Introduction to Statistics

Joao Lourenco and Rachel Marcone
6-9th February 2023

SIB
Swiss Institute of
Bioinformatics
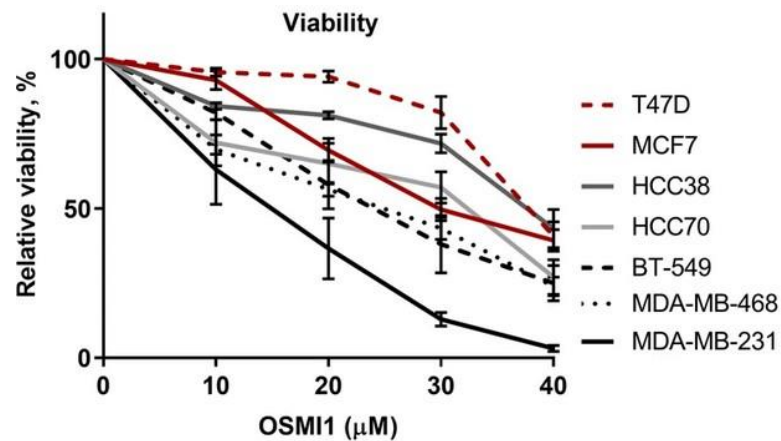
What type of graphics do you know ?

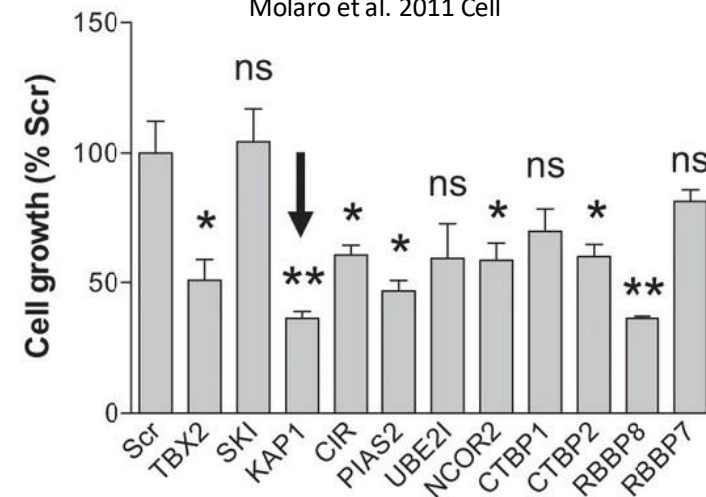# Error bars are ubiquitous in the scientific literature
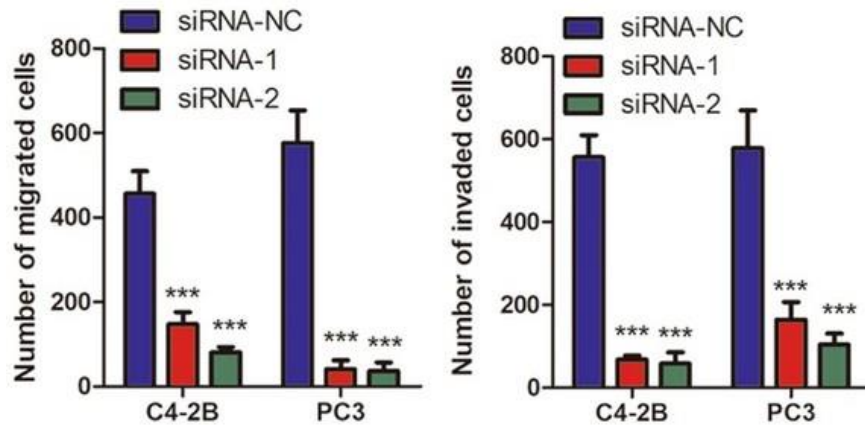
Cao et al. 2021 Cell Death & Disease
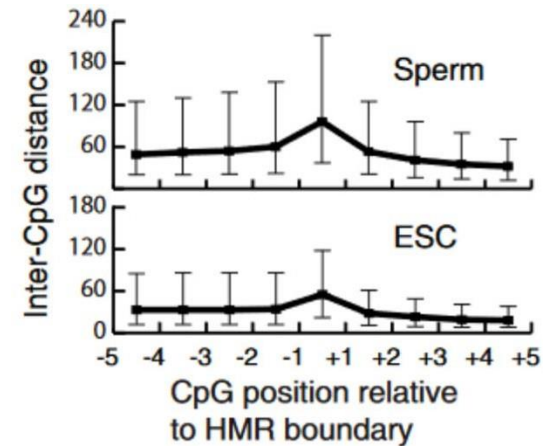
Molaro et al. 2011 Cell

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
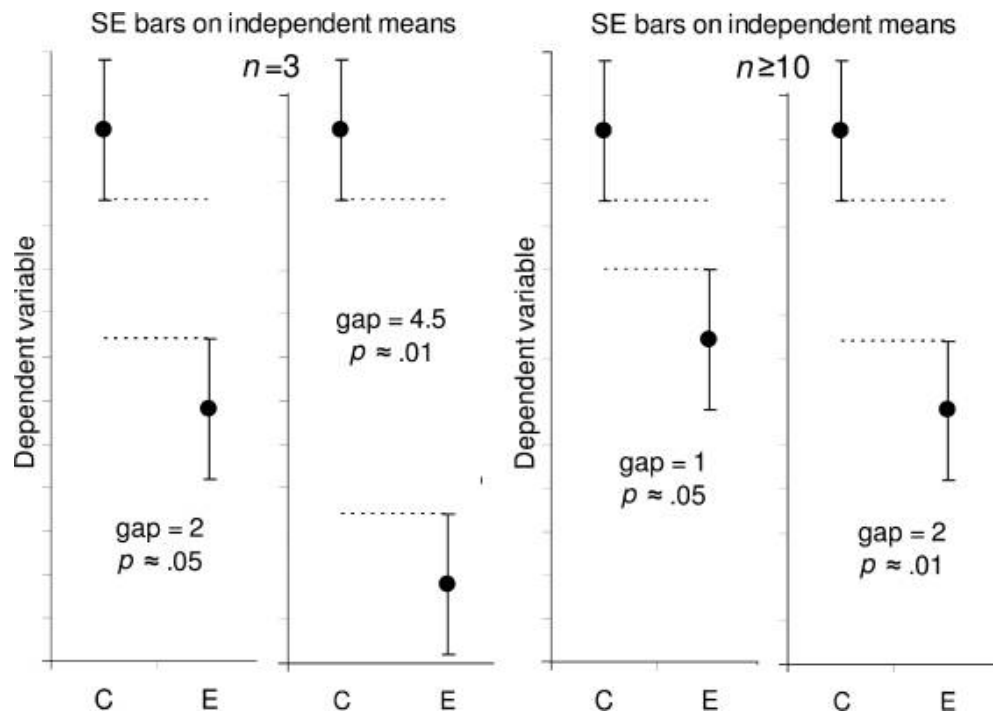from January 1, 2019 to March 31, 2019.

