



universität
wien

053614 VU Statistics for Data Science

Lukas Steinberger
University of Vienna

Winter 2023



- ▶ **Lectures:** Mondays 11:30–13:00 (SR 4) +
+ Thursdays 9:45–11:15 (PC-Room, bi-weekly)
- ▶ slides on Moodle
- ▶ **Homework/Lab sessions:** Thursdays 9:45–11:15
(bi-weekly)
- ▶ **Exception:** Homework session on Monday, October 30th
and January 29th.
- ▶ **mandatory** bi-weekly exercise presentations (Thursdays)
- ▶ grading: 50P exercises, 30P oral final exam or project
- ▶ At least 25P of all exercises completed for passing grade!

GRADING SCHEME



universität
wien

grade	points
1	80 – 71
2	70 – 61
3	60 – 51
4	50 – 41
5	≤ 40

MANDATORY HOMEWORK SESSIONS



universität
wien

- ▶ do the homework, upload solutions and flag solved problems (to get points)
- ▶ randomly selected students
- ▶ present homework solutions on the blackboard/beamer and bring your laptop for coding exercises
- ▶ use any programming language you like (typically R or Python)
- ▶ Your code has to run only on your machine!
- ▶ If you can't present, you lose all the points of that session!

FINAL EXAM



universität
wien

Choose either

- ▶ an oral exam (30min) about the lectures and homework
- ▶ a final project with 15min presentation and questions

THE CHALLENGE



universität
wien

- ▶ What is data science? What is statistics for data science?
- ▶ ⇒ learn classical concepts through modern challenges of statistics
- ▶ Statisticians, mathematicians, computer scientists, natural scientists and engineers from all around the world meet in one class room
- ▶ ⇒ self assessment test (Moodle)
- ▶ **mandatory** but does not count towards your grade
- ▶ do it no later than **Thursday, October 5, 9:45am.**

WHAT TYPE OF COURSE IS THIS?



universität
wien

- ▶ Part of the core theoretical section of DS Master program:
 - ▶ Introduction to Machine Learning (645, 910)
 - ▶ Mathematics for Data Science (645)
 - ▶ Optimisation Methods for Data Science (645)
 - ▶ **Statistics for Data Science** (645, 910)
- ▶ Analyze and understand statistical properties of classical and modern methods mathematically and by simulations.
- ▶ Prerequisites:
 - ▶ Basics of analysis and linear algebra
 - ▶ Basics of probability theory
 - ▶ Basics of statistical inference (very helpful)
 - ▶ No fear of formal mathematical manipulations
 - ▶ Working knowledge of some statistical programming language (R, Python)

WHAT WE WILL COVER...



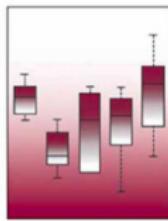
universität
wien

1. Introduction: Data and models (statistical thinking)
2. Simulation and bootstrap methods
3. Linear models
4. Inference for network data
5. Differential Privacy

A COMPREHENSIVE TEXTBOOK



universität
wien



Mathematical Statistics
and Data Analysis

THIRD EDITION

John A. Rice

DUXBURY ADVANCED SERIES

YOUR NEXT TASKS (MOODLE!!!)



universität
wien

- ▶ Do the self assessment test (until Thursday!).
- ▶ Check out the first exercise problem set this week.
- ▶ Upload your solutions and flag those problems you have been able to solve (until Wednesday, Oct. 11th).



How to get in touch:

- ▶ lukas.steinberger@univie.ac.at
- ▶ office hours: by appointment (Oskar-Morgenstern-Platz 1, Room 6.610)
- ▶ Open Moodle forum!

Hope you can enjoy this course!!!

Statistics for Data Science, Winter 2023

1. Introduction: Data and Models

OVERVIEW

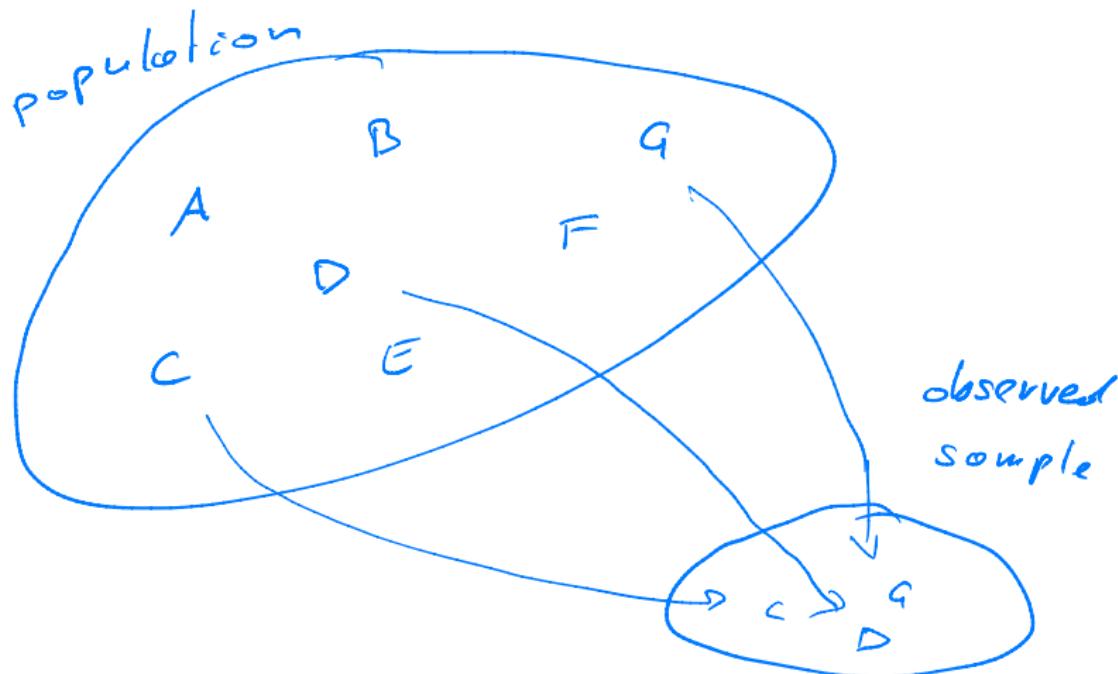
- ▶ Introduction (statistical perspective on data)
- ▶ Recap: Probability Theory
- ▶ Formalism of statistical modeling
- ▶ Estimators, tests and confidence intervals

LEARNING FROM DATA AKA. STATISTICAL INFERENCE



- ▶ data are everywhere!
- ▶ purely descriptive vs. learning/inference
- ▶ To learn (generalize, make inference) we need to know something about our data!
- ▶ ⇒ 'assumptions', statistical model, data generating process

THE STATISTICAL PERSPECTIVE ON DATA: SAMPLING FROM A POPULATION

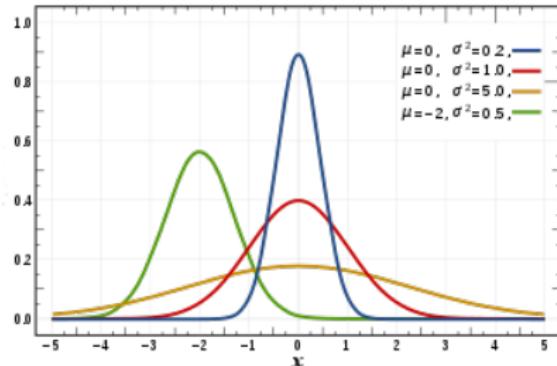


STATISTICAL MODEL VS. ML MODEL

Statistical Model

$$p_{\mu, \sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}\right)$$

$$\mu \in \mathbb{R}, \sigma^2 > 0$$

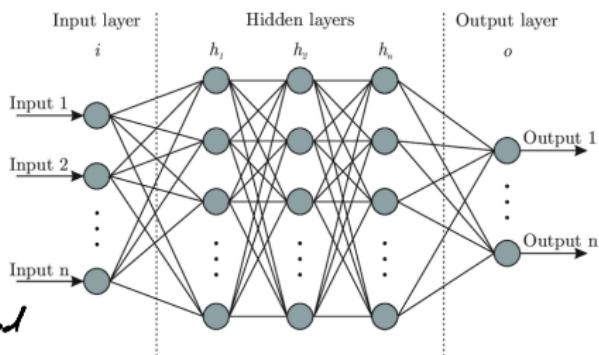


ML Model

$$\hat{Y}_{new} = g(X_{new})$$

ML... machine learning

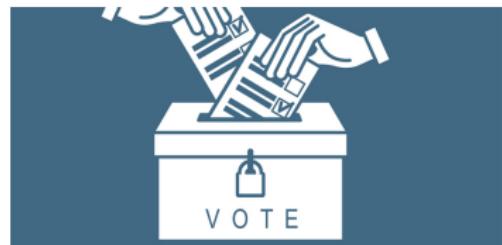
ML \neq maximum likelihood



EXAMPLE 1.1: ELECTORAL SURVEY

Data: *sample*

id	age	sex	party
1	37	m	A
2	59	f	B
:			:
500	25	m	B



Population: *all voters of a country*

EXAMPLE 1.1: ELECTORAL SURVEY

Data:

id	age	sex	party
1	37	m	A
2	59	f	B
:			:
500	25	m	B



Model:

- ▶ individuals are selected randomly from the population
- ▶ independent of each other
- ▶ everybody had the same probability to be selected
- ▶ every selected person gave a complete and truthful answer

EXAMPLE 1.1: ELECTORAL SURVEY

Data:

id	age	sex	party
1	37	m	A
2	59	f	B
:			:
500	25	m	B



Goal: draw conclusions about the unknown fraction of supporters of party A in the whole population

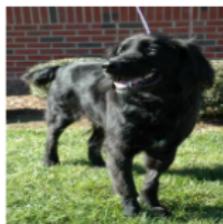
EXAMPLE 1.2: IMAGE CLASSIFICATION

Data:

$$_1 = \mathbf{x}_1$$



$$1$$



$$_0 = \mathbf{x}_2$$



$$0$$



$$1$$



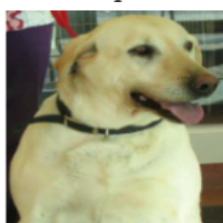
$$0$$



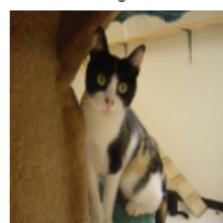
$$0$$



$$1$$



$$0$$



$$(x_1, y_1), \dots, (x_n, y_n)$$

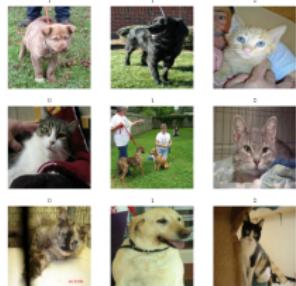
$$x_i \in \mathbb{R}^{3p}, \text{RGB-values}$$

$$y_i \in \{0, 1\}, \text{cat or dog}$$

$$P \dots \# \text{ pixels}$$

EXAMPLE 1.2: IMAGE CLASSIFICATION

Data:



$$(x_1, y_1), \dots, (x_n, y_n)$$

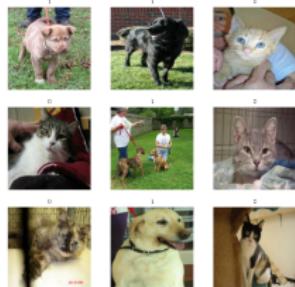
$$x_i \in \mathbb{R}^{3p}, \text{RGB-values}$$

$$y_i \in \{0, 1\}, \text{cat or dog}$$

Population: all images of cats or dogs
with p pixels.

EXAMPLE 1.2: IMAGE CLASSIFICATION

Data:



$$(x_1, y_1), \dots, (x_n, y_n)$$

$$x_i \in \mathbb{R}^{3p}, \text{RGB-values}$$

$$y_i \in \{0, 1\}, \text{cat or dog}$$

Model:

- ▶ Data are realizations of iid pairs of random variables $(X_i, Y_i)_{i=1}^n$.
- ▶ In particular: The function $x \mapsto P(Y_i = 1 | X_i = x)$ is the same for all $i = 1, \dots, n$.

EXAMPLE 1.2: IMAGE CLASSIFICATION

Goal:

- ▶ Find/learn/estimate the function (Bayes classifier)

$$g(x) := \begin{cases} 1, & \text{if } P(Y_1 = 1|X_1 = x) \geq \frac{1}{2}, \\ 0, & \text{if } P(Y_1 = 1|X_1 = x) < \frac{1}{2}. \end{cases}$$

- ▶ Predict the class Y_{new} of the unlabeled picture X_{new} by $\hat{g}(X_{new})$. \Rightarrow generalization

Notice:

- ▶ $g : \mathbb{R}^{3p} \rightarrow \{0, 1\}$ is an unknown/unobserved 'population' quantity
- ▶ We need to estimate/learn g from the sample $(X_i, Y_i)_{i=1}^n$
 $\Rightarrow \hat{g}$

DESCRIPTIVE STATISTICS VS. STATISTICAL INFERENCE

description	inference
summarize and visualize data	learn about population
describe	generalize/estimate
only data, no models	statistical modeling
no assumptions	idealizations/assumptions
all data sets are different/unique	data generating process? sampling error/statistical error uncertainty quantification quantify probability of error

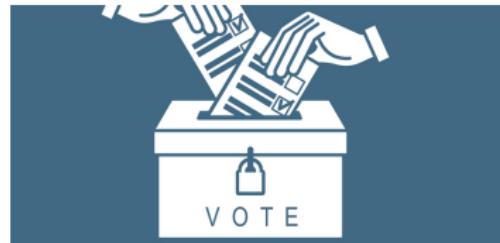
- ▶ Here: statistical inference For data visualization see:
VU Visual and Exploratory Data Analysis

sometimes: estimation vs. inference

EXAMPLE 1.1: ELECTORAL SURVEY

Data:

id	party	X_i
1	A	1
2	B	0
:	:	:
500	B	0



description:

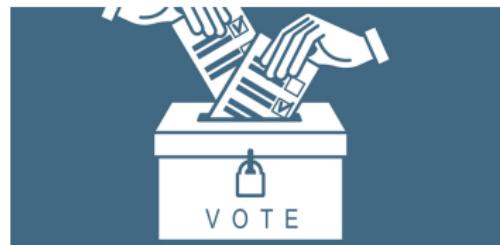
n	votes for A	votes for B
500	318	182

proportion of A votes: $p = \frac{318}{500} = 0.636$

EXAMPLE 1.1: ELECTORAL SURVEY

Data:

id	party	X_i
1	A	1
2	B	0
:	:	:
500	B	0



description:

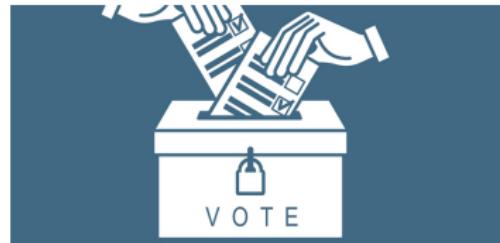
n	votes for A	votes for B
500	318	182

proportion of A votes: $p = \frac{318}{500} = 0.636$

EXAMPLE 1.1: ELECTORAL SURVEY

Data:

id	party	X_i
1	A	1
2	B	0
:	:	:
500	B	0



description:

n	votes for A	votes for B
500	293	207

proportion of A votes: $p = \frac{293}{500} = 0.586$

EXAMPLE 1.1: ELECTORAL SURVEY

Data:

id	party	X_i
1	A	1
2	B	0
:	:	:
500	B	0



Model:

- ▶ $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Bernoulli}(\theta)$
- ▶ i.e. $P(X_i = 1) = 1 - P(X_i = 0) = \theta \in [0, 1]$
- ▶ θ ... true proportion of supporters of party A in the population

EXAMPLE 1.1: ELECTORAL SURVEY

Model:

- $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Bernoulli}(\theta)$
- i.e. $P(X_i = 1) = 1 - P(X_i = 0) = \theta \in [0, 1]$
- θ ... true proportion of supporters of party A in the population

estimation:

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n X_i \quad (= 0.636, 0.586, \text{etc.})$$

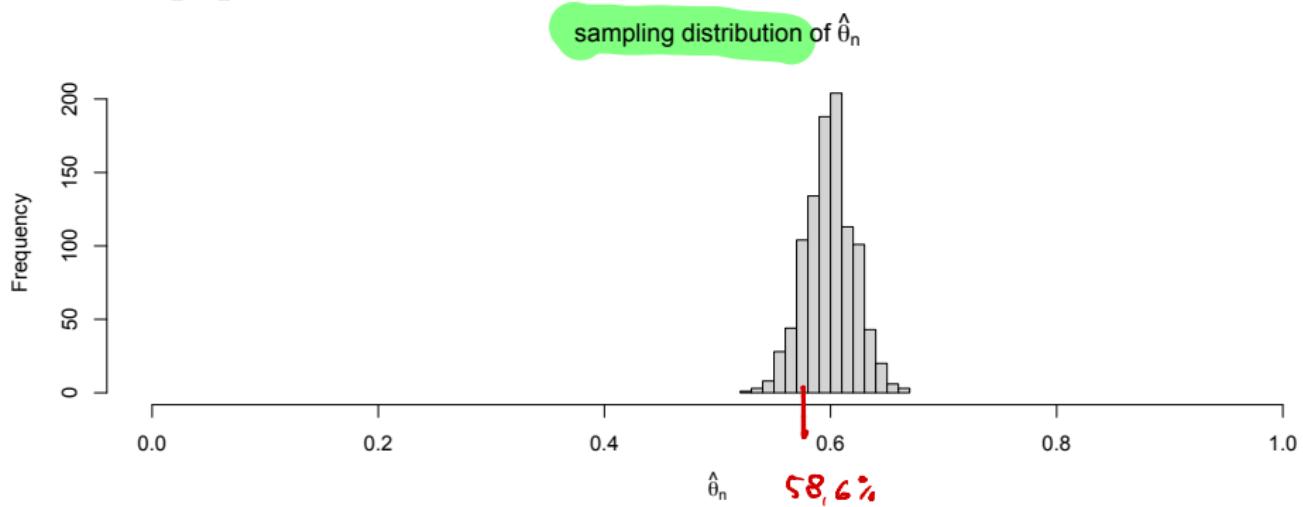
$$\begin{aligned}\mathbb{E}_{\theta}[\hat{\theta}_n] &= \frac{1}{n} \sum_{i=1}^n \underbrace{\mathbb{E}_{\theta}(X_i)}_{1 \cdot \theta + 0 \cdot (1-\theta)} = \frac{1}{n} \sum_{i=1}^n \theta = \theta\end{aligned}$$

“unbiased estimator”

EXAMPLE 1.1: ELECTORAL SURVEY

Model:

- ▶ $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Bernoulli}(\theta)$
- ▶ i.e. $P(X_i = 1) = 1 - P(X_i = 0) = \theta \in [0, 1]$
- ▶ θ ... true proportion of supporters of party A in the population



EXAMPLE 1.1: ELECTORAL SURVEY

Model:

- ▶ $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Bernoulli}(\theta)$
- ▶ i.e. $P(X_i = 1) = 1 - P(X_i = 0) = \theta \in [0, 1]$
- ▶ θ ... true proportion of supporters of party A in the population

inference: (approximate Gaussian level $1 - \alpha$ CI for θ)

$$CI_{\alpha} := \left[\hat{\theta}_n - q_{1-\frac{\alpha}{2}}^{(N)} \hat{\sigma}, \hat{\theta}_n + q_{1-\frac{\alpha}{2}}^{(N)} \hat{\sigma} \right] \quad (= [0.594, 0.678]), \alpha = 0.05$$

$$\hat{\sigma} := \sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}}, \quad q_{1-\frac{\alpha}{2}}^{(N)} : P\left(N(0, 1) \leq q_{1-\frac{\alpha}{2}}^{(N)}\right) = 1 - \frac{\alpha}{2}$$

$$P(\theta \in CI_{\alpha}) \approx 1 - \alpha \quad \text{if } n \text{ is large}$$

$$P(0.594 < \theta < 0.678)$$

“quantifies uncertainty of estimation”

EXAMPLE 1.2: IMAGE CLASSIFICATION

Model:

- ▶ Data are realizations of iid pairs of random variables $(X_i, Y_i)_{i=1}^n$ on $\mathbb{R}^{3p} \times \{0, 1\}$.
- ▶ Optimal predictor (Bayes classifier)

$$g(x) := \begin{cases} 1, & \text{if } P(Y_1 = 1|X_1 = x) \geq \frac{1}{2}, \\ 0, & \text{if } P(Y_1 = 1|X_1 = x) < \frac{1}{2}. \end{cases}$$

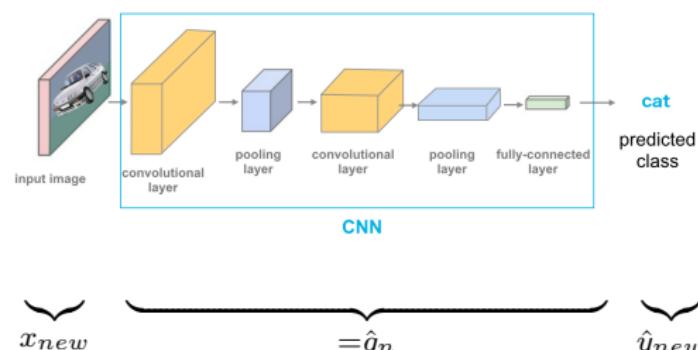
estimation/classification:

Estimate g by a CNN with SGD

$$\hat{g}_n : \mathbb{R}^{3p} \rightarrow \{0, 1\}$$

classification:

$$\hat{y}_{new} = \hat{g}_n(x_{new})$$



EXAMPLE 1.2: IMAGE CLASSIFICATION

- ▶ Data are realizations of iid pairs of random variables $(X_i, Y_i)_{i=1}^n$ on $\mathbb{R}^{3p} \times \{0, 1\}$.
- ▶ estimated/learned classifier $\hat{g}_n : \mathbb{R}^{3p} \rightarrow \{0, 1\}$

validation/error quantification:

split data $S_{train} \cup S_{val} = [n]$, $S_{train} \cap S_{val} = \emptyset$,
 $|S_{train}| = n_1 = n - |S_{val}|$.

train \hat{g}_{n_1} on S_{train}

estimate false positive rate of the classifier \hat{g}_{n_1} by

$$\hat{F}P = \frac{1}{n - n_1} \# \{i \in S_{val} : \hat{g}_{n_1}(X_i) = 1, Y_i = 0\}.$$

“quantifies uncertainty of classification”

Recap: Probability Theory

MATHEMATICAL PROBABILITY

- ▶ Ω (or \mathcal{X}) ... sample space
- ▶ \mathcal{A} ... collection of events/subsets of Ω (σ -algebra)
- ▶ $P : \mathcal{A} \rightarrow [0, 1]$... probability assignment

axioms of probability theory

1. $P(\Omega) = 1.$
2. $P(\emptyset) = 0.$
3. If $A_i \in \mathcal{A}, i \in \mathbb{N}$, are mutually disjoint, then

$\emptyset = \text{empty set}$

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

We say the events $A, B \subseteq \Omega$ are independent if

$$P(A \cap B) = P(A)P(B).$$

MATHEMATICAL PROBABILITY: DISCRETE (EXAMPLE)

Tossing a regular six sided die

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

\mathcal{A} = all subsets of Ω

$$= \{\emptyset, \{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{1, 2\}, \{1, 3\}, \dots, \{1, 2, 3, 4, 5\}, \Omega\}$$

$$P(\{i\}) := \frac{1}{6}, \quad i = 1, \dots, 6$$

compute probability of an odd number:

$$\begin{aligned} P(\{1, 3, 5\}) &= P(\{1\} \cup \{3\} \cup \{5\} \cup \emptyset \cup \emptyset \dots) \\ &\stackrel{3.}{=} P(\{1\}) + P(\{3\}) + P(\{5\}) + P(\emptyset) + P(\emptyset) + \dots \\ &\stackrel{2.}{=} \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + 0 + 0 + \dots = \frac{3}{6} = \frac{1}{2} \end{aligned}$$

MATHEMATICAL PROBABILITY: DISCRETE (EXAMPLE)

Tossing two fair coins

$$\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$$

\mathcal{A} = all subsets of Ω

$$\begin{aligned} &= \{\emptyset, \{(H, H)\}, \{(H, T)\}, \{(T, H)\}, \{(T, T)\}, \{(H, H), (H, T)\}, \dots \\ &\quad \{(H, H), (H, T), (T, H)\}, \Omega\} \end{aligned}$$

$$P(\{(a, b)\}) := \frac{1}{4}, \quad a, b \in \{H, T\}$$

The events $A = \{(H, H), (H, T)\}$ and $B = \{(H, T), (T, T)\}$ are independent:

$$P(A) = P(\{(H, H)\}) + P(\{(H, T)\}) = \frac{1}{2} = P(B)$$

$$P(A \cap B) = P(\{(H, T)\}) = \frac{1}{4} = P(A)P(B)$$

MATHEMATICAL PROBABILITY: DISCRETE

- ▶ \mathcal{X} = a finite or countably infinite set ($\{x_1, x_2, \dots\}$)
- ▶ \mathcal{A} = the collection of all subsets of \mathcal{X}
- ▶ $f : \mathcal{X} \rightarrow [0, 1]$ a **probability mass function (pmf)**, i.e.,

$$\underbrace{f(x) := P(\{x\}),}_{x \in \mathcal{X}}$$

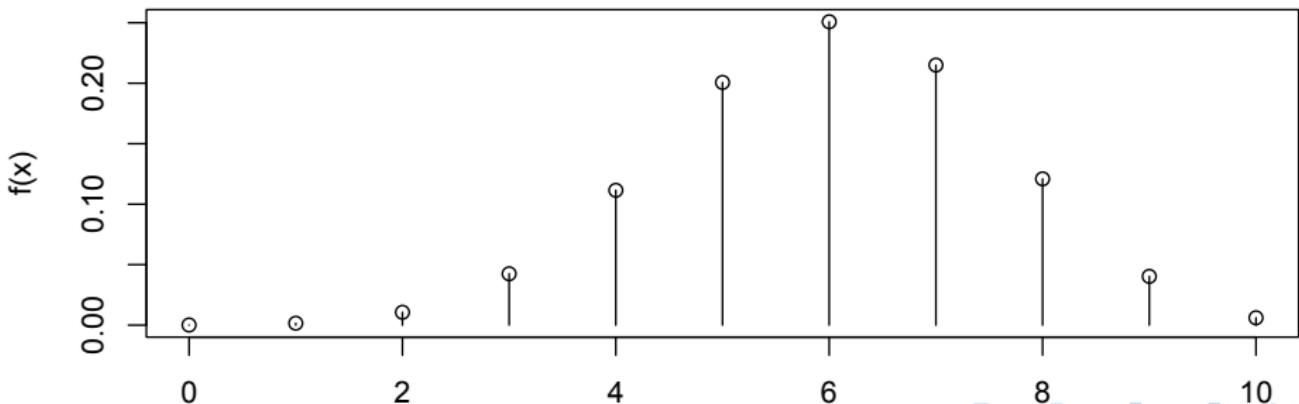
$$P(A) := \sum_{x \in A} f(x), \quad \text{for } A \in \mathcal{A}.$$

MATHEMATICAL PROBABILITY: DISCRETE (EXAMPLE)

- ▶ ~~$\mathcal{X} = \mathbb{N}_0 = \{0, 1, 2, \dots\}$~~ $\mathcal{X} = \{0, 1, 2, \dots, n\}$
- ▶ parameters $n \in \mathbb{N}, \theta \in [0, 1]$.

$$f(x) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}$$

pmf of binomial distribution



MATHEMATICAL PROBABILITY: CONTINUOUS

$$P(\{x\}) = ? \quad x \in \mathbb{R}$$

- ▶ $\mathcal{X} \subseteq \mathbb{R}^d$
- ▶ $\mathcal{A} =$ a collection of (measurable) subsets of ~~\mathbb{R}~~ \mathcal{X}
- ▶ $f : \mathcal{X} \rightarrow [0, \infty)$ a **probability density function (pdf)**, i.e.,

$$f(x) \geq 0, \quad \int_{\mathcal{X}} f(x) dx = 1.$$

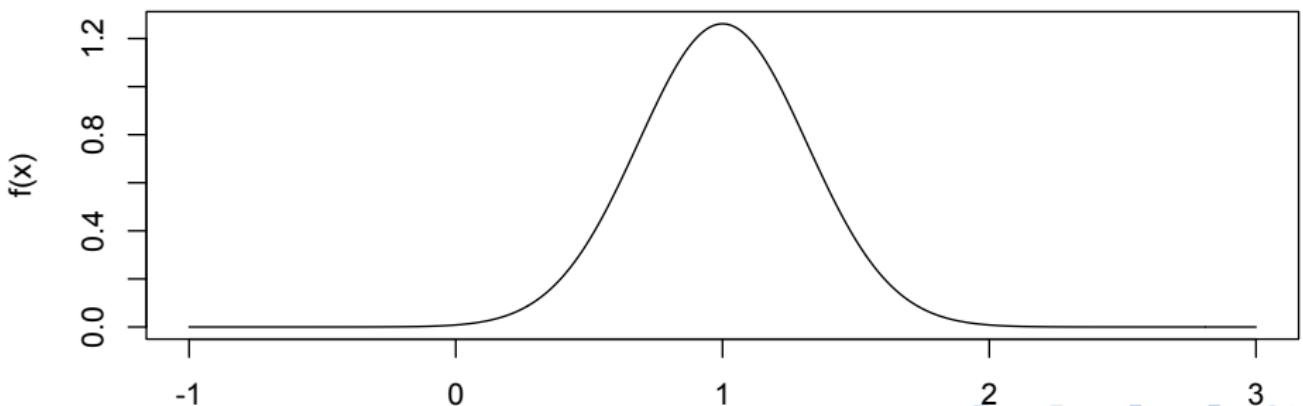
$$P(A) := \int_A f(x) dx, \quad \text{for } A \in \mathcal{A}.$$

MATHEMATICAL PROBABILITY: CONTINUOUS (EXAMPLE)

- ▶ $\mathcal{X} = \mathbb{R}$
- ▶ Parameters $\mu \in \mathbb{R}$ and $\sigma^2 \in (0, \infty)$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

density of normal distribution



RANDOM VARIABLES

Informally: A random variable X represents all potential realizations of a random experiment.

Formally: A random variable X is a (measurable) function $X : \Omega \rightarrow \mathbb{R}$ and $P(X \in A) := P(\{\omega \in \Omega : X(\omega) \in A\})$.

We say that X has pmf or pdf f (notation: $X \sim f$), if

$$P(X \in A) = \begin{cases} \sum_{x \in A} f(x), & \text{in the discrete case} \\ \int_A f(x) dx, & \text{in the continuous case.} \end{cases}$$

For $g : \mathbb{R} \rightarrow \mathbb{R}$,

$$Q.8 \quad g(x) = x \quad \mathbb{E} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = ?$$

$$\mathbb{E}[g(X)] = \begin{cases} \sum_{x \in \mathcal{X}} g(x)f(x), & \text{discrete} \\ \int_{\mathcal{X}} g(x)f(x)dx, & \text{continuous.} \end{cases}$$

$$:= \begin{pmatrix} \mathbb{E}(x_1) \\ \mathbb{E}(x_2) \end{pmatrix}$$

$$\text{Var}[g(X)] := \mathbb{E}[(g(X) - \mathbb{E}[g(X)])^2] = \mathbb{E}[g(X)^2] - \mathbb{E}[g(X)]^2$$

RANDOM VARIABLES

We say that the random variables X and Y are **independent** if

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B),$$

for all events A and B .

$$\begin{aligned} & P(\{\omega : X(\omega) \in A, Y(\omega) \in B\}) \\ &= P(\{\omega : X(\omega) \in A\} \cap \{\omega : Y(\omega) \in B\}) \end{aligned}$$

RANDOM VARIABLES

Can do abstract computations without fixing any particular number/realization/sample.

For X, Y real RVs and $a, b \in \mathbb{R}$:

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

$$\text{Var}[aX] = a^2 \text{Var}[X]$$

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] \quad \text{if } X \text{ and } Y \text{ are independent}$$

$$\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] \quad \text{if } X \text{ and } Y \text{ are independent}$$

RANDOM VARIABLES (EXAMPLE)

- ▶ data: X_1, \dots, X_n i.i.d. (= independent and identically distributed)
- ▶ represents all potential samples of size n
- ▶ sample mean $\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i$

$$\mathbb{E}(X_1) = \mu$$

$$= \mathbb{E}(X_2) = \dots = \mathbb{E}(X_n)$$

$$\begin{aligned}\mathbb{E}[\bar{X}_n] &= \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \mathbb{E}\left(\sum_{i=1}^n X_i\right) \\ &= \frac{1}{n} \sum_{i=1}^n \underbrace{\mathbb{E}(X_i)}_{=\mu} = \frac{1}{n} n \cdot \mu = \mu\end{aligned}$$

$$\begin{aligned}\text{Var}[\bar{X}_n] &= \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^n X_i\right) \\ &= \frac{1}{n^2} \sum_{i=1}^n \underbrace{\text{Var}(X_i)}_{=\sigma^2} = \frac{1}{n^2} n \cdot \sigma^2 = \frac{\sigma^2}{n}\end{aligned}$$

CUMULATIVE DISTRIBUTION FUNCTION (CDF)

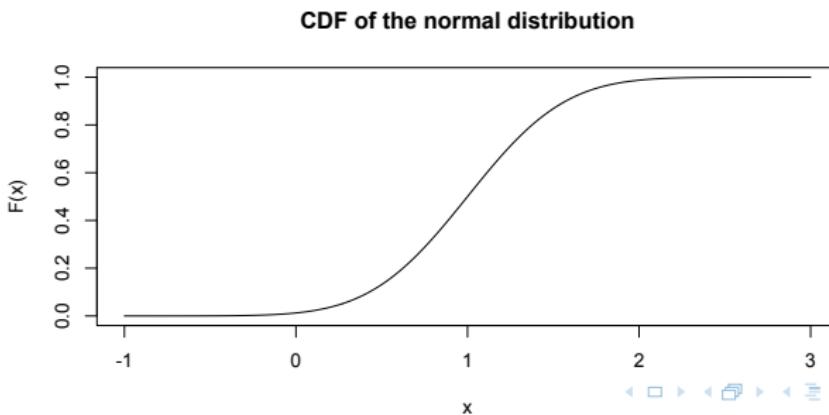
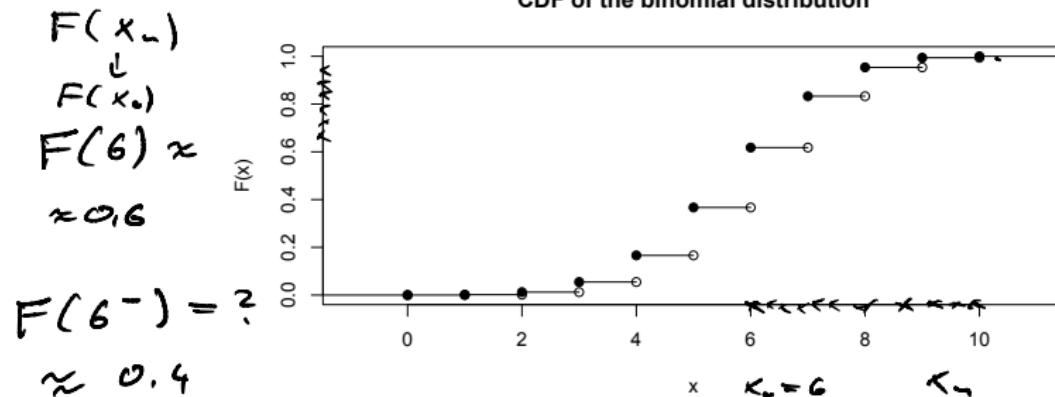
Unifying discrete and continuous distributions:

$$\underbrace{F_X(x) := P(X \leq x), \quad x \in \mathbb{R}}_{\swarrow}$$

Properties:

- ▶ $F : \mathbb{R} \rightarrow [0, 1]$
- ▶ F is non-decreasing
- ▶ $\lim_{x \rightarrow -\infty} F(x) = 0, \lim_{x \rightarrow +\infty} F(x) = 1$
- ▶ $F(x_0^+) := \lim_{\underline{x \downarrow x_0}} F(x) = F(x_0)$ (right-continuity)

CUMULATIVE DISTRIBUTION FUNCTION (CDF)



CUMULATIVE DISTRIBUTION FUNCTION (CDF)

The CDF encodes all information of a univariate distribution.

continuous:

$$F(x) = \int_{-\infty}^x f(u) du, \quad f(x) = \frac{d}{dx} F(x).$$

pdf *cdf*

discrete:

$$F(x) = \sum_{y \leq x} f(y), \quad f(x) = F(x) - F(x^-).$$

pmf *cdf* *jump size*

$$F(x^-) := \lim_{y \uparrow x} F(y)$$

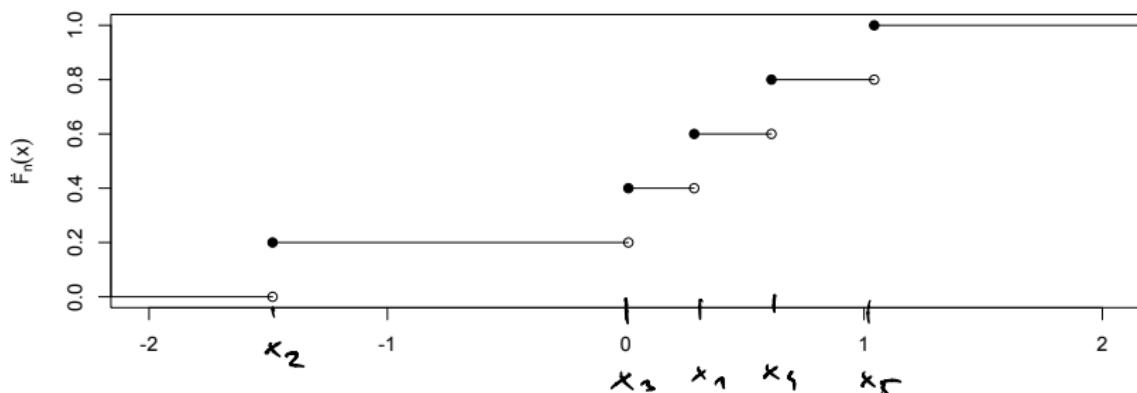
EMPIRICAL CDF

Given data $x_1, \dots, x_n \in \mathbb{R}$, the corresponding empirical cdf is given by

$$\hat{F}_n(x) := \frac{\#\{i : x_i \leq x\}}{n} = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty, x]}(x_i)$$

$$\mathbb{1}_A(x) := \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$$

empirical CDF



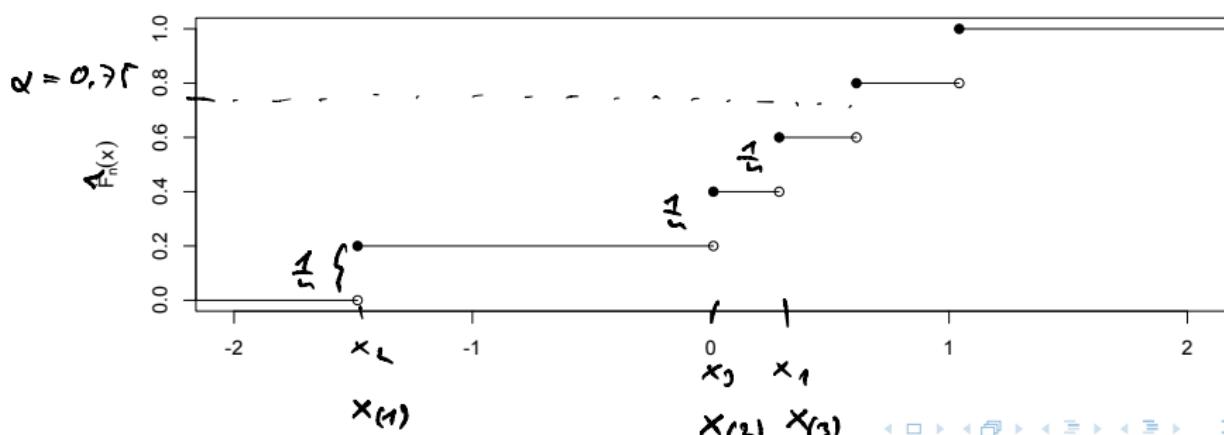
EMPIRICAL CDF

Given data $x_1, \dots, x_n \in \mathbb{R}$, let $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ be the ordered values. Then (assuming no ties)

$$\hat{F}_n(x) = \begin{cases} 0, & \text{if } x < x_{(1)}, \\ \frac{i}{n}, & \text{if } x_{(i)} \leq x < x_{(i+1)}, \quad i = 1, \dots, n-1, \\ 1, & \text{if } x_{(n)} \leq x. \end{cases}$$

$\hat{F}_n(q_\alpha) \neq 0.75$

empirical CDF



QUANTILES

Informally: For a given probability $\alpha \in [0, 1]$, the α -quantile of a distribution F is the number q_α such that exactly α of the probability mass lies at or below q_α , i.e.,

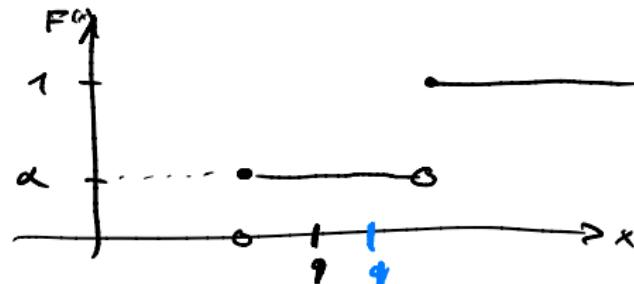
$$F(q_\alpha) = P(X \leq q_\alpha) = \alpha.$$

Note: q_α may not exist!

QUANTILES

$$\cdot F(q) = \alpha \geq \alpha$$

$$\cdot F(q^-) = F(q) = \alpha \leq \alpha$$



Formally: For a given probability $\alpha \in [0, 1]$ and a CDF $F : \mathbb{R} \rightarrow [0, 1]$, an α -quantile of F is an extended real number $q_\alpha \in \bar{\mathbb{R}} := [-\infty, \infty]$, such that

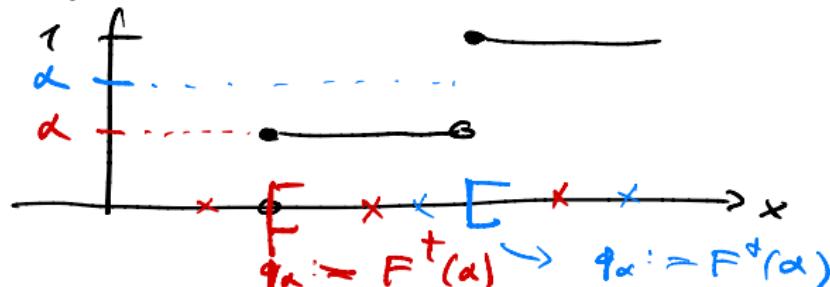
$$F(q_\alpha) \geq \alpha \quad \text{and} \quad F(q_\alpha^-) \leq \alpha.$$

Note: q_α may not be unique! How do we pick one?

QUANTILES

Quantile function aka. generalized inverse

$$F^\dagger(\alpha) := \inf\{x \in \mathbb{R} : F(x) \geq \alpha\}, \quad \alpha \in [0, 1]$$



- ▶ Is the smallest of all α -quantiles of F .
- ▶ If F is invertible, then $F^\dagger(\alpha) = F^{-1}(\alpha)$.
- ▶ For $i = 1, \dots, n$, we have $\hat{F}_n^\dagger(\alpha) = x_{(i)}$ if, and only if, $\frac{i-1}{n} < \alpha \leq \frac{i}{n}$. (empirical quantiles = order statistics)

Convention: $\inf \emptyset := +\infty$.

HW

THE LAW OF LARGE NUMBERS (WEAK)

$X_i \sim \text{Bernoulli}(\omega)$

$\mathbb{E} X_i = \omega$

Let X_1, \dots, X_n be i.i.d. random variables with $\mathbb{E}[X_1] = \mu \in \mathbb{R}$ and $\text{Var}[X_1] = \sigma^2 < \infty$. Then

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow[n \rightarrow \infty]{i.p.} \mu.$$

"in probability"

More precisely, for every $\varepsilon > 0$,

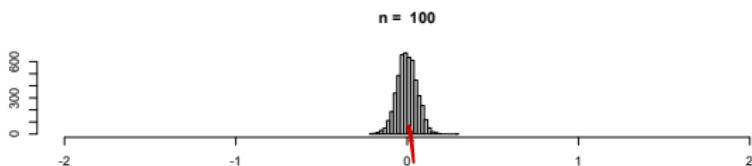
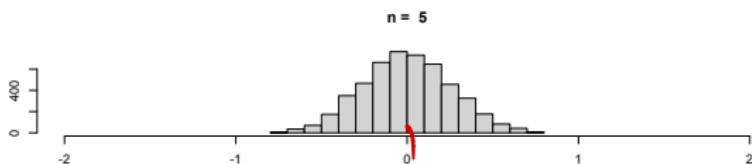
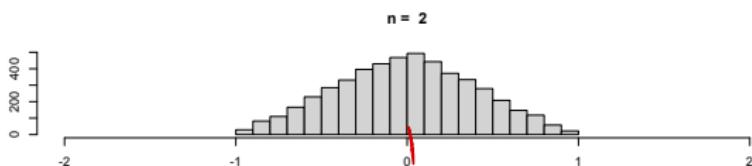
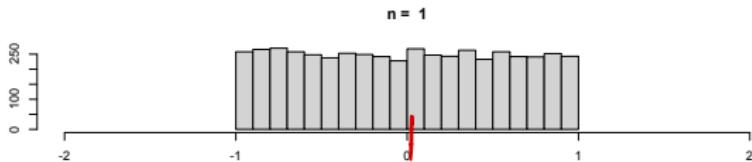
$$P(|\bar{X}_n - \mu| > \varepsilon) \xrightarrow{n \rightarrow \infty} 0.$$

Proof: on black board

THE LAW OF LARGE NUMBERS

histograms
of 1000

X



$$\mu = 0$$

THE CENTRAL LIMIT THEOREM

Let X_1, \dots, X_n be i.i.d. random variables with $\mathbb{E}[X_1] = \mu \in \mathbb{R}$ and $\text{Var}[X_1] = \sigma^2 < \infty$. Then

$$\sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} \xrightarrow[n \rightarrow \infty]{d.} \text{in distribution } N(0, 1).$$

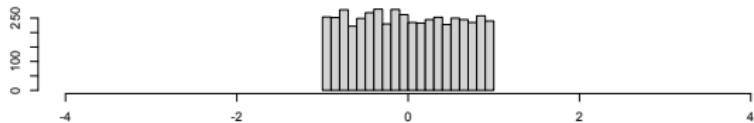
More precisely, for every $x \in \mathbb{R}$,

$$P \left(\sqrt{n} \frac{\bar{X}_n - \mu}{\sigma} \leq x \right) \xrightarrow[n \rightarrow \infty]{} \Phi(x) := P(N(0, 1) \leq x).$$

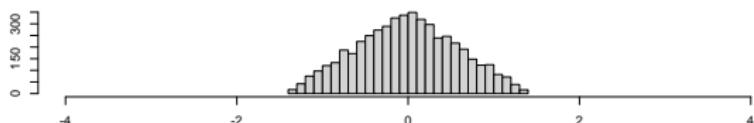
Note: $\text{Var}(\sqrt{n} \frac{\bar{X}_n - \mu}{\sigma}) = \frac{n}{\sigma^2} \text{Var}(\bar{X}_n) = 1$

THE CENTRAL LIMIT THEOREM

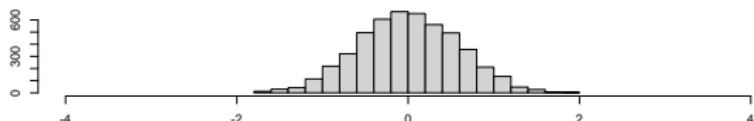
$n = 1$



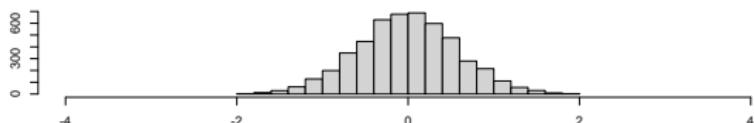
$n = 2$



$n = 5$



$n = 100$



MULTIVARIATE DISTRIBUTIONS

The notions of pmf, pdf, and cdf naturally extend to more than one dimension:

Let X and Y be random variables and $A \subseteq \mathcal{X}^2$:

joint pmf (\mathcal{X} discrete): $f_{X,Y} : \mathcal{X}^2 \rightarrow [0, 1]$

$$f_{X,Y}(x,y) = P((X,Y) \in A) = \sum_{(x,y) \in A} f_{X,Y}(x,y)$$

joint pdf ($\mathcal{X} = \mathbb{R}$): $f_{X,Y} : \mathbb{R}^2 \rightarrow [0, \infty)$

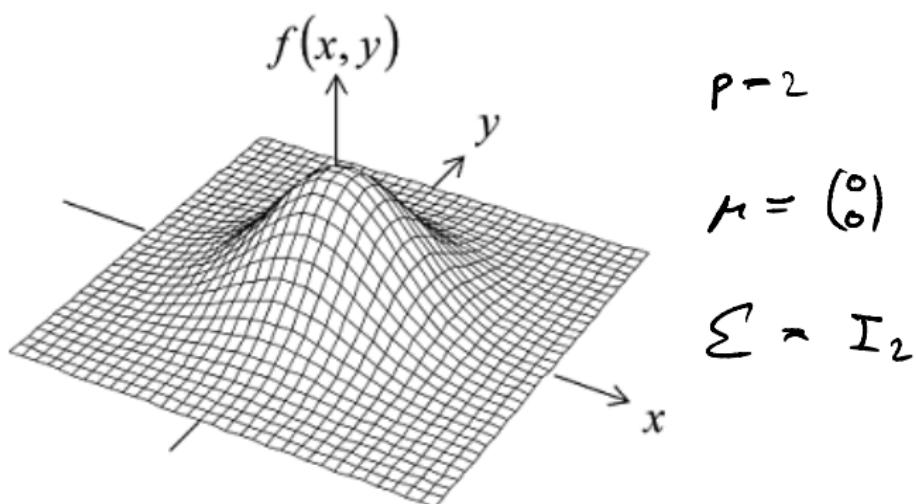
$$P((X,Y) \in A) = \int_A f_{X,Y}(x,y) dx dy$$

joint cdf: $F_{X,Y} : \mathbb{R}^2 \rightarrow [0, 1]$, $F_{X,Y}(x,y) = P(X \leq x, Y \leq y)$

MULTIVARIATE NORMAL DISTRIBUTION

- Parameters $\mu \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}^{p \times p}$ positive definite.

$$f_{\mu, \Sigma}(x) = (2\pi)^{-p/2} \det(\Sigma)^{-1/2} \exp\left(-\frac{1}{2}(x - \mu)' \Sigma^{-1} (x - \mu)\right), \quad x \in \mathbb{R}^p.$$



MULTIVARIATE DISTRIBUTIONS

Let X and Y be random variables with joint pdf/pmf $f_{X,Y}$.

