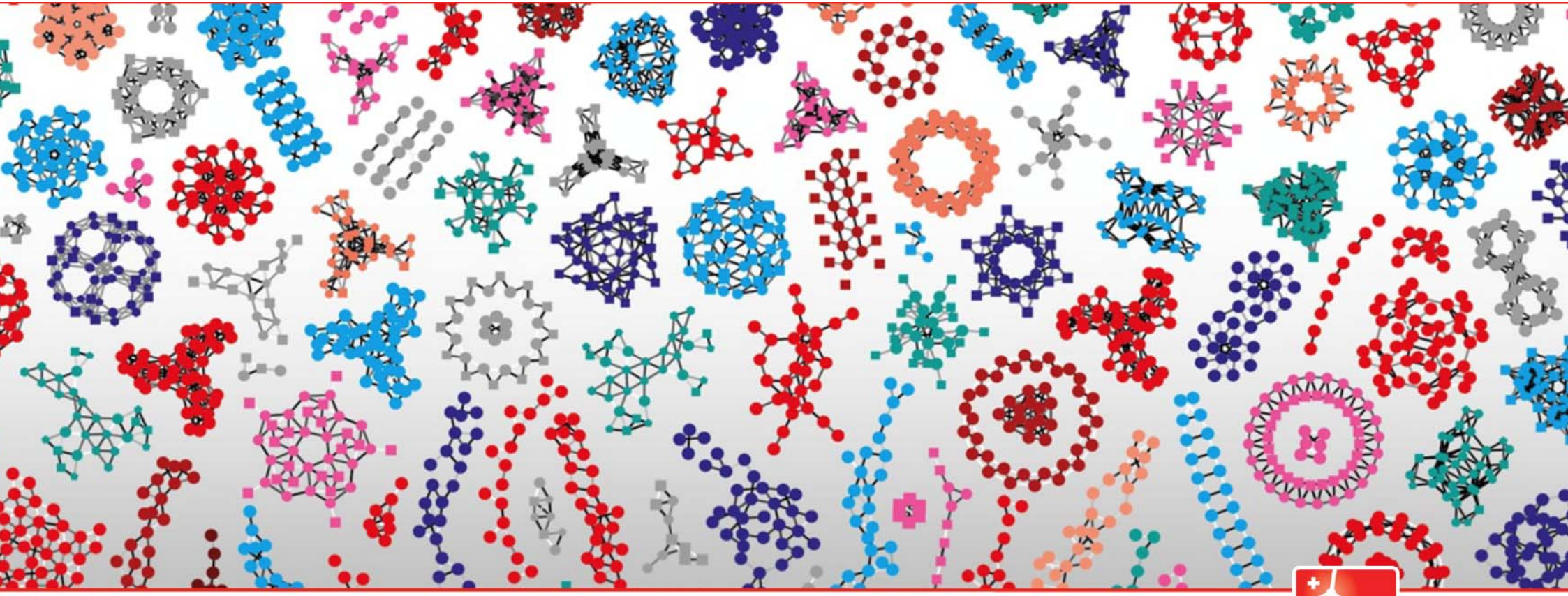


Introduction to Statistics

SIB Swiss Institute of Bioinformatics

Isabelle Dupanloup (isabelle.dupanloup@sib.swiss) and Rachel Marcone (rachel.jeitziner@sib.swiss)

8th-11th February 2021

The Bioinformatics Core Facility at SIB

https://bcf.sib.swiss

- Teaching and training
- Biostatistics and bioinformatics support
- Collaboration

https://www.sib.swiss/mauro-delorenzi-frederic-schutz-group

Course material and credits

- Moodle: https://edu.sib.swiss/course/view.php?id=480
- Login: is21
- Password: SIB-is21

Please, give us feedback at the end of the course !

- Exam: exercises for credits (1 ECTS)
- Send answers to isabelle.dupanloup@sib.swiss

First, tell us about yourself !

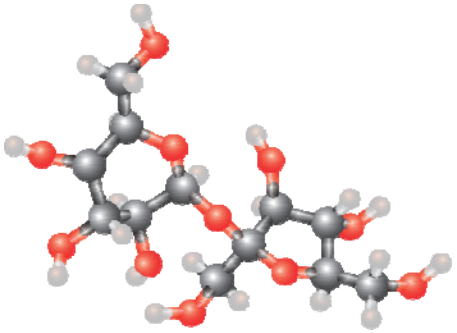
- Background and research area
- What you expect from this course, experience with R

Photo by National Cancer Institute, Unsplash

Photo by Scott Graham, Unsplash

Why biologists need sampling, experimental design and statistics

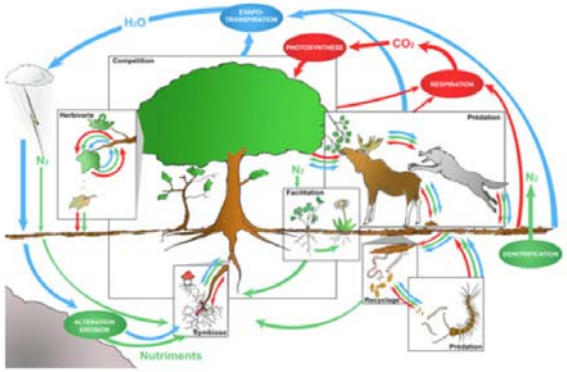
Biology: study of the living world



molecule

cell

population



ecosystem

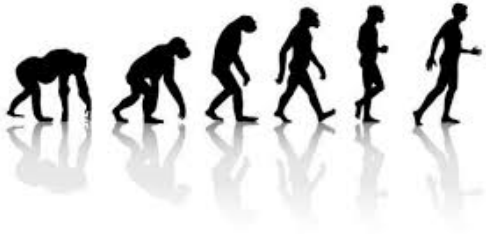
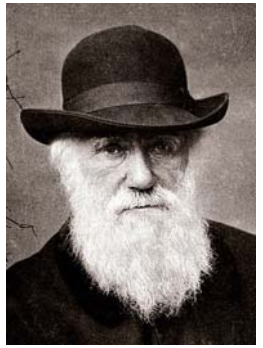
Our knowledge of the living world depends upon careful **observation** and **experimentation**, followed by **analysis** and **interpretation** of the results.

3 indispensable tools used in this process: **sampling**, **experimental design** and **statistical analysis**.

Biology: study of the living world

- Biology in the 19th century



- Biology is now subdivided in a variety of specialised areas defined by
  - the different levels of organization of biological systems
  - the type of organism being investigated
  - the type of biological process being studied

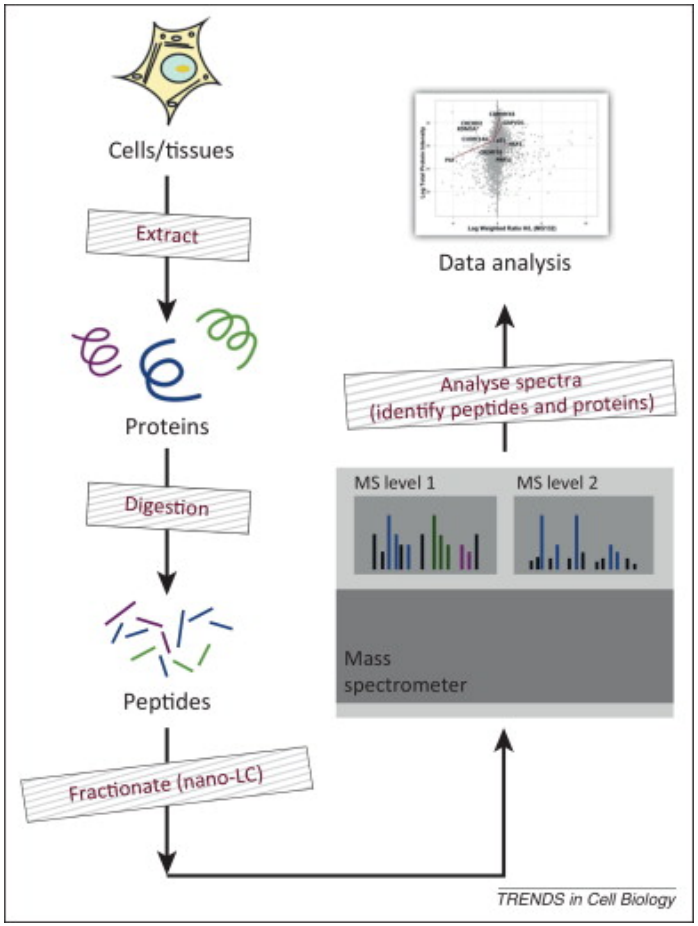
Biology: study of the living world

- What biologists do
  - **Describe** the characteristics of the objects of study
  - **Explain** what has been described
- Regardless of his field of activity, a biologist makes observations, describes them and then attempts to explain them.
- He measures numbers which are called **variables**, because these numbers vary for different reasons.
- A biologist seeks to characterize the observed variability.

