Statistical Data Analysis, Lecture 1

dr. Dennis Dobler

Vrije Universiteit Amsterdam

February 3, 2021

Lecture overview

1. course parameters
2. introduction
3. topic 1: summarizing data
4. R-demo

course parameters

People and literature

Teacher dr. Dennis Dobler, d.dobler@vu.nl
Second half: dr. Paulo Serra will take over

Assistants
Alexandra Vegelien, Anna Tsachouridi, Francesca Candelora, Nikki Kramer, Luminita Maxim, Misho Yanakiev.
No fixed office hours. Contact via Canvas; more details: course manual!

Literature On Canvas: Syllabus, R-manuals, lecture handouts

Pre-requisite Basic knowledge of statistics and probability theory (e.g. *Statistics (X_400004)* and *Probability Theory (X_400622)*)

Lectures and assignments

Lectures Study videos and in parallel Syllabus before Wednesdays.
Online open office hours (Zoom): Wed. from 10.00 (till 10.45 or earlier).
Also, use Canvas Discussion board for further information exchange with fellow students/teaching assistants/teacher.

Biweekly assignments (1st due in 1 week!)
in groups of 2 students, deadline Tue at 23.59.
No partner? Use discussion board to find one!

Assignment discussion every second Wednesday, in separate videos

Groups & practical classes (normally Fri, with some exceptions!)
100 previously created groups on Canvas; join one of them!

Assignments done in a previous year & want to keep assignment grade?
E-Mail me before Feb 2!

Solution format

Format: Read AssignmentFormat2021.pdf!

Concise.

Clear.

Complete.

Figures.

R code.

Software

R: `www.r-project.org`

RStudio: `www.rstudio.com`

Install on your computer, use them to solve the assignments!

Exam and course grade

Exam Two exams (March 26, May 26), or one resit exam (July ??)

Content exam The entire syllabus and lecture notes

Grade $= \frac{E+A}{2}$, where

- $E = \frac{E_1 + E_2}{2}$ or $E =$ resit exam grade,
- $A =$ average assignment grade.

Grade condition:
$\min(A, E) \geq 5.5$. Otherwise Grade$= \min(A, E)$.

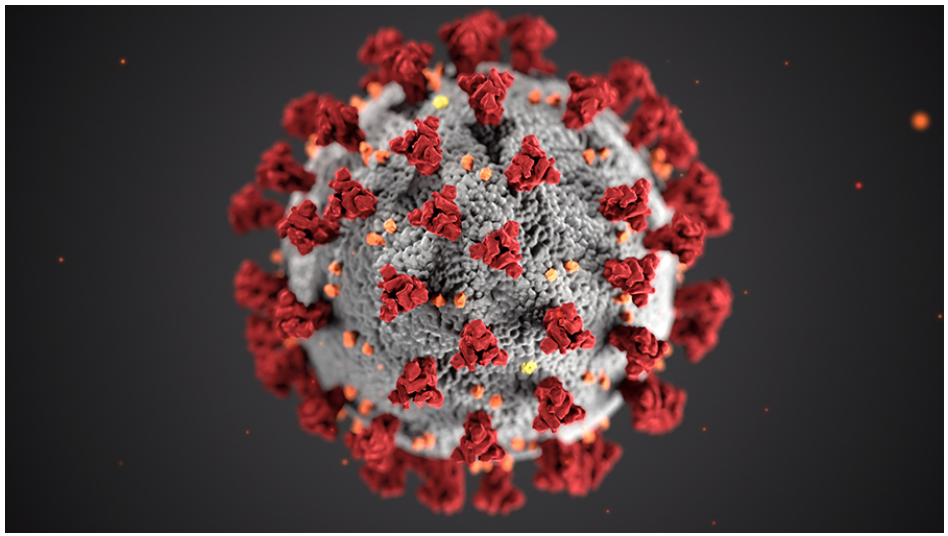
introduction

What is statistics?