- ▶ The **marginal pmf/pdf** of X is given by

$$f_X(x) = \begin{cases} \int_{\mathbb{R}} f_{X,Y}(x,y) dy, \\ \sum_{y \in \mathcal{X}} f_{X,Y}(x,y). \end{cases}$$

- ▶ X and Y are **independent** if, and only if,

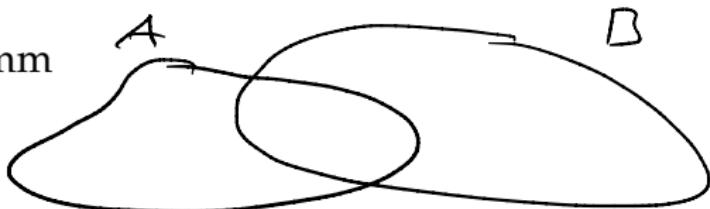
$$f_{X,Y}(x,y) = f_X(x)f_Y(y), \quad \forall x, y.$$

CONDITIONAL PROBABILITY

For events $A, B \subseteq \Omega$ with $P(B) > 0$, the conditional probability of A given B is defined as

$$P(A|B) := \frac{P(A \cap B)}{P(B)}$$

Intuition: Venn-Diagramm



Note: $A \mapsto P(A|B)$ is a probability assignment, i.e.,

- $P(\Omega | B) = 1$

- $P(\emptyset | B) = 0$

- $P(\bigcup_{i=1}^{\infty} A_i | B) = \sum_{i=1}^{\infty} P(A_i | B) \quad \text{if pairwise disjoint}$

CONDITIONAL DENSITY

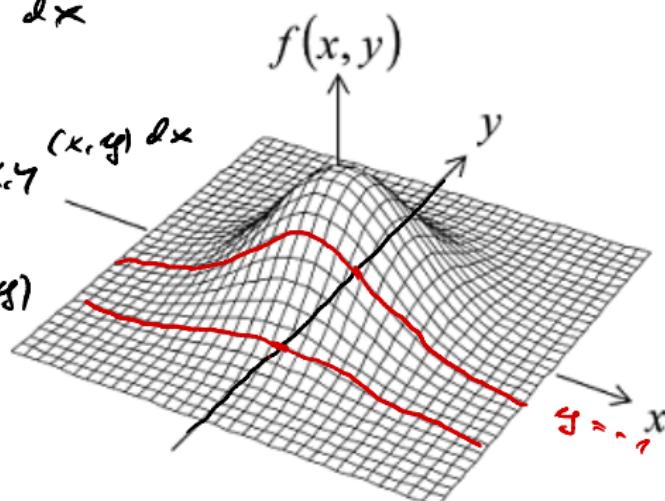
Let X and Y be random variables with joint pdf/pmf $f_{X,Y}$.

Then the function

$$f_{X|Y=y}(x) := \frac{f_{X,Y}(x,y)}{f_Y(y)},$$

is called the **conditional pdf/pmf** of X given $Y = y$.

$$\begin{aligned} & \int_{\mathbb{R}} f_{X|Y=y}(x) dx \\ &= \frac{1}{f_Y(y)} \int_{\mathbb{R}} f_{X,Y}(x,y) dx \\ &= \frac{1}{f_Y(y)} F_{X,Y}(y) \\ &= 1 \end{aligned}$$



BAYES THEOREM

$$\frac{P(A \cap B)}{P(A)}$$



$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$f_{X|Y=y}(x) = \frac{f_{Y|X=x}(y)f_X(x)}{f_Y(y)}$$

Formalism of statistical modeling

STATISTICAL MODELS

Definition 1.1

A statistical model is a triple $\mathcal{M} = (\mathcal{X}, \Theta, \{f_\theta : \theta \in \Theta\})$.

- ▶ \mathcal{X} is a set called the **sample space**.
- ▶ Θ is a set called the **parameter space**.
- ▶ $\{f_\theta : \theta \in \Theta\}$ is a family of pdf's or pmf's on \mathcal{X} indexed by Θ , that is, $f_\theta : \mathcal{X} \rightarrow [0, \infty)$ with either

$$\int_{\mathcal{X}} f_\theta(x) dx = 1, \quad \text{or} \quad \sum_{x \in \mathcal{X}} f_\theta(x) = 1.$$

EXAMPLE: NORMAL LOCATION MODEL

- ▶ $\mathcal{X} = \mathbb{R}^p$
- ▶ $\Theta = \mathbb{R}^p$
- ▶ $f_\theta(x) = (2\pi)^{-p/2} \exp\left(-\frac{1}{2}\|x - \theta\|_2^2\right), x \in \mathcal{X}, \theta \in \Theta.$

$$\mathcal{E} = \mathcal{I}_p$$

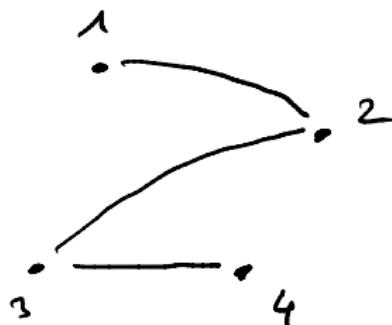
$$\rho = 1$$



EXAMPLE: ERDÖS-RÉNYI RANDOM GRAPH MODEL

- $\mathcal{X} = \{A \in \{0, 1\}^{n \times n} : A' = A, A_{ii} = 0 \text{ for all } i \in [n]\}$
- $\Theta = [0, 1]$
- For $A \in \mathcal{X}, \theta \in \Theta$,

$$f_\theta(A) = \prod_{\substack{i,j=1 \\ i < j}}^n \theta^{A_{ij}} (1 - \theta)^{1 - A_{ij}}.$$


$$A = \begin{array}{c|ccccc} & 1 & 2 & 3 & 4 \\ \hline 1 & 0 & 1 & 0 & 0 \\ 2 & 1 & 0 & 1 & 0 \\ 3 & 0 & 1 & 0 & 1 \\ 4 & 0 & 0 & 1 & 0 \end{array}$$

STATISTICAL MODELS

$$\mathcal{M} = (\mathcal{X}, \Theta, \{f_\theta : \theta \in \Theta\})$$

We assume the actual data have been generated from the distribution f_θ for some unknown $\theta \in \Theta$. To emphasize that the data are realizations of a **random process** and could have been also different, we describe them mathematically as **random variables**.

formally: X is a random variable taking values in \mathcal{X}

$$X \sim f_\theta \text{ for some } \theta \in \Theta$$

observed ↪ *unobserved*

THE IID MODEL

iid ... independent identically distributed

Often it makes sense to assume a product form:

$$\mathcal{X} = \mathcal{X}_0^n$$

e.g. $\mathcal{X}_0 = \mathbb{R}$

$$f_{\theta}(x) = \prod_{i=1}^n p_{\theta}(x_i), \quad \theta \in \Theta, x = (x_1, \dots, x_n)' \in \mathcal{X}_0^n,$$

p_{θ} a pdf or pmf on \mathcal{X}_0 n ... sample size.

For instance: (both our examples above)

- ▶ measuring the molecular weight of the same substance n times with a mass spectrometer
- ▶ taking an fMRI of n randomly selected individuals
- ▶ throwing n darts at a target

THE IID MODEL

$$\mathcal{X} = \mathcal{X}_0^n$$

$$f_\theta(x) = \prod_{i=1}^n p_\theta(x_i), \quad \theta \in \Theta, x = (x_1, \dots, x_n)' \in \mathcal{X}_0^n,$$

p_θ a pdf or pmf on \mathcal{X}_0 $n \dots$ sample size.

In this case:

$$X = (X_1, \dots, X_n)' \sim f_\theta, \quad \text{for some } \theta \in \Theta$$

in other words:

$$X_1, \dots, X_n \stackrel{iid}{\sim} p_\theta, \quad \text{for some } \theta \in \Theta$$

X_i takes values in \mathcal{X}_0

PARAMETERS OF INTEREST VS. NUISANCE PARAMETERS

Consider a statistical model

$$\mathcal{M} = (\mathcal{X}, \Theta, \{f_\theta : \theta \in \Theta\})$$

We are often only interested in some components of $\theta \in \Theta$.

E.g.:

$$f_\theta(x) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right)$$
$$\theta = (\mu, \sigma^2) \in \Theta := \mathbb{R} \times \mathbb{R}_+$$

But we want to estimate only $\mu \in \mathbb{R}$. Then we call $\sigma^2 > 0$ a
nuisance parameter.

PARAMETERS OF INTEREST VS. NUISANCE PARAMETERS

Consider a statistical model

$$\mathcal{M} = (\mathcal{X}, \Theta, \{f_\theta : \theta \in \Theta\})$$

We are often only interested in some components of $\theta \in \Theta$.

In general: Let $\psi : \Theta \rightarrow \Psi$ be a function. We may want to estimate and do inference on

$$\psi(\theta).$$

E.g.: $\psi : \mathbb{R} \times \mathbb{R}_+ \rightarrow \mathbb{R}$, $\psi(\mu, \sigma^2) = \mu$.

Estimators, tests and confidence intervals

STATISTIC AND ESTIMATOR

Recall: To learn something about the true unknown $\theta \in \Theta$ (or $\psi(\theta)$) we can only use the data $X \sim f_\theta$.

Definition 1.2

Let \mathcal{S} be any set. A (measurable) function $S : \mathcal{X} \rightarrow \mathcal{S}$ is called a **statistic**. If $\mathcal{S} = \Theta$ (or $\mathcal{S} = \Psi$) then S is called an **estimator** of θ (or of $\psi(\theta)$).

NOTATION

Let $\mathcal{M} = (\mathcal{X}, \Theta, \{f_\theta : \theta \in \Theta\})$ be a model.

$$\mathbb{E}(X) = ?$$

$$\text{Var} = ?$$

For an event $A \subseteq \mathcal{X}$ we write


$$\mathbb{P}_\theta(A) := \begin{cases} \int_A f_\theta(x) dx & \text{if } f_\theta \text{ is a pdf,} \\ \sum_{x \in A} f_\theta(x) & \text{if } f_\theta \text{ is a pmf.} \end{cases}$$

For a statistic $S : \mathcal{X} \rightarrow \mathbb{R}$ we write


$$\mathbb{E}_\theta[S] := \begin{cases} \int_{\mathcal{X}} S(x) f_\theta(x) dx & \text{if } f_\theta \text{ is a pdf,} \\ \sum_{x \in \mathcal{X}} S(x) f_\theta(x) & \text{if } f_\theta \text{ is a pmf.} \end{cases}$$

Note: $X \sim f_\theta \Rightarrow \mathbb{E}[S(X)] = \mathbb{E}_\theta[S]$.

Analogously for $\text{Var}_\theta[S] = \text{Var}[S(X)]$,
 $\text{Cov}_\theta(S_1, S_2) = \text{Cov}(S_1(X), S_2(X))$.

EXAMPLE

$X_1, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2), \quad \theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty) =: \Theta.$

that is, $f_\theta(x) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2} \sum_{i=1}^n \frac{(x_i - \mu)^2}{\sigma^2}\right)$, $\mathcal{X} = \mathbb{R}^n$.

$\hat{\mu}_n(x) := \frac{1}{n} \sum_{i=1}^n x_i, \quad x = (x_1, \dots, x_n)' \in \mathbb{R}^n$

$\hat{\sigma}_n^2(x) := \frac{1}{n-1} \sum_{i=1}^n (x_i - \underbrace{\hat{\mu}_n(x)}_{\text{mean}})^2$

$\mathbb{E}_\theta[\hat{\mu}_n] = \mathbb{E}[\hat{\mu}_n(X_1, \dots, X_n)] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] = \mu, \quad \forall \theta \in \Theta.$

“ $\hat{\mu}_n$ is an **unbiased estimator** of μ ”

$\mathbb{E}_\theta[\hat{\sigma}_n^2] = \dots = \sigma^2, \quad \forall \theta \in \Theta. \quad \text{HW}$

“ $\hat{\sigma}_n^2$ is an **unbiased estimator** of σ^2 ”

$\text{Cov}_\theta(\hat{\mu}_n, \hat{\sigma}_n^2) = \dots = 0, \quad \forall \theta \in \Theta$

MAXIMUM LIKELIHOOD ESTIMATION

Consider a statistical model

$$\mathcal{M} = (\mathcal{X}, \Theta, \{f_\theta : \theta \in \Theta\})$$

Then $\theta \mapsto L(\theta|x) := f_\theta(x)$ is called the **likelihood function** and

$$\hat{\theta}(x) := \operatorname{argmax}_{\theta \in \Theta} L(\theta|x), \quad x \in \mathcal{X},$$

is called the **maximum likelihood estimator**.

If \mathcal{M} is an **iid model** with sample size n , we often maximize the **log-likelihood**

$$\Theta \mapsto \ell_n(\theta|x) := \log L(\theta|x) = \log \prod_{i=1}^n p_\theta(x_i) = \sum_{i=1}^n \log p_\theta(x_i).$$

This is equivalent, because

$$L(\theta|x) \leq L(\hat{\theta}(x)|x) \quad \forall \theta \in \Theta \iff \log L(\theta|x) \leq \log L(\hat{\theta}(x)|x) \quad \forall \theta \in \Theta.$$

EXAMPLE: ERDÖS-RÉNYI MODEL

$$N = \frac{n(n-1)}{2}$$

$$f_{\theta}(A) = \prod_{\substack{i,j=1 \\ i < j}}^n \theta^{A_{ij}} (1-\theta)^{1-A_{ij}} = \Theta^{N \cdot \bar{A}_n} \cdot (1-\Theta)^{N - N \cdot \bar{A}_n}$$

$$\hat{\Theta} = \bar{A}_n = \frac{1}{N} \sum_{\substack{i,j=1 \\ i < j}} A_{ij}$$

$$L(\theta | A) = f_{\theta}(A)$$

$$\ell_N(\theta | A) = \log L(\theta | A) = N \cdot \bar{A}_n \log \theta +$$

$$\frac{\partial}{\partial \theta} \ell_N(\theta | A) = \frac{N \cdot \bar{A}_n}{\theta} - \frac{N(1-\bar{A}_n)}{1-\theta} \stackrel{?}{=} 0$$

$$\Rightarrow \hat{\Theta} = \bar{A}_n$$

$$\frac{\partial^2}{\partial \theta^2} \ell_N(\theta | A) \stackrel{?}{\leq} 0$$

\Leftrightarrow concave

HYPOTHESIS TESTS

null hypothesis
alternative
↓ ↓
hyp-

Let $\mathcal{M} = (\mathcal{X}, \Theta, \{f_\theta : \theta \in \Theta\})$ be a model and $\Theta = \Theta_0 \cup \Theta_1$ where $\Theta_0 \cap \Theta_1 = \emptyset$.

Based on data $X \sim f_\theta$, we want to decide whether $H_0 : \theta \in \Theta_0$ or $H_1 : \theta \in \Theta_1$.

\Rightarrow test function $\varphi : \mathcal{X} \rightarrow \{0, 1\}$

Can make a mistake!

probability of **type one error**: $P_\theta(\varphi = 1), \theta \in \Theta_0$

probability of **type two error**: $P_\theta(\varphi = 0), \theta \in \Theta_1$

SIGNIFICANCE TESTS

Let $\mathcal{M} = (\mathcal{X}, \Theta, \{f_\theta : \theta \in \Theta\})$ be a model and $\Theta = \Theta_0 \cup \Theta_1$ where $\Theta_0 \cap \Theta_1 = \emptyset$.

Common approach: Fix a significance level $\alpha \in (0, 1)$ to control the probability of a type one error.

Definition 1.3

$\varphi_\alpha : \mathcal{X} \rightarrow \{0, 1\}$ is a level α test if

$$P_\theta(\varphi_\alpha = 1) \leq \alpha, \quad \forall \theta \in \Theta_0.$$

trivial level α test:

$$\varphi_0(x) = 0, \forall x \in \mathcal{X} \Rightarrow P_\theta(\varphi_0 = 1) = 0 \leq \alpha, \forall \theta \in \Theta_0.$$

Power: $P_{\Theta_1}(\varphi_0 = 1) = P_{\Theta_1}(\emptyset) = 0$

CRITICAL VALUE

Most of the time, a level α test will be of the form

$$\varphi_\alpha(x) = \begin{cases} 1, & \text{if } S(x) \geq c_\alpha, \\ 0, & \text{else,} \end{cases} \quad (1)$$

for some **test statistic** $S : \mathcal{X} \rightarrow \mathbb{R}$ and some **critical value** $c_\alpha \in \mathbb{R}$. (practical + theoretical reasons)

Ideally, we would like to take c_α such that

$$\overbrace{\sup_{\theta \in \Theta_0} P_\theta(\varphi_\alpha = 1)}^{\leq \alpha} = \sup_{\theta \in \Theta_0} P_\theta(S \geq c_\alpha) \stackrel{!}{=} \alpha.$$

$$c < c_\alpha : P_\Theta(S \geq c) > P_\Theta(S \geq c_\alpha) = \alpha$$

$$c > c_\alpha : P_\Theta(S \geq c) < P_\Theta(S \geq c_\alpha) = \alpha \quad \Theta \in \Theta_1$$

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812

NOVEMBER 5, 2020

VOL. 383 NO. 19

Remdesivir for the Treatment of Covid-19 — Final Report

J.H. Beigel, K.M. Tomashek, L.E. Dodd, A.K. Mehta, B.S. Zingman, A.C. Kalil, E. Hohmann, H.Y. Chu, A. Luetkemeyer, S. Kline, D. Lopez de Castilla, R.W. Finberg, K. Dierberg, V. Tapson, L. Hsieh, T.F. Patterson, R. Paredes, D.A. Sweeney, W.R. Short, G. Touloumi, D.C. Lye, N. Ohmagari, M. Oh, G.M. Ruiz-Palacios, T. Benfield, G. Fätkenheuer, M.G. Kortepeter, R.L. Atmar, C.B. Creech, J. Lundgren, A.G. Babiker, S. Pett, J.D. Neaton, T.H. Burgess, T. Bonnett, M. Green, M. Makowski, A. Osinusi, S. Nayak, and H.C. Lane, for the ACTT-1 Study Group Members*

ABSTRACT

BACKGROUND

Although several therapeutic agents have been evaluated for the treatment of coronavirus disease 2019 (Covid-19), no antiviral agents have yet been shown to be efficacious.

METHODS

We conducted a double-blind, randomized, placebo-controlled trial of intravenous remdesivir in adults who were hospitalized with Covid-19 and had evidence of lower respiratory tract infection. Patients were randomly assigned to receive either remdesivir (200 mg loading dose on day 1, followed by 100 mg daily for up to 9 additional days) or placebo for up to 10 days. The primary outcome was the time to recovery, defined by either discharge from the hospital or hospitalization for infection-control purposes only.

RESULTS

A total of 1062 patients underwent randomization (with 541 assigned to remdesivir and 521 to placebo). Those who received remdesivir had a median recovery time of 10 days (95% confidence interval [CI], 9 to 11), as compared with 15 days (95% CI, 13 to 18) among those who received placebo (rate ratio for recovery, 1.29; 95% CI, 1.12 to 1.49; $P < 0.001$, by a log-rank test). In an analysis that used a proportion-al-odds model with an eight-category ordinal scale, the patients who received remdesivir were found to be more likely than those who received placebo to have clinical improvement at day 15 (odds ratio, 1.5; 95% CI, 1.2 to 1.9, after adjustment for actual disease severity). The Kaplan-Meier estimates of mortality were 6.7% with remdesivir and 11.9% with placebo by day 15 and 11.4% with remdesivir and 15.2% with placebo by day 29 (hazard ratio, 0.73; 95% CI, 0.52 to 1.03). Serious adverse events occurred in 31.3% of the patients assigned to remdesivir and 33.5% of those assigned to placebo.

The authors' full names, academic degrees, and affiliations are listed in the Appendix. Address reprint requests to Dr. Beigel at the National Institute of Allergy and Infectious Diseases, National Institutes of Health, 5601 Fishers Ln., Rm. 7E60, MSC 9826, Rockville, MD 20892-9826, or at jbeigel@niaid.nih.gov.

*A complete list of members of the ACTT-1 Study Group is provided in the Supplementary Appendix, available at NEJM.org.

A preliminary version of this article was published on May 22, 2020, at NEJM.org. This article was published on October 8, 2020, and updated on October 9, 2020, at NEJM.org.

N Engl J Med 2020;383:1813-26.
DOI: 10.1056/NEJMoa2007764
Copyright © 2020 Massachusetts Medical Society.

EXAMPLE: THE TWO SAMPLE z -TEST

Consider the (simple) model:

- ▶ Treatment group: $X_1, \dots, X_{n_1} \stackrel{iid}{\sim} N(\mu_{treat}, 1), \mu_{treat} \geq 0$
- ▶ Control group: $Y_1, \dots, Y_{n_2} \stackrel{iid}{\sim} N(\mu_{cont}, 1), \mu_{cont} \geq 0$
- ▶ Treatment and control groups are independent
- ▶ Equal and known variances (for simplicity!). 

X_i, Y_j are observed times to recovery.

$$H_0 : \mu_{cont} \leq \mu_{treat} \quad \text{vs.} \quad H_1 : \mu_{cont} > \mu_{treat}$$

$\mathcal{X}, \Theta, f_\theta, \Theta_0, \Theta_1$?

EXAMPLE: THE TWO SAMPLE z -TEST

- ▶ Treatment group: $X_1, \dots, X_{n_1} \stackrel{iid}{\sim} N(\mu_t, 1), \mu_t \geq 0$
- ▶ Control group: $Y_1, \dots, Y_{n_2} \stackrel{iid}{\sim} N(\mu_c, 1), \mu_c \geq 0$

$$H_0 : \mu_c \leq \mu_t \quad \text{vs.} \quad H_1 : \mu_c > \mu_t$$

Test statistic:

$$S = \frac{\bar{Y}_{n_2} - \bar{X}_{n_1}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

S follows a normal distribution with mean

$$\Delta_\mu := \frac{\mu_c - \mu_t}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

and unit variance:

$$S \sim N(\Delta_\mu, 1).$$

EXAMPLE: THE TWO SAMPLE z -TEST

$$H_0 : \mu_c \leq \mu_t \quad \text{vs.} \quad H_1 : \mu_c > \mu_t$$

or

$$H_0 : \Delta_\mu \leq 0 \quad \text{vs.} \quad H_1 : \Delta_\mu > 0$$

$$S \sim N(\Delta_\mu, 1).$$

Test φ_α : Reject H_0 if $S \geq q_{1-\alpha}^{(N)}$ $1 - \alpha$ quantile

Type-I error probability: $H_0 : \theta \in \Theta_0 \quad (\iff \Delta_\mu \leq 0)$

$$\sup_{\theta \in \Theta_0} P_\theta(\varphi_\alpha = 1) = \alpha?$$

EXAMPLE: THE TWO SAMPLE z -TEST

$$H_0 : \mu_c \leq \mu_t \quad \text{vs.} \quad H_1 : \mu_c > \mu_t$$

or

$$H_0 : \Delta_\mu \leq 0 \quad \text{vs.} \quad H_1 : \Delta_\mu > 0$$

$$S \sim N(\Delta_\mu, 1).$$

Test φ_α : Reject H_0 if $S \geq q_{1-\alpha}^{(N)}$ $1 - \alpha$ quantile

Power: $\theta \in \Theta_1 \iff \Delta_\mu > 0$

$$\begin{aligned} P_\theta(S \geq q_{1-\alpha}^{(N)}) &= P_\theta(S - \Delta_\mu \geq \underbrace{q_{1-\alpha}^{(N)} - \Delta_\mu}_{< 0}) \\ &> P_\theta(S - \Delta_\mu \geq q_{1-\alpha}^{(N)}) = \alpha. \end{aligned}$$

p-VALUE

The *p*-value is the smallest significance level at which the test still rejects H_0 .

Definition 1.4

If, for every $\alpha \in (0, 1)$, φ_α is a level α test for $H_0 : \theta \in \Theta_0$, then the associated ***p*-value** is defined to be

$$p(x) := \inf\{\alpha \in (0, 1) : \varphi_\alpha(x) = 1\}, \quad x \in \mathcal{X}.$$

p-VALUE

Theorem 1.5

If, for every $\alpha \in (0, 1)$, φ_α is a level α test of $H_0 : \theta \in \Theta_0$ and $\varphi_\alpha(x) \leq \varphi_{\alpha'}(x)$ for all $x \in \mathcal{X}$ and all $\alpha \leq \alpha'$, then

$$\sup_{\theta \in \Theta_0} P_\theta(p \leq u) \leq u, \quad \forall u \in (0, 1).$$

Thus, the *p*-value can be used to perform a significance test at arbitrary level $u \in (0, 1)$:

$$\tilde{\varphi}_u(x) := \begin{cases} 1, & \text{if } p(x) \leq u, \\ 0, & \text{else.} \end{cases}$$

p-VALUE

Theorem 1.6

If the level α test φ_α of $H_0 : \theta \in \Theta_0$ is of the form (1) with $\sup_{\theta \in \Theta_0} P_\theta(S \geq c_\alpha) = \alpha$ for all $\alpha \in (0, 1)$ and $\alpha \mapsto c_\alpha$ is continuous and strictly decreasing, then

$$p(x) = \sup_{\theta \in \Theta_0} P_\theta(S \geq S(x)), \quad x \in \mathcal{X}.$$

In particular, if $\Theta_0 = \{\theta_0\}$, then

$$p(x) = P_{\theta_0}(S \geq S(x)), \quad x \in \mathcal{X}.$$

Thus, the *p*-value is the probability (under the null hypothesis) to observe the same or a more extreme value of the test statistic S than we actually did.

Note: The *p*-value quantifies the evidence against the null hypothesis.

EXAMPLE: THE TWO SAMPLE z -TEST

$$H_0 : \mu_c \leq \mu_t \quad \text{vs.} \quad H_1 : \mu_c > \mu_t$$

or

$$H_0 : \Delta_\mu \leq 0 \quad \text{vs.} \quad H_1 : \Delta_\mu > 0$$

$$S \sim N(\Delta_\mu, 1).$$

Test φ_α : Reject H_0 if $S \geq q_{1-\alpha}^{(N)}$ $1 - \alpha$ quantile

$$\sup_{\theta \in \Theta} P_\theta(S \geq s)$$

p-value: (cf. Theorem 1.6)
 $\neq \Delta_\mu$

$$p = \sup_{\theta \in \Theta_0} P_\theta(S \geq s) |_{s=S} = \sup_{\Delta_\mu \leq 0} P(N(0, 1) \geq s - \Delta_\mu) |_{s=S}$$

$$= P(N(0, 1) \geq s) |_{s=S} = 1 - \Phi(s)$$

$$\rho(x) = 1 - \Phi(S(x))$$

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812

NOVEMBER 5, 2020

VOL. 383 NO. 19

Remdesivir for the Treatment of Covid-19 — Final Report

J.H. Beigel, K.M. Tomashek, L.E. Dodd, A.K. Mehta, B.S. Zingman, A.C. Kalil, E. Hohmann, H.Y. Chu, A. Luetkemeyer, S. Kline, D. Lopez de Castilla, R.W. Finberg, K. Dierberg, V. Tapson, L. Hsieh, T.F. Patterson, R. Paredes, D.A. Sweeney, W.R. Short, G. Touloumi, D.C. Lye, N. Ohmagari, M. Oh, G.M. Ruiz-Palacios, T. Benfield, G. Fätkenheuer, M.G. Kortepeter, R.L. Atmar, C.B. Creech, J. Lundgren, A.G. Babiker, S. Pett, J.D. Neaton, T.H. Burgess, T. Bonnett, M. Green, M. Makowski, A. Osinusi, S. Nayak, and H.C. Lane, for the ACTT-1 Study Group Members*

ABSTRACT

BACKGROUND

Although several therapeutic agents have been evaluated for the treatment of coronavirus disease 2019 (Covid-19), no antiviral agents have yet been shown to be efficacious.

METHODS

We conducted a double-blind, randomized, placebo-controlled trial of intravenous remdesivir in adults who were hospitalized with Covid-19 and had evidence of lower respiratory tract infection. Patients were randomly assigned to receive either remdesivir (200 mg loading dose on day 1, followed by 100 mg daily for up to 9 additional days) or placebo for up to 10 days. The primary outcome was the time to recovery, defined by either discharge from the hospital or hospitalization for infection-control purposes only.

RESULTS

A total of 1062 patients underwent randomization (with 541 assigned to remdesivir and 521 to placebo). Those who received remdesivir had a median recovery time of 10 days (95% confidence interval [CI], 9 to 11), as compared with 15 days (95% CI, 13 to 18) among those who received placebo (rate ratio for recovery, 1.29; 95% CI, 1.12 to 1.49; $P<0.001$, by a log-rank test). In an analysis that used a proportion-al-odds model with an eight-category ordinal scale, the patients who received remdesivir were found to be more likely than those who received placebo to have clinical improvement at day 15 (odds ratio, 1.5; 95% CI, 1.2 to 1.9, after adjustment for actual disease severity). The Kaplan-Meier estimates of mortality were 6.7% with remdesivir and 11.9% with placebo by day 15 and 11.4% with remdesivir and 15.2% with placebo by day 29 (hazard ratio, 0.73; 95% CI, 0.52 to 1.03). Serious adverse events occurred in 31.4% of the patients assigned to remdesivir and 33.3% of those assigned to placebo.

The authors' full names, academic degrees, and affiliations are listed in the Appendix. Address reprint requests to Dr. Beigel at the National Institute of Allergy and Infectious Diseases, National Institutes of Health, 5601 Fishers Ln., Rm. 7E60, MSC 9826, Rockville, MD 20892-9826, or at jbeigel@niaid.nih.gov.

*A complete list of members of the ACTT-1 Study Group is provided in the Supplementary Appendix, available at NEJM.org.

A preliminary version of this article was published on May 22, 2020, at NEJM.org. This article was published on October 8, 2020, and updated on October 9, 2020, at NEJM.org.

N Engl J Med 2020;383:1813-26.
DOI: 10.1056/NEJMoa2007764
Copyright © 2020 Massachusetts Medical Society.

CONFIDENCE INTERVALS (CI)

Definition 1.7

Let $\mathcal{M} = (\mathcal{X}, \Theta, \{f_\theta : \theta \in \Theta\})$ be a model and $\alpha \in (0, 1)$ be an **error probability**. A data dependent set $CS_\alpha(x) \subseteq \Theta$, $x \in \mathcal{X}$, is called a **level $1 - \alpha$ confidence set** for θ if

$$P_\theta(\theta \in CS_\alpha) \geq 1 - \alpha, \quad \forall \theta \in \Theta.$$

If $\Theta \subseteq \mathbb{R}$ and $CS_\alpha(x) = [l_\alpha(x), u_\alpha(x)]$ is an interval, then $CI_\alpha := CS_\alpha$ is called a **confidence interval** for θ .

trivial level $1 - \alpha$ CI: ($\Theta \subseteq \mathbb{R}$)

$$CI_0(x) := \mathbb{R} \quad \Rightarrow \quad P_\theta(\theta \in CI_0) = P_\theta(\mathcal{X}) = 1 \geq 1 - \alpha, \quad \forall \theta \in \Theta.$$

Length of the CI: $length(CI_0(x)) = \infty$

DUALITY OF TESTS AND CONFIDENCE SETS

Theorem 1.8

Let $\mathcal{M} = (\mathcal{X}, \Theta, \{f_\theta : \theta \in \Theta\})$ be a model and $\alpha \in (0, 1)$ be an error probability.

- a) If, for every $\theta_0 \in \Theta$, $\varphi_{\alpha, \theta_0} : \mathcal{X} \rightarrow \{0, 1\}$ is a level α test for $H_0 : \theta = \theta_0$ (i.e., $\Theta_0 = \{\theta_0\}$), then

$$CS_\alpha(x) := \{\theta \in \Theta : \varphi_{\alpha, \theta}(x) = 0\}$$

is a level $1 - \alpha$ confidence set for θ .

- b) If $CS_\alpha(x) \subseteq \Theta$ is a level $1 - \alpha$ confidence set for θ , then

$$\varphi_\alpha(x) := \begin{cases} 1, & \text{if } \Theta_0 \cap CS_\alpha(x) = \emptyset, \\ 0, & \text{else,} \end{cases}$$

is a level α test for $H_0 : \theta \in \Theta_0$.

THE TWO SAMPLE z -TEST

- ▶ Treatment group: $X_1, \dots, X_{n_1} \stackrel{iid}{\sim} N(\mu_t, 1), \mu_t \geq 0$
- ▶ Control group: $Y_1, \dots, Y_{n_2} \stackrel{iid}{\sim} N(\mu_c, 1), \mu_c \geq 0$

$$H_0 : \mu_c \leq \mu_t \quad \text{vs.} \quad H_1 : \mu_c > \mu_t$$

Test statistic:

$$S = \frac{\bar{Y}_{n_2} - \bar{X}_{n_1}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(\Delta_\mu, 1), \quad \Delta_\mu := \frac{\mu_c - \mu_t}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \in \mathbb{R}.$$

$$H_0 : \Delta_\mu \leq 0 \quad \text{vs.} \quad H_1 : \Delta_\mu > 0$$

Δ_μ ... effect size

Want a confidence interval for Δ_μ .

CI FOR THE MEAN OF A NORMAL DISTRIBUTION

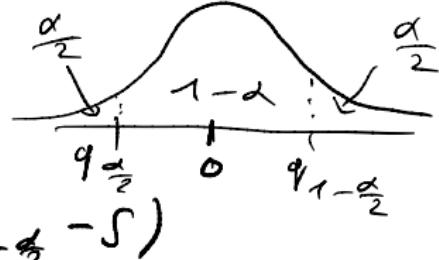
Model

$$S \sim N(\Delta, 1), \quad \Delta \in \mathbb{R}.$$

Want a confidence interval for the effect size Δ .

$$S - \Delta \sim N(0, 1)$$

$$1 - \alpha = P_{\Delta} \left(q_{\alpha/2}^{(N)} \leq S - \Delta \leq q_{1-\alpha/2}^{(N)} \right)$$



$$= P_{\Delta} \left(q_{\alpha/2}^{(N)} - S \leq -\Delta \leq q_{1-\alpha/2}^{(N)} - S \right)$$

$$= P_{\Delta} \left(S - q_{\alpha/2}^{(N)} \geq \Delta \geq S - q_{1-\alpha/2}^{(N)} \right)$$

$$= -q_{1-\alpha/2}^{(N)}$$

$$= P_{\Delta} (\Delta \in CI_{\alpha})$$

$$CI_{\alpha} = [S - q_{1-\alpha/2}^{(N)}, S + q_{1-\alpha/2}^{(N)}]$$

CI FOR THE MEAN OF A NORMAL DISTRIBUTION

$$S \sim N(\Delta, 1), \quad \Delta \in \mathbb{R}.$$

$$CI_{\alpha}(S) = [S - q_{1-\alpha/2}^{(N)}, S + q_{1-\alpha/2}^{(N)}]$$

Turn this into a test for

$$H_0 : \Delta \leq 0 \quad \text{vs.} \quad H_1 : \Delta > 0,$$

using Theorem 1.8:

$$\varphi_{\alpha}(S) = \begin{cases} 1, & \text{if } \Theta_0 \cap CI_{\alpha}(S) = \emptyset \\ 0, & \text{else,} \end{cases}$$

CI FOR THE MEAN OF A NORMAL DISTRIBUTION

$$S \sim N(\Delta, 1), \quad \Delta \in \mathbb{R}.$$

$$CI_{\alpha}(S) = [S - q_{1-\alpha}^{(N)}, \infty)$$

Turn this into a test for

$$H_0 : \Delta \leq 0 \quad \text{vs.} \quad H_1 : \Delta > 0,$$

using Theorem 1.8:

$$\varphi_{\alpha}(S) = \begin{cases} 1, & \text{if } \Theta_0 \cap CI_{\alpha}(S) = \emptyset \\ 0, & \text{else,} \end{cases}$$

Statistics for Data Science, Winter 2023

Chapter 2:

Montecarlo and Bayesian Methods



Analyze statistical properties of inference methods (e.g., precision of estimates, coverage of CIs, power of tests) ...

- ▶ mathematically
- ▶ numerically

Use numerical/simulation methods for data analysis, e.g., ...

- ▶ compute quantiles of analytically intractable distributions
- ▶ MCMC for Bayesian data analysis
- ▶ Bootstrap and resampling



- ▶ Montecarlo methods
 - ▶ Numerical integration
 - ▶ Random number generation
- ▶ Introduction to Bayesian analysis
- ▶ Refined MC methods

Montecarlo Methods

MONTECARLO INTEGRATION

Suppose we want to (approximately) compute the mean of a real random variable X with pdf or pmf f .

If we can generate a sample $X_1, \dots, X_N \stackrel{iid}{\sim} f$, then by the LLN,

$$\frac{1}{N} \sum_{j=1}^N X_j \xrightarrow[N \rightarrow \infty]{i.p.} \mathbb{E}[X].$$

If X takes values in \mathcal{X} and $S : \mathcal{X} \rightarrow \mathbb{R}$, then $S(X_1), \dots, S(X_N)$ are also iid and

$$\frac{1}{N} \sum_{j=1}^N S(X_j) \xrightarrow[B \rightarrow \infty]{i.p.} \mathbb{E}[S(X)] = \mathbb{E}_f[S],$$

provided that the expectation $\mathbb{E}[S(X)]$ is finite.

Idea: For large N , generate (pseudo) random numbers $X_1, \dots, X_N \stackrel{iid}{\sim} f$ and compute $\frac{1}{N} \sum_{j=1}^N S(X_j)$ to approximate $\mathbb{E}[S(X)]$.

To compute the probability $P(X \in A)$ we can use

$$\frac{1}{N} \sum_{j=1}^N \mathbb{1}_A(X_j) \xrightarrow[N \rightarrow \infty]{i.p.} \underbrace{\mathbb{E}[\mathbb{1}_A(X)]}_{=\int 1, \begin{matrix} X \in A \\ 0, \text{ else} \end{matrix}} = P(X \in A).$$

How to approximately evaluate the variance of a real $X \sim f$?

$$\text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E} X)^2$$

$$\uparrow \qquad \qquad \uparrow$$

$$\frac{1}{N} \sum_{i=1}^N x_i^2 - \left(\frac{1}{N} \sum_{i=1}^N x_i \right)^2$$



The trick also works for quantiles of f , provided that they are unique.

- ▶ Suppose the corresponding cdf $F(x) := \int_{-\infty}^x f(y)dy$ is invertible.
- ▶ $X_1, \dots, X_N \stackrel{iid}{\sim} f$
- ▶ $\hat{F}_N^\dagger(\alpha) \xrightarrow[N \rightarrow \infty]{i.p.} F^{-1}(\alpha) = q_\alpha$ HW

R Example:

Simulate the quantile function of the χ^2 -distribution.



The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812

NOVEMBER 5, 2020

VOL. 383 NO. 19

Remdesivir for the Treatment of Covid-19 — Final Report

J.H. Beigel, K.M. Tomashek, L.E. Dodd, A.K. Mehta, B.S. Zingman, A.C. Kalil, E. Hohmann, H.Y. Chu, A. Luetkemeyer, S. Kline, D. Lopez de Castilla, R.W. Finberg, K. Dierberg, V. Tapson, L. Hsieh, T.F. Patterson, R. Paredes, D.A. Sweeney, W.R. Short, G. Touloumi, D.C. Lye, N. Ohmagari, M. Oh, G.M. Ruiz-Palacios, T. Benfield, G. Fätkenheuer, M.G. Kortepeter, R.L. Atmar, C.B. Creech, J. Lundgren, A.G. Babiker, S. Pett, J.D. Neaton, T.H. Burgess, T. Bonnett, M. Green, M. Makowski, A. Osinusi, S. Nayak, and H.C. Lane,
for the ACTT-1 Study Group Members*

ABSTRACT

BACKGROUND

Although several therapeutic agents have been evaluated for the treatment of coronavirus disease 2019 (Covid-19), no antiviral agents have yet been shown to be efficacious.

METHODS

We conducted a double-blind, randomized, placebo-controlled trial of intravenous remdesivir in adults who were hospitalized with Covid-19 and had evidence of lower respiratory tract infection. Patients were randomly assigned to receive either remdesivir (200 mg loading dose on day 1, followed by 100 mg daily for up to 9 additional days) or placebo for up to 10 days. The primary outcome was the time to recovery, defined by either discharge from the hospital or hospitalization for infection-control purposes only.

RESULTS

A total of 1062 patients underwent randomization (with 541 assigned to remdesivir and 521 to placebo). Those who received remdesivir had a median recovery time of 10 days (95% confidence interval [CI], 9 to 11), as compared with 15 days (95% CI, 13 to 18) among those who received placebo (rate ratio for recovery, 1.29; 95% CI, 1.12 to 1.49; $P<0.001$, by a log-rank test). In an analysis that used a proportional-odds model with an eight-category ordinal scale, the patients who received remdesivir were found to be more likely than those who received placebo to have clinical improvement at day 15 (odds ratio, 1.5; 95% CI, 1.2 to 1.9, after adjustment for actual disease severity). The Kaplan-Meier estimates of mortality were 6.7% with remdesivir and 11.9% with placebo by day 15 and 11.4% with remdesivir and 15.2% with placebo by day 29 (hazard ratio, 0.73; 95% CI, 0.52 to 1.03). Serious adverse events occurred in 11.2% of the patients assigned to remdesivir and 12.5% of those assigned to placebo.

The authors' full names, academic degrees, and affiliations are listed in the Appendix. Address reprint requests to Dr. Beigel at the National Institute of Allergy and Infectious Diseases, National Institutes of Health, 5601 Fishers Ln, Rm. 7E60, MSC 9826, Rockville, MD 20892-9826, or at jbeigel@niaid.nih.gov.

*A complete list of members of the ACTT-1 Study Group is provided in the Supplementary Appendix, available at NEJM.org.

A preliminary version of this article was published on May 22, 2020, at NEJM.org. This article was published on October 8, 2020, and updated on October 9, 2020, at NEJM.org.

N Engl J Med 2020;383:1813-26.
DOI: [10.1056/NEJMoa2007764](https://doi.org/10.1056/NEJMoa2007764)
Copyright © 2020 Massachusetts Medical Society.

Consider the model:

- ▶ Treatment group: $X_1, \dots, X_{n_1} \stackrel{iid}{\sim} N(\mu_t, \sigma_t^2)$, $\mu_t \geq 0$
- ▶ Control group: $Y_1, \dots, Y_{n_2} \stackrel{iid}{\sim} N(\mu_c, \sigma_c^2)$, $\mu_c \geq 0$
- ▶ Treatment and control groups are independent

X_i, Y_j are observed times to recovery.

$$H_0 : 0 \leq \mu_c \leq \mu_t, \sigma_t^2 > 0, \sigma_c^2 > 0 \quad \text{vs.}$$

$$H_1 : \mu_c > \mu_t \geq 0, \sigma_t^2 > 0, \sigma_c^2 > 0$$

EXAMPLE: BEHRENS-FISHER PROBLEM



- ▶ Treatment group: $X_1, \dots, X_{n_1} \stackrel{iid}{\sim} N(\mu_t, \sigma_t^2), \mu_t \geq 0$
- ▶ Control group: $Y_1, \dots, Y_{n_2} \stackrel{iid}{\sim} N(\mu_c, \sigma_c^2), \mu_c \geq 0$

$$H_0 : 0 \leq \mu_c \leq \mu_t, \sigma_t^2 > 0, \sigma_c^2 > 0 \quad \text{vs.}$$

$$H_1 : \mu_c > \mu_t \geq 0, \sigma_t^2 > 0, \sigma_c^2 > 0$$

$$\bar{Y}_{n_2} - \bar{X}_{n_1} \sim N \left(\mu_c - \mu_t, \frac{\sigma_t^2}{n_1} + \frac{\sigma_c^2}{n_2} \right)$$

Test statistic:

$$Z = \frac{\bar{Y}_{n_2} - \bar{X}_{n_1}}{\sqrt{\frac{\sigma_t^2}{n_1} + \frac{\sigma_c^2}{n_2}}} \quad ?! \quad S = \frac{\bar{Y}_{n_2} - \bar{X}_{n_1}}{\sqrt{\frac{\hat{\sigma}_t^2}{n_1} + \frac{\hat{\sigma}_c^2}{n_2}}}.$$

$$\hat{\sigma}_t^2 = \frac{1}{n_1} \sum_{i=1}^{n_1} \underbrace{(X_i - \bar{X}_{n_1})^2}_{\sim \sigma_t^2} \chi^2_{n_1-1}, \quad \hat{\sigma}_c^2 = \frac{1}{n_2} \sum_{j=1}^{n_2} (Y_j - \bar{Y}_{n_2})^2$$

EXAMPLE: BEHRENS-FISHER PROBLEM



universität
wien

- ▶ Treatment group: $X_1, \dots, X_{n_t} \stackrel{iid}{\sim} N(\mu_t, \sigma_t^2)$, $\mu_t \geq 0$
- ▶ Control group: $Y_1, \dots, Y_{n_c} \stackrel{iid}{\sim} N(\mu_c, \sigma_c^2)$, $\mu_c \geq 0$

$$H_0 : 0 \leq \mu_c \leq \mu_t, \sigma_t^2 > 0, \sigma_c^2 > 0 \quad \text{vs.}$$

$$H_1 : \mu_c > \mu_t \geq 0, \sigma_t^2 > 0, \sigma_c^2 > 0$$

$$\sigma_t = \sigma_c = 1$$

Test statistic:

$$S = \frac{\bar{Y}_{n_c} - \bar{X}_{n_t}}{\sqrt{\frac{\hat{\sigma}_t^2}{n_t} + \frac{\hat{\sigma}_c^2}{n_c}}}.$$

$$S \sim N(\Delta_{\mu}, 1)$$

$$\hat{\sigma}_t = \hat{\sigma}_c = 1$$

No closed form pdf of S exists! Depends on $\mu_t, \mu_c, \sigma_t^2, \sigma_c^2$!

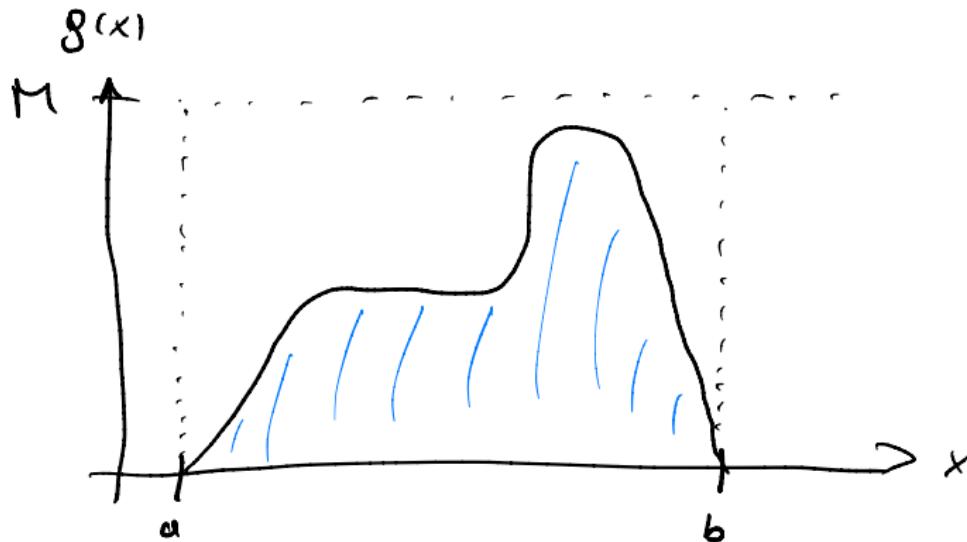
Reject H_0 if $S \geq q_{1-\alpha}$

Simulate $1 - \alpha$ quantile! But for which $\mu_t, \mu_c, \sigma_t^2, \sigma_c^2$?

