Longitudinal data analysis in Python:
examples and common biases

Filippo Biscarini

(Biostatistician, bioinformatician and quantitative geneticist) **CNR**, Milan (Italy)

It's been a long way to modern statistics

- 1661: John Graunt: first statistician ("Natural and political observations made upon the bills of mortality")
- 1870's: Francis Galton: linear regression
- ~1900: Karl Pearson: correlation
- 1925: Ronald Fisher's "*Statistical Methods for Research Workers*" (he later regretted the 0.05 p-value threshold) → **frequentist statistics**
- **Bayesian** resurgence: 1980s → **MCMC** (1986: Gibbs sampling by Geman & Geman)
- Non-parametric statistics & resampling methods
- The **statistical** (machine) **learning** paradigm

A lot of math!

Increasing computer power

Big data

Types of longitudinal data

1. treatments and timepoints
2. repeated records
3. censored data
4. time series data

Types of longitudinal data

1. treatments and timepoints
2. repeated records
3. censored data
4. time series data

Q: can you think of other types of longitudinal data?

A few examples from literature

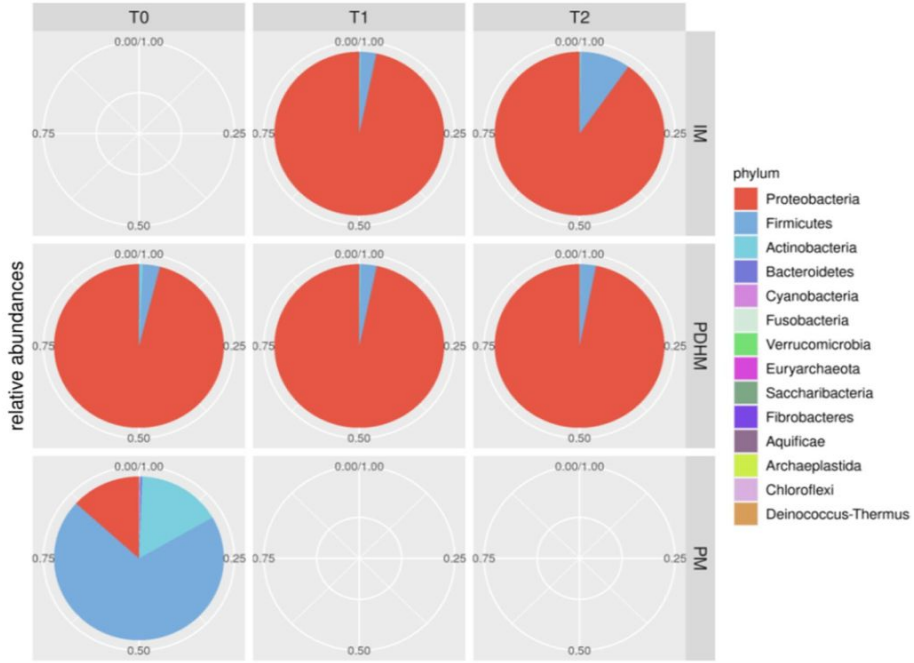


Fig. 3 Pie-charts of phylum relative abundances in the PDHM, PM and IM samples. Pie charts showing the distribution of the dominant bacterial phyla in the PDHM, PM and IM samples. The numbers around the pie-charts indicate the percentage of abundance. *PM* preterm milk samples, *PDHM* pasteurized donor human milk samples, *IM* inoculated milk samples. T0: baseline (before inoculum); T1: 2 h after inoculation; T2: 4 h after inoculation

