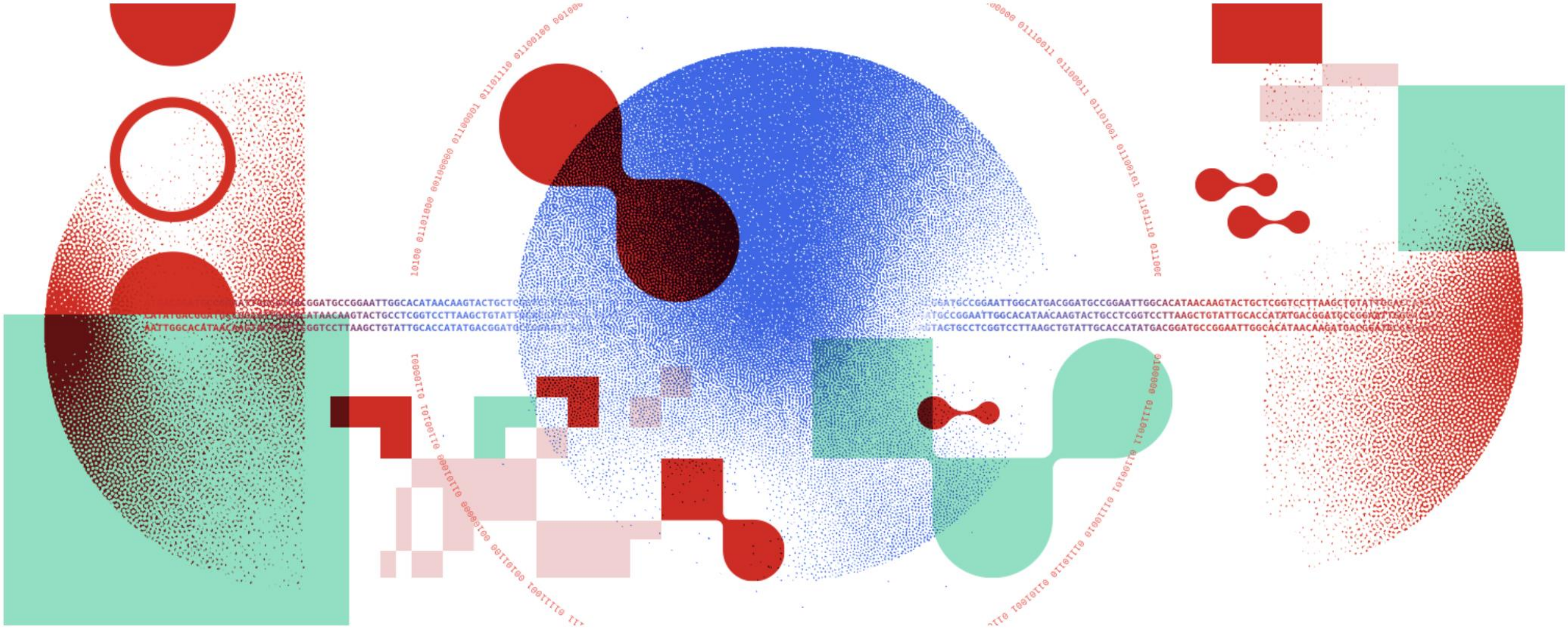


Introduction to Statistics and Data Visualisation with R

Lausanne, January 2026

Joao Lourenço and Rachel Marcone

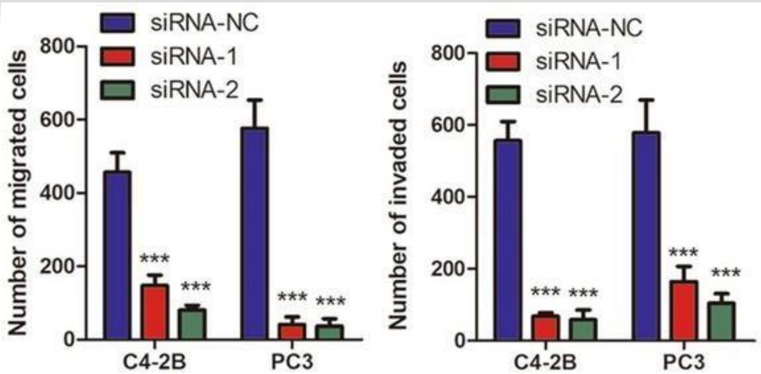
Graphics and summary

Summary and visualisation of data

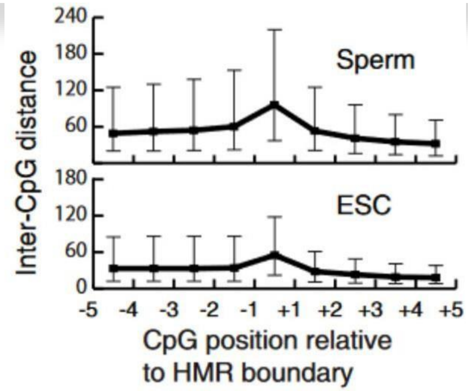
Learning objectives:

- Summarise your data
- Learn about different graphics (which ones do you know ?)
- Learn about error bars and difference with confidence intervals

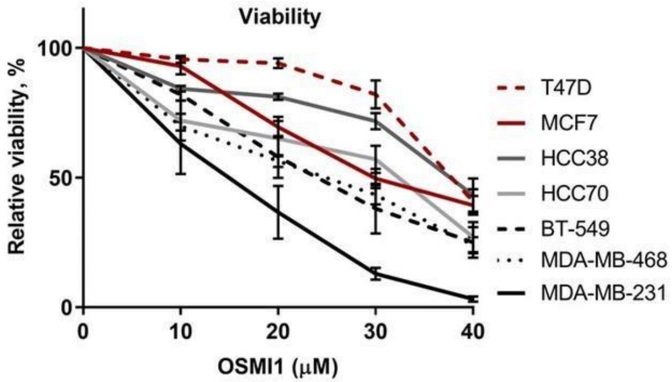
Error bars are ubiquitous in the scientific literature



