

# Statistics in the AI era: Obsolete discipline or essential tool?

Aldo Gardini

University of Bologna, Department of Statistical Sciences “P. Fortunati”

DIFA Summer School on Physical Sensing and Processing – VI edition  
Bologna, July 9, 2024

1 Introduction

2 Conundrum 1

3 Conundrum 2

4 Conundrum 3

5 Conundrum 4

6 Wrapping up

# Why statistics in a summer school of Physics?

- **Data** is a crucial resource in the modern era.
- **Analyzing data** is a highly sought-after task, involving professionals from various backgrounds such as engineers, computer scientists, mathematicians, physicists, and *sometimes* statisticians.
- The term **data science** often refers to a set of tools used for empirical analysis.
- While significant attention is given to methods, the overall workflow and some key aspects of an analysis are frequently guided by **common sense** and **intuition**.

# What does a statistician deal with?

# What does a statistician deal with?

- Development of new methods relying on mathematical and probabilistic tools.
- Addressing computational challenges through numerical and simulation methods.
- Implementing methodologies in computer programs.

# What does a statistician deal with?

- Development of new methods relying on mathematical and probabilistic tools.
- Addressing computational challenges through numerical and simulation methods.
- Implementing methodologies in computer programs.
- **Designing experiments for data collection.**
- **Selecting and developing methods suited to the available data and the research question.**

# Starting point

*“To consult the statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of.”*



**Ronald Fisher (1890-1962).**  
Creator of modern statistics.

## Getting things in order

*The outcomes of a statistical analysis can include:*

- Prediction or supervised classification.
- Dimensionality reduction.
- **Exploring relationships among variables.**



## Getting things in order

*The outcomes of a statistical analysis can include:*

- Prediction or supervised classification.
- Dimensionality reduction.
- **Exploring relationships among variables.**

*A proper statistical analysis requires:*

- Data.
- Understanding of the framework that generated the data.
- Knowledge of how to obtain meaningful results.
- Awareness of the limitations associated with the available data.

# Goal of today

- Demonstrate how the mantra “*Let the data speak for themselves*” can fail dramatically in certain settings.
- Raise awareness that **analyzing data** involves more than just running code and interpreting results; it requires a comprehensive understanding of the relevant frameworks.
- To this end, we will explore four situations where intuition alone is insufficient for a proper analysis, leading to possible misleading conclusions.

① Introduction

② Conundrum 1

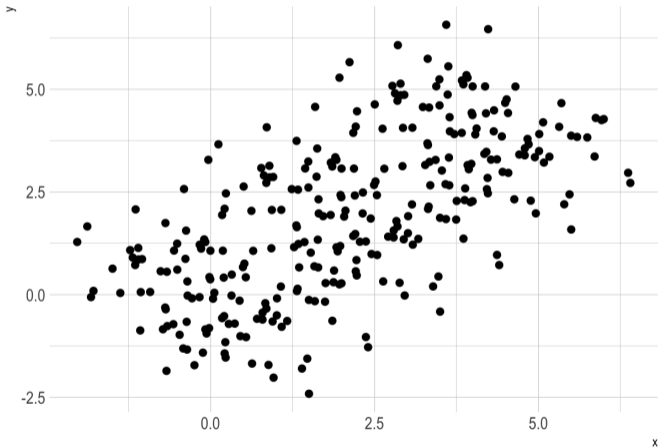
③ Conundrum 2

④ Conundrum 3

⑤ Conundrum 4

⑥ Wrapping up

# What can you say about this plot?



What can you say about this table?<sup>1</sup>

	Vaccinated (V)	Unvaccinated ( $\bar{V}$ )	Total
Death (D)	838	839	1,677
Survived ( $\bar{D}$ )	90,528	139,838	230,366
Total	91,366	140,677	232,043

# What can you say about this table?<sup>1</sup>

	Vaccinated (V)	Unvaccinated ( $\bar{V}$ )	Total
Death (D)	838	839	1,677
Survived ( $\bar{D}$ )	90,528	139,838	230,366
Total	91,366	140,677	232,043