From Mallardi et al. 2021

A few examples from literature

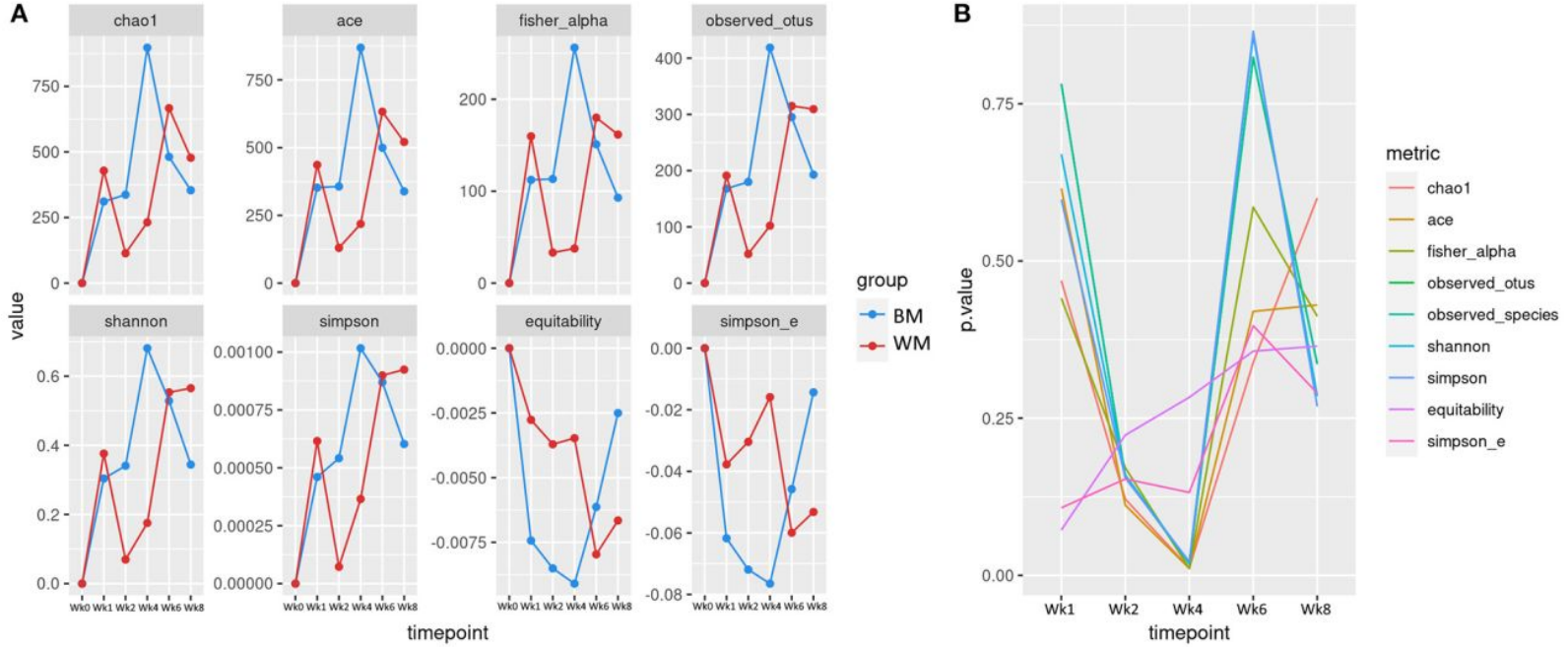


FIGURE 2 | (A) Alpha diversity indices at the different trial time points for the two groups: bulk tank milk (BM) and waste milk (WM). Index values are indicated as differences from baseline (Wk0 = 0). **(B)** Statistical significance of alpha diversity indices at the various time points.