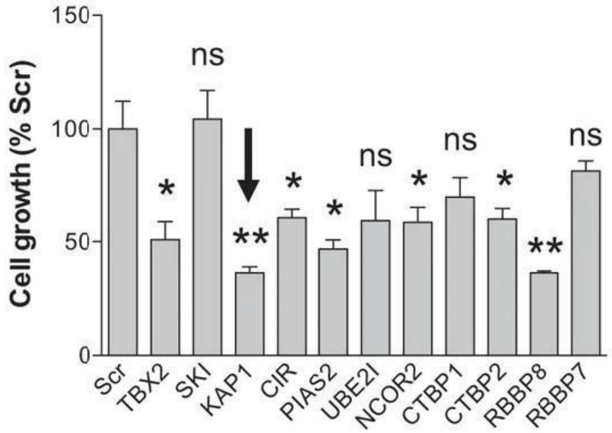
Cao et al. 2021 Cell Death & Disease



Molaro et al. 2011 Cell

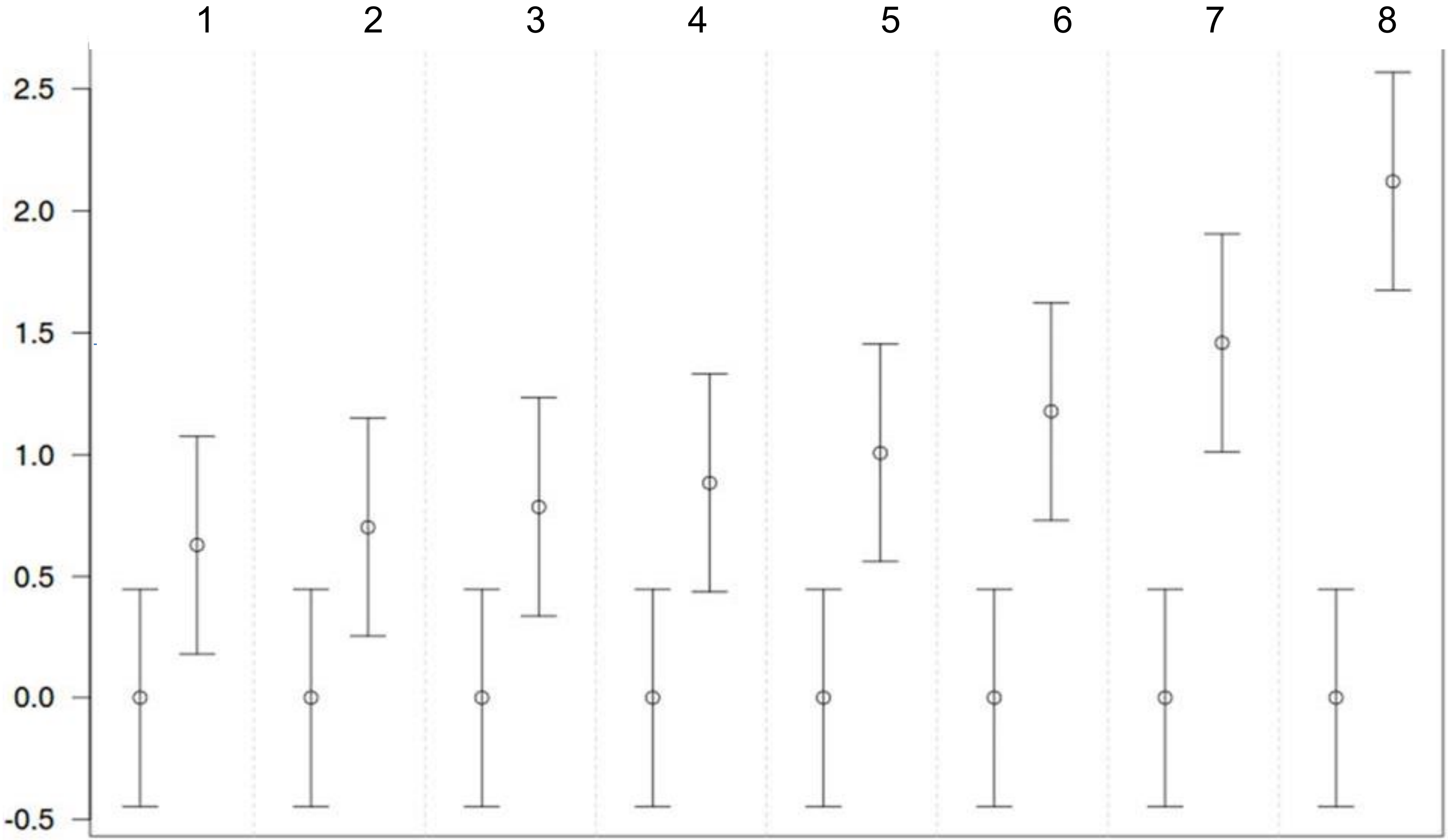


Barkovskaya et al. 2019 Scientific Reports

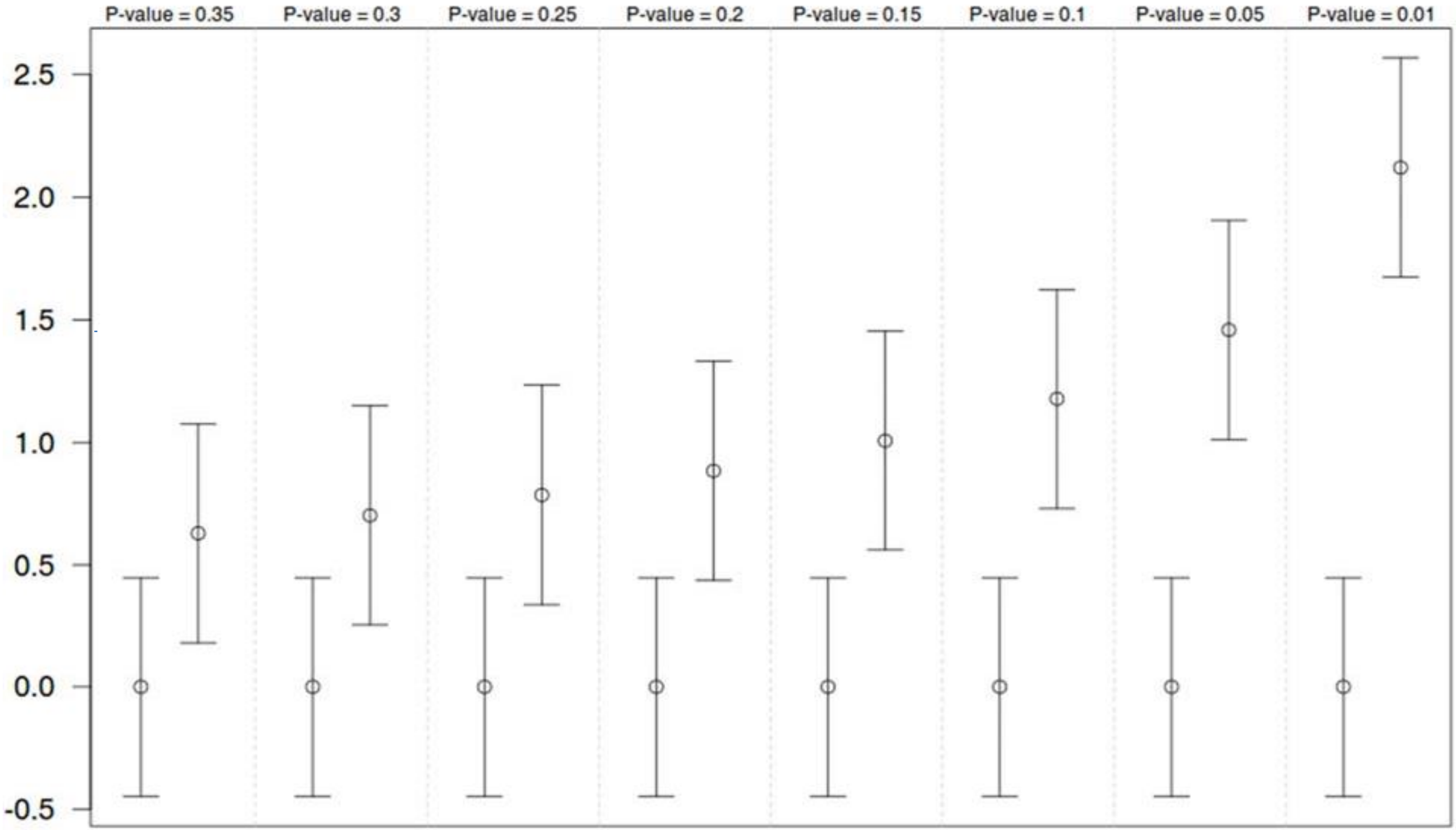


Crawford et al. 2019 Oncogene

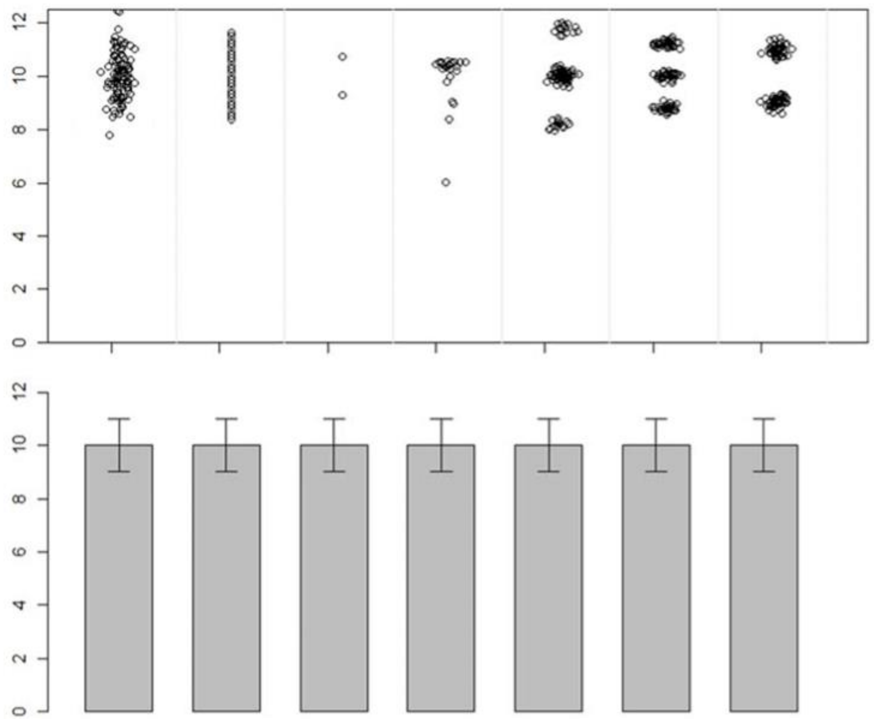
Quiz: When is it significant



Quiz: When is it significant



Be aware of error bars, hiding the data!



	Counts of articles by error bar types				
Journals	SD	SEM	Others*	Unidentified	Total counts†
Science	20	29	15	7	71
Nature	43	47	19	5	114
Cell	30	34	4	3	71
New England Journal of Medicine	0	4	9	2	15
Journal of the American Medical Association	0	2	14	0	16
The Lancet	1	1	17	2	21

SD = standard deviation, SEM = standard error of the mean.

* Other measures shown as error bars.

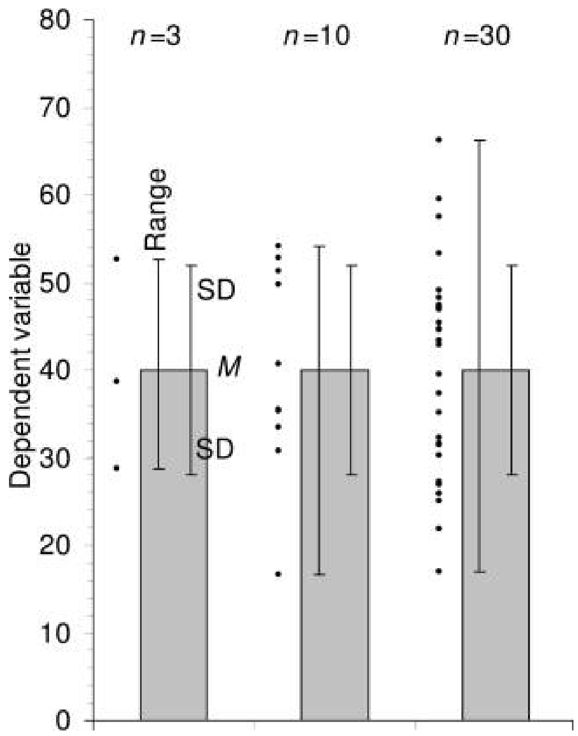
† These data represent the total number of articles that appeared in the publication during the review period that used error bars in figures. The articles using 2 or more types of error bars were counted in each category but only once in the total category.

Error bars are ubiquitous in the scientific literature

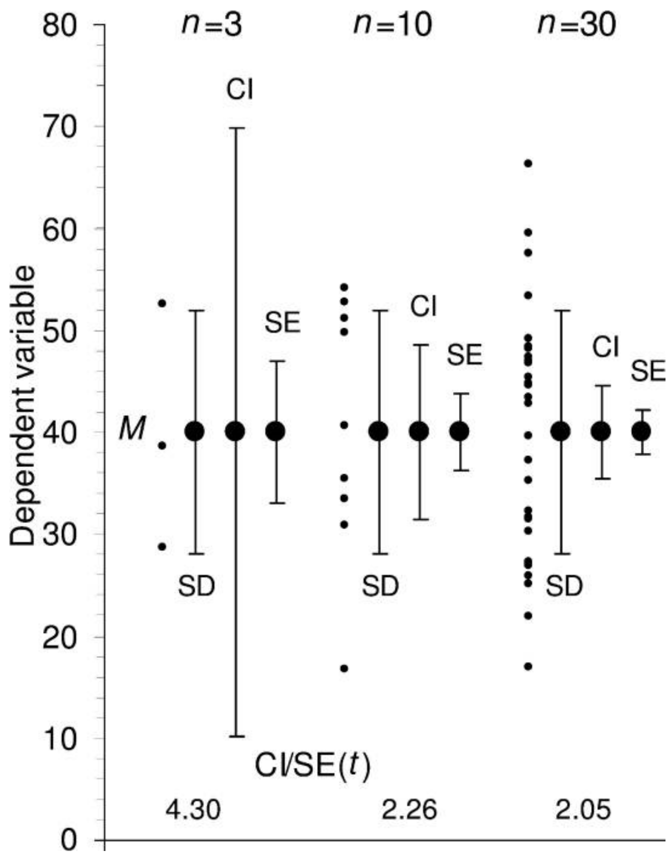
- Counts of articles by types of error bars published in representative scientific journals
- from January 1, 2019 to March 31, 2019.