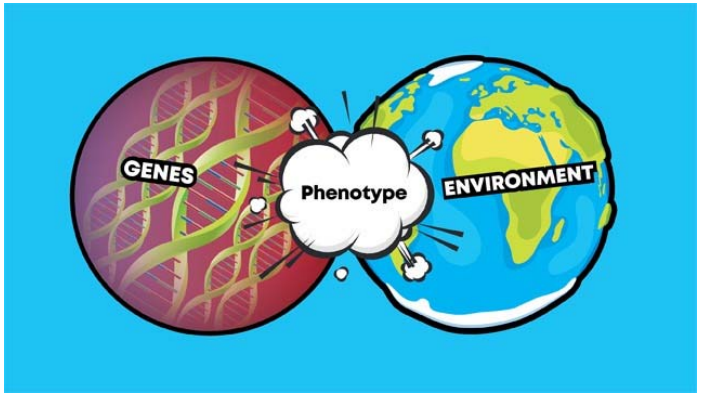
Sources of variability

experimental variation

experimental error



Sources of variability

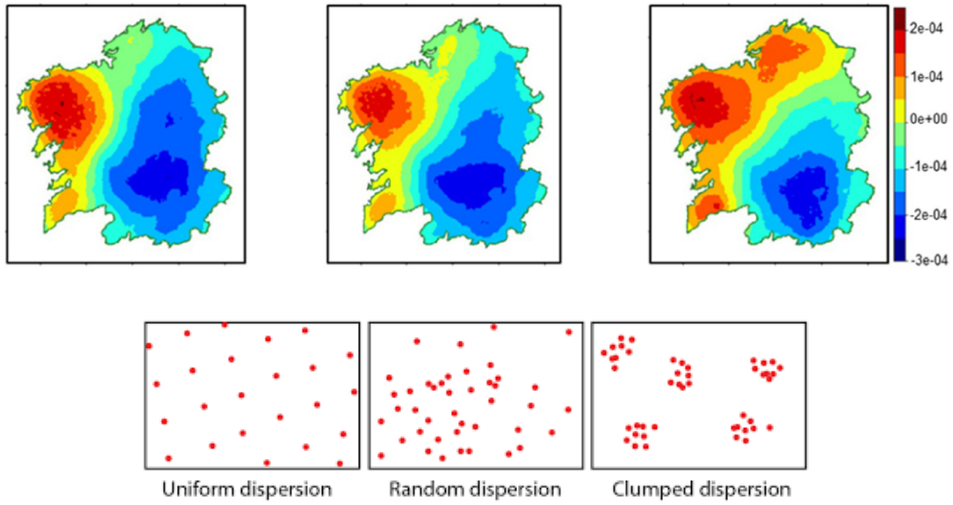


Variability in genotypes and phenotypes

```
Swiss1        ATGCGTGATTGGTGAGACTTTGATTAGGGA
Swiss2        ATGCGTGATTGGTGAGACTTTGATTAGGGA
Nigerian1     ATGCGAGATTGGTGAGACTTTGATTAGGGA
German1       ATGCGTGATTGGTGAGAGTTTGATTAGGGA
Chinese2      ATGCGTGATTGGTGAGAGTTTGATTAGGGA
Inuit3        ATGCGTGATTGGTGAGACTTTGATTAGGGA
Swedish1      ATGCGTGACTGGTGAGACTTTGATTAGGGA
```

Sources of variability

Variability in space and time



Abies alba

Why biologists need statistics ?

- Estimating
  - Use of a subset of all possible observations: a **sample**
  - Set of all possible observations: **population**
  - Inferences on the characteristics of the population
  - **Sampling variation, sampling error**
  - Notion of **bias**

Why biologists need statistics ?

- How to solve those issues ?
  - Bias ?
  - ➢ Design an objective sampling strategy
  - Sampling variation ?
  - ➢ Get a measure of reliability of the estimate
  - ➢ **Statistical analysis**

Why biologists need statistics ?

- Detect the differences between populations

➢ we take a sample for each of the populations

- Differences between samples: 2 possibilities

▪ the original populations are different

▪ the differences observed are due to sampling

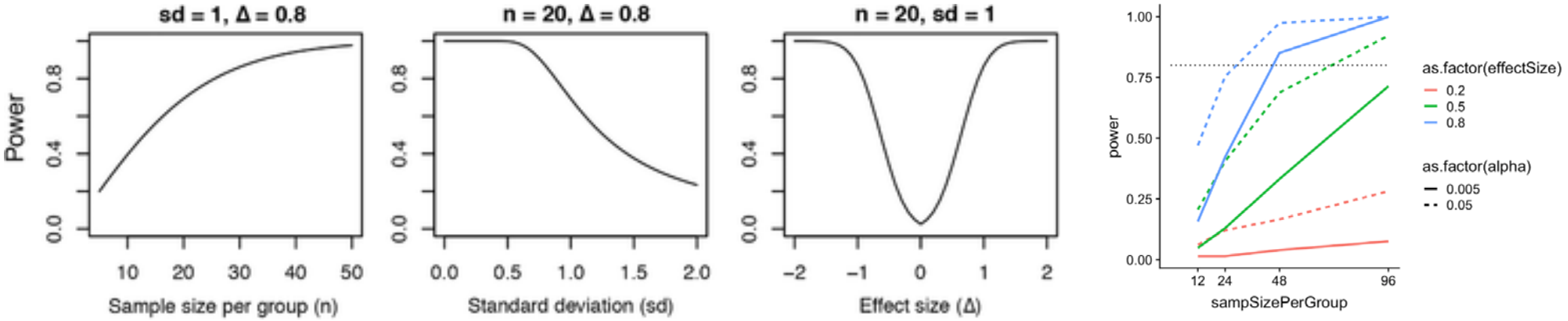
➢ **statistical analysis**

Guideline for using statistics in biology

1. Specify the biological question of interest.
2. Put the question in the form of a **biological null hypothesis** and **alternate hypothesis**.
3. Put the question in the form of a **statistical null hypothesis** and **alternate hypothesis**.
4. Determine which **variables** are relevant to the question and what kind of variable each one is.
5. **Design an experiment** that controls or randomizes the **confounding variables**.
6. Based on the number of variables, the kinds of variables, the expected fit to the parametric assumptions, and the hypothesis to be tested, **choose the best statistical test to use**.
7. If possible, do a **power analysis** to determine a good sample size for the experiment.
8. Do the experiment.
9. **Examine the data** (explore variation and check if the assumptions of the statistical test you chose (primarily normality and homoscedasticity for tests of measurement variables) are met (if it doesn't, choose a more appropriate test)).
10. **Apply the statistical test** you chose, and **interpret** the results.
11. **Communicate** your results **effectively**.

Statistical power

- probability of rejecting a null hypothesis when it is false = 1 - β
  probability of a Type II error
- **common target = 0.8**
- depends on: number of measurements, variability of those measurements, and effect size



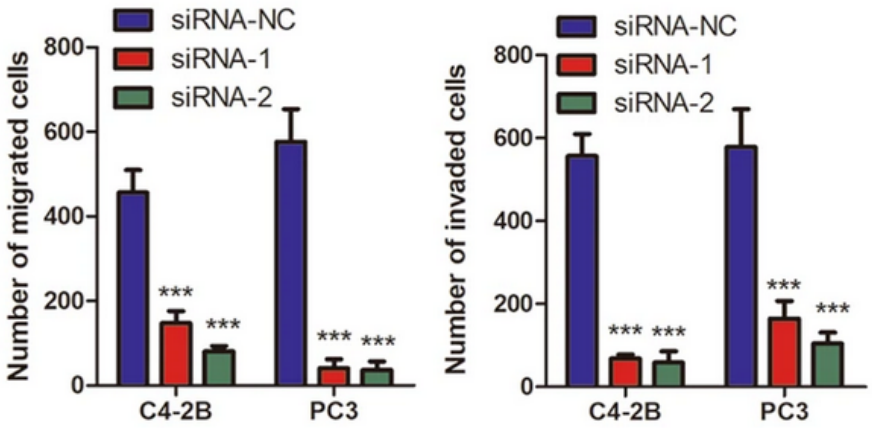
- probability of rejecting a null hypothesis when it is true = α
  probability of a Type I error

Guideline for using statistics in biology

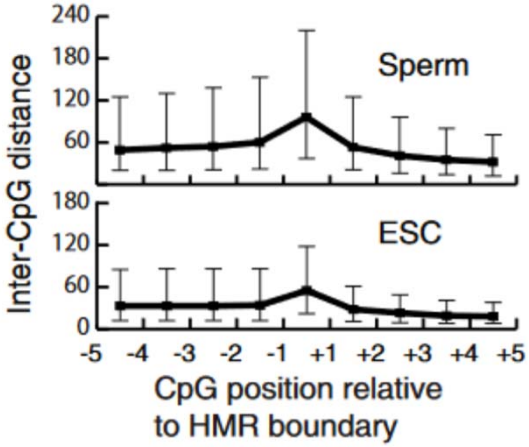
1. Specify the biological question of interest.
2. Put the question in the form of a **biological null hypothesis** and **alternate hypothesis**.
3. Put the question in the form of a **statistical null hypothesis** and **alternate hypothesis**.
4. Determine which **variables** are relevant to the question and what kind of variable each one is.
5. **Design an experiment** that controls or randomizes the **confounding variables**.
6. Based on the number of variables, the kinds of variables, the expected fit to the parametric assumptions, and the hypothesis to be tested, **choose the best statistical test to use**.
7. If possible, do a **power analysis** to determine a good sample size for the experiment.
8. Do the experiment.
9. **Examine the data** (explore variation and check if the assumptions of the statistical test you chose (primarily normality and homoscedasticity for tests of measurement variables) are met (if it doesn't, choose a more appropriate test)).
10. **Apply the statistical test** you chose, and **interpret** the results.
11. **Communicate** your results **effectively**.