Statistics: collecting, analyzing, and interpreting data

Present in

- industry
- medical studies
- scientific research
- politics
- climate change
- ...



A statistical study

Stages:

- Research question
- Experimental design
- Data collection
- Data analysis
- Interpretation of results
- Presentation of results & conclusions

Conditions

- Theoretical (lectures, syllabus)
- Practical (assignments, R)

Course overview

Data analysis:

1. Summarizing data
2. Exploring distributions
3. Density estimation
4. Bootstrap methods
5. Nonparametric tests
6. Analysis of categorical data
7. Multiple linear regression

data types

Chapter 2: Summarizing data

Contents of Chapter 2

- data types
- summary types
  - univariate data
    - numerical summary
    - graphical summary
  - bivariate data
    - numerical summary
    - graphical summary
  - multivariate data

First read some data about Corona virus

```
> setwd("C:/Users/Dennis/...")
> # data from ourworldindata.org/coronavirus-data (Jan 21, 2021)
> covid_data <- read.table("owid-covid-data.csv", sep=";", header=T, dec=",")
> attach(covid_data)
> covid_data_select <- covid_data[date=="2021-01-20",c("continent", "location",
    "total_cases_per_million", "gdp_per_capita", "human_development_index")]
```

First 6 rows:

```
  > head(covid_data_select)
        continent            location total_cases_per_million gdp_per_capita human_development_index
332          Asia         Afghanistan                1394.306       1803.987                   0.498
650        Europe             Albania               24059.351      11803.431                   0.785
981        Africa             Algeria                2385.485      13913.839                   0.754
1306       Europe             Andorra              120468.517             NA                   0.858
1613       Africa              Angola                 580.93       5819.495                   0.581
1927 North America Antigua and Barbuda                1940.201      21490.943                    0.78
```

Data types

data: quantified measurements, stored in variables

variable: varying outcome of a characteristic

Variables

- scales,
- univariate, bivariate, or multivariate,
- (in)dependent.

Measurement scales of variables

qualitative

- nominal (e.g. `continent`: Asia, `location`: Afghanistan)
- ordinal (e.g. `human_development_index`: 0.498)

quantitative

- discrete
  - interval (e.g. `date`: 2021-01-20)
  - ratio (e.g. `total_cases`)
- continuous
  - interval (e.g. `date_exact_time`)
  - ratio (e.g. `gdp_per_capita`: 1803.987)

Other partitions of variables

No. of characteristics

1: univariate

2: bivariate

$\geq$ 2: multivariate

Role

- dependent: variable of interest
- independent: background information

univariate summaries

Example

Example `total_cases_per_million`

1394.306, 24059.351, 2385.485, 120468.517, 580.930, 1940.201 . . .

- Scale?
- Good summary?

Data summaries

- location, scale
- range, extremes
- "holes", modes
- symmetry

Additionally:

- rounded?
- known distribution?
- divide into groups?
- time influence?
- relationships?

Univariate data — graphical summaries

- histogram
- stem-and-leaf-plot
- empirical distribution function
- boxplot (also numerical)

Univariate data — numerical summaries

sample size		n
location	*mean*	$\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$
	α-trimmed mean	$\frac{1}{n-2[\alpha n]}\sum_{j=[n\alpha]+1}^{n-[\alpha n]} x_{(j)}, \qquad 0 \leq \alpha < \frac{1}{2}$
	median	$\text{med}(x) = \begin{cases} x_{((n+1)/2)}, & \text{if } n \text{ odd} \\ \frac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right), & \text{if } n \text{ even} \end{cases}$
scale	*variance*	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
	standard deviation	$s = \sqrt{s^2}$
	coefficient of variation	$cv = s/\bar{x}$
	median absolute deviation	$\frac{1}{\Phi^{-1}(3/4)}\text{med}\left(\lvert x_i - \text{med}(x_1, \ldots, x_n)\rvert\right)$
	range	$(x_{(1)}, x_{(n)})$
	quartiles	$\text{quart}(x)$, $3\text{quart}(x)$
	interquartile range	$3\text{quart}(x) - \text{quart}(x)$
skewness	*skewness*	$b_1 = \frac{\sqrt{n}\sum_{j=1}^{n}(x_j-\bar{x})^3}{\{\sum_{j=1}^{n}(x_j-\bar{x})^2\}^{3/2}}$
size of tails	*curtosis*	$b_2 = \frac{n\sum_{j=1}^{n}(x_j-\bar{x})^4}{\{\sum_{j=1}^{n}(x_j-\bar{x})^2\}^{2}}$

