Introduction

Causality

Christina Heinze-Deml

Spring 2019

Overview

- *Lecturer:*
  Christina Heinze-Deml (heinzedeml@stat.math.ethz.ch)

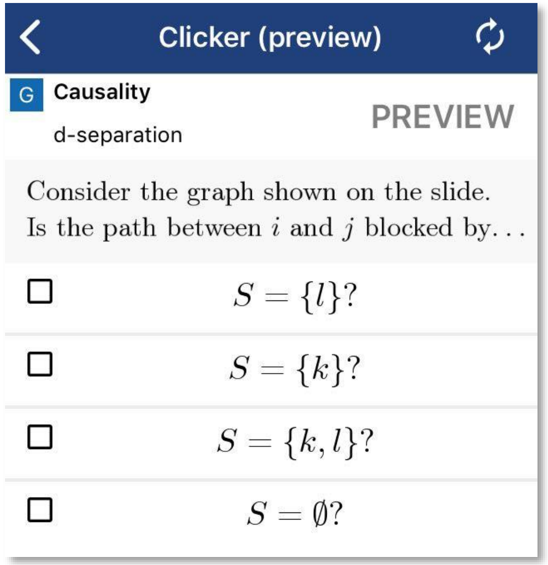
- *Assistant:*
  Niklas Pfister (niklas.pfister@stat.math.ethz.ch)

- Office hours upon request

- Course website: https://stat.ethz.ch/lectures/ss19/causality.php

Lecture style

- Typically: two-hour lecture per week
- Will use ”clicker questions” – please install the ETH EduApp



Lecture style

- Typically: two-hour lecture per week
- Will use ”clicker questions” – please install the ETH EduApp
- R scripts

Take-home exercises

- Take-home exercises available but no separate exercise classes
- Mandatory for PhD students who need ETH credit points
  - Please email me if this applies to you
  - For ECTS credits need to take exam
- Solutions will be provided but no individual corrections

In-class exercises

- Every few weeks in-class exercise session instead of a lecture
- Will use R and Jupyter Notebooks
- Installation requirements are detailed on the website

Further announcements

- Course materials
  - Slides and R scripts used during the lecture will be made available
  - Literature
    - Peters, Janzing and Schölkopf (2017). Elements of Causal Inference.
    - Script from spring semester 2018
    - More links to literature on course website

Further announcements

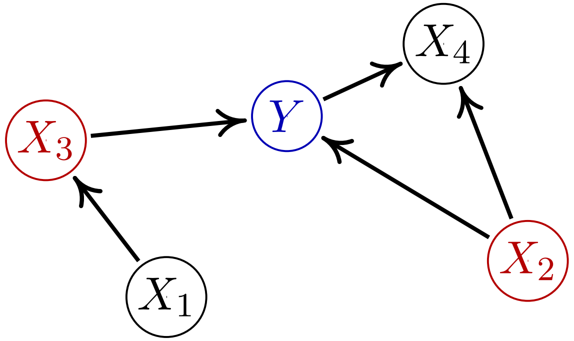
- Course materials
  - Slides and R scripts used during the lecture will be made available
  - Literature
    - Peters, Janzing and Schölkopf (2017). Elements of Causal Inference.
    - Script from spring semester 2018
    - More links to literature on course website

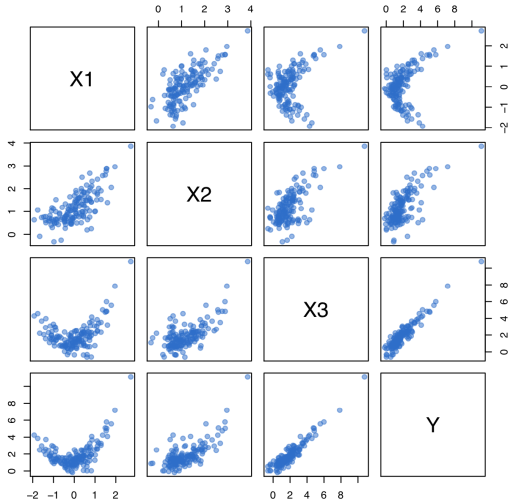
- Exam
  - Two-hour written exam
  - Questions similar to exercises but multiple choice

Questions?

Tentative course outline

- Background and framework
- Methods using the known causal structure
- Learning the causal structure



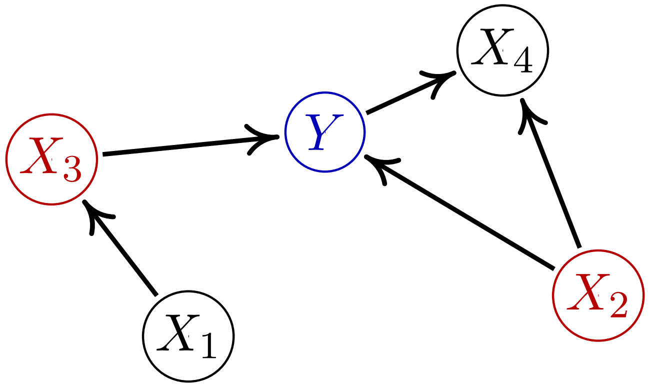


Tentative course outline

- Background and framework
  - Controlled experiments vs. observational studies
  - Simpson’s paradox
  - Graphical models
  - Causal graphical models
  - Structural equation models
  - Interventions
  - …

Tentative course outline

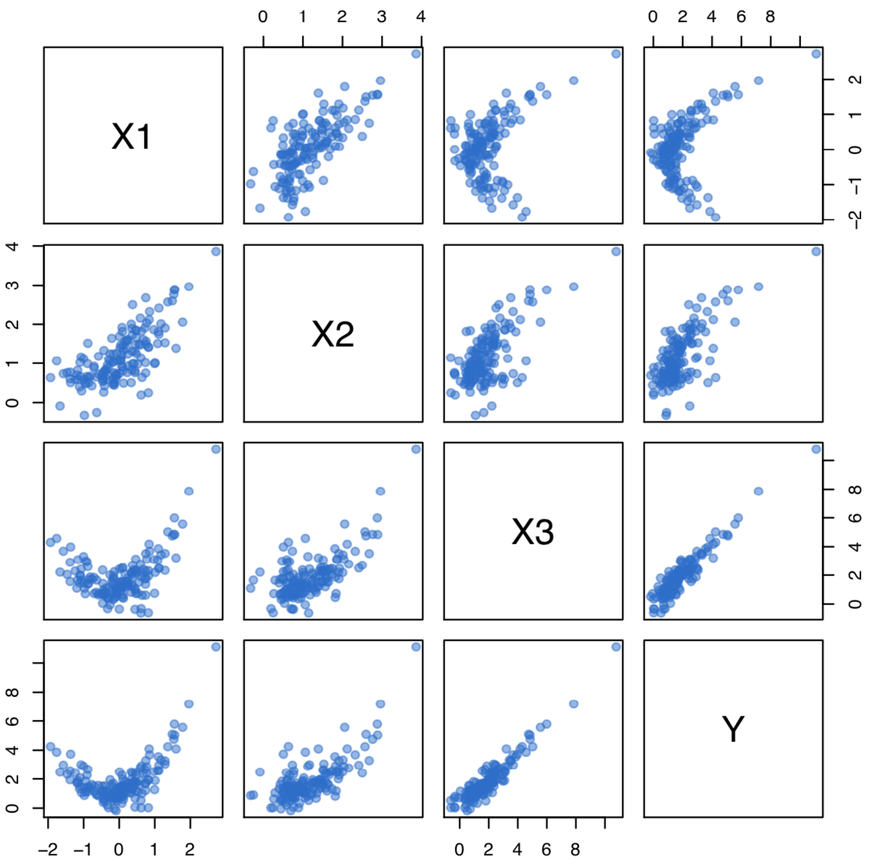
- Methods using the known causal structure
  - Covariate adjustment
  - Instrumental variables
  - Counterfactuals
  - …



$$
\begin{aligned}
Y &= f_Y(\text{parents}(Y), \text{noise}_Y)\\
X_1 &= f_1(\text{parents}(X_1), \text{noise}_1)\\
X_2 &= f_2(\text{parents}(X_2), \text{noise}_2)\\
&\dots\\
X_p &= f_p(\text{parents}(X_p), \text{noise}_p)
\end{aligned}
$$