What type of graphics
do you know ?

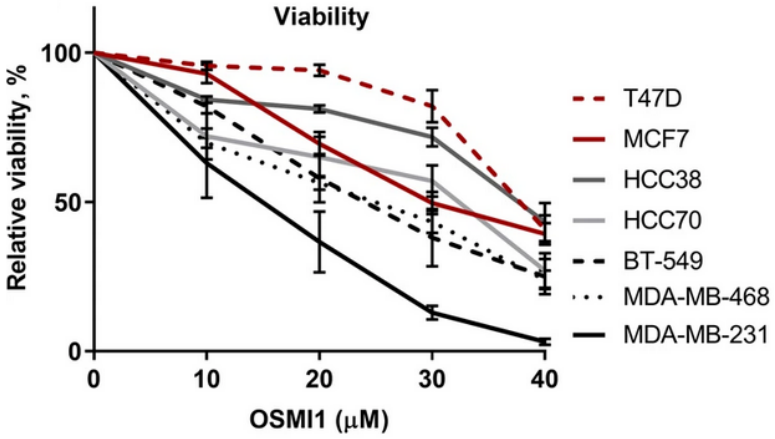
Error bars are ubiquitous in the scientific literature



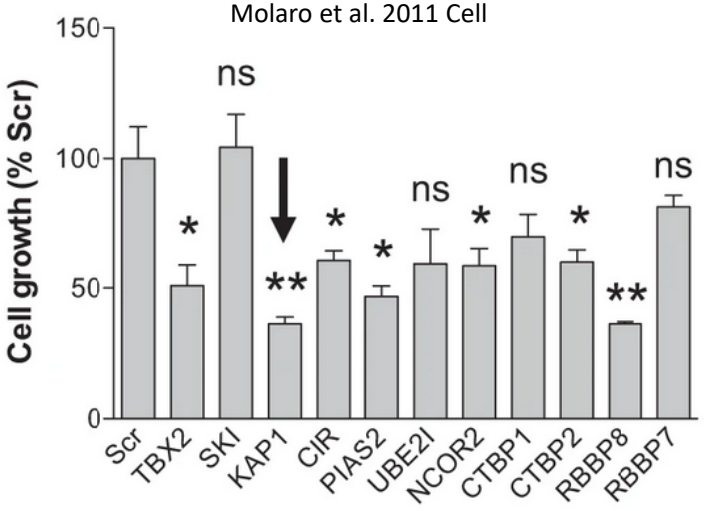
Cao et al. 2021 Cell Death & Disease



Molaro et al. 2011 Cell



Barkovskaya et al. 2019 Scientific Reports



Crawford et al. 2019 Oncogene

Error bars are ubiquitous in the scientific literature

Journals	Counts of articles by error bar types				Total counts[†]
	SD	SEM	Others[*]	Unidentified	
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.

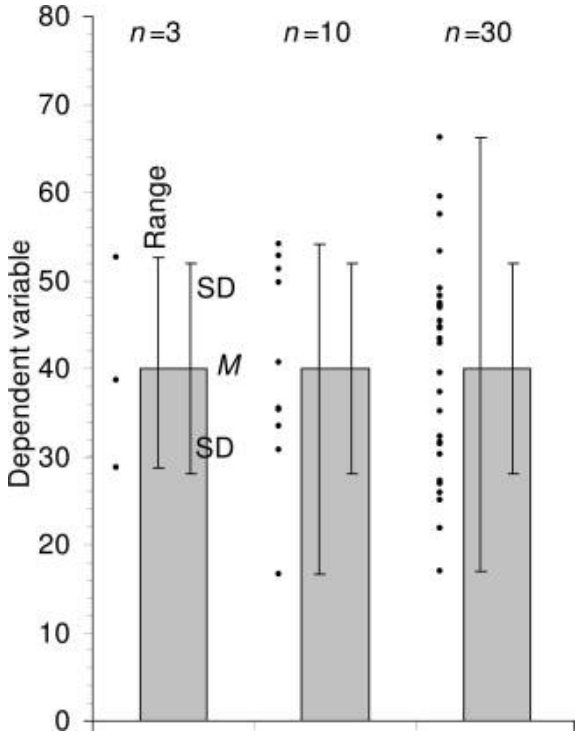
[†] These data represent the total number of articles that appeared in the publication during the review period that used error bars in figures. The articles using 2 or more types of error bars were counted in each category but only once in the total category.

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

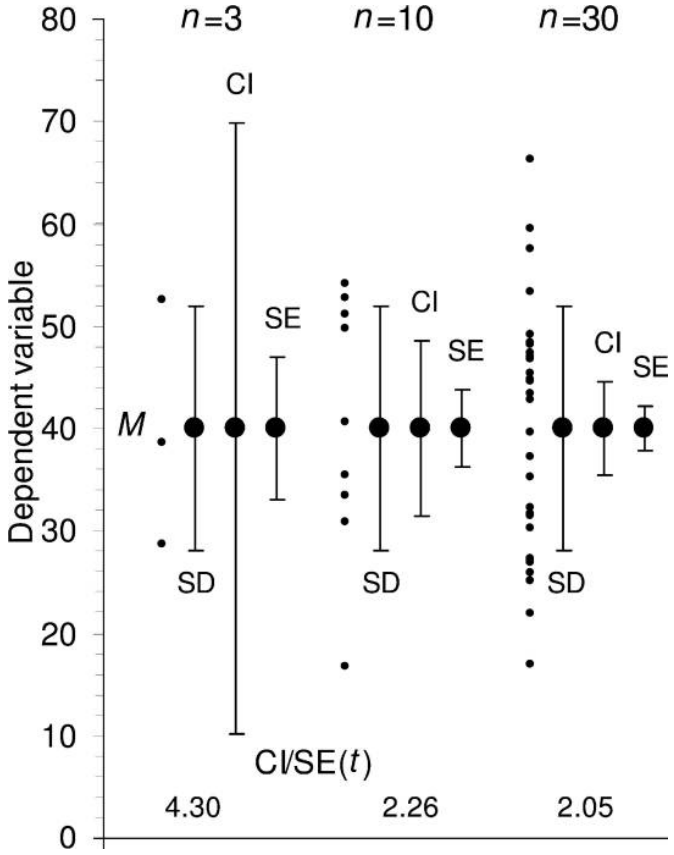
Error bars are ubiquitous in the scientific literature

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	$M \pm t_{(n-1)} \times \mathrm{SEM}$, where $t_{(n-1)}$ is a critical value of t. If n is 10 or more, the 95% CI is approximately $M \pm 2 \times \mathrm{SEM}$.