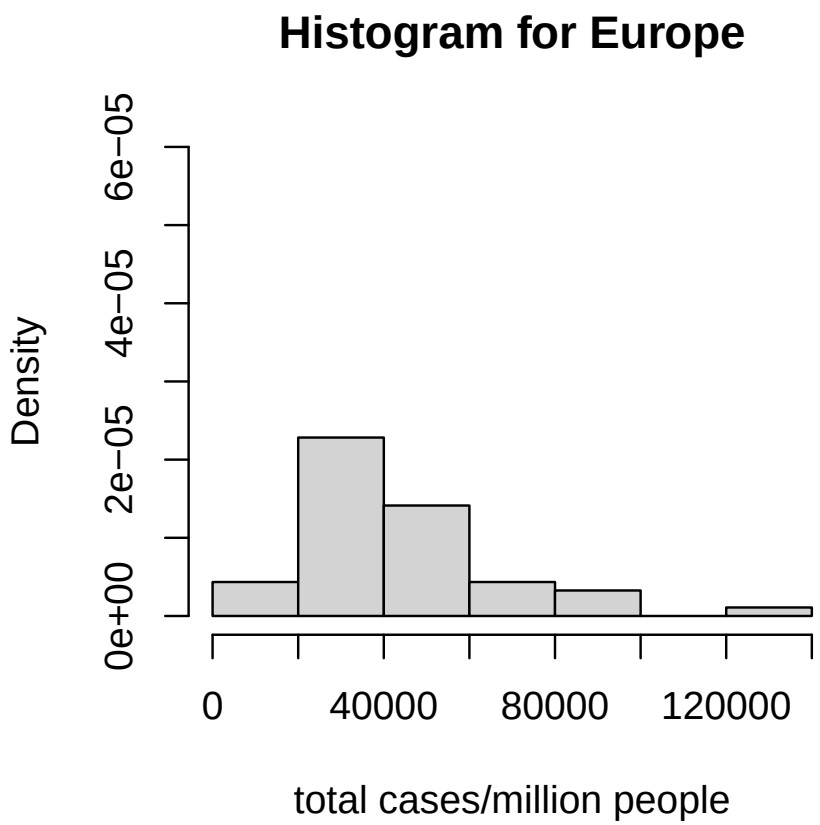
Univariate examples: total_cases_per_million