- $\mathbb{P}[D|V] = 9.1719 \times 10^{-3}$
- $\mathbb{P}[D|\bar{V}] = 5.9640 \times 10^{-3}$
- $Odds_V = 9.2568 \times 10^{-3}$
- $Odds_{\bar{V}} = 5.9998 \times 10^{-3}$

**Measure of association:** Odds Ratio

$$OR = \frac{Odds_V}{Odds_{\bar{V}}} = 1.5489$$

<sup>1</sup>Data related to Covid-19 cases in December 2021 from Istituto Superiore di Sanità. Discussed in Crupi et al. (2022)

# Conditional analysis

		$V$	$\bar{V}$
$D$	Age < 80	207	440
	Age $\geq$ 80	631	399
$\bar{D}$	Age < 80	81,396	136,684
	Age $\geq$ 80	9,132	3,154

## Considering people < 80

- $Odds_{V|<80} = 2.5431 \times 10^{-3}$
- $Odds_{\bar{V}|<80} = 3.2191 \times 10^{-3}$

Association:  $OR_{<80} = 0.7900$

## Considering people $\geq$ 80

- $Odds_{V|\geq 80} = 69.0977 \times 10^{-3}$
- $Odds_{\bar{V}|\geq 80} = 126.5060 \times 10^{-3}$

Association:  $OR_{\geq 80} = 0.5462$

## Which is the the correct analysis?

- From a *formal* statistical perspective, both marginal and conditional analyses are correct.
- Incorporating knowledge about the phenomenon is necessary to achieve a causal interpretation of the problem.

This situation exemplifies **Simpson's paradox** (Simpson, 1951), where the direction of an association changes when moving from marginal to conditional analyses.



## Which is the the correct analysis?

- From a *formal* statistical perspective, both marginal and conditional analyses are correct.
- Incorporating knowledge about the phenomenon is necessary to achieve a causal interpretation of the problem.

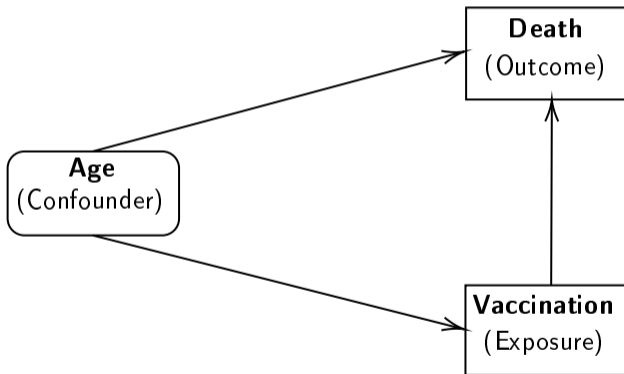
This situation exemplifies **Simpson's paradox** (Simpson, 1951), where the direction of an association changes when moving from marginal to conditional analyses.

**More formally:**

*"Good for men, good for women yet bad for people"*

# Reversal association under confounding

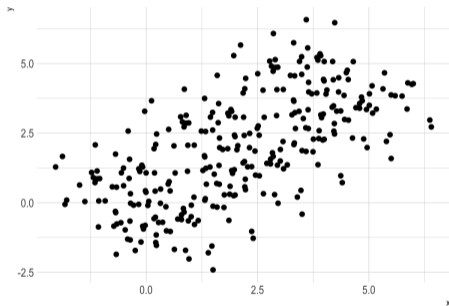
It is a paradox **only** if the causal structure of the problem is ignored:



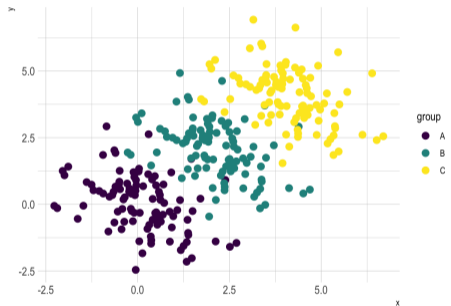
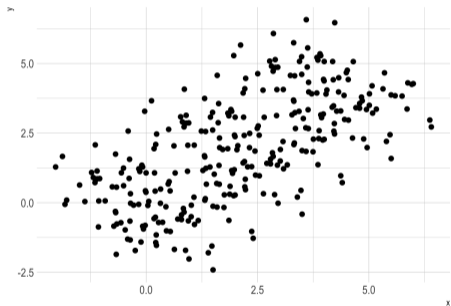
## How to deal with confounding?

