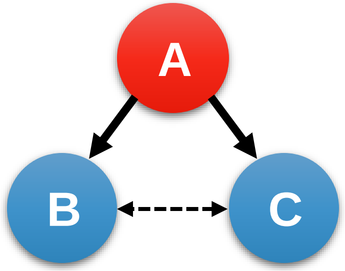
3. Eat more chocolate!

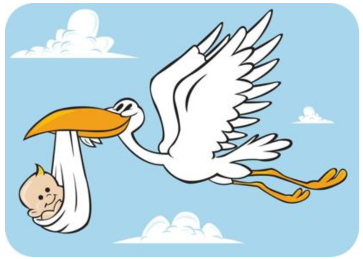


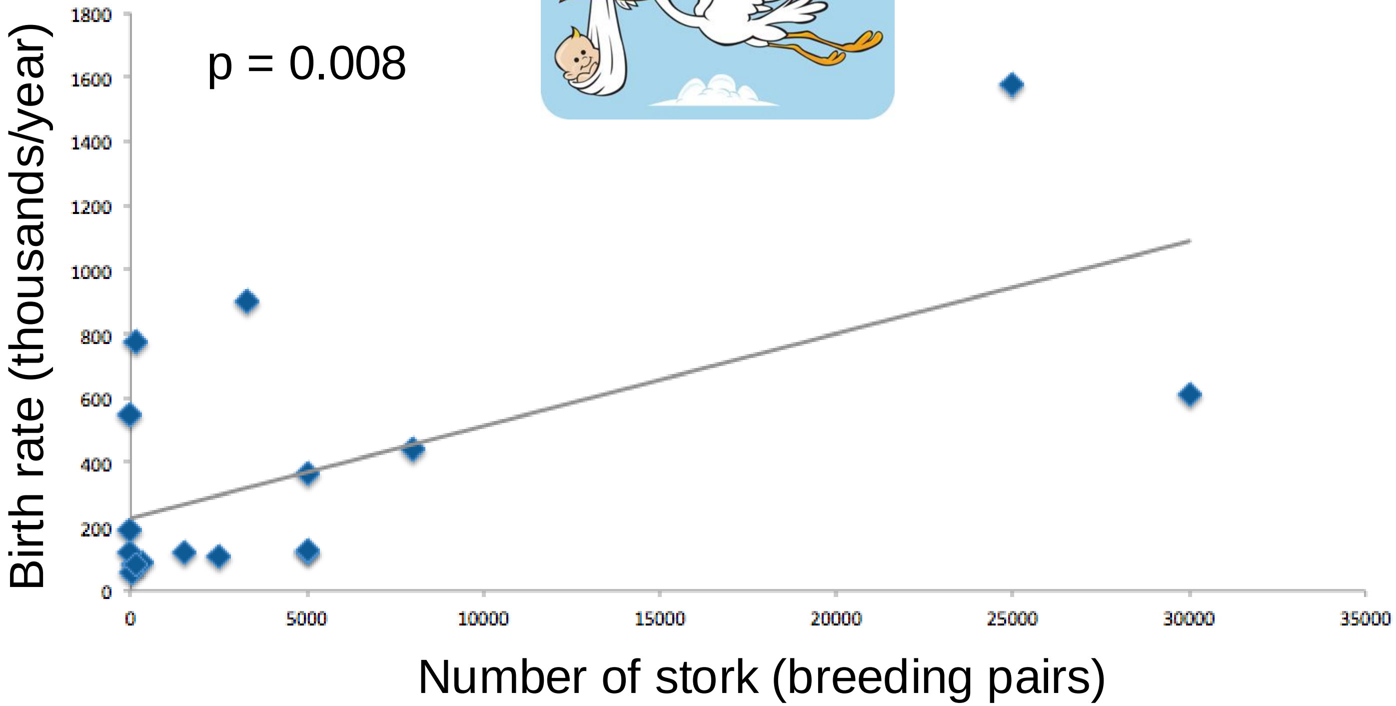
"The slope of the regression line allows us to estimate that it would take about 0.4 kg of chocolate per capita per year to increase the number of Nobel laureates in a given country by 1."