Tentative course outline

- Learning the causal structure
  - Constraint-based methods
  - Score-based methods
  - Invariant causal prediction
  - …



Today

- Controlled experiments vs. observational studies
- Simpson’s paradox

Controlled experiments

- Setting:
  - E.g. a new drug is introduced
  - Investigators decide who receives it = controlled

- Question: How can we measure its effectiveness in the real world?

- Example: Polio and the Salk Vaccine Field Trial

Salk Vaccine Field Trial

- Polio claimed hundreds of thousands of victims from 1916-1956
  - Mainly children

- By ~1950, several vaccines had been discovered
  - Successful in the lab
  - Most promising one from Jonas Salk

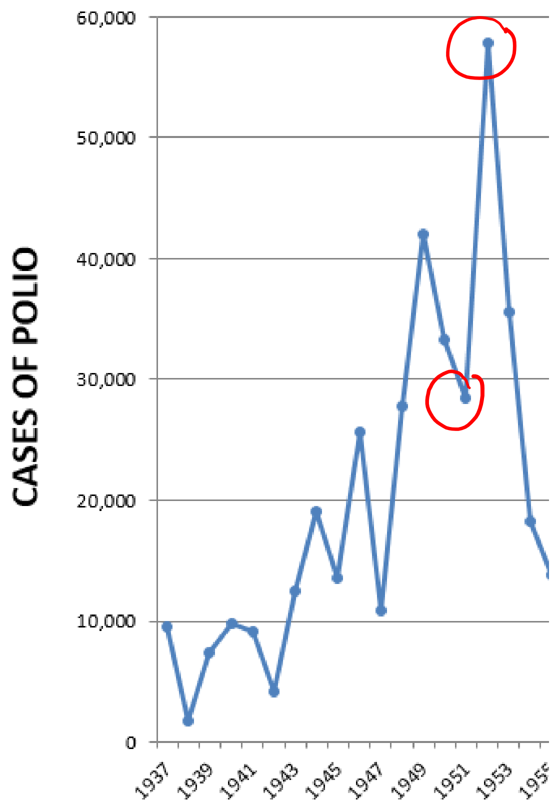
- By 1954 public health service was ready to try the vaccine in the real word
  - I.e. outside the lab on patients

Salk Vaccine Field Trial

- Design 1
  - Give vaccine to a large number of children
  - Compare incidence rate to previous year
  - Caveat: Polio is an epidemic disease

Salk Vaccine Field Trial

- Design 1
  - Give vaccine to a large number of children
  - Compare incidence rate to previous year
  - Caveat: Polio is an epidemic disease

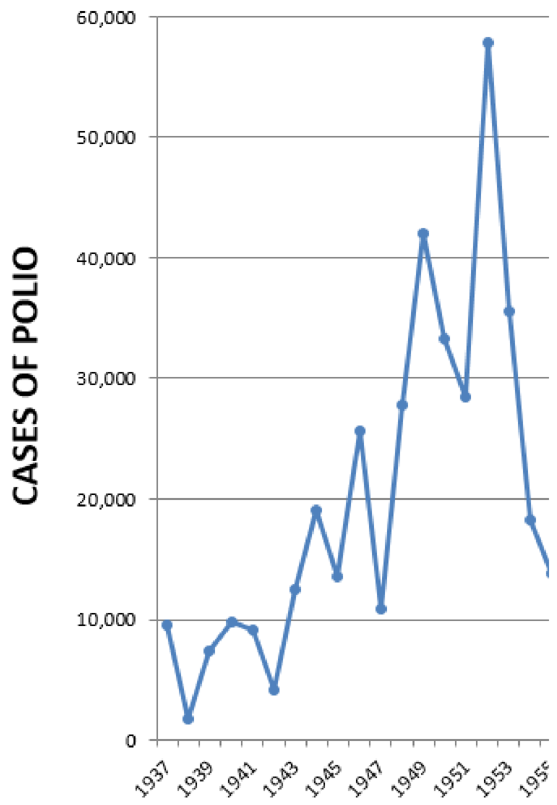


https://vaccines.procon.org/view.additional-resource.php?resourceID=005964

Salk Vaccine Field Trial

- Design 1
  - Give vaccine to a large number of children
  - Compare incidence rate to previous year
  - Caveat: Polio is an epidemic disease

  - Cannot say whether the effect is due to the year, the vaccine or both
    - The two effects are confounded
    - Need to leave some children unvaccinated and use them as a control group
    - Then compare rates at which children get polio in the two groups (treatment vs. control)



https://vaccines.procon.org/view.additional-resource.php?resourceID=005964

Salk Vaccine Field Trial

- Design 2
  - Grade 2: vaccine if parents consent (treatment)
  - Grade 2: no vaccine if no parental consent (control)
  - Grades 1 + 3: no vaccine (control)

Salk Vaccine Field Trial

- Design 2
  - Grade 2: vaccine if parents consent (treatment)
  - Grade 2: no vaccine if no parental consent (control)
  - Grades 1 + 3: no vaccine (control)

  - Caveat 1: polio is contagious, incidence could have been higher in grade 2 vs. 1 & 3
  - Caveat 2:
    - Higher-income parents more likely to consent
    - Children of higher-income parents are more vulnerable to polio (effect of hygiene)

Salk Vaccine Field Trial

- Design 2
  - Grade 2: vaccine if parents consent (treatment)
  - Grade 2: no vaccine if no parental consent (control)
  - Grades 1 + 3: no vaccine (control)

Population

Treatment

Control

No consent

$\boxed{\text{Consent}}$ ~> at risk (higher income)

↳ hard to draw conclusions

Salk Vaccine Field Trial

- Design 2
  - Higher-income parents more likely to consent
  - Children of higher-income parents are more vulnerable to polio (effect of hygiene)
  - Outcome would be biased against the vaccine
  - Family background is confounded with the effect of the vaccine

  - Lesson: Treatment and control groups should be as similar as possible

Salk Vaccine Field Trial

- Lesson: Treatment and control groups should be as similar as possible

Population

Consent

No consent

Treatment

Control

control is chosen from the same population as the treatment group

Salk Vaccine Field Trial

- Design 3
  - Need a control and a treatment group from the same population
  - Only consider children of consenting parents
  - Randomize: 50% chance of being put in the control or the treatment group

Salk Vaccine Field Trial

- Design 3
  - Need a control and a treatment group from the same population
  - Only consider children of consenting parents
  - Randomize: 50% chance of being put in the control or the treatment group