Error bar	Type	Description	Formula
Range	Descriptive	Amount of spread between the extremes of the data	Highest data point minus the lowest
Standard deviation (SD)	Descriptive	Typical or (roughly speaking) average difference between the data points and their mean	$\mathrm{SD} = \sqrt{\frac{\sum(X - M)^2}{n-1}}$
Standard error of the mean (SEM)	Inferential	A measure of how variable the mean will be, if you repeat the whole study many times	$SEM = \frac{SD}{\sqrt{n}}$
Confidence interval (CI), usually 95% CI	Inferential	A range of values you can be 95% confident contains the true mean	$M \pm t_{(n-1)} \times$ SEM, where $t_{(n-1)}$ is a critical value of t. If n is 10 or more, the 95% CI is approximately $M \pm 2 \times$ SEM.

Error bars are ubiquitous in the scientific literature

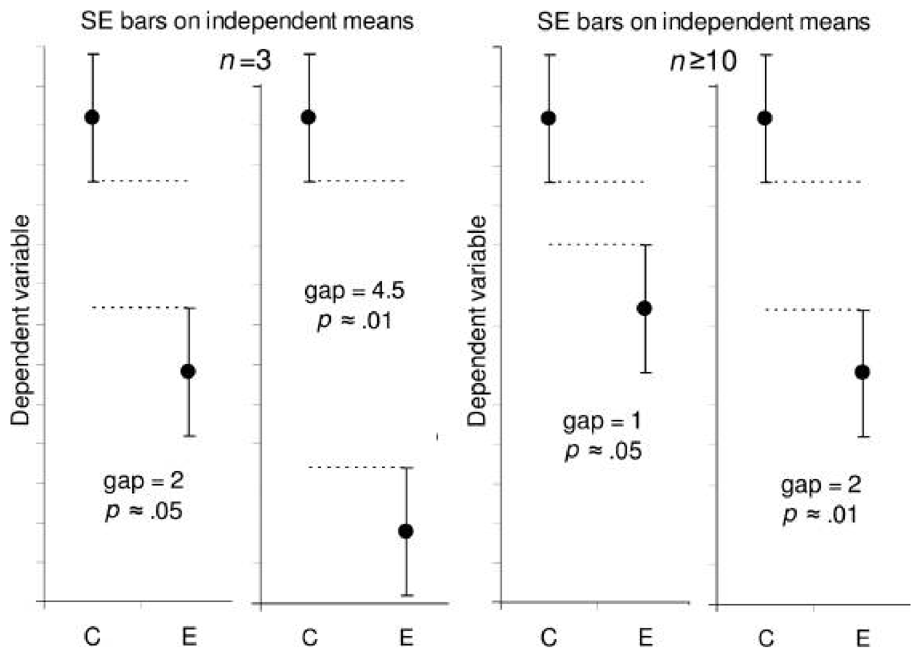


Descriptive error bars

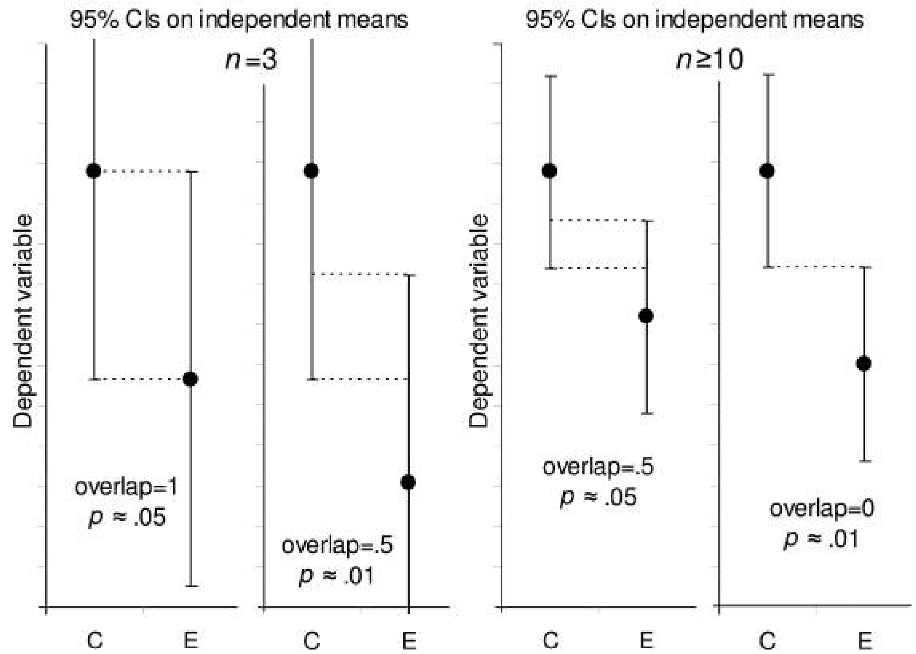


Inferential error bars

Error bars are ubiquitous in the scientific literature



Estimating statistical significance using the overlap rule for SE bars



Estimating statistical significance using the overlap rule for 95% CI bars

Error bars are ubiquitous in the scientific literature

> Psychol Methods. 2005 Dec;10(4):389-96. doi: 10.1037/1082-989X.10.4.389.

Researchers misunderstand confidence intervals and standard error bars

Sarah Belia [1], Fiona Fidler, Jennifer Williams, Geoff Cumming

Affiliations + expand

PMID: 16392994 DOI: 10.1037/1082-989X.10.4.389

Abstract

Little is known about researchers' understanding of confidence intervals (CIs) and standard error (SE) bars. Authors of journal articles in psychology, behavioral neuroscience, and medicine were invited to visit a Web site where they adjusted a figure until they judged 2 means, with error bars, to be just statistically significantly different ($p < .05$). Results from 473 respondents suggest that many leading researchers have severe misconceptions about how error bars relate to statistical significance, do not adequately distinguish CIs and SE bars, and do not appreciate the importance of whether the 2 means are independent or come from a repeated measures design. Better guidelines for researchers and less ambiguous graphical conventions are needed before the advantages of CIs for research communication can be realized.

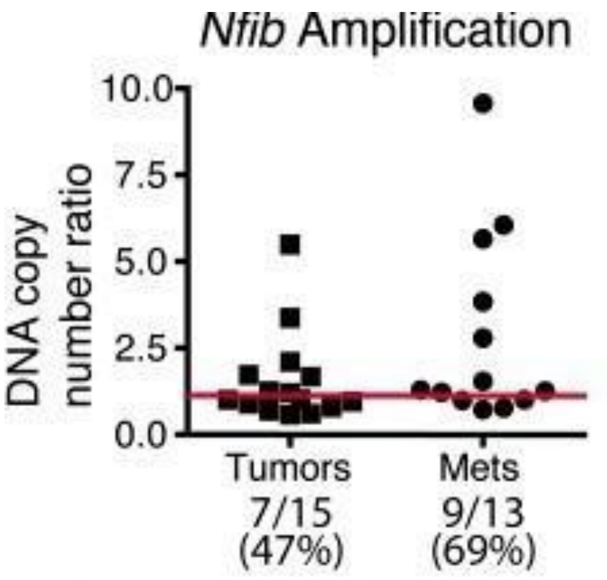
Error bars are ubiquitous in the scientific literature

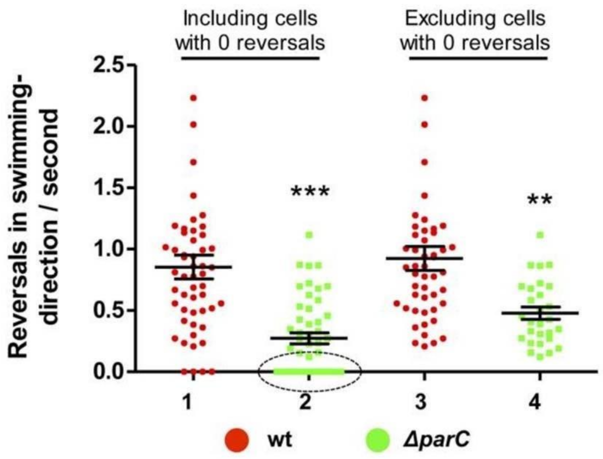
Take home message on error bars

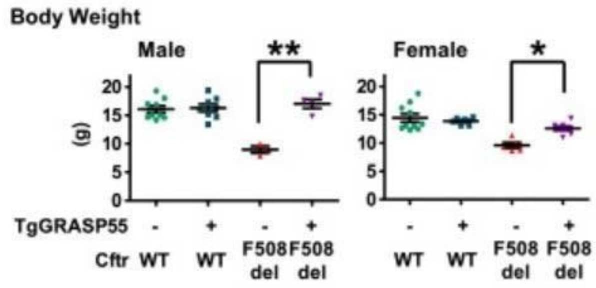
- **SD** is a variation among **individual data points in your data**
- **SEM** is a measure of the uncertainty in the **estimated mean in your data**
- **CI** is a **range**, based on the SEM and the statistics, telling you where the true mean likely is

- Avoid error bars if possible
- If you have to use them, document them, and try not to use them alone.

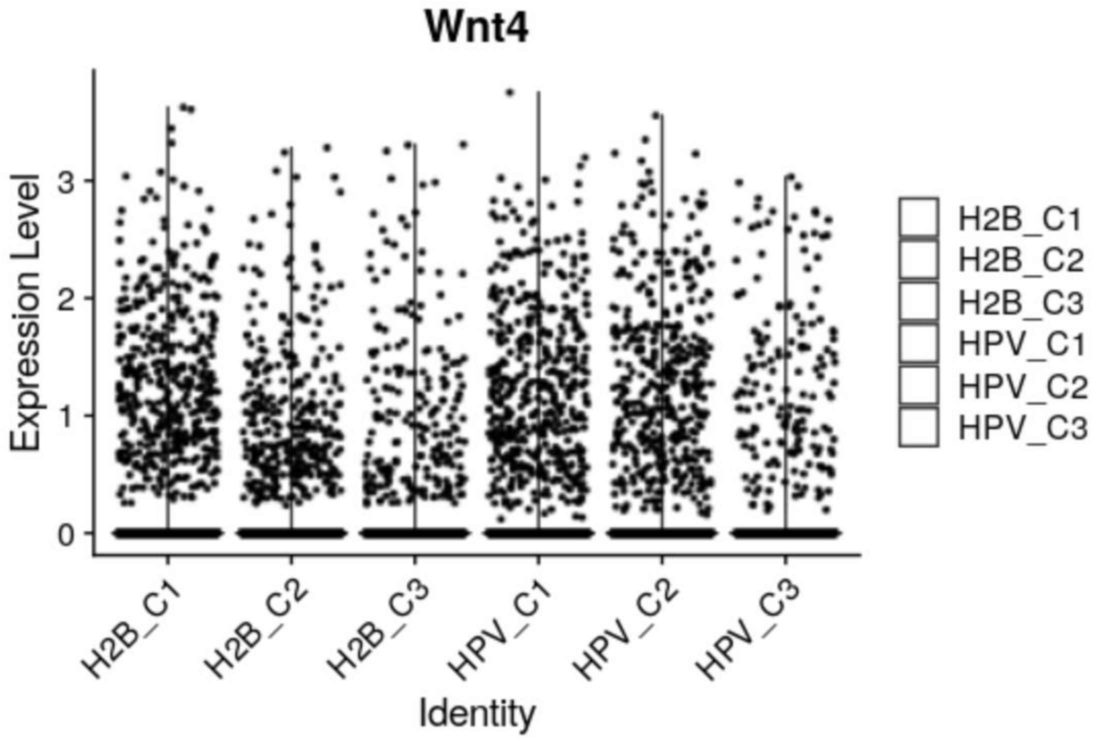
- What are the alternatives ?