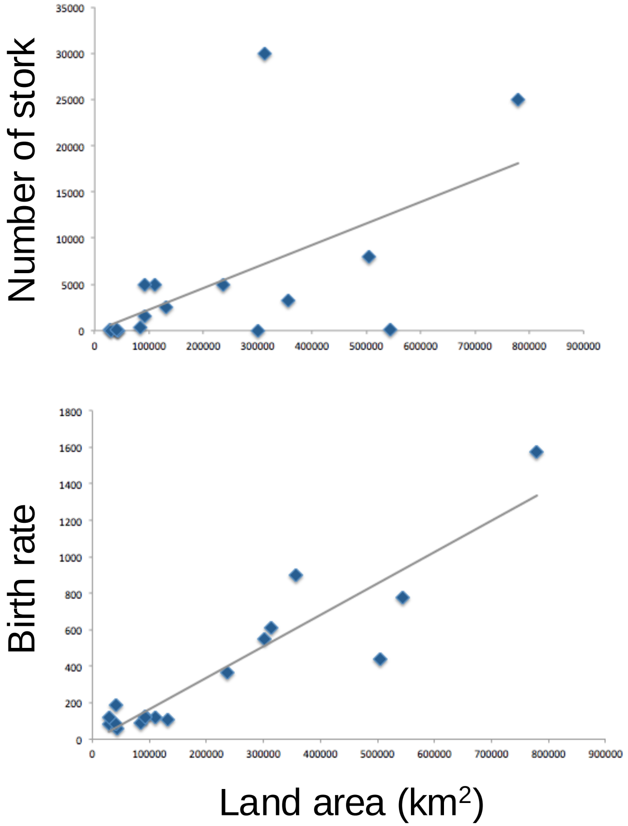
"Chocolate consumption, cognitive function and Nobel laureates" doi: 10.1056/NEJMon1211064

3. Do Storks deliver babies?



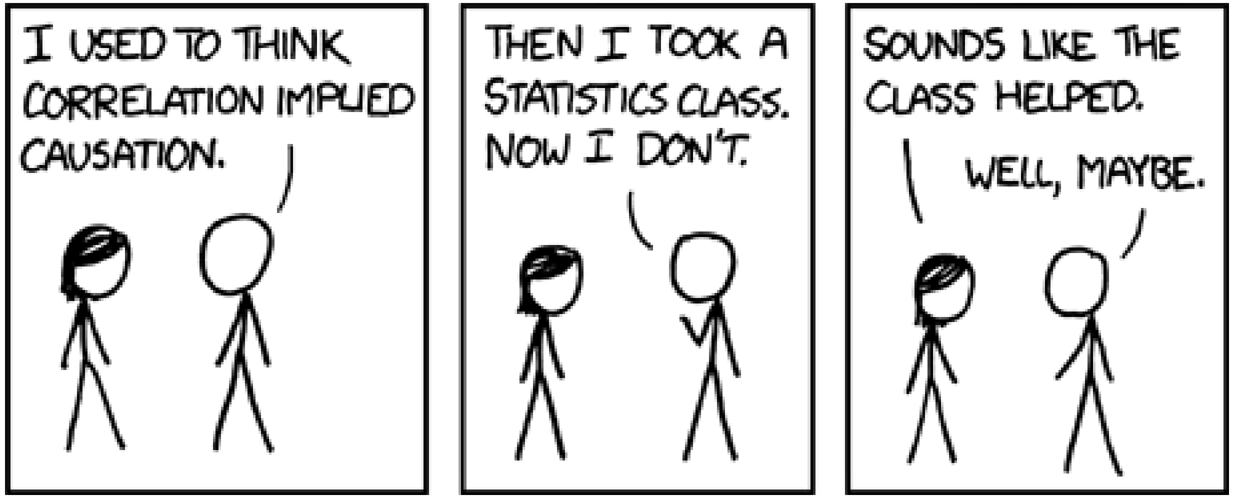






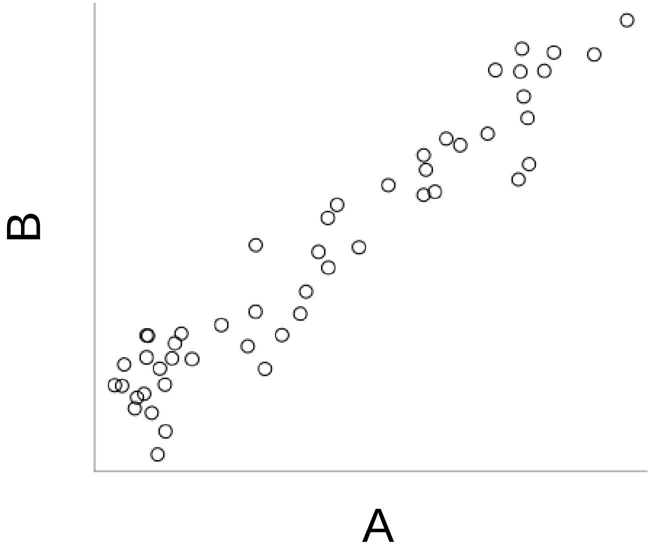
"Storks deliver babies (p=0.008)" doi: 10.1111/1467-9639.00013
priceonomics.com/do-storks-deliver-babies/