Note: $\mathbb{E}[S(X)] = \int S(x) f(x) dx.$

How about general (definite) integrals?

$$\int_a^b g(x) dx, \quad 0 \leq g(x) \leq M$$



How about general (definite) integrals?

$$I := \int_a^b g(x) dx, \quad 0 \leq g(x) \leq M$$

- ▶ Generate $(X_i, Y_i) \stackrel{iid}{\sim} \text{Unif}([a, b] \times [0, M]), i = 1, \dots, N$.
- ▶ Compute $\hat{I}_N := \frac{M(b-a)}{N} \sum_{i=1}^N \mathbb{1}_{\{Y_i \leq g(X_i)\}} \xrightarrow[N \rightarrow \infty]{i.p.} I$

$$\begin{aligned} \mathbb{E}[\mathbb{1}\{\gamma \leq g(x)\}] &= P(\gamma \leq g(x)) \\ &= \int_0^b \int_0^{g(x) \wedge M} \frac{1}{M(b-a)} dy dx \\ &= \frac{1}{M(b-a)} \int_0^b \int_0^{g(x)} 1 dy dx = \frac{1}{M(b-a)} \int_0^b g(x) dx \end{aligned}$$

IMPORTANCE SAMPLING

We want to approximately evaluate $\mathbb{E}_f[S] = \int S(x) f(x) dx$

Let $g : \mathcal{X} \rightarrow [0, \infty)$ be a pdf or pmf such that

$$g(x) = 0 \Rightarrow S(x)f(x) = 0.$$

Define

$$T(x) := \begin{cases} S(x) \frac{f(x)}{g(x)}, & \text{if } g(x) > 0, \\ 0, & \text{if } g(x) = 0. \end{cases}$$

- ▶ Generate $Y_1, \dots, Y_N \stackrel{iid}{\sim} g$
- ▶ $\tilde{\mu}_N := \frac{1}{N} \sum_{i=1}^N T(Y_i) \xrightarrow[N \rightarrow \infty]{i.p.} \mathbb{E}_g[T] = \mathbb{E}_f[S]$

Claim: $\mathbb{E}_g[T] = \mathbb{E}_f[S]$.

IMPORTANCE SAMPLING



universität
wien

We want to approximately evaluate $\mathbb{E}_f[S]$.

Let $g : \mathcal{X} \rightarrow [0, \infty)$ be a pdf or pmf such that

$$g(x) = 0 \quad \Rightarrow \quad S(x)f(x) = 0.$$

Define

$$T(x) := \begin{cases} S(x) \frac{f(x)}{g(x)}, & \text{if } g(x) > 0, \\ 0, & \text{if } g(x) = 0. \end{cases}$$

- ▶ Generate $Y_1, \dots, Y_N \stackrel{iid}{\sim} g$
- ▶ $\tilde{\mu}_N := \frac{1}{N} \sum_{i=1}^N T(Y_i) \xrightarrow[N \rightarrow \infty]{i.p.} \mathbb{E}_g[T]$.
- ▶ Useful when it is easier to sample from g than from f .
- ▶ Can lead to improved efficiency.

EXAMPLE: EVALUATE $P(N(0, 1) \geq 4)$



By classical MC integration:

- ▶ $f(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$, $S(x) = \mathbb{1}_{[4,\infty)}(x)$
- ▶ $X_1, \dots, X_N \stackrel{iid}{\sim} f$
- ▶ $\hat{\mu}_N = \frac{1}{N} \sum_{i=1}^N S(X_i).$

By importance sampling:

- ▶ $f(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$, $S(x) = \mathbb{1}_{[4,\infty)}(x)$
- ▶ $g(x) = e^{-(x-4)}$, if $x \geq 4$, and $g(x) = 0$ else.
- ▶ $Y_1, \dots, Y_N \stackrel{iid}{\sim} g$
- ▶ $\tilde{\mu}_N = \frac{1}{N} \sum_{i=1}^N T(Y_i).$

RANDOM NUMBER GENERATION



$$P(U \leq t) = \begin{cases} 0 & t < 0 \\ t & 0 \leq t \leq 1 \\ 1 & t > 1 \end{cases}$$



1. The inversion method

We want to sample from a distribution with cdf $F: \mathbb{R} \rightarrow [0, 1]$.

Claim: If $U \sim \text{Unif}(0, 1)$ then $F^\dagger(U)$ has cdf F .

Claim: $\underbrace{F^\dagger(\alpha) \leq x}_{\substack{\alpha \in [0, 1] \\ F^\dagger(\alpha) \leq x}} \iff \alpha \leq F(x) \quad \checkmark \quad \begin{matrix} t = F(x) \\ \in [0, 1] \end{matrix}$

$$P(F^\dagger(U) \leq x) = P(U \leq F(x)) = F(x)$$

i.e. cdf of $F^\dagger(U)$ is F .

recall: if F is invertible, then $F^+ = F^{-1}$

2. Rejection sampling

We want to sample from a distribution with pdf $f : \mathcal{X} \rightarrow [0, \infty)$, where possibly $f(x) = c \cdot h(x)$ for unknown norming constant $c > 0$.

$$c = \left(\int_{\mathcal{X}} h(x) dx \right)^{-1}$$

Pick a pdf $g : \mathcal{X} \rightarrow [0, \infty)$ and $M > 0$ such that $h(x) \leq Mg(x)$ for all $x \in \mathcal{X}$. Define

$$T(x) := \begin{cases} \frac{h(x)}{Mg(x)}, & \text{if } g(x) > 0, \\ 0, & \text{if } g(x) = 0. \end{cases} \leq 1$$

Algorithm:

1. Generate $U \sim \text{Unif}(0, 1)$ and $Y \sim g$ indep.
2. If $U \leq T(Y)$, return $X = Y$, otherwise discard Y .
3. Repeat sufficiently many times to produce X_1, \dots, X_N .

Claim: $X_1 \sim f$

Introduction to Bayesian Data Analysis

THE BAYESIAN PARADIGM



statistical model: $(\mathcal{X}, \Theta, (p_\theta)_{\theta \in \Theta})$, $\Theta \subseteq \mathbb{R}^p$

prior distribution: $\pi : \Theta \rightarrow [0, \infty)$ a pdf or pmf

“prior knowledge about the parameter θ ”

The ‘likelihood’

$$p(x|\theta) = p_\theta(x)$$

is understood as the conditional distribution of the data X_1, \dots, X_n given θ , and $\theta \sim \pi$.

Goal: update our belief about θ using the data, i.e., compute (features of) posterior distribution

$$p(\theta|x) := \frac{p(x|\theta)\pi(\theta)}{\int_{\Theta} p(x|\theta)\pi(\theta)d\theta}$$

Bayes formula

Notation 'proportional to' omitting norming constants:

$$p(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{\int_{\Theta} p(x|\theta)\pi(\theta)d\theta} = c(x)p(x|\theta)\pi(\theta)$$

$$p(\theta|x) \propto p(x|\theta)\pi(\theta)$$

$$c(x) = \left(\int_{\Theta} p(x|\theta)\pi(\theta) d\theta \right)^{-1}$$



- ▶ Very popular in applications because of conceptual simplicity and powerful simulation techniques (see later).
- ▶ What about the true parameter?
- ▶ How to choose the prior?
 - ▶ cheating?
 - ▶ non-informative prior
 - ▶ conjugate prior
 - ▶ Jeffrey's prior
- ▶ Conclusions of frequentist and Bayesian analysis often almost identical.



Likelihood:

$$p(x|\theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \theta)^2\right), \quad \theta \in \Theta = \mathbb{R}$$

prior:

$$\pi_\mu(\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\theta - \mu)^2\right)$$

$\mu \in \mathbb{R}$... prior mean

Find posterior distribution!

EXAMPLE: GAUSSIAN MEAN ESTIMATION



Likelihood:

$$p(x|\theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \theta)^2\right), \quad \theta \in \Theta = \mathbb{R}$$

Gaussian prior:

$$\pi_\mu(\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\theta - \mu)^2\right)$$

$$\mathcal{N}(\hat{\theta}_n, \hat{\sigma}_n^2)$$

$\mu \in \mathbb{R}$... prior mean

Gaussian posterior distribution: (π_μ ... 'conjugate prior')

$$p(\theta|x, \mu) \propto p(x|\theta)\pi(\theta) \propto \frac{1}{\sqrt{2\pi\hat{\sigma}_n^2}} \exp\left(-\frac{1}{2}\frac{(\theta - \hat{\theta}_n)^2}{\hat{\sigma}_n^2}\right)$$

$$\hat{\theta}_n = \frac{n}{n+1}\bar{x}_n + \frac{1}{n+1}\mu \quad \dots \text{posterior mean } \hat{\theta}_n(x) = \int_{\mathbb{R}} \theta \cdot p(\theta|x, \mu) d\theta$$

$$\hat{\sigma}_n^2 = \frac{1}{n+1}$$



Likelihood:

$$p(x|\theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \theta)^2\right), \quad \theta \in \Theta = \mathbb{R}$$

non-informative prior / improper prior ($N(0, \infty)$):

$$\pi(\theta) = 1, \quad \theta \in \Theta = \mathbb{R} \quad \text{not a pdf !!!}$$

Can still find Gaussian posterior density

$$p(\theta|x) \propto p(x|\theta) \cdot 1 \propto \frac{1}{\sqrt{2\pi\hat{\sigma}_n^2}} \exp\left(-\frac{1}{2}\frac{(\theta - \bar{x}_n)^2}{\hat{\sigma}_n^2}\right), \quad \hat{\sigma}_n^2 = \frac{1}{n}$$

Here: posterior mean = \bar{x}_n = frequentist MLE

For $\alpha \in (0, 1)$ and a sample $x \in \mathcal{X}$, a $1 - \alpha$ Bayesian credible interval is an interval

$$BI_\alpha(x) = [L_\alpha(x), U_\alpha(x)]$$

such that

$$P(\theta \in BI_\alpha(x) | X = x) \geq 1 - \alpha.$$



Note: Here, the probability is over the randomness in θ given the data, i.e., w.r.t. the posterior distribution,

$$P(L_\alpha(x) \leq \theta \leq U_\alpha(x) | X = x) = \int_{L_\alpha(x)}^{U_\alpha(x)} p(\theta | x) d\theta.$$

Thus, $L_\alpha(x)$ and $U_\alpha(x)$ are quantiles of the posterior distribution.

*posterior
distrib.*



Gaussian Likelihood and Gaussian prior

$$p(x|\theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \theta)^2\right),$$

$$\pi_\mu(\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\theta - \mu)^2\right)$$

Gaussian posterior distribution: ($\hat{\theta}_n = \frac{n}{n+1}\bar{x}_n + \frac{1}{n+1}\mu$, $\hat{\sigma}_n^2 = \frac{1}{n+1}$)

$$p(\theta|x, \mu) = \frac{1}{\sqrt{2\pi\hat{\sigma}_n^2}} \exp\left(-\frac{1}{2}\frac{(\theta - \hat{\theta}_n)^2}{\hat{\sigma}_n^2}\right)$$

$$BI_\alpha(x) = ?$$



Point estimation: compute posterior mean (or median or mode), e.g., $\hat{\theta}_n = \frac{n}{n+1}\bar{x}_n + \frac{1}{n+1}\mu$.

Credible intervals: compute quantiles of the posterior distribution.

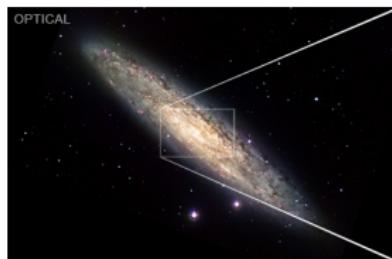
All of Bayesian statistics is concerned with computing features of the posterior distribution!

$$p(\theta|x) := \frac{p(x|\theta)\pi(\theta)}{\int_{\Theta} p(x|\theta)\pi(\theta)d\theta}$$

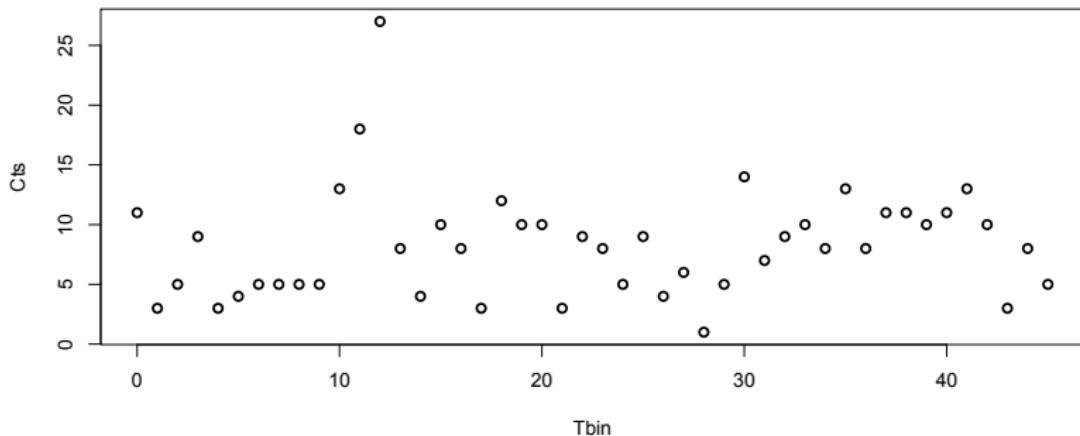
Can be hard in complex models! But all we need is to be able to sample from the posterior distribution!

EXAMPLE:

BAYESIAN CHANGE POINT DETECTION



Chandra: Orion solar flares



EXAMPLE:

BAYESIAN CHANGE POINT DETECTION



The raw data gives the **individual photon arrival times (in seconds)** and their energies (in keV). The processed data we consider here is obtained by **grouping the events into evenly-spaced time bins (10,000 seconds width)**.

Our goal for this data analysis is to **identify the change point** and estimate the intensities of the process before and after the change point.

Source:

<http://www.iiap.res.in/astrostat/School07/R/MCMC.html>

EXAMPLE:

BAYESIAN CHANGE POINT DETECTION

Data:

$Y_i \in \mathbb{N}$... counts of events in time interval $i = 1, \dots, n$
 ↵ $k \in \{1, \dots, n - 1\}$... change point ↵ *not observed!*

How to model this?

$$Y_i = \sum_{t=1}^{10000} X_t$$

$$X_t = \begin{cases} 1, & \text{photon arrival at time } t, \\ 0, & \text{no arrival at time } t. \end{cases}$$

$p = P(X_t = 1)$ is small! Counting rare events.

$$Y_i \sim \text{Binomial}(10000, p)$$



EXAMPLE:

BAYESIAN CHANGE POINT DETECTION

Data:

- „ $Y_i \in \mathbb{N}$... counts of events in time interval $i = 1, \dots, n$
- „ $k \in \{1, \dots, n-1\}$... change point “ *not observed*

Model:

$$\begin{aligned} Y_1, \dots, Y_k | k, \theta, \lambda &\stackrel{iid}{\sim} \text{Poisson}(\theta) \\ Y_{k+1}, \dots, Y_n | k, \theta, \lambda &\stackrel{iid}{\sim} \text{Poisson}(\lambda) \end{aligned} \left. \right\} \text{indep.}$$

prior:

$$\begin{aligned} \theta | b_1, b_2 &\stackrel{iid}{\sim} \text{Exp}(b_1) = \text{Gamma}(1, b_1) \\ \lambda | b_1, b_2 &\stackrel{iid}{\sim} \text{Exp}(b_2) \\ k | b_1, b_2 &\stackrel{iid}{\sim} \text{Unif}(\{1, \dots, n-1\}) \end{aligned} \left. \right\} \text{indep.}$$

hyper-prior : $b_1, b_2 \stackrel{iid}{\sim} \text{Exp}(1)$

Advanced MC methods



Goal: sample from a multivariate pdf $f(x), x \in \mathbb{R}^p$.

Algorithm:

$$\mathbf{x} = (x_1 \dots x_p)$$

Choose a starting value $X^{(0)} \in \mathbb{R}^p$ with $f(X^{(0)}) > 0$.

For $t = 1, 2, \dots$

- 1.) generate $X_1^{(t)}$ from density $x_1 \mapsto f_1(x_1 | X_2^{(t-1)}, \dots, X_p^{(t-1)})$
- k.) generate $X_k^{(t)}$ from density

$$x_k \mapsto f_k(x_k | X_1^{(t)}, \dots, X_{k-1}^{(t)}, X_{k+1}^{(t-1)}, \dots, X_p^{(t-1)})$$

- p.) generate $X_p^{(t)}$ from density $x_p \mapsto f_p(x_p | X_1^{(t)}, \dots, X_{p-1}^{(t)})$

... to get $X^{(t)} = (X_1^{(t)}, \dots, X_p^{(t)})^T \in \mathbb{R}^p$.



Theoretical properties:

- $(X^{(t)})_{t \in \mathbb{N}}$ is a Markov chain in \mathbb{R}^p , i.e.,

$$P(X^{(t)} \in A | X^{(0)}, \dots, X^{(t-1)}) = P(X^{(t)} \in A | X^{(t-1)}), \quad \forall A \subseteq \mathbb{R}^p.$$

- $X^{(t)} \xrightarrow[t \rightarrow \infty]{d.} X \sim f$

-

$$\frac{1}{T} \sum_{t=1}^T h(X^{(t)}) \xrightarrow[T \rightarrow \infty]{i.p.} \mathbb{E}_f[h]$$

EXAMPLE:

BAYESIAN CHANGE POINT DETECTION



posterior distribution:

$$p(\theta, \lambda, k, b_1, b_2 | Y) \propto \\ \propto \theta^{\sum_{i=1}^k Y_i} e^{-k\theta} \lambda^{\sum_{i=k+1}^n Y_i} e^{-(n-k)\lambda} b_1 e^{-n_1 \theta} b_2 e^{-b_2 \lambda} e^{-b_1} e^{-b_2}.$$

For Gibbs we need all five univariate conditional distributions:

$$p(\theta | \lambda, k, b_1, b_2, Y), p(\lambda | \theta, k, b_1, b_2, Y), p(k | \theta, \lambda, b_1, b_2, Y), \\ p(b_1 | \theta, \lambda, k, b_2, Y), p(b_2 | \theta, \lambda, k, b_1, Y).$$

But that is easy!



Goal: sample from a multivariate pdf $f(x)$, $x \in \mathbb{R}^p$.

Algorithm:

Choose a starting value $X^{(0)} \in \mathbb{R}^p$ with $f(X^{(0)}) > 0$.

For $t = 1, 2, \dots$

- A.) generate Z from a proposal density $y \mapsto q(z|X^{(t-1)})$.
- B.) compute

$$\alpha := \alpha(Z, X^{(t-1)}) := \min \left(1, \frac{f(Z)}{f(X^{(t-1)})} \frac{q(X^{(t-1)}|Z)}{q(Z|X^{(t-1)})} \right)$$

- C.) set

$$X^{(t)} = \begin{cases} Z, & \text{with probability } \alpha, \\ X^{(t-1)}, & \text{with probability } 1 - \alpha. \end{cases}$$

Statistics for Data Science, Winter 2023

Chapter 3:
Bootstrap and Jackknife

Bootstrap and Jackknife are simulation based methods for inference/uncertainty quantification.

CIs and testing

pros:

- ▶ Conceptually simple
- ▶ No need for mathematically complex probability calculations

cons:

- ▶ Computationally expensive
- ▶ Requires large sample size (as for the normal approximation)
- ▶ Can go wrong even in large samples!
- ▶ Use with care!

RECALL: SAMPLING DISTRIBUTION



$X_1, \dots, X_n \stackrel{iid}{\sim} \text{Bernoulli}(\theta), \theta \in (0, 1), \hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n X_i$

$$\frac{\hat{\theta}_n - \theta}{se_\theta} \xrightarrow[n \rightarrow \infty]{d.} N(0, 1), \quad se_\theta = \sqrt{\text{Var}_\theta[\hat{\theta}_n]} = \sqrt{\frac{\theta(1-\theta)}{n}}$$

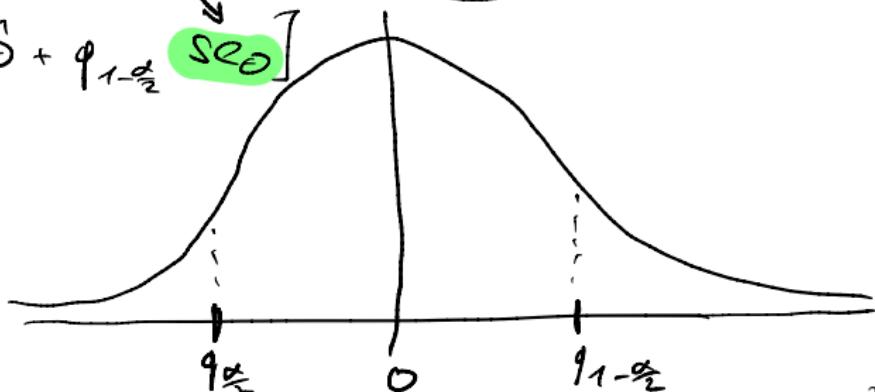
estimate se_θ by plug-in rule $se_{\hat{\theta}_n} = \sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}}$

$$P\left(q_{\frac{\alpha}{2}} \leq \frac{\hat{\theta}_n - \theta}{se_{\hat{\theta}_n}} \leq q_{1-\frac{\alpha}{2}}\right) \approx 1-\alpha \sqrt{\frac{\hat{\theta}_n - \theta}{se_{\hat{\theta}_n}}} \approx N(0, 1)$$

\nwarrow unknown \searrow

$$[\hat{\theta}_n - q_{1-\frac{\alpha}{2}} se_{\hat{\theta}_n}, \hat{\theta}_n + q_{1-\frac{\alpha}{2}} se_{\hat{\theta}_n}]$$

$$\begin{aligned}\hat{se}_{\hat{\theta}_n} &= se_{\hat{\theta}_n} = \\ &= \sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}}\end{aligned}$$



IN GENERAL



$$X_1, \dots, X_n \stackrel{iid}{\sim} f_\theta, \theta \in \Theta,$$

We often have a CLT:

$$\frac{\hat{\theta}_n - \theta}{se_\theta} \xrightarrow[n \rightarrow \infty]{d.} N(0, 1)$$

But the standard error $se_\theta := \sqrt{\text{Var}_\theta[\hat{\theta}_n]}$ may be hard to compute or the plug-in idea may not be feasible.

First Bootstrap idea: *se bootstrap*

Estimate se_θ (without knowing its exact analytical form) and rely on the CLT.

Second Bootstrap idea: *pivotal bootstrap*

Estimate quantiles of the sampling distribution of $\hat{\theta}_n - \theta$. No Gaussian approximation required.

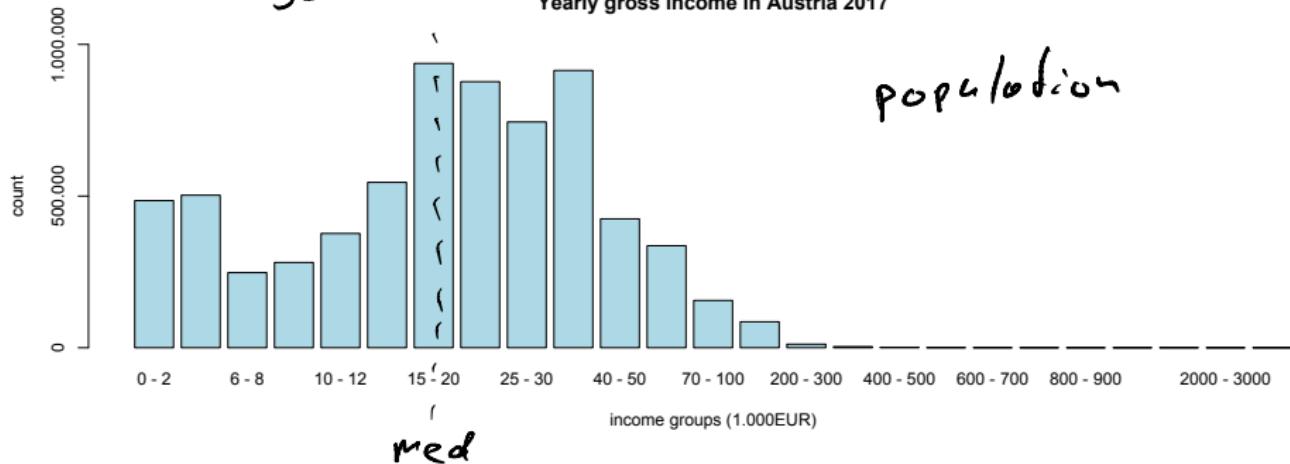
EXAMPLE: MEDIAN INCOME



50% ! 50%

Yearly gross income in Austria 2017

population

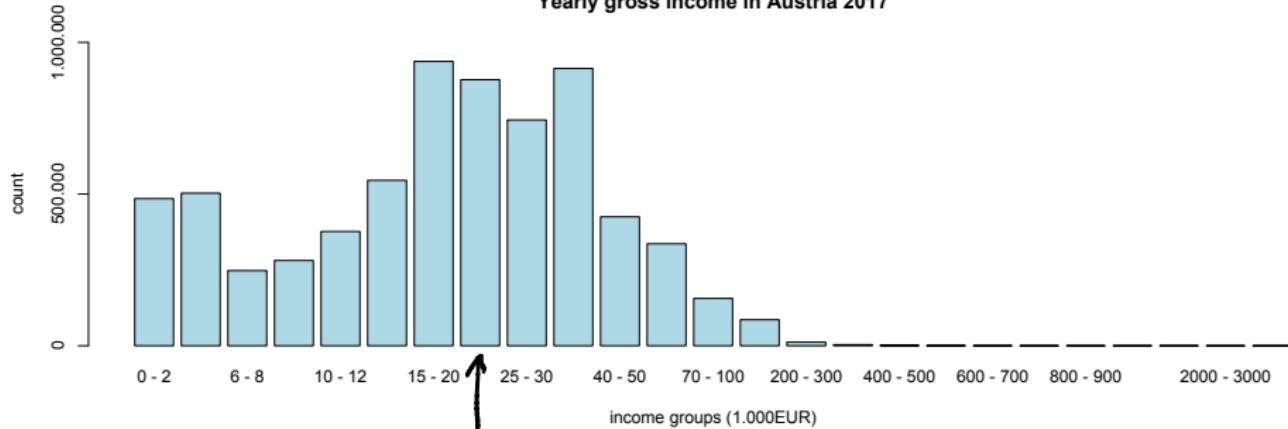


Recall: Median = 50% quantile = smallest number $m \in \mathbb{R}$
such that 50% of data are below or equal to m

EXAMPLE: MEDIAN INCOME



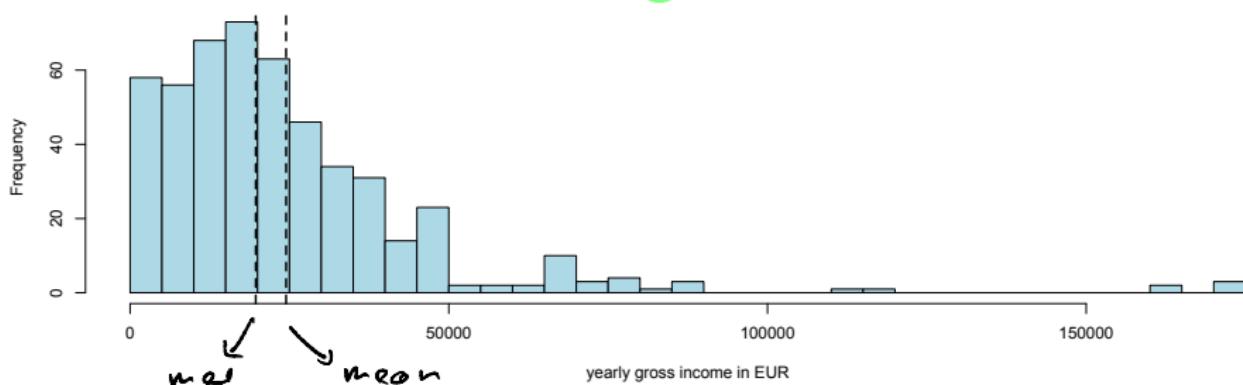
Yearly gross income in Austria 2017



EXAMPLE: MEDIAN INCOME



Histogram of sample of size 500 of yearly gross income



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
8	10716	19712	24475	31148	174629

6

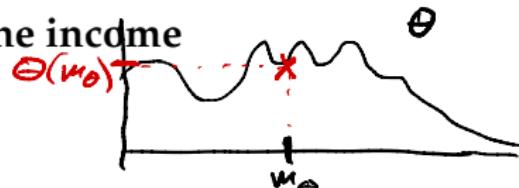
Median should be in the 20k – 25k group!?

EXAMPLE: MEDIAN INCOME



universität
wien

A realistic (nonparametric) model for the income distribution:



$$\mathcal{X} = [0, \infty)^n,$$

$$\Theta = \{\text{all pdfs } \theta \in C^1([0, \infty)) : \theta(m_\theta) > 0\}, m_\theta := \text{med}(\theta)$$

$$f_\theta(x) = \prod_{i=1}^n \theta(x_i), \quad x = (x_1, \dots, x_n)' \in \mathcal{X}, \theta \in \Theta.$$

True population is discrete! Should use pmfs?

→ Model is an approximation/idealization!

EXAMPLE: MEDIAN INCOME

A realistic (nonparametric) model for the income distribution:

$$\mathcal{X} = [0, \infty)^n,$$

$$\Theta = \{\text{all pdfs } \theta \in C^1([0, \infty)) : \theta(m_\theta) > 0\}, m_\theta := \text{med}(\theta)$$

$$f_\theta(x) = \prod_{i=1}^n \theta(x_i), \quad x = (x_1, \dots, x_n)' \in \mathcal{X}, \theta \in \Theta.$$

Use sample median $\hat{m}_n := \hat{F}_n^\dagger(1/2)$ to estimate population median $m_\theta := \psi(\theta) := F_\theta^\dagger(1/2)$, $F_\theta(x) := \int_{-\infty}^x \theta(y) dy$.

It is well known that for every $\theta \in \Theta$,

plug-in for θ ?
estimate $\hat{\theta}$
non-parametrically

$$\frac{\hat{m}_n - m_\theta}{se_\theta} \xrightarrow[n \rightarrow \infty]{d.} N(0, 1)$$

where $se_\theta = \frac{1}{2\sqrt{n}\theta(m_\theta)}$.

How do we estimate that?!

BOOTSTRAP ESTIMATION OF SE



Given data $X = (X_1, \dots, X_n)'$, $X_1, \dots, X_n \stackrel{iid}{\sim} p_\theta$, $\theta \in \Theta$, and an estimator $\hat{\theta}_n : \mathcal{X}_0^n \rightarrow \mathbb{R}$ of θ (or $\psi(\theta)$), we want to estimate

$$se_\theta(\hat{\theta}_n) = \sqrt{\text{Var}_\theta[\hat{\theta}_n]}.$$

1. Draw a large number B of random samples of size n (with replacement) from the sample!

$$X_1^* = (X_{1,1}^*, \dots, X_{n,1}^*)'$$

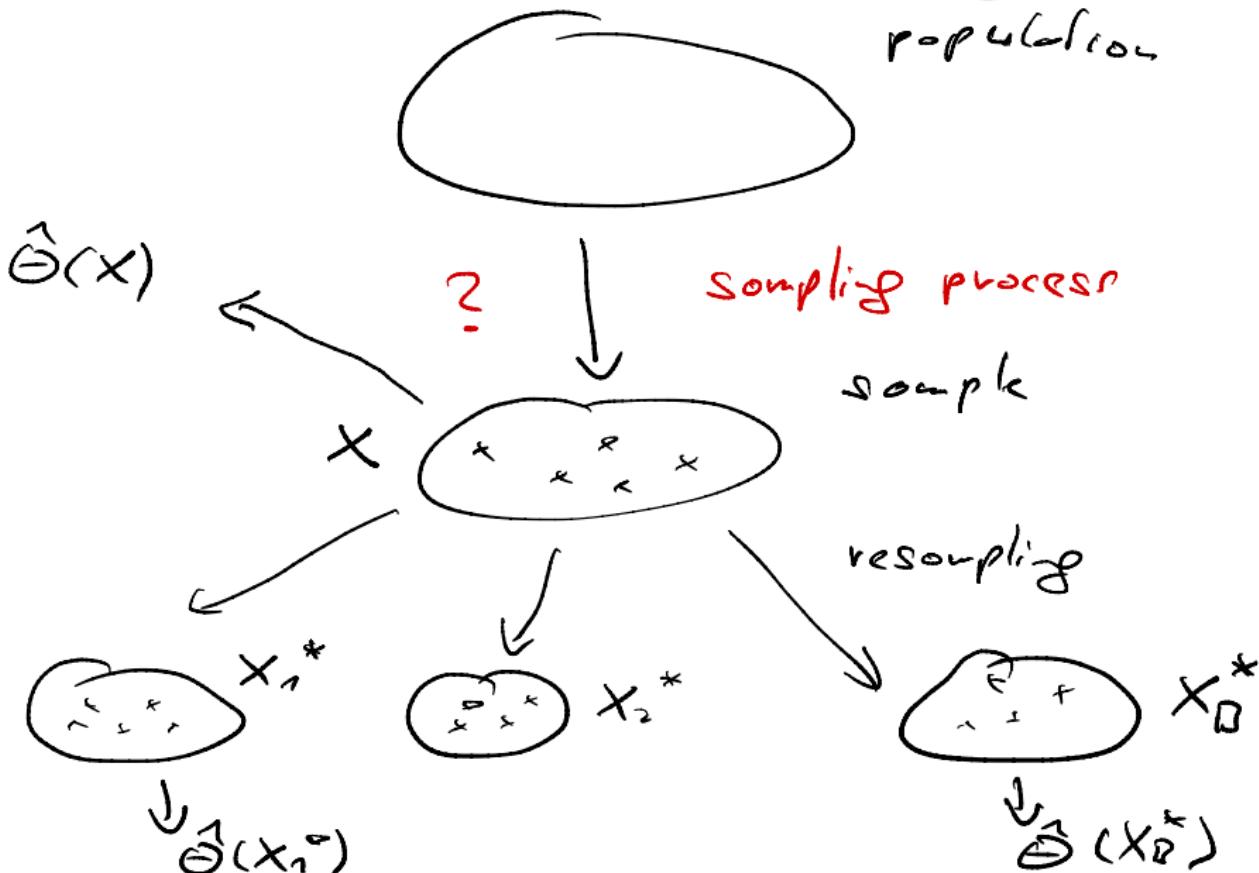
$$X_2^* = (X_{1,2}^*, \dots, X_{n,2}^*)'$$

⋮

$$X_B^* = (X_{1,B}^*, \dots, X_{n,B}^*)'$$

2. Compute $\hat{\theta}_n(X_1^*), \dots, \hat{\theta}_n(X_B^*)$.
3. Approximate $\text{Var}_\theta[\hat{\theta}_n]$ using the LLN (MC idea)

$$\hat{s}e_{boot}^2 = \frac{1}{B} \sum_{j=1}^B \hat{\theta}_n(X_j^*)^2 - \left(\frac{1}{B} \sum_{j=1}^B \hat{\theta}_n(X_j^*) \right)^2.$$





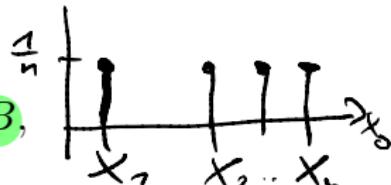
Pretend that the sample is the population and simulate (imitate) the sampling process by 'resampling' from the sample.

Why should this work?

Assume that the original data sample $X = (X_1, \dots, X_n)'$ from sample space $\mathcal{X} = \mathcal{X}_0^n$ is given and non-random.

- ▶ Then the Bootstrap draws

$$X_{i,j}^*, i = 1, \dots, n, j = 1, \dots, B,$$



are iid from the empirical distribution \hat{F}_n of the data, which has pmf

$$\hat{p}_n(x_0) := \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i\}}(x_0), \quad x_0 \in \{X_1, \dots, X_n\}.$$

- ▶ Thus, also the X_1^*, \dots, X_B^* are iid from the n -fold product

$$\hat{f}_n(x) := \prod_{i=1}^n \hat{p}_n(x_i), \quad x = (x_1, \dots, x_n)' \in \{X_1, \dots, X_n\}^n.$$

- But if $X_1^*, \dots, X_B^* \stackrel{iid}{\sim} \hat{f}_n$, then (by the LLN)

$$\hat{s}e_{boot}^2 = \frac{1}{B} \sum_{j=1}^B \hat{\theta}_n(X_j^*)^2 - \left(\frac{1}{B} \sum_{j=1}^B \hat{\theta}_n(X_j^*) \right)^2 \xrightarrow[B \rightarrow \infty]{i.p.} \text{Var}_{\hat{f}_n}[\hat{\theta}_n].$$

- Because $X_1^* = (X_{1,1}^*, \dots, X_{n,1}^*)' \sim \hat{f}_n$ follows an iid model with marginal pmf \hat{p}_n , we also write $\text{Var}_{\hat{f}_n}[\hat{\theta}_n] = \text{Var}_{\hat{p}_n}[\hat{\theta}_n]$.
- Since the sample $X_1, \dots, X_n \stackrel{iid}{\sim} p_\theta$, we have $\hat{F}_n \approx F_\theta$ (recall HW3), $\hat{p}_n \approx p_\theta$ and one can often show that

$$\text{Var}_{\hat{p}_n}[\hat{\theta}_n] \approx \text{Var}_{p_\theta}[\hat{\theta}_n] = \text{Var}_\theta[\hat{\theta}_n], \quad \text{if } n \text{ is large.}$$

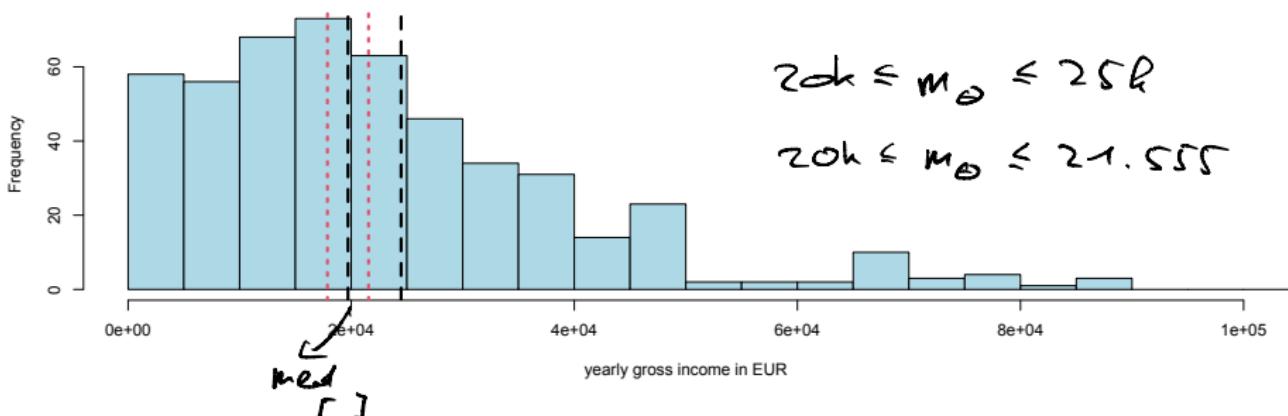
$$\underbrace{\hat{s}e_{boot}^2 \approx \text{Var}_{\hat{p}_n}[\hat{\theta}_n]}_{\text{if } B \text{ is large}}, \quad \underbrace{\text{Var}_{\hat{p}_n}[\hat{\theta}_n] \approx \text{Var}_\theta[\hat{\theta}_n] = se_\theta^2(\hat{\theta}_n)}_{\text{if } n \text{ is large}}$$

EXAMPLE: MEDIAN INCOME

Compute an approximate 95% ($\alpha = 0.05$) bootstrap confidence interval (by bootstrap estimation of the standard error)

$$CI_{\alpha} = [\hat{m}_n - q_{1-\frac{\alpha}{2}}^{(N)} \hat{s}e_{boot}, \hat{m}_n + q_{1-\frac{\alpha}{2}}^{(N)} \hat{s}e_{boot}]$$

Histogram of sample of size 500 of yearly gross income



```
> c(CI_l, CI_u)  
[1] 17870 21555
```

```
> CI_u - CI_l  
[1] 3685
```

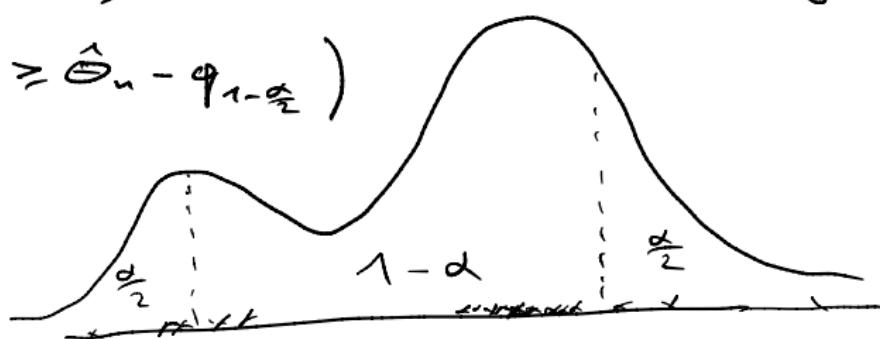
Given data $X = (X_1, \dots, X_n)'$, $X_1, \dots, X_n \stackrel{iid}{\sim} p_\theta$, $\theta \in \Theta$, and an estimator $\hat{\theta}_n : \mathcal{X}_0^n \rightarrow \mathbb{R}$ of θ (or $\psi(\theta)$), we want to estimate the $\alpha/2$ and $1 - \alpha/2$ quantiles of the distribution of $\hat{\theta}_n(X) - \theta$.

Why? Say $\hat{\theta}_n - \theta \sim g_n$ is the unknown sampling distribution.

$$P_\theta(q_{\frac{\alpha}{2}} \leq \hat{\theta}_n - \theta \leq q_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

$$\hat{\theta}_n - \theta \sim g_n$$

$$= P_\theta(\hat{\theta}_n - q_{\frac{\alpha}{2}} \geq \theta \geq \hat{\theta}_n - q_{1-\frac{\alpha}{2}})$$



$$CI_\alpha = [\hat{\theta}_n - q_{1-\frac{\alpha}{2}}, \hat{\theta}_n - q_{\frac{\alpha}{2}}]$$

Given data $X = (X_1, \dots, X_n)'$, $X_1, \dots, X_n \stackrel{iid}{\sim} p_\theta$, $\theta \in \Theta$, and an estimator $\hat{\theta}_n : \mathcal{X}_0^n \rightarrow \mathbb{R}$ of θ (or $\psi(\theta)$), we want to estimate the $\alpha/2$ and $1 - \alpha/2$ quantiles of the distribution of $\hat{\theta}_n(X) - \theta$.

1. Draw a large number B of random samples of size n (with replacement) from the sample!

$$X_1^*, X_2^*, \dots, X_B^*$$

2. Compute empirical (bootstrap) quantiles $\hat{q}_{\alpha/2}^*$ and $\hat{q}_{1-\frac{\alpha}{2}}^*$ of $\hat{\theta}_n(X_1^*) - \hat{\theta}_n(X)$, ..., $\hat{\theta}_n(X_B^*) - \hat{\theta}_n(X)$.
3. $CI_\alpha = [\hat{\theta}_n(X) - \hat{q}_{1-\frac{\alpha}{2}}^*, \hat{\theta}_n(X) - \hat{q}_{\alpha/2}^*]$

EXAMPLE: MEDIAN INCOME



Comparison of 95% bootstrap CIs:

$B = 100$

method	lower	median	upper	length
bootstrap se	17870	19712	21555	3685
pivotal bootstrap	18074	19712	21452	3378

$B = 1000$

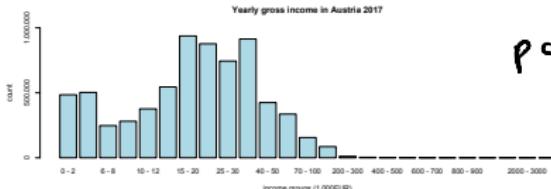
method	lower	median	upper	length
bootstrap se	18100	19712	21324	3224
pivotal bootstrap	18161	19712	21388	3227

EXAMPLE: MEDIAN INCOME



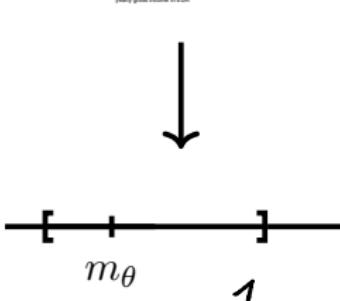
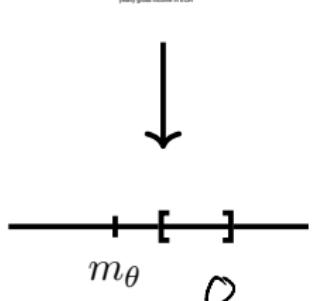
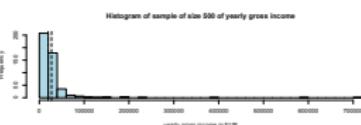
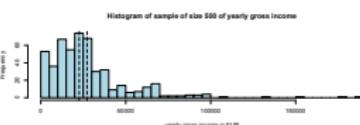
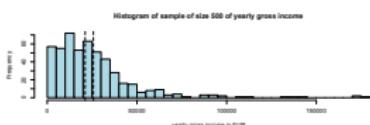
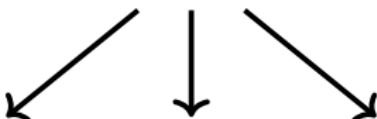
Can we actually trust this CI? Check by simulation!

$$m_\theta = F_\theta^\dagger(1/2)$$



population

samples of size
 $n = 500$



EXAMPLE: MEDIAN INCOME



Simulation results:

$$1 - \alpha = 0.95$$

number of (Montecarlo) samples drawn = 1000

$$B = 1000$$

sample size $n = 100$

method	coverage prob.	average length
se bootstrap	0.932	7601
pivotal bootstrap	0.871	7404

$$\geq 1 - \alpha = 0.95$$

sample size $n = 1000$

method	coverage prob.	average length
se bootstrap	0.935	2412
pivotal bootstrap	0.917	2377

Failure of the Bootstrap



- ▶ Regard your data sample as the population.
- ▶ Draw B iid random (bootstrap) samples from the sample (re-sample) like in a MC simulation.
- ▶ Compute your estimator on each of the bootstrap samples.
- ▶ Use this resulting 'bootstrap distribution' of your estimator as an approximation of its true unknown sampling distribution.



- ▶ In the median income example we simulated the actual coverage probability of the bootstrap CIs and found

$$P_\theta(m_\theta \in CI_\alpha) \approx 1 - \alpha.$$

- ▶ We did that for the true data generating parameter $\theta \in \Theta$.
(we cheated!)
- ▶ We concluded that the bootstrap works relatively well for n large.
- ▶ In practice we would try many different choices for θ .