- **Expensive approach:** Randomization. Although it requires significant time and resources, designing a randomized study effectively controls for confounding variables.
- **Alternative approach:** When only observational data are available, controlling for confounding necessitates specialized statistical methods that depend on a deep understanding of the causal diagram. Various methods are available to adjust for confounding in such cases.

# The issue does not concern only tables...



# The issue does not concern only tables...



① Introduction

② Conundrum 1

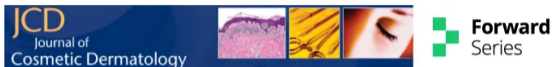
③ Conundrum 2

④ Conundrum 3

⑤ Conundrum 4

⑥ Wrapping up

## In Covid times...

LETTER TO THE EDITOR | [Full Access](#)

## A preliminary observation: Male pattern hair loss among hospitalized COVID-19 patients in Spain – A potential clue to the role of androgens in COVID-19 severity

Andy Goren MD, Sergio Vaño-Galván MD, Carlos Gustavo Wambier MD, PhD [✉](#) John McCoy PhD, Alba Gomez-Zubiaur MD, Oscar M. Moreno-Arrones MD, Jerry Shapiro MD ... [See all authors](#) ▾

First published: 16 April 2020 | <https://doi.org/10.1111/jocd.13443> | Citations: 134

A-Link Universita di Bologna

☰ SECTIONS

PDF TOOLS SHARE

### Abstract

A preliminary observation of high frequency of male pattern hair loss among admitted COVID-19 patients and suggest that androgen expression might be a clue to COVID-19 severity.


... Scopus like <https://www.tylervigen.com/spurious-correlations>

Apr 21, 2020  CC BY

<https://doi.org/10.32388/WPP19W.3>

## Low incidence of daily active tobacco smoking in patients with symptomatic COVID-19 Preprint v3

Makoto Miyara<sup>1</sup> , Florence Tubach<sup>1</sup> , Valérie Pourcher<sup>1</sup> , Capucine Morelot-Panzini<sup>1</sup> , Julie Pernet<sup>1</sup>, Julien Haroche<sup>1</sup>, Said Lebbah<sup>1</sup>, Elise Morawiec , Guy Gorochoy<sup>2</sup> , Eric Caumes<sup>1</sup>, Pierre Hausfater<sup>1</sup> , Alain Combes<sup>1</sup> , Thomas Similowski , Zahir Amoura<sup>1</sup>

Affiliations  [Highly-cited researchers](#)

**Participants:** We estimated the rates of daily current smokers in COVID-19-infected patients in a large French university hospital between February 28th, 2020 and March 30th, 2020 for outpatients and from March 23rd, till April 9th, 2020 for inpatients.

**Design:** The rates from both groups were compared to those of daily current smokers in the 2018 French general population, established in 2018, after standardization of the data for sex and age.

**Conclusions and relevance:** Our cross sectional study in both COVID-19 out- and inpatients strongly suggests that daily smokers have a very much lower probability of developing symptomatic or severe SARS-CoV-2 infection as compared to the general population.



## Main problem: collider bias

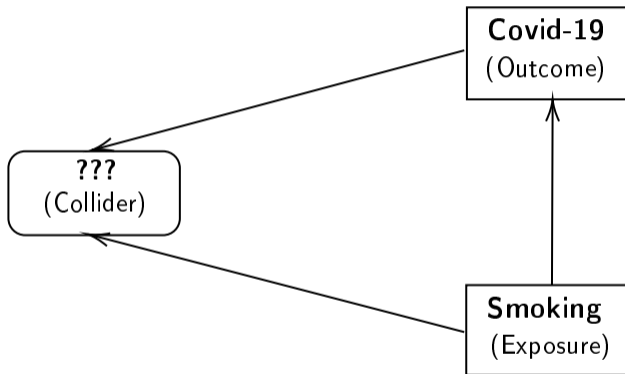
Almost all studies on Covid-19 rely on **non-random samples**. This is often overlooked when selecting methods and interpreting results.