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	$SD = \sqrt{\dfrac{\sum(X - M)^2}{n - 1}}$
Standard error of the mean (SEM)	Inferential	A measure of how variable the mean will be, if you repeat the whole study many times	$SEM = \dfrac{SD}{\sqrt{n}}$
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
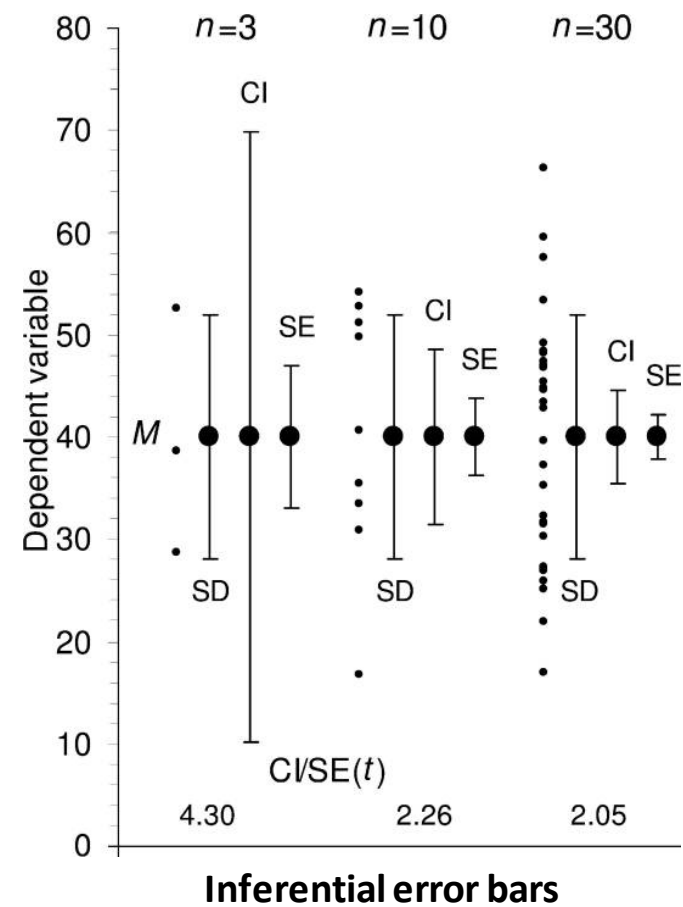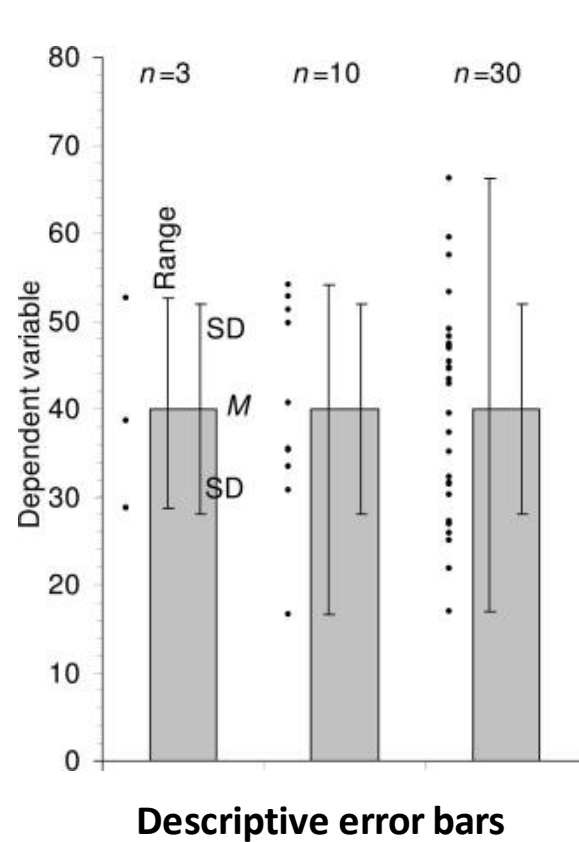
Error bars are ubiquitous in the scientific literature

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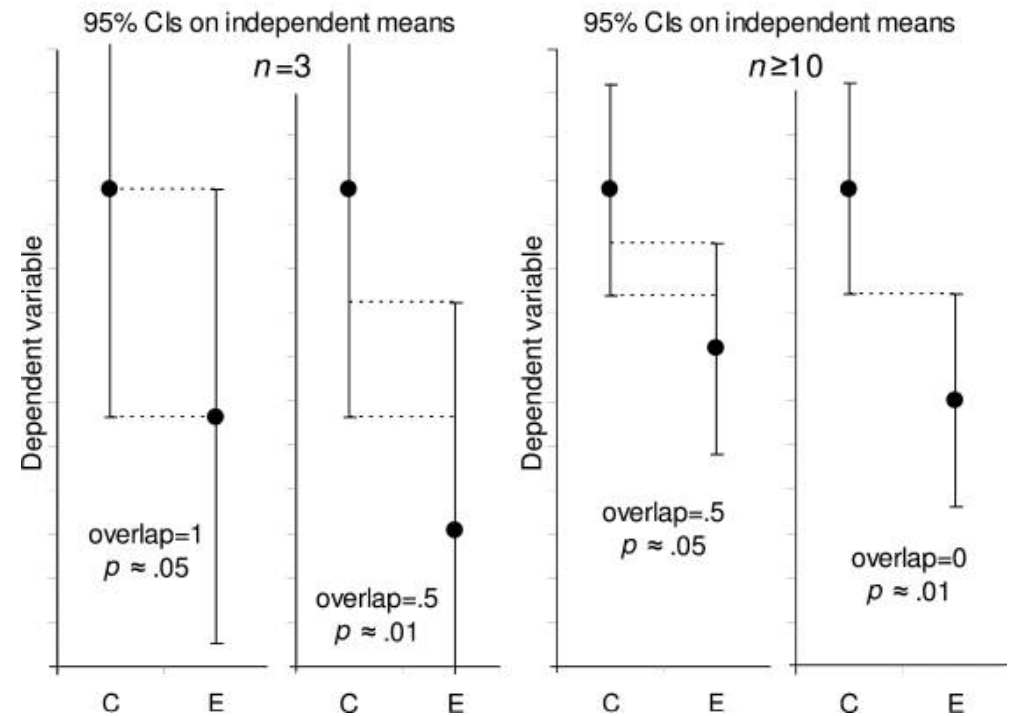
Estimating statistical significance using the overlap rule for 95% CI bars