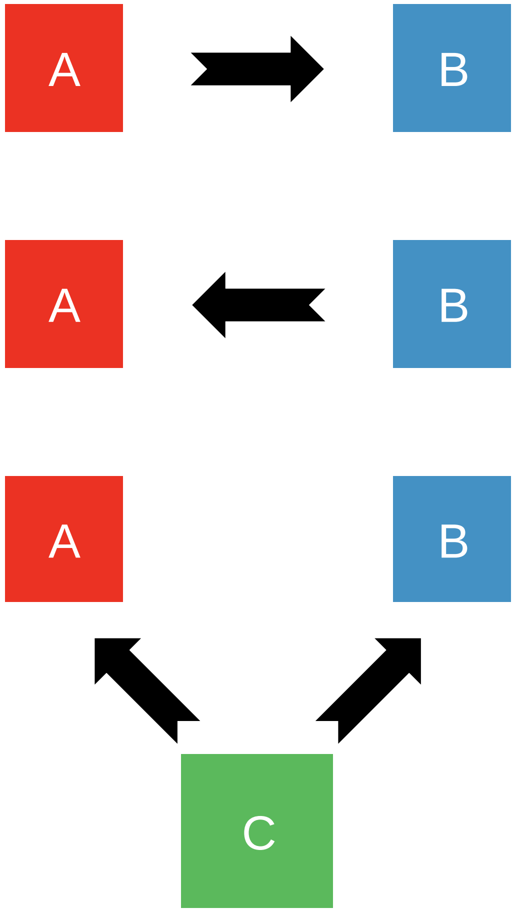
3. Interpreting correlations in general



xkcd.com/552/

3. Interpreting correlations in general





Episodes

1. Statistical significance
2. Multiple testing correction
3. Correlation and causation (or “Eat more chocolate!”)
4. Good practices for data visualization

4. Good practices for data visualization

A good visualization…

- has a clear message and is focused
- is easy to interpret (summarizes the data)
- is an honest and true reflection of the data

Visualizations are made up of…

- geometrical primitives (lines, points, …) that represent data sets
- scales (length, position, color, angle, …) that encode the data values

(“The grammar of graphics”, ggplot2)

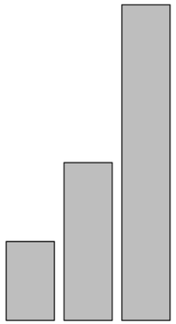
Encode the information with the most effective channel

“Statistical relevance—relevant statistics, part II: presenting experimental data” doi: 10.15252/embj.201694659
http://www.bioinformatics.babraham.ac.uk/training.html#figuredesign

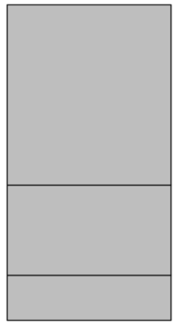
4. Scales encode data values

All the scales below encode the same data: (1, 2, 4)

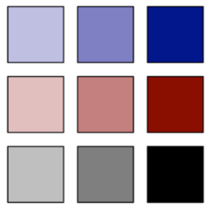
points (location)



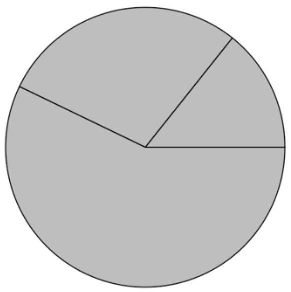
bars (height)



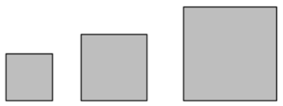
stacked bars (height)



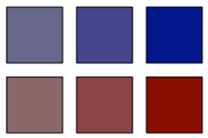
color (brightness)