The problem of collider bias is closely related to **sample collection methods** and is far more subtle than confounding!

For example, selection bias frequently occurs in:

- Observational studies,
- Voluntary sampling.

# What does a collider is?



Collider bias occurs when the collider variable controls the sampling process.  
In our case, being in the sample means **being tested!**

# Why collider is '???' ?

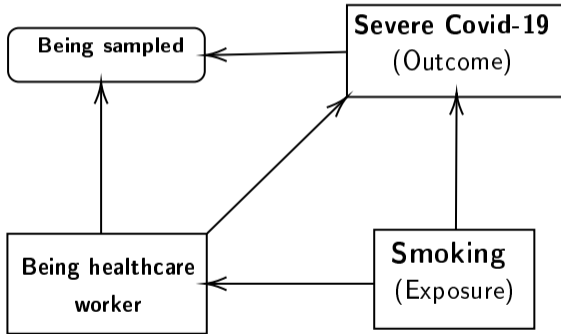
## Situation 1: late pandemic phase (Tattan-Birch et al., 2021)

??? = Cough!

Cough represented a symptom that increased the probability of being tested. Smokers are more subjected to cough, so they are **over-represented** in the sample, with a large number of negative tests.

# Why collider is '???'

## Situation 2: early pandemic phase (Fenton, 2020)



- Swab tests performed on people with severe illness or healthcare workers.
- Healthcare workers subjected to higher viral load. High probability of severe illness.
- Healthcare workers tend to smoke less than the remaining population.

## How to overcome collider bias?

- Utilize conditional (network) models that are based on causal schemes developed by researchers.
- Examine the composition of the available sample (e.g., demographic or socio-economic status). If systematic discrepancies are observed with respect to the target population, apply post-stratification procedures.
- Again, randomization!

1 Introduction

2 Conundrum 1

3 Conundrum 2

4 Conundrum 3

5 Conundrum 4

6 Wrapping up

## A simple research question...

### Problem from Lord (1967)

*A university is interested in investigating the effects on the students of the diet provided in the university dining halls and any sex difference in these effects.*

For each student we know:

- Weight at the time of arrival ( $W_I$ )
  - Weight at June ( $W_F$ )
  - Sex ( $S$ )
- } Gain:  $Y = W_F - W_I$

How would you exploit the data to answer the question?

# Solution 1

You consider the gain  $Y$  and you simply test for differences in mean in the two sub-populations individuated by sex  $S$ .

**Basic statistical tool:** t-test (equivalent to ANOVA when 2 groups are present).

**Results:**

	mean( $Y$ )	SD( $Y$ )
Male	0.043	0.174
Female	0.039	0.171

p-value: 0.869

**No significant differences in the gain comparing the two groups.**



## Solution 2

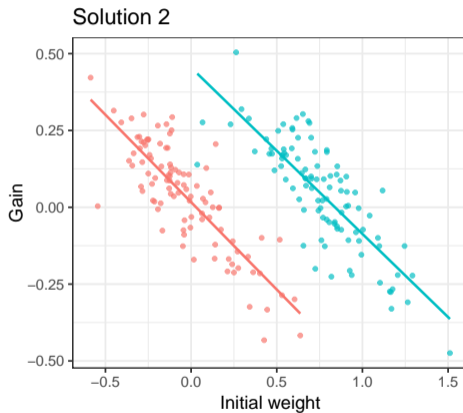
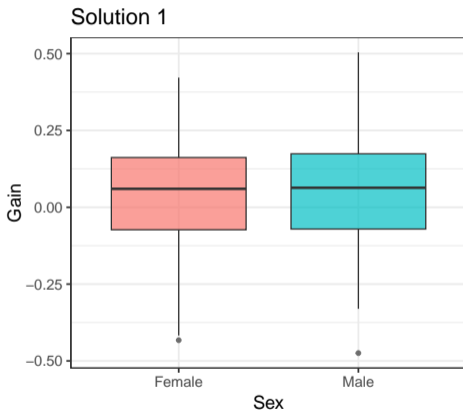
You compare the gain in the two sub-populations individuated by  $S$ , **considering the initial weight  $W_I$** .