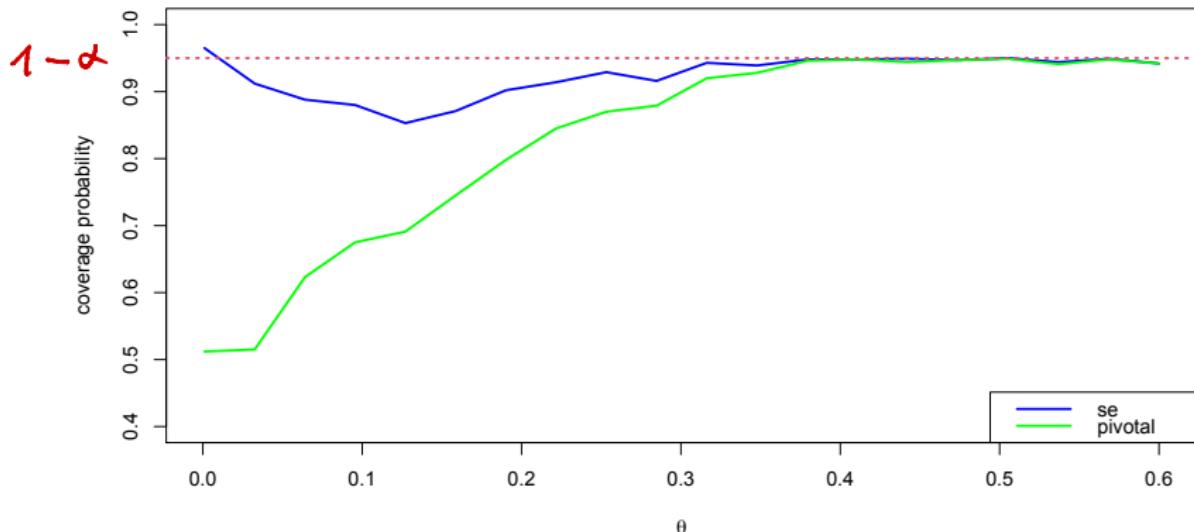
- ▶ Consider $X_1, \dots, X_n \stackrel{iid}{\sim} N(\theta, 1)$, $\theta \in \Theta = [0, \infty)$.
- ▶ Goal: Bootstrap inference on θ .
- ▶ The classical estimator (MLE) is

$$\hat{\theta}_n = \max\{0, \bar{X}_n\}, \quad \bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i.$$

FAILURE OF THE BOOTSTRAP



- ▶ $n = 100, \alpha = 0.05, B = 1000$
- ▶ 1000 MC iterations



FAILURE OF THE BOOTSTRAP

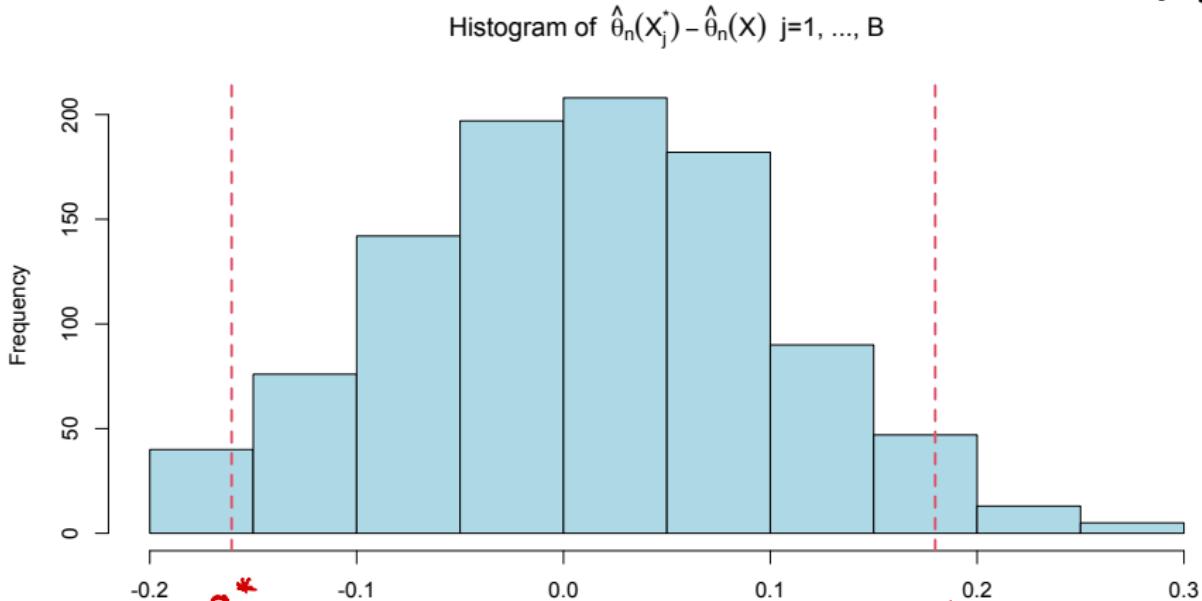


What went wrong?

Look at the 'bootstrap distribution' for a sample

$X = (X_1, \dots, X_n)$ with $\bar{X}_n = 0.185$.

$$\hat{\theta}_n(x_i^*) = \max\{\theta_n(\bar{X}_j^*)\}$$



$$[\hat{\theta}_n(x) - q_{1-\alpha}^*, \hat{\theta}_n(x) - q_{\alpha}^*]$$

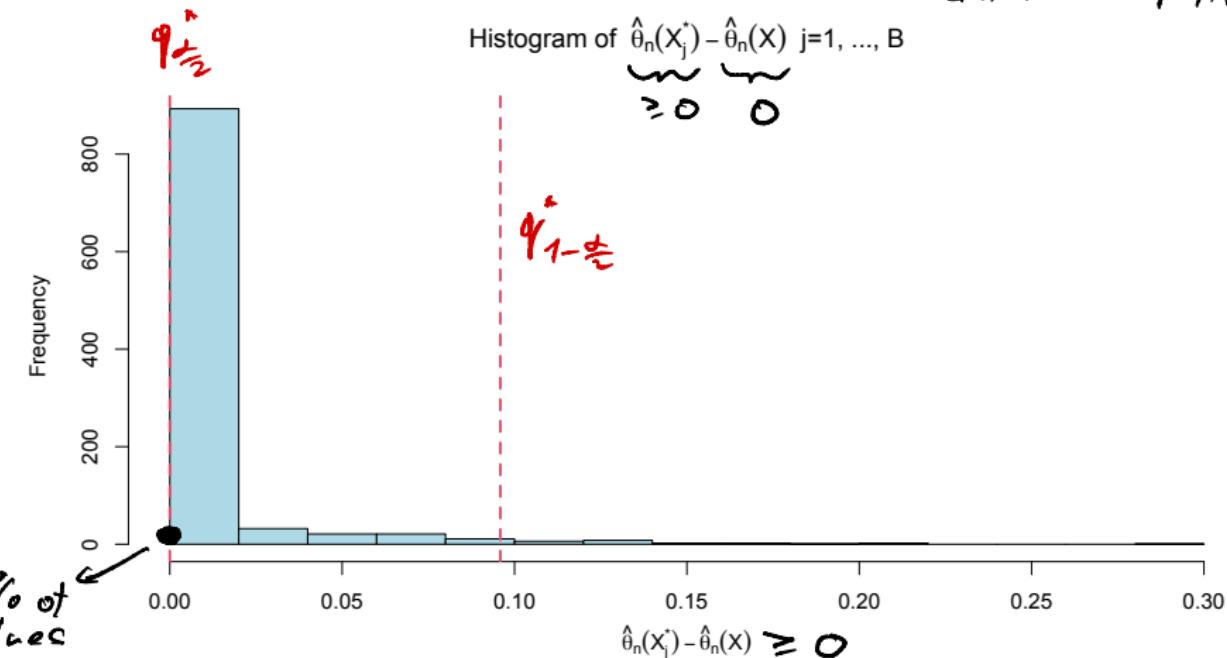
FAILURE OF THE BOOTSTRAP



Look at the 'bootstrap distribution' for a sample with

$$\bar{X}_n = -0.122.$$

$$\hat{\Theta}_-(X) = \max\{0, \hat{X}_-\}$$



Here, actually 86% of bootstrap samples X_j^* produce
 $\hat{\theta}_n(X_j^*) - \hat{\theta}_n(X) = 0.$

- ▶ Look at the ‘bootstrap distribution’ for a sample with $\bar{X}_n = -0.122$.
- ▶ Notice: for every $j \in \{1, \dots, B\}$, we have

$$\hat{\theta}_n(X_j^*) - \hat{\theta}_n(X) = \max\{0, \bar{X}_j^*\} - \max\{0, \bar{X}_n\} \geq 0,$$
$$\Rightarrow 0 \leq \hat{q}_{\alpha/2}^* \leq \hat{q}_{1-\alpha/2}^*.$$

- ▶ Therefore,

$$CI_\alpha = [\underbrace{\hat{\theta}_n(X) - \hat{q}_{1-\alpha/2}^*}_{=0}, \underbrace{\hat{\theta}_n(X) - \hat{q}_{\alpha/2}^*}_{=0}] \subseteq (-\infty, 0] \quad \forall \theta > 0.$$

FAILURE OF THE BOOTSTRAP



Recall: $X_1, \dots, X_n \stackrel{iid}{\sim} N(\theta, 1)$

$\alpha = 0, 05$

Thus, intuitively, for very small $\theta > 0$,

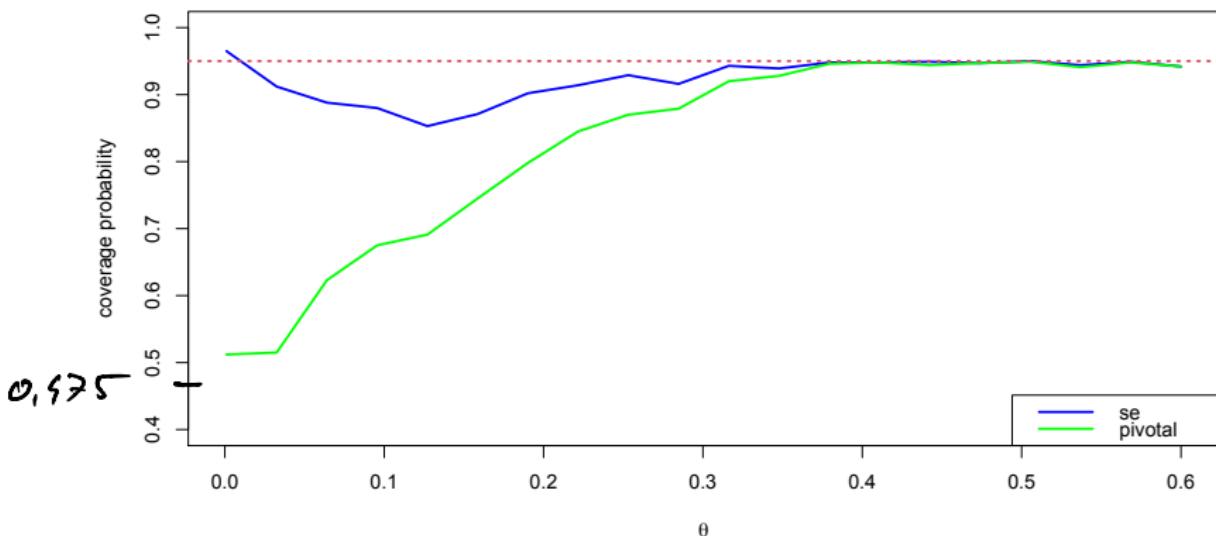
$$P_\theta(\theta \in CI_{0.05}) = P_\theta(\theta \in CI_{0.05} | \bar{X}_n < 0)P_\theta(\bar{X}_n < 0) \\ = 0 \approx \frac{1}{2}$$

$$\bar{X}_n \sim N(\theta, \frac{1}{n}) + P_\theta(\theta \in CI_{0.05} | \bar{X}_n \geq 0)P_\theta(\bar{X}_n \geq 0) \\ \approx 0, 95 \leq 1 \approx \frac{1}{2} \\ \approx 0, 475$$

FAILURE OF THE BOOTSTRAP



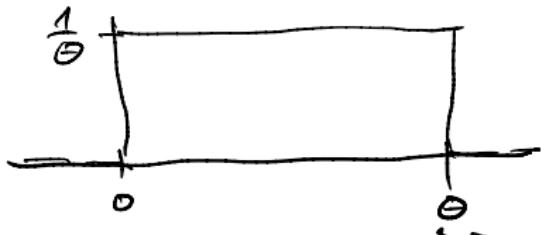
- ▶ $n = 100, \alpha = 0.05, B = 1000$
- ▶ 1000 MC iterations



Is the se-bootstrap always superior to the pivotal method?

- ▶ Consider $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Unif}[0, \theta]$, $\theta \in \Theta = (0, \infty)$.
- ▶ Goal: Bootstrap inference on θ .
- ▶ The classical estimator (MLE) is

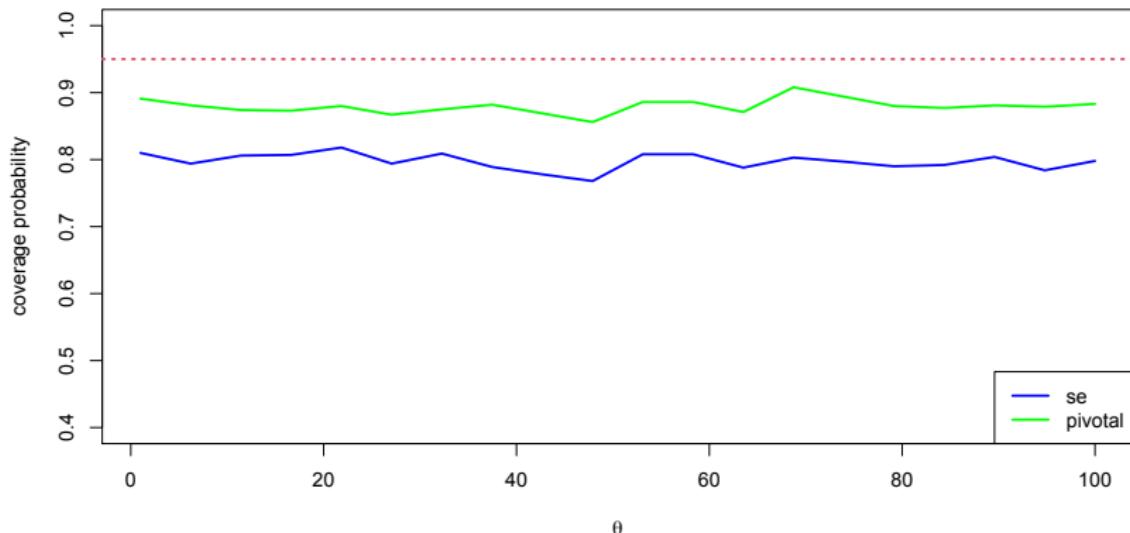
$$\hat{\theta}_n = \max\{X_1, \dots, X_n\}.$$



FAILURE OF THE BOOTSTRAP



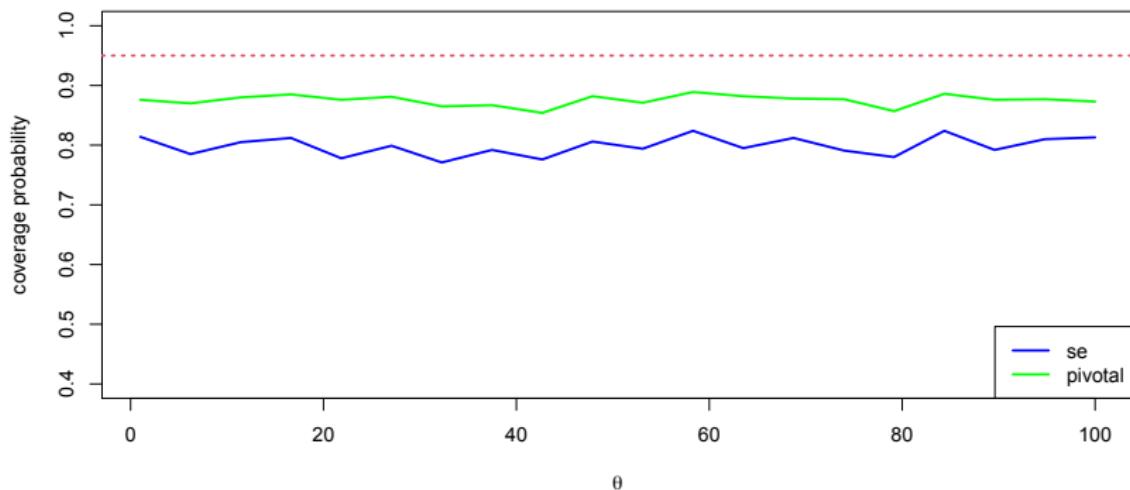
- ▶ $n = 100, \alpha = 0.05, B = 1000$
- ▶ 1000 MC iterations



FAILURE OF THE BOOTSTRAP



- ▶ $n = 1000, \alpha = 0.05, B = 1000$
- ▶ 1000 MC iterations

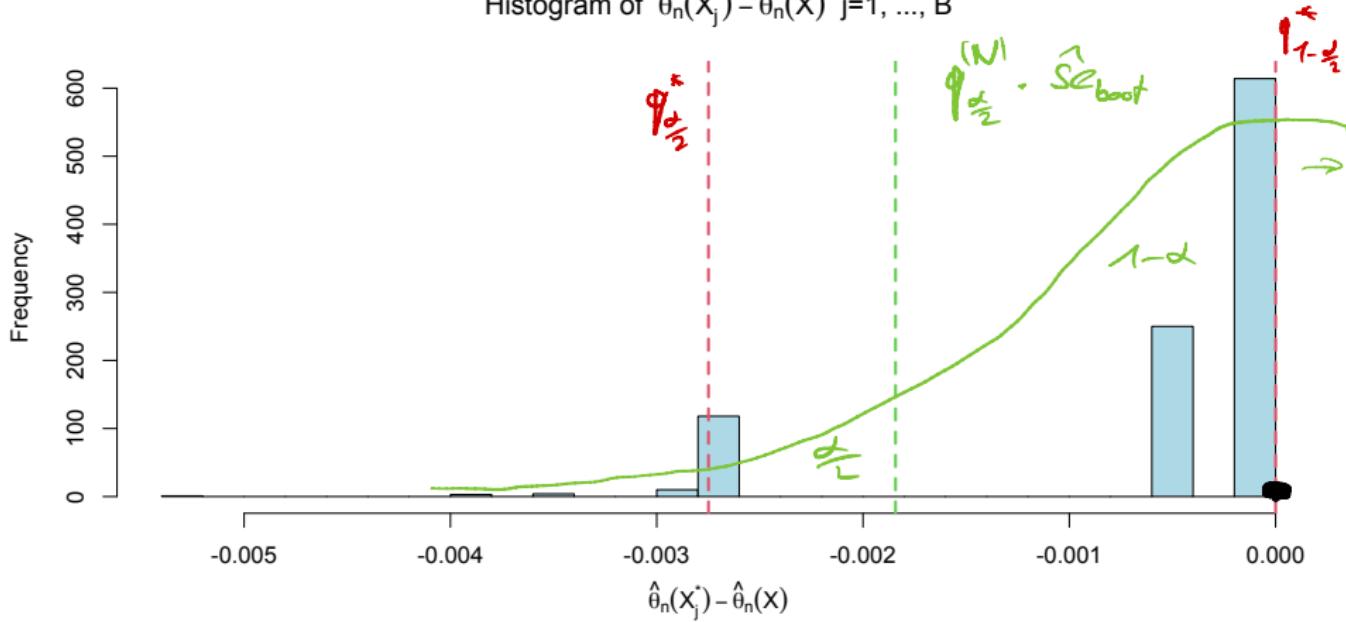


FAILURE OF THE BOOTSTRAP



Look at the bootstrap distribution of one given sample of size $n = 1000$ with $\theta = 1$.

Histogram of $\hat{\theta}_n(X_j^*) - \hat{\theta}_n(X)$ $j=1, \dots, B$



Here, 62% of all bootstrap samples X_j^* produce $\hat{\theta}_n(X_j^*) - \hat{\theta}_n(X) = 0$.

Here, 62% of all bootstrap samples X_j^* produce

$$\hat{\theta}_n(X_j^*) - \hat{\theta}_n(X) = 0.$$

Why?
assume no ties

$$\hat{\theta}_n(X_1^*) - \hat{\theta}_n(X) = \max\{X_{1,1}^*, \dots, X_{n,1}^*\} - \max\{X_1, \dots, X_n\}$$

- ▶ This is equal to 0 if, and only if, in our bootstrap sample $X_{1,1}^*, \dots, X_{n,1}^*$ we happen to draw the largest sample point from $\{X_1, \dots, X_n\}$.
- ▶ What is the probability of that happening?

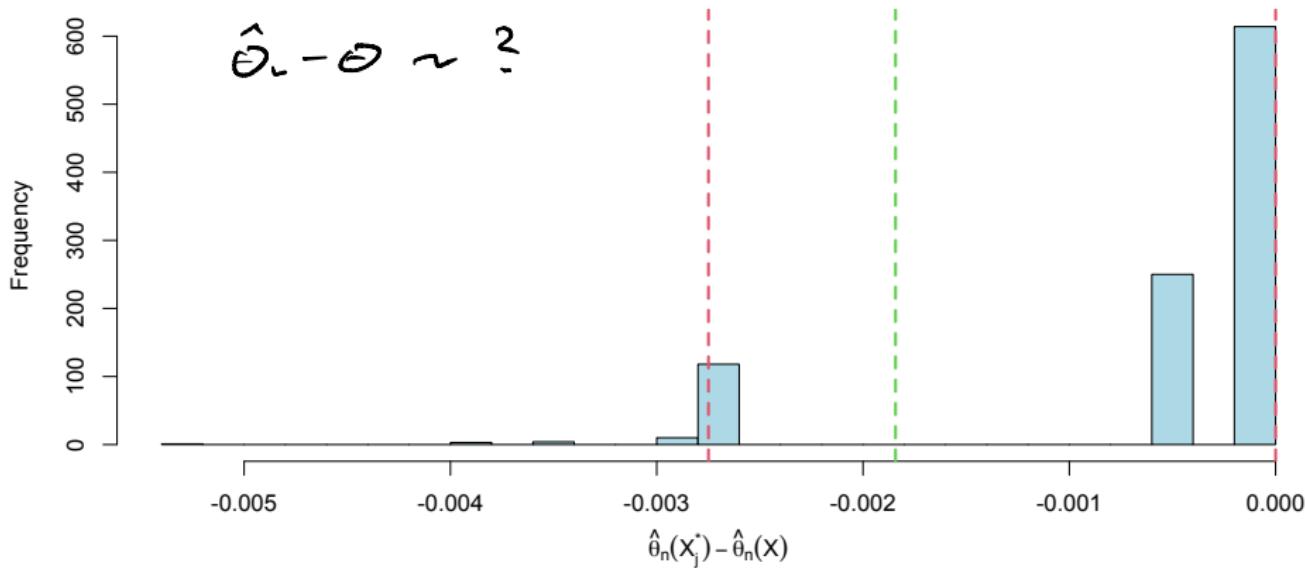
$$1 - \left(\frac{n-1}{n}\right)^n = 1 - \left(1 - \frac{1}{n}\right)^n \xrightarrow{n \rightarrow \infty} 1 - e^{-1} \approx 0.63$$

FAILURE OF THE BOOTSTRAP



Look at the bootstrap distribution of one given sample of size $n = 1000$ with $\theta = 1$.

Histogram of $\hat{\theta}_n(X_j^*) - \hat{\theta}_n(X)$ $j=1, \dots, B$



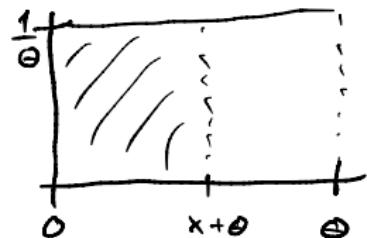
Here, 62% of all bootstrap samples X_j^* produce $\hat{\theta}_n(X_j^*) - \hat{\theta}_n(X) = 0$.

What is the actual sampling distribution of $\hat{\theta}_n - \theta$?

Recall:

$$X_1, \dots, X_n \stackrel{iid}{\sim} \text{Unif}[0, \theta], \theta \in \Theta = (0, \infty), \hat{\theta}_n = \max\{X_1, \dots, X_n\}.$$

$$\begin{aligned} P_\theta(\hat{\theta}_n - \theta \leq x) &= P_\theta\left(\max_{1 \leq i \leq n} \{X_i\} \leq x + \theta\right) \\ &= P_\theta(X_1 \leq x + \theta \text{ and } \dots \text{ and } X_n \leq x + \theta) \\ &\stackrel{\text{indep.}}{=} \prod_{i=1}^n P_\theta(X_i \leq x + \theta) \end{aligned}$$



$$= \begin{cases} 0 & \text{if } x + \theta < 0, \text{ or } x < -\theta \\ \frac{x+\theta}{\theta} & \text{if } 0 \leq x + \theta \leq \theta, -\theta \leq x \leq 0 \\ 1 & \text{if } x + \theta > \theta, \text{ or } x > 0 \end{cases}$$

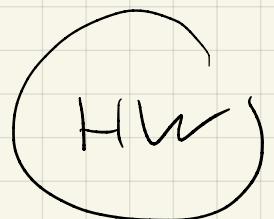
$$P_0(\bar{\Theta}_n - \Theta \leq x) = \begin{cases} 0 & x < -\Theta \\ \left(1 + \frac{x}{\Theta}\right)^n & -\Theta \leq x \leq 0 \\ 1 & x > 0 \end{cases}$$

$$g_n(x) = \frac{d}{dx} P_0(\bar{\Theta}_n - \Theta \leq x)$$

$$= \begin{cases} 0 & x < -\Theta \\ n \left(1 + \frac{x}{\Theta}\right)^{n-1} \frac{1}{\Theta}, & -\Theta \leq x \leq 0 \\ 0, & x > 0 \end{cases}$$

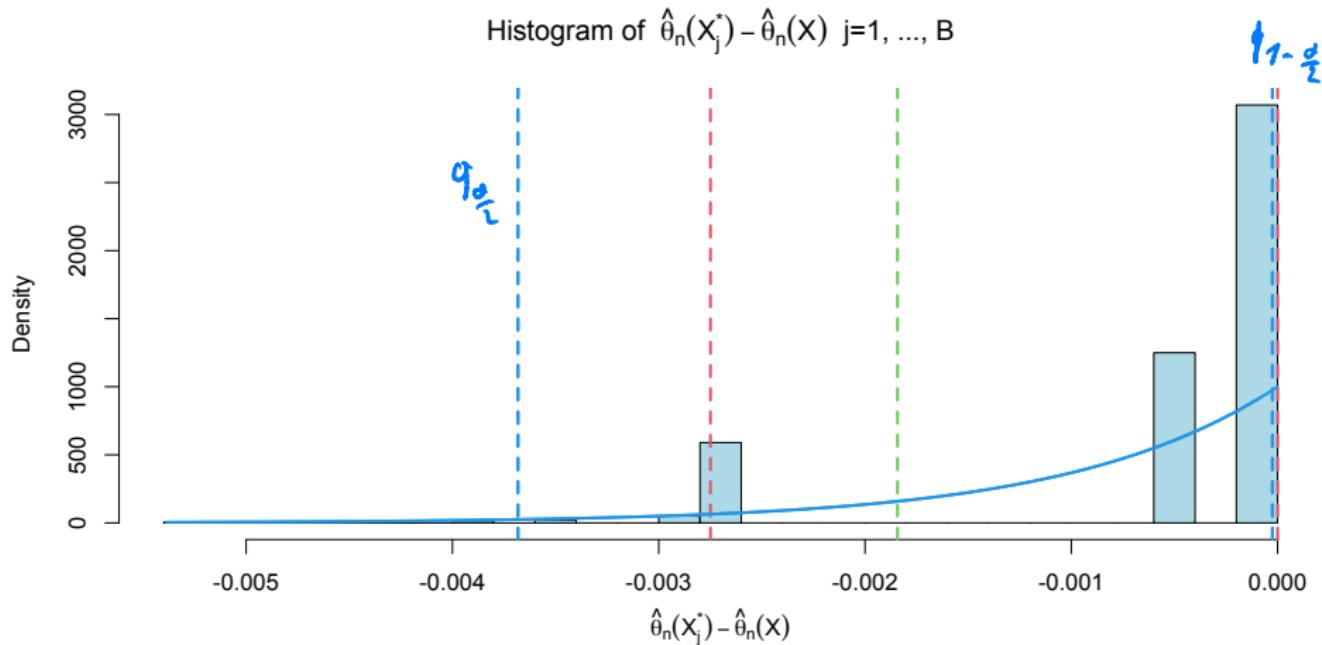
pdf of sampling distribution

$$f_x(g_n) = \left[x^{\frac{1}{n}} - 1 \right] \Theta$$



FAILURE OF THE BOOTSTRAP

Bootstrap vs. true sampling distribution of $\hat{\theta}_n - \theta$. $\theta = 1, n = 1000.$



Here, 62% of all bootstrap samples X_j^* produce $\hat{\theta}_n(X_j^*) - \hat{\theta}_n(X) = 0.$



**Always take a look at the
bootstrap distribution!!!**



Summing up:

- ▶ Computationally expensive simulation method.
- ▶ Very flexible, generic, no need for problem specific formulas.
- ▶ Produces only approximate inference; requires large samples.
- ▶ Can go wrong!
- ▶ In general: Works, if the normal approximation works.

The Jackknife

... is a more specific resampling plan based on a leave-one-out idea:

- ▶ Let $\hat{\theta}_n : \mathcal{X}_0^n \rightarrow \mathbb{R}$ be an estimator.
- ▶ For $x = (x_1, \dots, x_n)' \in \mathcal{X}_0^n$ and $i \in \{1, \dots, n\}$, let

$$\hat{\theta}_{(i)}(x) := \hat{\theta}_{n-1}(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$$

be the estimator computed without the i -th observation.

- ▶ Write $\hat{\theta}_{(\cdot)}(x) := \frac{1}{n} \sum_{i=1}^n \hat{\theta}_{(i)}(x)$.
- ▶ The Jackknife estimate of the squared standard error (=variance) of $\hat{\theta}_n$ is given by:

$$\hat{s}e^2(x) := \frac{n-1}{n} \sum_{i=1}^n \left(\hat{\theta}_{(i)}(x) - \hat{\theta}_{(\cdot)}(x) \right)^2.$$

$$\hat{se}^2 := \frac{n-1}{n} \sum_{i=1}^n \left(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)} \right)^2.$$

This does not look quite right! Why is it not a sample variance?

Consider the case $\hat{\theta}_n(x) = \frac{1}{n} \sum_{i=1}^n x_i$. Then

$$\hat{\theta}_{(i)} = \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^n x_j = \frac{1}{n-1} \left(n \bar{x}_n - x_i \right)$$

$$\begin{aligned} \hat{\theta}_{(\cdot)} &= \frac{1}{n} \sum_{i=1}^n \hat{\theta}_{(i)} = \frac{1}{n-1} \cdot \frac{1}{n} \sum_{i=1}^n \left(n \bar{x}_n - x_i \right) \\ &= \frac{1}{n-1} \left(n \bar{x}_n - \bar{x}_n \right) = \bar{x}_n \end{aligned}$$

$$\hat{se}^2 := \frac{n-1}{n} \sum_{i=1}^n \left(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)} \right)^2.$$

$$\hat{se}^2 = \frac{n-1}{n} \sum_{i=1}^n \underbrace{\left(\frac{1}{n-1} (\bar{x}_n - x_i) - \bar{x}_n \right)^2}_{\textcircled{*}}$$

$$\textcircled{*} = \frac{n}{n-1} \bar{x}_n - \bar{x}_n - \frac{1}{n-1} x_i = \frac{1}{n-1} (\bar{x}_n - x_i)$$

$$\hat{\sigma}_{\hat{\theta}}^2 = \frac{n-1}{n} \underbrace{\frac{1}{(n-1)^2} \sum_{i=1}^n (x_i - \bar{x}_n)^2}_{\hat{\sigma}_n^2} = \frac{\hat{\sigma}_n^2}{n}$$

$$\mathbb{E}_{\theta} \left[\hat{\sigma}_{\hat{\theta}}^2 \right] = \sigma_{\theta}^2$$

Recall HW 2.1a: In the iid model $se_{\theta}^2 := \text{Var}_{\theta}[\hat{\theta}_n] = \frac{\sigma^2}{n}$ and $\mathbb{E}_{\theta}[\hat{\sigma}_n^2] = \sigma^2$.



Unfortunately, however, the Jackknife does not always produce good estimates for the standard error!

e.g., sample quantiles

H W

For these kinds of parametric problems, the Jackknife idea is kind of outdated.

However, ...

UNCERTAINTY QUANTIFICATION IN STATISTICAL LEARNING



- ▶ We observe iid pairs $Z_i = (X_i, Y_i)$, $i = 1, \dots, n$ from (marginal) sample space $\mathcal{X}_0 = \mathbb{R}^p \times \mathbb{R}$.
- ▶ Let (X_0, Y_0) be another independent pair with identical distribution (prediction period).
- ▶ We observe X_0 but not Y_0 . Want to predict the value of Y_0 .
- ▶ Use a predictor/learning algorithm $\hat{m}_n : \mathbb{R}^p \rightarrow \mathbb{R}$ to predict the value of Y_0 by $\hat{m}_n(X_0)$.
- ▶ Actually \hat{m}_n depends also on the training data! So $\hat{m}_n : \mathcal{X}_0^n \times \mathbb{R}^p \rightarrow \mathbb{R}$, $\hat{m}_n(X_0) = \hat{m}_n(Z_1, \dots, Z_n; X_0)$.
- ▶ For example:
 - ▶ $\hat{m}_n(x) = x' \hat{\beta}_n$ with $\hat{\beta}_n = (X'X + \lambda I_p)^{-1} X'Y$,
 $X = [X_1, \dots, X_n]', Y = (Y_1, \dots, Y_n)'$
 - ▶ \hat{m}_n is a CNN with weights obtained from SGD.

UNCERTAINTY QUANTIFICATION IN STATISTICAL LEARNING



- ▶ We would like to quantify the uncertainty associated with predicting the new label/response Y_0 .
- ▶ Prediction interval: $PI_\alpha \subseteq \mathbb{R}$

$$P(Y_0 \in PI_\alpha) \geq 1 - \alpha.$$

- ▶ Would like to know the distribution of the prediction error

$$P\left(q_{\frac{\alpha}{2}} \leq Y_0 - \hat{m}_n(X_0) \leq q_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

- ▶ Could use theoretical quantiles $q_{\alpha/2}$ and $q_{1-\alpha/2}$ to construct

$$PI_\alpha = [\hat{m}_n(X_0) + q_{\alpha/2}, \hat{m}_n(X_0) + q_{1-\alpha/2}].$$

PREDICTIVE INFERENCE BY SAMPLE SPLITTING

- ▶ How to estimate/approximate the distribution of the prediction error

$$Y_0 - \hat{m}_n(Z_1, \dots, Z_n; X_0)$$

- ▶ Traditional approach: Split the sample into $S_{train} \cup S_{val} = \{1, \dots, n\}$, $S_{train} \cap S_{val} = \emptyset$, $n_1 = |S_{train}|$, $n_2 = |S_{val}|$, $n_1 + n_2 = n$.
- ▶ Train your algorithm on S_{train} and validate it on S_{val} , i.e., compute

$$\text{n}_2 \quad \boxed{R_j^{ss} := Y_j - \hat{m}_{n_1}(\{Z_i : i \in S_{train}\}; X_j), \quad j \in S_{val}.}$$

- ▶ Use empirical quantiles $\hat{q}_{\alpha/2}$ and $\hat{q}_{1-\alpha/2}$ of $R_j^{ss}, j \in S_{val}$ and compute

$$PI_\alpha = [\hat{m}_{n_1}(X_0) + \hat{q}_{\alpha/2}, \hat{m}_{n_1}(X_0) + \hat{q}_{1-\alpha/2}].$$

PREDICTIVE INFERENCE BY SAMPLE SPLITTING

- ▶ Conditional on the data in S_{train} , the residuals

$$R_j^{ss} := Y_j - \hat{m}_{n_1}(\{Z_i : i \in S_{train}\}; X_j); \quad j \in S_{val}.$$

are an iid sample with distribution equal to that of

$$R^{ss} := Y_0 - \hat{m}_{n_1}(\{Z_i : i \in S_{train}\}, X_0).$$

- ▶ Thus, $\hat{q}_\alpha \xrightarrow{p.} q_\alpha^{(R^{ss})}$ as $n_2 \rightarrow \infty$.

$$Y_0 \in PI_\alpha = [\hat{m}_{n_1}(X_0) + \hat{q}_{\alpha/2}, \hat{m}_{n_1}(X_0) + \hat{q}_{1-\alpha/2}]$$

$$\iff \hat{q}_{\alpha/2} \leq Y_0 - \hat{m}_{n_1}(X_0) \leq \hat{q}_{1-\alpha/2}$$

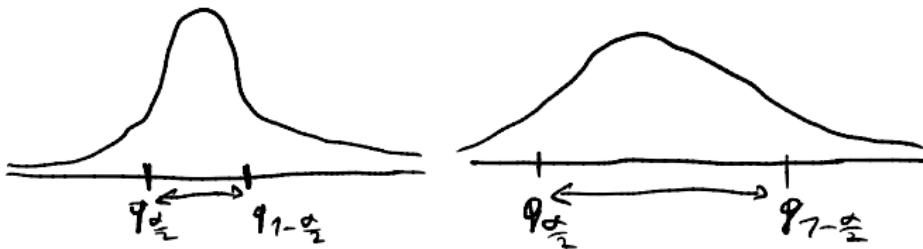
$$P(Y_0 \in PI_\alpha | S_{train}) = P(\hat{q}_{\alpha/2} \leq R^{ss} \leq \hat{q}_{1-\alpha/2} | S_{train}) \approx 1 - \alpha$$

PREDICTIVE INFERENCE BY SAMPLE SPLITTING



- ▶ Sample splitting works very well when n is large relative to p .
- ▶ Otherwise, $\hat{m}_{n_1} : \mathbb{R}^p \rightarrow \mathbb{R}$ may be much less accurate than \hat{m}_n .
- ▶ Recall: We need n_2 large, so $n_1 = n - n_2 \ll n$.

$$Y_0 - \hat{m}_n(X_0) \quad \text{vs.} \quad Y_0 - \hat{m}_{n_1}(X_0)$$



PREDICTIVE INFERENCE WITH THE JACKKNIFE



- ▶ How to estimate/approximate the distribution of the prediction error

$$R := Y_0 - \hat{m}_n(X_0)$$

- ▶ Let $R_i^{l1o} := Y_i - \hat{m}_{(i)}(X_i)$, $i = 1, \dots, n$ where

$$\hat{m}_{(i)}(X_i) = \hat{m}_{n-1}(Z_1, \dots, Z_{i-1}, Z_{i+1}, \dots, Z_n; X_i)$$

is the prediction at X_i of the learning algorithm trained on the data set with the i -th observation pair removed.

- ▶ If $\hat{m}_n \approx \hat{m}_{(i)}$, then, approximately $R_i^{l1o} \approx R$.
- ▶ The $R_1^{l1o}, \dots, R_n^{l1o}$ are (usually) identically distributed but not independent.
- ▶ We still use empirical quantiles $\hat{q}_{\alpha/2}^{l1o}$ and $\hat{q}_{1-\alpha/2}^{l1o}$ to compute...

PREDICTIVE INFERENCE WITH THE JACKKNIFE



- ▶ $R_i^{l1o} := Y_i - \hat{m}_{(i)}(X_i)$, $i = 1, \dots, n$
- ▶ $\hat{q}_{\alpha/2}^{l1o}$ and $\hat{q}_{1-\alpha/2}^{l1o}$ empirical quantiles.

$$PI_{\alpha}^{l1o} = [\hat{m}_n(X_0) + \hat{q}_{\alpha/2}^{l1o}, \hat{m}_n(X_0) + \hat{q}_{1-\alpha/2}^{l1o}]$$

Under some regularity assumptions, one can show

$$\mathbb{E} \left[\left| P(Y_0 \in PI_{\alpha}^{l1o} | Z_1, \dots, Z_n) - (1 - \alpha) \right| \right] \xrightarrow[n,p \rightarrow \infty]{} 0.$$

$$P(Y_0 \in PI_{\alpha}^{l1o}) \approx 1 - \alpha$$

$$| P(Y_0 \in PI_\alpha^{(1)}) - (1-\alpha) |$$

$$= | E[P(Y_0 \in PI_\alpha^{(1)} | Z_1, \dots, Z_n)] - (1-\alpha) |$$

$$= | E \left[P(Y_0 \in PI_\alpha^{(1)} | Z_1, \dots, Z_n) - (1-\alpha) \right] |$$

$$\leq E \left[| P(Y_0 \in PI_\alpha^{(1)} | Z_1, \dots, Z_n) - (1-\alpha) | \right]$$

$$\xrightarrow{\hspace{1cm}} 0$$

$$n, p \rightarrow \infty$$

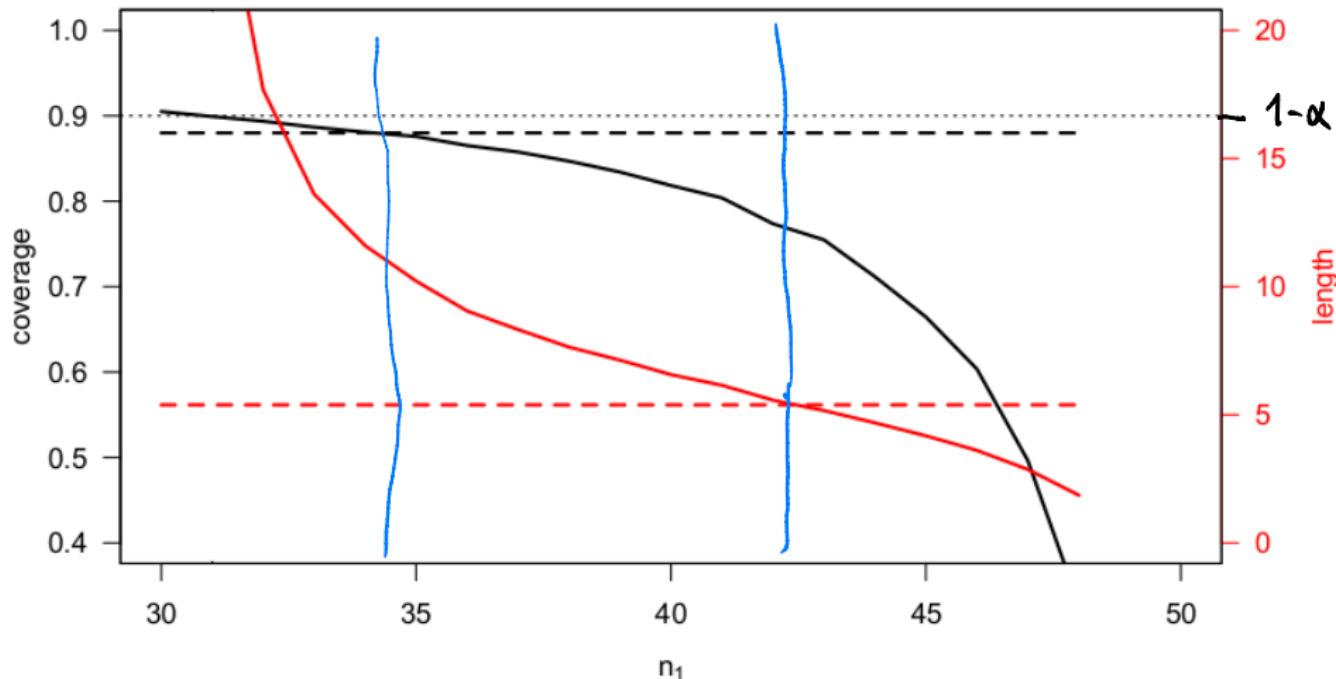
PREDICTIVE INFERENCE:

SAMPLE SPLITTING VS. JACKKNIFE



$$\hat{m}_n(x_1) = x_1^T \hat{\beta}$$
$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$n = 50 \quad p = 30$



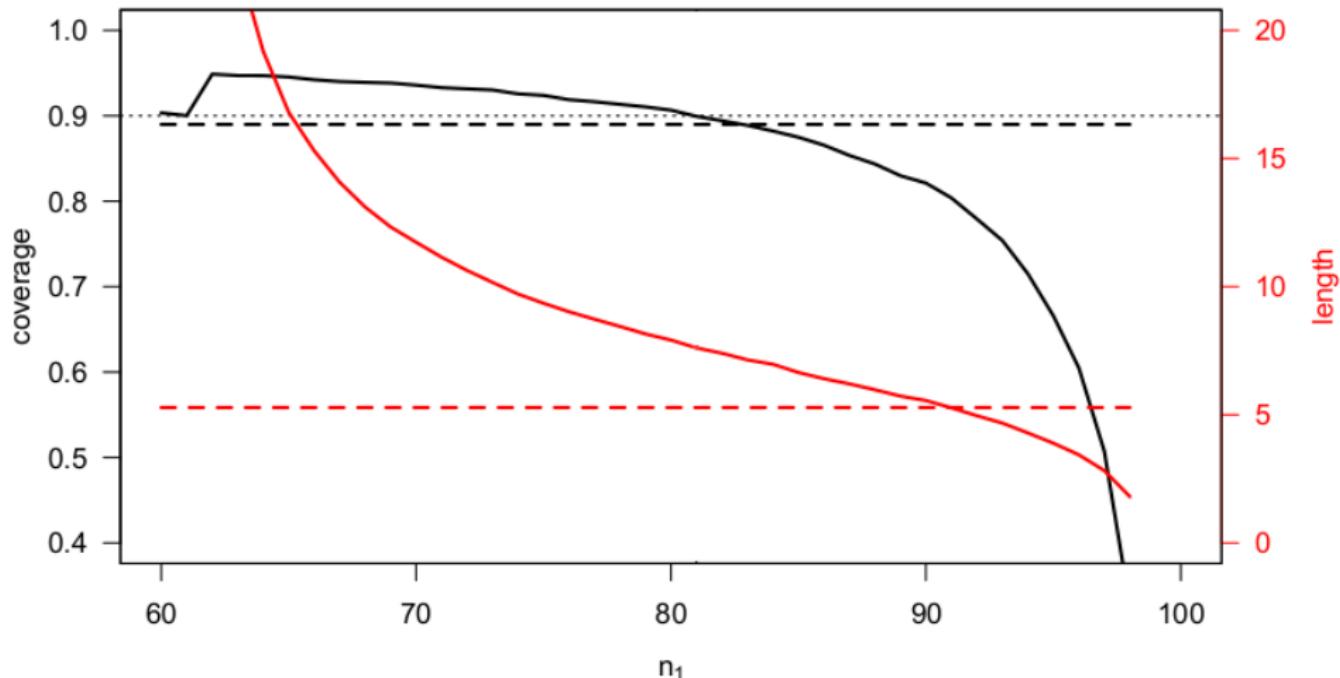
PREDICTIVE INFERENCE:

SAMPLE SPLITTING VS. JACKKNIFE



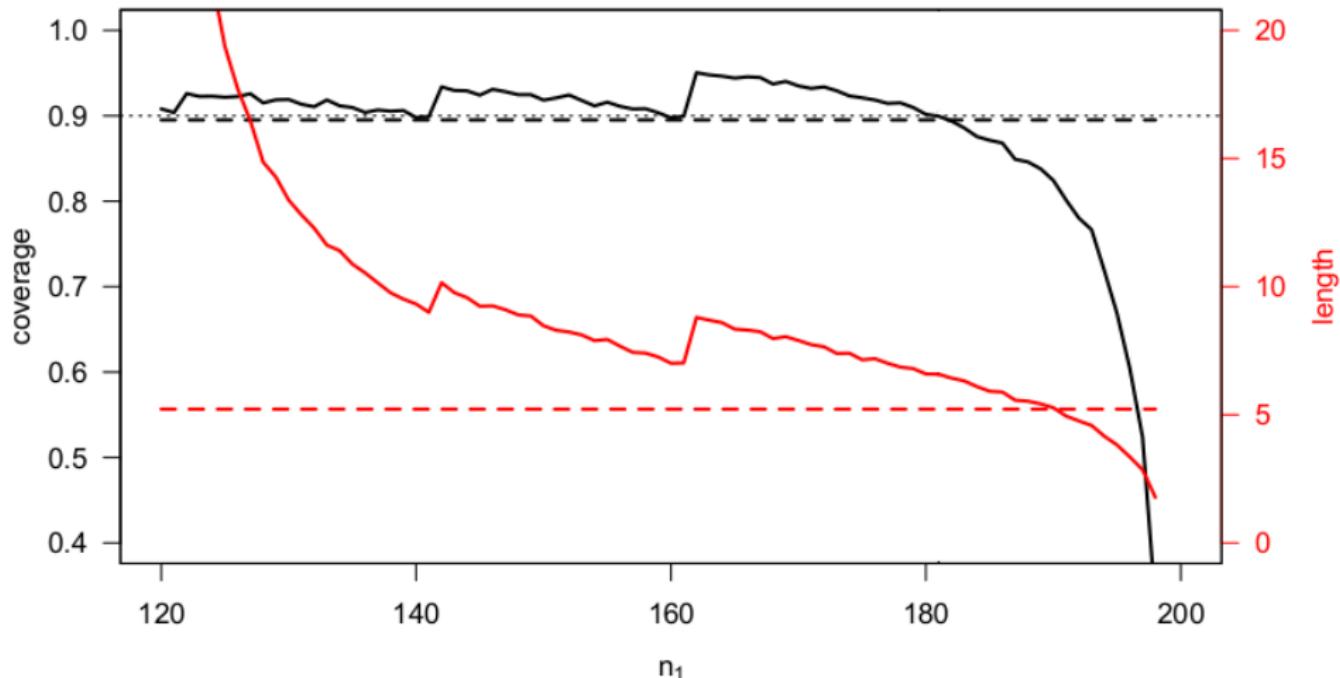
universität
wien

$n = 100 \ p = 60$





SAMPLE SPLITTING VS. JACKKNIFE

 $n = 200 \ p = 120$ 

PREDICTIVE INFERENCE WITH THE JACKKNIFE



universität
wien

Why use leave-one-out residuals

and not simply

$$\left. \begin{array}{l} Y_i - \hat{m}_{(i)}(X_i) \\ Y_i - \hat{m}_n(X_i) \end{array} \right\} \approx \mathcal{Z};$$

Does it make a big difference?

Hw

Would be computationally much cheaper!!!