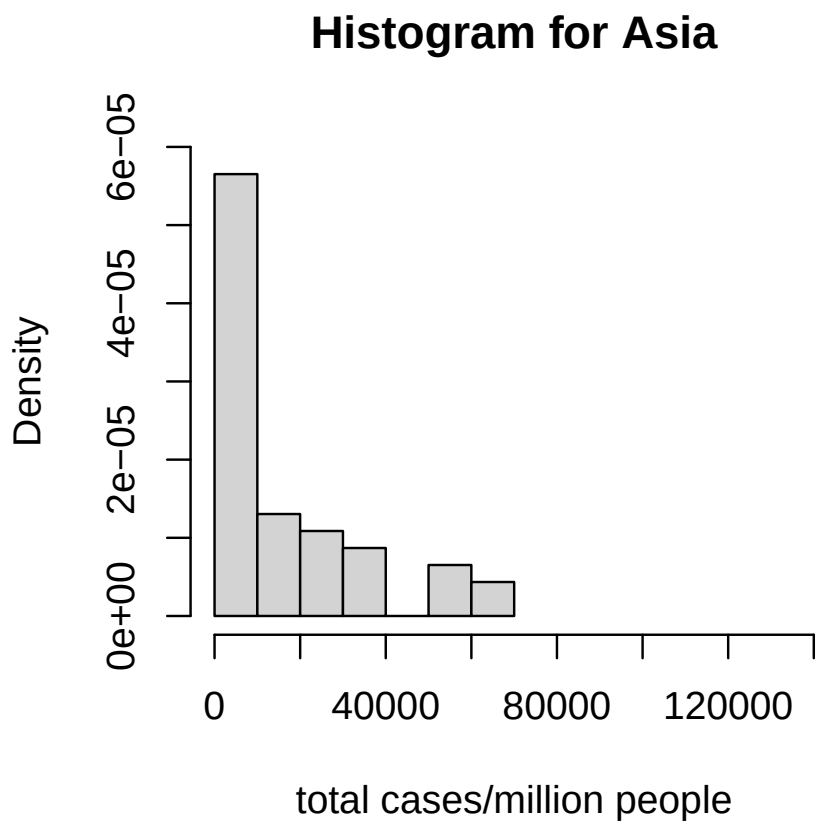
Numerical summaries

Europe		Asia	
sample size	46	sample size	46
mean	44,686	mean	15,043.98
sd	22,390.09	sd	19,155.01
var	501,316,041	var	366,914,298
min	7,430	min	5.64
1st qu.	30,664	1st qu.	1,174.77
median	38,269	median	4,925.13
3rd qu.	55,398	3rd qu.	26,476.58
max	120,469	max	66,528.71
IQR	24,734	IQR	25,301.81
NA's	0	NA's	0

Univar. examples: total_cases_per_million (Europe, Asia)

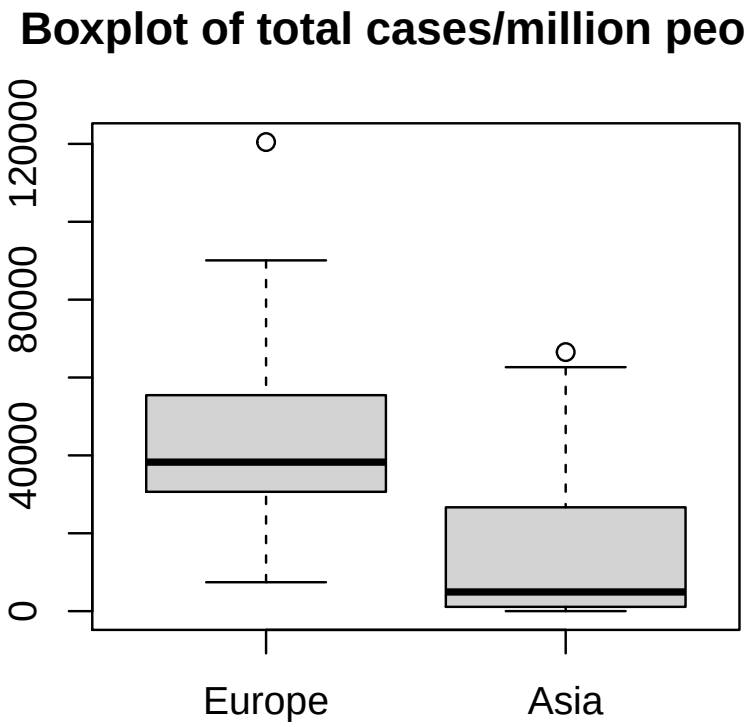
Graphical summaries (1)