Error bars are ubiquitous in the scientific literature

Researchers misunderstand confidence intervals and standard error bars

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

Affiliations + expand

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
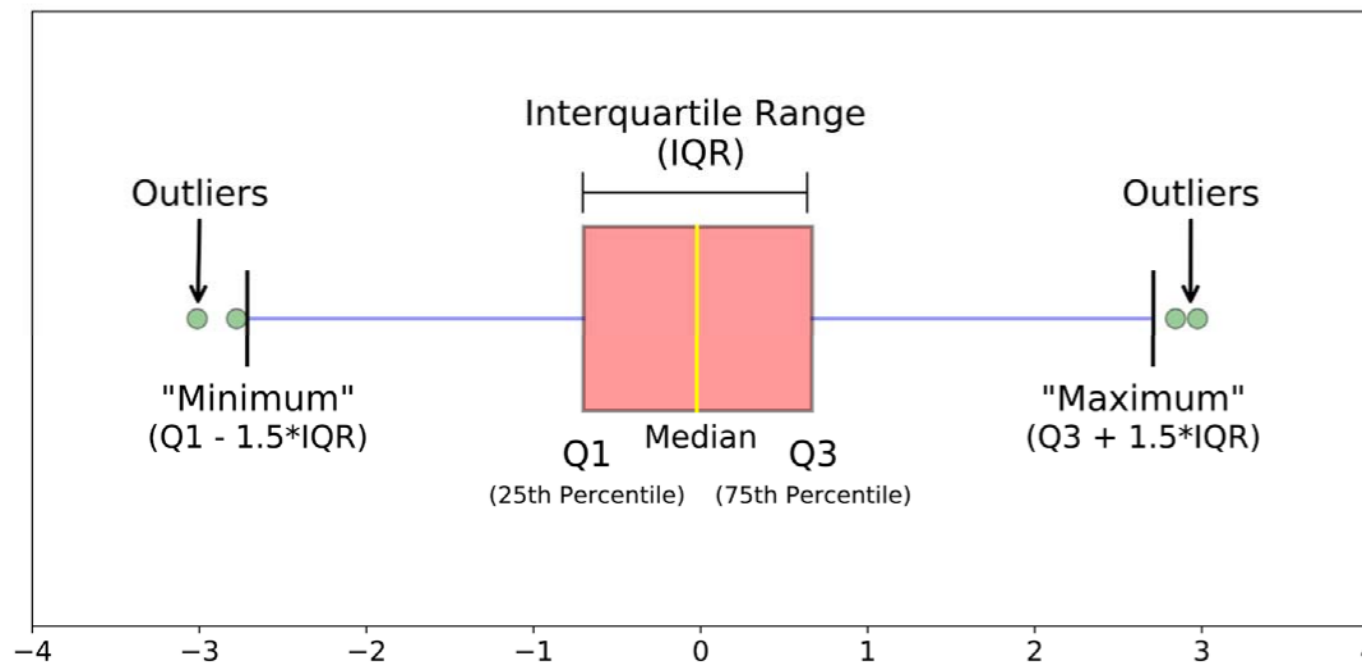
Avoid error bars if possible