angle



area

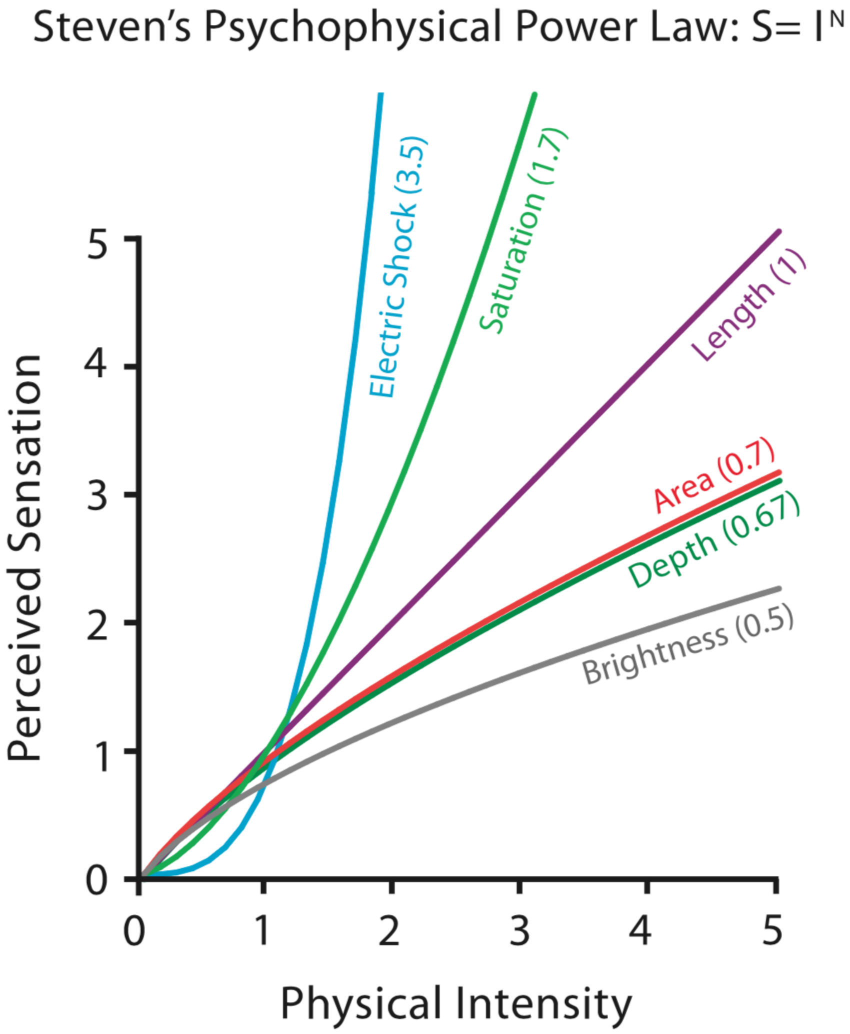


color (saturation)

Which one best represents the 1-2-4 ratios?

4. Human perception is not linear

… except for length

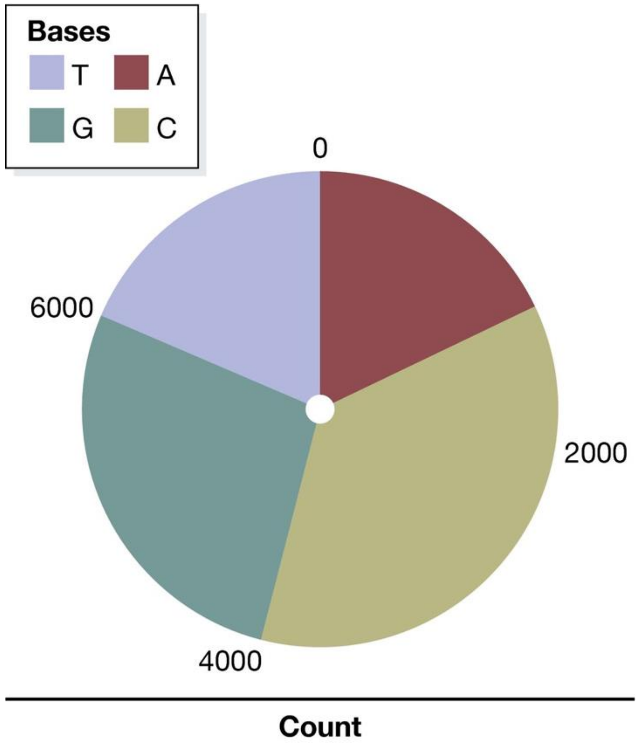


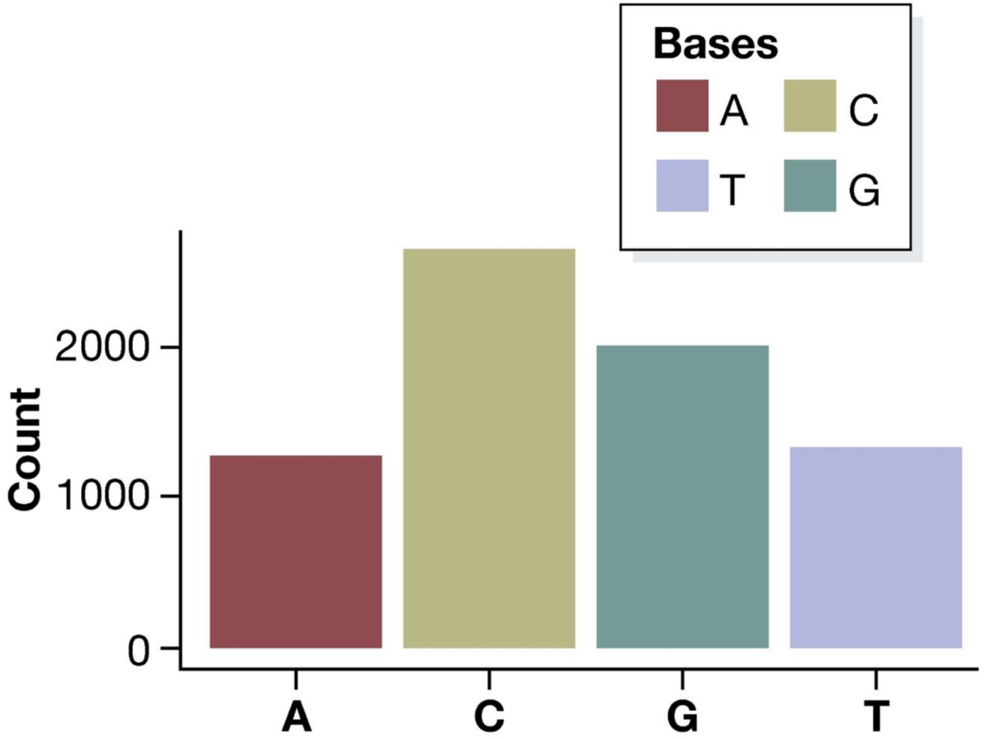
Stevens, S. S. (1957) PMID: 13441853
en.wikipedia.org/wiki/Stevens%27s_power_law
Munzner T. (2014) ISBN 9781466508910

