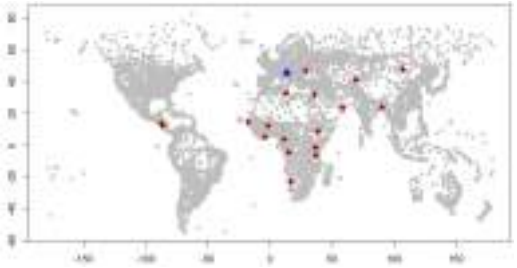
P5.2 Statistics for Medicine

Massimo Borelli

Master of Advanced Studies in Medical Physics

welcome, introduction and objectives



welcome, introduction and objectives

`github.com/MassimoBorelli/ictpmmp`

- free copy of the Lecture Notes
- all the slides
- homeworks (for the final exam)

welcome, introduction and objectives

brief Syllabus

- Descriptive Statistics
- Probability and Medicine
- Sampling and Inference
- the linear model
- the generalized linear model

welcome, introduction and objectives

goal to achieve at the end of the course

- to be able to summarize a biomedical dataset, by means of properly choosen statistical indicators
- to be able to provide basic statistical inference, choosing the proper statistical test or regression model
- to be able to properly interpret the frequentist and bayesian reporting

welcome, introduction and objectives

About the exam

- 'homework' assignments
- final exam
  - median vote of the homeworks

welcome, introduction and objectives

Safety, first

my advice: to study and to work together

P5.2 Statistics for Medicine

Massimo Borelli

Master of Advanced Studies in Medical Physics

shifting Statistics from Physics to Medicine /1 of 2

- $N \longrightarrow \infty$?
- $j \in \{1, ..., N\}$!

shifting Statistics from Physics to Medicine /2 of 2

Softwares used by Statisticians /1

standard console of R /2

working with scripts in R /3

best interface: R Studio /4

Helping beginners: R Commander /5

Integrating spreadsheet: Jamovi /5

a promising 'new entry': JASP /6

The iris dataset

- 150 rows

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
...	...	...	...	...
6.2	2.9	4.3	1.3	versicolor
5.1	2.5	3.0	1.1	versicolor
5.7	2.8	4.1	1.3	versicolor
...	...	...	...	...
6.5	3.0	5.2	2.0	virginica
6.2	3.4	5.4	2.3	virginica
5.9	3.0	5.1	1.8	virginica

The breastioert dataset

- 37 rows

Energy	Diameter	Angle	Difference	Area
9	6.5	0	-2	NA
9	7	15	-2	NA
9	5.5	0	-3	NA
9	5.5	0	-5	NA
6	5.5	15	-1	NA
..	..	..	..	..
6	5.5	15	-1	0
9	6.5	0	-5	11.4
9	5.5	15	-5	1.5
9	5.5	30	-1	4.7

The otitis dataset

- 1000 rows

episodes
0
0
0
...
1
1
...
6
6
6

The roma dataset

- 210 rows, 7 columns

logHE4	logCA125	...	AgePatient	Menopause	Histology
3.58	4.25	...	34	ante	benign
3.42	5.45	...	21	ante	benign
5.68	4.72	...	64	post	malignant
4.14	3.96	...	58	post	malignant
...	...	...	...	...	...
4.06	2.20	...	55	post	benign
3.96	4.03	...	63	post	malignant

The cholesterol dataset

- 1025 rows

idanag	sex	TOTchol	HDLchol
id537	m	243	64
id15956	m	168	49
...	...	...	...
id1060787	f	186	57
id1060796	m	146	48
...	...	...	...
id1060888	f	193	74
id1061003	m	151	60

The gossett dataset

nkd	kd	difference
1903	2009	106
1935	1915	-20
1910	2011	101
2496	2463	-33
2108	2180	72
1961	1925	-36
2060	2122	62
1444	1482	38
1612	1542	-70
1316	1443	127
1511	1535	24