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

What are the alternatives ?
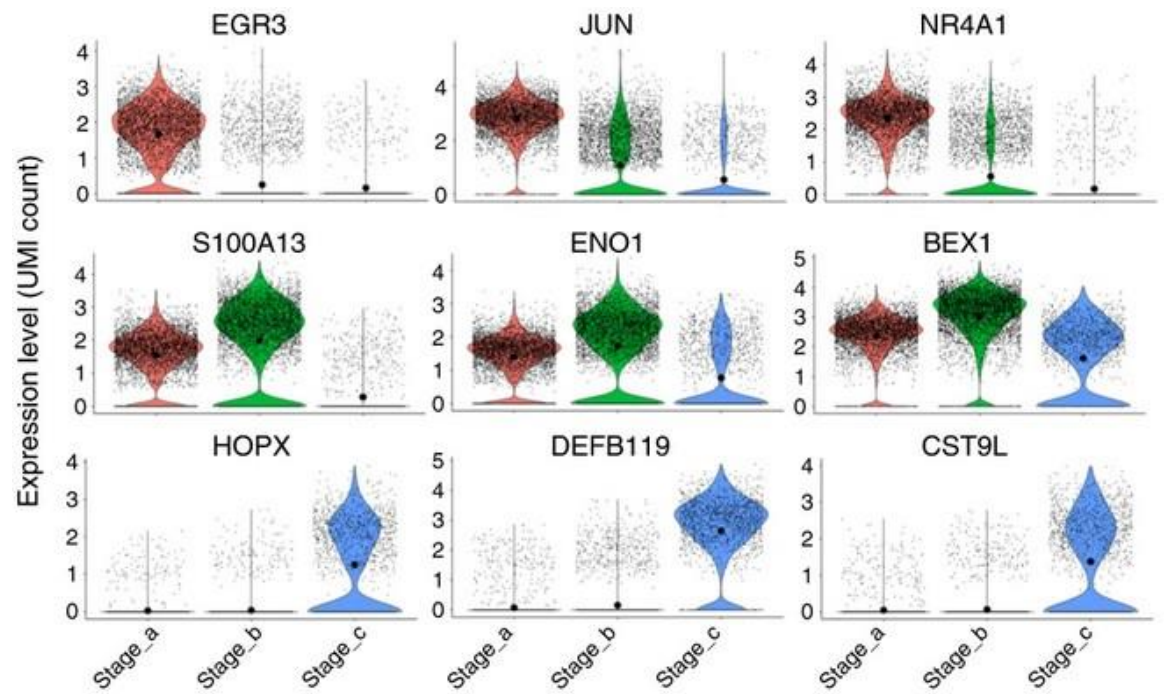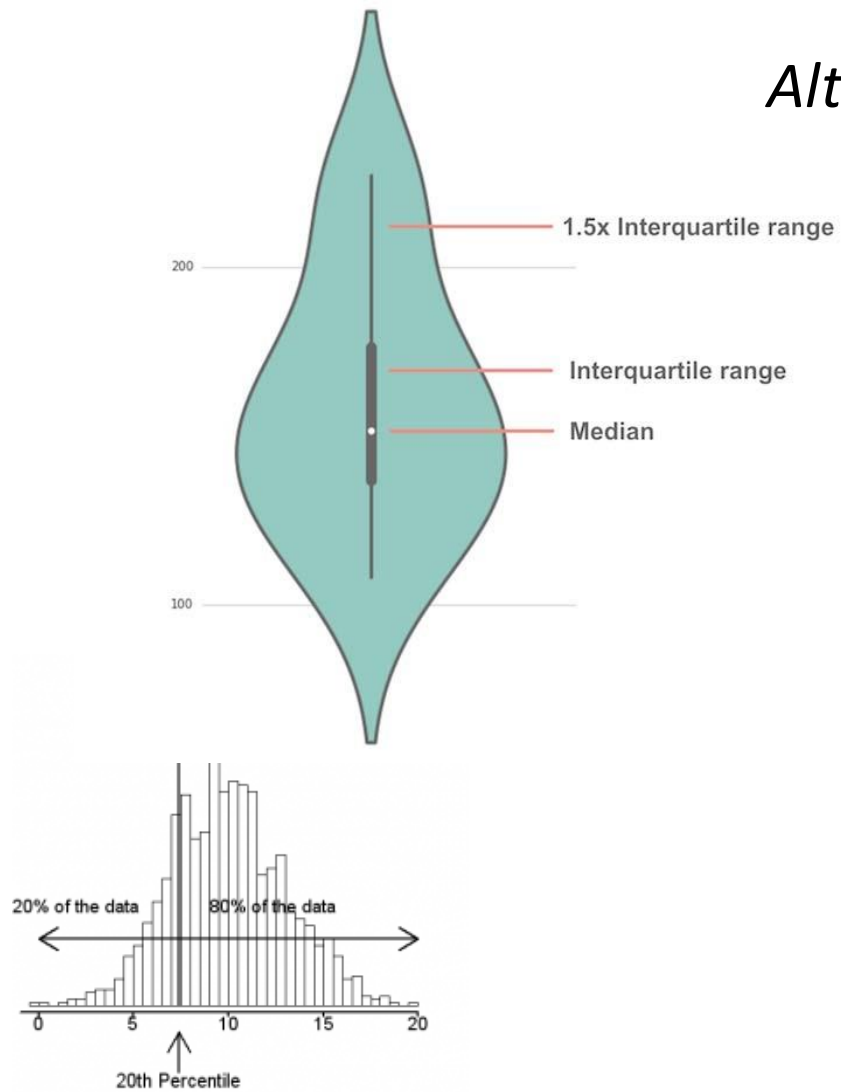
Alternative: show your data !
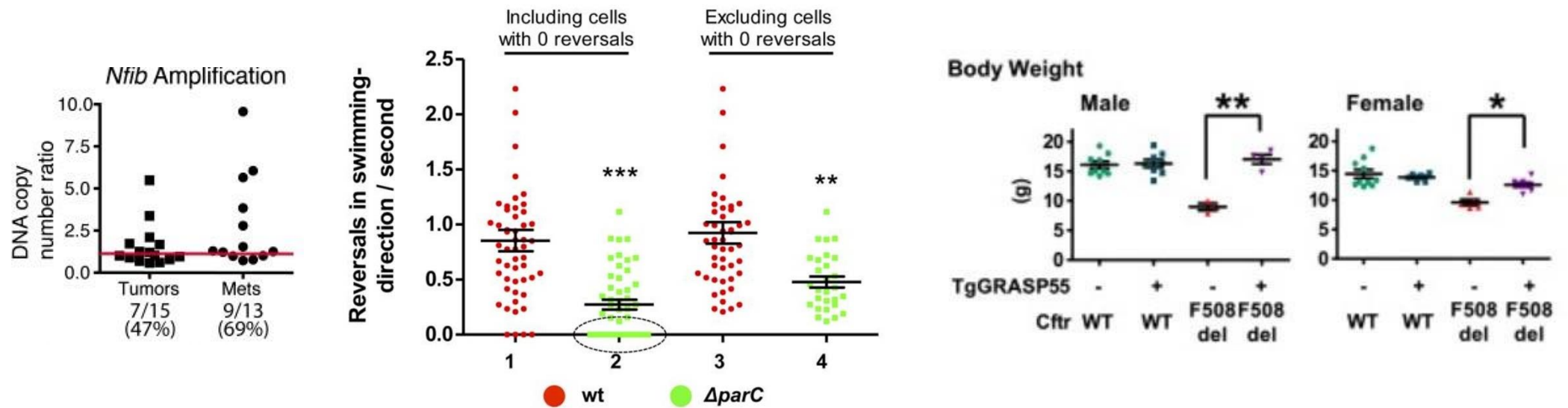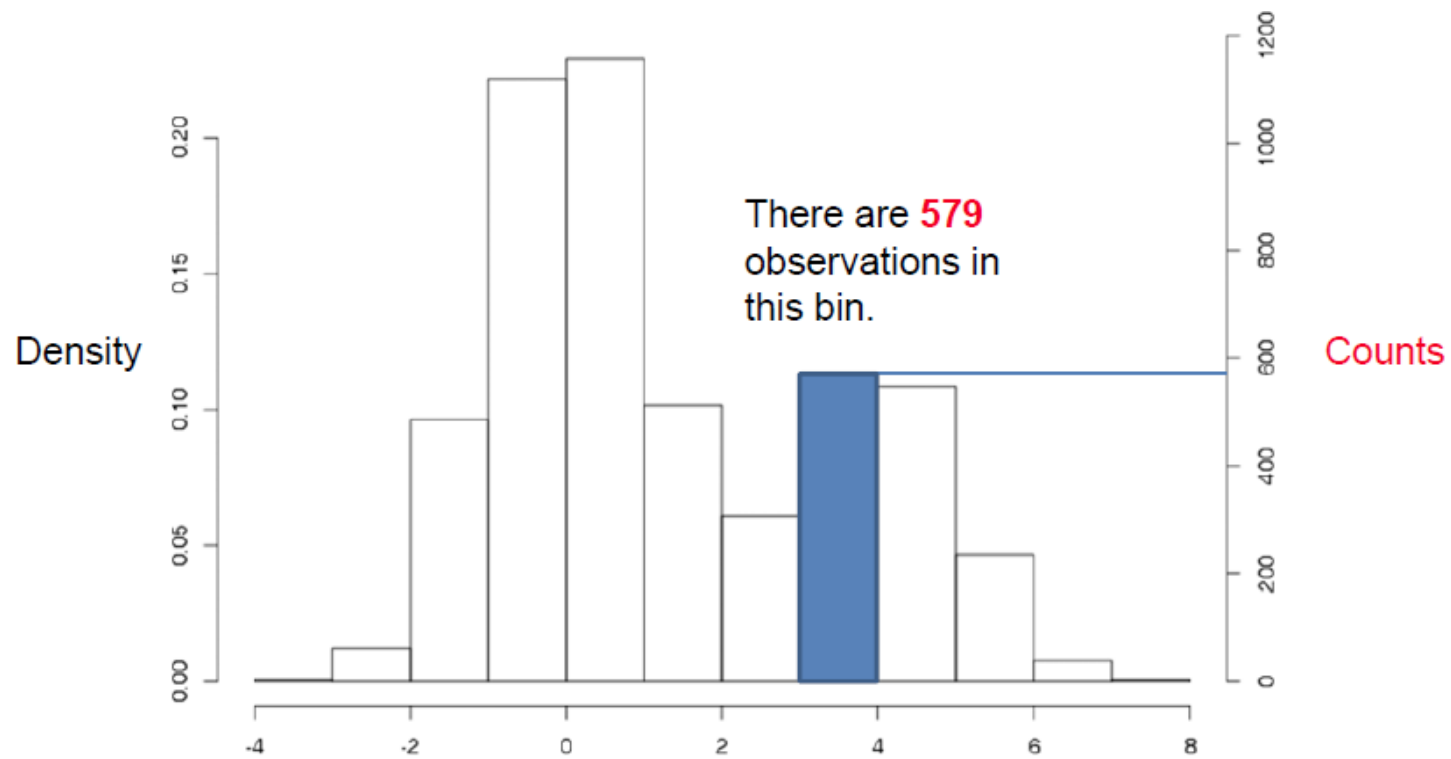
Alternative: boxplots (box and whiskers plots)
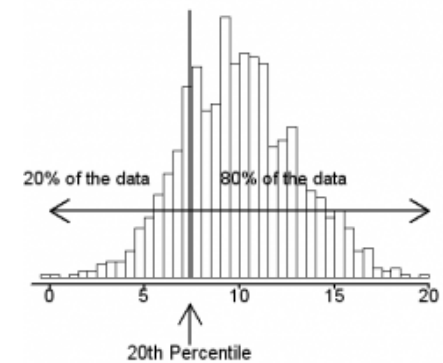
In R: boxplot(data)