4. What wrong with pie charts?

base composition in the Zyxin gene

A	1,285
C	2,635
G	2,013
T	1,332

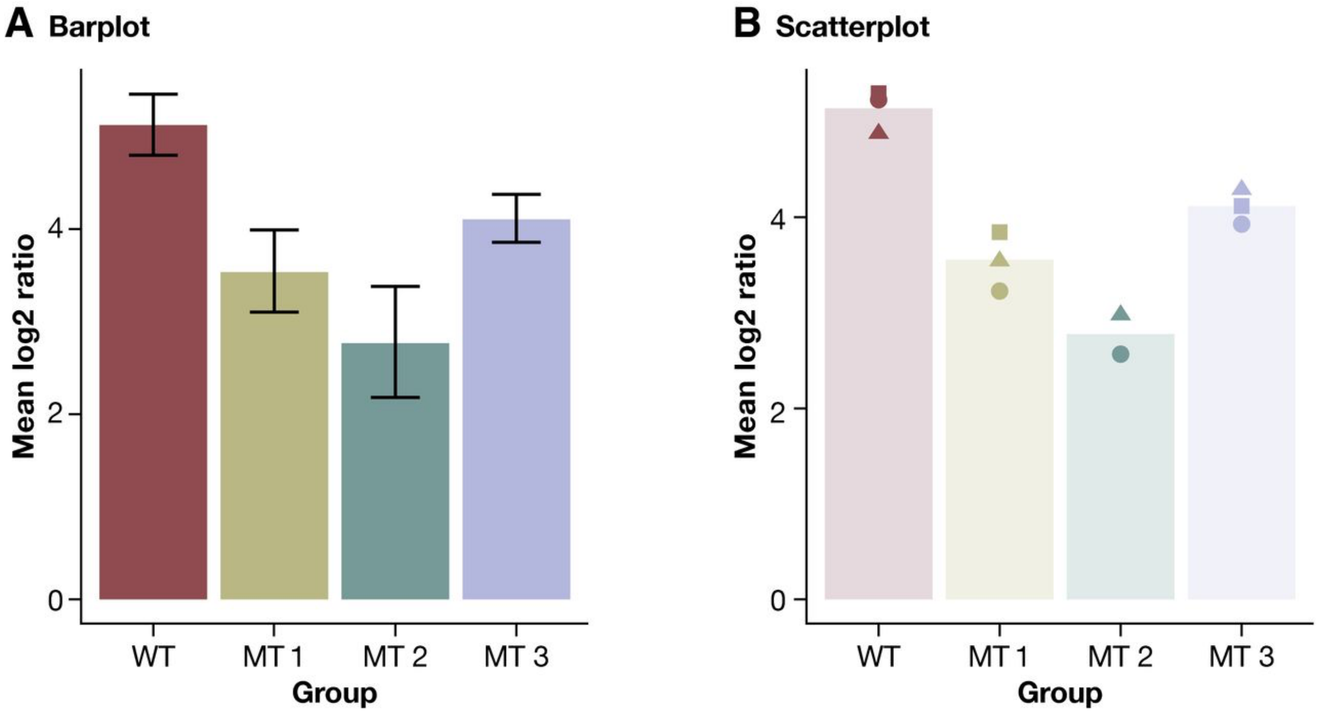




human perception: better to use length than area

“Statistical relevance—relevant statistics, part II: presenting experimental data” doi: 10.15252/embj.201694659