The fresher dataset

- 65 rows

gnd	hght	wght	shsz	smoke	gym	fc	syst	diast
f	155	53	36	no	not	62	90	60
f	157	50	37	no	occasional	64	120	70
f	158	48	36	no	occasional	74	95	75
..	..	..	..	..	..	..	..	..
m	187	73	45	no	sporty	66	135	100
m	191	75	44	no	sporty	60	135	110
m	194	79	46	yes	sporty	66	120	65

data quality

- In hospitals, **spreadsheets** are routinary
- poor **data quality** is an issue
  - multicenter trials are routinary

0th homework (optional, not compulsory)

- Very often data not properly masked

protecting privacy in a spreadsheet

As an exercise, download on your computer the `privacy` dataset (at `https://github.com/MassimoBorelli/ictpmmp`), explore it with your favourite spreadsheet and create a new column of data by means of a text function (or joining together the outputs of different text functions) in order to provide a unique identifier for each row ('record') of the dataset.

P5.2 Statistics for Medicine

Massimo Borelli

Master of Advanced Studies in Medical Physics

Background

To describe the features of a quantitative dataset:

- the **location** of the data
- and their **variability**

Elise Whitley, Jonathan Ball.
Statistics review 1: Presenting and summarising data
`https://ccforum.biomedcentral.com/articles/10.1186/cc1455`

Alla Katsnelson.
Colour me better: fixing figures for colour blindness
`https://www.nature.com/articles/d41586-021-02696-z`

historical example: the `iris` dataset

- *setosa*
- *versicolor*
- *virginica*

- petal length, petal width
- sepal length, sepal width

preview

`iris` is already stored in JASP

menu Descriptives

lab guided activity /1

Example (position and dispersion measures)

Are we able to understand?

- **measures of central tendency / location**
- **measures of shapes / dispersions**
- the concepts of quantiles
- a **balanced** dataset
- a **complete** dataset

Jonathan Blitzstein, Jessica Hwang.
Introduction to Probability.
`https://projects.iq.harvard.edu/stat110/home`

lab guided activity /2

'A picture is worth a thousand words'

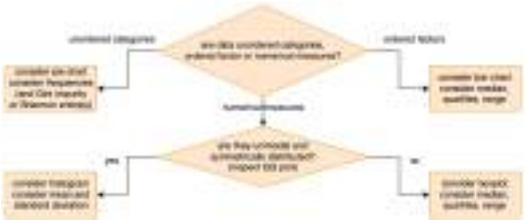
Example (graphs)

- dot plots
- distribution plots
- boxplots (quantiles and outliers?)
- scatter plots

1st Homework Activity (final exam)

Mario de Denaro and Mara Severgnini (Radiation Oncology)

suggested algorithm



parametric vs. non-parametric

Table 1. Characteristics of Infants with Very Low Birth Weight and Those Born at Term,*

Characteristic	Study Participants	Study Nonpart
Very low birth weight		
No. of subjects	166	89
Gestational age — wk	29.17±2.22	28.17±2.1
Birth weight — g	1120±221	1130±28

	n	Mean (SD) g/ week	Median g/ week

funny ;-)

Results

Reliability of automated analysis

For the analysis of the datasets, the two expert analysts manually detected, on average, 4,562 (range 4,439 to 4,686) events (EAds or P_s events). ICCs for the Neurolyte

retrieving descriptive information / 1 of 2

Can we guess the mean from the median?

$$\mu \approx \frac{a+2m+b}{4} + \frac{a-2m+b}{4n} \approx \frac{a+2m+b}{4}$$

$$\mu \approx \frac{a+2Q_1+2m+2Q_3+b}{8}$$

$$\mu \approx \frac{Q_1+m+Q_3}{3}$$

retrieving descriptive information / 2 of 2

Can we guess the standard deviation from the quartiles?

$$\sigma \approx \frac{b-a}{\xi(n)}$$

$$\sigma \approx \frac{Q_3-Q_1}{\eta(n)}$$

$$\sigma \approx \frac{1}{2}\left(\frac{b-a}{\xi(n)} + \frac{Q_3-Q_1}{\eta(n)}\right)$$