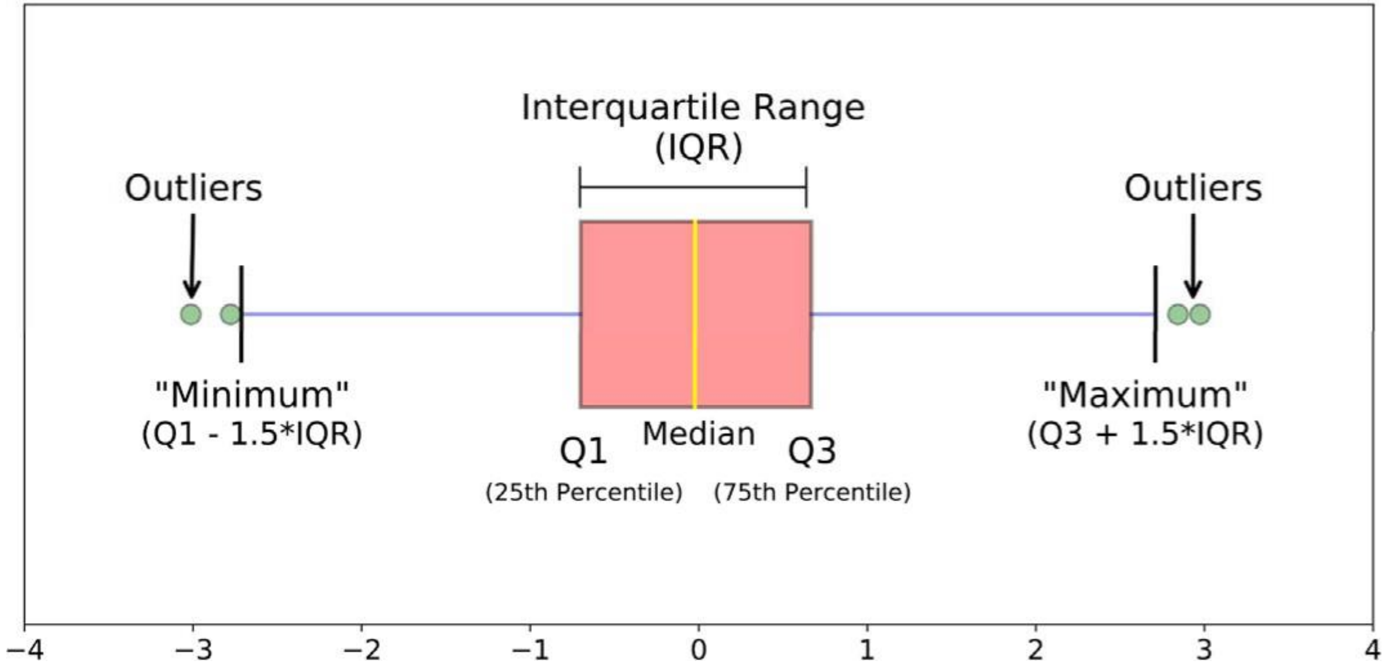


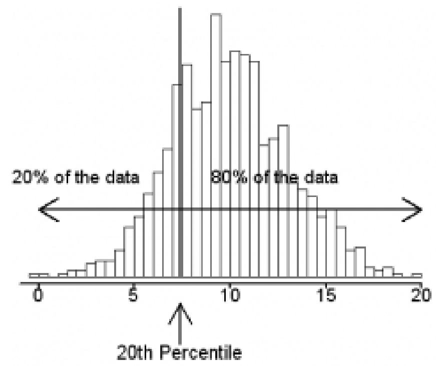


Alternative: show your data !



Alternative: show your data, ***if you can***

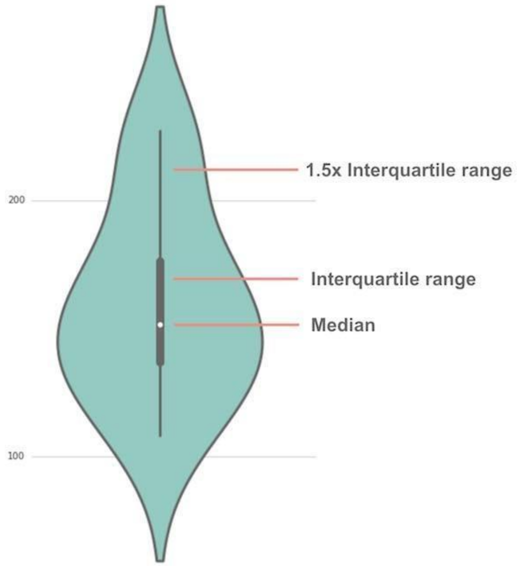


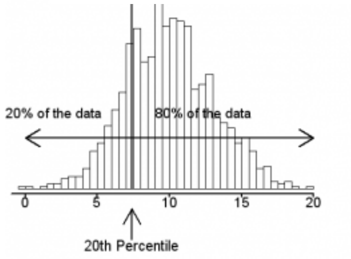


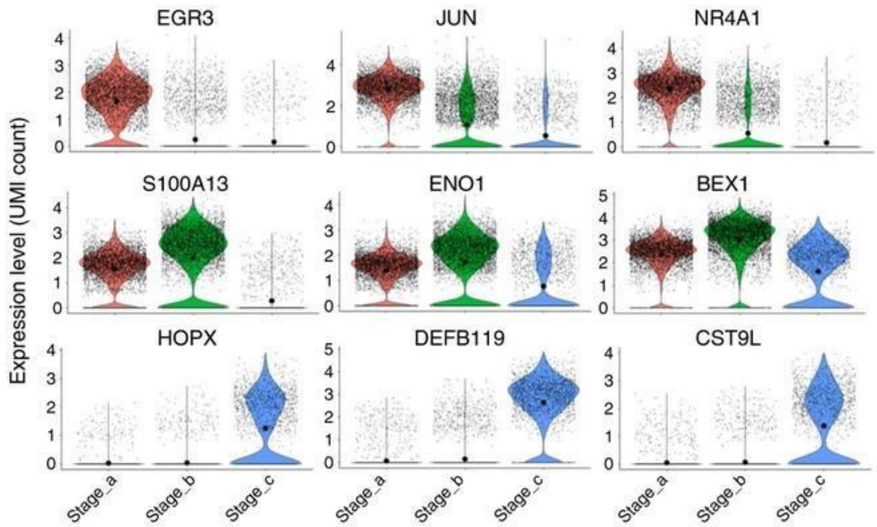
In R: boxplot(data)

Alternative: boxplots (box and whiskersplots)