Alternative: violin plots

1.5x Interquartile range

Interquartile range

Median

20% of the data   80% of the data

20th Percentile

Expression level (UMI count)

EGR3   JUN   NR4A1

S100A13   ENO1   BEX1

HOPX   DEFB119   CST9L

Stage_a   Stage_b   Stage_c

In R: library(vioplot)
vioplot(data)

Alternative: histograms

There are **579** observations in this bin.

Density

Counts

In R: hist(data, freq=TRUE)

Alternative: histograms

The area of this bin is 0.1158, so it contains **11.6%** of all observations
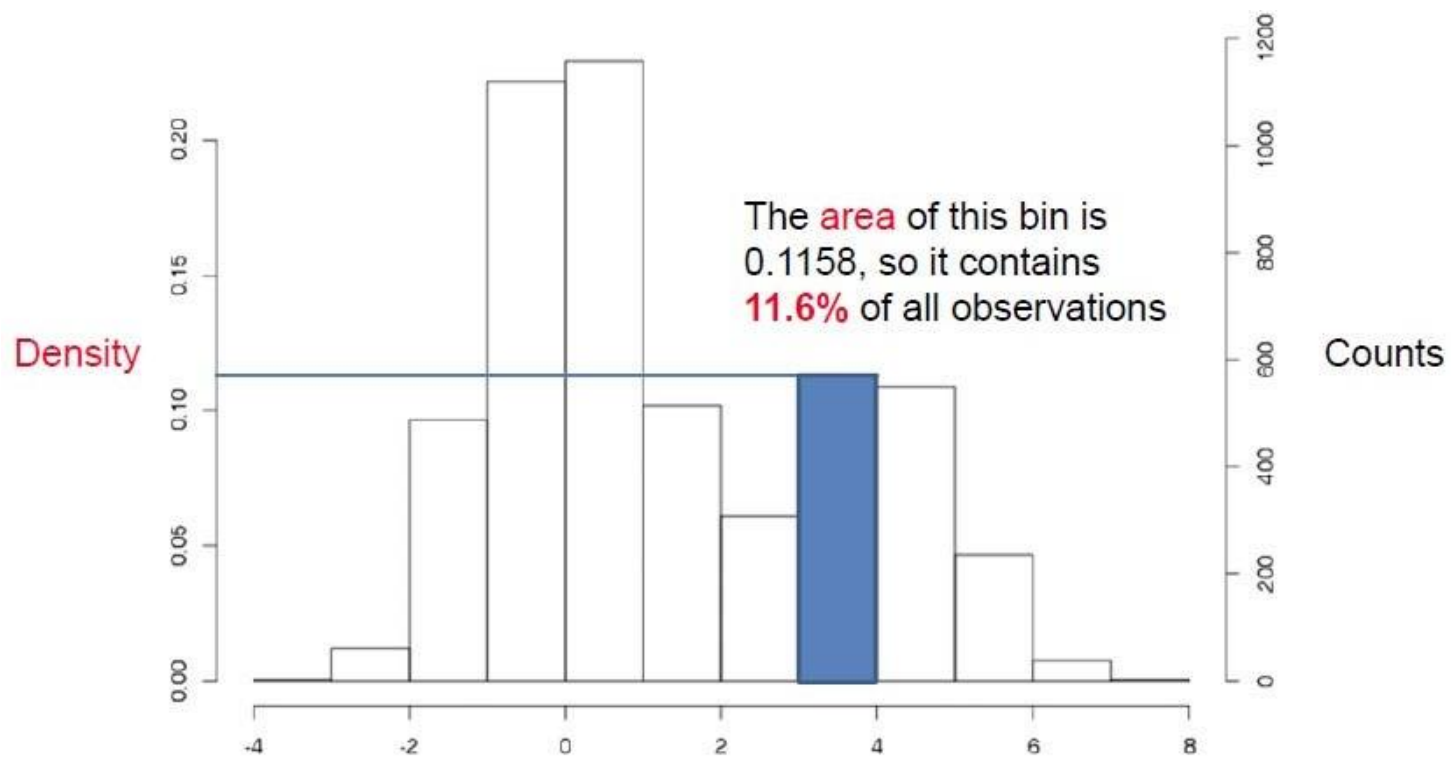
In R: hist(data, freq=FALSE)