Error bars are ubiquitous in the scientific literature

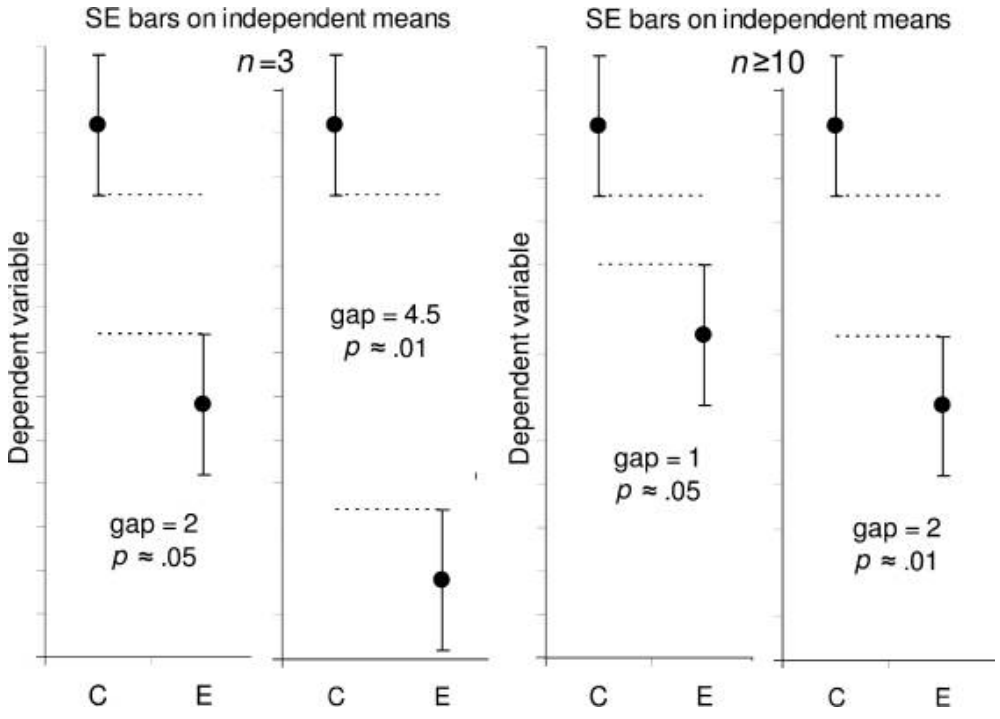


Descriptive error bars

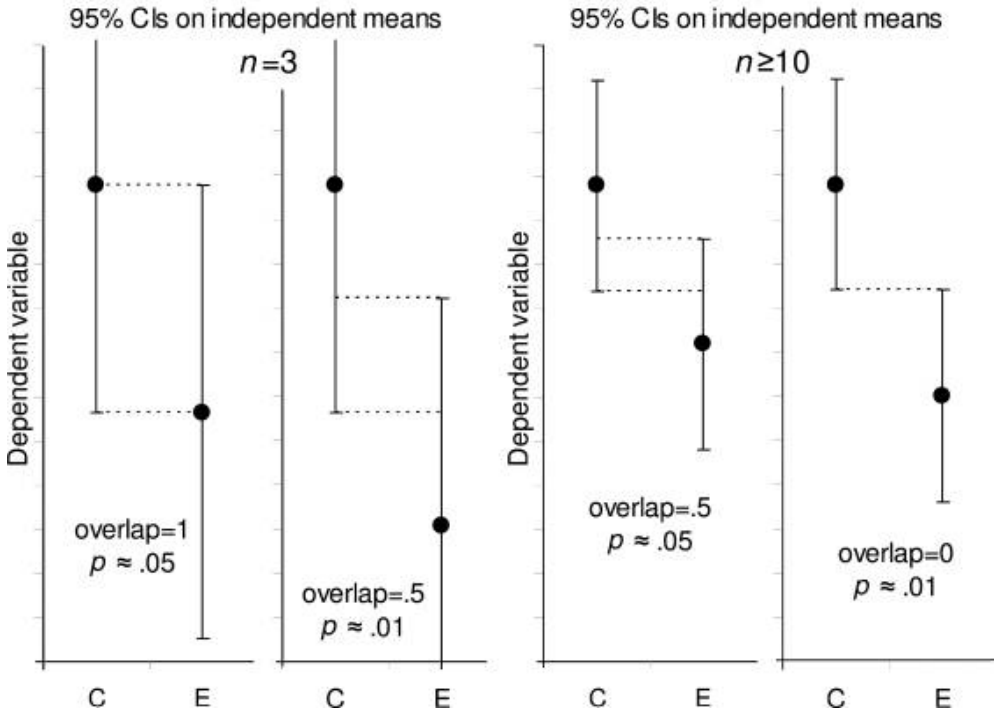


Inferential error bars

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.

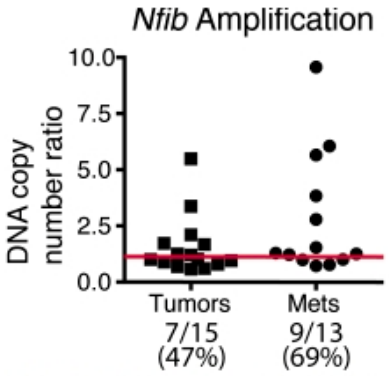
Error bars are ubiquitous in the scientific literature

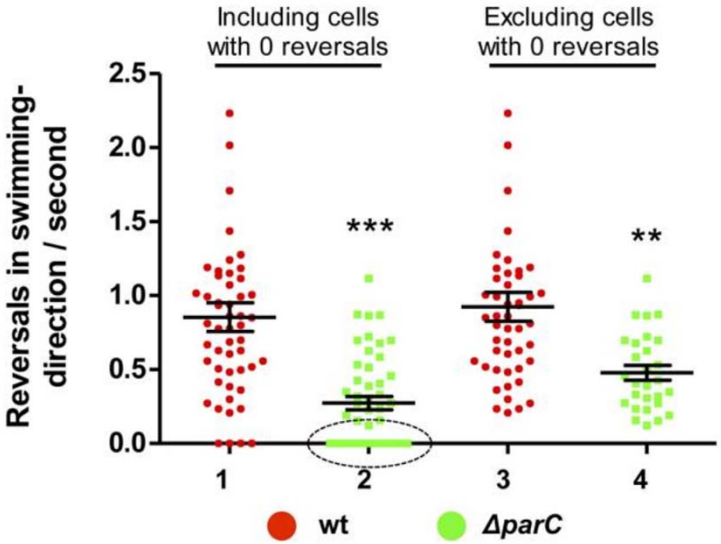
Avoid error bars if possible

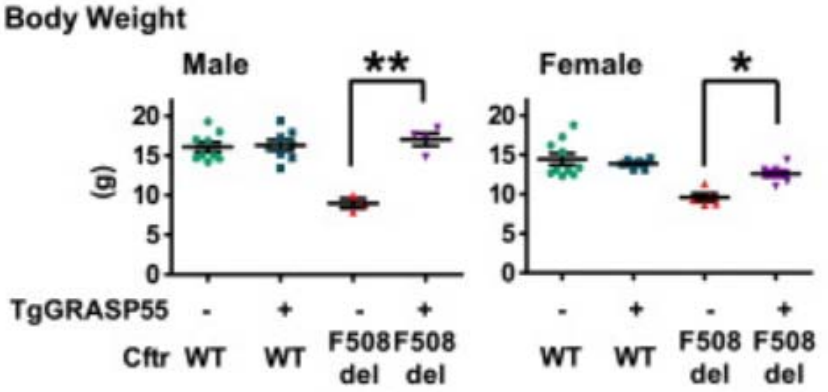
If you have to use them, document them, and try not to use them alone.

What are the alternatives ?

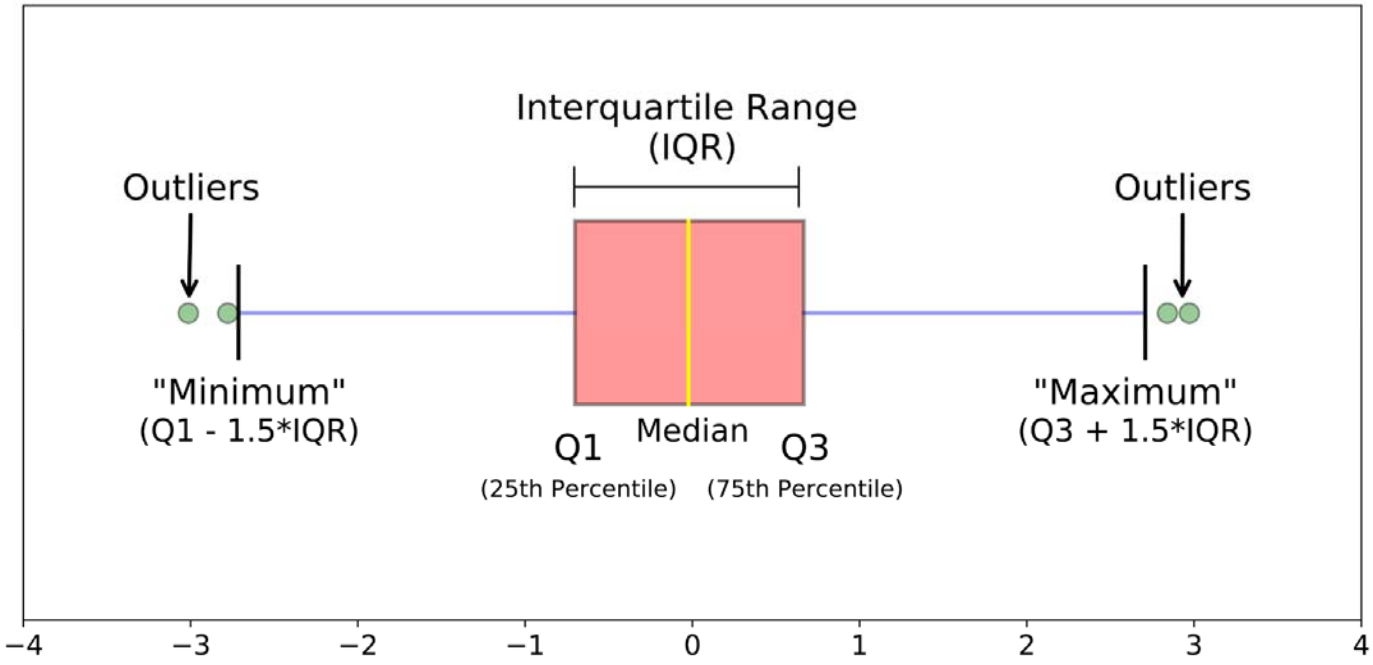
Alternative: show your data !