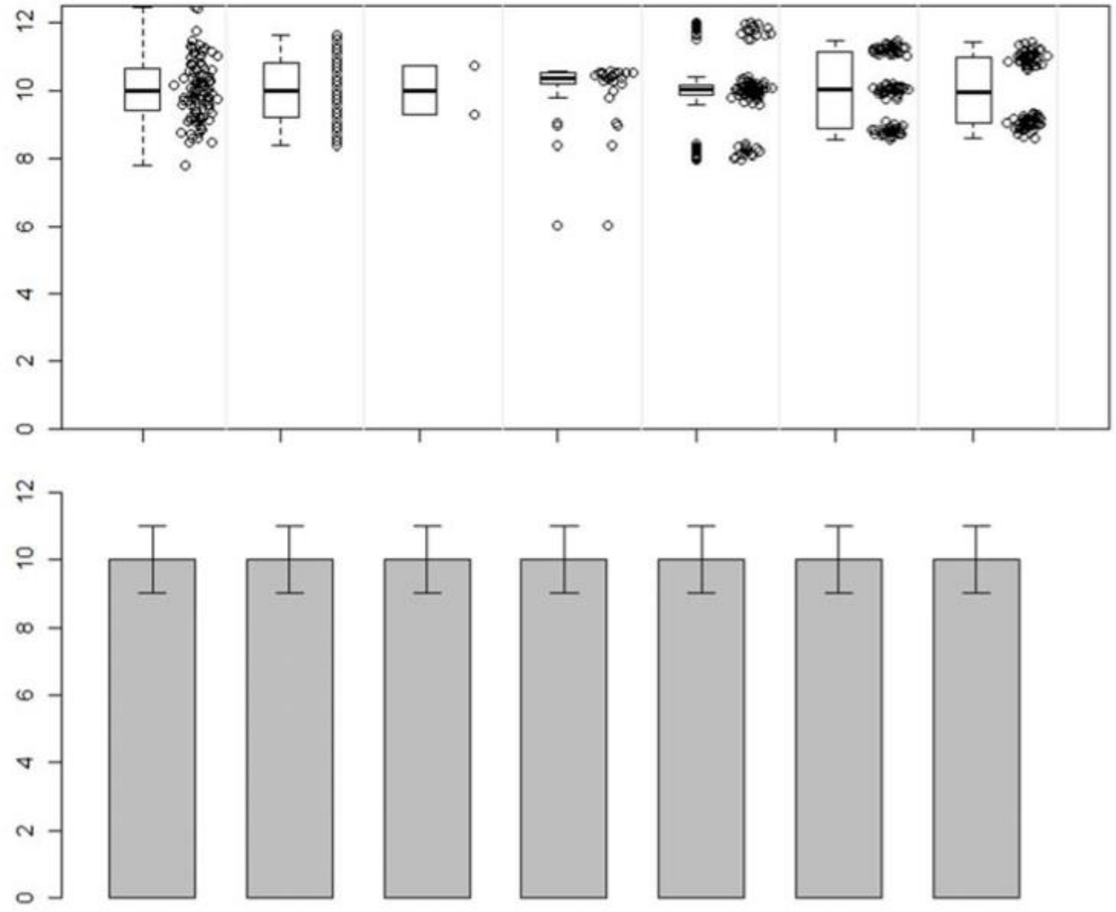
Alternative: violin plots



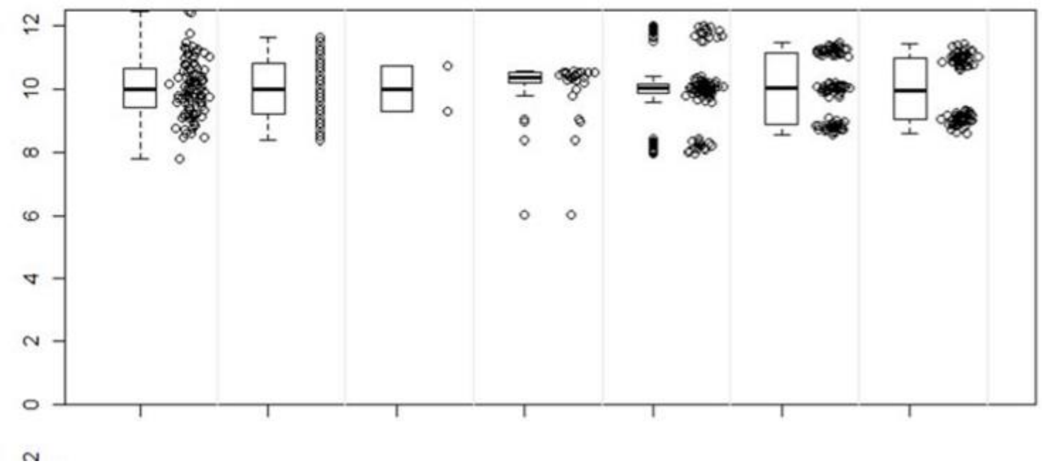


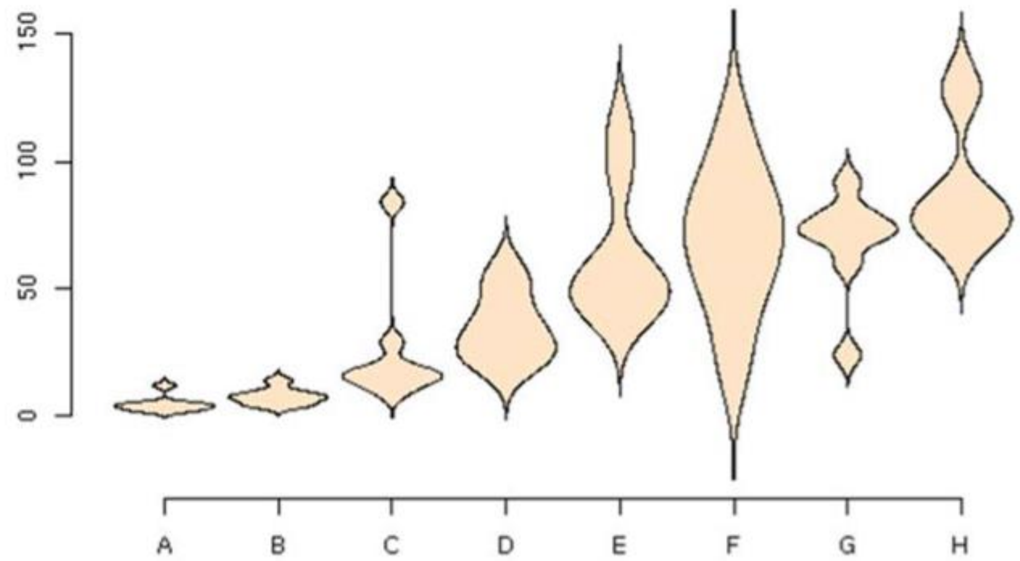


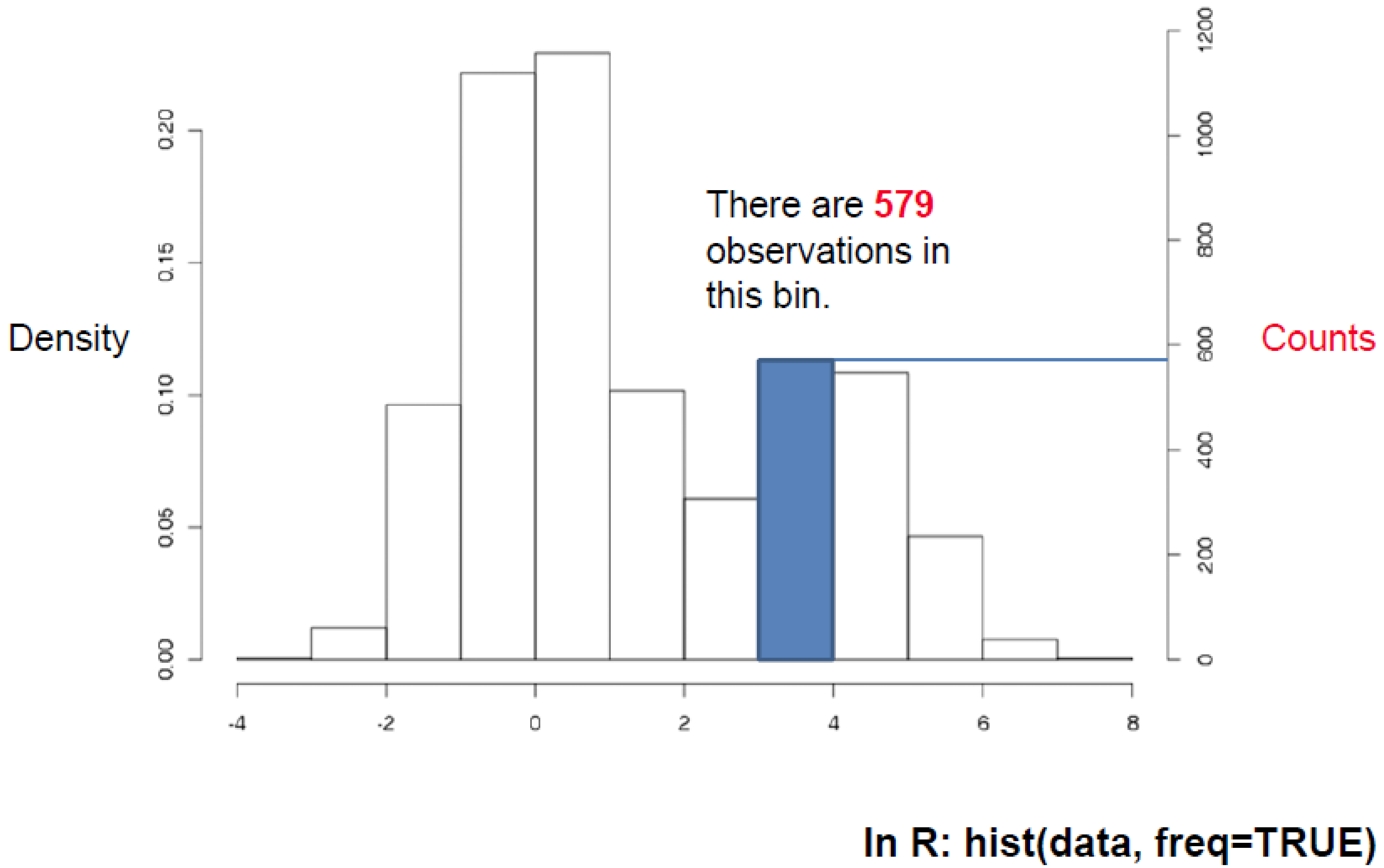
In R: library(vioplot)
vioplot(data)



The associated boxplots

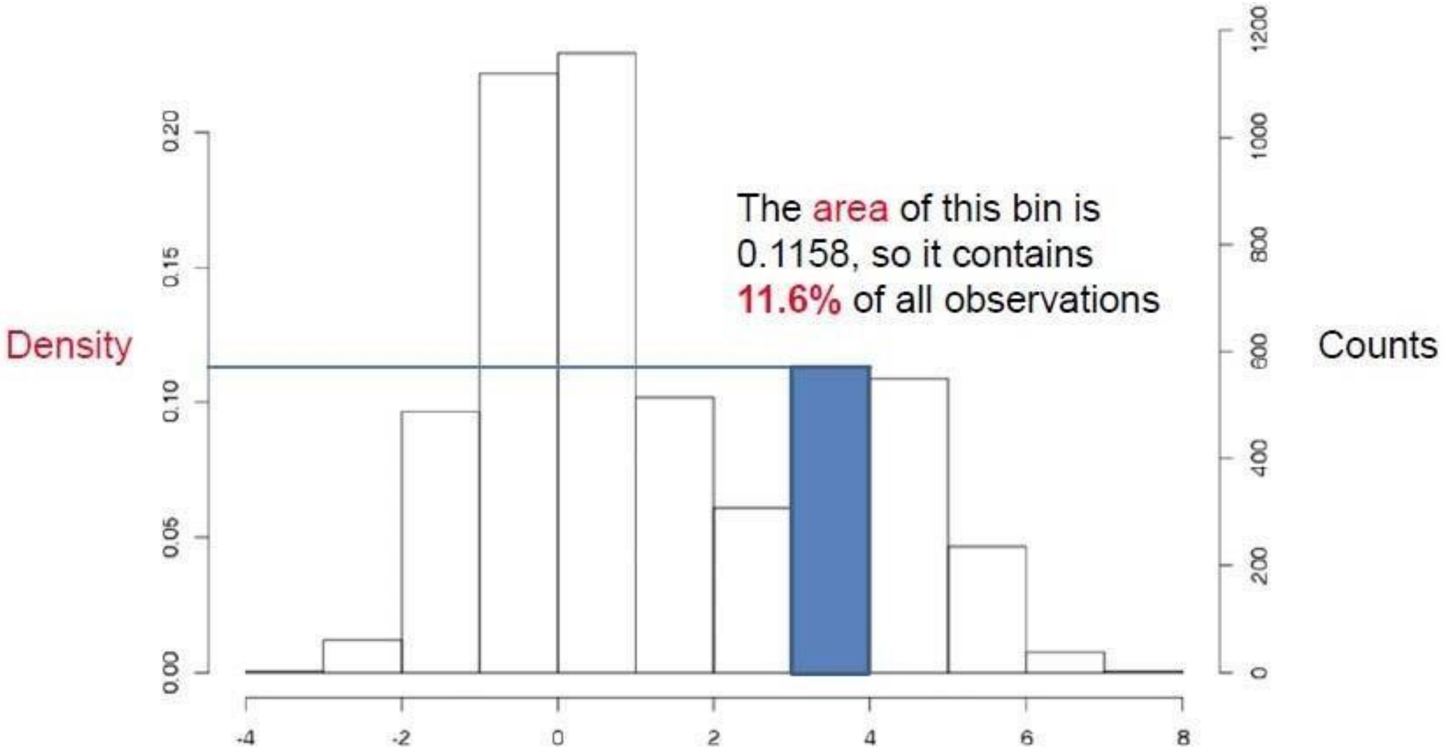

12
10
8
6
4
2
0


150
100
50
0
A
B
C
D
E
F
G
H



Density
There are 579 observations in this bin.
Counts
0.00
0.05
0.10
0.15
0.20
0
200
400
600
800
1000
1200
-4
-2
0
2
4
6
8
In R: hist(data, freq=TRUE)