**Still a basic statistical tool:** ANCOVA, i.e. analysis of covariance.

### Result:

In average, a male have a gain higher of 0.447 than a female, keeping constant the initial weight (significant difference with  $p\text{-value} < 10^{-16}$ ).

# What it is happening?

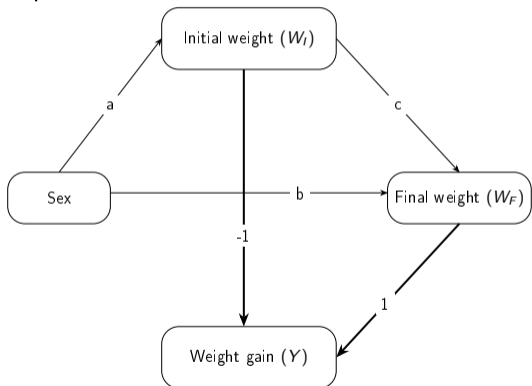


## Which is the correct solution?

As pointed out by Pearl (2016), again the solution is in the causal structure of the phenomenon!

## Which is the correct solution?

As pointed out by Pearl (2016), again the solution is in the causal structure of the phenomenon!



**Solution 1.** It evaluates the **total effect** of  $S$  on  $Y$ :

$$TE = b + ac - a.$$

**Solution 2.** It evaluates the **direct effect** of  $S$  on  $Y$ , considering  $W_I$  as mediator effect:

$$DE = b.$$

① Introduction

② Conundrum 1

③ Conundrum 2

④ Conundrum 3

**⑤ Conundrum 4**

⑥ Wrapping up

## Empirical experiment

- An experiment produced  $n = 11$  observations of two numerical variables  $X$  and  $Y$ , with the following descriptive statistics.

	Mean	S.D.
$X$	9.00	3.32
$Y$	7.50	2.03

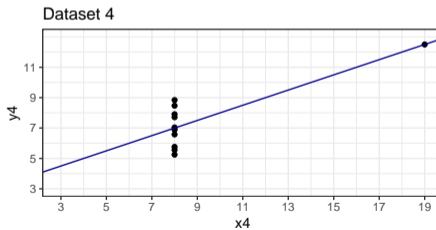
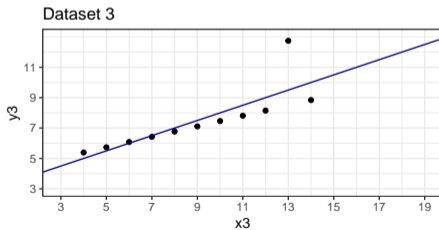
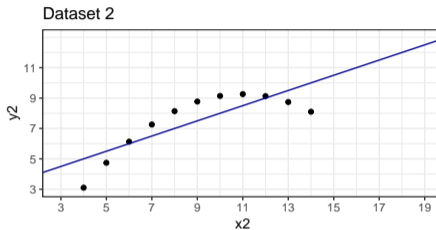
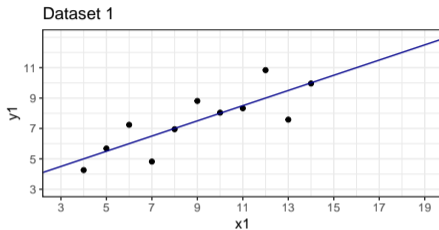
- A simple linear model is fitted:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i; \quad \varepsilon_i \stackrel{ind}{\sim} \mathcal{N}(0, \sigma^2), \quad \forall i.$$

- The following outcomes are observed

$$\hat{\beta}_0 = 3.00, \quad \hat{\beta}_1 = 0.50, \quad R^2 = 0.67.$$

# The quartet by Anscombe (1973)



## How to deal with this?

When applying a statistical model (or test) to data, certain assumptions are implicitly made and must be **carefully assessed**.