From Penati et al. 2021

A few examples from literature

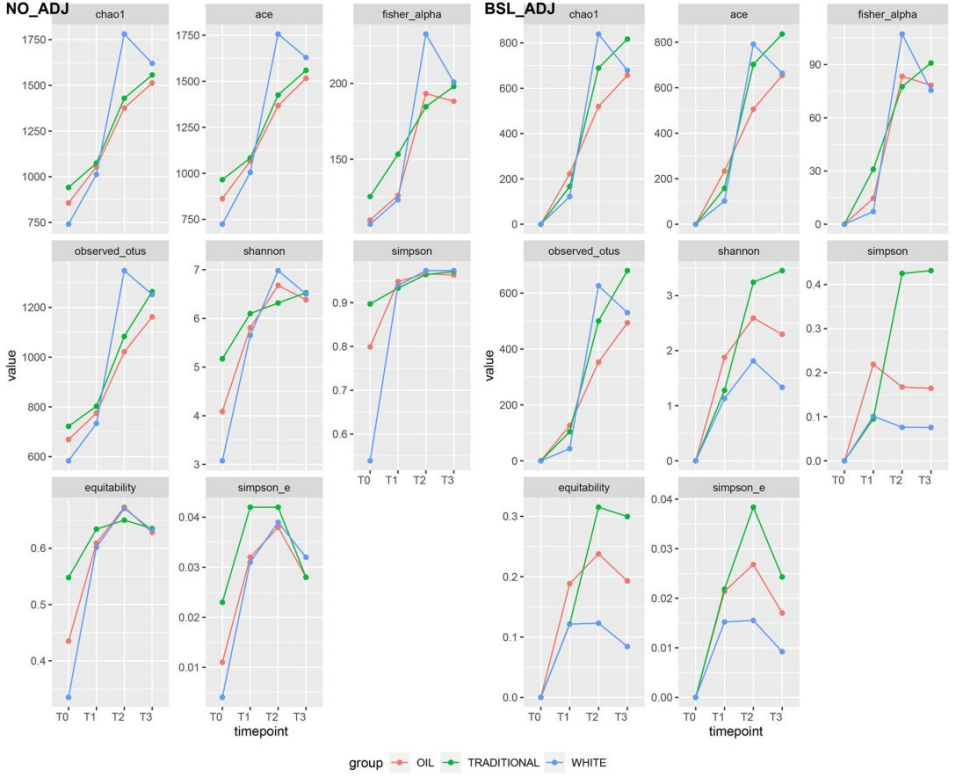


Fig 5. Alpha diversity. Average alpha diversity indices per group over timepoints. Non-adjusted (left) and baseline-adjusted (right) values.

From Cremonesi et al. 2022

A few examples from literature

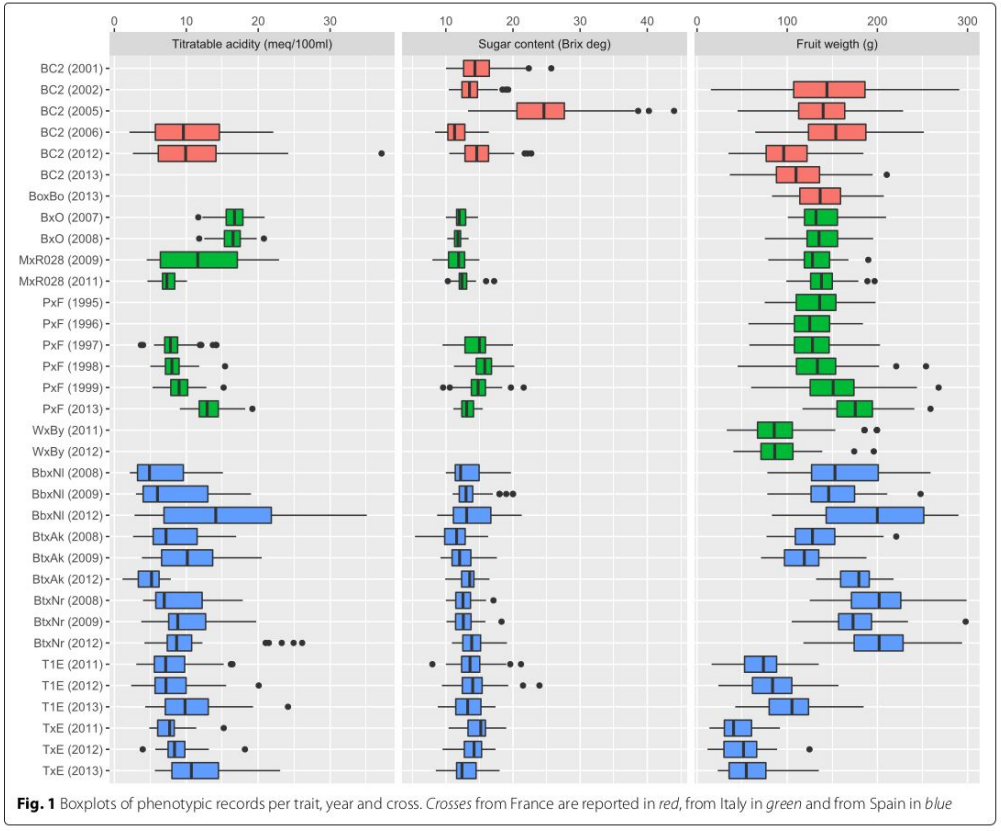


Fig. 1 Boxplots of phenotypic records per trait, year and cross. *Crosses* from France are reported in *red*, from Italy in *green* and from Spain in *blue*