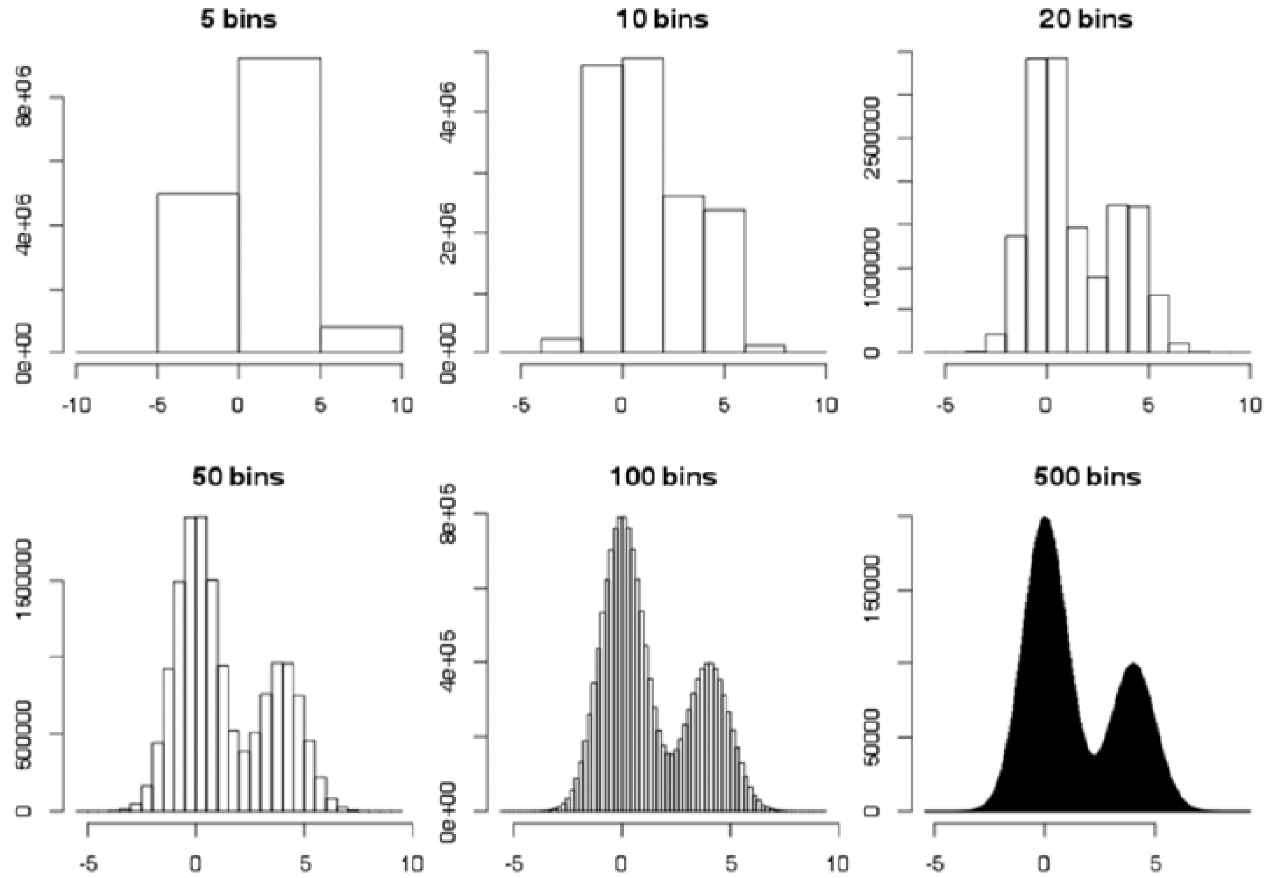
Histograms



In R: hist(data, freq=FALSE)

Histograms

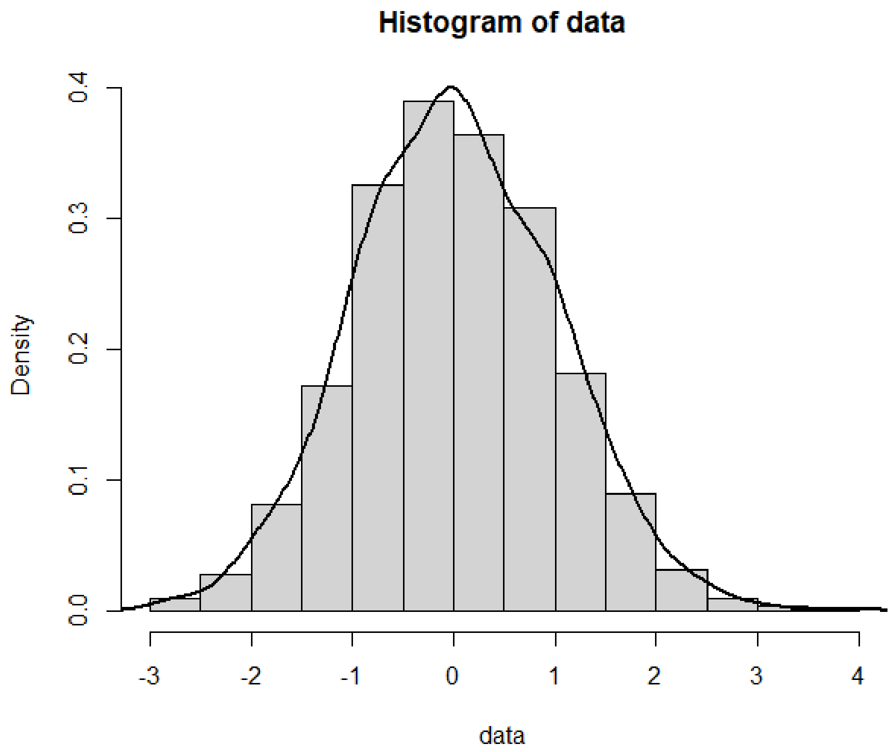
Alternative: histograms



In R: hist(data, breaks=20)

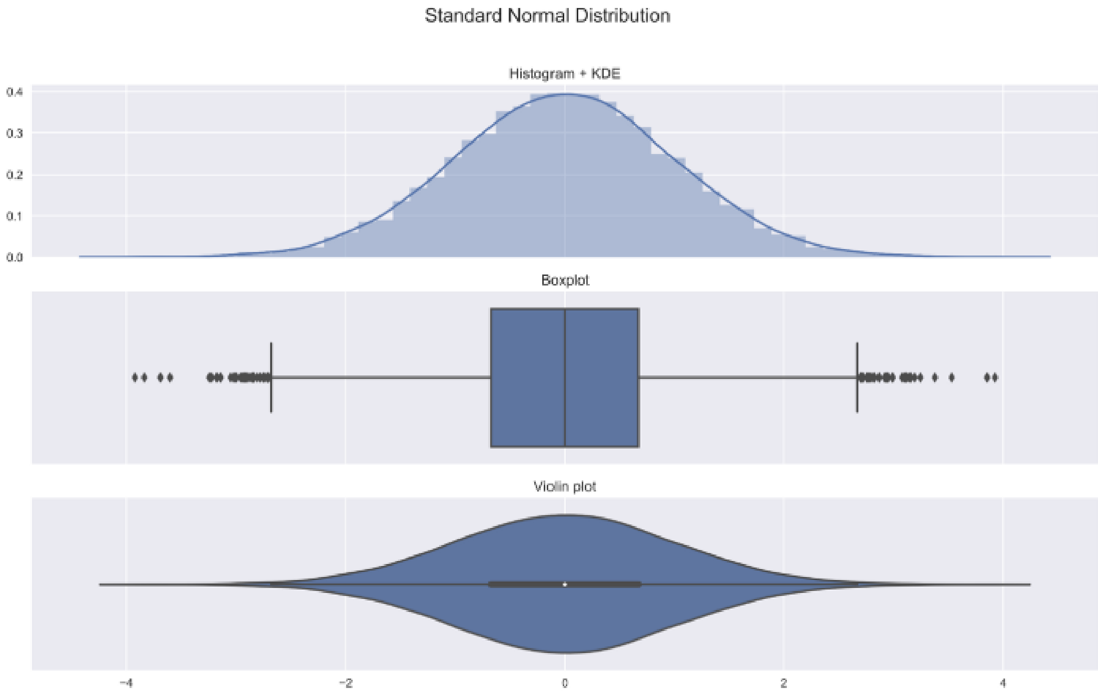
Alternative: histograms with density

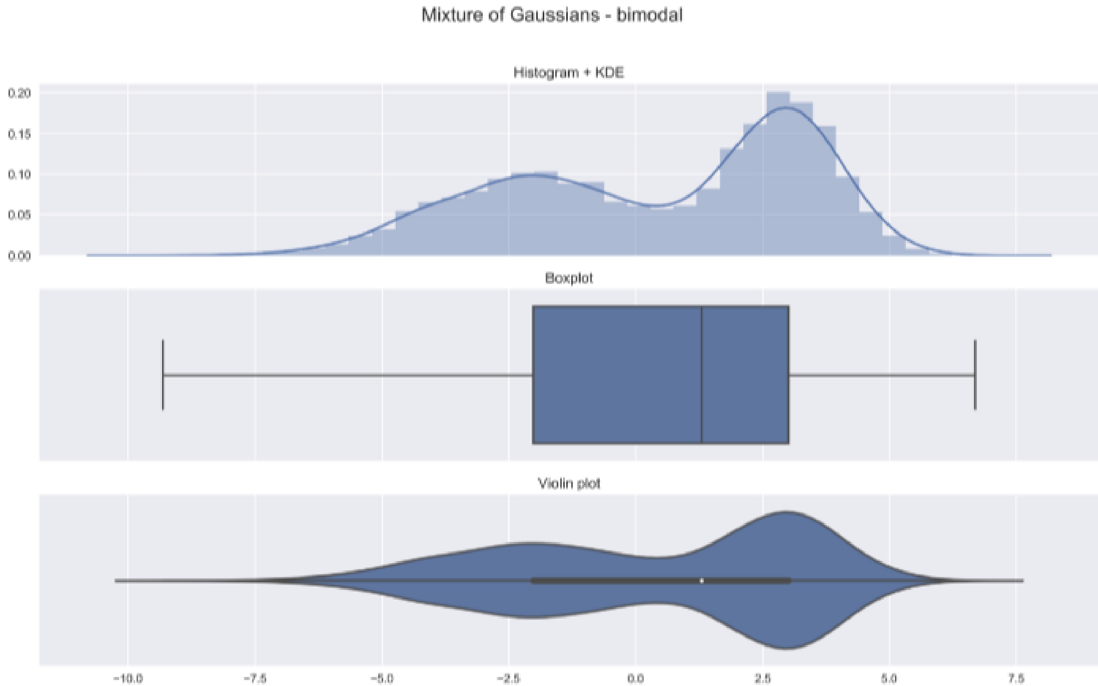
- The density describes the theoretical probability distribution of a variable
- Conceptually, it is obtained in the limit of infinitely many data points
- When we estimate it from a finite set of data, we usually assume that the density is a smooth function
- You can think of it as a "smoothed histogram"