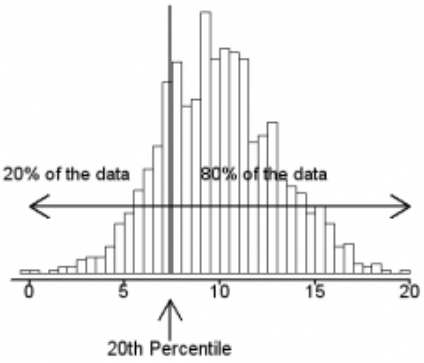






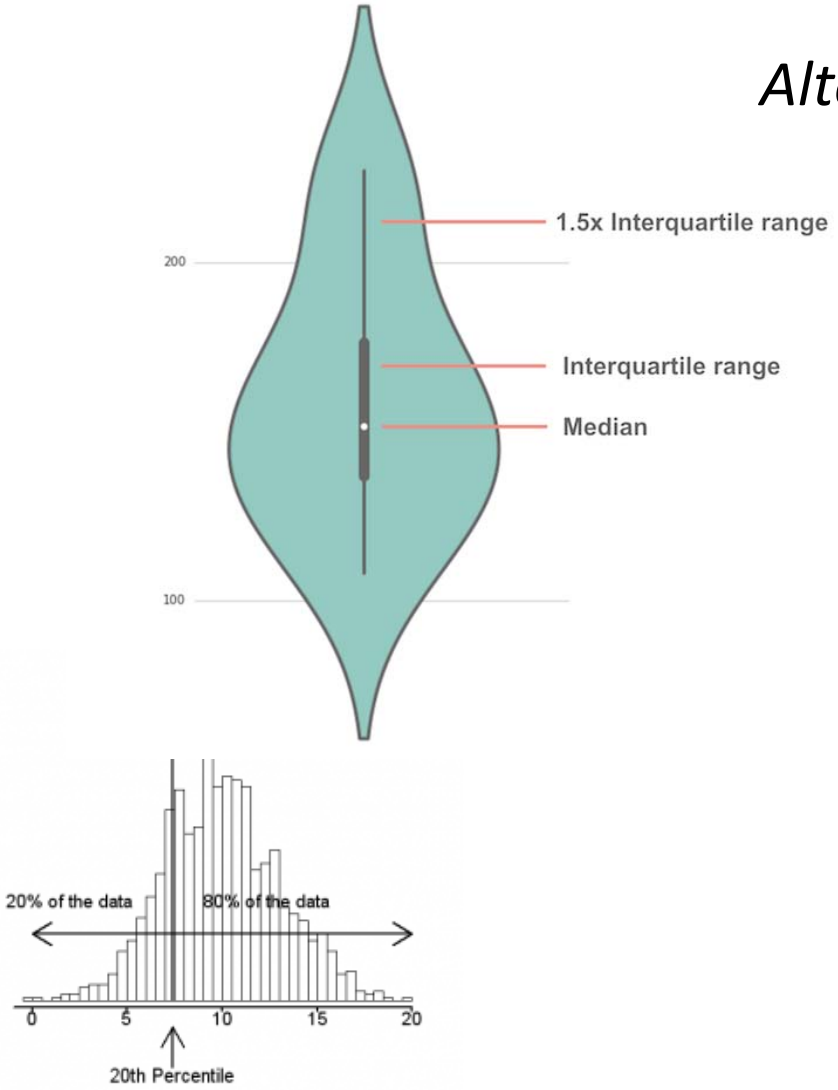
Alternative: boxplots (box and whiskers plots)

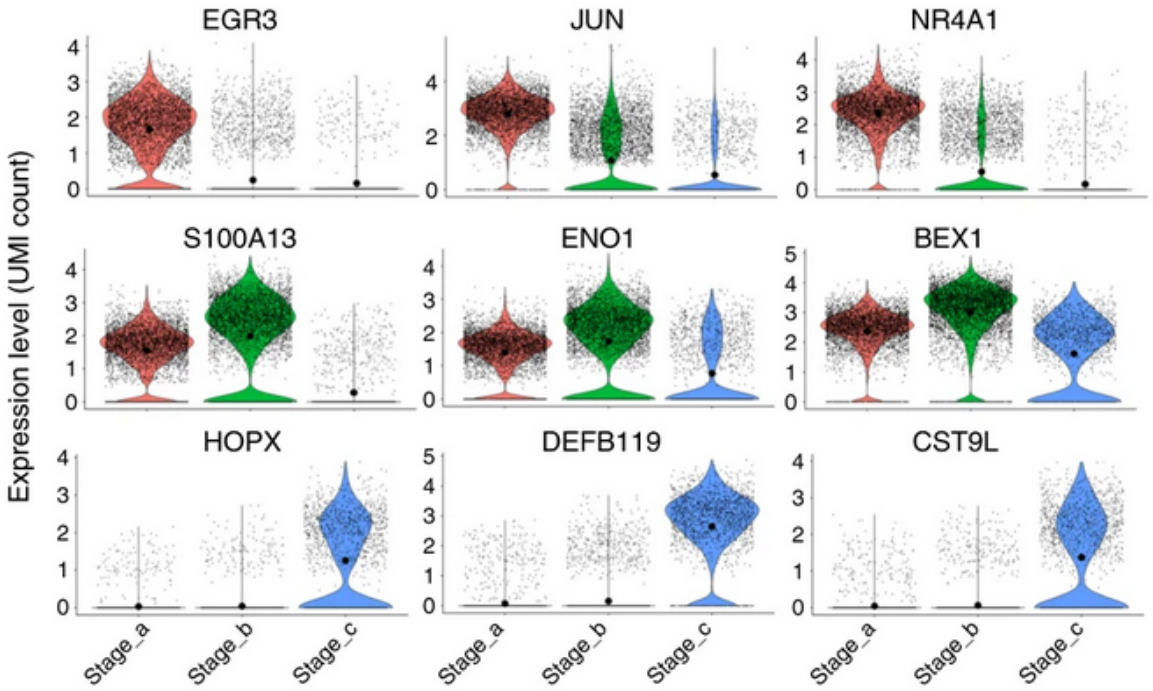




In R: boxplot(data)

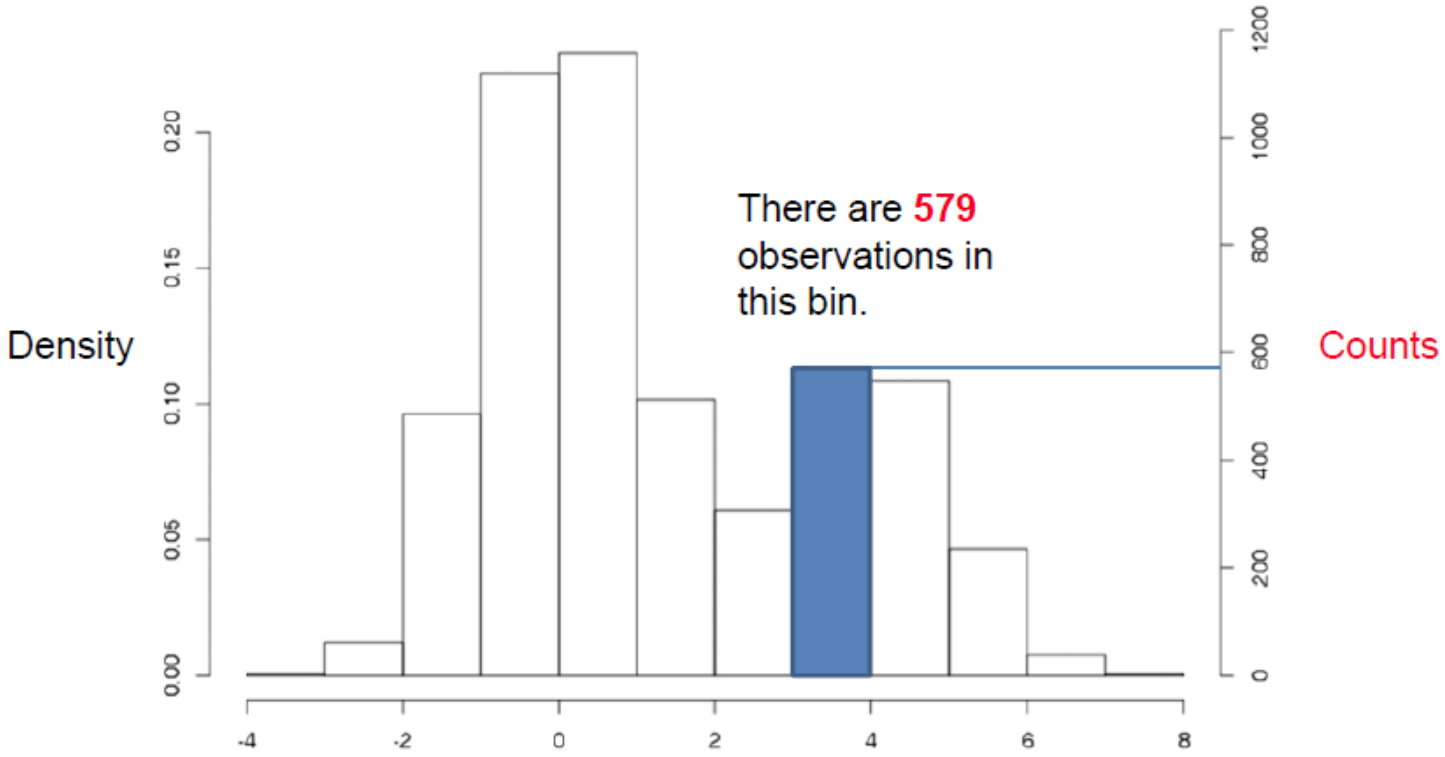
Alternative: violin plots





In R: library(vioplot)
vioplot(data)