  - Double-blinding:
    - Give placebo to control group and don't tell anyone whether they are in control or treatment group
    - Ensure that effect is due to vaccine and not due to the "idea of getting treatment"
    - Doctors (who decide whether child contracted polio during the experiment) were not told whether a child got real vaccine or placebo

  - Randomized controlled double-blind experiment

Salk Vaccine Field Trial

RCT

Design 2:

	Size	Rate
Grade 2 (consent)	225’000	25
Grades 1 & 3	725’000	54
Grade 2 (no consent)	125’000	44

also children of parents who would not have consented

Design 3:

	Size	Rate
Treatment (consent)	200’000	28
Control (consent)	200’000	71
No consent	350’000	46

contains more children from poorer families → less affected by polio

- Design 2 biased against the vaccine
- Design 3 shows effectiveness of vaccine

Summary

- Method of comparison: treatment vs. control
- If control group is like the treatment group except for the treatment, then any different in outcomes is likely to be caused by the treatment.
- If groups differ wrt factors: danger of confounding

- Best design: double-blind randomized controlled trial (RCT)
- RCT not always possible

ETH zürich

Observational studies

- Setting:
  - No control (or no idea) of the mechanism that assigned “subjects” to different “treatments”
  - Investigators just watch what happens

- Example:
  - Smoking is associated with disease
  - But does it cause diseases?

" correlation does not imply causation "

Observational studies

- Setting:
  - No control (or no idea) of the mechanism that assigned “subjects” to different “treatments”
  - Investigators just watch what happens
- Example:
  - Smoking is associated with disease
  - But does it cause diseases?
  - Cannot force people to smoke
  - Potential confounders: gender, …

confounder

$$\begin{tikzcd} & \boxed{\text{Gender}} \arrow[dl] \arrow[dr] & \\ \text{Smoke} & & \text{Heart disease} \end{tikzcd}$$

Observational studies

- Example:
  - Smoking is associated with disease
  - But does it cause diseases?
  - Cannot force people to smoke
  - Potential confounders:
    - Gender
    - Age
    - Social class
    - Health consciousness
    - Genes

eg Age
Social class
Health
conscious. ~

Confounder

Smoking Disease

Observational studies

- Example:
  - Smoking is associated with disease
  - But does it cause diseases?
  - Cannot force people to smoke
  - Potential confounders: Gender, age, …

- What to do?
  - Compare similar subgroups
    - i.e. males who smoke vs. males who don't
    - "Controlling for confounders"
  - What should we control for?
    - Covered in detail later

Controlled experiments vs. observational studies

Studies

↙ Controls
- ↙ Contemporary
  - ↙ Controlled exp.
    - ↙ Randomized
      - ↙ Double blind *
      - ↘ Other
    - ↘ Not randomized
  - ↘ Observational study
- ↘ Historical

↘ No controls

Observational study → study causal inference based on observational data

* "Gold standard" for causal inference

Example: Chocolate – Nobel Prizes

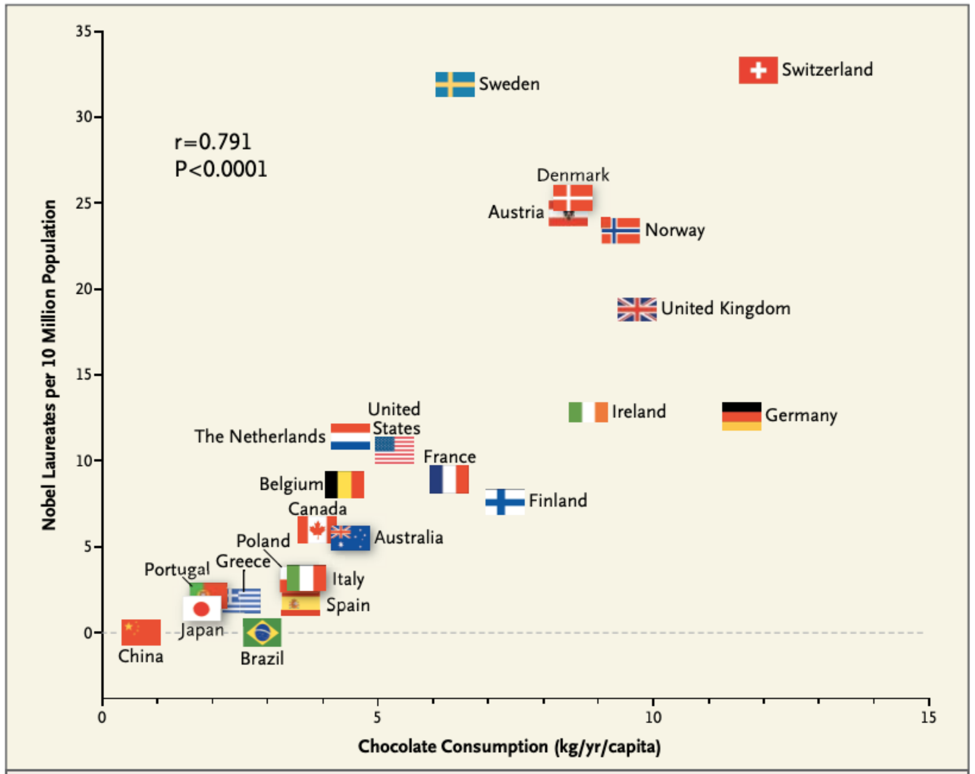


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

- Significant correlation between a country's chocolate consumption & # of Nobel prizes
- This correlation is a property of some observational distribution

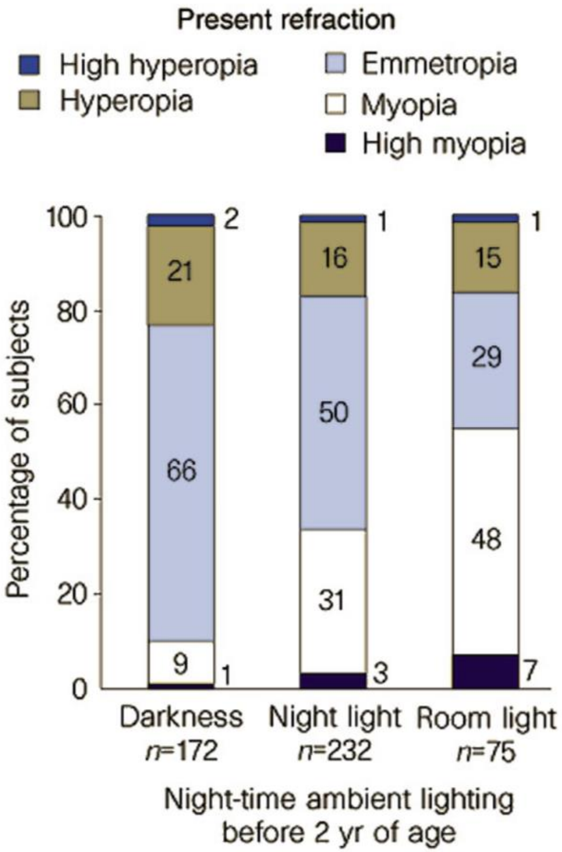
Messerli (2012)

Example: Chocolate – Nobel Prizes

- Significant correlation between a country's chocolate consumption & # of Nobel prizes
- This correlation is a property of some observational distribution
- Must be careful with causal conclusions
- Concern different distributions
  - E.g. scenario where citizens are forced to eat chocolate
- Using background knowledge: correlation stems from hidden variables

Messerli (2012)