Alternative: histograms
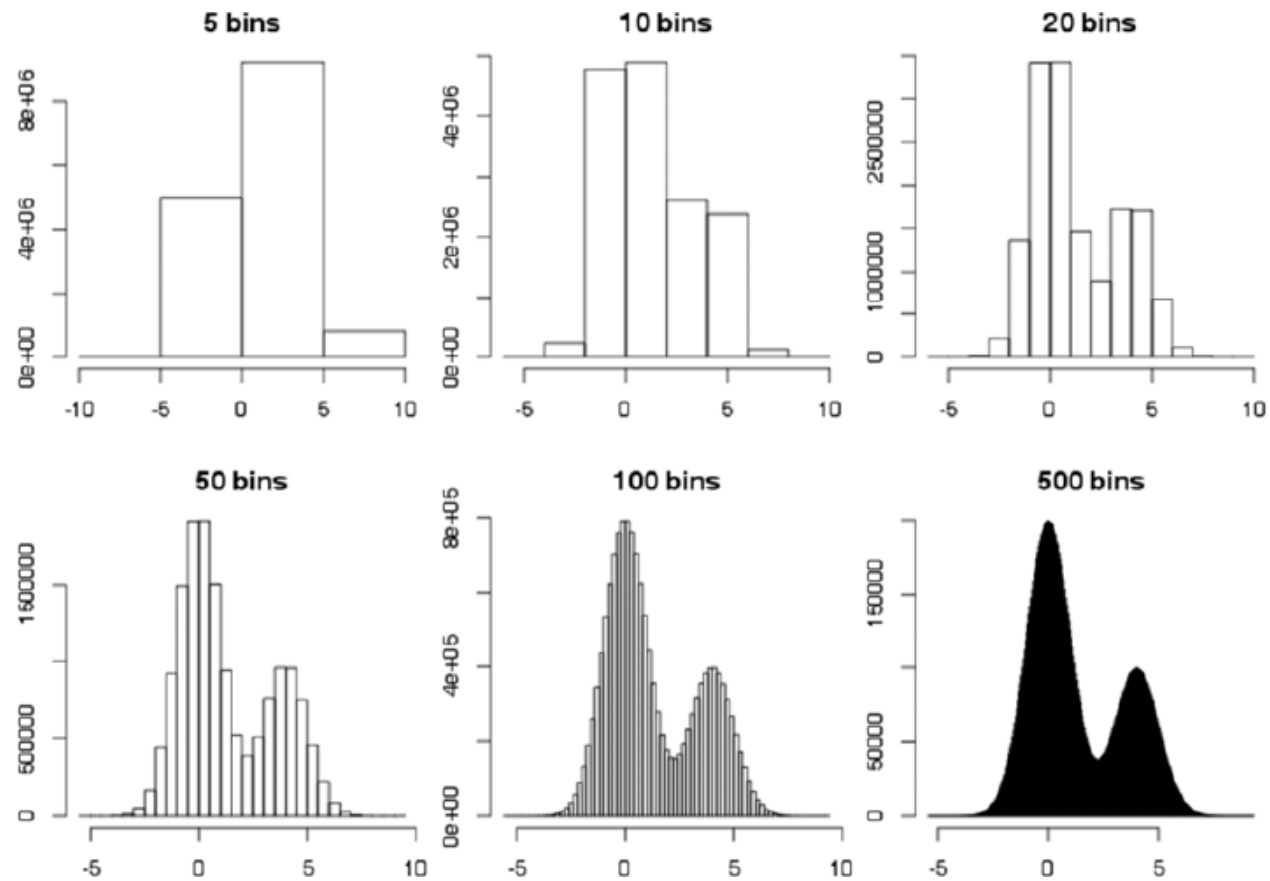
In R: hist(data, breaks=20)

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"

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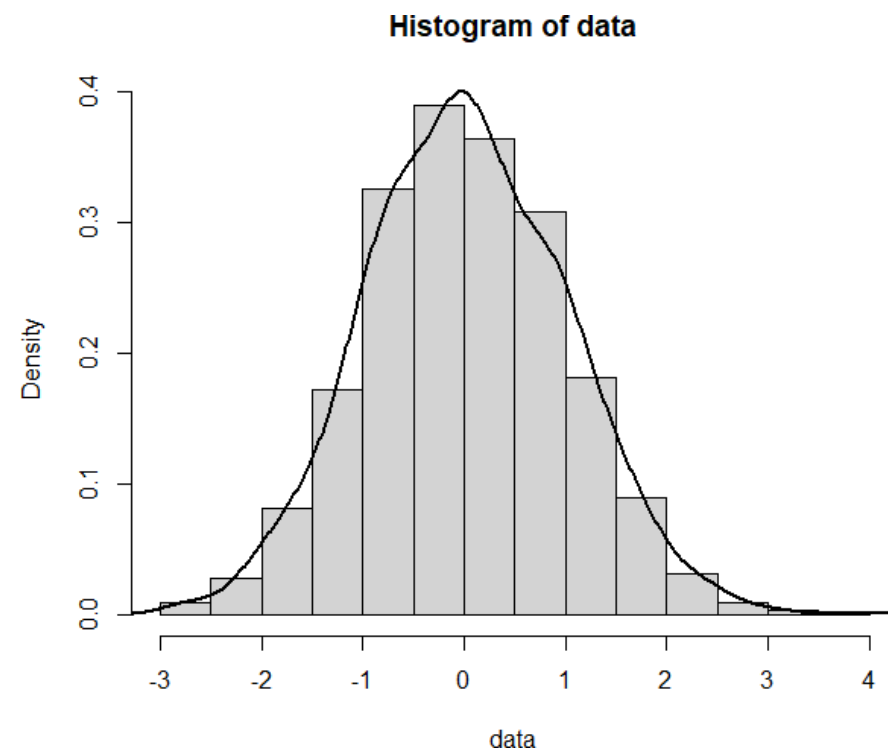
In R: hist(data, freq=F)
        lines(density(data), lwd=2)