Alternative: histograms

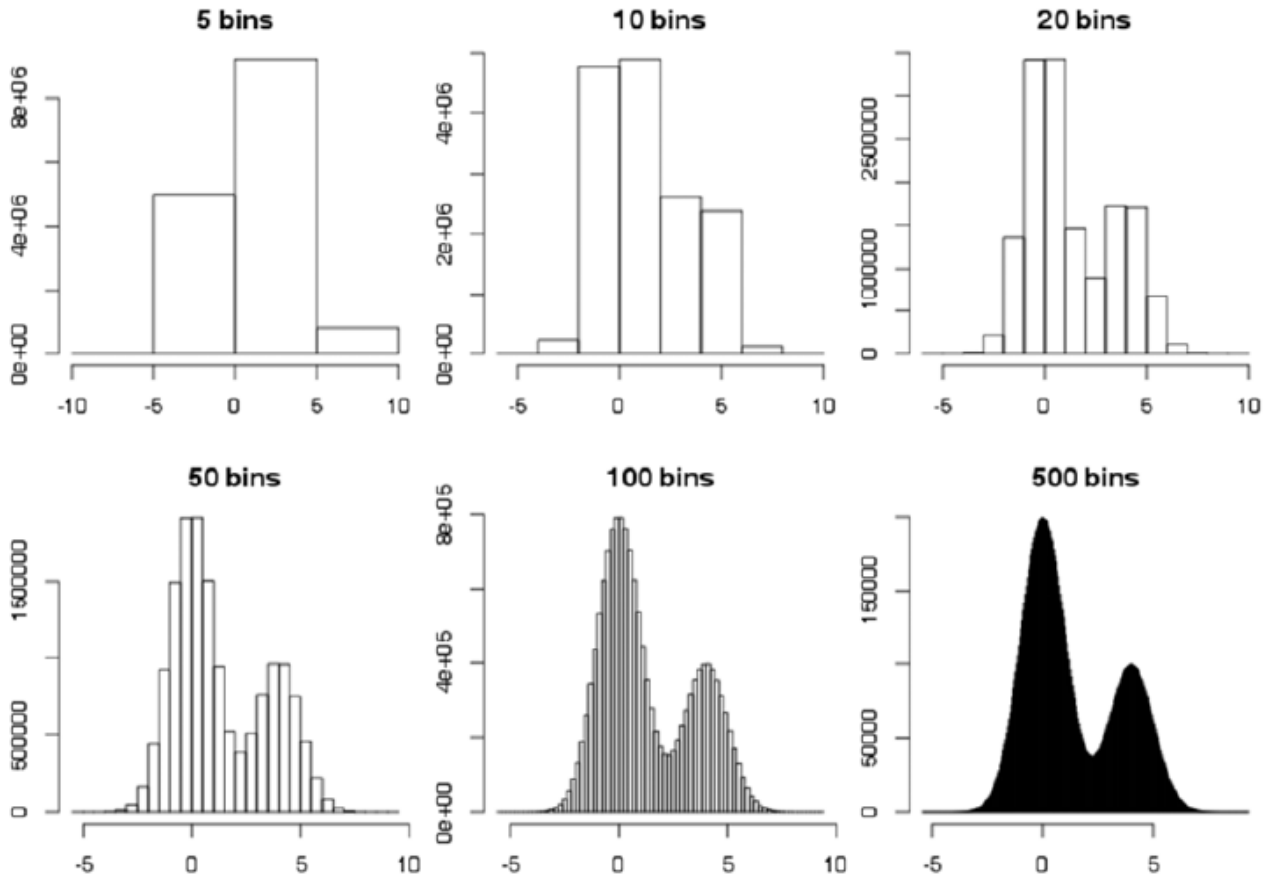


In R: hist(data, freq=TRUE)

Alternative: histograms

In R: hist(data, freq=FALSE)

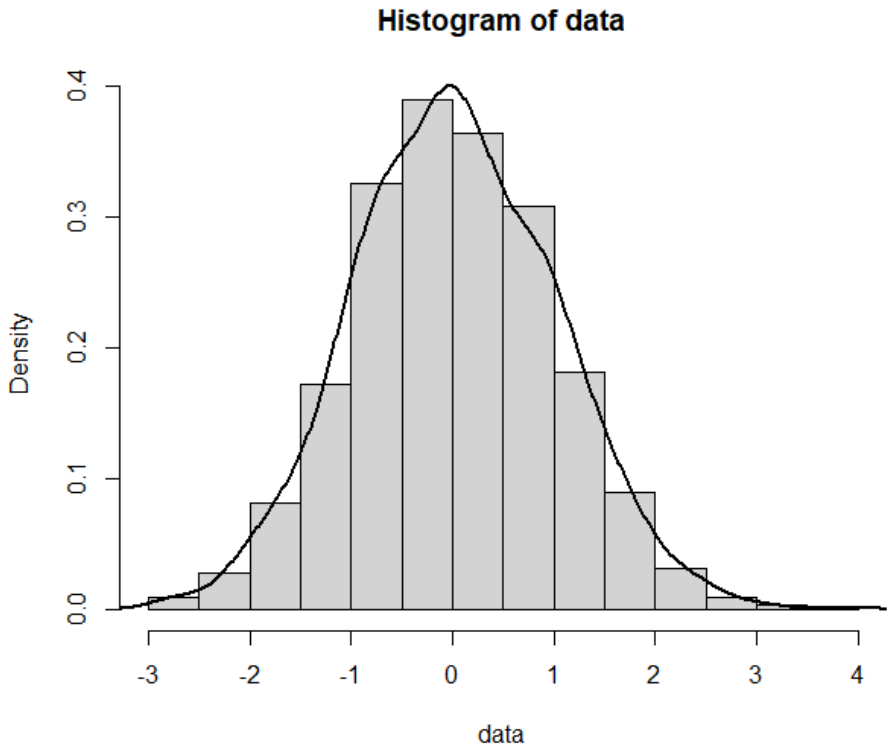
Alternative: histograms



In R: hist(data, breaks=20)