From Biscarini et al. 2017

A few examples from literature

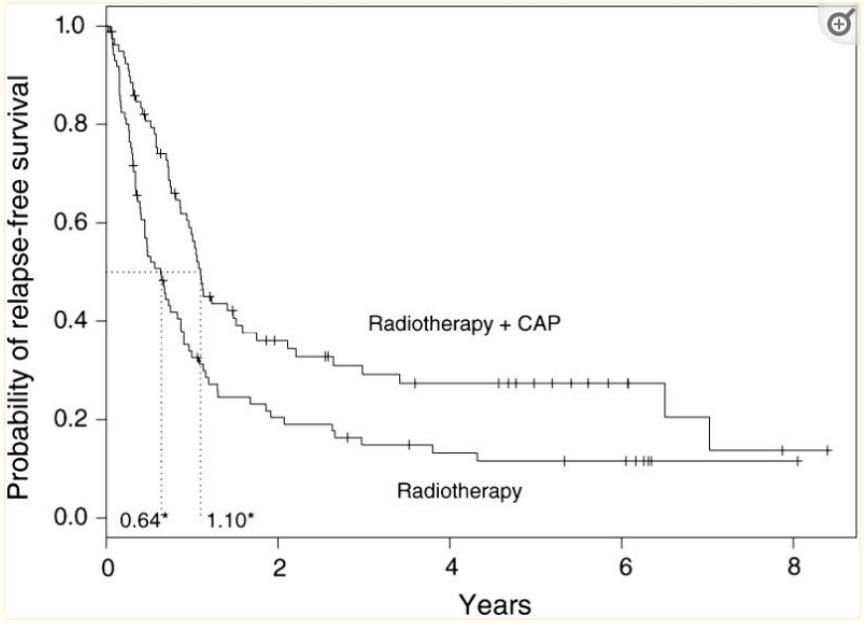


Figure 2

Relapse-free survival curves for the lung cancer trial. * Median relapse-free survival time for each arm, + censoring times, CAP=cytoxan, doxorubicin and platinum-based chemotherapy.

Lung cancer, radiotherapy vs radiotherapy + chemotherapy

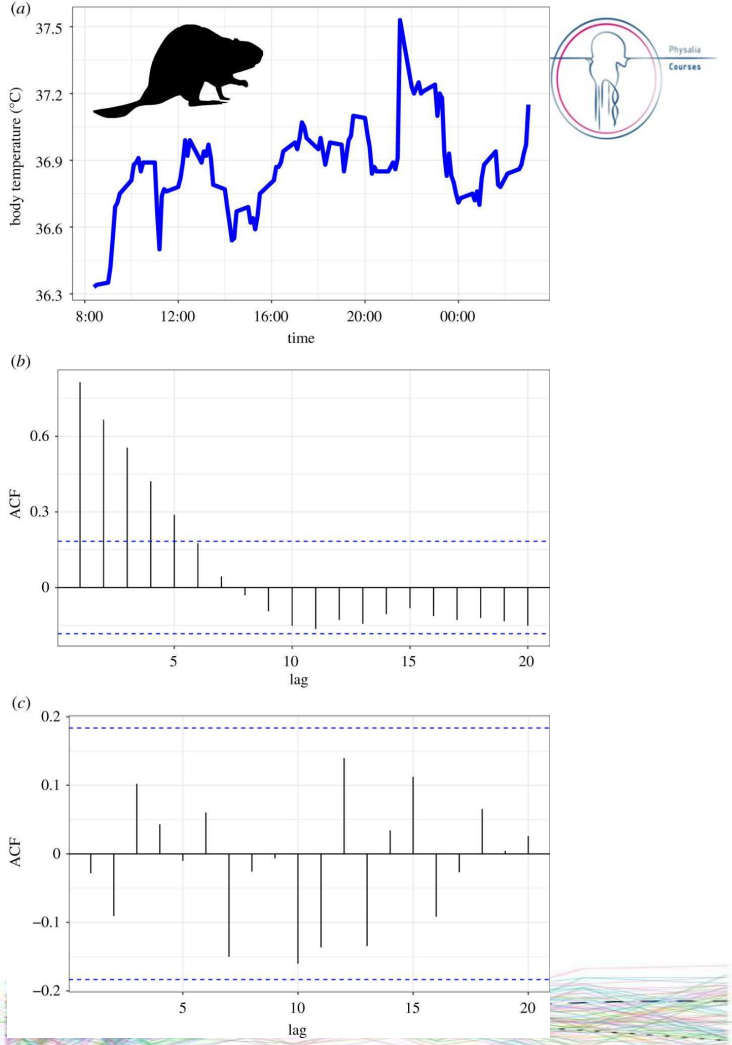
Q: does anybody know how to read this graph?

Clark et al 2003; https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2394262/

A few examples from literature

Time-series data on beaver body temperature:

- (*a*) Beaver body temperature data from a single individual recorded every 10 min over approximately 24 h
- (*b*) autocorrelation plot of the residuals; helps identify that there is still unmodelled autocorrelation in these data
- (*c*) Applying a more complex temporal autocorrelation model resolves these issues and produces a satisfactory autocorrelation plot



Harrison 2021; https://royalsocietypublishing.org/doi/10.1098/rstb.2020.0227

Beware of the bias!