Statistics for Data Science, WS2023

Chapter 4:

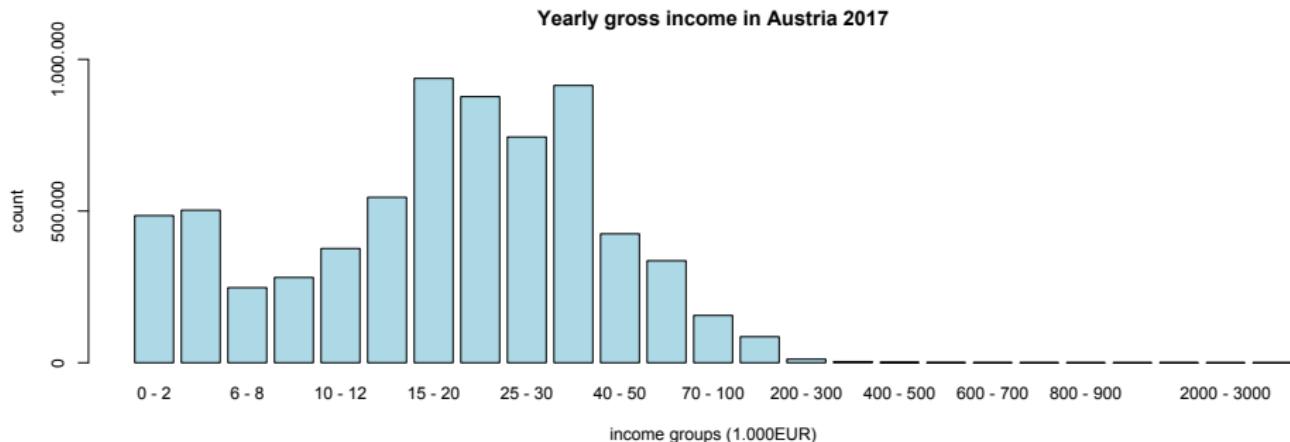
Linear Models and Model Selection

The Gaussian linear model

THE GAUSSIAN LINEAR MODEL: MOTIVATION

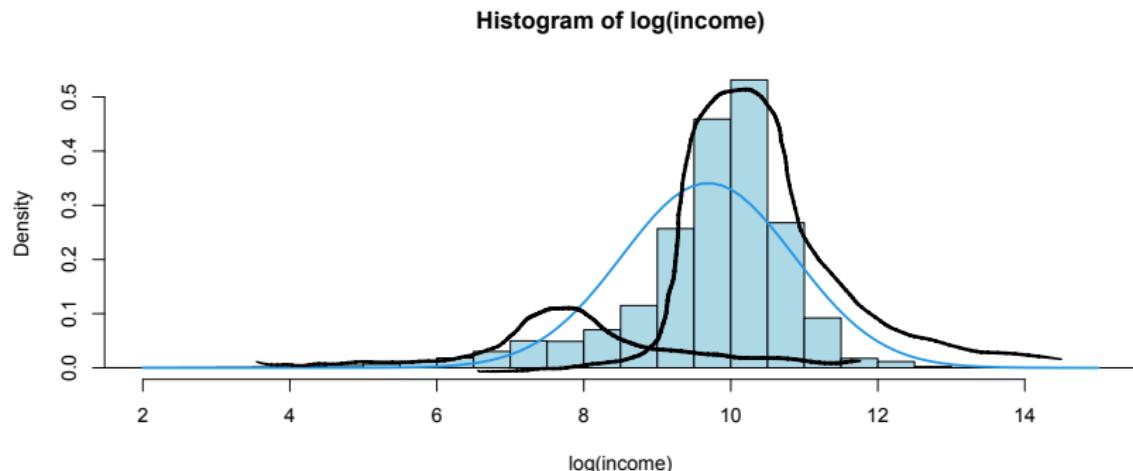


Recall our income example:



What are the most important factors that influence a persons income? How large is the gender pay gap?

Often, data are (nearly) Gaussian after an appropriate transformation.



They shouldn't really be! We are looking at many different sub-populations.

THE GAUSSIAN LINEAR MODEL: MOTIVATION



We want to ‘explain’ the (log) income using other variables,
e.g., gender, age, education, etc.

log income	intercept	gender	age
$Y = \begin{pmatrix} 8.23 \\ 11.54 \\ 10.02 \\ \vdots \\ 7.78 \end{pmatrix}$	$X_{\cdot 1} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$	$X_{\cdot 2} = \begin{pmatrix} 0 \\ 1 \\ 1 \\ \vdots \\ 0 \end{pmatrix}$	$X_{\cdot 3} = \begin{pmatrix} 37 \\ 25 \\ 62 \\ \vdots \\ 18 \end{pmatrix}$

Y is called the **response or dependent** variable.

The $X_{\cdot 1}, \dots, X_{\cdot p}$ are called: **covariates, predictor-, regressor-, explanatory-, feature- or independent** variables.

THE GAUSSIAN LINEAR MODEL



- ▶ $Y \sim N(X\beta, \sigma^2 I_n)$, $\beta \in \mathbb{R}^p$, $\sigma^2 \in (0, \infty)$. $\textcircled{H} = \mathbb{R}^p \times (0, \infty)$
matrix with independent vars.
- ▶ Low-dimensional case: $p < n$
- ▶ X is an $n \times p$ (non-random) design matrix with $\text{rank}(X) = p$ (e.g., analysis conditional on X)

“The mean of Y is assumed to be a linear function of our explanatory variables $X_{.1}, \dots, X_{.p}$, i.e.,

$$\mathbb{E}[Y] = X\beta = \beta_1 X_{.1} + \dots + \beta_p X_{.p} \in \mathbb{R}^n$$

or

$$\mathbb{E}[Y_i] = X_{i \cdot} \beta = \beta_1 X_{i1} + \dots + \beta_p X_{ip} \in \mathbb{R}.$$

“ β_k is the expected change of the response variable when the regressor $X_{.k}$ increases by one unit and all the others stay the same.”

- ▶ $Y \sim N(X\beta, \sigma^2 I_n)$, $\beta \in \mathbb{R}^p$, $\sigma^2 \in (0, \infty)$.
- ▶ Low-dimensional case: $p < n$
- ▶ X is an $n \times p$ (non-random) design matrix with $\text{rank}(X) = p$ (e.g., analysis conditional on X)

Ordinary least squares estimators:

- ▶ $\hat{\beta} := \underset{b \in \mathbb{R}^p}{\text{argmin}} \|Y - Xb\|_2^2 = (X'X)^{-1}X'Y$
- ▶ $\hat{\sigma}^2 := \frac{1}{n-p} \|Y - X\hat{\beta}\|_2^2$
- ▶ $\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$
- ▶ $\frac{\hat{\sigma}^2(n-p)}{\sigma^2} \sim \chi_{n-p}^2$ independent of $\hat{\beta}_n$



$$\hat{\beta} := \operatorname{argmin}_{b \in \mathbb{R}^p} \underbrace{\|Y - Xb\|_2^2}_{=: L(b)}$$

$$\nabla L(b) = \nabla \sum_{i=1}^n (Y_i - X_{i \cdot} b)^2 = -2 \sum_{i=1}^n X'_{i \cdot} (Y_i - X_{i \cdot} b) = -2X'(Y - Xb)$$

$\nabla^2 L(b) = 2X'X$ is positive definite

Normal equations: $-2X'(Y - Xb) = 0 \iff X'Xb = X'Y$

$$\Rightarrow \hat{\beta} = \underbrace{(X'X)^{-1}}_{P \times P} X'Y$$

THE GAUSSIAN LINEAR MODEL: STATISTICAL INFERENCE



$$Y \sim N(X\beta, \sigma^2 I_n), \quad \mathbb{E}[Y] = X\beta = \sum_{k=1}^p \beta_k X_{\cdot k},$$

$$\hat{\beta} = (X'X)^{-1}X'Y \sim N(\beta, \sigma^2(X'X)^{-1})$$

$$\ell_{\hat{\beta}} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \leftarrow \mathbf{q}$$

Want to do statistical inference on individual effects.

E.g.: $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$

$$\hat{\beta}_k = e'_k \hat{\beta} \sim N(e'_k \beta, e'_k [\sigma^2(X'X)^{-1}] e_k) = N(\beta_k, \sigma^2 [(X'X)^{-1}]_k)$$

$$se(\hat{\beta}_k) = \sigma \sqrt{[(X'X)^{-1}]_k}$$

where $[(X'X)^{-1}]_k$ is the k -th diagonal entry of $(X'X)^{-1}$.

One can show that

$$\frac{\hat{\beta}_k - \beta_k}{\hat{\sigma} \sqrt{[(X'X)^{-1}]_k}} \sim t_{n-p}, \quad \text{Student-t distribution}$$

THE GAUSSIAN LINEAR MODEL: STATISTICAL INFERENCE



universität
wien

$$\frac{\hat{\beta}_k - \beta_k}{\hat{\sigma} \sqrt{[(X'X)^{-1}]_k}} \sim t_{n-p}, \quad \text{Student-t distribution}$$

$$T_k := \frac{\hat{\beta}_k - b}{\hat{\sigma} \sqrt{[(X'X)^{-1}]_k}} \sim t_{n-p} \quad \text{e.g. } b = 0$$

under the null hypothesis $H_0 : \beta_k = b$. Thus

$$P_{H_0} \left(|T_k| > q_{1-\frac{\alpha}{2}}^{(t_{n-p})} \right) = 1 - P \left(-q_{1-\frac{\alpha}{2}}^{(t_{n-p})} \leq t_{n-p} \leq q_{1-\frac{\alpha}{2}}^{(t_{n-p})} \right) = \alpha$$

Test: Reject H_0 if $|T_k| > q_{1-\frac{\alpha}{2}}^{(t_{n-p})}$.

THE GAUSSIAN LINEAR MODEL: STATISTICAL INFERENCE



$$\frac{\hat{\beta}_k - \beta_k}{\hat{\sigma} \sqrt{[(X'X)^{-1}]_k}} \sim t_{n-p}, \quad \text{Student-t distribution}$$

Thus, with $\hat{s}e_k := \hat{\sigma} \sqrt{[(X'X)^{-1}]_k}$,

$$\begin{aligned} P\left(\hat{\beta}_k - q_{1-\frac{\alpha}{2}}^{(t_{n-p})} \hat{s}e_k \leq \beta_k \leq \hat{\beta}_k + q_{1-\frac{\alpha}{2}}^{(t_{n-p})} \hat{s}e_k\right) \\ = P\left(-q_{1-\frac{\alpha}{2}}^{(t_{n-p})} \leq \frac{\hat{\beta}_k - \beta_k}{\hat{s}e_k} \leq q_{1-\frac{\alpha}{2}}^{(t_{n-p})}\right) = 1 - \alpha \end{aligned}$$

$$CI_\alpha = \hat{\beta}_k \pm q_{1-\frac{\alpha}{2}}^{(t_{n-p})} \hat{s}e_k$$

Consider $p = 2$:

$$Y \sim N(X\beta, \sigma^2 I_n), \quad \hat{\beta} = (X'X)^{-1} X' Y \sim N(\beta, \sigma^2 (X'X)^{-1})$$

with

$$X = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \quad \left. \begin{array}{c} n_1 \\ n_2 \end{array} \right\} \quad \begin{array}{l} n_1 + n_2 = n \\ n_1 \\ n_2 \end{array}$$

$$X\beta = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \left\{ \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_2 \end{pmatrix} \right\}^{n_1} \quad \left\{ \begin{pmatrix} \beta_2 \\ \vdots \\ \beta_2 \end{pmatrix} \right\}^{n_2}$$

$$Y \sim N(X\beta, \sigma^2 I_n) \Rightarrow Y_1, \dots, Y_{n_1} \stackrel{\text{iid}}{\sim} N(\beta_1, \sigma^2)$$

$$Y_{n_1+1}, \dots, Y_n \stackrel{\text{iid}}{\sim} N(\beta_2, \sigma^2)$$

$$\hat{\beta} = (X'X)^{-1}X'Y = \begin{pmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_2 \end{pmatrix}$$

$$X'X = \begin{pmatrix} 1 & \dots & 1 & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & \dots & 1 \end{pmatrix} X$$

$$(X'X)^{-1} = \begin{pmatrix} \frac{1}{n_1} & 0 \\ 0 & \frac{1}{n_2} \end{pmatrix} = \begin{pmatrix} n_1 & 0 \\ 0 & n_2 \end{pmatrix}$$

$$X'Y = \begin{pmatrix} 1 & \dots & 1 & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} Y_1 \\ \vdots \\ Y_{n_1} \\ \vdots \\ Y_{n_1+1} \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n_1} Y_i \\ \vdots \\ \sum_{i=n_1+1}^n Y_i \end{pmatrix}$$

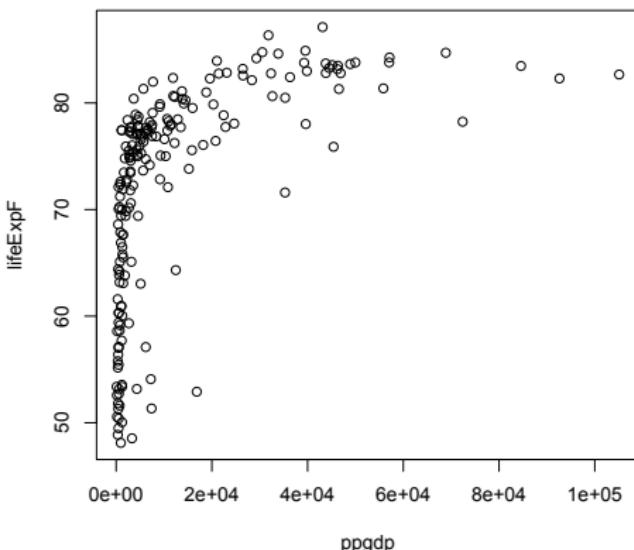
$$\hat{\beta} = \begin{pmatrix} \frac{1}{n_1} & 0 \\ 0 & \frac{1}{n_2} \end{pmatrix} \begin{pmatrix} \sum Y_i \\ \sum Y_i \end{pmatrix} = \begin{pmatrix} \frac{1}{n_1} \sum_{i=1}^{n_1} Y_i \\ \frac{1}{n_2} \sum_{i=n_1+1}^n Y_i \end{pmatrix}$$

EXAMPLE: UN DATA, 2009

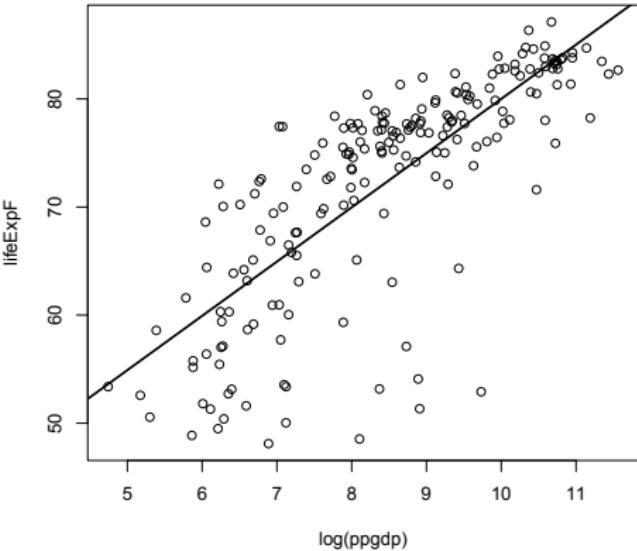
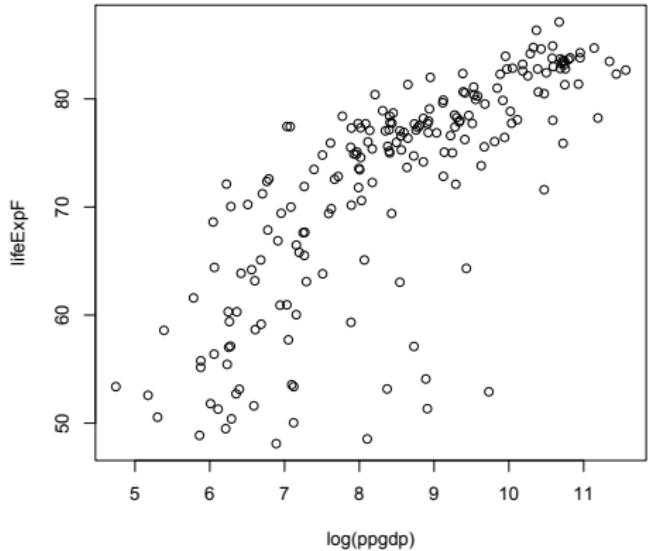


	country	region	group	fertility	ppgdp	lifeExpF	pctUrban
1	Afghanistan	Asia	other	5.968	499.0	49.49	23
2	Albania	Europe	other	1.525	3677.2	80.40	53
3	Algeria	Africa	africa	2.142	4473.0	75.00	67
4	Angola	Africa	africa	5.135	4321.9	53.17	59
5	Anguilla	Caribbean	other	2.000	13750.1	81.10	100
6	Argentina	Latin Amer	other	2.172	9162.1	79.89	93

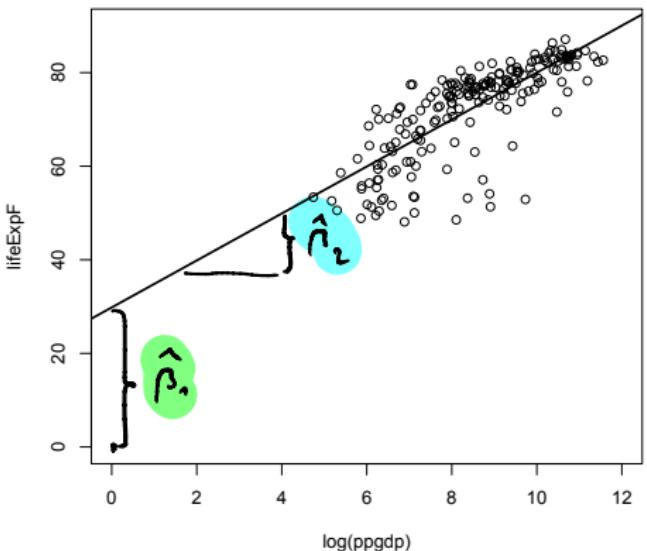
$n = 199$



EXAMPLE: UN DATA



EXAMPLE: UN DATA



$$X = \begin{pmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{n1} \end{pmatrix}$$

$\log(\text{ppgdp})$

$$E[Y_i] = X_i \cdot \beta = \beta_1 + X_{i2} \beta_2$$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
β_1 (Intercept)	29.8148	2.5314	11.78	<2e-16
β_2 log(ppgdp)	5.0188	0.2942	17.06	<2e-16

$\hat{\beta}$

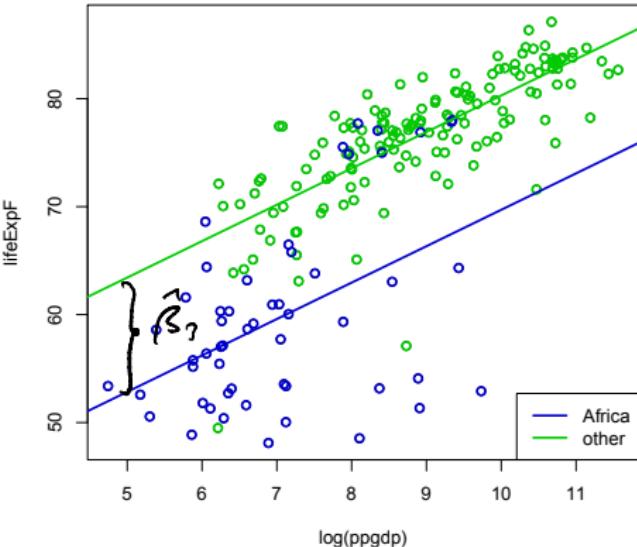
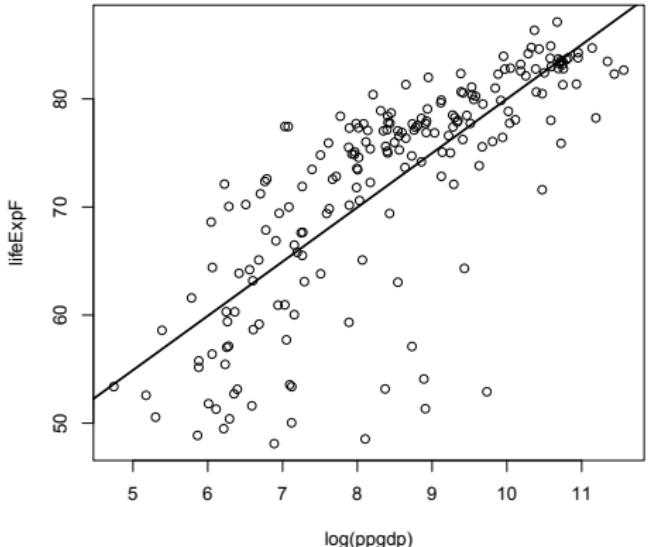
$\hat{s_e}$

T_k

$H_0: \beta_k = 0$



EXAMPLE: UN DATA

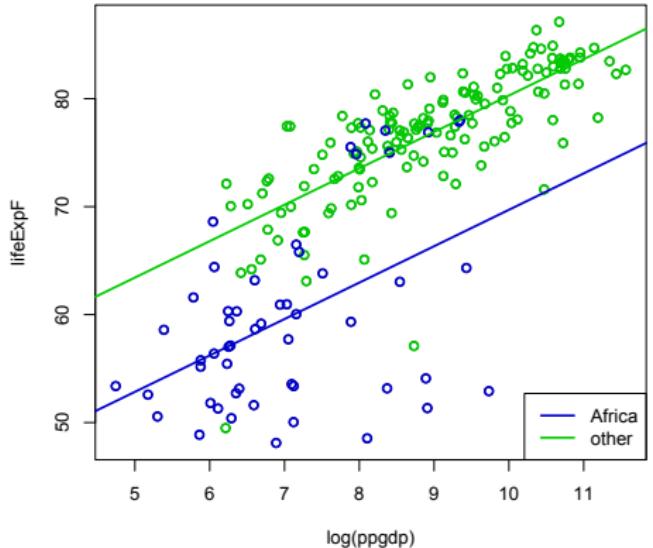


	Estimate	Std. Error	t value	Pr(> t)
--	----------	------------	---------	----------

β_1 (Intercept)	35.9798	2.0889	17.22	<2e-16
β_2 log(ppgdp)	3.3728	0.2788	12.10	<2e-16
β_3 groupother	10.5859	0.9802	10.80	<2e-16



EXAMPLE: UN DATA



$$X = \begin{pmatrix} 1 & CDP_1 & 1 \\ 1 & CDP_2 & 0 \\ \vdots & \vdots & \vdots \\ 1 & CDP_n & 0 \end{pmatrix}$$

group

$$E[Y_i] = X_i \cdot \beta = \beta_1 + X_{i2}\beta_2 + X_{i3}\beta_3$$

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	35.9798	2.0889	17.22	<2e-16
log(ppgdp)	3.3728	0.2788	12.10	<2e-16
groupother	10.5859	0.9802	10.80	<2e-16



EXAMPLE: UN DATA



✓ *Slowp*

$$\mathbb{E}[Y_i] = X_i \cdot \beta = \beta_1 + X_{i2}\beta_2 + X_{i3}\beta_3$$

X_{i3} is the 'group' variable where 0 = Africa, 1 = other. If the i -th country is in Africa we have

$$\mathbb{E}[Y_i] = \beta_1 + X_{i2}\beta_2 + 0,$$

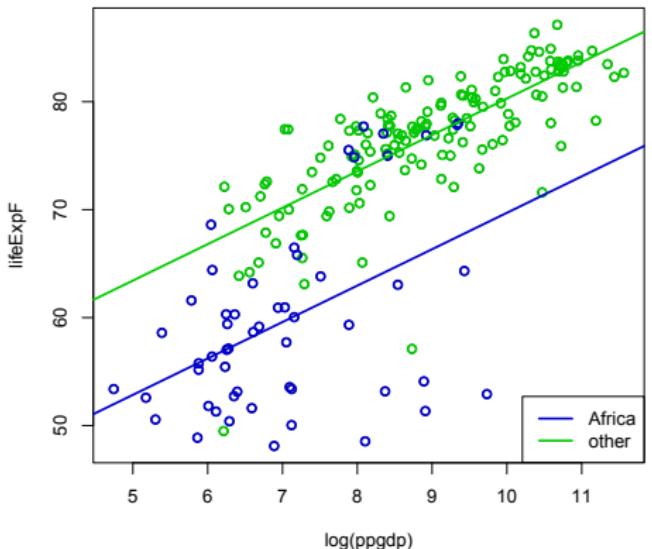
whereas if the j -th country is outside of Africa, we have

$$\mathbb{E}[Y_j] = \beta_1 + X_{j2}\beta_2 + \beta_3.$$

β_3 is the expected additional life expectancy of women in non-African countries, given that $\log(ppgdp)$ is the same ($X_{i2} = X_{j2}$).

test $H_0 : \beta_3 = 0, \beta_1, \beta_2 \in \mathbb{R}, \sigma^2 > 0$

UN DATA: INTERACTION EFFECT

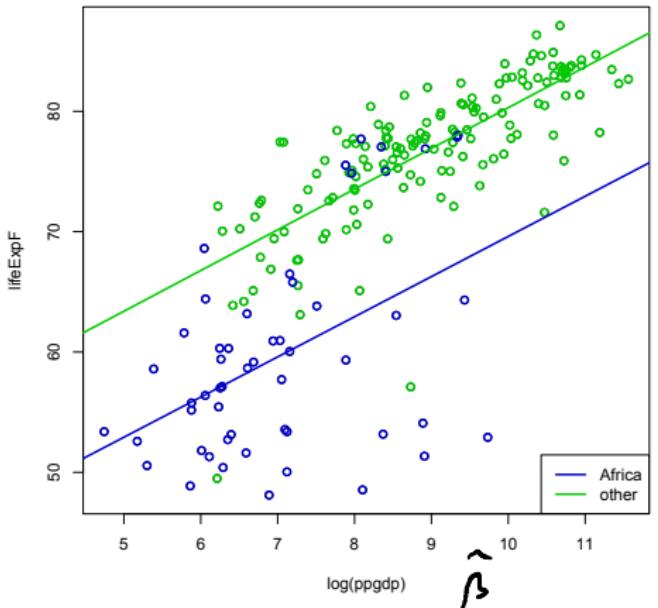


Here we forced the two regression lines to be parallel!

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	35.9798	2.0889	17.22	<2e-16
log(ppgdp)	3.3728	0.2788	12.10	<2e-16
groupother	10.5859	0.9802	10.80	<2e-16



UN DATA: INTERACTION EFFECT



$$X = \begin{pmatrix} 1 & GDP_1 & 0 & 0 \\ 1 & GDP_2 & 1 & GDP_1 \\ 1 & GDP_3 & 0 & 0 \\ \vdots & & & \\ 1 & GDP_n & 1 & GDP_n \end{pmatrix}$$

Africa:

$$\mathbb{E}[Y_i] = \beta_1 + X_{i2}\beta_2$$

other:

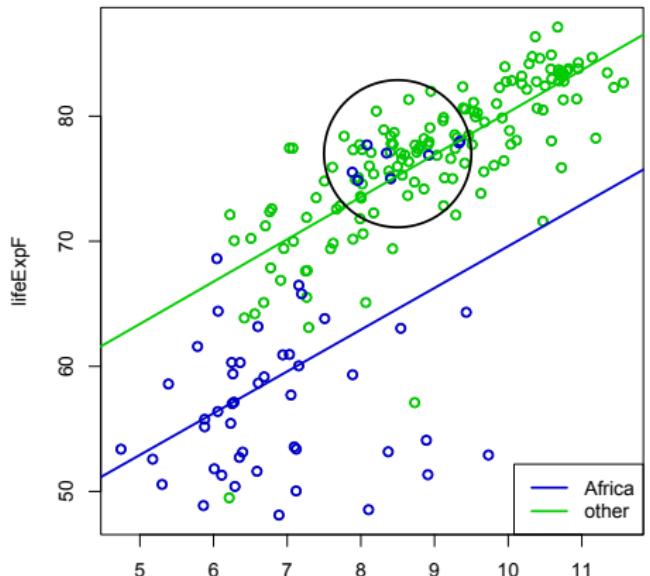
$$\mathbb{E}[Y_j] = \beta_1 + X_{j2}\beta_2 + \beta_3 + X_{j2}\beta_4 = \beta_1 + \beta_3 + X_{j2}(\beta_2 + \beta_4)$$

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	36.22882	4.18145	8.664	1.73e-15
log(ppgdp)	3.33752	0.58428	5.712	4.12e-08
groupother	10.24281	5.08235	2.015	0.0452
log(ppgdp):groupother	0.04578	0.66539	0.069	0.9452

Test: Are the slopes different?

$$H_0: \beta_4 = 0$$



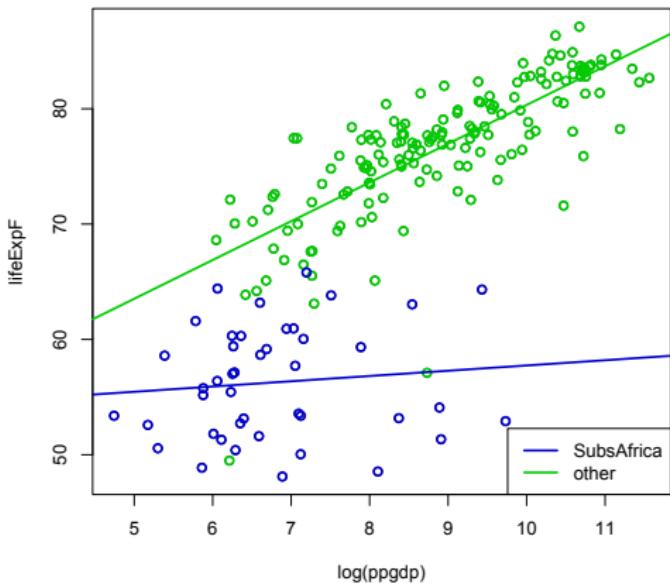


Algeria, Cape Verde, Egypt, Libya,
Mauritius, Morocco, Seychelles,
Tunisia

North African countries or islands!

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	36.22882	4.18145	8.664	1.73e-15
log(ppgdp)	3.33752	0.58428	5.712	4.12e-08
groupother	10.24281	5.08235	2.015	0.0452
log(ppgdp):groupother	0.04578	0.66539	0.069	0.9452





sub-Saharan Africa:

$$\mathbb{E}[Y_i] = \beta_1 + X_{i2}\beta_2$$

other:

$$\mathbb{E}[Y_j] = \beta_1 + X_{j2}\beta_2 + \beta_3 + X_{j2}\beta_4 = \beta_1 + \beta_3 + X_{j2}(\beta_2 + \beta_4)$$

Test: Are the slopes different?

$$H_0: \beta_4 = 0$$

Does GDP actually influence female life expectancy in sub-Saharan Africa?

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	β_1	53.4042	3.9035	13.681 < 2e-16
log(ppgdp)	β_2	0.4330	0.5656	0.766 0.4448
groupSSother	β_3	-7.9811	4.4871	-1.779 0.0769
log(ppgdp):groupSSother	β_4	3.0627	0.6167	4.967 1.48e-06

$$\alpha = 0.05 \quad \delta = 0.01$$





Looking at the data and then changing the model or the question to ask is actually not allowed!!

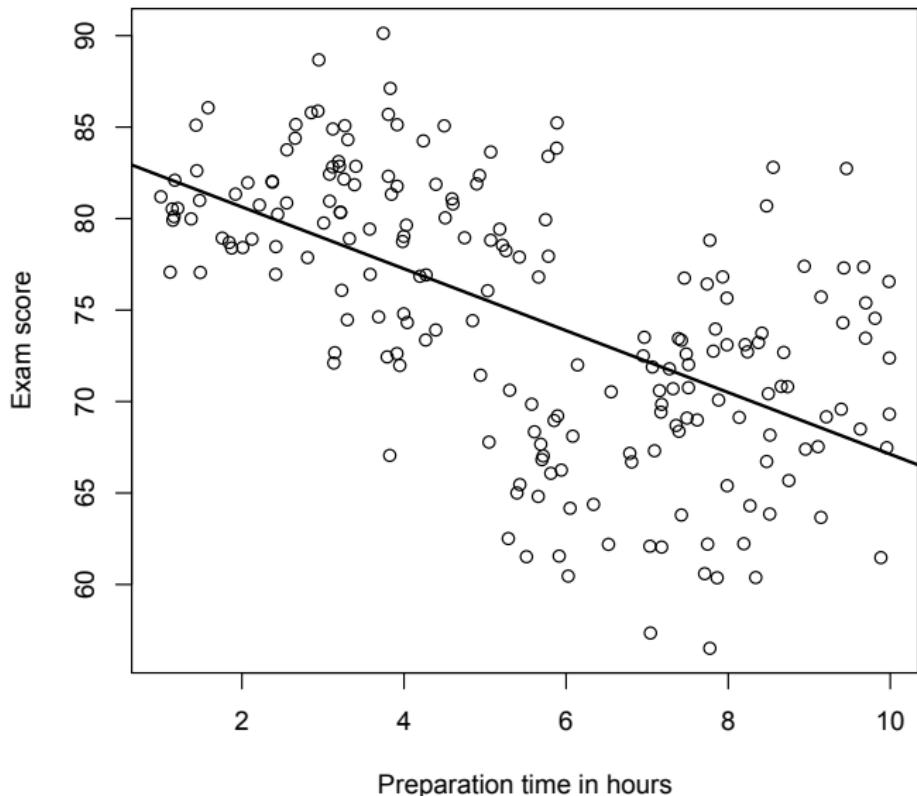
More on that later...

Exploratory data analysis vs. statistical inference!

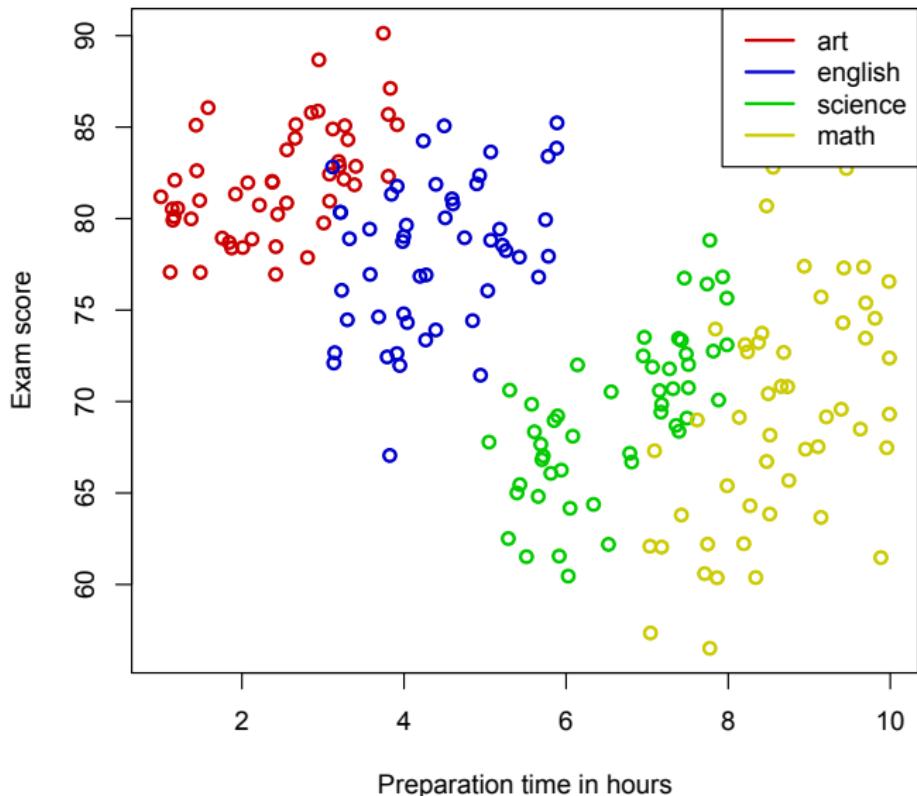


Philosophical question: What actually is the population here?

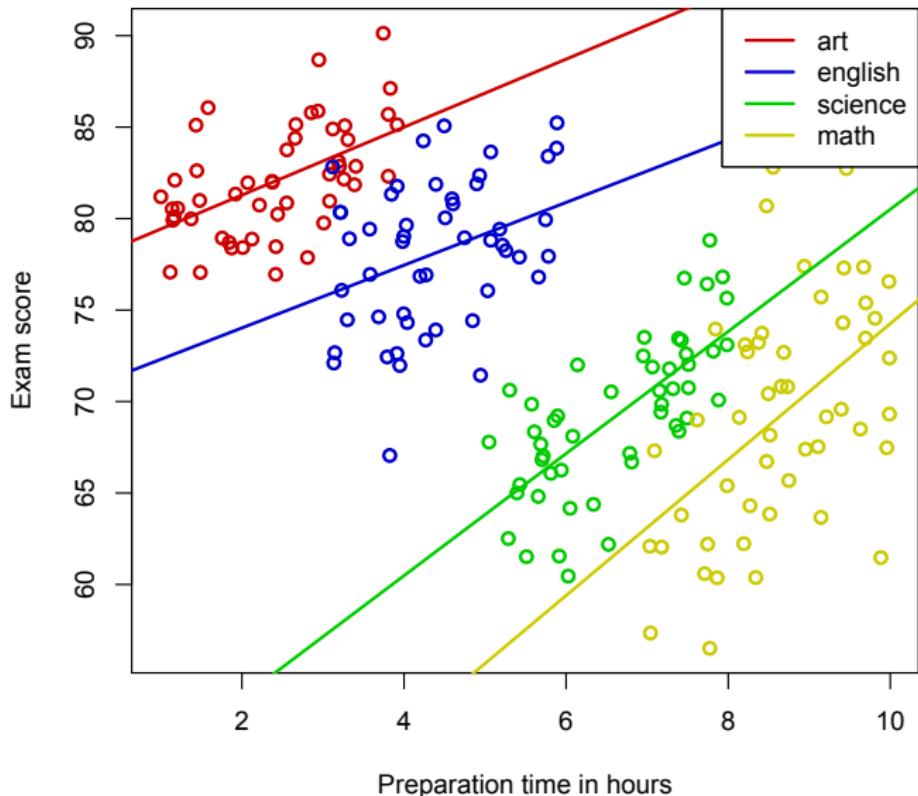
SIMPSON'S PARADOX



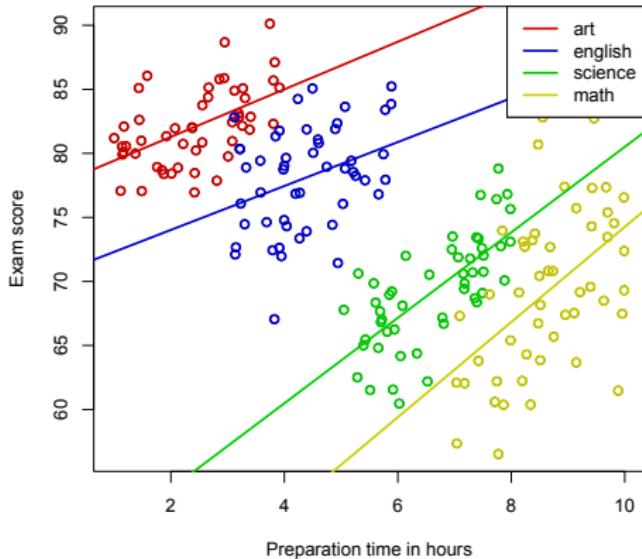
SIMPSON'S PARADOX



SIMPSON'S PARADOX



CATEGORICAL VARIABLES AKA FACTORS



Are the four slopes actually different?

Here 'subject' is a categorical variable (*factor*) with **four levels**. How to model that, i.e., how to construct X ?

CATEGORICAL VARIABLES AKA FACTORS



How to model a factor variable?

1.) Use a code: art=1, english=2, science=3, math=4.

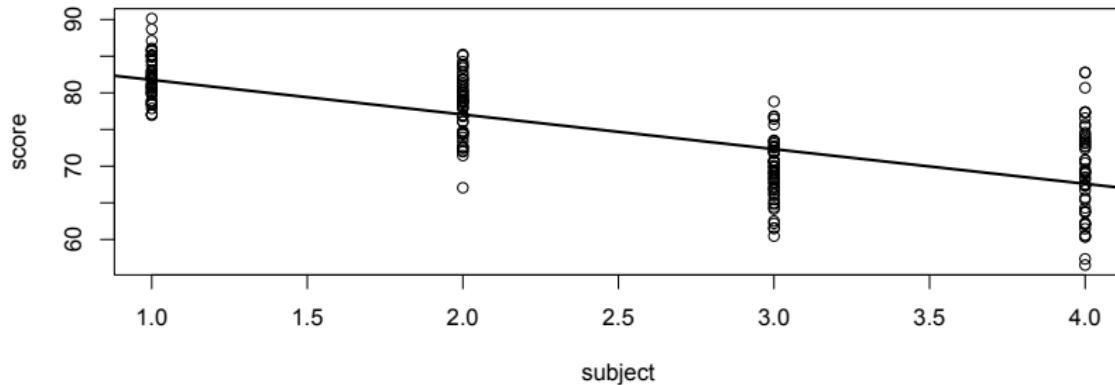
$$X = \begin{pmatrix} 1 & 2.5 & 1 \\ 1 & 3.5 & 2 \\ 1 & 6 & 4 \\ 1 & 7.5 & 3 \\ 1 & 5 & 3 \\ 1 & 1 & 1 \\ 1 & 1.5 & 2 \\ \vdots & & \\ 1 & 9 & 3 \end{pmatrix}$$

time → ← *subject*

CATEGORICAL VARIABLES AKA FACTORS



1a.) Use a code: art=1, english=2, science=3, math=4.

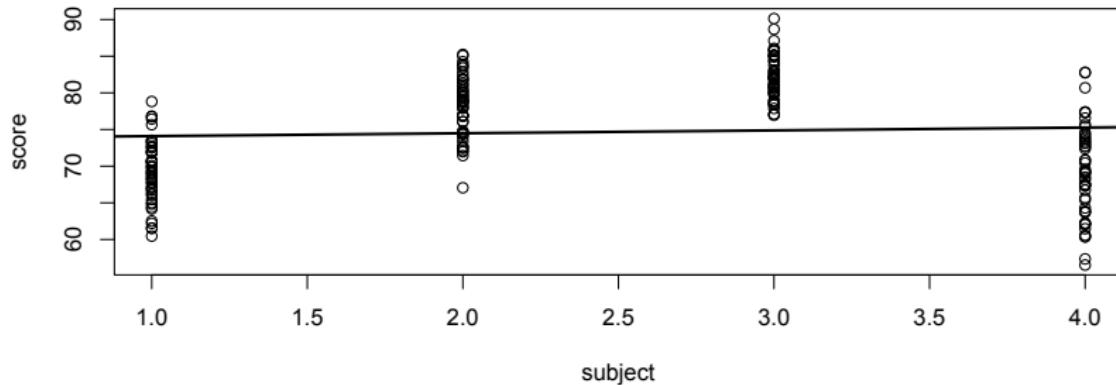


	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	86.4877	0.8512	101.6	<2e-16
subj	-4.7225	0.3108	-15.2	<2e-16

CATEGORICAL VARIABLES AKA FACTORS



1b.) Use a different code: art=3, english=2, science=1, math=4.



	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	73.7040	1.2504	58.946	<2e-16
subj	0.3911	0.4566	0.856	0.393

CATEGORICAL VARIABLES AKA FACTORS



2.) Introduce 'dummy variables'.

$$X_{.1} + X_{.4} + X_{.5} + X_{.6} = X_{.1}$$

$$X = \begin{pmatrix} 1 & 2.5 & 1 & 0 & 0 & 0 & 2.5 & 0 & 0 & 0 \\ 1 & 3.5 & 0 & 1 & 0 & 0 & 0 & 3.5 & 0 & 0 \\ 1 & 6 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 6 \\ 1 & 7.5 & 0 & 0 & 1 & 0 & 0 & 0 & 7.5 & 0 \\ 1 & 5 & 0 & 0 & 1 & 0 & 0 & 0 & 5 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1.5 & 0 & 1 & 0 & 0 & 0 & 1.5 & 0 & 0 \\ \vdots & \vdots & & & & & \vdots & & & \\ 1 & 9 & 0 & 0 & 1 & 0 & 0 & 0 & 9 & 0 \end{pmatrix}$$

time ↗
 sat ↗
 english ↗
 science ↗
 math ↗



Test if the slopes are all the same, i.e. $H_0 : \beta_7 = \beta_8 = \beta_9 = \beta_{10}$

CATEGORICAL VARIABLES

AKA FACTORS



$$\text{out: } E(Y_i) = \beta_1 + X_{i2} \beta_2$$

$$3a.) \text{ Avoiding the dummy variable trap: engl.: } E(Y_i) = \beta_1 + X_{i2} \beta_2 + \beta_3 +$$

$$X = \left(\begin{array}{ccccccc} 1 & 2.5 & 0 & 0 & 0 & 0 & 0 \\ 1 & 3.5 & 1 & 0 & 0 & 3.5 & 0 \\ 1 & 6 & 0 & 0 & 1 & 0 & 0 \\ 1 & 7.5 & 0 & 1 & 0 & 0 & 7.5 \\ 1 & 5 & 0 & 1 & 0 & 0 & 5 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1.5 & 1 & 0 & 0 & 1.5 & 0 \\ \vdots & \vdots & & & & \vdots & \\ 1 & 9 & 0 & 1 & 0 & 0 & 9 \end{array} \right) \quad \begin{aligned} &+ X_{i2} \beta_6 \\ &= \beta_1 + \beta_3 + X_{i2} (\beta_2 + \beta_6) \\ \text{math: } &E(Y_i) = \beta_1 + X_{i2} \beta_2 + \beta_5 \\ &+ X_{i2} \beta_8 \\ &= \beta_1 + \beta_5 + X_{i2} (\beta_2 + \beta_8) \end{aligned}$$

english ↑ ↑ math
 france ↘

Test if the slopes are all the same, i.e., $H_0 : \beta_6 = \beta_7 = \beta_8 = 0$.

CATEGORICAL VARIABLES AKA FACTORS



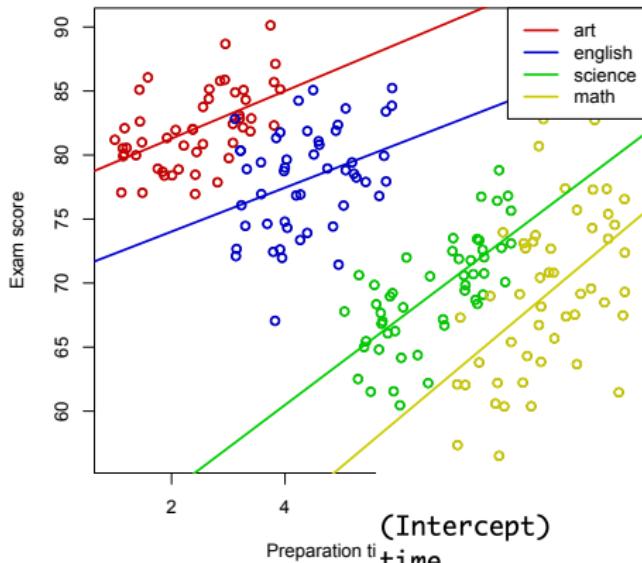
universität
wien

3b.) Avoiding the *dummy variable trap*:

$$X = \begin{pmatrix} 1 & 0 & 0 & 0 & 2.5 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 3.5 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 6 \\ 0 & 0 & 1 & 0 & 0 & 0 & 7.5 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 5 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1.5 & 0 & 0 \\ \vdots & & & & & \vdots & & \\ 0 & 0 & 1 & 0 & 0 & 0 & 9 & 0 \end{pmatrix}$$

Test if the slopes are all the same, i.e., $H_0 : \beta_5 = \beta_6 = \beta_7 = \beta_8$.