In our example, we consider a linear model with two variables, where simple scatter plots can provide valuable insights.



## What to do when many covariates are in the model?

Graphical inspection of residuals can be highly informative! Being defined as

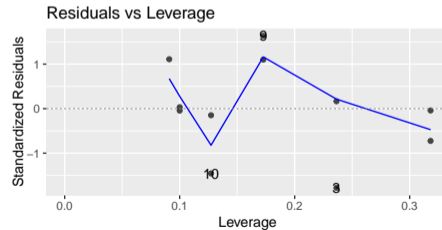
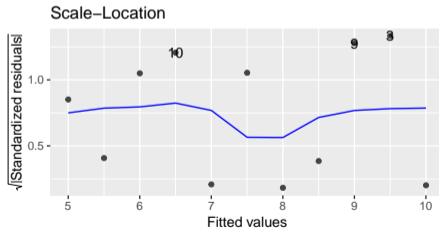
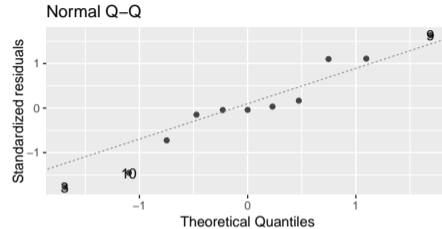
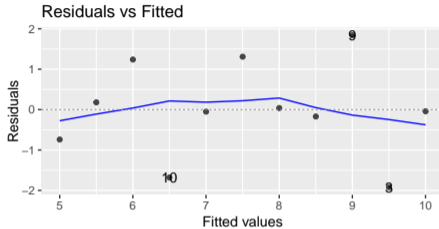
$$\hat{\epsilon}_i = y - \beta_0 - \beta_1 x_i - (\beta_2 x_{2i} + \dots),$$

they are valid also when multiple regressors are included.

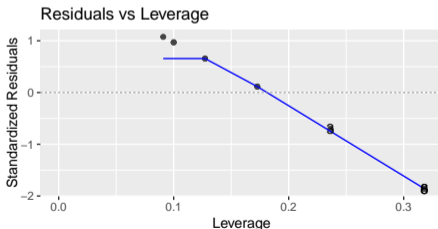
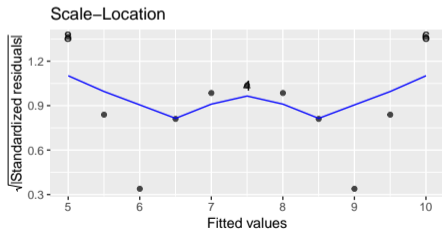
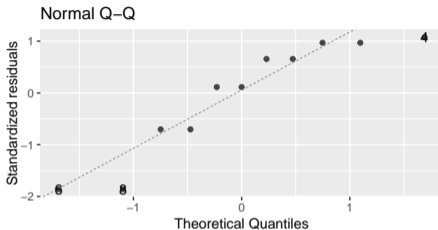
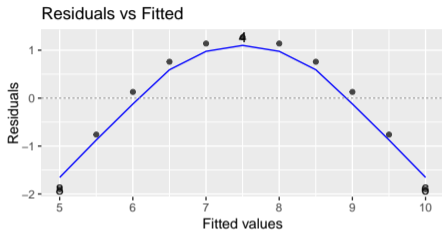
Through 4 basic plots, we can verify:

- Homoscedasticity
- Presence of non-linear relationships
- Presence of outliers
- Departures from Normality assumption