Example: Myopia



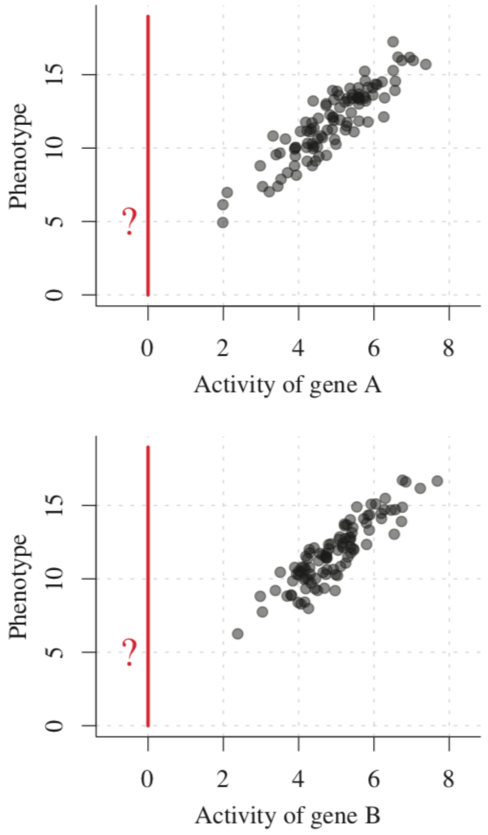
- Dependence between usage of a night light in a child's room and myopia
- False conclusion drawn that absence of darkness is a "potential precipitating factor in the development of myopia"
- Correlation due to parents' myopia
  - More likely to put a night light
  - More likely for child to inherit myopia

Parents' myopia

Night light

child's myopia

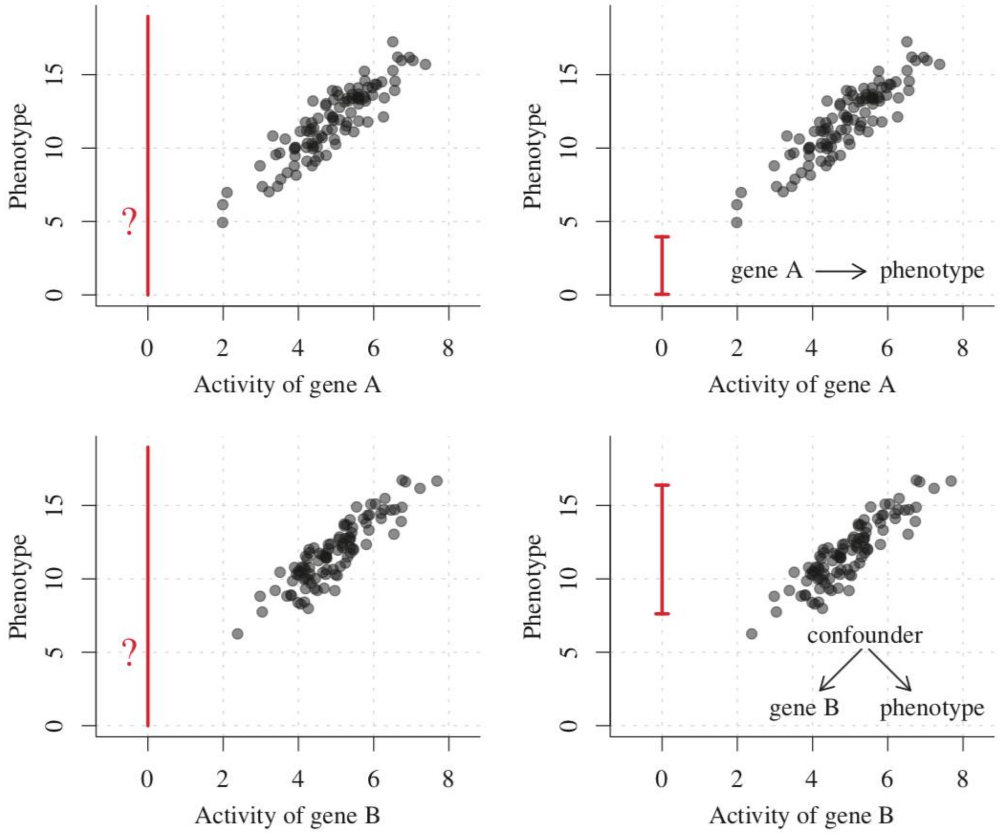
Example: Gene activity



- Strong correlation between gene activity and phenotype
- Can be exploited for classical prediction
- Causal question: What is the phenotype after deleting gene A?
- Cannot answer without knowledge of the causal structure

Peters et al. (2017)

Example: Gene activity



- Top right:
  - Gene A has a causal influence on the phenotype
  - Expect change after the intervention
- Bottom right:
  - Confounder
  - Intervention on gene B will have no effect on the phenotype
- In general, cannot distinguish these two cases based on purely observational data (even with infinite data)

Peters et al. (2017)

Simpson's paradox

	Treatment	Placebo
Male	50/100	150/500
Female	50/500	0/100
Total	100/600	150/600

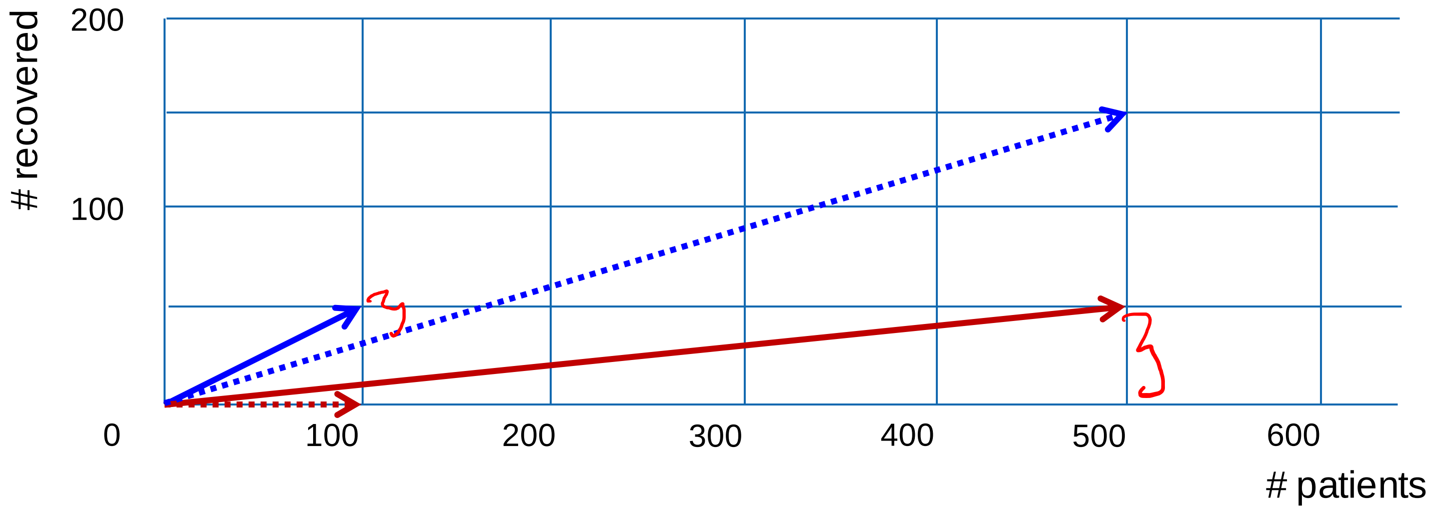
Hypothetical recovery rates, separated by gender