In R: hist(data, freq=F)
lines(density(data), lwd=2)

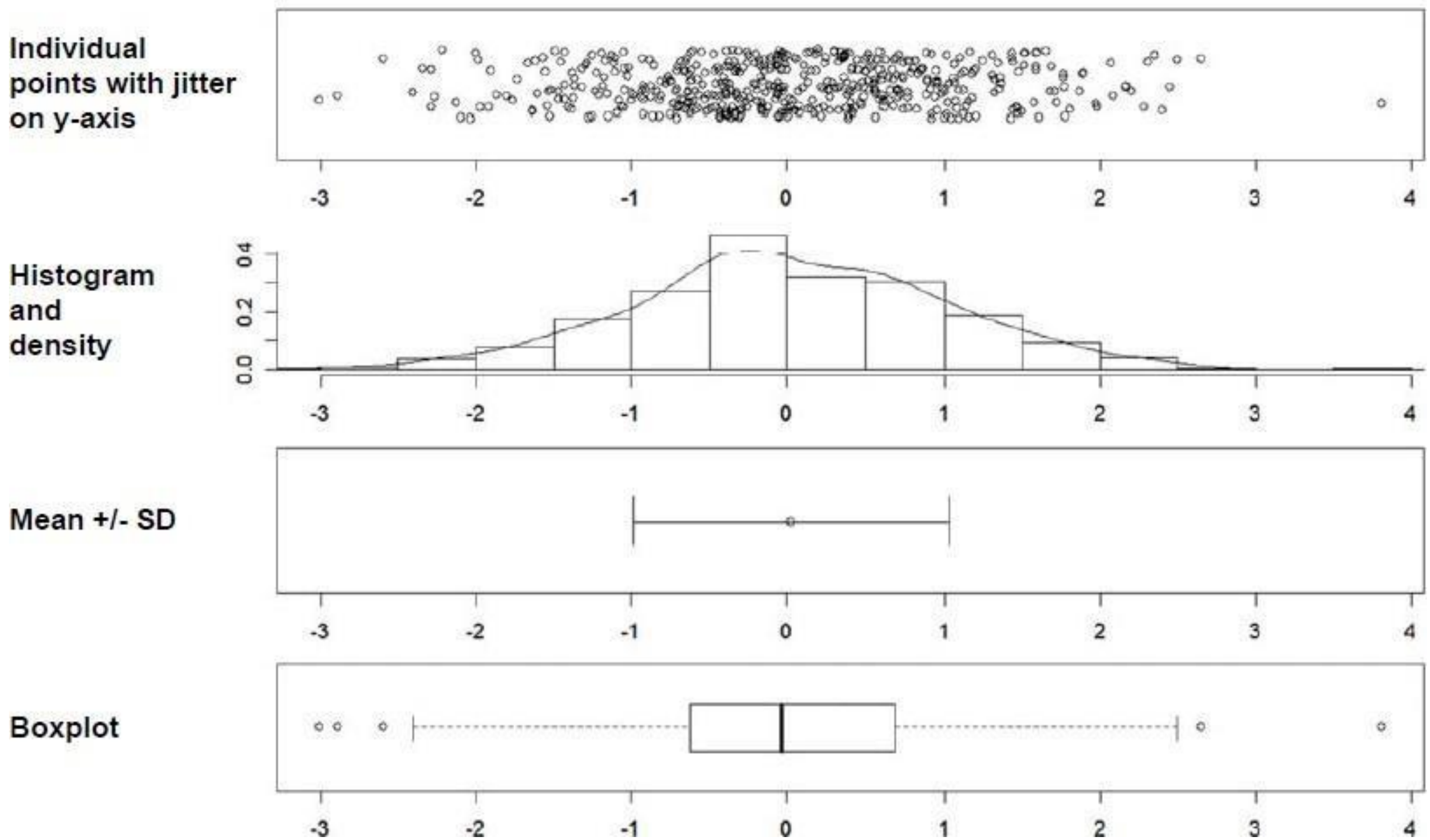
Comparisons of some graphs





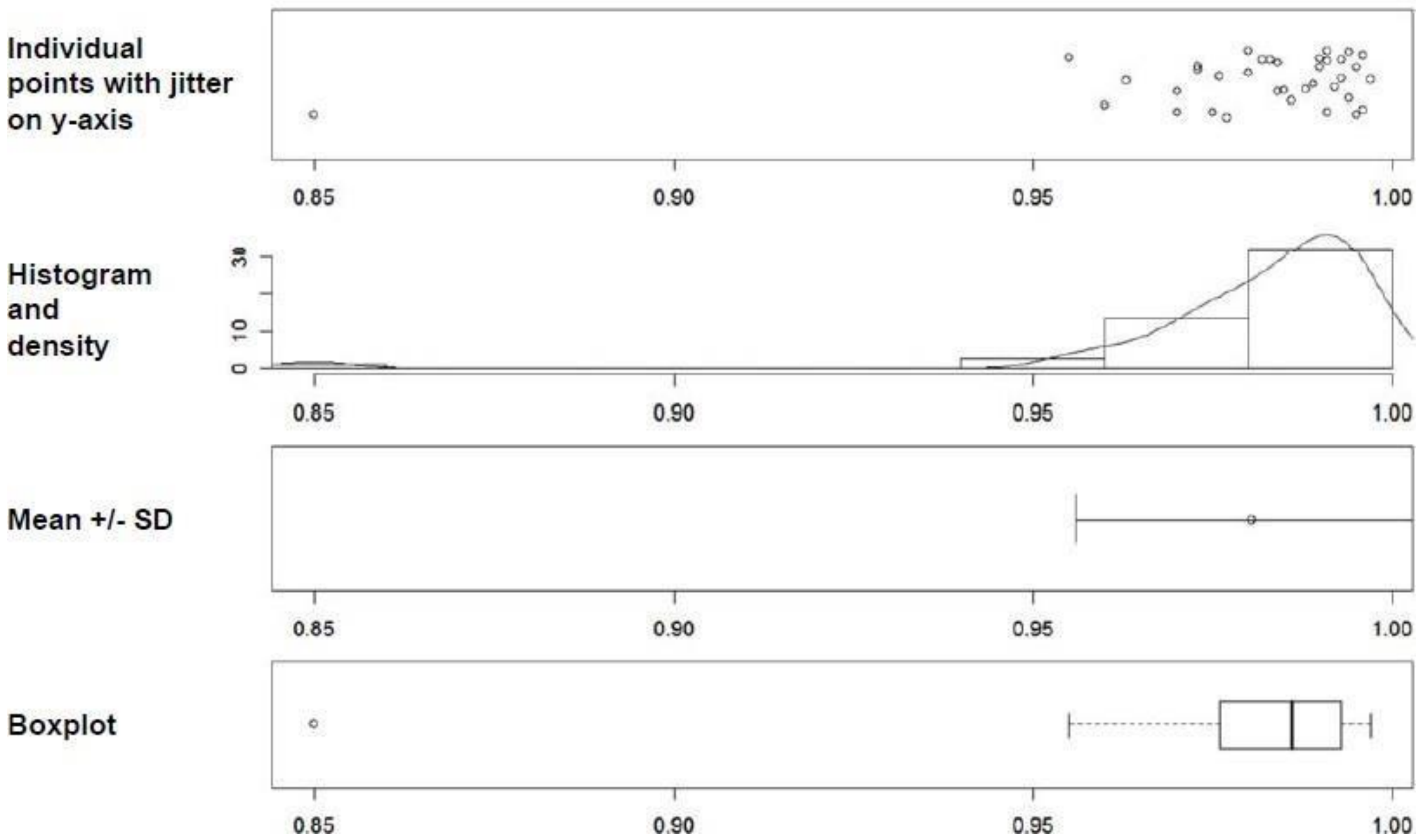
Comparisons of some graphs

Dataset 1 (500 points)