# Residuals of dataset 1

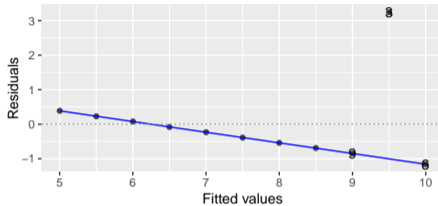


# Residuals of dataset 2

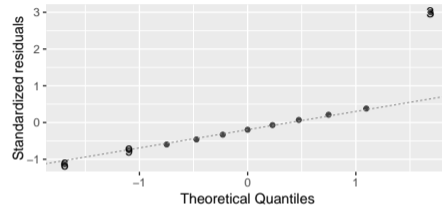


## Residuals of dataset 3

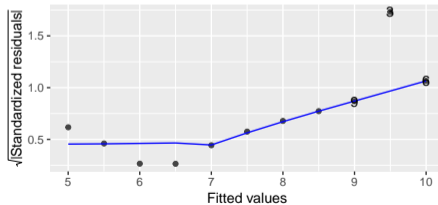
Residuals vs Fitted



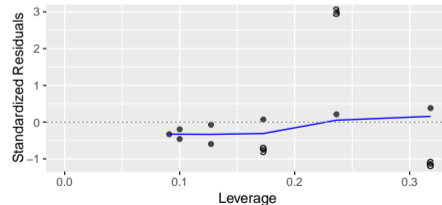
Normal Q-Q



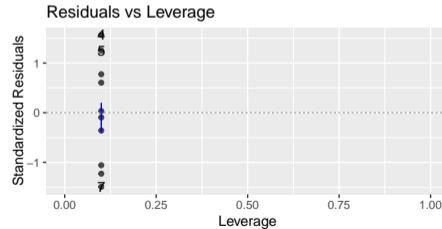
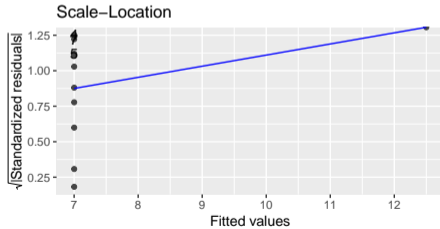
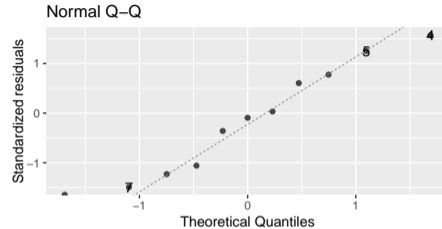
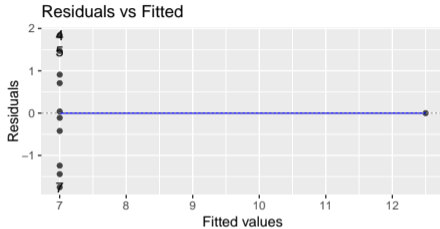
Scale-Location



Residuals vs Leverage



# Residuals of dataset 4



① Introduction

② Conundrum 1

③ Conundrum 2

④ Conundrum 3

⑤ Conundrum 4

⑥ Wrapping up

## So, is statistics still useful?

- Statistics offers a set of **tools** for analysing data, nowadays largely used (and also developed) by non-statisticians.
- The diffusion of such methods (and their user-friendliness) is not supported by the diffusion of statistical culture. **The discipline is not appealing!**
- Anyone can use statistics, without being a statistician.

- Anscombe, F. J. (1973). Graphs in statistical analysis. *The american statistician*, 27(1):17–21.
- Crupi, V., Calderisi, M., Pighin, S., and Tentori, K. (2022). Probabilità pandemiche: tre pezzi non troppo facili. *Sistemi intelligenti*, 34(2):329–341.
- Fenton, N. (2020). A note on 'collider bias undermines our understanding of covid-19 disease risk and severity' and how causal bayesian networks both expose and resolve the problem. *arXiv preprint arXiv:2005.08608*.
- Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological bulletin*, 68(5):304.
- Pearl, J. (2016). Lord's paradox revisited—(oh lord! kumbaya!). *Journal of Causal Inference*, 4(2):20160021.
- Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241.
- Tattan-Birch, H., Marsden, J., West, R., and Gage, S. H. (2021). Assessing and addressing collider bias in addiction research: the curious case of smoking and covid-19. *Addiction (Abingdon, England)*, 116(5):982.