- Among males, treatment is better
- Among females, treatment is better
- Overall, placebo is better

Simpson's paradox

	Treatment	Placebo
Male	50/100	150/500
Female	50/500	0/100
Total	100/600	150/600

Hypothetical recovery rates, separated by gender

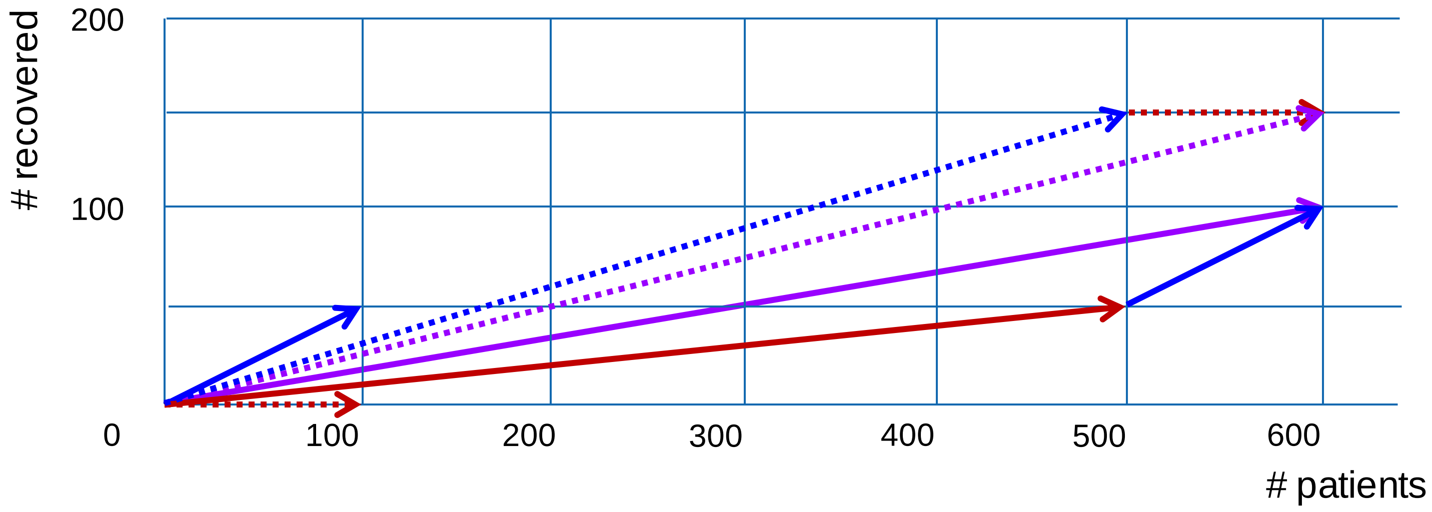


Vector representation: slope is proportion recovered

Simpson's paradox

	Treatment	Placebo
Male	50/100	150/500
Female	50/500	0/100
Total	100/600	150/600

Hypothetical recovery rates, separated by gender



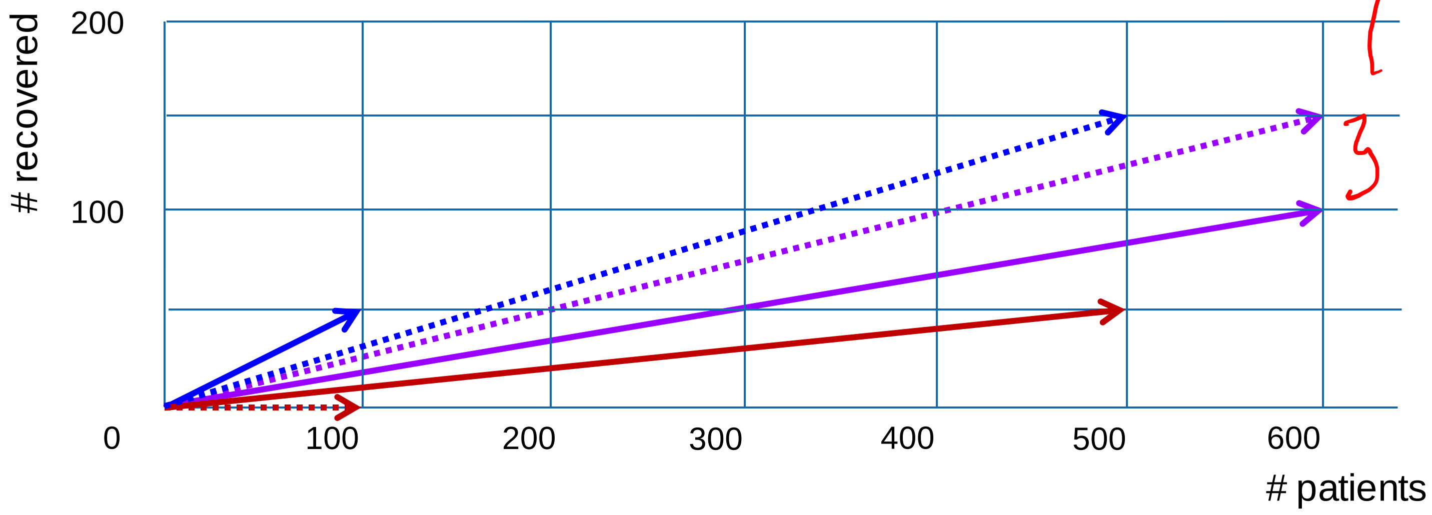
Vector representation: slope is proportion recovered

Simpson's paradox

	Treatment	Placebo
Male	50/100	150/500
Female	50/500	0/100
Total	100/600	150/600

Hypothetical recovery rates, separated by gender

overall placebo is better



Vector representation: slope is proportion recovered

ETH zürich

Simpson's paradox

	Treatment	Placebo
Male	50/100	150/500
Female	50/500	0/100
Total	100/600	150/600

Simpson (1951), in an example similar to this one:
"*The treatment can hardly be rejected as valueless to the race when it is beneficial when applied to males and to females.*"

"control for gender, use the treatment"

Simpson's paradox

	Treatment	Placebo
Male	50/100	150/500
Female	50/500	0/100
Total	100/600	150/600

Simpson (1951), in an example similar to this one:
"*The treatment can hardly be rejected as valueless to the race when it is beneficial when applied to males and to females.*"

replace gender by blood pressure

	Treatment	Placebo
High BP	50/100	150/500
Low BP	50/500	0/100
Total	100/600	150/600

ETH zürich

Simpson's paradox

	Treatment	Placebo
Male	50/100	150/500
Female	50/500	0/100
Total	100/600	150/600

Simpson (1951), in an example similar to this one:
"*The treatment can hardly be rejected as valueless to the race when it is beneficial when applied to males and to females.*"

	Treatment	Placebo
High BP	50/100	150/500
Low BP	50/500	0/100
Total	100/600	150/600

Simpson (1951), in an example similar to this one:
"…, *yet it is the combined table which provides what we would call the sensible answer…*"

"don't control for BP,
don't use the treatment"

Simpson’s paradox

- Same numbers, different conclusions …
- When should we look at the aggregated data, and when at the disaggregated data?