CATEGORICAL VARIABLES AKA FACTORS



Are the four slopes actually different?

$$H_0 : \beta_6 = \beta_7 = \beta_8 = 0$$

$$\alpha = 0.05$$

Can we test this
only by looking
at the standard
output? **no!**

	(Intercept)	time	subjectenglish	subjectmath	subjectscience	β_6 time:subjectenglish	β_7 time:subjectmath	β_8 time:subjectscience
β_0	77.5733	1.6652	46.585	< 2e-16				
β_1	1.8569	0.6487	2.863	0.00467				
	-6.9800	3.4766	-2.008	0.04608				
	-40.4671	5.9648	-6.784	1.41e-10				
	-30.4375	4.5206	-6.733	1.87e-10				
	-0.1419	0.9432	-0.150	0.88053				
	1.8583	0.9243	2.010	0.04579				
	1.4801	0.9026	1.640	0.10270				

	Estimate	Std. Error	t value	Pr(> t)
β_0	77.5733	1.6652	46.585	< 2e-16
β_1	1.8569	0.6487	2.863	0.00467
	-6.9800	3.4766	-2.008	0.04608
	-40.4671	5.9648	-6.784	1.41e-10
	-30.4375	4.5206	-6.733	1.87e-10
	-0.1419	0.9432	-0.150	0.88053
	1.8583	0.9243	2.010	0.04579
	1.4801	0.9026	1.640	0.10270

We want to test a general linear hypothesis:

$$H_0 : R\beta = r$$

where $R \in \mathbb{R}^{q \times p}$, $q \leq p$, $\text{rank } R = q$ and $r \in \mathbb{R}^q$.

$$R\beta = \begin{pmatrix} \beta_6 \\ \beta_7 \\ \beta_8 \end{pmatrix}$$

For instance, with

$$R = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_6 \\ \vdots \\ \beta_8 \end{pmatrix} \quad \text{and} \quad r = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

we have

$$H_0 : R\beta = r \iff H_0 : \beta_6 = 0, \beta_7 = 0, \beta_8 = 0$$

We want to test a general linear hypothesis:

$$H_0 : R\beta = r$$

where $R \in \mathbb{R}^{q \times p}$, $q \leq p$, $\text{rank } R = q$ and $r \in \mathbb{R}^q$.

$$R\beta = \begin{pmatrix} \beta_5 - \beta_6 \\ \beta_6 - \beta_7 \\ \beta_7 - \beta_8 \end{pmatrix}$$

Or, with

$$R = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix} \quad \text{and} \quad r = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

we have

$$H_0 : R\beta = r \iff H_0 : \beta_5 = \beta_6, \beta_6 = \beta_7, \beta_7 = \beta_8$$

$$\beta_5 = \beta_6 = \beta_7 = \beta_8$$



$$H_0 : R\beta = r$$

Notice:

$$\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1}) \Rightarrow R\hat{\beta} - r \sim N(\underbrace{R\beta - r}_{= 0} \underbrace{\sigma^2 R(X'X)^{-1} R'}_{\text{under } H_0})$$

Under H_0 , we therefore have

$$\frac{[R(X'X)^{-1}R']^{-1/2}(R\hat{\beta} - r)}{\sigma} \sim N(0, I_q)$$

and

$$\frac{(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)}{\sigma^2} \sim \chi_q^2.$$



$$H_0 : R\beta = r$$

$$\frac{(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)}{\sigma^2} \sim \chi_q^2.$$

Recall, $\hat{\sigma}^2 := \frac{1}{n-p} \|Y - X\hat{\beta}\|_2^2$ satisfies $\frac{\hat{\sigma}^2(n-p)}{\sigma^2} \sim \chi_{n-p}^2$ independent of $\hat{\beta}_n$.

Definition:

If $S_1 \sim \chi_{d_1}^2$ independent of $S_2 \sim \chi_{d_2}^2$, then $\frac{S_1/d_1}{S_2/d_2} \sim F_{d_1, d_2}$ follows an F -distribution with d_1 and d_2 degrees of freedom.

Hence, under H_0 ,

$$F := \frac{(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)/q}{\hat{\sigma}^2} \sim F_{q, n-p}$$

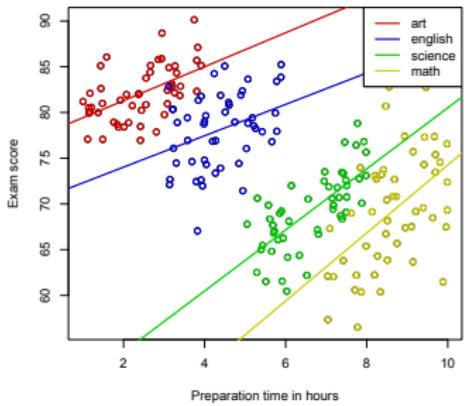
$$H_0 : R\beta = r$$

Under H_0 ,

$$F := \frac{(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)/q}{\hat{\sigma}^2} \sim F_{q,n-p}$$

Test: Reject H_0 if $F > c_\alpha := q_{1-\alpha}^{(F_{q,n-p})}$.

Hence, $P_{H_0}(F > c_\alpha) = 1 - P_{H_0}(F \leq c_\alpha) = 1 - (1 - \alpha) = \alpha$.



Are the four slopes actually different?

$$H_0 : \beta_6 = \beta_7 = \beta_8 = 0 \iff H_0 : R\beta = r$$

$$R = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, r = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\alpha = 0.05$$

$$q$$

$$n-p$$

$$F = 2.3877, \text{ df1} = 3, \text{ df2} = 192, \text{ p-value} = 0.07029$$

Is there a contradiction with the marginal t-tests?

What if we used another R matrix?

H_W

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	77.5733	1.6652	46.585	< 2e-16
time	1.8569	0.6487	2.863	0.00467
subjectenglish	-6.9800	3.4766	-2.008	0.04608
subjectmath	-40.4671	5.9648	-6.784	1.41e-10
subjectscience	-30.4375	4.5206	-6.733	1.87e-10
time:subjectenglish	-0.1419	0.9432	-0.150	0.88053
time:subjectmath	1.8583	0.9243	2.010	0.04579
time:subjectscience	1.4801	0.9026	1.640	0.10270

$$H_0:$$

$$\beta_6 = \beta_7 = \beta_8 = 0$$



Test $H_0 : \beta_6 = \beta_7 = \beta_8 = 0$ at level $\alpha = 0.05$.

$F = 2.3877$, $df1 = 3$, $df2 = 192$, p-value = 0.07029

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	77.5733	1.6652	46.585	< 2e-16
time	1.8569	0.6487	2.863	0.00467
subjectenglish	-6.9800	3.4766	-2.008	0.04608
subjectmath	-40.4671	5.9648	-6.784	1.41e-10
subjectscience	-30.4375	4.5206	-6.733	1.87e-10
time:subjectenglish	-0.1419	0.9432	-0.150	0.88053
time:subjectmath	1.8583	0.9243	2.010	0.04579
time:subjectscience	1.4801	0.9026	1.640	0.10270

} $\leq \alpha$?

Is there a contradiction with the marginal t-tests?

$\leq \frac{\alpha}{3}$?

$$\alpha = 0.05 \rightarrow \frac{\alpha}{3} = 0.0166$$



$$H_0: \beta_6 = \beta_7 = \beta_8 = 0$$

$$\varphi_a(y) := \begin{cases} 1, & \text{if } \varphi_a^{(k)}(y) = 1 \text{ for some } k \in \{6, 7, 8\} \\ 0, & \text{else} \end{cases}$$

where $\varphi_a^{(k)}(y) = \begin{cases} 1, & \text{if } |T_k(y)| > q_{1-\frac{\alpha}{2}}^{(t_{n-p})} \\ 0, & \text{else} \end{cases}$

is the t-test for $H_0^{(k)}: \beta_k = 0$ which satisfies

$$P_{H_0^{(k)}}(\varphi_a^{(k)} = 1) = \alpha.$$

Is φ_a a level- α test for H_0 ?

$$P_{H_0}(\varphi_\alpha = 1) \leftarrow P_{H_0}(\varphi_\alpha^{(G)} = 1 \cup \text{or } \varphi_\alpha^{(7)} = 1 \cup \text{or } \varphi_\alpha^{(8)} = 1)$$

$$\leq \underbrace{\sum_{k=G}^8 P_{H_0}(\varphi_\alpha^{(k)} = 1)}_{= \lambda \text{ because } H_0 \subseteq H_0^{(k)}} = 3 \cdot \lambda$$

For many X matrices we can have

$$P_{H_0}(\varphi_\alpha = 1) > \lambda.$$

note: $P_{H_0}(\varphi_{\alpha_{1/2}} = 1) \leq 3 \cdot \frac{\lambda}{3} = \lambda$

Theorem (Bonferroni correction)

Let $(\mathcal{X}, \Theta, \{f_\theta : \theta \in \Theta\})$ be a statistical model, $\alpha \in (0, 1)$ and $\Theta_0^{(1)}, \dots, \Theta_0^{(K)} \subseteq \Theta$ be hypotheses with corresponding level- α tests $\varphi_\alpha^{(1)}, \dots, \varphi_\alpha^{(K)}$, i.e., for every $k \in \{1, \dots, K\}$, we have $P_\theta(\varphi_\alpha^{(k)} = 1) \leq \alpha$ for all $\theta \in \Theta_0^{(k)}$. Then the test

$$\varphi_\alpha(x) := \begin{cases} 1, & \text{if } \exists k \in \{1, \dots, K\} : \varphi_\alpha^{(k)}(x) = 1, \\ 0, & \text{else,} \end{cases}$$

satisfies

$$P_\theta(\varphi_{\frac{\alpha}{K}} = 1) \leq \alpha \quad \forall \theta \in \Theta_0 := \bigcap_{k=1}^K \Theta_0^{(k)}.$$

MULTIPLE TESTING



(Intercept)	2.299346	1.307025	1.759	0.08010
Xpopulation	-3.477014	2.323128	-1.497	0.13608
Xhouseholdszie	1.118523	0.598059	1.870	0.06294
XracePctblack	-0.042317	0.205327	-0.206	0.83693
XracePctWhite	0.022103	0.290556	0.076	0.93944
XracePctAsian	-0.084169	0.139117	-0.605	0.54587
XracePctHisp	0.151692	0.204243	0.743	0.45855
XagePct12t21	-0.239641	0.598155	-0.401	0.68913
XagePct12t29	-0.870986	0.773690	-1.126	0.26165
XagePct16t24	0.381891	0.895346	0.427	0.67019
XagePct65up	0.199093	0.585702	0.340	0.73428
XnumUrban	3.802921	2.348860	1.619	0.10704
XpctUrban	-0.199798	0.167798	-1.191	0.23521
XmedIncome	-0.386196	0.864797	-0.447	0.65568
XpctWWage	-0.596946	0.551386	-1.083	0.28030
XpctWFarmSelf	-0.308686	0.148121	-2.084	0.03846
XpctWInvInc	-0.751876	0.291920	-2.576	0.01074
XpctWSocSec	-0.721007	0.574259	-1.256	0.21078
XpctWPubAsst	0.238305	0.214464	1.111	0.26786
XpctWRetire	0.064647	0.194944	0.332	0.74053
XmedFamInc	0.871977	0.641222	1.360	0.17543
XperCapInc	-0.876057	0.690411	-1.269	0.20598
XwhitePerCap	0.355919	0.418602	0.850	0.39622
XblackPerCap	-0.302640	0.226430	-1.337	0.18291
XindianPerCap	0.016152	0.120405	0.134	0.89342
XAsianPerCap	-0.147430	0.123883	-1.190	0.23546
XOtherPerCap	0.137916	0.128604	1.072	0.28486
XHispanicPerCap	-0.072551	0.146598	-0.495	0.62123
XNumUnderPov	-0.392185	0.314491	-1.247	0.21387
XPctPopUnderPov	0.525168	0.363595	1.444	0.15023
XPctLess9thGrade	-0.110777	0.317642	-0.349	0.72765
XPctNotHSGrad	-0.696948	0.439727	-1.585	0.11459
XPctBSorMore	-0.376918	0.385047	-0.979	0.32884

Looking at an output like that and asking, *is there something significant*, is not a good idea!

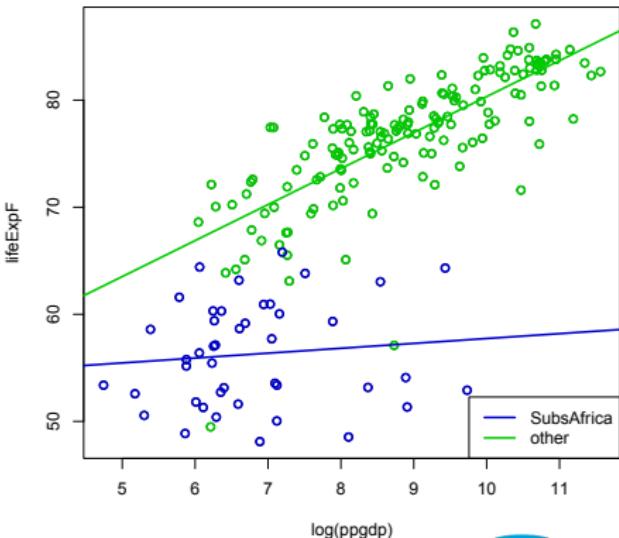
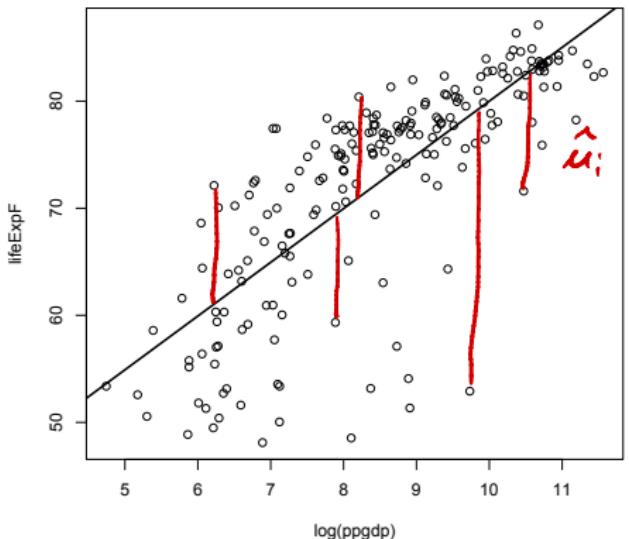
$$H_0 : \forall k : \beta_k = 0$$
$$H_1 : \exists k : \beta_k \neq 0$$

What's the probability of finding something even though there is nothing?

MODEL DIAGNOSTIC PLOTS



Gaussian linear Model: $Y \sim N(X\beta, \sigma^2 I_n)$



What if the model is too complex to visualize it like that? Can we still get a graphical representation of model fit?



MODEL DIAGNOSTIC PLOTS



Gaussian linear Model:

$$Y \sim N(X\beta, \sigma^2 I_n)$$

or, equivalently

$$Y_i = X_{i \cdot} \beta + u_i, \quad \text{with} \quad u = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} \sim N(0, \sigma^2 I_n)$$

Recall: OLS

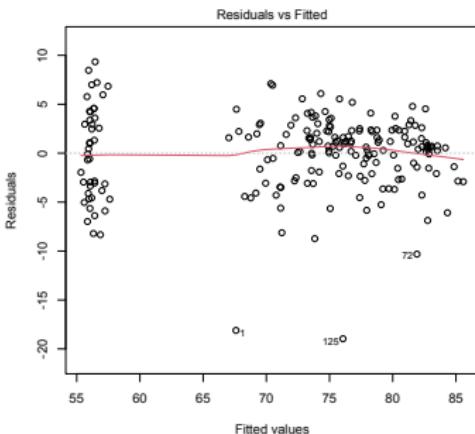
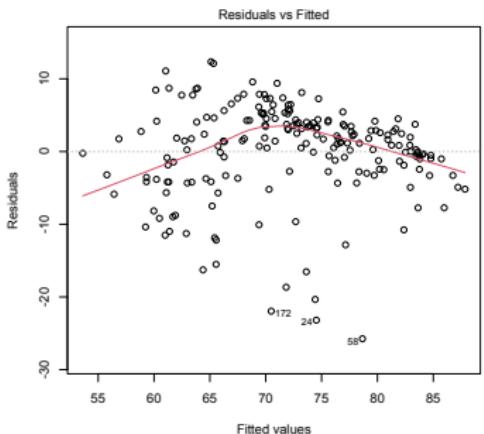
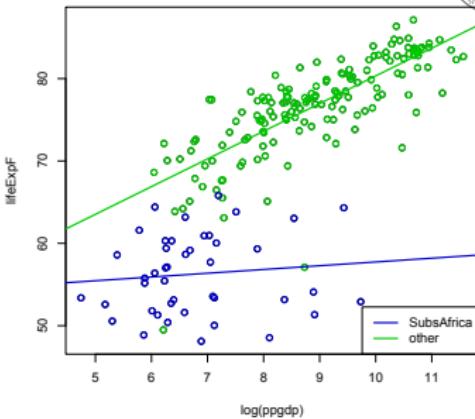
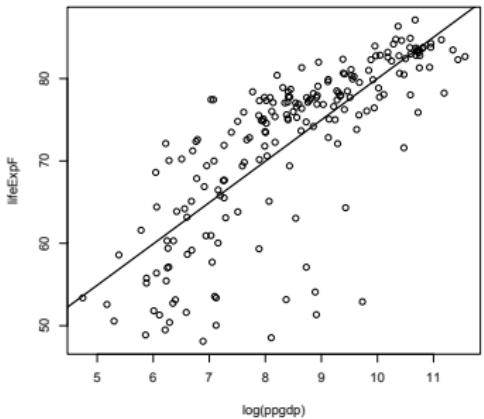
$$\hat{\beta} = \operatorname{argmin}_{b \in \mathbb{R}^p} \|Y - Xb\|_2^2 = \operatorname{argmin}_{b \in \mathbb{R}^p} \sum_{i=1}^n (Y_i - X_{i \cdot} b)^2$$

Idea: The *residuals* $\hat{u}_i = Y_i - X_{i \cdot} \hat{\beta}$ should be (approximately) iid Gaussian and evenly distributed around the regression line (independent of their 'location').

$\hat{Y}_i = X_{i \cdot} \hat{\beta}$ are the *fitted* or *predicted* values 'on' the regression line.

- 1.) plot \hat{Y}_i against \hat{u}_i
- 2.) check \hat{u}_i for Gaussianity

RESIDUAL PLOT



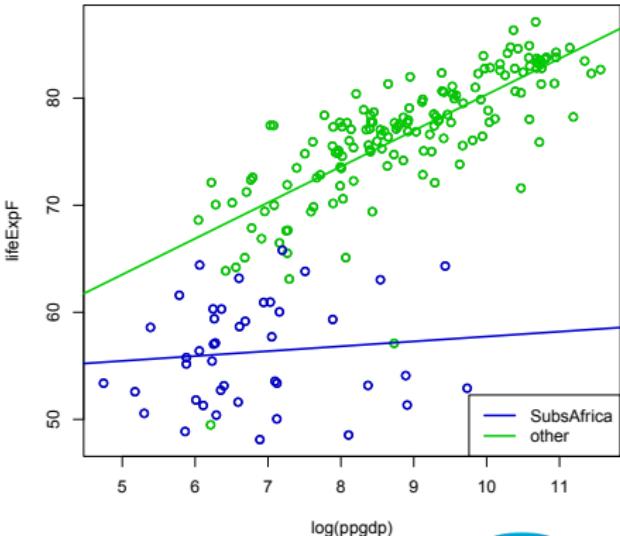
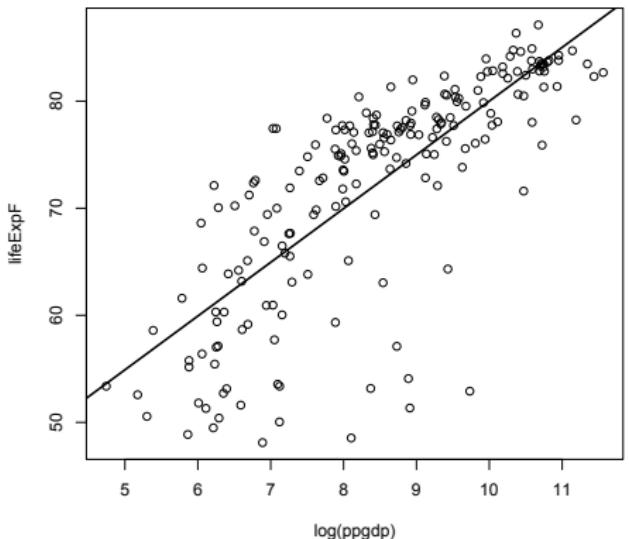
1 = Afghanistan
72 = Greenland
125 = Nauru

MODEL DIAGNOSTIC PLOTS



universität
wien

Gaussian linear Model: $Y \sim N(X\beta, \sigma^2 I_n)$



What if the model is too complex to visualize it like that? Can we still get a graphical representation of model fit?



MODEL DIAGNOSTIC PLOTS



universität
wien

Gaussian linear Model:

$$Y \sim N(X\beta, \sigma^2 I_n)$$

or, equivalently

$$Y_i = X_{i \cdot} \beta + u_i, \quad \text{with} \quad u = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} \sim N(0, \sigma^2 I_n)$$

Recall: OLS

$$\hat{\beta} = \operatorname{argmin}_{b \in \mathbb{R}^p} \|Y - Xb\|_2^2 = \operatorname{argmin}_{b \in \mathbb{R}^p} \sum_{i=1}^n (Y_i - X_{i \cdot} b)^2$$

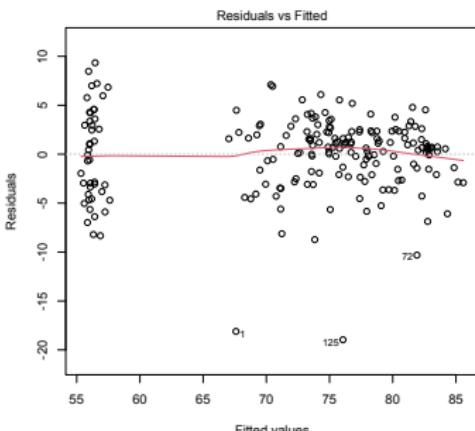
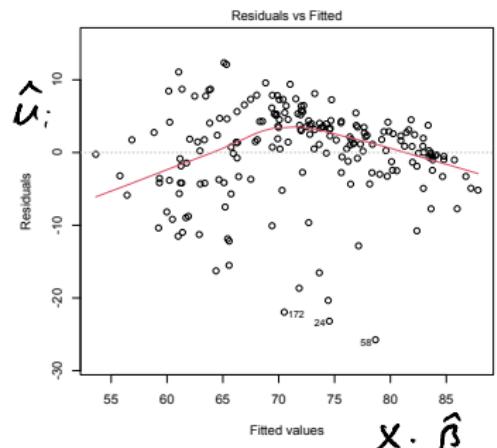
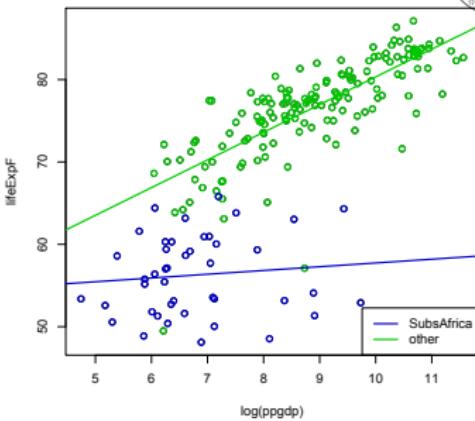
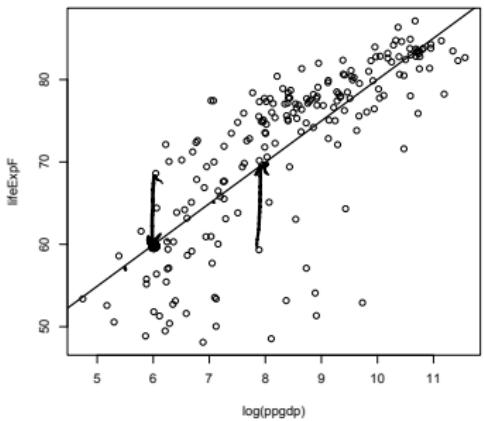
Idea: The *residuals* $\hat{u}_i = Y_i - X_{i \cdot} \hat{\beta}$ should be (approximately) iid Gaussian and evenly distributed around the regression line (independent of their 'location' on the regression line).

$\hat{Y}_i = X_{i \cdot} \hat{\beta}$ are the *fitted* or *predicted* values 'on' the regression line.

1.) plot \hat{Y}_i against \hat{u}_i

2.) check \hat{u}_i for Gaussianity

RESIDUAL PLOT

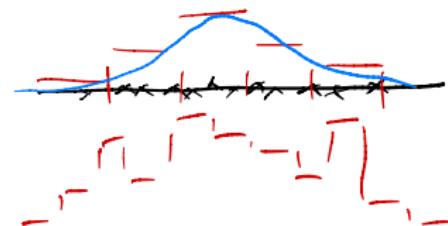


$1 = \text{Afghanistan}$
 $72 = \text{Greenland}$
 $125 = \text{Nauru}$

How do we check Gaussianity of the sample of residuals \hat{u}_i ,
 $i = 1, \dots, n$?

Construct a normal QQ-plot:

- ▶ Compute $p_i := \frac{\text{rank}(\hat{u}_i)}{n+1} \approx \frac{\text{rank}(\hat{u}_i)}{n}$.
- ▶ Compute $y_i := \Phi^{-1}(p_i)$.



Plot the y_i against the **standardized residuals \tilde{u}_i** and draw the 45-degree line.

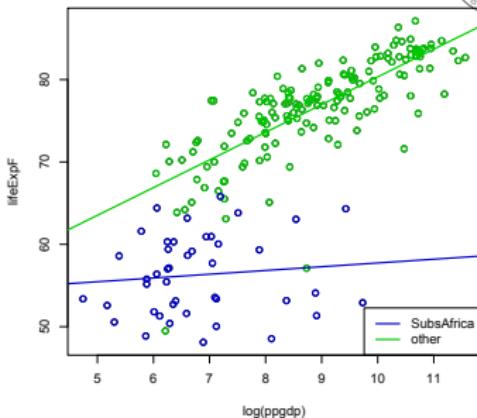
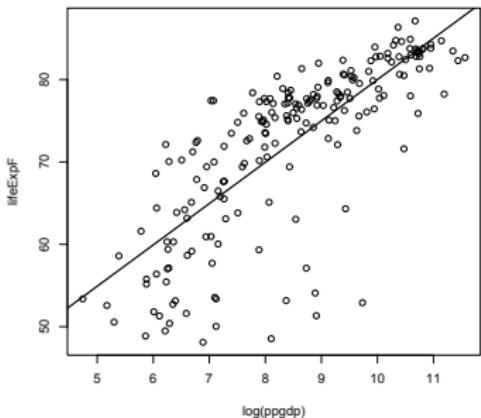
$$\tilde{u}_i = \frac{\hat{u}_i - \bar{\hat{u}}_n}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (\hat{u}_i - \bar{\hat{u}}_n)^2}}, \quad \bar{\hat{u}}_n = \frac{1}{n} \sum_{i=1}^n \hat{u}_i$$

$$y_i = \tilde{u}_i$$

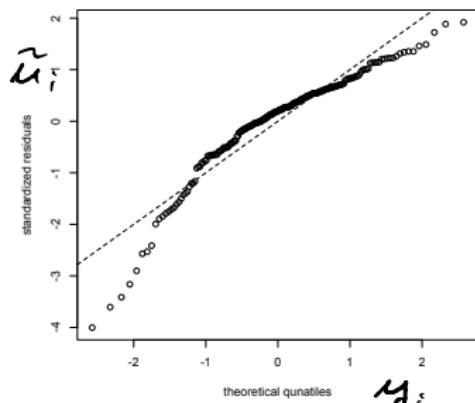
QQ PLOT



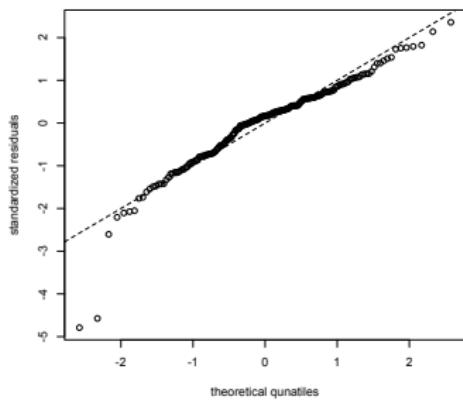
universität
wien



Normal QQ-plot

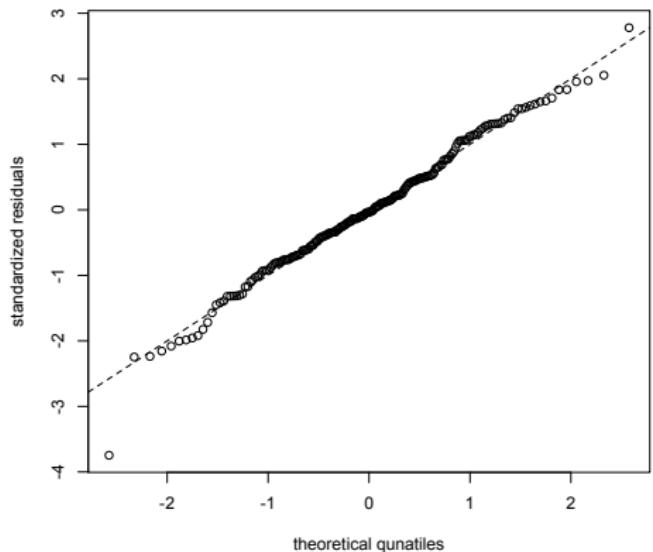


Normal QQ-plot

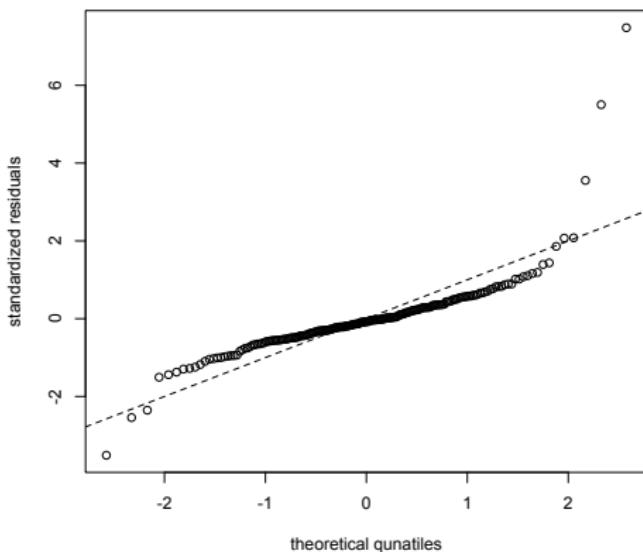


μ_f

Normal QQ-plot

 $N(0, 1)$ -data

Normal QQ-plot

 t_2 -data



FINDINGS ARE FALSE (IOANNIDIS, 2005)

Should we trust a study or interpretation that is based only on a single hypothesis test?

For simplicity: all tests have level α and power γ .

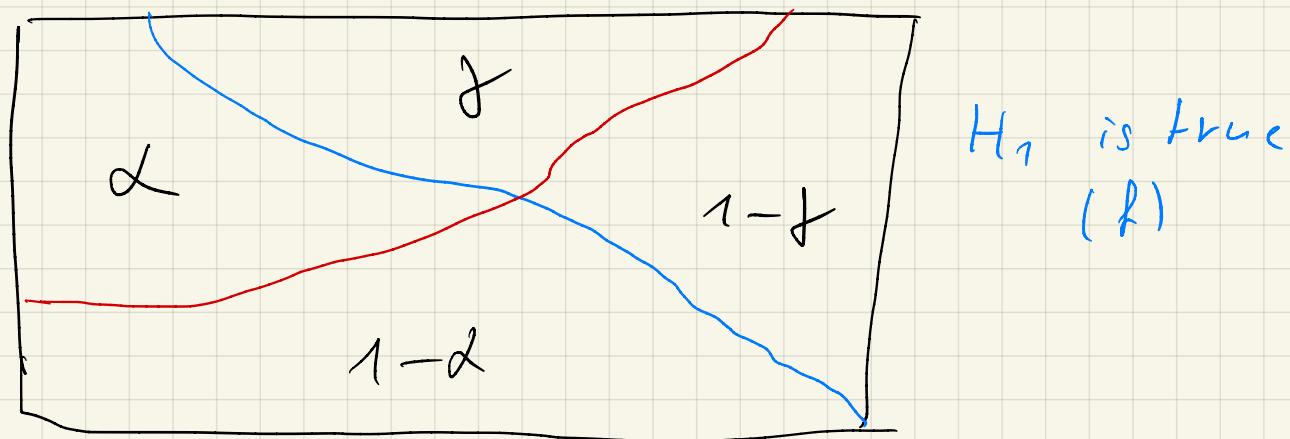
f is the fraction of all hypothesis tests conducted in a certain area of science (= the population) for which H_0 is actually false.

$$P(\text{true discovery} | \text{discovery}) = ?$$

$$P(\text{true discovery} \mid \text{discovery}) = \frac{P(\text{true discovery}, \text{discovery})}{P(\text{discovery})} = \frac{P(\text{true disc.})}{P(\text{disc.})}$$

H_0 rejected

all tests



H_0 is true

(1 - f)

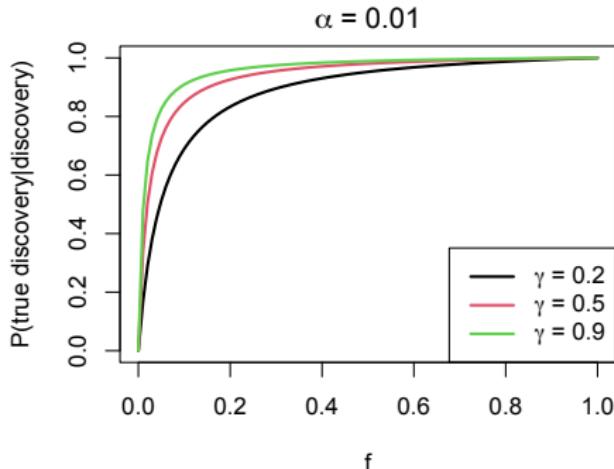
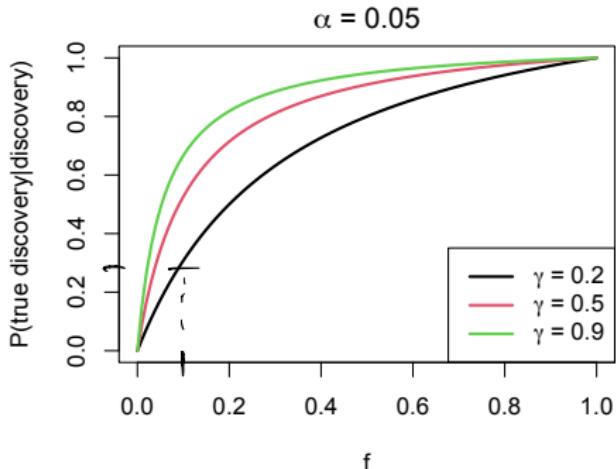
H_0 is accepted

H_1 is true
(f)

$$P(\text{true discovery}) = P(H_0 \text{ rejected}, H_1 \text{ true}) = f \cdot \beta$$

$$P(\text{discovery}) = f \cdot \beta + (1-f) \cdot \alpha$$

WHY MOST PUBLISHED RESEARCH FINDINGS ARE FALSE (IOANNIDIS, 2005)



- ▶ Ask: Is f in our case reasonably large?
Test hypotheses that are based on well established theories.
Just asking lots of questions will not give us more reliable answers.
- ▶ Try to reproduce your findings on independent follow-up studies!



The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812

NOVEMBER 5, 2020

VOL. 383 NO. 19

Remdesivir for the Treatment of Covid-19 — Final Report

J.H. Beigel, K.M. Tomashek, L.E. Dodd, A.K. Mehta, B.S. Zingman, A.C. Kalil, E. Hohmann, H.Y. Chu, A. Luettkemeyer, S. Kline, D. Lopez de Castilla, R.W. Finberg, K. Dierberg, V. Tapson, L. Hsieh, T.F. Patterson, R. Paredes, D.A. Sweeney, W.R. Short, G. Touloumi, D.C. Lye, N. Ohmagari, M. Oh, G.M. Ruiz-Palacios, T. Benfield, G. Fätkenheuer, M.G. Kortepeter, R.L. Atmar, C.B. Creech, J. Lundgren, A.G. Babiker, S. Pett, J.D. Neaton, T.H. Burgess, T. Bonnett, M. Green, M. Makowski, A. Osinusi, S. Nayak, and H.C. Lane,
for the ACTT-1 Study Group Members*

ABSTRACT

BACKGROUND

Although several therapeutic agents have been evaluated for the treatment of coronavirus disease 2019 (Covid-19), no antiviral agents have yet been shown to be efficacious.

METHODS

We conducted a double-blind, randomized, placebo-controlled trial of intravenous remdesivir in adults who were hospitalized with Covid-19 and had evidence of lower respiratory tract infection. Patients were randomly assigned to receive either remdesivir (200 mg loading dose on day 1, followed by 100 mg daily for up to 9 additional days) or placebo for up to 10 days. The primary outcome was the time to recovery, defined by either discharge from the hospital or hospitalization for infection-control purposes only.

RESULTS

A total of 1062 patients underwent randomization (with 541 assigned to remdesivir and 521 to placebo). Those who received remdesivir had a median recovery time of 10 days (95% confidence interval [CI], 9 to 11), as compared with 15 days (95% CI, 13 to 18) among those who received placebo (rate ratio for recovery, 1.29; 95% CI, 1.12 to 1.49; $P<0.001$, by a log-rank test). In an analysis that used a proportional-odds model with an eight-category ordinal scale, the patients who received remdesivir were found to be more likely than those who received placebo to have clinical improvement at day 15 (odds ratio, 1.5; 95% CI, 1.2 to 1.9, after adjustment for actual disease severity). The Kaplan-Meier estimates of mortality were 6.7%

The authors' full names, academic degrees, and affiliations are listed in the Appendix. Address reprint requests to Dr. Beigel at the National Institute of Allergy and Infectious Diseases, National Institutes of Health, 5601 Fishers Ln, Rm. 7E60, MSC 9826, Rockville, MD 20892-9826, or at jbeigel@niaid.nih.gov.

*A complete list of members of the ACTT-1 Study Group is provided in the Supplementary Appendix, available at NEJM.org.

A preliminary version of this article was published on May 22, 2020, at NEJM.org. This article was published on October 8, 2020, and updated on October 9, 2020, at NEJM.org.

N Engl J Med 2020;383:1813-26.

DOI: 10.1056/NEJMoa2007764

Copyright © 2020 Massachusetts Medical Society.

THE FILE DRAWER EFFECT



Can we use a published p -value just like that?

- ▶ Suppose we are doing a meta-analysis collecting 100 papers on clinical trials that investigated the efficacy of Remdesivir for Covid-19 treatment.
- ▶ Suppose they all applied sound statistical methodology.
⇒ Their p -values are valid!
- ▶ Suppose Remdesivir is really ineffective for treating Covid-19.
- ▶ In how many of our 100 papers do you expect to see a p -value of less than or equal to $\alpha_0 = 0.05$? $\chi \approx 5$

$$P_{H_0} (p \leq \alpha_0) \leq \alpha_0$$

THE FILE DRAWER EFFECT



Can we use a published p -value just like that?

- ▶ Suppose we are doing a meta-analysis collecting 100 papers on clinical trials that investigated the efficacy of Remdesivir for Covid-19 treatment.
- ▶ Suppose they all applied sound statistical methodology.
⇒ Their p -values are valid!
- ▶ Suppose Remdesivir is really ineffective for treating Covid-19.
- ▶ Suppose the peer review system only allows statistically significant results to be published.
- ▶ Suppose a result is considered significant if $p \leq \alpha_0 = 0.05$.
- ▶ In how many of our 100 papers do you expect to see a p -value of less than or equal to $\alpha_0 = 0.05$? *all of them*
- ▶ In how many of our 100 papers do you expect to see a p -value of less than or equal to $\alpha = 0.01$?



We are really looking at a subpopulation of all Remdesivir trials, namely those where $p \leq \alpha_0$.

$$P_{H_0}(p \leq \alpha | p \leq \alpha_0) = \alpha \quad ?$$

Recall Ex.2.2: A valid p-value of a simple null hypothesis is uniformly distributed under H_0 .

$$P_{H_0} (p \leq \alpha \mid p \leq \alpha_0) = \frac{P_{H_0}(p \leq \alpha, p \leq \alpha_0)}{P_{H_0}(p \leq \alpha_0)}$$

$$= \frac{\min(\alpha, \alpha_0)}{\alpha_0} \gg \alpha_0$$

$$\frac{0,01}{0,05} = 0,2$$

THE FILE DRAWER EFFECT



- ▶ Suppose a result is considered significant if $p \leq \alpha_0 = 0.05$.
- ▶ In how many of our 100 papers do you expect to see a p -value of less than or equal to $\alpha = 0.01$?

in approx. 20

Hence, the current peer review system may produce a lot more spurious discoveries than we would actually expect from controlling type-one error probabilities!

$$\begin{aligned} P_{\text{corr}} &:= \frac{P}{\alpha_0} \Rightarrow P_{H_0}(P_{\text{corr}} \leq \alpha | p \leq \alpha_0) \\ &= \frac{0.001}{0.05} \quad = P_{H_0}(p \leq \alpha \cdot \alpha_0 | p \leq \alpha_0) \\ &= \frac{\min(\alpha \cdot \alpha_0, \alpha_0)}{\alpha_0} \quad = \frac{\alpha \cdot \alpha_0}{\alpha_0} = \alpha. \end{aligned}$$

The misspecified Gaussian linear model

THE GAUSSIAN LINEAR MODEL



- ▶ $Y \sim N(X\beta, \sigma^2 I_n)$, $\beta \in \mathbb{R}^p$, $\sigma^2 \in (0, \infty)$.
- ▶ Low-dimensional case: $p < n$
- ▶ X is an $n \times p$ (non-random) design matrix with $\text{rank}(X) = p$ (e.g., analysis conditional on X)

These assumptions may be violated!

- ▶ That Y has Gaussian distribution can often be removed by appealing to the CLT approximation, i.e.,

$$\frac{\hat{\beta}_k - \beta_k}{\hat{\sigma} \sqrt{[(X'X)^{-1}]_k}} \xrightarrow[n \rightarrow \infty]{D} N(0, 1),$$

- ▶ or by applying the bootstrap.
- ▶ But why would $\mathbb{E}[Y] = X\beta$ and $p < n$?

THE MISSPECIFIED GAUSSIAN LINEAR MODEL



- ▶ $Y \sim N(\mu, \sigma^2 I_n)$, $\mu \in \mathbb{R}^n$, $\sigma^2 \in (0, \infty)$.
- ▶ Low-dimensional case: $p < n$ (for now)
- ▶ X is an $n \times p$ (non-random) design matrix with $\text{rank}(X) = p$.
- ▶ For simplicity: $\sigma^2 = 1$.

We may still use OLS:

- ▶ $\hat{\beta} := \underset{b \in \mathbb{R}^p}{\text{argmin}} \|Y - Xb\|_2^2 = (X'X)^{-1}X'Y$
- ▶ $\hat{\beta} \sim N(\beta^*, \sigma^2(X'X)^{-1})$, where
- ▶ $\beta^* = \mathbb{E}[\hat{\beta}] = (X'X)^{-1}X'\mu = \underset{b \in \mathbb{R}^p}{\text{argmin}} \|\mu - Xb\|_2^2$.

“best approximation”

THE MISSPECIFIED GAUSSIAN LINEAR MODEL



$\beta = ?$ does not exist!

$$Y \sim N(\mu, I_n), \quad \mathbb{E}[Y] = \mu \approx X\beta^*$$

Suppose X_2 is 'gender' where 0 = female, 1 = male. If the i -th individual is female we have

$$\mathbb{E}[Y_i] = \mu_i \approx \beta_1^* + 0 + \sum_{k=3}^p \beta_k^* X_{ik},$$

whereas if the j -th individual is male, we have

$$\mathbb{E}[Y_j] = \mu_j \approx \beta_1^* + \beta_2^* + \sum_{k=3}^p \beta_k^* X_{jk}.$$

β_2^* is the additional income of men over women, given that all other variables are the same, in the best linear approximation to the expected income using the variables in X .

THE MISSPECIFIED GAUSSIAN LINEAR MODEL: STATISTICAL INFERENCE



universität
wien

$$Y \sim N(\mu, I_n), \quad \mathbb{E}[Y] = \mu \approx X\beta^* = \sum_{k=1}^p \beta_k^* X_{\cdot k}$$

Then

$$\frac{\hat{\beta}_k - \beta_k^*}{\sqrt{[(X'X)^{-1}]_k}} \sim N(0, 1),$$

where $[(X'X)^{-1}]_k$ is the k -th diagonal entry of $(X'X)^{-1}$.

Use this to construct tests or confidence intervals for β_k^* .

not for β !

THE MISSPECIFIED GAUSSIAN LINEAR MODEL



universität
wien

$$Y \sim N(\mu, I_n), \quad \mathbb{E}[Y] = \mu \approx X\beta^*$$

If we happen to choose the “correct” variables, i.e.,
 $\mu \in \text{span}(X)$, ...

$$\mu \in \text{span}(X) := \left\{ \begin{matrix} Xb \\ \in \mathbb{R}^n \end{matrix} : b \in \mathbb{R}^p \right\}$$

$$\Rightarrow \exists b_0 \in \mathbb{R}^p : \mu = Xb_0$$

$$\Rightarrow \beta^* = \underset{b \in \mathbb{R}^n}{\arg \min} \| \mu - Xb \|_2^2 = b_0$$

$$\Rightarrow Y \sim N(Xb_0, I_n)$$

Statistics for Data Science, WS2023

Chapter 5:

Statistical Network Analysis

NETWORKS ARE EVERYWHERE

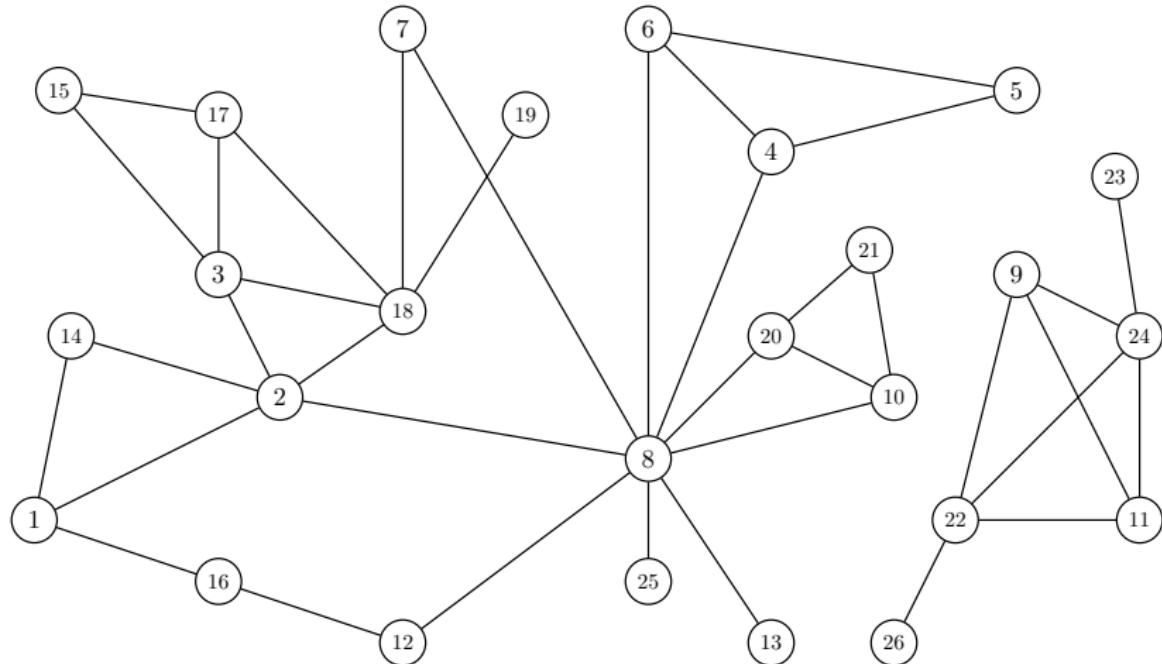
- ▶ social networks
- ▶ computer networks (WWW)
- ▶ electricity grid
- ▶ street maps
- ▶ gene regulatory networks
- ▶ etc.