Comparisons of some graphs

Standard Normal Distribution
- Histogram + KDE
- Boxplot
- Violin plot
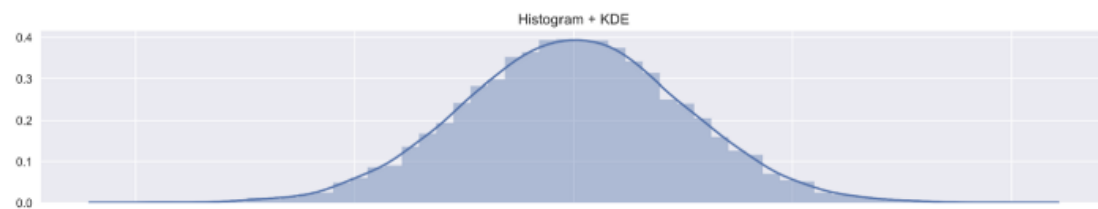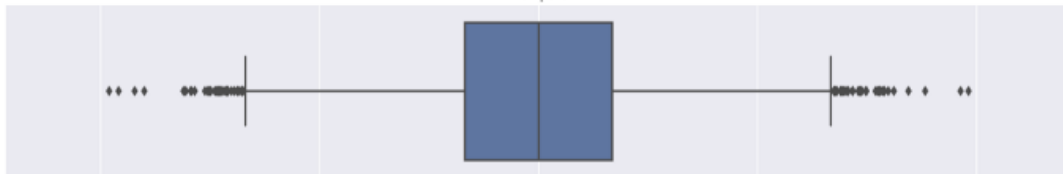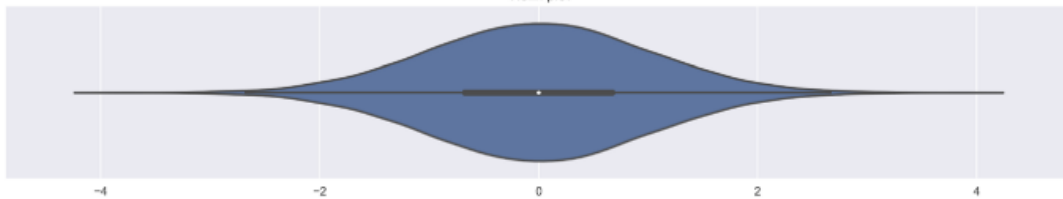
Mixture of Gaussians - bimodal
- Histogram + KDE
- Boxplot
- Violin plot
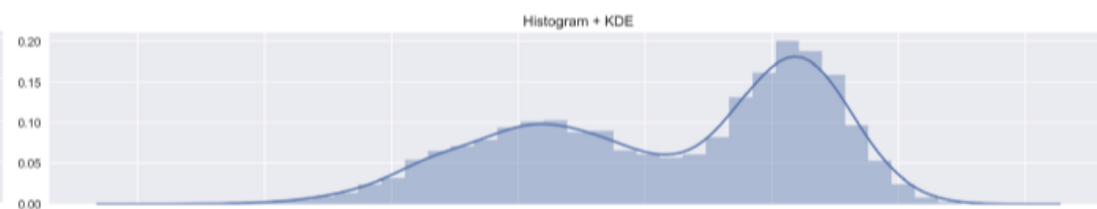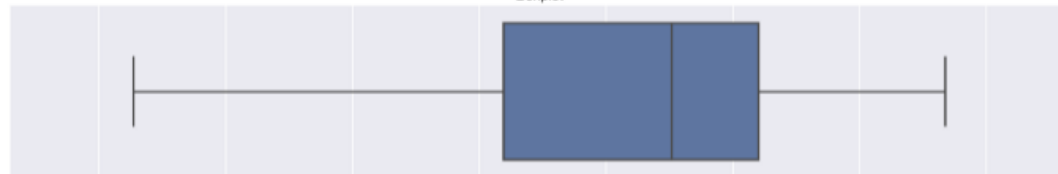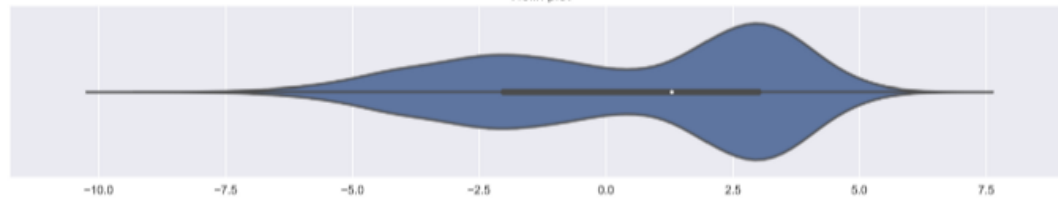
Comparisons of some graphs

Dataset 1 (500 points)
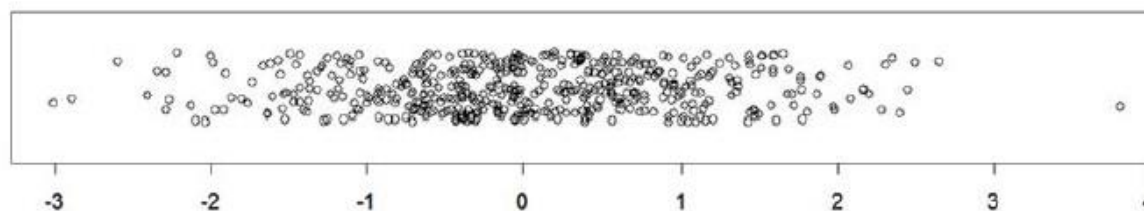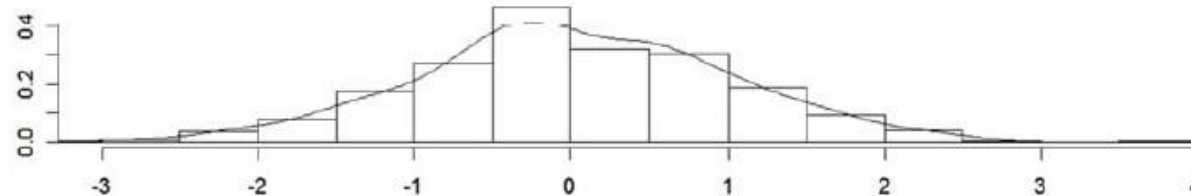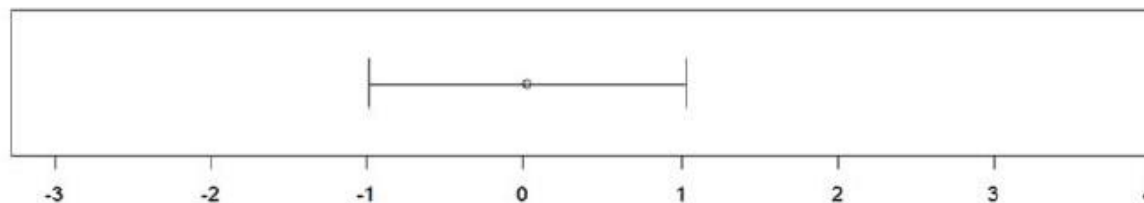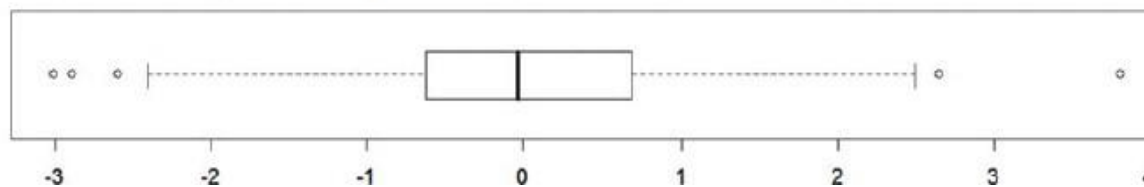
Comparisons of some graphs