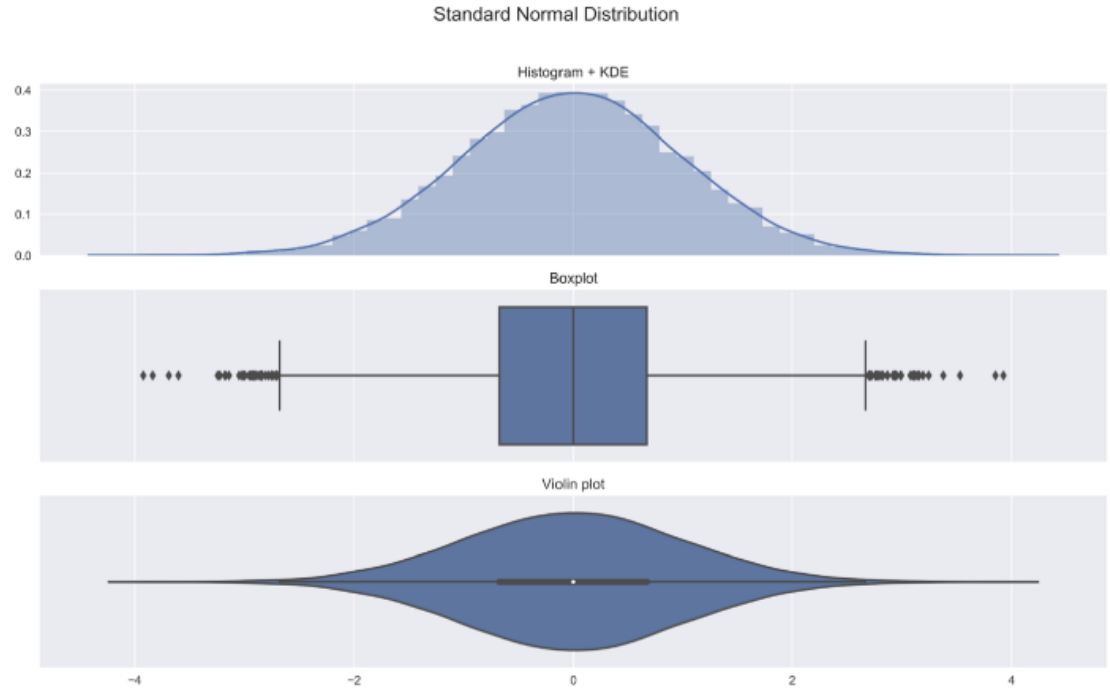
Alternative: histograms with density

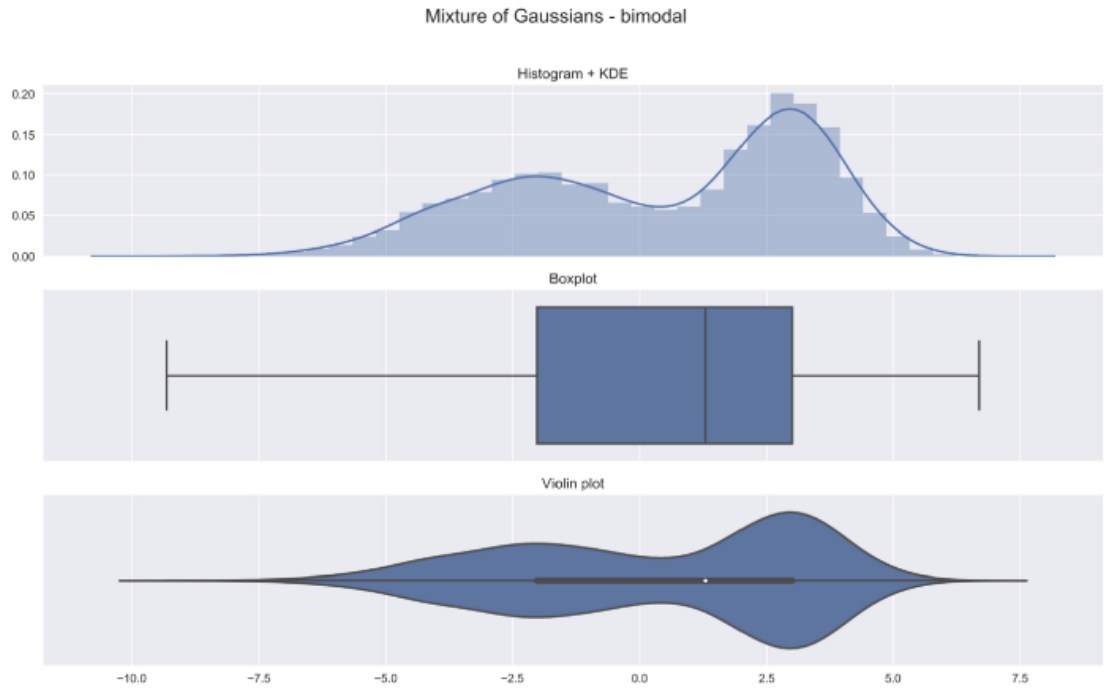
- The density describes the theoretical probability distribution of a variable
- Conceptually, it is obtained in the limit of infinitely many data points
- When we estimate it from a finite set of data, we usually assume that the density is a smooth function
- You can think of it as a “smoothed histogram”



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

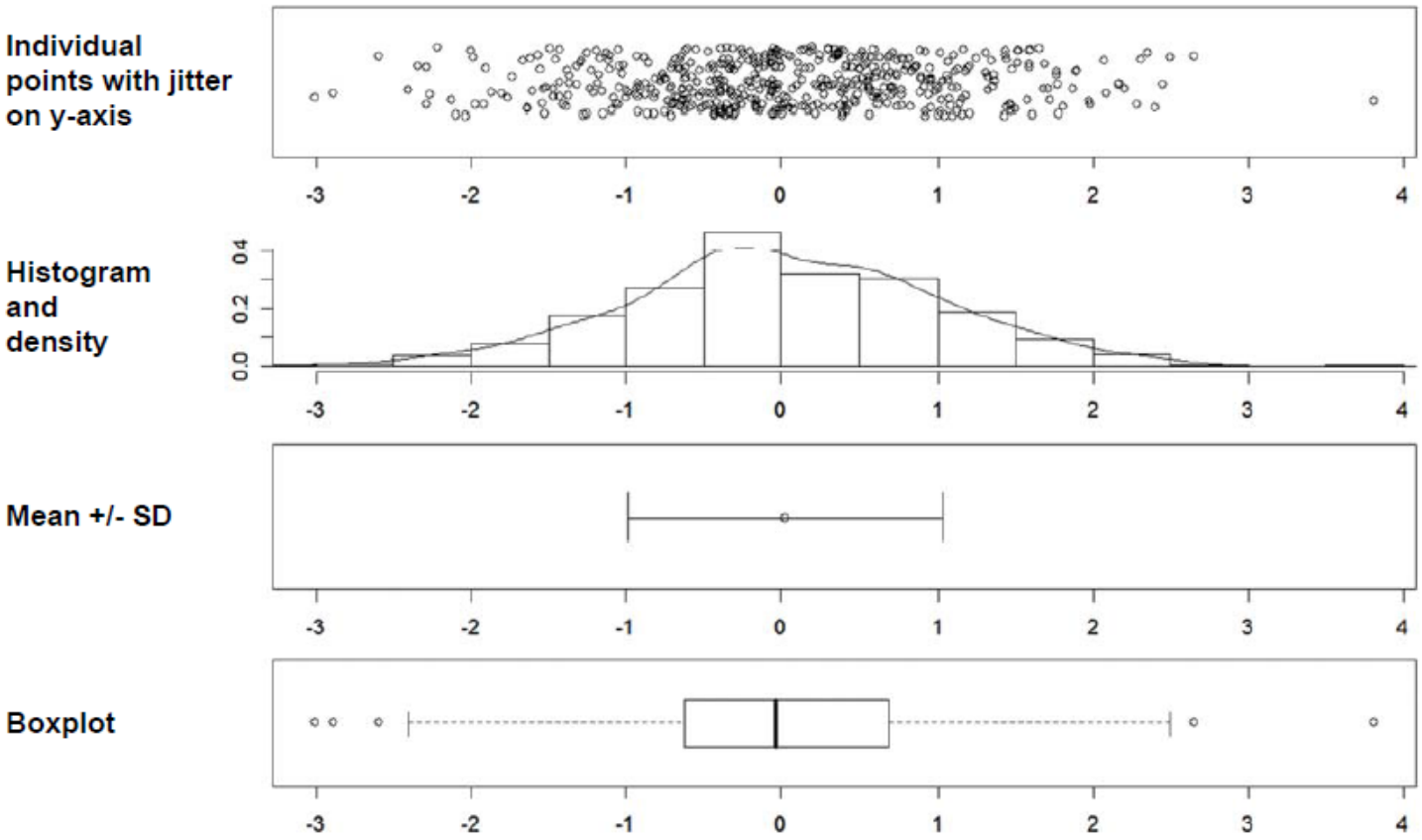
Comparisons of some graphs





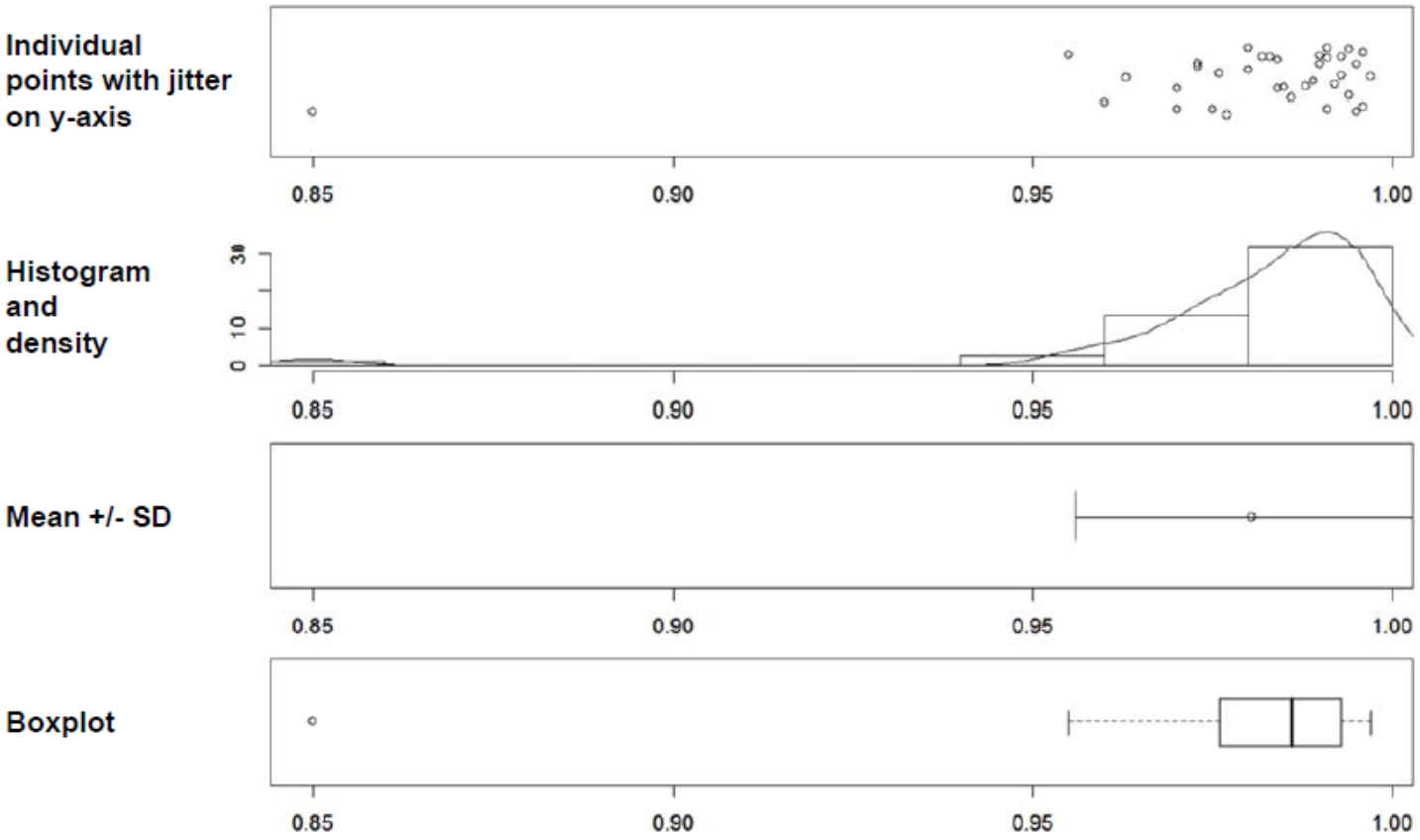
Comparisons of some graphs

Dataset 1 (500 points)