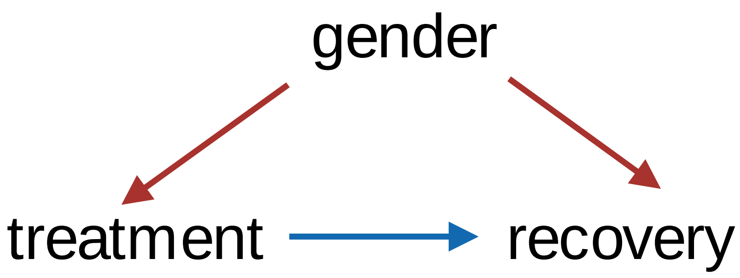
- Perhaps you have seen Simpson’s paradox in intro stats class:
  - Emphasis on numerical phenomenon
  - Take home message: Be careful with conditioning, no clear guidance given

- We should use causal diagrams

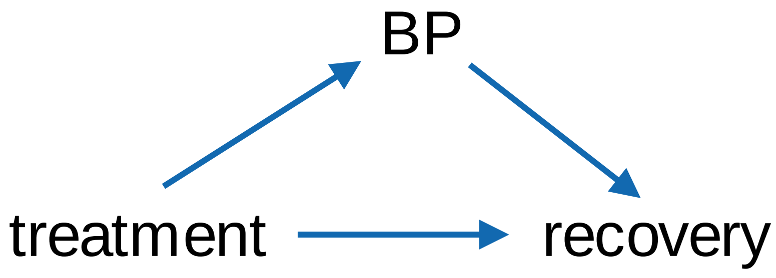
Simpson’s paradox and causal diagrams

- Same numbers, different conclusions….
  - Must use additional information: “story behind the data”, causal assumptions

- Consider total causal effect of treatment on recovery
  - Possible scenarios:



gender is a confounder;
control for gender

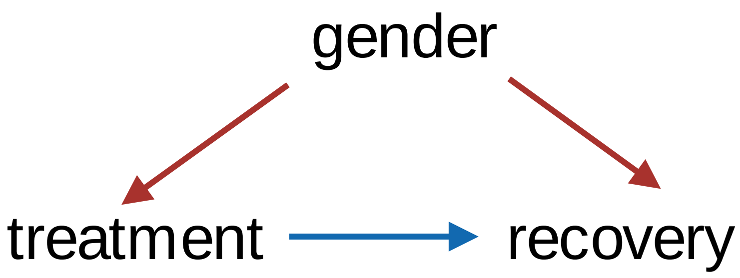


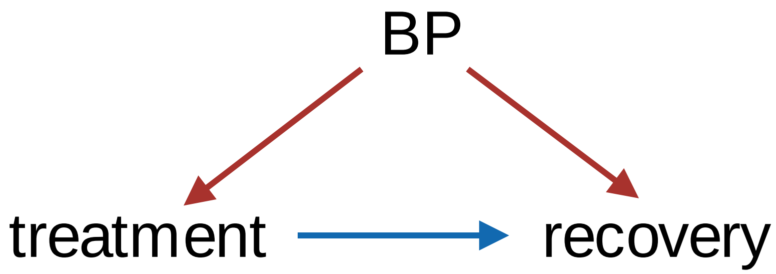
BP is an intermediate variable;
don’t control for BP

Or…..

Simpson's paradox and causal diagrams

- Same numbers, different conclusions….
  - Must use additional information: "story behind the data", causal assumptions
- Consider total causal effect of treatment on recovery
  - Possible scenarios:

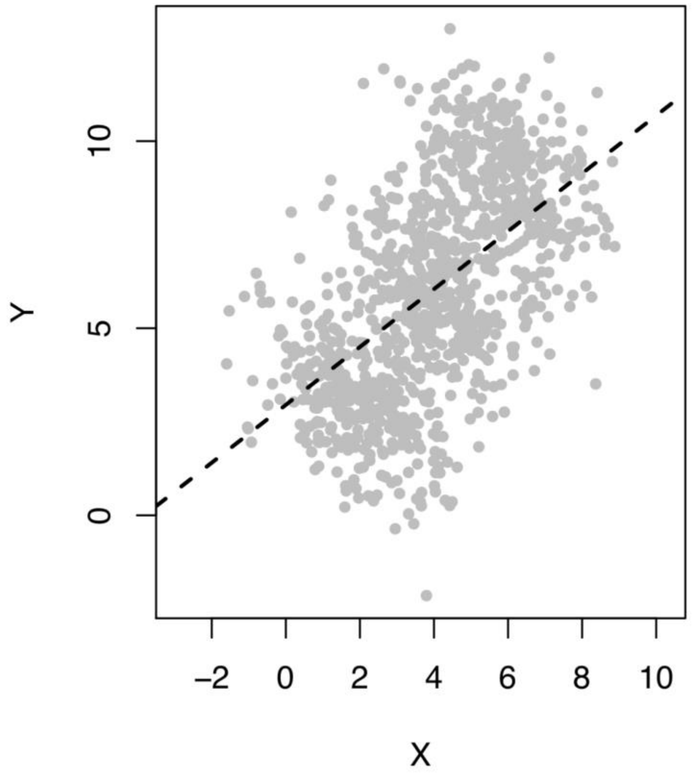




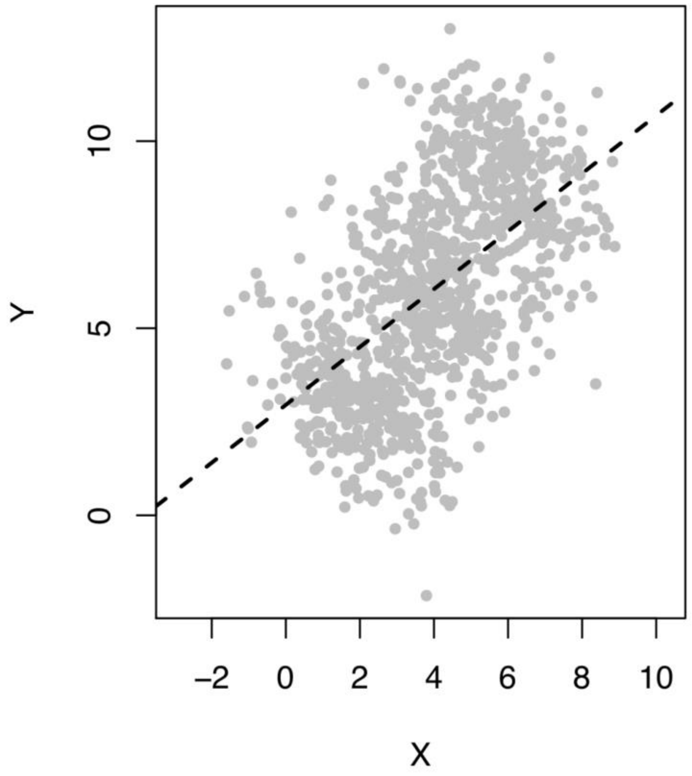
gender is a confounder;
control for gender