(common biases to look out for)

Time is pervasive

Confounding

- What it is: a third (omitted) variable that has effect both on the independent (X) and dependent (Y) variables, and can mask a potential relationship (correlation/association)
- Examples: smoking (X), lung cancer (Y), age (confounder)
- How to deal with it: i) at experimental design (e.g. randomization); ii) at statistical analysis (e.g. stratification)

Collider?

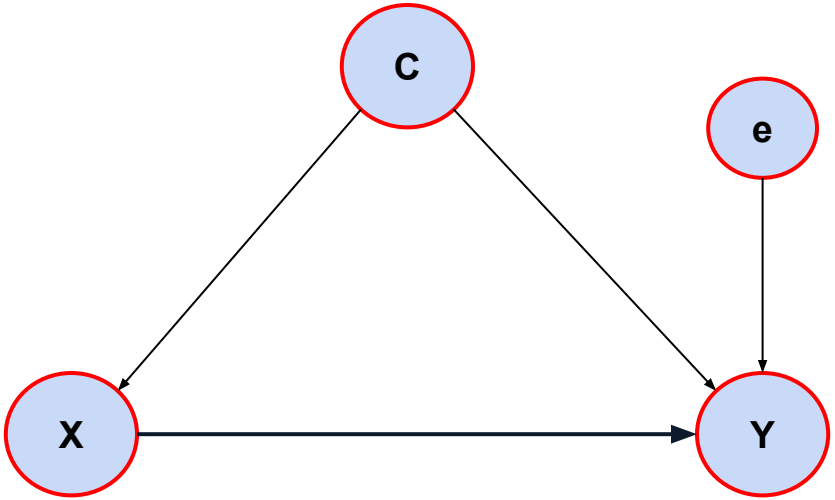
- What it is: a third variable which is influenced by both X and Y, and can generate a spurious relationship between them
- Examples: IQ (Y), study load (X), college acceptance (collider) →considering college acceptance in the analysis may result in a spurious correlation
- How to deal with it: i) experimental design (e.g. representative sample); ii) analysis (e.g. don't include colliders in the model)

Mediator?

- What it is: a proxy variable which is between X and Y
- Examples: fire → smoke → alarm
- How to deal with it: i) experiment!; ii) careful interpretation of results

More later on with notebook and code!

Confounding



DAG

- X, Y: variables under investigation
- C: confounder
- e: error / residual

Is X associated with Y?
Or does X cause Y? (but we won't go down the causality path)

Confounding

Experiment: we want to test whether a fertilizer has a positive effect on the growth of maize

- Two fields, one with the fertilizer, one without (control)
- After some time we measure growth

Q: do you notice a problem with this experimental setting?

[Generated with AI · August 25, 2024 at 1:23 PM (Copilot)]

Confounding

Experiment: we want to test whether a fertilizer has a positive effect on the growth of maize

- Two fields, one with the fertilizer, one without (control)
- After some time we measure growth

Treatment and field (physical location) are confounded:

- One field receives more sun
- One field is in the shadow side of the hill

The effect of the fertilizer will be masked by that of the field (e.g. exposure to sunlight)

[Generated with AI · August 25, 2024 at 1:23 PM (Copilot)]

Confounding

Experiment: we want to test whether a fertilizer has a positive effect on the growth of maize

- Two fields, one with the fertilizer, one without (control)
- After some time we measure growth

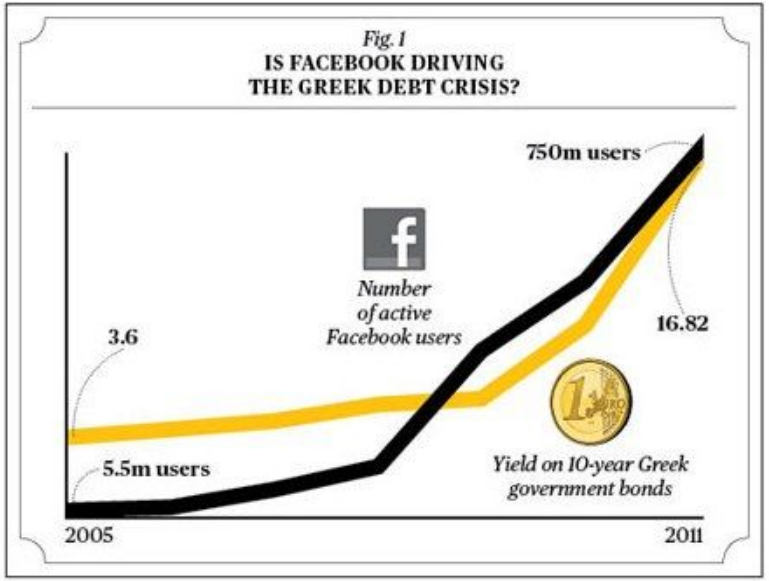
[Generated with AI · August 25, 2024 at 1:23 PM (Copilot)]

What about **time**?