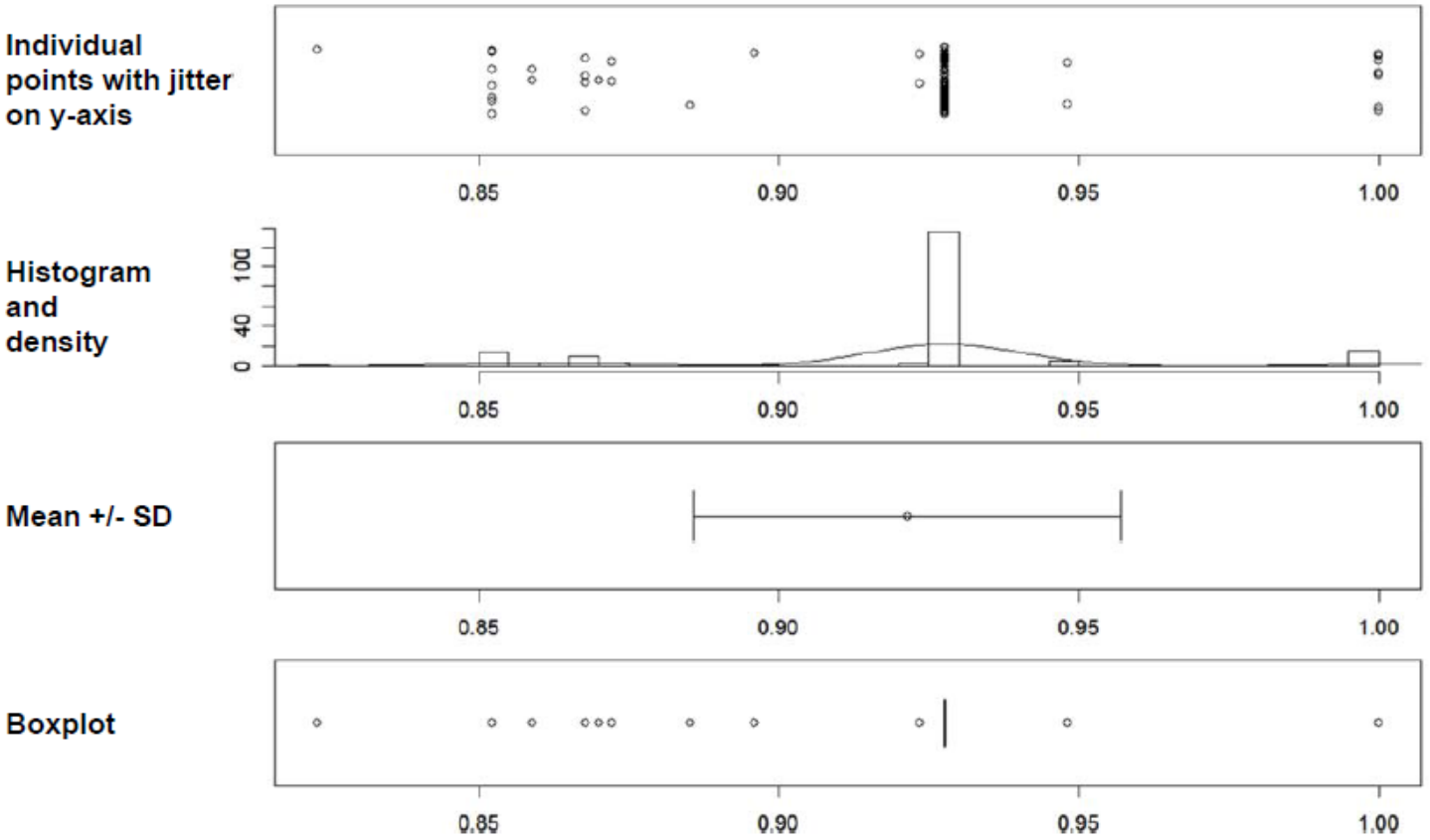
Comparisons of some graphs

Dataset 2 (37 points)



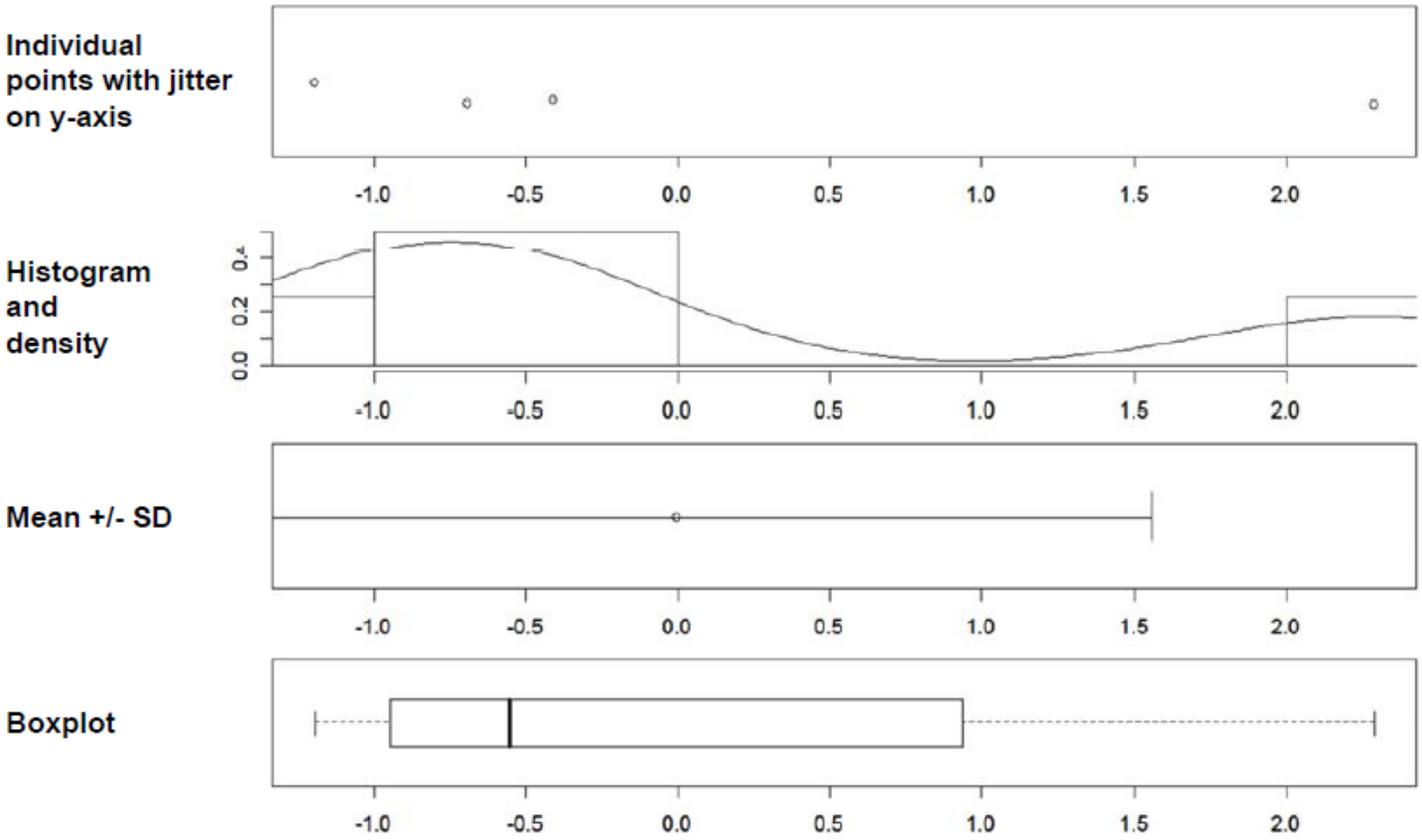
Comparisons of some graphs

Dataset 3 (100 points)



Comparisons of some graphs

Dataset 4 (4 points)



Bivariate and multivariate data