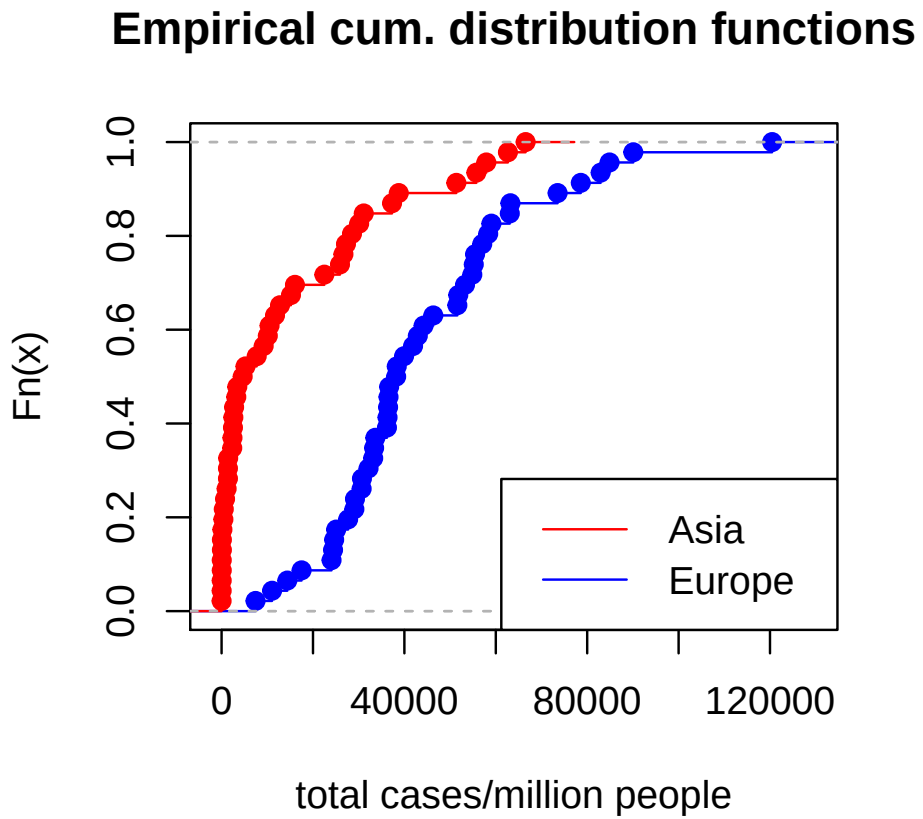




Univar. examples: total_cases_per_million (Europe, Asia)

Graphical summaries (2)





multivariate summaries

Bivariate example: total_cases_per_million & gdp_per_capita (Europe)

location	total/million	GDP/capita
Albania	24,059.351	11,803.431
Andorra	120,468.517	NA
Austria	44,201.457	45,436.686
Belarus	24,392.652	17,167.967
Belgium	59,040.438	42,658.576
Bosnia and Herzegovina	36,185.216	11,713.895
	⋮	⋮

- Good summary?

Bivariate data — graphical summaries

- scatter plot
- time plot
- contingency table (also numerical)

Bivariate data — numerical summaries

Bivariate data $(x_1, y_1), \ldots, (x_n, y_n)$.

$(r_1, \ldots, r_n)$ & $(t_1, \ldots, t_n)$: ranks.

For $z > 0$: $\operatorname{sgn}(z) = 1,\ \operatorname{sgn}(-z) = -1,\ \operatorname{sgn}(0) = 0$.