Let's say that we measure growth in the control field on day 15; then we rest, and measure growth in the treated field on the next day → confounding between time and treatment (even if the two fields had the same exposure to sunlight)

Confounding

- Positive correlation (strong) between the number of Facebook users and the amount of Greek government debt
- Behind the scenes, there's a common cause for both variables: **time!** →confounder
- [spurious correlations archive]



From: https://b2bstorytelling.wordpress.com/2014/09/18/the-incredible-lightness-of-numbers/

Are you “confounded”?

A little thought exercise

Take 5 minutes, and think of possible confounding effects in your research field/data

How to deal with confounding

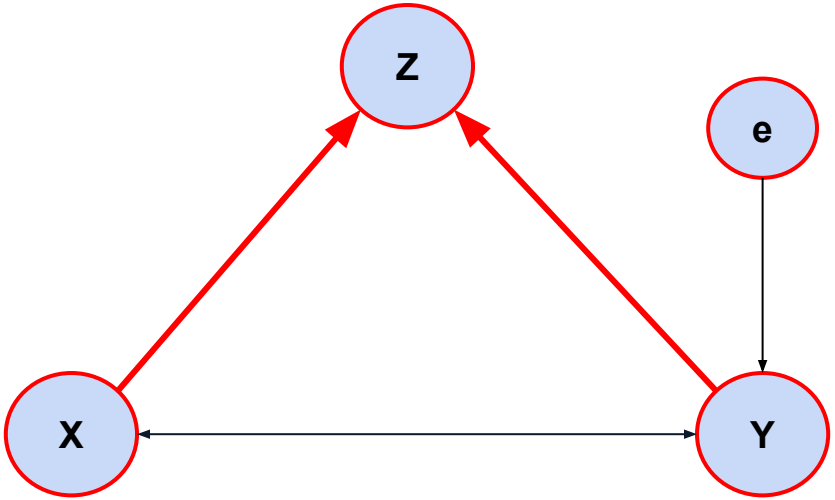
Confounding can be controlled with the experimental design through:

1. **randomization**: e.g. RCT
2. **restriction**: i.e. use only one value of the potential confounder to recruit subjects (e.g. only females)
3. **matching**: i.e. match (balance) exposed / non-exposed by values of the confounder

Confounding can be managed during data analysis by:

1. **standardization**: e.g. normalize by age, body weight etc. (useful with continuous confounders)
2. **stratification**: e.g. stratified analysis (e.g. shark attacks and ice cream sales appear to be correlated: if you condition on season/temperature, the correlation disappears)
3. **multivariable regression (linear, logistic, cox, etc.) models**: include confounders in the statistical model

Collider bias: the opposite of confounding!

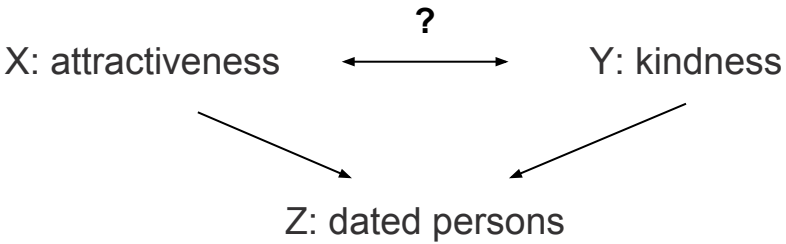


DAG

- X, Y: variables under investigation
- Z: collider
- e: error / residual

Is X associated with Y?
Or does X cause Y?

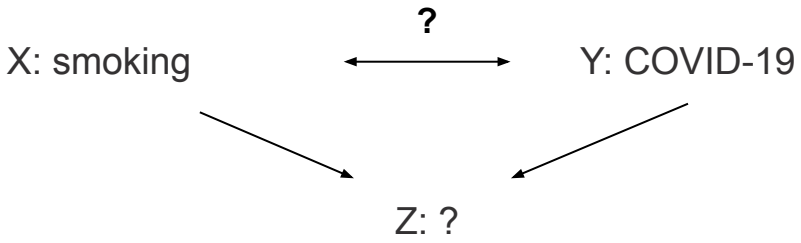
Collider bias



- you will tend to date: i) both pretty and kind (few); ii) pretty but rude (more); iii) kind but plain (more)
- you will NOT date: rude and plain

This bias will produce an artificial negative correlation between attractiveness and kindness, that is not present in reality

Collider bias



We conduct an experiment within a large hospital, recruiting healthcare workers as subjects for this observational study (doctors, nurses, lab technicians etc.)