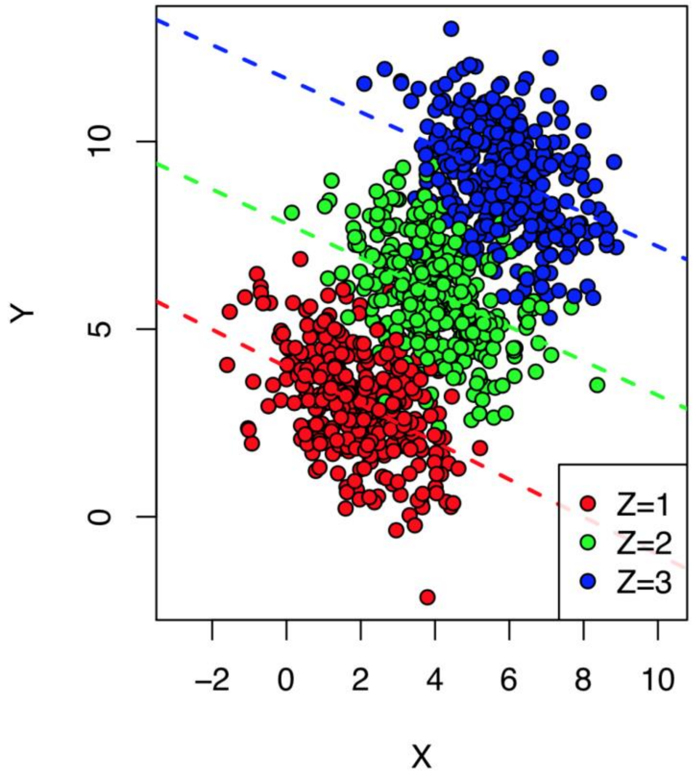
BP is a confounder;
control for BP

Simpson's paradox in regression



Simpson’s paradox in regression






```
n <- 1000
Z <- sample(c(1,2,3),
         size=n,
         replace=TRUE)
X <- 2*Z + rnorm(n)
Y <- 4*Z - 0.5*X + rnorm(n)
```

Simpson's paradox in regression

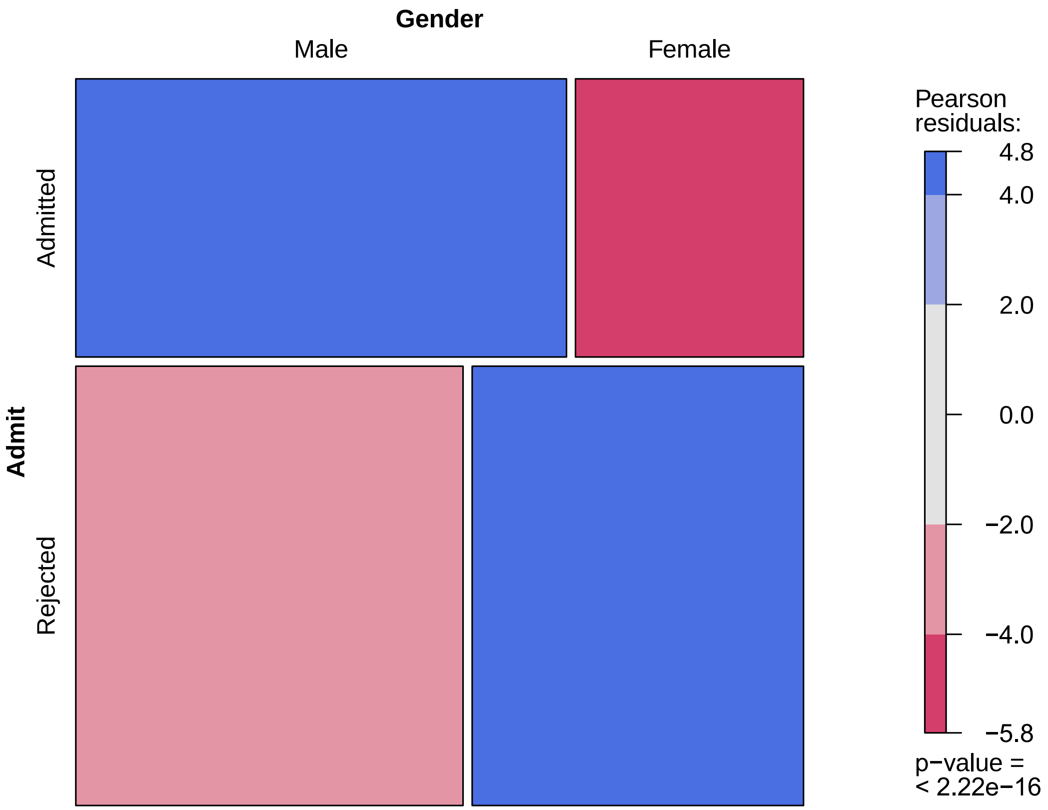
- Different variables in the model can lead to different conclusions
- Simpson's paradox is an extreme case, where we get sign flips

- Multiple regression analysis:
  - Interpretation of regression coefficients depends on model
  - $\beta_j =$ "effect" of X_j on Y when all other variables in the model are "held constant"
  - Little guidance about the choice of variables in the model, apart from standard model selection techniques