4. What wrong with bar plots?

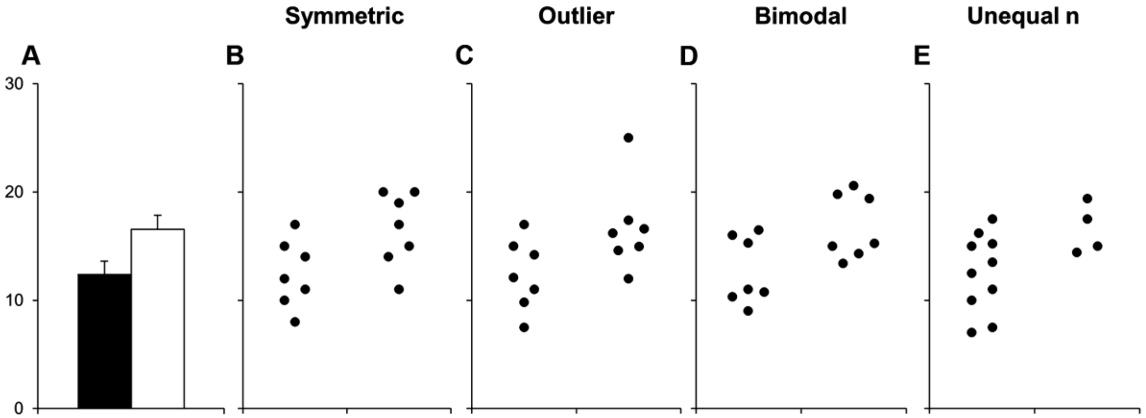


- barplot only show one number per group (e.g. mean)
- can be used to hide information (e.g. outlier, multimodal distribution)
- many types of error bars (s.d., s.e.m., 95%-confidence intervals)

"Statistical relevance—relevant statistics, part II: presenting experimental data" doi: 10.15252/embj.201694659

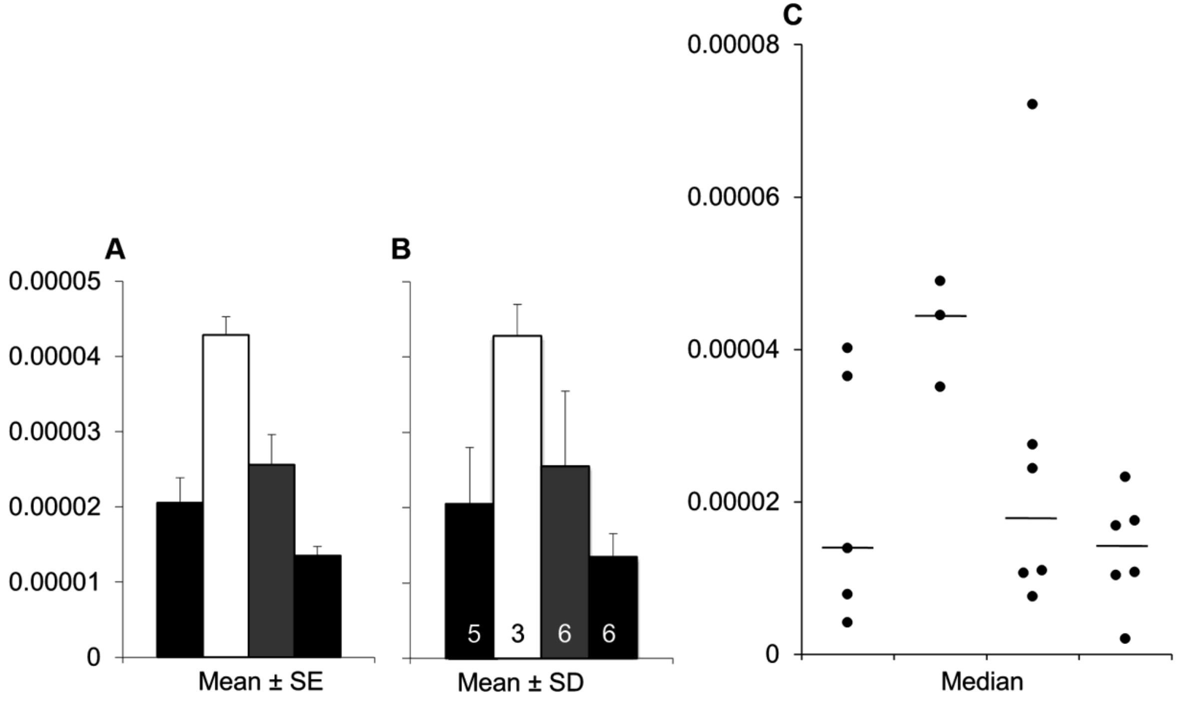
4. Examples of misleading bar plots

Various data produce the same barplot



Test	p value			
T-test: Equal var.	0.035	0.050	0.026	0.063
T-test: Unequal var.	0.035	0.050	0.026	0.035
Wilcoxon	0.054	0.073	0.128	0.103