Your task: work out what Z is in this case and how the collider bias can produce a spurious associations

Collider bias

Field data: milk samples, microbiota

- We assume there are only two types of microbes, *Lactobacillus* and *Acinetobacter*
- Sequencing →normalised counts / ml

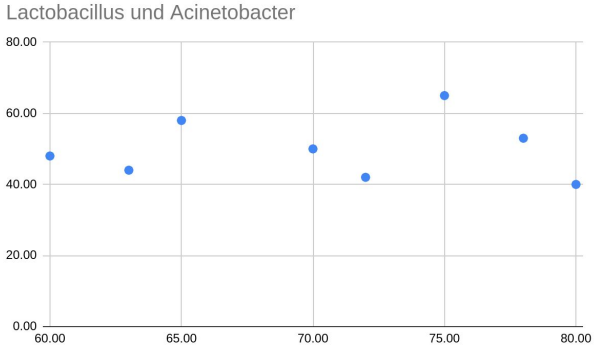
Generated with AI · August 25, 2024 at 1:51 PM (Copilot)

milk sample (ml)	Lactobacillus (#/ml)	Acinetobacter (#/ml)
100	75	65
150	70	50
200	78	53
175	72	42
225	60	48
125	80	40
130	65	58
180	63	44

Collider bias

Field data: milk samples, microbiota

- We assume there are only two types of microbes, *Lactobacillus* and *Acinetobacter*
- Sequencing →normalised counts / ml
- No relationship between #L and #A (r = 0.01)



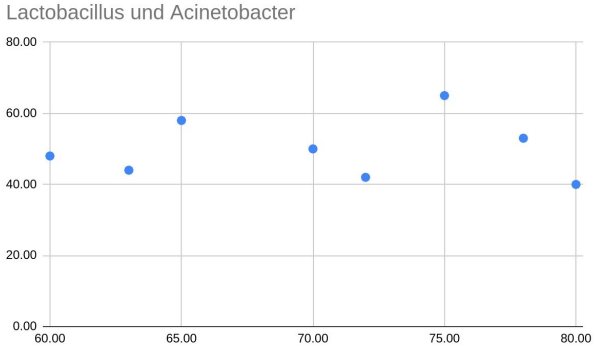
milk sample (ml)	Lactobacillus (#/ml)	Acinetobacter (#/ml)
100	75	65
150	70	50
200	78	53
175	72	42
225	60	48
125	80	40
130	65	58
180	63	44

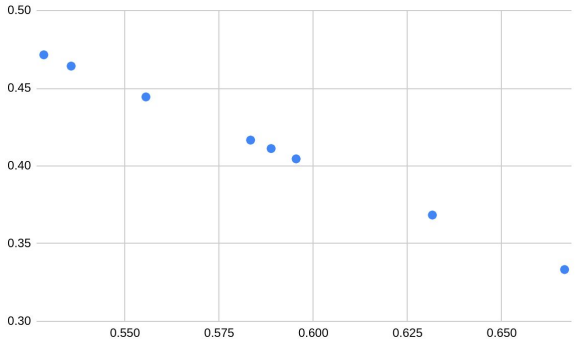
Collider bias

Field data: milk samples, microbiota

- We assume there are only two types of microbes, *Lactobacillus* and *Acinetobacter*
- Sequencing →normalised counts / ml
- No relationship between #L and #A (r = 0.01)
- If we derive relative abundances (new variables), we get correlation! (r = -1)

milk sample (ml)	Lactobacillus (#/ml)	Acinetobacter (#/ml)
100	75	65
150	70	50
200	78	53
175	72	42
225	60	48
125	80	40
130	65	58
180	63	44





Collider bias

- 1000 old coins: 300 pretty, 100 rare, 30 pretty and rare

	rare	common	
pretty	30	270	300
plain	70	630	700
	100	900	1000

Are “prettiness” and rarity associated?

Question: how would you check this?

Collider bias

- 1000 old coins: 300 pretty, 100 rare, 30 pretty and rare

	rare	common	
pretty	30	270	300
plain	70	630	700
	100	900	1000

Are “prettiness” and rarity associated?

Question: how would you check this?