Dataset 2 (37 points)

Comparisons of some graphs

Dataset 3 (100 points)
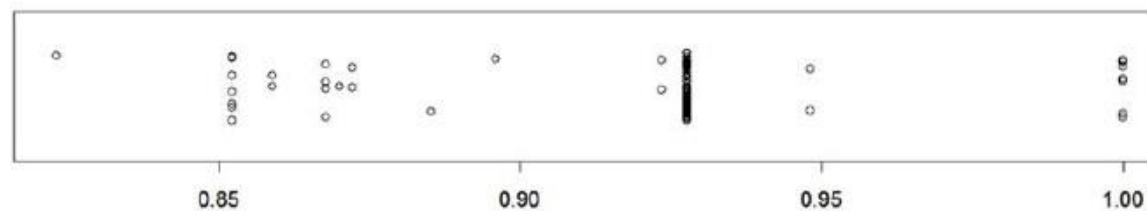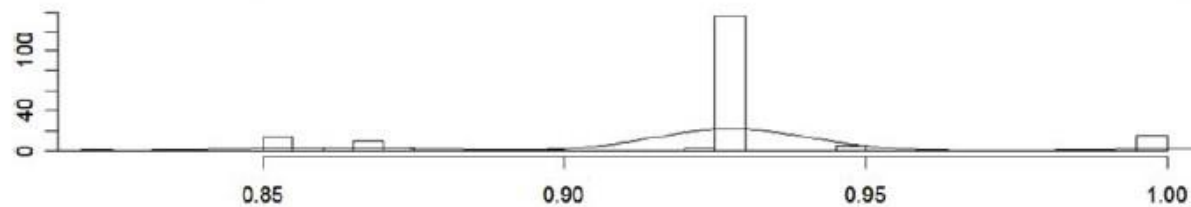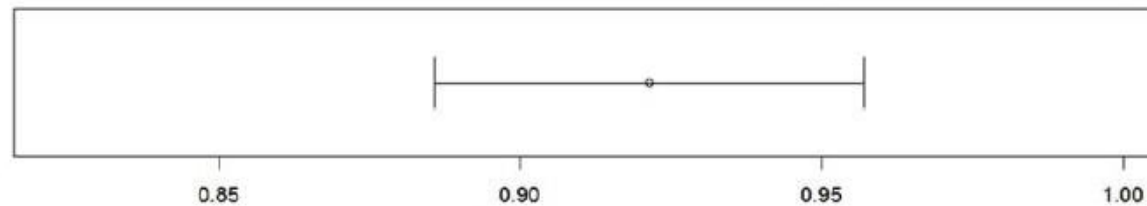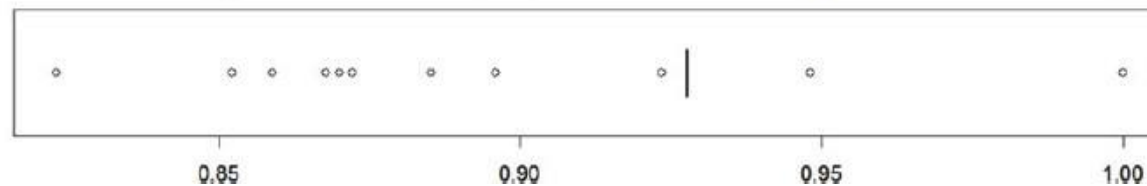
Comparisons of some graphs

Dataset 4 (4 points)
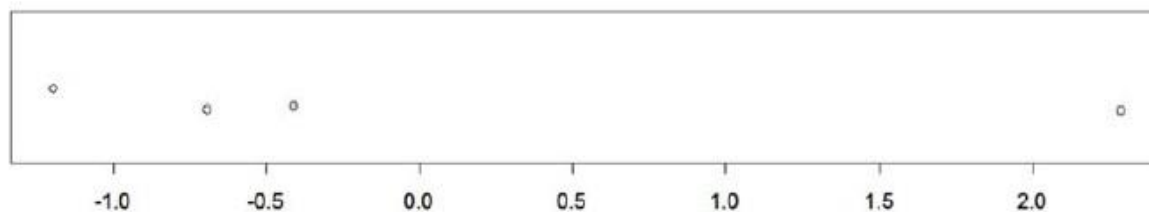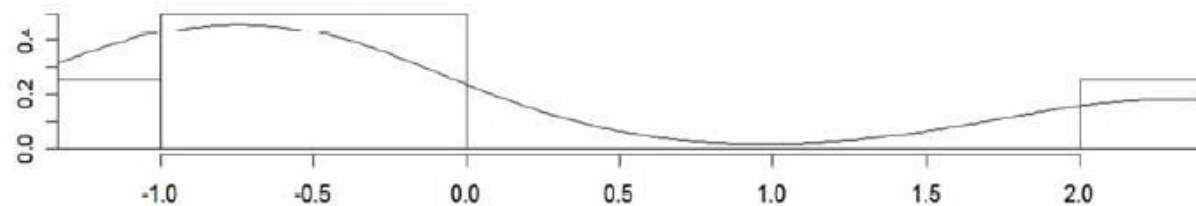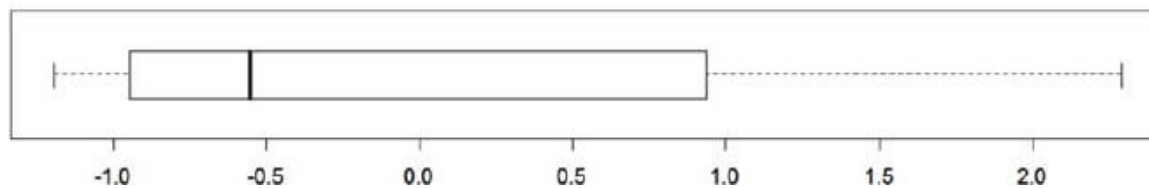
Bivariate and multivariate data