- We can use causal reasoning to decide about variables in the model

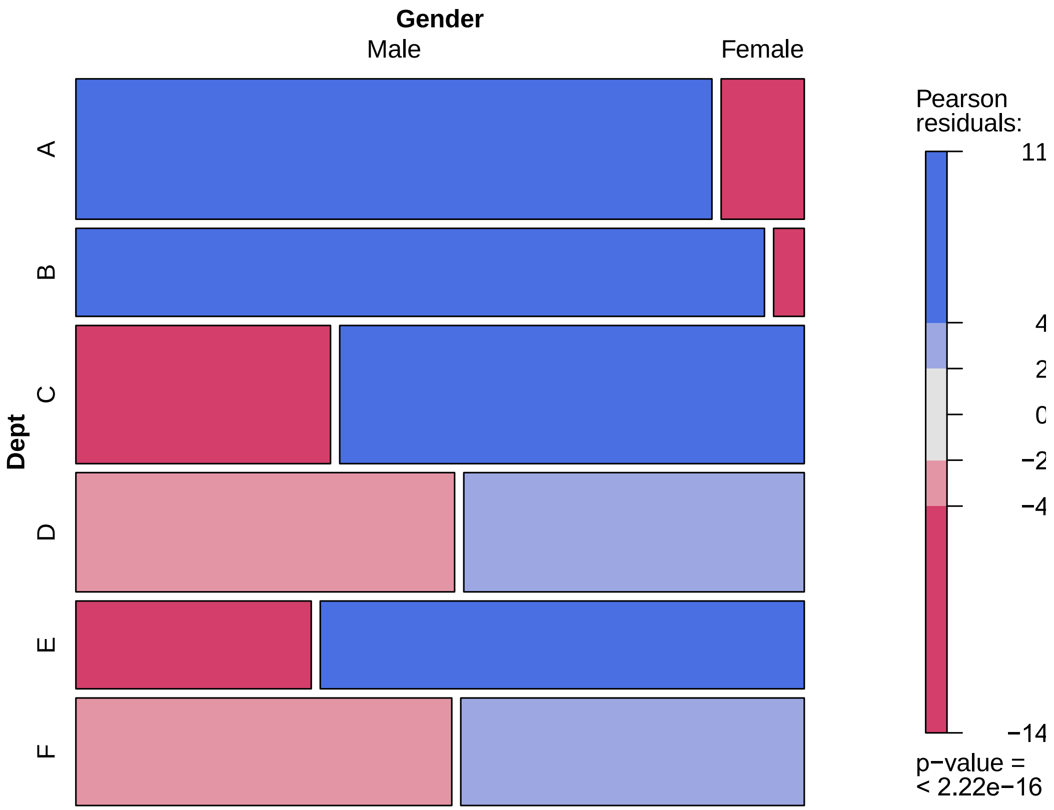
Example: UC Berkeley college admissions

- Claimed gender discrimination in UC Berkeley college admissions in 1973



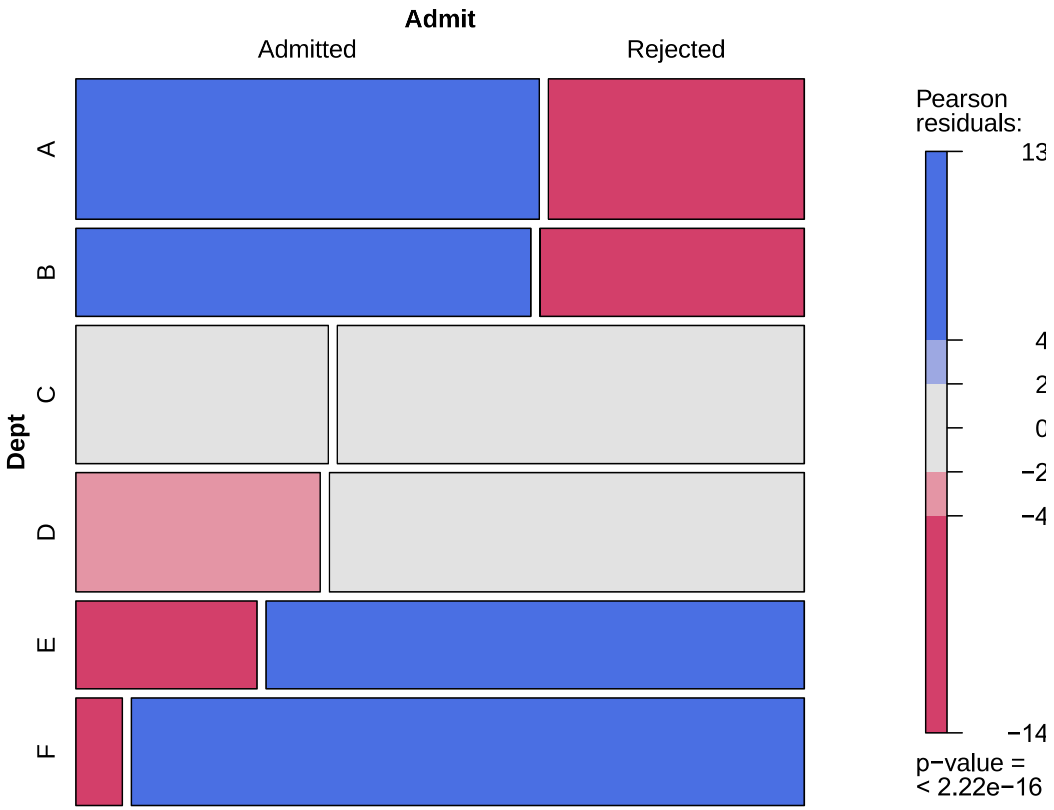
Example: UC Berkeley college admissions

- Where do people apply?



Example: UC Berkeley college admissions

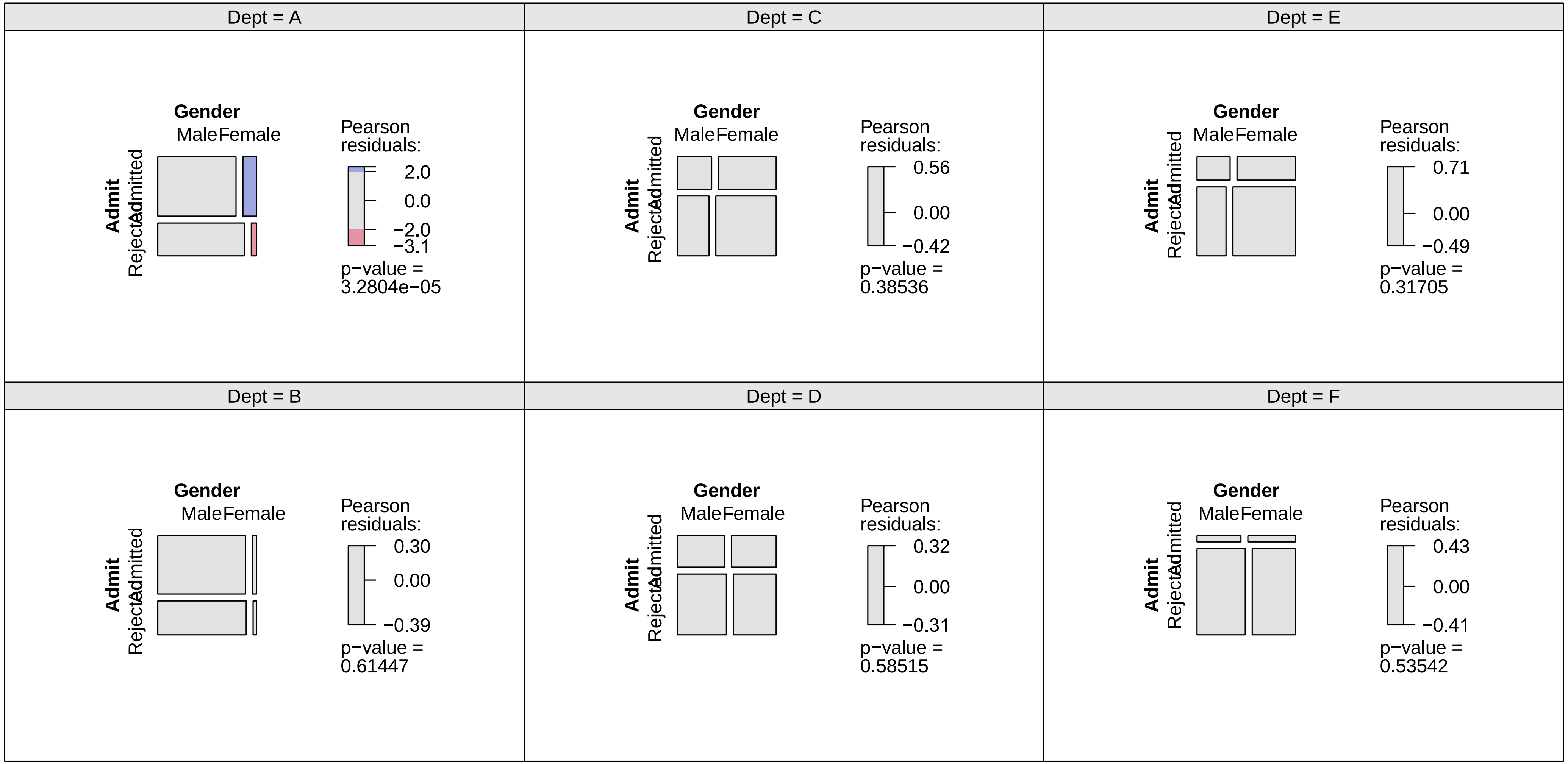
- How selective are departments?



Example: UC Berkeley college admissions

Gender ?

Dpt choice → Admission



Recap

- Concepts to know:
  - Controlled experiments vs. observational studies
  - Simpson's paradox

- Admin:
  - Install EduApp
  - Install R and Jupyter
  - Email me if you are a PhD student who needs ETH credit points

References and acknowledgments

- Salk Vaccine Field Trial
  - Freedman, Pisani and Purves (2007). Statistics. Fourth edition. Chapters 1-2.
  - Slides partly adapted from Lukas Meier
- Examples
  - “Chocolate – Nobel prizes” and “Myopia”: from script by J. Peters & N. Meinshausen
  - “Gene Activity” from Peters, Janzing and Schölkopf (2017). Elements of Causal Inference.
- Simpson’s paradox
  - Slides adapted from M. Maathuis