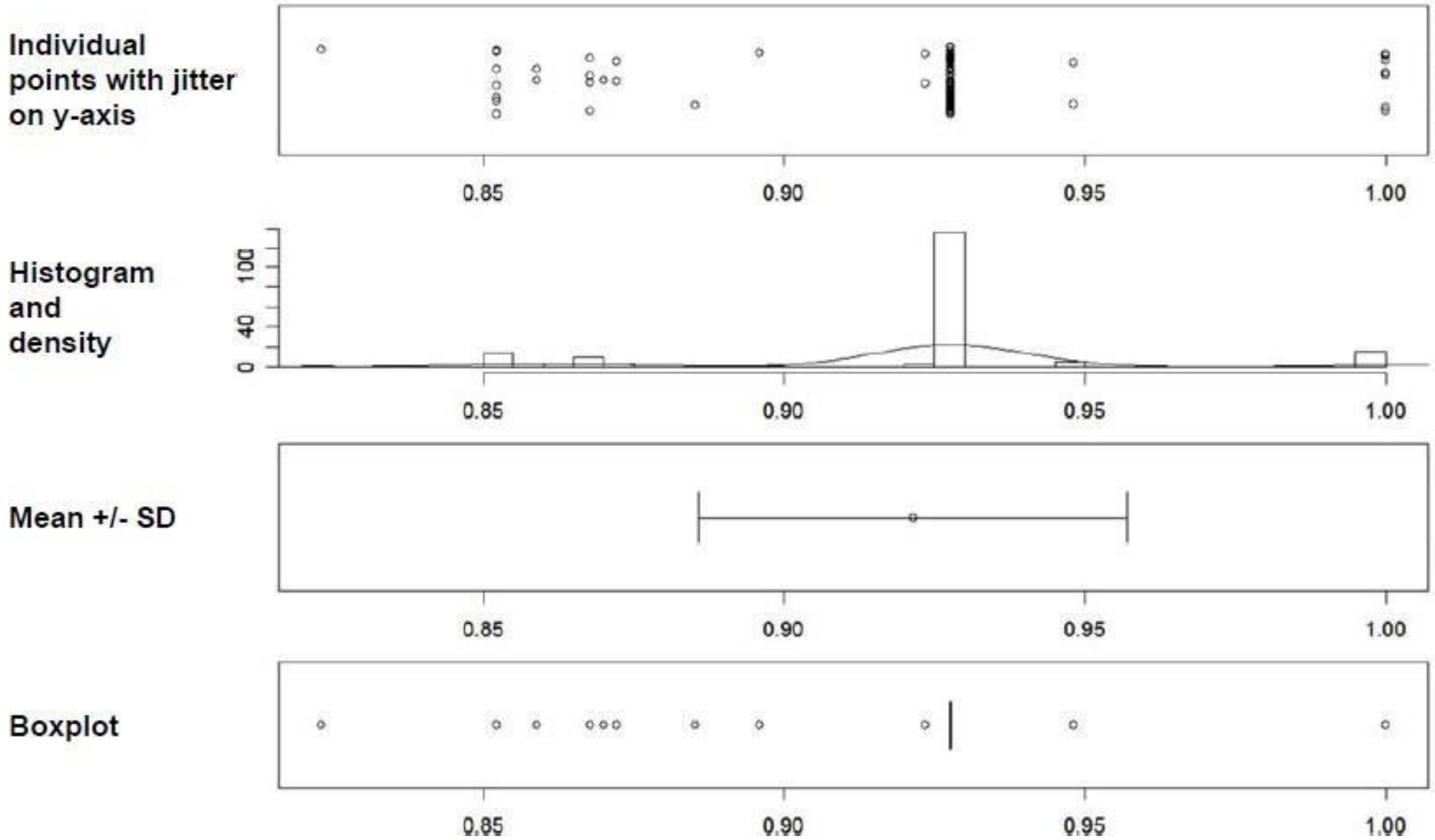
Comparisons of some graphs

Dataset 2 (37 points)



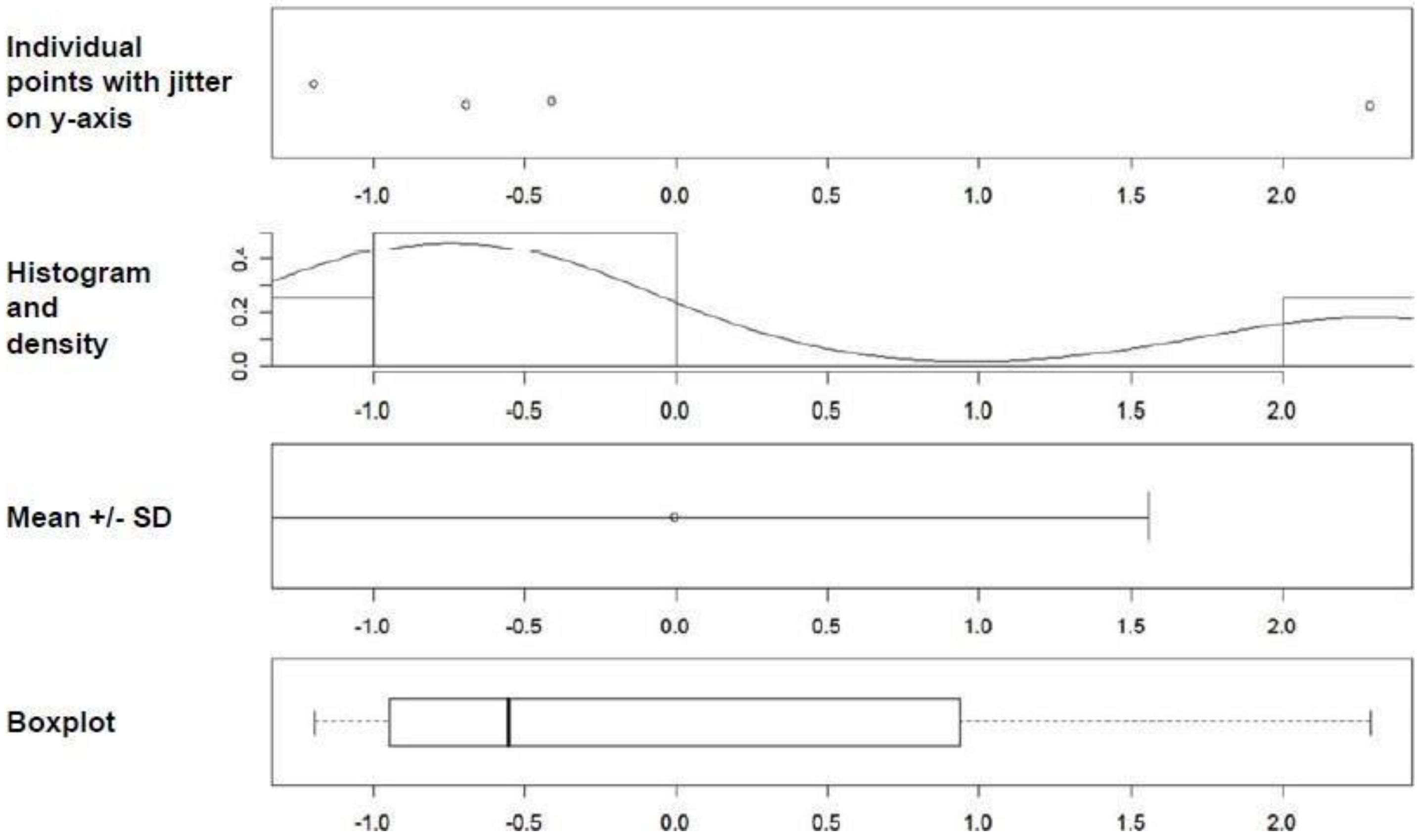
Comparisons of some graphs

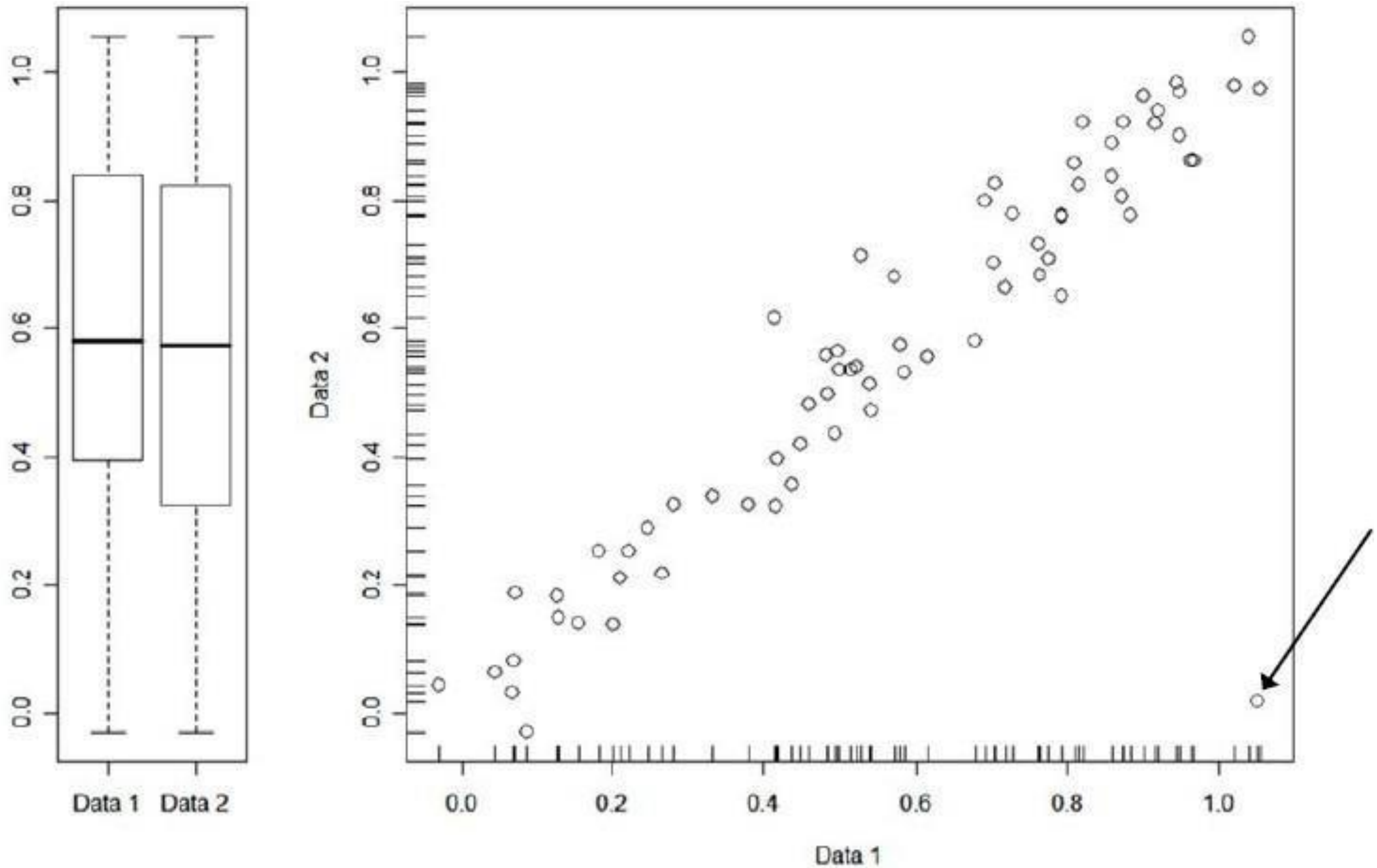
Dataset 3 (100 points)



Comparisons of some graphs

Dataset 4 (4 points)





Bivariate and multivariate data

- **scatterplot**

Summary

- Error bars of different types:
  - Range in R: range(data)
  - SD in R: sd(data)
  - SEM in R: stderror <- function(x) sd(x)/sqrt(length(x)) ; stderror(data)
  - CI
- Histograms in R: hist(data,freq=F)
- With density curve in R:lines(density(data),lwd=2)
- Violin plots in R with the library vioplot: library(vioplot) vioplot(data)
- Boxplot with the meaning of all the lines in R: boxplot(data)
- … and the best way to look at your data, would be to look at it using scatterplots in R: plot(data), if multidimensional, visualise 2 by 2, with pairs(data) and if too big data, use projections.