Often the full network is unknown or too big to compute or extract the characteristics of interest

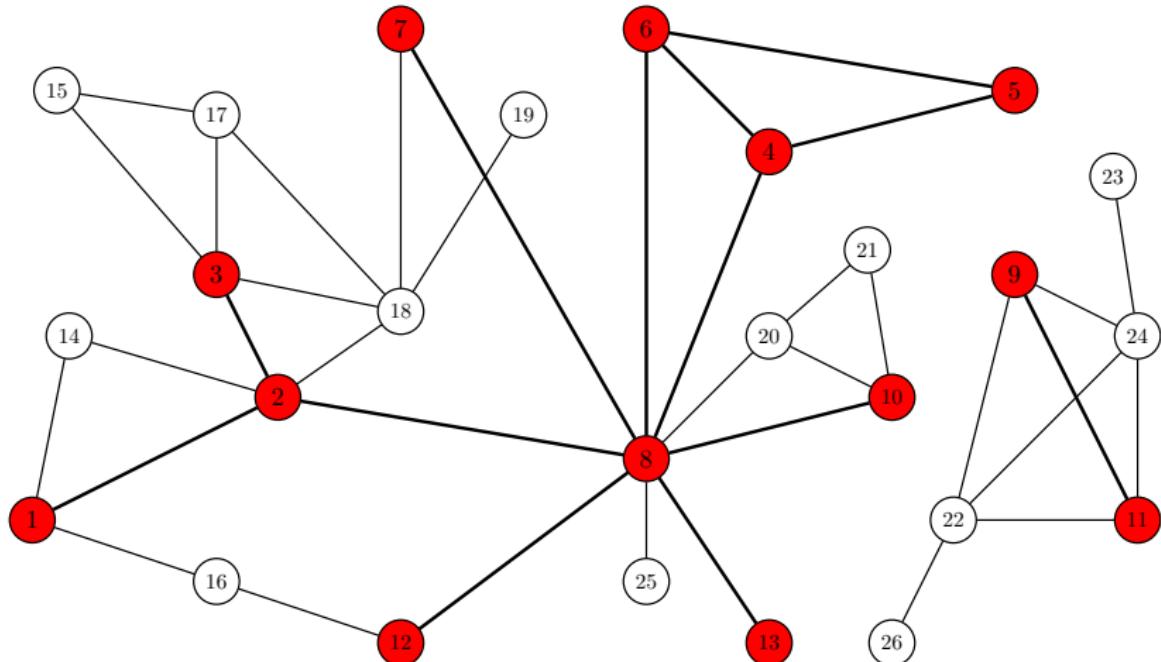


Random sampling + statistical inference

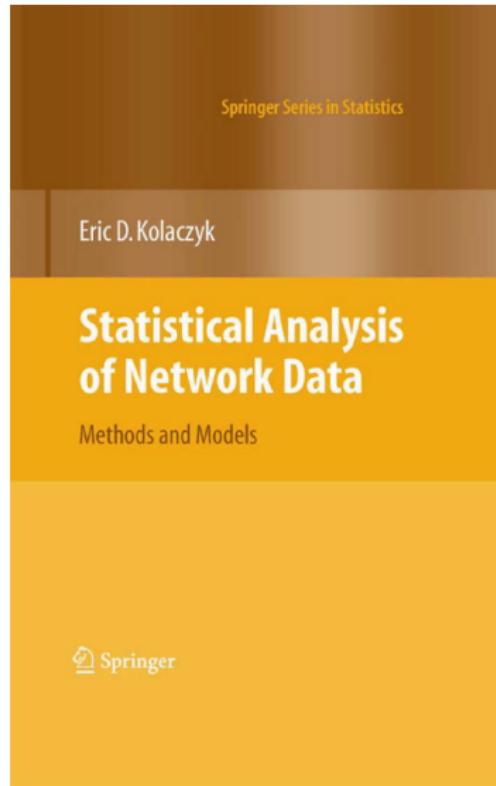
SAMPLING FROM A HIDDEN NETWORK



SAMPLING FROM A HIDDEN NETWORK



FOLLOWING KOLACZYK (2009)



GRAPH NOTATION

Definition 5.1

An undirected graph is given by a pair $G = (V, E)$ where $V \subseteq \mathbb{N}$ is the set of *vertices* (or *nodes*) and

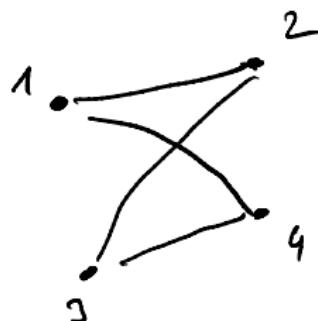
$E \subseteq V^{(2)} := \{\{v, u\} : v, u \in V, v \neq u\}$ is the set of (undirected) *edges* or *links*. We write $N_V := |V|$, $N_E := |E|$ and $V = \{v_1, \dots, v_{N_V}\}$.

$$\{v, u\} = \{u, v\}$$

GRAPH NOTATION

For (an undirected) graph $G = (V, E)$, its *adjacency matrix* $A \in \mathbb{R}^{N_V \times N_V}$ is given by

$$A_{ij} = \begin{cases} 1, & \text{if } \{v_i, v_j\} \in E, \\ 0, & \text{else.} \end{cases}$$



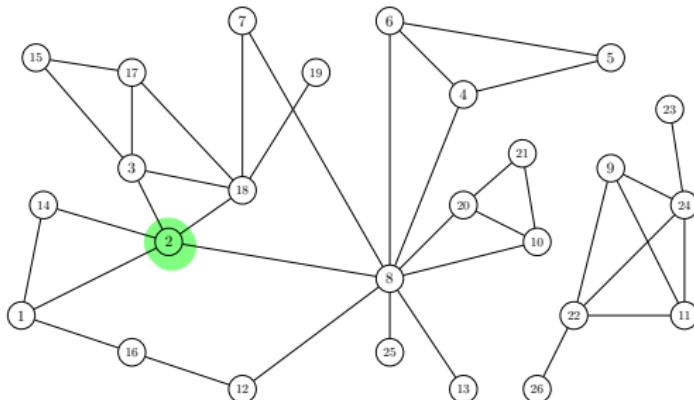
$$\begin{array}{ccccc} & 1 & 2 & 3 & 4 \\ 1 & 0 & 1 & 0 & 1 \\ 2 & 1 & 0 & 1 & 0 \\ 3 & 0 & 1 & 0 & 1 \\ 4 & 1 & 0 & 1 & 0 \end{array}$$

DEGREE

For $G = (V, E)$ and $v \in V$, the *degree* d_v of v is given by the number of vertices adjacent to v , i.e.,

$$d_v := \sum_{u \in V} \mathbf{1}_E(\{u, v\}) = \sum_{j=1}^{N_V} A_{ij},$$

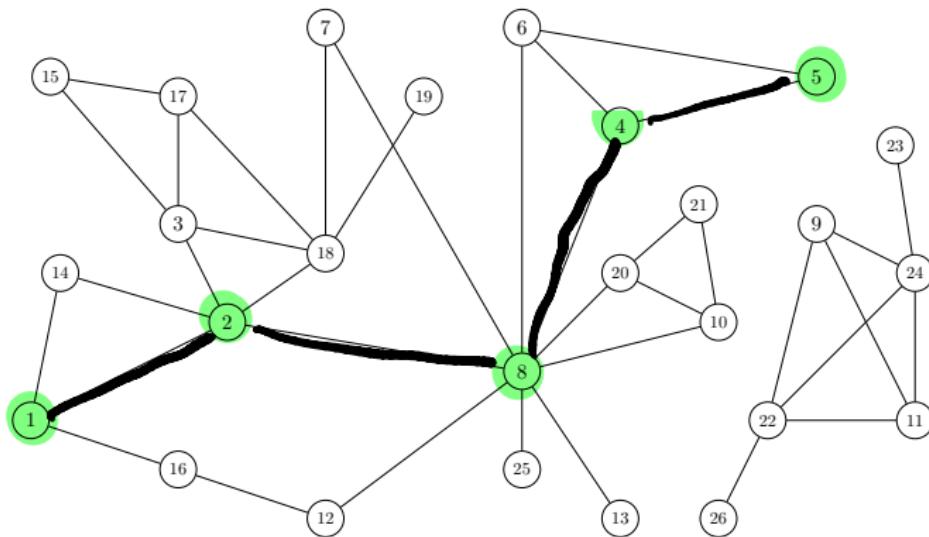
where $i \in [N_V]$ is such that $v_i = v$.



$$d_2 = 5$$

PATH

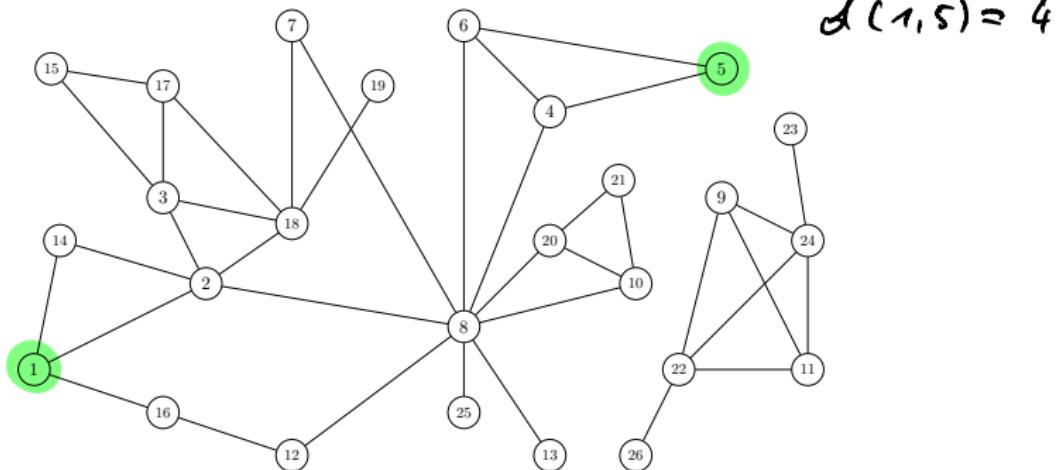
For $G = (V, E)$ a *path* is a sequence of adjacent vertices in which no vertex occurs twice, i.e., $(v_1, v_2, \dots, v_l) \in V^l$ is a path of length $l - 1$ if $\{v_i, v_{i+1}\} \in E$ for $i = 1, \dots, l - 1$ and $v_i \neq v_j$ for all $i \neq j$.



GEODESIC DISTANCE

For $G = (V, E)$ and $v, u \in V$, the geodesic distance $\text{dist}(v, u)$ between v and u is the length of the (or a) shortest path starting in v and ending in u , and $\text{dist}(v, v) := 0$.

If there is no path from v to u , we set $\text{dist}(v, u) = \infty$.



CENTRALITY

For $G = (V, E)$ and $v \in V$, the *closeness centrality* $c_{Cl}(v)$ of v is defined as

$$c_{Cl}(v) := \frac{1}{\sum_{u \in V} \text{dist}(v, u)}.$$

For a vertex $v \in V$, the *betweenness centrality* $c_B(v)$ of v is defined as

$$c_B(v) := \sum_{\substack{\{s,t\} \in V^{(2)} \\ v \notin \{s,t\}}} \frac{\sigma(s, t|v)}{\sigma(s, t)},$$

where $\sigma(s, t)$ is the number of shortest paths between s and t , and $\sigma(s, t|v)$ is the number of all those paths that also pass through v . By convention we set $\frac{0}{0} = 0$.

Sampling from a finite population

Course evaluation!

SAMPLING FROM FINITE POPULATION

- ▶ Population or universe $\mathcal{U} = \{1, \dots, N\}$, with N known
- ▶ Characteristics of interest $y_i \in \mathbb{R}, i \in \mathcal{U}$
- ▶ Population total and average $\tau := \sum_{i \in \mathcal{U}} y_i, \mu := \frac{\tau}{N}$
- ▶ Draw a **random** sample $S = (i_1, \dots, i_n) \in \mathcal{U}^n$
- ▶ We **observe** y_{i_1}, \dots, y_{i_n} (duplicates possible!) $(1, 2) \neq$

Goal: Estimate τ and/or μ .

$(2, 1)$

SAMPLING FROM FINITE POPULATION

$$y_i \in \mathbb{R}, i \in \mathcal{U} = \{1, \dots, N\}, \quad \tau := \sum_{i \in \mathcal{U}} y_i, \quad \mu := \frac{\tau}{N}$$

random sample $S = (i_1, \dots, i_n) \in \mathcal{U}^n$

natural choice: (why?)

$$\tilde{\mu} = \frac{1}{n} \sum_{j=1}^n y_{i_j}, \quad \tilde{\tau} = N \tilde{\mu}$$

How to compute $\mathbb{E}[\tilde{\tau}]$ and $\text{Var}[\tilde{\tau}]$?

SAMPLING FROM FINITE POPULATION

$$y_i \in \mathbb{R}, i \in \mathcal{U} = \{1, \dots, N\}, \quad \tau := \sum_{i \in \mathcal{U}} y_i, \quad \mu := \frac{\tau}{N}$$

random sample $S = (i_1, \dots, i_n) \in \mathcal{U}^n$

For $i \in \mathcal{U}$, define

$$Z_i := \sum_{j=1}^n \mathbb{1}_{\{i_j\}}(i)$$

number of times individual i is sampled

and $\pi_i := \mathbb{E}[Z_i]$, $\pi_{ij} := \mathbb{E}[Z_i Z_j]$.

$$\pi_{ij} \neq \pi_{ij}$$

If $\pi_i > 0$, define

$$\hat{\tau} := \sum_{j=1}^n \frac{y_{i_j}}{\pi_{ij}}.$$

Horvitz-Thompson estimate

SAMPLING FROM FINITE POPULATION

$$y_i \in \mathbb{R}, i \in \mathcal{U} = \{1, \dots, N\}, \quad \tau := \sum_{i \in \mathcal{U}} y_i, \quad \mu := \frac{\tau}{N}$$

random sample $S = (i_1, \dots, i_n) \in \mathcal{U}^n$

For $i \in \mathcal{U}$, define

$$Z_i := \sum_{j=1}^n \mathbb{1}_{\{i_j\}}(i) \quad \text{number of times individual } i \text{ is sampled}$$

and $\pi_i := \mathbb{E}[Z_i]$, $\pi_{ij} := \mathbb{E}[Z_i Z_j]$.

Because of

$$\sum_{i \in \mathcal{U}} \frac{y_i}{\pi_i} Z_i = \sum_{j=1}^n \sum_{i \in \mathcal{U}} \frac{y_i}{\pi_i} \mathbb{1}_{\{i_j\}}(i) = \sum_{j=1}^n \frac{y_{i_j}}{\pi_{i_j}} = \hat{\tau},$$

we have

$$\mathbb{E}[\hat{\tau}] = \mathbb{E} \left[\sum_{i \in \mathcal{U}} \frac{y_i}{\pi_i} Z_i \right] = \sum_{i \in \mathcal{U}} \frac{y_i}{\pi_i} \mathbb{E}[Z_i] = \sum_{i \in \mathcal{U}} y_i = \tau.$$

SAMPLING FROM FINITE POPULATION

$$\begin{aligned}\text{Var}[\hat{\tau}] &= \text{Var}\left(\sum_{i \in U} \frac{y_i}{\pi_i} z_i\right) = \sum_{i,j \in U} \text{Cov}\left(\frac{y_i}{\pi_i} z_i, \frac{y_j}{\pi_j} z_j\right) \\&= \sum_{i,j \in U} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \underbrace{\text{Cov}(z_i, z_j)}_{= E(z_i z_j) - E(z_i) E(z_j)} \\&\quad = \frac{\pi_{ij}}{\pi_i \cdot \pi_j} \\&= \sum_{i,j \in U} y_i y_j \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\right)\end{aligned}$$

HW: Find an unbiased estimator for $\text{Var}[\hat{\tau}]$.

SAMPLING WITH REPLACEMENT

$$y_i \in \mathbb{R}, i \in \mathcal{U} = \{1, \dots, N\}, \quad \tau := \sum_{i \in \mathcal{U}} y_i$$

draw n times uniformly from \mathcal{U} with replacement to obtain
 $S = (i_1, \dots, i_n) \in \mathcal{U}^n$

For $i \in \mathcal{U}$, define

$$Z_i := \sum_{j=1}^n \mathbb{1}_{\{i_j\}}(i) \quad \text{number of times individual } i \text{ is sampled.}$$

We have

$$(Z_1, \dots, Z_N) \sim \text{Multinomial} \left(n; \underbrace{\frac{1}{N}, \dots, \frac{1}{N}}_{N \text{ times}} \right).$$

In particular, $\pi_i = \mathbb{E}[Z_i] = \frac{n}{N}$, $\pi_{ii} = \mathbb{E}[Z_i^2] = \frac{n(N+n-1)}{N^2}$, and
 $\pi_{ij} = \mathbb{E}[Z_i Z_j] = \frac{n(n-1)}{N^2}$ for $i \neq j$.

Multinomial Distribution

$$\text{P.m.f.} : P(Z_1 = z_1, Z_2 = z_2, \dots, Z_N = z_N) = \cancel{\star}$$

$$\sum_{i=1}^N z_i = n, \quad z_i \in N_0$$

↖ draw n-times with replacement

$$U = \{1, 2, \dots, N\} \quad p_i = \frac{1}{N}$$

$$p_1 \quad p_2 \quad \quad \quad p_N$$

$$\text{Example: } S = (1, 3, 5, 3, 3, 1)$$

$$Z_1 \quad Z_2 \quad Z_3 \quad Z_4 \quad Z_5 \quad \dots \quad Z_N \quad n=6$$

$$z = 2 \quad 0 \quad 3 \quad 0 \quad 1 \quad \quad \quad 0$$

$$\text{probability of drawing } S = p_1^2 \cdot p_3^3 \cdot p_5^1$$

probability of observing z ? Different samples can produce the same z !

How many?

$$P_1^2 \cdot P_3^3 \cdot P_5^1 \cdot \frac{6!}{2! 3! 1!}$$

$$= n! \prod_{i=1}^N \frac{P_i^{z_i}}{z_i!} = \textcircled{X}$$

$$P_i = \frac{1}{N} \Rightarrow \textcircled{X} = n! \prod_{i=1}^N \frac{\left(\frac{1}{N}\right)^{z_i}}{z_i!}$$

$$= n! \frac{\left(\frac{1}{N}\right)^{\sum_{i=1}^N z_i}}{z_1! \cdots z_N!} \quad \sum z_i = n$$

$$= \frac{n!}{N^n} \frac{1}{z_1! \cdots z_N!}$$

SAMPLING WITHOUT REPLACEMENT

$$y_i \in \mathbb{R}, i \in \mathcal{U} = \{1, \dots, N\}, \quad \tau := \sum_{i \in \mathcal{U}} y_i$$

draw n times uniformly from \mathcal{U} without replacement to obtain
 $S = (i_1, \dots, i_n) \in \mathcal{U}^n$

For $i \in \mathcal{U}$, define $Z_i := \sum_{j=1}^n \mathbb{1}_{\{i_j\}}(i) \in \{0, 1\}$. We have

$$P(Z_i = 1) = 1 - \frac{\# \text{ samples of size } n \text{ not containing individual } i}{\# \text{ samples of size } n}$$

$$= 1 - \frac{(N-1)(N-2) \cdots (N-n)}{N(N-1) \cdots (N-n+1)} = \frac{n}{N}$$

$$\Rightarrow \pi_i = \mathbb{E}[Z_i] = P(Z_i = 1) = \frac{n}{N} \text{ and } \pi_{ii} = \mathbb{E}[Z_i^2] = \mathbb{E}[Z_i] = \pi_i.$$

SAMPLING WITHOUT REPLACEMENT

For $i \neq j$, we have

$$\begin{aligned}\pi_{ij} &= \mathbb{E}[Z_i Z_j] = \mathbb{E}[Z_i + Z_j - 1 + \underbrace{(1 - Z_i)(1 - Z_j)}_l] \\ &= P(Z_i = 1) + P(Z_j = 1) - 1 + P(Z_i = 0 = Z_j) \\ &= 2\frac{n}{N} - 1 + \frac{(N-2)\cdots(N-n-1)}{N \cdot (N-1)\cdots(N-n+1)} \\ &= \frac{2n(N-1) - N(N-1) + (N-n)(N-1-n)}{N(N-1)} \\ &= \frac{2n(N-1) - N(N-1) + N(N-1) - n(N-1) - Nn + n^2}{N(N-1)} \\ &= \frac{n(N-1) - Nn + n^2}{N(N-1)} = \frac{n(n-1)}{N(N-1)}.\end{aligned}$$

SAMPLING FROM FINITE POPULATION

$$y_i \in \mathbb{R}, i \in \mathcal{U} = \{1, \dots, N\}, \quad \tau := \sum_{i \in \mathcal{U}} y_i, \quad \mu := \frac{\tau}{N}$$

random sample $S = (i_1, \dots, i_n) \in \mathcal{U}^n$

For $i \in \mathcal{U}$, define

$$Z_i := \sum_{j=1}^n \mathbb{1}_{\{i_j\}}(i) \quad \text{number of times individual } i \text{ is sampled}$$

and $\pi_i := \mathbb{E}[Z_i]$, $\pi_{ij} := \mathbb{E}[Z_i Z_j]$.

If $\pi_i > 0$, define

$$\hat{\tau} := \sum_{j=1}^n \frac{y_{i_j}}{\pi_{i_j}}$$

Horvitz-Thompson estimate

Graph sampling designs

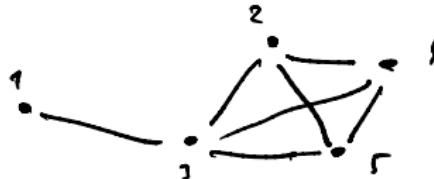
GRAPH CHARACTERISTICS AS TOTALS

- ▶ Population graph $G = (V, E)$
- ▶ Without loss of generality, write $V = [N_V] = \{1, \dots, N_V\}$

We want to estimate a population total, for example:

- ▶ $\mathcal{U} := V, (y_u)_{u \in \mathcal{U}}, \tau = \sum_{u \in \mathcal{U}} y_u, \mu = \frac{\tau}{N_V}.$
 - ▶ vertex characteristics, e.g., y_i is gender, age, etc.
 - ▶ degree $y_i = d_i \Rightarrow \tau = 2N_E$
- ▶ $\mathcal{U} := V^{(2)}, (y_u)_{u \in \mathcal{U}}, \tau = \sum_{u \in \mathcal{U}} y_u.$
 - ▶ $y_{\{i,j\}}$ is the proportion of shortest paths between i and j passing through a given vertex $k \in V$ and $y_{\{i,j\}} = 0$ if $k \in \{i,j\} \Rightarrow \tau = c_B(k)$
 - ▶ edge characteristics/weights, e.g., number of phone calls between two phone numbers $\Rightarrow \tau$ is the total number of phone calls
 - ▶ $y_{\{i,j\}} = \mathbb{1}_E(\{i,j\}) \Rightarrow \tau = \sum_{e \in E} 1 = N_E (\mathcal{U} = E)$
 - ▶ $y_{\{i,j\}} = \mathbb{1}_E(\{i,j\}) \mathbb{1}_{y_i = y_j}$ (e.g., y_i gender) $\Rightarrow \tau$ = number of same sex friendships ($\mathcal{U} = E$)

GRAPH CHARACTERISTICS AS TOTALS



We want to estimate a population total, for example:

- ▶ Number of connected triangles in the graph:

$$U = V^{(3)} := \{ \{i, j, k\} : i+j, j+k, i+k \}$$

$$\begin{aligned} y_u &= \mathbb{1}_E(\{i, j\}) \cdot \mathbb{1}_E(\{j, k\}) \cdot \mathbb{1}_E(\{i, k\}) \\ &= \begin{cases} 1 & \text{if triangle} \\ 0 & \text{if not} \end{cases} \end{aligned}$$

$$\{2, 3, 5\} = \{3, 2, 5\} = \{5, 3, 2\} = \dots$$

GRAPH SAMPLING AND ESTIMATION

- ▶ Population graph $G = (V, E)$
- ▶ Without loss of generality, write $V = [N_V] = \{1, \dots, N_V\}$

Either $(y_u)_{u \in \mathcal{U}}$ is unobserved or G is too big/complicated to compute $\tau = \sum_{u \in \mathcal{U}} y_u$:

- ▶ Randomly sample a subgraph $G^* = (V^*, E^*)$ from G **without replacement/duplicates**:
- ▶ That is, draw $V^* \subseteq V$ and $E^* \subseteq E$ according to some sampling scheme (see below) to get a random sample $S \subseteq \mathcal{U}$.
- ▶ Use Horvitz-Thompson approach

$$\hat{\tau} = \sum_{u \in S} \frac{y_u}{\pi_u}$$

for inclusion probabilities π_u , $u \in \mathcal{U}$.



GRAPH SAMPLING AND ESTIMATION

- ▶ Population graph $G = (V, E)$
- ▶ Without loss of generality, write $V = [N_V] = \{1, \dots, N_V\}$

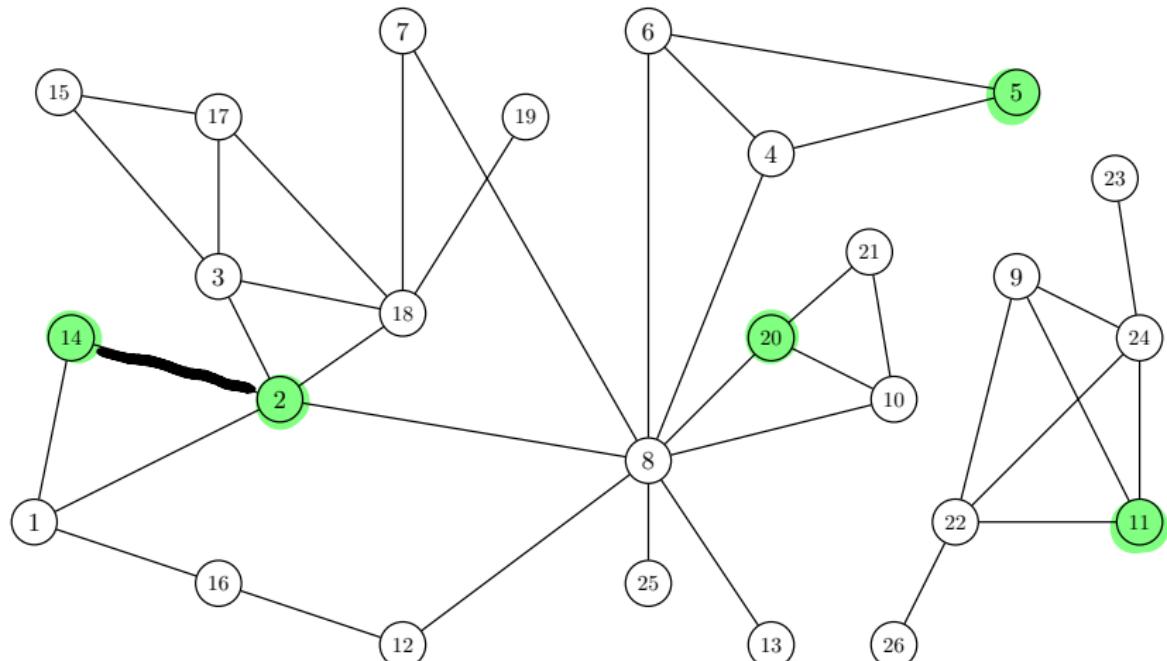
Either $(y_u)_{u \in \mathcal{U}}$ is unobserved or G is too big/complicated to compute $\tau = \sum_{u \in \mathcal{U}} y_u$:

- ▶ Randomly sample a subgraph $G^* = (V^*, E^*)$ from G **without replacement/duplicates**:
- ▶ That is, draw $V^* \subseteq V$ and $E^* \subseteq E$ according to some sampling scheme (see below) to get a random sample $S \subseteq \mathcal{U}$.
- ▶ Use Horvitz-Thompson approach

$$\hat{\tau} = \sum_{u \in S} \frac{y_u}{\pi_u}$$

for inclusion probabilities $\pi_u, u \in \mathcal{U}$.

INDUCED SUBGRAPH SAMPLING



$$V^* = \{14, 2, 20, 5, 11\}$$

$$E^* = \{\{14, 2\}\}$$

INDUCED SUBGRAPH SAMPLING

1. Sample n times **without replacement** from V to obtain V^* .
2. Add all edges joining vertices in V^* , i.e.,
$$E^* := \{\{u, v\} \in E : u, v \in V^*\}.$$

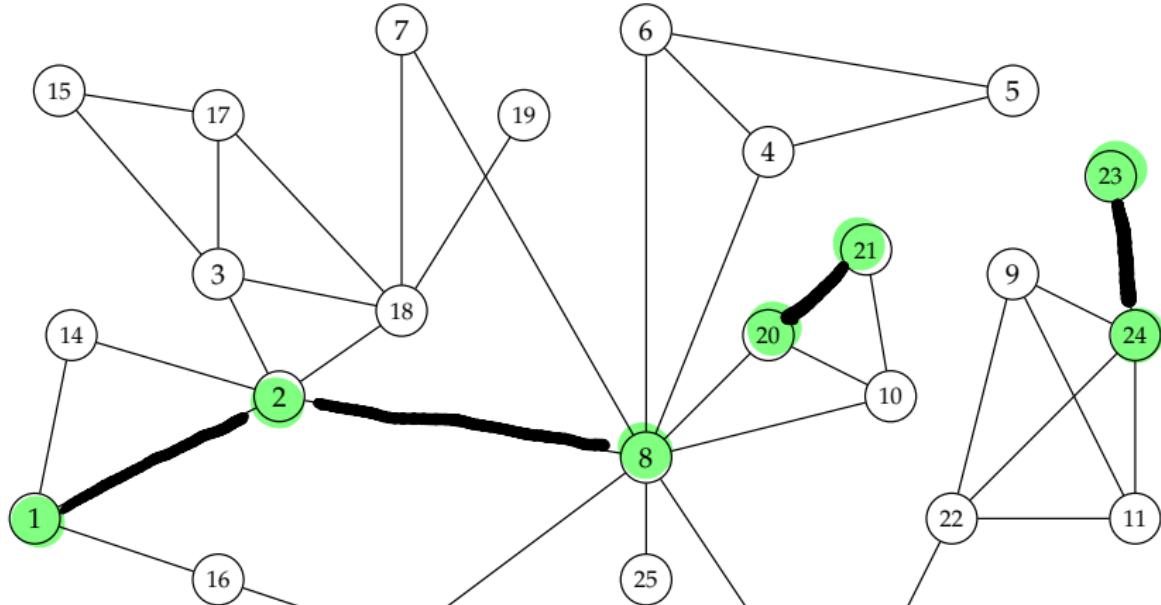
Vertex inclusion probabilities (as above):

- For $i \in V$, we have $\pi_i = \frac{n}{N_V}$.
- For $\{i, j\} \in V^{(2)}$, we have $\pi_{\{i, j\}} = \frac{n(n-1)}{N_V(N_V-1)}$.

in general :

$$e = \{i, j\}, \pi_e = P(\text{edge } e \text{ is sampled}) \neq \pi_{\{i, j\}}$$

INCIDENT SUBGRAPH SAMPLING



$$E^* = \{ \{1, 2\}, \{2, 8\}, \{20, 21\}, \{23, 25\} \}$$

$$V^* = \{1, 2, 8, 20, 21, 23, 25\}$$

INCIDENT SUBGRAPH SAMPLING

1. Sample n times **without replacement** from E to obtain E^* .
2. Add all incident vertices

$$V^* := \{v \in V : \exists e \in E^* \text{ such that } v \in e\}.$$

Edge inclusion probabilities:

- For $e \in E \subseteq V^{(2)}$, $\pi_e = P(\text{edge } e \text{ is sampled}) = \frac{n}{N_E}$.

Note: If $\mathcal{U} = V^{(2)}$, we can only use this if

$$\tau = \sum_{\{i,j\} \in V^{(2)}} y_{\{i,j\}} = \sum_{e \in E} y_e$$

is an edge total! Otherwise, we would need to compute
 $\pi_{\{i,j\}} := P(\text{vertex pair } \{i,j\} \text{ is sampled}), \text{ for all } \{i,j\} \in V^{(2)}$.

INCIDENT SUBGRAPH SAMPLING

1. Sample n times **without replacement** from E to obtain E^* .
2. Add all incident vertices

$$V^* := \{v \in V : \exists e \in E^* \text{ such that } v \in e\}.$$



Vertex inclusion probabilities:

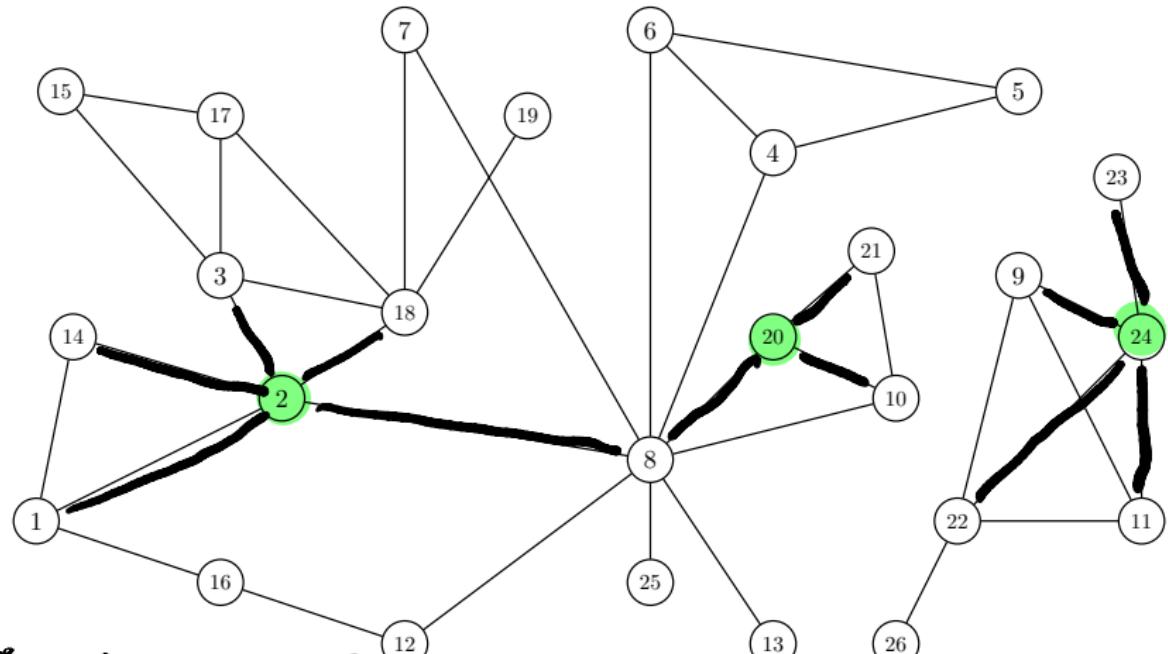
- ▶ For $i \in V$, $\pi_i = P(\text{vertex } i \text{ is sampled})$
- ▶ If $d_i > N_E - n$, then $\pi_i = 1$
- ▶ If $d_i \leq N_E - n$, then for $E_i := \{e \in E : i \in e\}$, $|E_i| = d_i$ and

$$\pi_i = 1 - P(\text{no edge incident to } i \text{ is sampled})$$

$$= 1 - \frac{\# \text{ (unordered) samples of size } n \text{ drawn from } E \setminus E_i}{\# \text{ (unordered) samples of size } n \text{ drawn from } E}$$

$$= 1 - \frac{\binom{N_E - d_i}{n}}{\binom{N_E}{n}}$$

UNLABELED STAR SAMPLING



$$V^* = \{ 2, 20, 24 \}$$
$$E^* = \{ \{ 1, 13, 15, 16, 17, 18 \}, \dots \}$$

UNLABELED STAR SAMPLING

1. Sample n times **without replacement** from V to obtain V^* .
2. Add incident edges

$$E^* := \{e \in E : \exists v \in V^* \text{ such that } v \in e\}.$$

Vertex inclusion probabilities:

- ▶ For $i \in V$, $\pi_i = \frac{n}{N_V}$.

UNLABELED STAR SAMPLING

1. Sample n times **without replacement** from V to obtain V^* .
2. Add incident edges

$$E^* := \{e \in E : \exists v \in V^* \text{ such that } v \in e\}.$$

- ▶ For $e = \{i, j\} \in E \subseteq V^{(2)}$, (**edge (!)** inclusion probabilities)

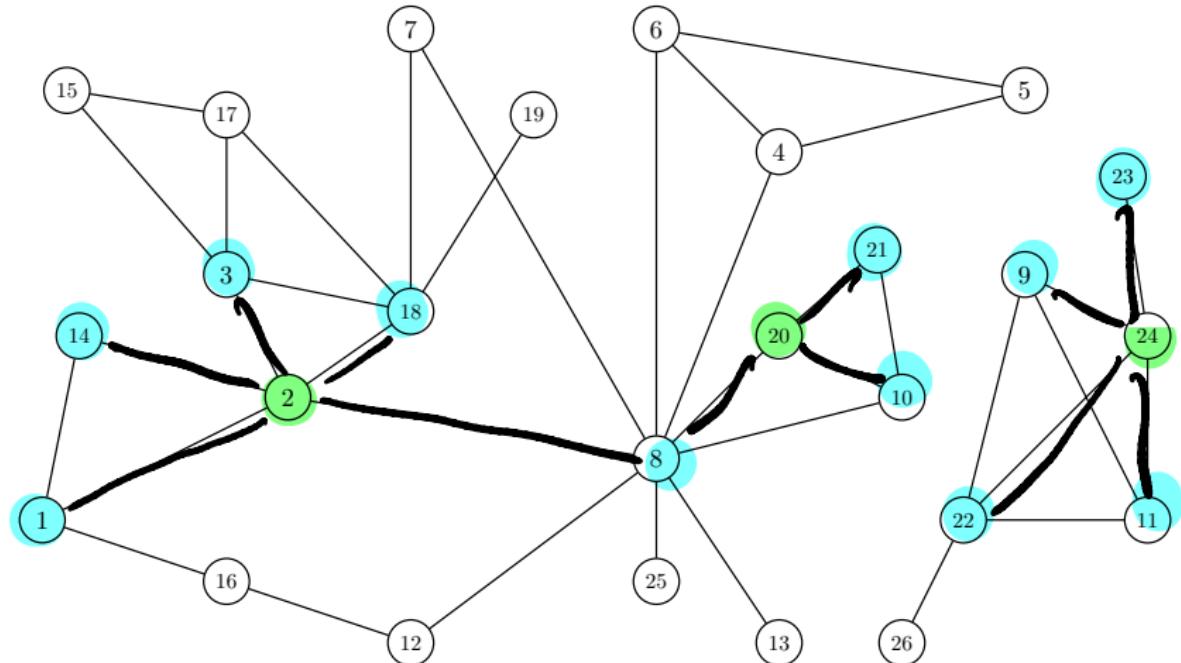
$$\begin{aligned}\pi_e &= P(\text{edge } e = \{i, j\} \text{ is sampled}) \\ &= 1 - P(\text{neither vertex } i \text{ nor vertex } j \text{ is sampled}) \\ &= 1 - \frac{\binom{N_V-2}{n}}{\binom{N_V}{n}}.\end{aligned}$$

- ▶ **However, for vertex pairs** $\{i, j\} \in V^{(2)}$, we have

$$\pi_{\{i,j\}} = P(\text{vertex pair } \{i, j\} \text{ is sampled}) = \frac{n(n-1)}{N_V(N_V-1)},$$

as in induced subgraph sampling.

LABELED STAR SAMPLING



LABELED STAR SAMPLING

1. Sample n times **without replacement** from V to obtain V_0^* .

2. Add incident edges

$$E^* := \{e \in E : \exists v \in V_0^* \text{ such that } v \in e\}.$$

3. Add the vertices incident to E^*

$$V^* := \{v \in V : \exists e \in E^* \text{ such that } v \in e\}.$$

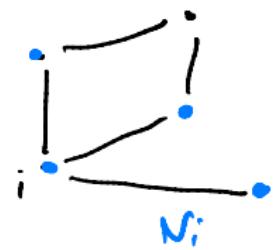
Edge inclusion probabilities (as in the unlabeled case):

► For $e \in E$,

$$\pi_e = 1 - \frac{\binom{N_V-2}{n}}{\binom{N_V}{n}}.$$

LABELED STAR SAMPLING

1. Sample n times **without replacement** from V to obtain V_0^* .
2. Add incident edges
 $E^* := \{e \in E : \exists v \in V_0^* \text{ such that } v \in e\}.$
3. Add the vertices incident to E^*
 $V^* := \{v \in V : \exists e \in E^* \text{ such that } v \in e\}.$



Vertex inclusion probabilities:

- For $i \in V$, define $\underline{N}_i := \{j \in V : \text{dist}(i, j) \leq 1\}$. Thus,
 $|N_i| = d_i + 1$ and

$$\begin{aligned}\pi_i &= P(\text{vertex } i \text{ is sampled}) \\ &= 1 - P(\text{no vertex from } N_i \text{ is contained in } V_0^*) \\ &= 1 - \frac{\binom{N_V - (d_i + 1)}{n}}{\binom{N_V}{n}}.\end{aligned}$$

Example: How political are TV-shows?

EXAMPLE: HOW POLITICAL ARE TV-SHOWS?

UCI 

Machine Learning Repository
Center for Machine Learning and Intelligent Systems

About Citation Policy Donate a Data Set Contact
 Search
 Repository Web Google

[View ALL Data Sets](#)

Facebook Large Page-Page Network Data Set

[Download](#) [Data Folder](#) [Data Set Description](#)

Abstract: This webgraph is a page-page graph of verified Facebook sites. Nodes represent official Facebook pages while the links are mutual likes between sites.

Data Set Characteristics:	Multivariate	Number of Instances:	22470	Area:	Social
Attribute Characteristics:	N/A	Number of Attributes:	4714	Date Donated	2020-07-22
Associated Tasks:	Classification	Missing Values?	N/A	Number of Web Hits:	27883

Source:

Benedek Rozemberczki
benedek.rozemberczki@gmail.com
The University of Edinburgh

Data Set Information:

Node features are extracted from the site descriptions that the page owners created to summarize the purpose of the site. This graph was collected through the Facebook Graph API in November 2017 and restricted to pages from 4 categories which are defined by Facebook. These categories are: politicians, governmental organizations, television shows and companies. The task related to this dataset is multi-class node classification for the 4 site categories. Provide all relevant information about your data set.

Attribute Information:

<https://archive.ics.uci.edu/ml/datasets/>

Facebook+Large+Page–Page+Network

EXAMPLE:

POLITIZATION OF TV-SHOWS

- ▶ Population graph: 22 470 Facebook pages of politicians, government agencies, TV-shows and companies.
- ▶ Links/edges are mutual likes (sparse).
- ▶ Year: 2019
- ▶ Question: How strongly connected are TV-shows and political agents (politicians or government agencies)?
- ▶ Formally: For an edge $e \in E$, set $y_e = 1$ if e connects a TV-show with a political agent, and set $y_e = 0$ otherwise.
Compute

$$\tau := \sum_{e \in E} y_e, \quad \mu := \frac{\tau}{N_E}.$$

- ▶ $\mathcal{U} = E$
- ▶ Assume that we don't have access to the edge set E directly, but only through querying vertex neighbors.
- ▶ Consider both known and unknown N_E ($N_E = 171\,002$).

1.) INDUCED SUBGRAPH SAMPLING

- ▶ Draw V^* of size n uniformly from V without replacement.
- ▶ Choose $E^* = \{\{u, v\} \in E : u \in V^* \wedge v \in V^*\}$.
- ▶ Here, the vertex pair inclusion probability $\pi_{\{i,j\}}$ equals the edge inclusion probability $\pi_e = \pi = \frac{n(n-1)}{N_V(N_V-1)}$.
- ▶ Horvitz-Thompson estimate

$$\begin{aligned}\hat{\tau} &= \sum_{e \in E^*} \frac{y_e}{\pi_e} = \frac{1}{\pi} \sum_{e \in E^*} y_e \\ &= \frac{N_V(N_V - 1)}{n(n - 1)} \cdot \# \{e \in E^* : e \text{ is a TV/politics pair}\}.\end{aligned}$$

For moderately large n we often find $\hat{\tau} = 0$.

ESTIMATING μ AND N_E

- If N_E is known,

$$\hat{\mu} = \frac{\hat{\tau}}{N_E}$$

is an unbiased estimator for μ .

- If $N_E = \sum_{e \in E} 1$ is unknown, we can estimate it as a population total using the Horvitz-Thompson approach

$$\hat{N}_E := \sum_{e \in E^*} \frac{1}{\pi_e} = \frac{|E^*|}{\pi}, \quad y_e = 1$$

with

$$\mathbb{E}[\hat{N}_E] = \sum_{e \in E} y_e = N_E,$$

and set

$$\hat{\mu} = \frac{\hat{\tau}}{\hat{N}_E} = \frac{\# \{e \in E^* : e \text{ is a TV/politics pair}\}}{|E^*|}.$$

2.) LABELED STAR SAMPLING

- ▶ Draw V_0^* of size n uniformly from V without replacement.
- ▶ Choose $E^* = \{\{u, v\} \in E : u \in V_0^* \vee v \in V_0^*\}$.
- ▶ Take $V^* := \{v \in V : \exists e \in E^* \text{ such that } v \in e\}$.
- ▶ Edge inclusion probability
$$\pi_e = \pi = 1 - \frac{\binom{N_V-2}{n}}{\binom{N_V}{n}} = 1 - \frac{(N_V-n)(N_V-n-1)}{N_V(N_V-1)}.$$
- ▶ Horvitz-Thompson estimate

$$\begin{aligned}\hat{\tau} &= \sum_{e \in E^*} \frac{y_e}{\pi_e} = \frac{1}{\pi} \sum_{e \in E^*} y_e \\ &= \frac{\# \{e \in E^* : e \text{ is a TV/politics pair}\}}{\pi}.\end{aligned}$$

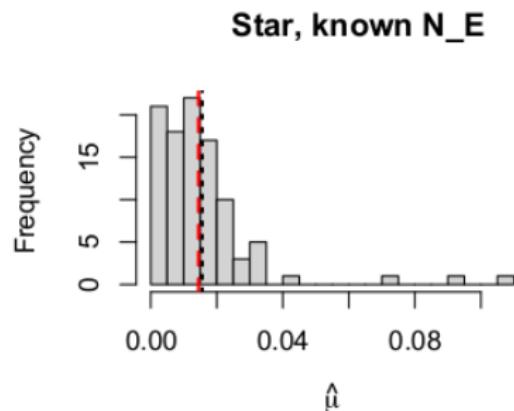
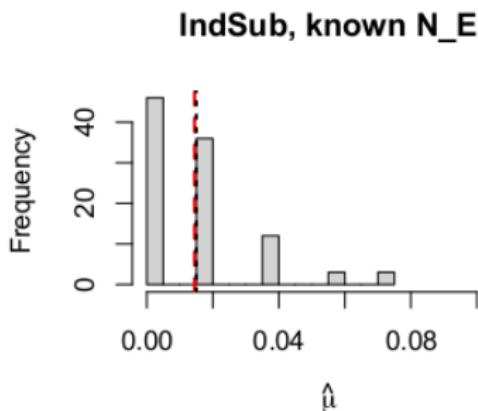
For the same n as in 1.) this searches through a lot more edges and thus has much higher chance of discovering any TV/politics edges.

MC SIMULATIONS

- ▶ Investigate the sampling distribution of $\hat{\mu}$ in our different scenarios.
- ▶ For the same $n = |V_{\bullet}^*|$, unlabeled star sampling is computationally more expensive than induced subgraph sampling.
- ▶ We simulate such that both sampling designs have similar runtime and compare their statistical performance.