E.g. chi-square test (of independence)

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

χ^2 = chi squared

O_i = observed value

E_i = expected value

Expected values? Proportions of row/column sums over grand total

Collider bias

- 1000 old coins: 300 pretty, 100 rare, 30 pretty and rare

	rare	common	
pretty	30	270	300
plain	70	630	700
	100	900	1000

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

χ^2 = chi squared
O_i = observed value
E_i = expected value

Let's do the math!

(100*300)/1000 =
(100*700)/1000 =
(900*300)/1000 =
(900*700)/1000 =

Collider bias

- 1000 old coins: 300 pretty, 100 rare, 30 pretty and rare

	rare	common	
pretty	30	270	300
plain	70	630	700
	100	900	1000

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

χ^2 = chi squared
O_i = observed value
E_i = expected value

Let's do the math!

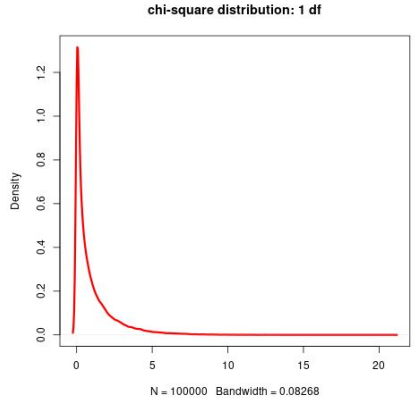
(100*300)/1000 = 30
(100*700)/1000 = 70
(900*300)/1000 = 270
(900*700)/1000 = 630

$X^2 = 0$
d.f. = (n. rows - 1) * (n. columns - 1) = 1

p-value = 1

no deviation from expectations (under H_0)

→no association between prettiness and rarity



Collider bias

old coins: selection of coins for display

- 370 coins:
  - pretty (270)
  - rare (100)

	rare	common	
pretty	30	270	300
plain	70	0	70
	100	270	370

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

χ^2 = chi squared
O_i = observed value
E_i = expected value

Let's redo the math!

(100*300)/370 = 81.1
(100*70)/370 = 18.9
(270*300)/370 = 218.9
(270*70)/370 = 51.1

$X^2 = (30-81.1)^2/81.1 + (70-18.9)^2/18.9 + (270-218.9)^2/218.9 + (0 - 51.1)^2 = 233.38$
d.f. = (n. rows - 1) * (n. columns - 1) = 1

p-value = ~0 → (negative) association between prettiness and rarity

selection bias

(think of unpublished negative results)

Collider bias

- Collider bias typically happens when the data over/under represent some groups of subjects
- The dangers of non-representative samples!
- a.k.a. Berkson's paradox

Are you “collided”?

A little thought exercise

Take 5 minutes, an think of possible colliding effects in your research data

The smoking “gun”

smoking during pregnancy improves survival of the offspring: true?

let’s work this out together

How to deal with the collider bias?

This mainly requires careful:

- i) **experimental design**
- ii) **data analysis** approach
- iii) **interpretation** of results

- avoid inadvertently restricting data/population (e.g. selection bias);
- stratification: analyse data on the entire population and by stratum
- check for potential colliders (implicit/explicit conditioning variables)
- are there characteristics/variables that might be influenced by both the exposure and outcome?
- understand data-generating mechanisms (how the data is generated) help identify potential collider variables
- use Directed Acyclic Graphs (DAGs): visual tool to represent the (causal) relationships between variables → can help identify potential colliders

particularly important in Mendelian Randomization studies (e.g. see here, Coscia et al 2022)

Mediators

$X \rightarrow Z \rightarrow Y$

Do lighters “cause” lung cancer?

- strong positive association between number of lighters and occurrence of lung cancer
- missing variable! Smoking is actually related with lung cancer, but smokers own lighters
- # lighters is a mediator (in-between variable)

Mediators

$X \rightarrow Z \rightarrow Y$

Do lighters “cause” lung cancer?

- strong positive association between number of lighters and occurrence of lung cancer
- missing variable! Smoking is actually related with lung cancer, but smokers own lighters
- # lighters is a mediator (in-between variable)

Can you think of other examples of “mediators”?

Collider bias - resources

For R

- https://jeangoldinginstitute.blogs.bristol.ac.uk/2019/10/28/a-brief-introduction-to-colliders/
- https://jamanetwork.com/journals/jama/fullarticle/2790247
- https://rpubs.com/akhilr/677414
- https://www.r-bloggers.com/2023/06/simulating-confounders-colliders-and-mediators-by-ellis2013nz/
- https://github.com/Osmahmoud/SlopeHunter
- https://observablehq.com/@herbps10/collider-bias