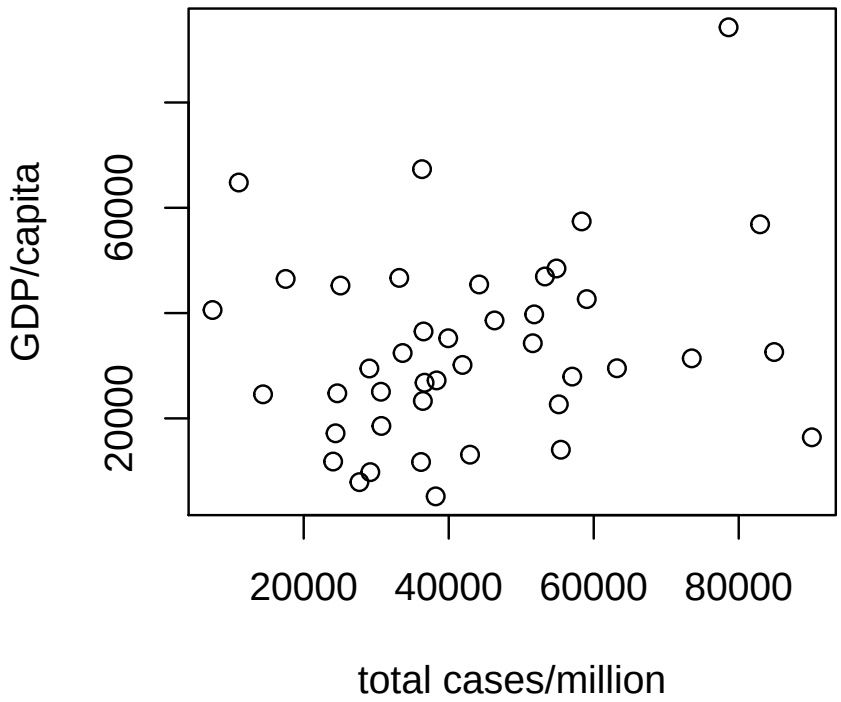
mean	$(\bar{x}, \bar{y})$
covariance	$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$
correlation coefficient	$r_{xy} = \frac{s_{xy}}{s_x s_y}$
covariance matrix	$\Sigma = \begin{pmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{pmatrix}$
Spearman's rank correlation coefficient	$r_s = \frac{\sum_{i=1}^{n} (r_i - \frac{1}{2}(n+1))(t_i - \frac{1}{2}(n+1))}{\sqrt{\sum_{i=1}^{n} (r_i - \frac{1}{2}(n+1))^2 \sum_{i=1}^{n} (t_i - \frac{1}{2}(n+1))^2}}$
Kendall's rank correlation coefficient	$\tau = \frac{\sum\sum_{i \neq j} \operatorname{sgn}(r_i - r_j)\operatorname{sgn}(t_i - t_j)}{n(n-1)} = \frac{4N_\tau}{n(n-1)} - 1$