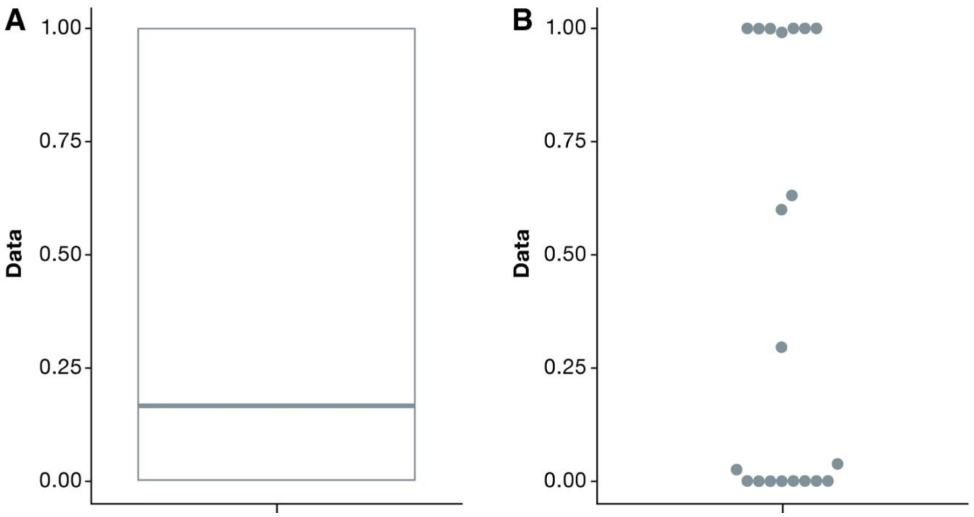
Placental endothelin 1 (EDN1) mRNA data



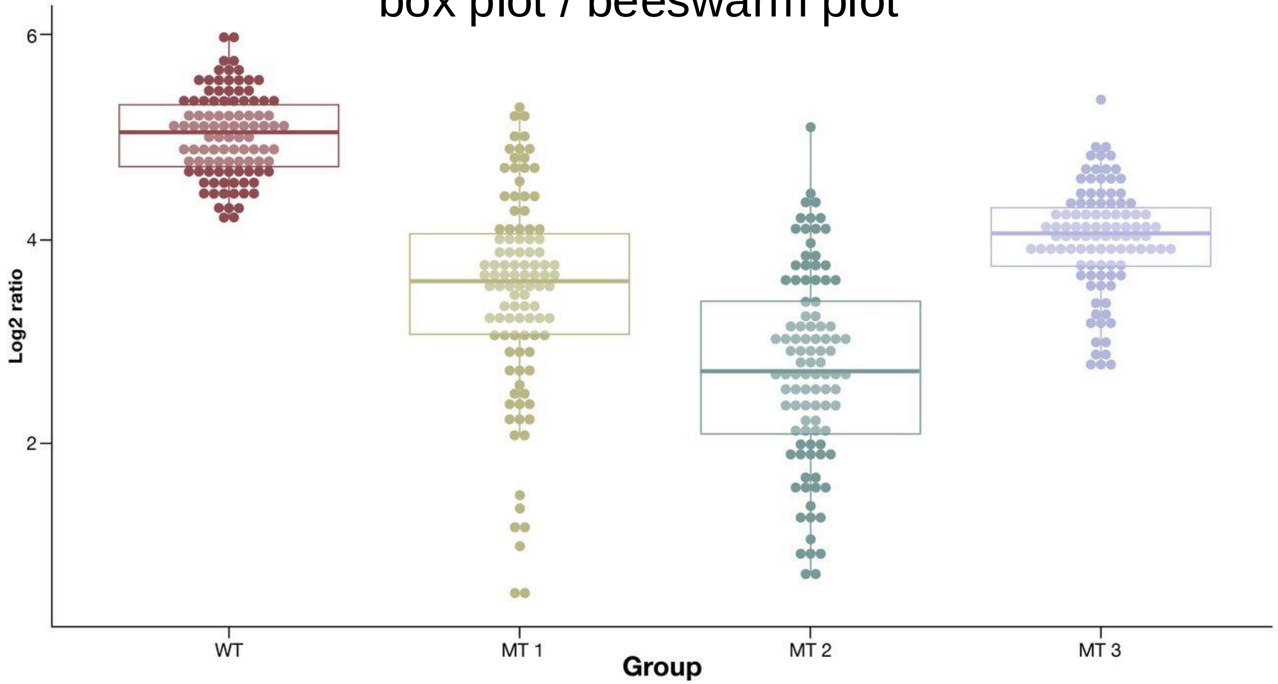
“Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm” doi: 10.1371/journal.pbio.1002128

4. Plots for more data points

boxplot may also hide information



box plot / beeswarm plot

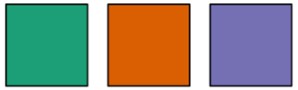


“Statistical relevance—relevant statistics, part II: presenting experimental data” doi: 10.15252/embj.201694659