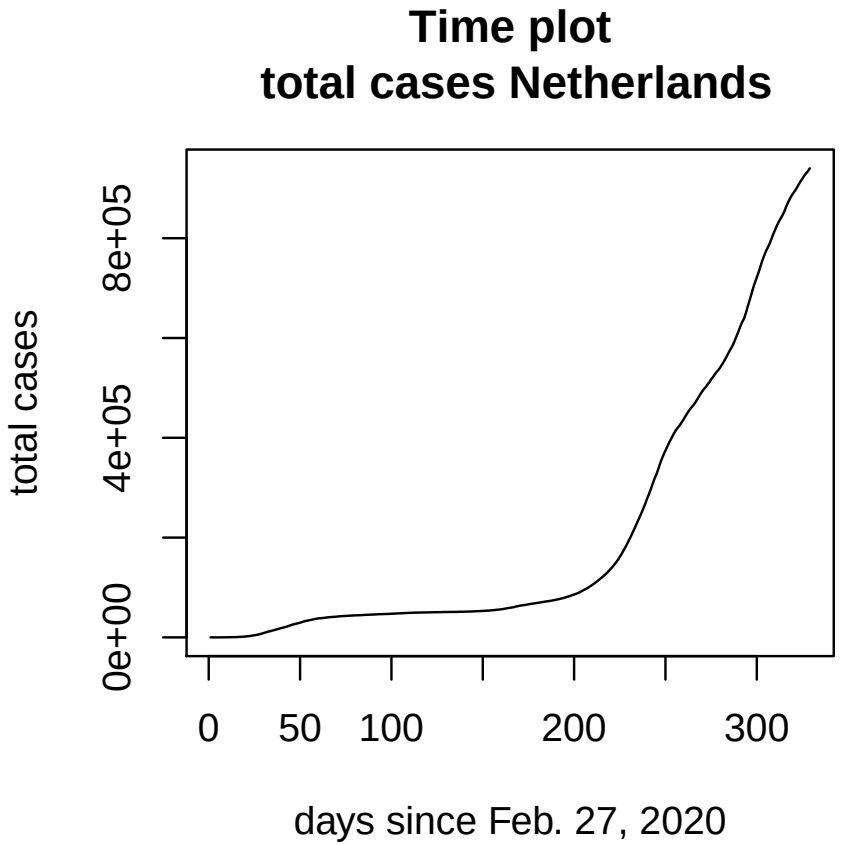
Bivariate example: matrix bivariate contains total_cases_per_million & gdp_per_capita (Europe)

```
> colMeans(bivariate, na.rm=T)
[1] 44685.52 33360.62

> bivariate_woNA <- bivariate[-which(is.na(covid_data_europe$gdp_per_capita)),]
> cov(bivariate_woNA)
          [,1]      [,2]
[1,] 392093636  74109142
[2,]  74109142 325076858
> cor(bivariate_woNA)
          [,1]      [,2]
[1,] 1.0000000 0.2075792
[2,] 0.2075792 1.0000000
> cor(bivariate_woNA, method="spearman")
          [,1]      [,2]
[1,] 1.0000000 0.2039543
[2,] 0.2039543 1.0000000
> cor(bivariate_woNA, method="kendall")
          [,1]      [,2]
[1,] 1.0000000 0.1660859
[2,] 0.1660859 1.0000000
```

Bivariate example: total_cases_per_million vs. gdp_per_capita (Europe) & total_cases by time (NL)



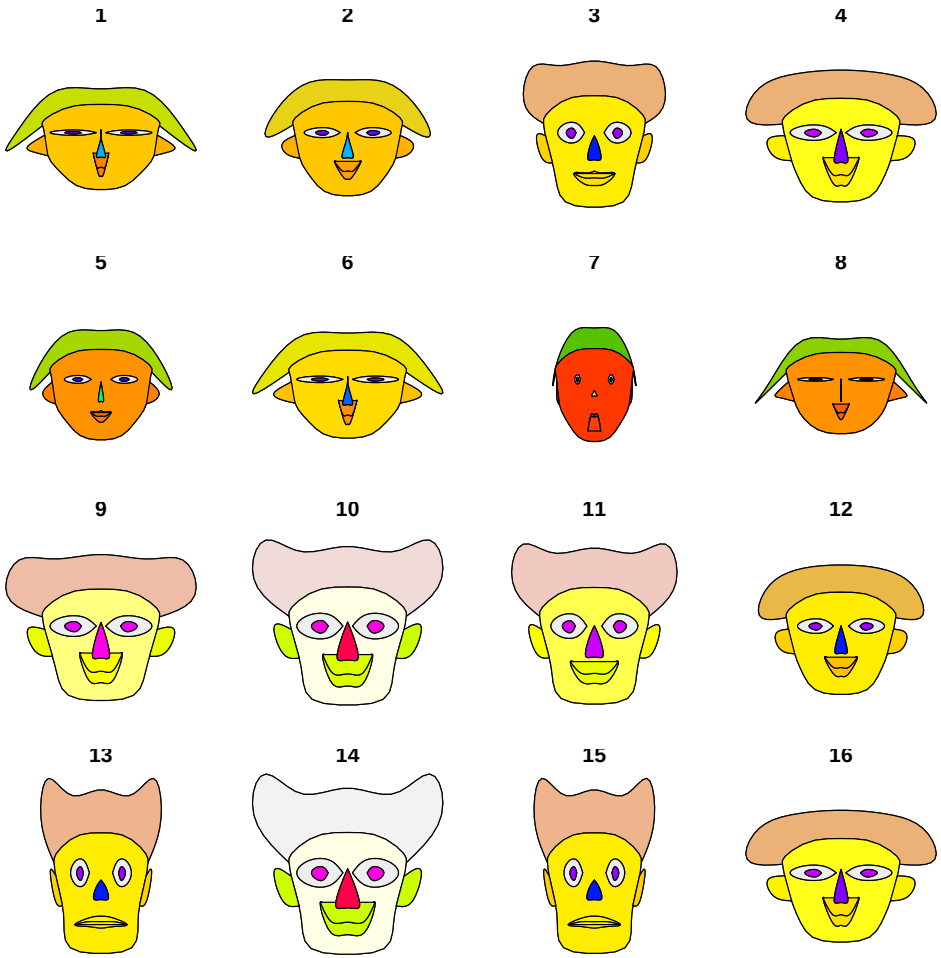


Multivariate data: graphical summary

Chernoff faces

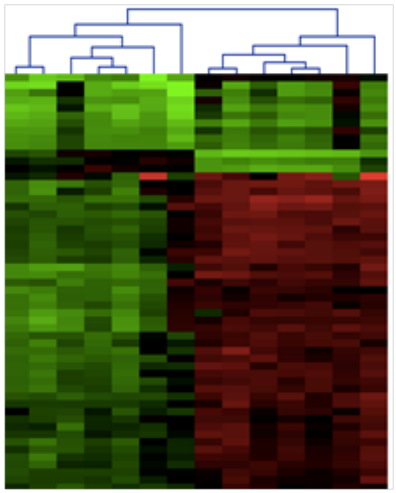
`wikipedia.org/wiki/Chernoff_face`

Idea: human face recognition.
Not to be taken too seriously. ;-)

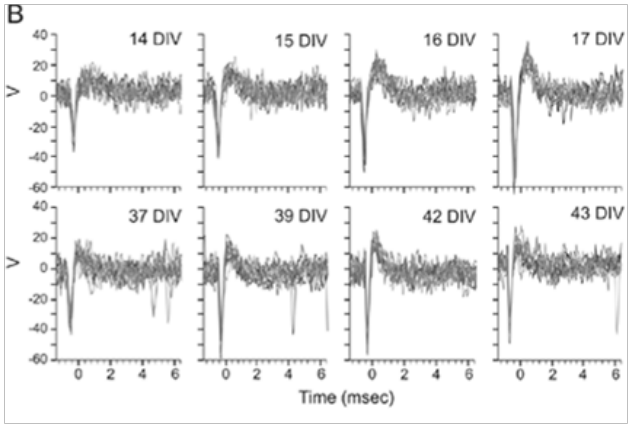


Multivariate neuroscience data (high dimensional)

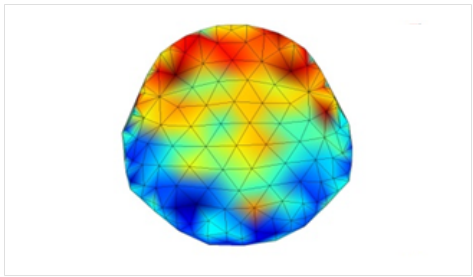
micro array data



action potentials data



MEG data



to finish

To wrap up

Today we discussed

- data types
- summary types
  - univariate data
  - bivariate data
  - multivariate data

Assignment 1: practice these summaries!
Make proper reports, i.e. proper language, neat pictures, etc.
Make nice numerical and graphical summaries,
and always describe them in words!

Friday: exercise classes via Zoom (with partner). Come prepared!
You are the one responsible for your progress, and not the TA!

Next week: exploring distributions