4. On the use of colors

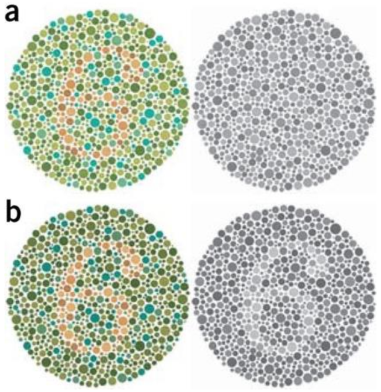
- poor in encoding quantitative data

- great to represent categorical data



- don’t forget the colorblind

Ishihara color-vision test



Colors optimized for color-blind individuals

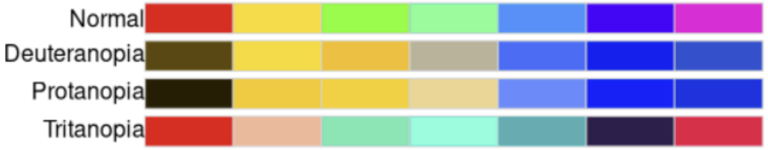
Color	Color name	RGB (1–255)	CMYK (%)	P	D
	Black	0, 0, 0	0, 0, 0, 100		
	Orange	230, 159, 0	0, 50, 100, 0		
	Sky blue	86, 180, 233	80, 0, 0, 0		
	Bluish green	0, 158, 115	97, 0, 75, 0		
	Yellow	240, 228, 66	10, 5, 90, 0		
	Blue	0, 114, 178	100, 50, 0, 0		
	Vermillion	213, 94, 0	0, 80, 100, 0		
	Reddish purple	204, 121, 167	10, 70, 0, 0		

P and D indicate simulated colors as seen by individuals with protanopia (red-) and deuteranopia (green-).

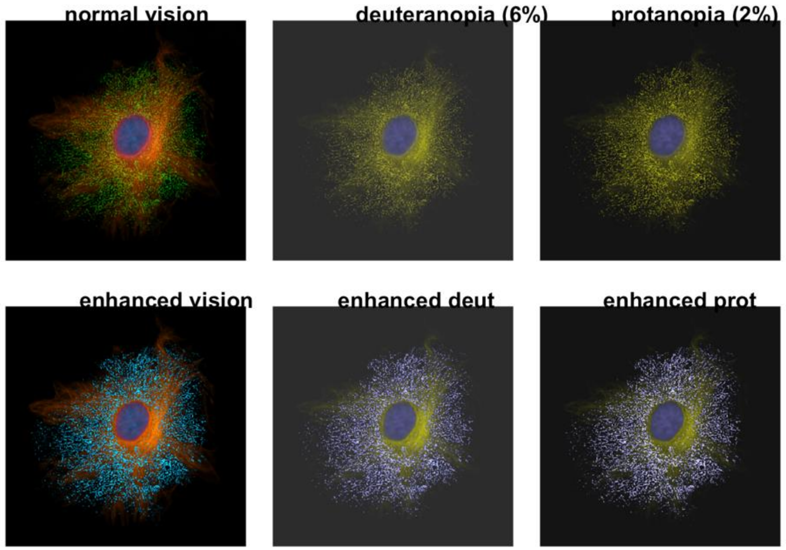
nowosad.github.io/colorblindcheck/

```
library(colorblindcheck)
rainbow_pal = rainbow(n = 7)
rainbow_pal
#> [1] "#FF0000" "#FFDB00" "#49FF00" "#00FF92" "#0092FF" "#4900FF" "#FF00DB"
```

```
palette_check(rainbow_pal, plot = TRUE)
```




```
#>          name n tolerance ncp ndcp  min_dist mean_dist  max_dist
#> 1      normal 7  12.13226  21   21 12.132257  61.06471 107.63470
#> 2 deuteranopia 7  12.13226  21   18  7.725825  50.11732  91.56339
#> 3  protanopia 7  12.13226  21   19  2.355309  55.41310  88.34820
#> 4  tritanopia 7  12.13226  21   19  8.216194  51.53678  83.10000
```



“Points of view: Color blindness” doi: 10.1038/nmeth.1618

colorbrewer2.org   colororacle.org (app for macOS/Windows/Linux)

cran.r-project.org/package=colorBlindness

Further information

- Nature Methods “Points of Significance”, web collection: https://www.nature.com/collections/qghhqm/pointsofsignificance
- Klaus B. Statistical relevance – relevant statistics, part I. EMBO J. 2015;34(22):2727-30. doi: 10.15252/embj.201592958.
- Klaus B. Statistical relevance – relevant statistics, part II. EMBO J. 2016;35(16):1726-9. doi: 10.15252/embj.201694659.
- Scientific Figure Design (1-day-course, Babraham). http://www.bioinformatics.babraham.ac.uk/training.html#figuredesign
- Munzner T. Visualization Analysis and Design. 2014 by A. K. Peters/CRC Press. ISBN 9781466508910