MC SIMULATIONS

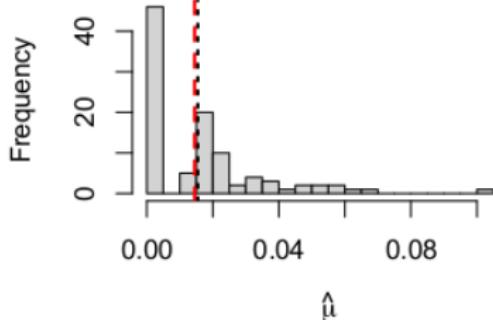
$MC = 100$	induced subgraph	labeled star
$n = V_0^* $	400	50
avg. $ E^* $	53.17	768.49
avg. runtime	0.49	0.38



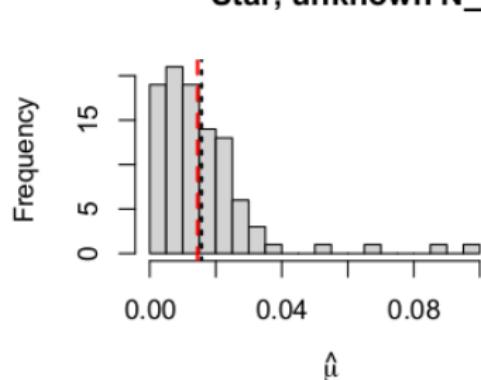
MC SIMULATIONS

$MC = 100$	induced subgraph	labeled star
$n = V^* $	400	50
avg. $ E^* $	53.17	768.49
avg. runtime	0.49	0.38

IndSub, unknown N_E

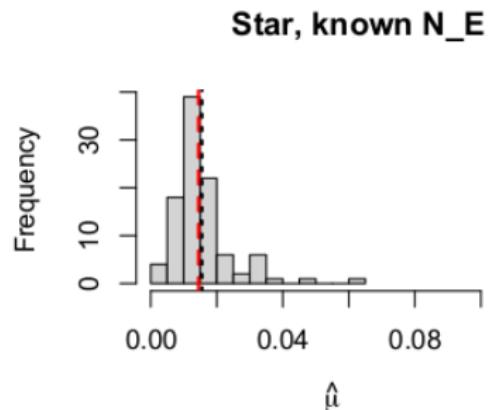
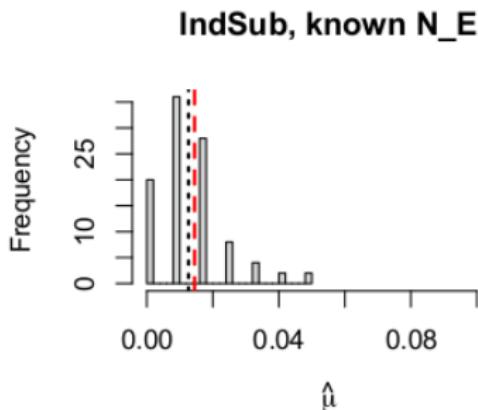


Star, unknown N_E



MC SIMULATIONS

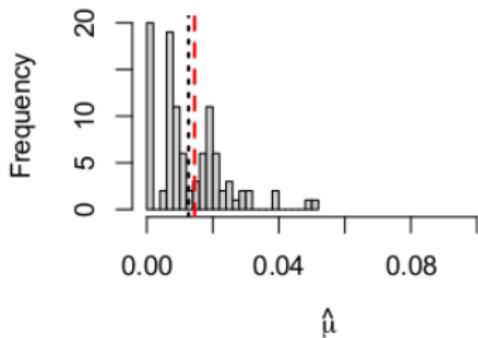
$MC = 100$	induced subgraph	labeled star
$n = V^* $	600	100
avg. $ E^* $	124.11	1550.64
avg. runtime	1.27	1.74



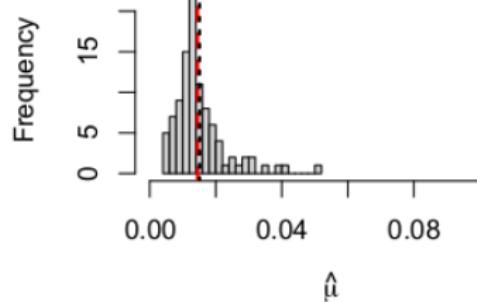
MC SIMULATIONS

$MC = 100$	induced subgraph	labeled star
$n = V^* $	600	100
avg. $ E^* $	124.11	1550.64
avg. runtime	1.27	1.74

IndSub, unknown N_E

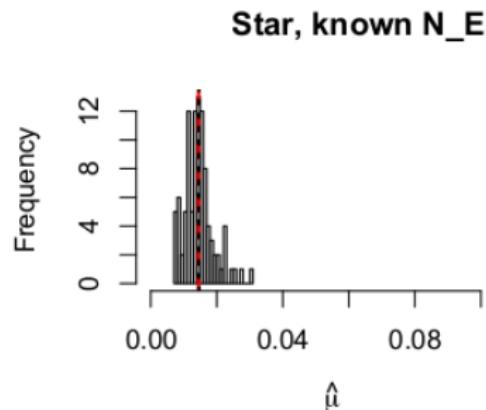
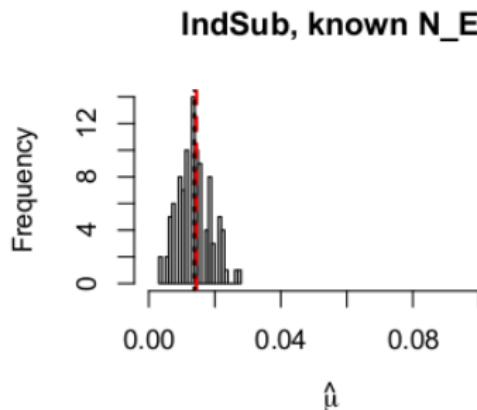


Star, unknown N_E



MC SIMULATIONS

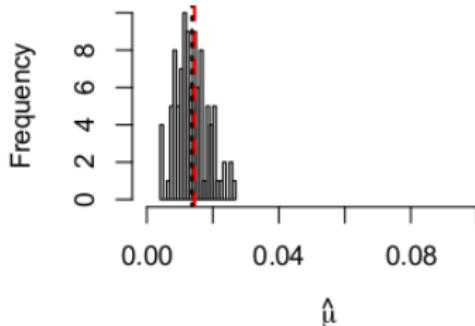
$MC = 100$	induced subgraph	labeled star
$n = V^* $	1500	300
avg. $ E^* $	767.68	4505.11
avg. runtime	17.51	16.51



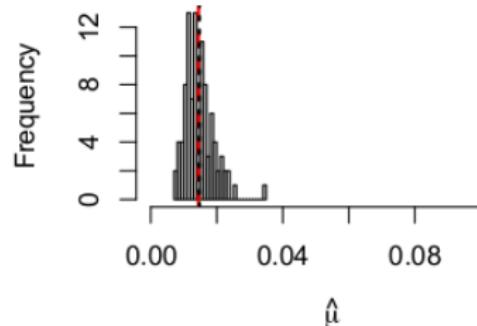
MC SIMULATIONS

$MC = 100$	induced subgraph	unlabeled star
$n = V^* $	1500	300
avg. $ E^* $	767.68	4505.11
avg. runtime	17.51	16.51

IndSub, unknown N_E

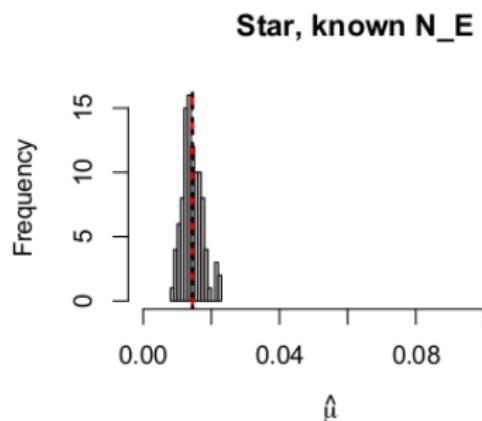
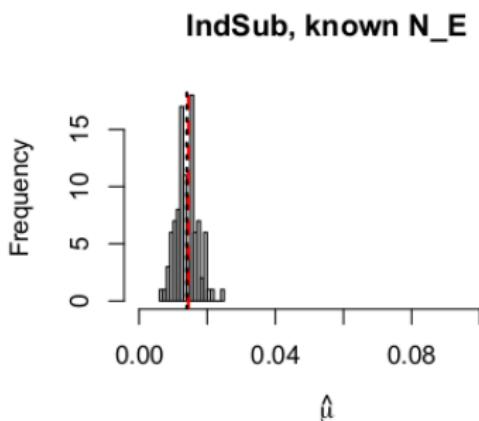


Star, unknown N_E



MC SIMULATIONS

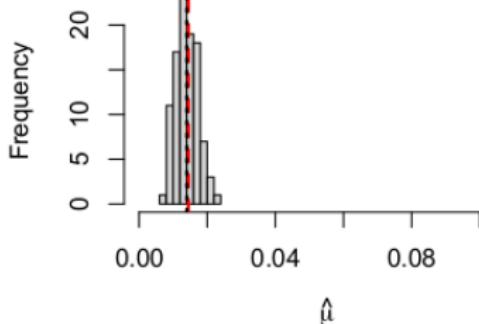
$MC = 100$	induced subgraph	labeled star
$n = V^* $	2800	800
avg. $ E^* $	2663.96	11839.38
avg. runtime	87.41	91.21



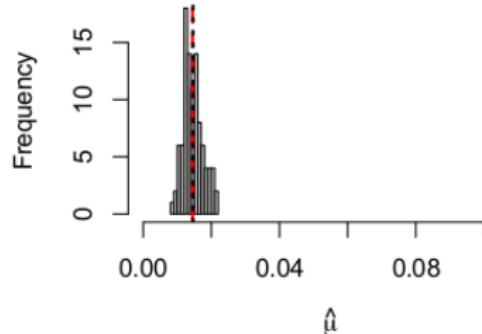
MC SIMULATIONS

$MC = 100$	induced subgraph	labeled star
$n = V^* $	2800	800
avg. $ E^* $	2663.96	11839.38
avg. runtime	87.41	91.21

IndSub, unknown N_E



Star, unknown N_E



Towards uncertainty quantification for graph sampling

TOWARDS UNCERTAINTY QUANTIFICATION

- ▶ We continue with the Facebook example.
- ▶ For sufficiently large n , the sampling distributions of our estimators look symmetric and bell shaped with no serious outliers.
- ▶ This motivates a normal approximation.
- ▶ A rigorous mathematical motivation is beyond the scope of this course.
- ▶ Recall the approximate Gaussian CI

$$CI_{\alpha} = \hat{\mu} \pm q_{1-\frac{\alpha}{2}}^{(N)} \hat{s}e.$$

- ▶ We need an estimate $\hat{s}e = \hat{s}e(\hat{\mu})$ of the standard error of our estimator $\hat{\mu}$.

TOWARDS UNCERTAINTY QUANTIFICATION

- ▶ Recall homework: For $\mathcal{U} = [N]$ and $S = (i_1, \dots, i_n) \in \mathcal{U}^n$,

$$\hat{s}e^2 = \sum_{k=1}^n \sum_{l=1}^n y_{i_k} y_{i_l} \left(\frac{1}{\pi_{i_k} \pi_{i_l}} - \frac{1}{\pi_{i_k i_l}} \right)$$

is unbiased for $\text{Var}[\hat{\tau}]$, where $\hat{\tau} = \sum_{j=1}^n \frac{y_{i_j}}{\pi_{i_j}}$.

- ▶ Therefore,

$$\hat{s}e^2 / N^2 = \frac{1}{N^2} \sum_{k=1}^n \sum_{l=1}^n y_{i_k} y_{i_l} \left(\frac{1}{\pi_{i_k} \pi_{i_l}} - \frac{1}{\pi_{i_k i_l}} \right)$$

is unbiased for $\text{Var}[\hat{\mu}] = \text{Var}[\hat{\tau}/N] = \text{Var}[\hat{\tau}]/N^2$.

TOWARDS UNCERTAINTY QUANTIFICATION

- Now: $\boxed{\mathcal{U} = E}$, $N = N_E$, $S = E^*$ and

$$\widehat{se}^2(\hat{\mu}) \quad \cancel{se^2} := \frac{1}{N_E^2} \sum_{e \in E^*} \sum_{f \in E^*} y_e y_f \left(\frac{1}{\pi_e \pi_f} - \frac{1}{\pi_{ef}} \right)$$

is unbiased for $\text{Var}[\hat{\mu}]$, where $\hat{\mu} := \frac{1}{N_E} \sum_{e \in E^*} \frac{y_e}{\pi_e}$.

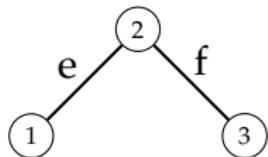
- If N_E is unknown, we estimate it as before by $\hat{N}_E = |E^*|/\pi$.
- We have already computed edge inclusion probabilities π_e .
- We need also edge-pair inclusion probabilities π_{ef} for $e, f \in E$.

EDGE PAIR INCLUSION PROBABILITIES

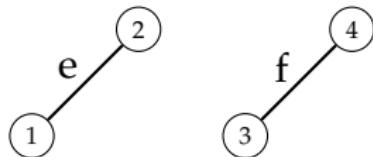
$$\begin{aligned}\Pi_{ef} &= P(e \text{ and } f \text{ are sampled}) \\ &= P(e, f \in E^*)\end{aligned}$$

- There are two different kinds of edge pairs (e, f) :

A



B

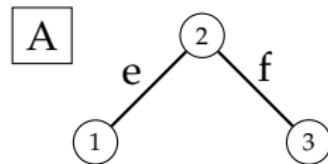


EDGE PAIR INCLUSION PROBABILITIES

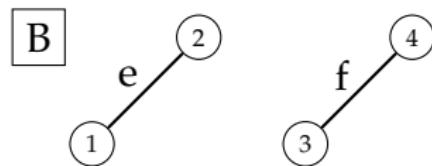
Induced subgraph sampling:

$$n = |V^*|$$

$$\pi_{ef} = P(e \in E^* \text{ and } f \in E^*) =$$



$$= P(1, 2, 3 \in V^*) = \frac{\binom{N_V - 3}{n-3}}{\binom{N_V}{n}} = \frac{n(n-1)(n-2)}{N_V(N_V-1)(N_V-2)}.$$



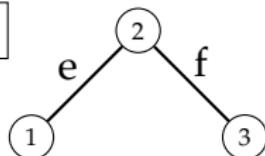
$$= P(1, 2, 3, 4 \in V^*) = \frac{\binom{N_V - 4}{n-4}}{\binom{N_V}{n}} = \frac{n(n-1)(n-2)(n-3)}{N_V(N_V-1)(N_V-2)(N_V-3)}.$$

EDGE PAIR INCLUSION PROBABILITIES

Labeled star sampling:

$$\pi_{ef} = P(e \in E^* \text{ and } f \in E^*) =$$

A



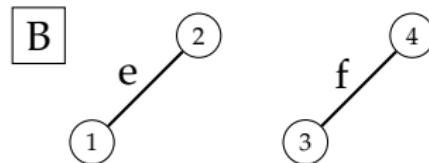
$$= P(\underbrace{\{1, 2, 3 \in V^*\}}_{\text{all three}}, \underbrace{\{1, 2 \in V^*, 3 \notin V^*\}}_{\text{e only}}, \underbrace{\{1, 3 \in V^*, 2 \notin V^*\}}_{\text{f only}} \\ \cup \underbrace{\{2, 3 \in V^*, 1 \notin V^*\}}_{\text{neither}}, \underbrace{\{2 \in V^*, 1, 3 \notin V^*\}}_{\text{f only}})$$

$$\pi = \frac{\binom{N_v - 3}{n-3} + 3 \cdot \binom{N_v - 3}{n-2} + \binom{N_v - 3}{n-1}}{\binom{N_v}{n}}$$

EDGE PAIR INCLUSION PROBABILITIES

Labeled star sampling:

$$\pi_{ef} = P(e \in E^* \text{ and } f \in E^*) =$$



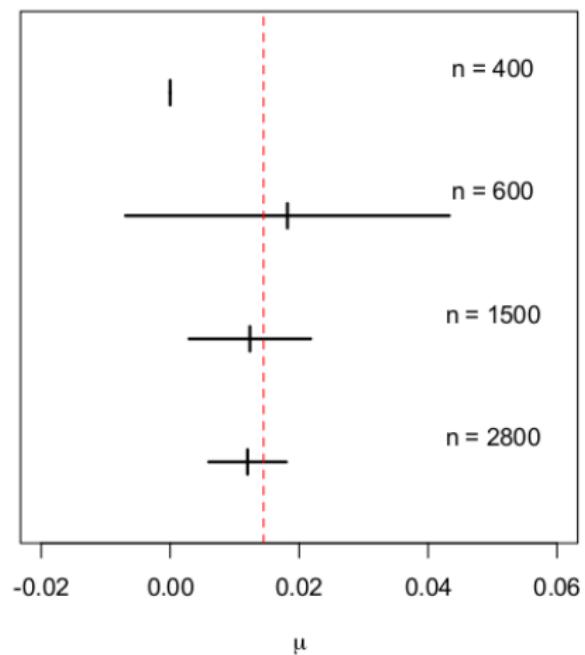
$$= P(\{1, 2, 3, 4 \in V^*\}) + 4 \cdot P(\{1, 2, 3 \in V^*, 4 \notin V^*\}) \\ + 4 \cdot P(\{1, 3 \in V^*, 2, 4 \notin V^*\})$$

$$= \frac{\binom{N_V - 4}{n-4} + 4 \cdot \binom{N_V - 4}{n-3} + 4 \cdot \binom{N_V - 4}{n-2}}{\binom{N_V}{n}}$$

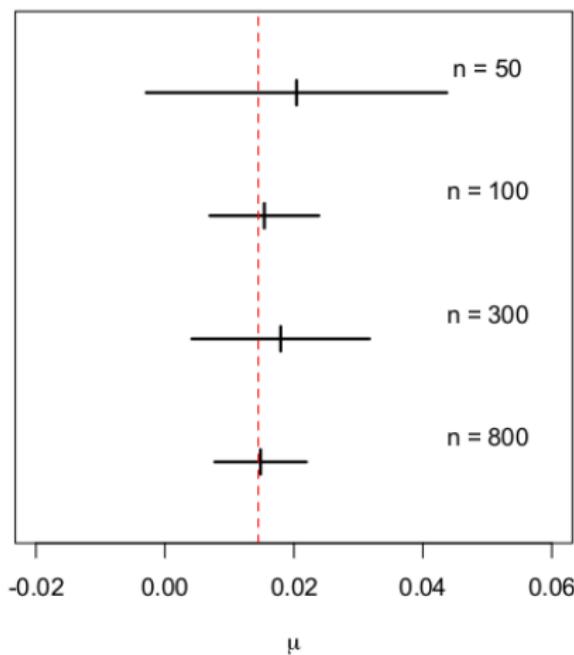
95% CI FOR μ

μ = proportion of edges that connect a TV show with a politician or government institution.

induced subgraph sampling



labeled star sampling



Statistics for Data Science, WS2023

Chapter 6:
Differential Privacy

OVERVIEW

Issues of data privacy protection

Definition of Differential Privacy

Designing ϵ -DP mechanisms

Approximate differential privacy

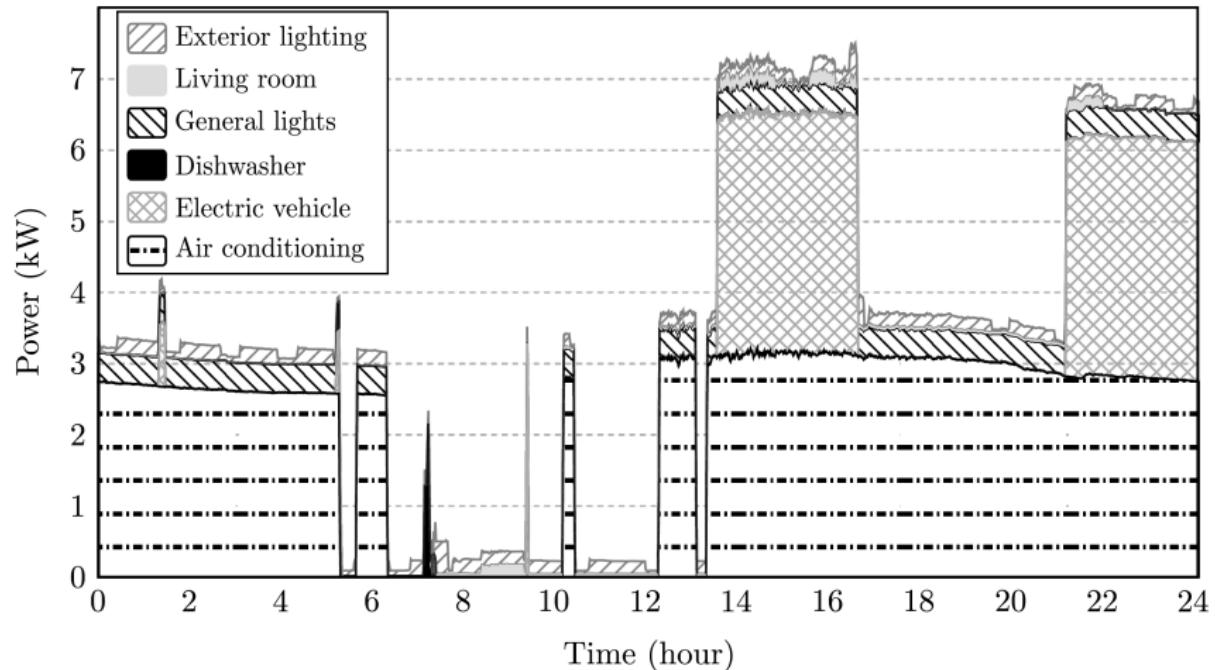
Issues of data privacy protection

ISSUES OF DATA PRIVACY PROTECTION

This is an old problem with increasing relevance in the modern era of big data. For instance:

- ▶ official statistics
- ▶ large scale medical research
- ▶ smart phone user data
- ▶ social media data
- ▶ IoT
- ▶ etc.

EXAMPLE: DATA FROM SMART METER



ISSUES OF DATA PRIVACY PROTECTION

Traditional solutions:

- ▶ anonymize
- ▶ aggregate

ANONYMIZATION -> 'RE-IDENTIFICATION ATTACKS'

Example: Student survey

ID	age	sex	#sib.	firstSem.	best	worst	cheated	drugs
01622490	21	f	3	SS2017	1	5	no	no
10628491	23	m	1	WS2017	1	5	yes	yes
14937612	24	m	1	WS2017	1	4	no	no
11274513	23	f	1	SS2017	1	3	yes	yes
09663822	20	f	0	WS2017	1	2	no	yes
⋮								
07257738	21	m	0	WS2017	1	1	no	yes

$$n = 24$$

ANONYMIZATION -> 'RE-IDENTIFICATION ATTACKS'

Example: Student survey

ID	age	sex	#sib.	firstSem.	best	worst	cheated	drugs
_____	21	f	3	SS2017	1	5	no	no
_____	23	m	1	WS2017	1	5	yes	yes
_____	24	m	1	WS2017	1	4	no	no
_____	23	f	1	SS2017	1	3	yes	yes
_____	20	f	0	WS2017	1	2	no	yes
⋮								
_____	21	m	0	WS2017	1	1	no	yes

$$n = 24$$

ANONYMIZATION -> 'RE-IDENTIFICATION ATTACKS'

```
> # use only age  
> agg <- aggregate(data$age, by=data[ "age" ], length)  
> agg  
age x  
1 20 1  
2 21 6  
3 22 4  
4 23 4  
5 24 1  
6 25 2  
7 26 1  
8 27 4  
9 31 1
```

ANONYMIZATION -> 'RE-IDENTIFICATION ATTACKS'

```
> (agg <- aggregate(data$age, by=data[c("sex",  
"age")], length))  
    sex age x  
1     f   20 1  
2     f   21 5  
3     m   21 1  
4     f   22 2  
5     m   22 2  
6     m   23 4  
7     f   24 1  
8     f   25 1  
9     m   25 1  
10    m   26 1  
11    f   27 2  
12    m   27 2  
13    f   31 1  
> sum(agg$x==1)/n # fraction uniquely identified  
[1] 0.2916667
```

ANONYMIZATION -> 'RE-IDENTIFICATION ATTACKS'

```
> # sex, age, first semester
> agg <- aggregate(data$age, by=data[c("sex", "age",
  "start")], length)
> sum(agg$x==1)/n # fraction uniquely identified
[1] 0.625

> # sex, age, first semster, worst grade
> agg <- aggregate(data$age, by=data[c("sex", "age",
  "start", "worst")], length)
> sum(agg$x==1)/n # fraction uniquely identified
[1] 0.6666667

> # sex, age, first semester, worst grade, #siblings
> agg <- aggregate(data$age, by=data[c("sex", "age",
  "start", "siblings", "worst")], length)
> sum(agg$x==1)/n # fraction uniquely identified
[1] 0.75
```

ANONYMIZATION -> 'RE-IDENTIFICATION ATTACKS'

- ▶ personal identifiers may look unsuspicious (e.g., age)
- ▶ **sets** of attributes/variables can be personal identifiers
- ▶ **auxiliary information** may be available
- ▶ the problem worsens for **high-dimensional data**

Real world examples:

Narayanan, A. and Shmatikov, V. (2006). How to break anonymity of the netflix prize dataset. arXiv preprint cs/0610105.

Sweeney L, Abu A, and Winn J. (2013). Identifying Participants in the Personal Genome Project by Name. Harvard University. Data Privacy Lab. White Paper 1021-1.

AGGREGATION -> DE-AGGREGATION

- ▶ Publish only summary statistics: $S_n = \sum_{i=1}^n X_i$.

Statistical agencies compute sensitivity/privacy measures:
e.g., p-percent rule: for $X_i \geq 0$, (e.g., revenue of companies)

$$\frac{X_{(n)}}{\sum_{i \neq n-1} X_{(i)}} > p.$$

Worst case: $S_n - \sum_{i=2}^n X_i = X_1$.

Whether S_n is publishable depends on the original data X_1, \dots, X_n . What is the 'correct' sensitivity measure?

AGGREGATION -> DIFFERENCING

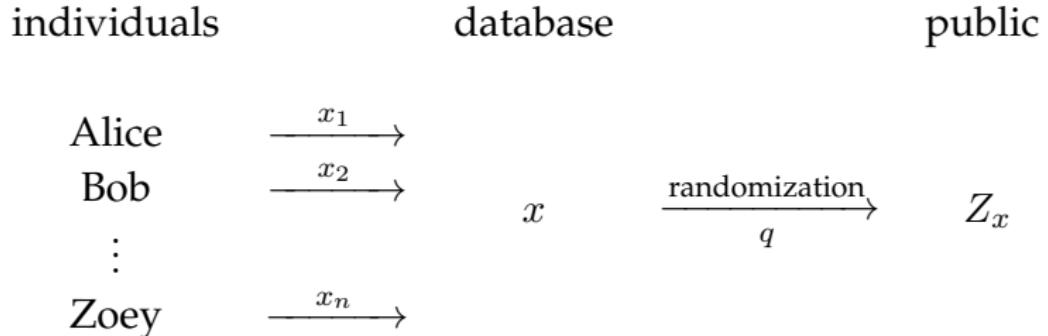
- ▶ Publish only answers to counting queries.

How many students are 21 and male? Answer: 1

How many students are 21, male and took drugs? Answer: 1

Definition of Differential Privacy

DEFINITION OF DIFFERENTIAL PRIVACY



For each possible database $x \in \mathcal{X}^n$ with n rows, we specify a randomization mechanism, that is, a random variable Z_x taking values in some output space \mathcal{Z} . Typically $\mathcal{Z} \subseteq \mathbb{R}^m$.

Z_x should not depend too much on any individual contribution x_i .

DEFINITION OF DIFFERENTIAL PRIVACY

For $x, x' \in \mathcal{X}^n$, define the Hamming distance

$$d_0(x, x') := |\{i : x_i \neq x'_i\}|.$$

Definition (Dwork et al. 2006)

Fix a privacy parameter $\varepsilon \in (0, \infty)$. The randomization mechanism outputting Z_x on \mathcal{Z} for a given $x \in \mathcal{X}^n$, is called ε -differentially private if for all $x, x' \in \mathcal{X}^n$ with $d_0(x, x') \leq 1$, we have

$$\mathbb{P}(Z_x \in A) \leq e^\varepsilon \cdot \mathbb{P}(Z_{x'} \in A), \quad \forall A \subseteq \mathcal{Z} \text{ (measurable)}.$$

We call Z_x an ε -differentially private view of $x \in \mathcal{X}^n$.

DEFINITION OF DIFFERENTIAL PRIVACY

The idea is the following:

- ▶ If the true database is $x \in \mathcal{X}^n$, the distribution of the output Z_x (in case $\mathcal{Z} = \mathbb{R}$) has cdf
 $F_x(t) := \mathbb{P}(Z_x \leq t) = \mathbb{P}(Z_x \in (-\infty, t]).$
- ▶ If I decide not to contribute my data x_i and the corresponding row of x is erased (x_i set to an arbitrary value), we obtain a new database, $x' \in \mathcal{X}^n$, say, with $x_i \neq x'_i$, that is, $d_0(x, x') = 1$.
- ▶ If Z_x is ε -DP, then

$$e^{-\varepsilon} \leq \frac{F_x(t)}{F_{x'}(t)} \leq e^\varepsilon, \quad \forall t \in \mathbb{R}.$$

- ▶ If ε is close to 0, this means that $F_x \approx F_{x'}$.
- ▶ Thus, the distribution of the output Z_x is almost the same, no matter if I contribute my data or not.

VERIFYING DIFFERENTIAL PRIVACY USING A PDF OR PMF

For given $x \in \mathcal{X}^n$, let $q(\cdot|x)$ be a pdf or pmf of Z_x satisfying

$$q(z|x) \leq e^\varepsilon q(z|x'), \quad \forall z \in \mathcal{Z}, \forall x, x' \in \mathcal{X}^n : d_0(x, x') \leq 1.$$

Then, for every (measurable) $A \subseteq \mathcal{Z}$ and every $x, x' \in \mathcal{X}^n$ with $d_0(x, x') \leq 1$,

$$\mathbb{P}(Z_x \in A) = \int_A q(z|x) dz \leq \int_A e^\varepsilon q(z|x') dz = e^\varepsilon \mathbb{P}(Z_{x'} \in A), \quad (\text{pdf})$$

$$\mathbb{P}(Z_x \in A) = \sum_{z \in A} q(z|x) \leq \sum_{z \in A} e^\varepsilon q(z|x') = e^\varepsilon \mathbb{P}(Z_{x'} \in A), \quad (\text{pmf})$$

EXAMPLE: SAMPLE MEAN

- ▶ Data: $x = (x_1, \dots, x_n)^T \in \mathbb{R}^n$, $-M \leq x_i \leq M$
- ▶ We want to publish $f(x) := \bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$.
- ▶ Let $W \sim Laplace(\theta)$, with pdf $g_\theta(z) := \frac{\theta}{2} e^{-\theta|z|}$, $\theta > 0$.
- ▶ Publish $Z_x = f(x) + Laplace\left(\frac{n\varepsilon}{2M}\right)$.

$$q(z|x) = \frac{n\varepsilon}{4M} e^{-\frac{n\varepsilon}{2M}|z-f(x)|}$$

PROPERTIES OF DIFFERENTIAL PRIVACY

Proposition 5.1 (post-processing)

If Z_x is an ε -differentially private view of $x \in \mathcal{X}^n$ and $h : \mathcal{Z} \rightarrow \mathcal{Z}'$, then $h(Z_x)$ is also an ε -differentially private view of x .

Proposition 5.2 (sequential composition)

If $Z_x^{(1)}$ is an ε_1 -DP view of $x \in \mathcal{X}^n$ and $Z_x^{(2)}$ is an ε_2 -DP view of $x \in \mathcal{X}^n$, independent of $Z_x^{(1)}$, then $Z_x = (Z_x^{(1)}, Z_x^{(2)})$ is an $\varepsilon_1 + \varepsilon_2$ -DP view of x .

Proposition 5.3 (parallel composition)

For $x = (x_1, \dots, x_n)^T \in \mathcal{X}^n$, write $\xi = (x_1, \dots, x_{n_1})^T \in \mathcal{X}^{n_1}$ and $\zeta = (x_{n_1+1}, \dots, x_n)^T \in \mathcal{X}^{n-n_1}$. If Z_ξ is an ε_1 -DP view of ξ and Z_ζ is an ε_2 -DP view of ζ , independent of Z_ξ , then $Z_x = (Z_\xi, Z_\zeta)$ is a $\max(\varepsilon_1, \varepsilon_2)$ -DP view of x .

PROPERTIES OF DIFFERENTIAL PRIVACY

Designing ε -DP mechanisms

SENSITIVITY OF QUERY FUNCTIONS

- ▶ Data: $x \in \mathcal{X}^n$
- ▶ Analyst would like to know $f(x)$ for some *query function* $f : \mathcal{X}^n \rightarrow \mathbb{R}$ (e.g., $f(x) = \bar{x}_n$).
- ▶ Define the (*global*) *sensitivity* of f by

$$\Delta_f := \sup_{\substack{x, x' \in \mathcal{X}^n \\ d_0(x, x') \leq 1}} |f(x) - f(x')|.$$

- ▶ Publish $Z_x = f(x) + \text{Laplace}\left(\frac{\varepsilon}{\Delta_f}\right)$.

$$q(z|x) = \frac{\varepsilon}{2\Delta_f} \exp\left(-\frac{\varepsilon}{\Delta_f}|z - f(x)|\right)$$

EXAMPLES OF (GLOBAL) SENSITIVITIES

Data: $x = (x_1, \dots, x_n)^T \in \mathbb{R}^n, -M \leq x_i \leq M.$

- ▶ $f(x) = \bar{x}_n$
- ▶ $\Delta_f = \frac{2M}{n}$

- ▶ $f(x) = \max\{x_1, \dots, x_n\}$
- ▶ $\Delta_f = 2M$

- ▶ $f(x) = \text{med}(x), n \text{ odd } (\text{med}(x) = x_{(\frac{n+1}{2})}).$
- ▶ $\Delta_f = 2M$

Attention: $M = M(x) := \max_i |x_i|$ is not allowed!!!

LOCAL SENSITIVITY OF QUERY FUNCTIONS

- ▶ Data: $x \in \mathcal{X}^n$
- ▶ Analyst would like to know $f(x)$ for some *query function* $f : \mathcal{X}^n \rightarrow \mathbb{R}$ (e.g., $f(x) = \bar{x}_n$).
- ▶ Define the *global sensitivity* of f by

$$\Delta_f := \sup_{\substack{x, x' \in \mathcal{X}^n \\ d_0(x, x') \leq 1}} |f(x) - f(x')|.$$

- ▶ Define the *local sensitivity* of f at $x \in \mathcal{X}^n$ by

$$\Delta_f(x) := \sup_{\substack{x' \in \mathcal{X}^n \\ d_0(x, x') \leq 1}} |f(x) - f(x')|.$$

EXAMPLES OF LOCAL SENSITIVITIES

Data: $x = (x_1, \dots, x_n)^T \in \mathbb{R}^n$, $-M \leq x_i \leq M$.

- ▶ $f(x) = \bar{x}_n$
- ▶ $\Delta_f(x) = \frac{1}{n} \max_i \max\{|-M - x_i|, |M - x_i|\} \in [\frac{M}{n}, \frac{2M}{n}]$
- ▶ $f(x) = \max\{x_1, \dots, x_n\} = x_{(n)}$
- ▶ $\Delta_f(x) = \max\{M - x_{(n)}, x_{(n)} - x_{(n-1)}\} \in [0, 2M]$
- ▶ $f(x) = \text{med}(x)$, n odd ($\text{med}(x) = x_{(\frac{n+1}{2})}$).
- ▶ $\Delta_f(x) = \max\{x_{(\frac{n+1}{2}+1)} - x_{(\frac{n+1}{2})}, x_{(\frac{n+1}{2})} - x_{(\frac{n+1}{2}-1)}\} \in [0, 2M]$

LOCAL SENSITIVITY OF QUERY FUNCTIONS

- ▶ Define the *local sensitivity* of f at $x \in \mathcal{X}^n$ by

$$\Delta_f(x) := \sup_{\substack{x' \in \mathcal{X}^n \\ d_0(x, x') \leq 1}} |f(x) - f(x')|.$$

Releasing

$$Z_x = f(x) + \text{Laplace} \left(\frac{\varepsilon}{\Delta_f(x)} \right)$$

is not ε -DP!

See “*approximate DP*” below!

INVERSE SENSITIVITY OF QUERY FUNCTIONS

- ▶ Data: $x \in \mathcal{X}^n$
- ▶ Analyst would like to know $f(x)$ for some *query function* $f : \mathcal{X}^n \rightarrow \mathbb{R}$.
- ▶ Define the *range* of f by $\mathcal{F} := f(\mathcal{X}^n) := \{f(x) : x \in \mathcal{X}^n\}$.
- ▶ Define the *inverse local sensitivity* of f at $(x, z) \in \mathcal{X}^n \times \mathcal{F}$ by

$$\Delta_f^{-1}(x, z) := \min\{d_0(x, x') : f(x') = z, x' \in \mathcal{X}^n\}.$$

INVERSE SENSITIVITY FOR FINITE \mathcal{F}

- ▶ Define the *inverse local sensitivity* of f at $(x, z) \in \mathcal{X}^n \times \mathcal{F}$ by

$$\Delta_f^{-1}(x, z) := \min\{d_0(x, x') : f(x') = z, x' \in \mathcal{X}^n\}.$$

Then Z_x distributed with pmf

$$q(z|x) := \frac{\exp(-\frac{\varepsilon}{2}\Delta_f^{-1}(x, z))}{\sum_{u \in \mathcal{F}} \exp(-\frac{\varepsilon}{2}\Delta_f^{-1}(x, u))}, \quad z \in \mathcal{F}, x \in \mathcal{X}^n,$$

is ε -DP and

$$q(f(x)|x) \geq q(z|x), \quad \forall z \in \mathcal{F}.$$

INVERSE SENSITIVITY FOR DISCRETE \mathcal{F}

Proof:

EXAMPLE: INVERSE SENSITIVITY OF COUNTING QUERY

$$f(x) = \sum_{i=1}^n \mathbb{1}_A(x_i) \in \mathcal{F} = [n]$$

$$\Delta_f^{-1}(x, z) = \min\{d_0(x, x') : f(x') = z, x' \in \mathcal{X}^n\} =$$

REPEATED QUERIES OF THE SAME DATABASE

- ▶ Data: $x = (x_1, \dots, x_n)^T \in [-M, M]^n$
- ▶ m Analysts want to compute $f(x) = \bar{x}_n$.
- ▶ Let $W^{(1)}, \dots, W^{(m)} \stackrel{i.i.d.}{\sim} \text{Laplace}(1)$.
- ▶ Publish $Z_x^{(j)} = f(x) + \frac{2M}{n\varepsilon} W^{(j)}$, $j = 1, \dots, m$.

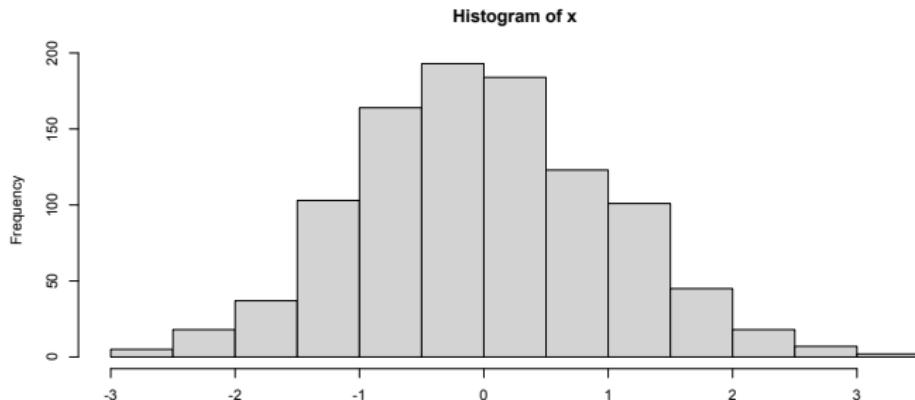
Adversary computes aggregate (recall sequential composition)

$$Z_x = \frac{1}{m} \sum_{j=1}^m Z_x^{(j)} = f(x) + \frac{2M}{n\varepsilon} \frac{1}{m} \sum_{j=1}^m W^{(j)} \xrightarrow[m \rightarrow \infty]{LLN} f(x).$$

Would like to release an ε -DP synthetic multi-purpose database once and for all.

RELEASING A PRIVATE HISTOGRAM

- ▶ Data: $x = (x_1, \dots, x_n)^T \in [L, U]^n$



$$k \in \mathbb{N}, h = (U - L)/k, \quad B_j := L + [(j - 1)h, jh), \quad j \in [k],$$
$$\hat{c}_j := |\{i \in [n] : x_i \in B_j\}|,$$

RELEASING A PRIVATE HISTOGRAM

- ▶ Data: $x = (x_1, \dots, x_n)^T \in [L, U]^n$
- ▶ $k \in \mathbb{N}, h = (U - L)/k, \quad B_j := L + [(j - 1)h, jh), \quad j \in [k],$
- ▶ $\hat{c}_j(x) := |\{i \in [n] : x_i \in B_j\}|$

randomize:

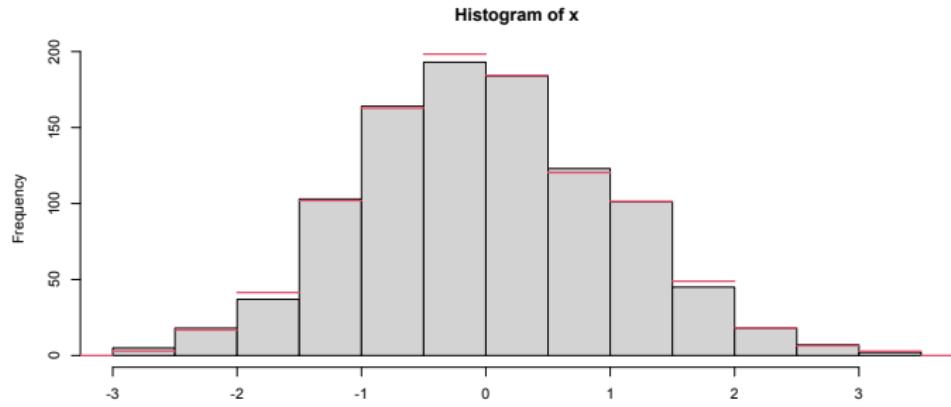
- ▶ $\tilde{c}_j := \hat{c}_j(x) + W_j, \quad W_j \stackrel{iid}{\sim} Laplace(\varepsilon/2).$
- ▶ $Z_x = (\tilde{c}_1, \dots, \tilde{c}_k)^T.$

$$q(z|x) = \prod_{j=1}^k \left[\frac{\varepsilon}{4} \exp \left(-\frac{\varepsilon}{2} |z_j - \hat{c}_j(x)| \right) \right], \quad z_j \in \mathbb{R}.$$

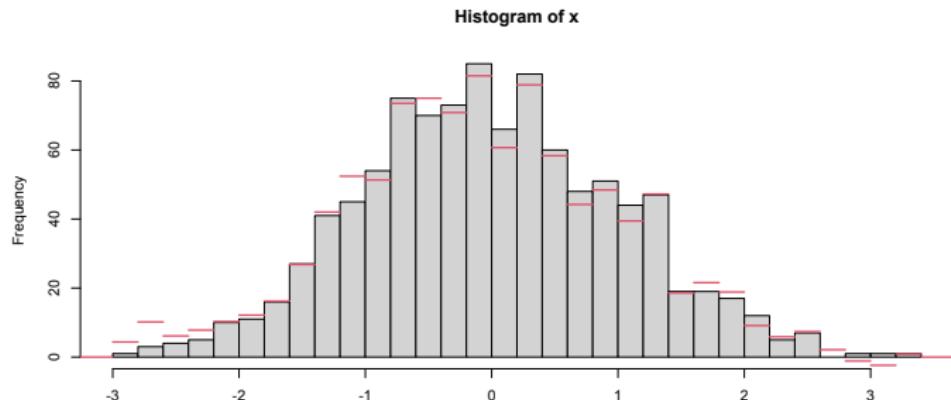
RELEASING A PRIVATE HISTOGRAM

$$q(z|x) = \left(\frac{\varepsilon}{4}\right)^k \exp\left(-\frac{\varepsilon}{2}\|z - \hat{c}(x)\|_1\right), \quad z \in \mathbb{R}^k$$

RELEASING A PRIVATE HISTOGRAM

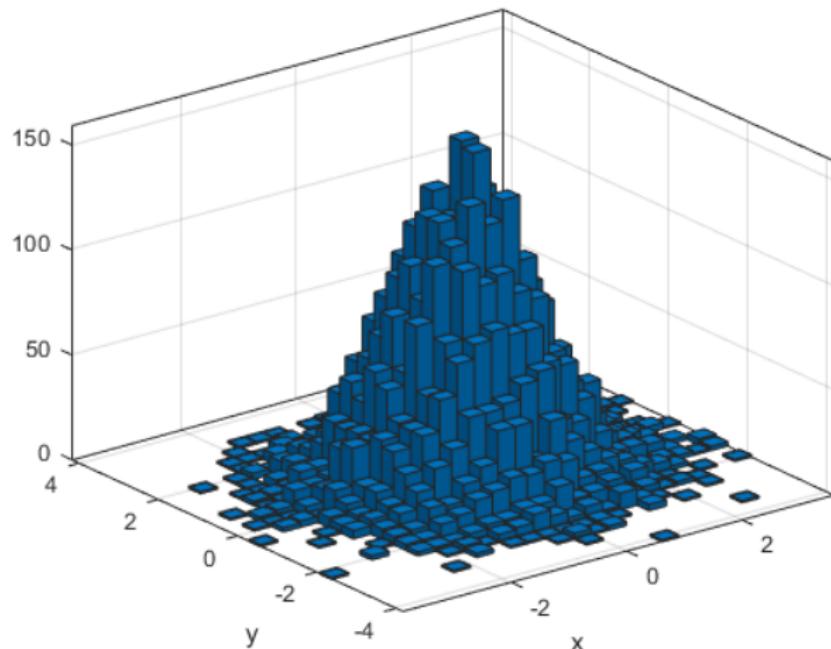


$$\varepsilon = 1$$



RELEASING A PRIVATE HISTOGRAM

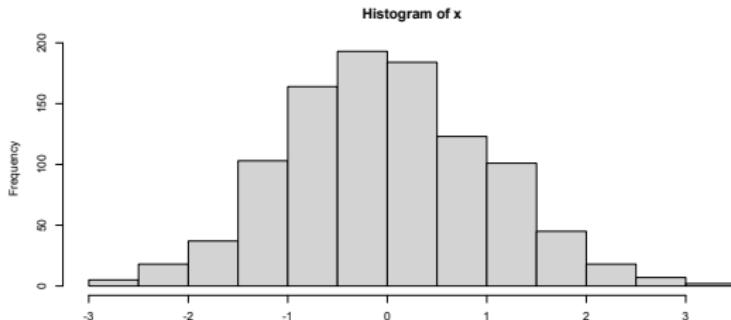
Can do the same for multivariate data $x = (x_1, \dots, x_n)^T$,
 $x_i \in \mathbb{R}^p$.



RELEASING A PRIVATE HISTOGRAM

Suppose an analyst wants to compute \bar{x}_n .

Idea: Treat the histogram as a probability density function.



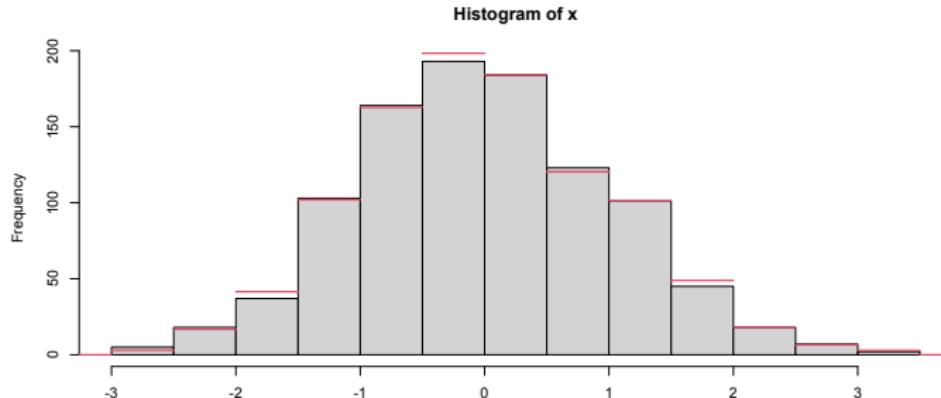
$$\hat{p}_n(y) := \sum_{j=1}^k \frac{\hat{c}_j}{nh} \mathbb{1}_{B_j}(y) \geq 0, \quad \int_{-\infty}^{\infty} \hat{p}_n(y) dy = \sum_{j=1}^k \frac{\hat{c}_j}{nh} h = 1$$

RELEASING A PRIVATE HISTOGRAM

$$\begin{aligned}B_j &= [l_j, l_j + h) = [L + (j - 1)h, L + jh) \\ \bar{x}_n &\approx \mathbb{E}_n[X] := \int_{-\infty}^{\infty} y \cdot \hat{p}_n(y) dy = \sum_{j=1}^k \frac{\hat{c}_j}{nh} \int_{l_j}^{l_j+h} y dy \\ &= \sum_{j=1}^k \frac{\hat{c}_j}{2nh} ((l_j + h)^2 - l_j^2) = \sum_{j=1}^k \frac{\hat{c}_j}{2nh} (2l_j h + h^2) \\ &= \sum_{j=1}^k \frac{\hat{c}_j(l_j + \frac{h}{2})}{n}\end{aligned}$$

Here $\bar{x}_n = -0.0075$, $\mathbb{E}_n[X] = -0.0095$.

RELEASING A PRIVATE HISTOGRAM

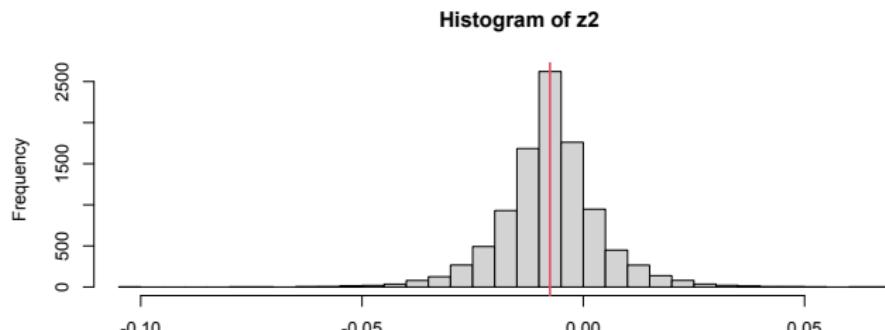
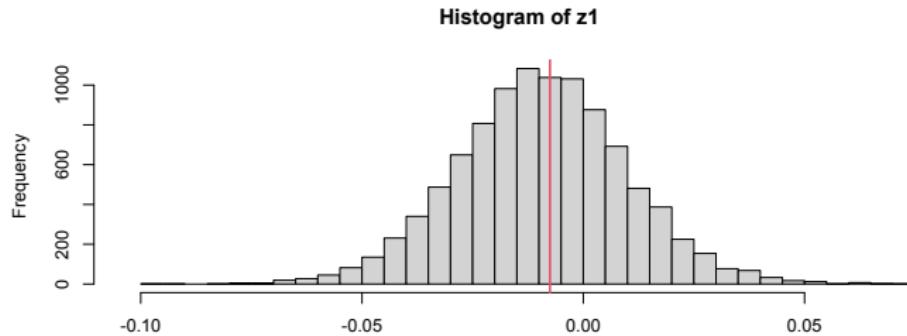


$$\mathbb{E}_n[X] = \sum_{j=1}^k \frac{\hat{c}_j(l_j + \frac{h}{2})}{n}, \quad \mathbb{E}_n[Z_x] := \sum_{j=1}^k \frac{\tilde{c}_j(l_j + \frac{h}{2})}{n}$$

Here $\mathbb{E}_n[Z_x] = -0.0845$ and $\bar{x}_n + \text{Laplace}(\frac{n\varepsilon}{2M}) = -0.03176$.
($M = 4$)

RELEASING A PRIVATE HISTOGRAM: SIMULATION

Compare mean from private histogram (\bar{Z}_1) with perturbed sample mean (\bar{Z}_2). $x \in \mathbb{R}^n$ fixed, $\bar{x}_n = -0.0075$



Approximate differential privacy

DEFINITION OF DIFFERENTIAL PRIVACY

For $x, x' \in \mathcal{X}^n$, define the Hamming distance

$$d_0(x, x') := \#\{i : x_i \neq x'_i\}.$$

Definition (Dwork et al. 2006)

Fix a privacy level $\varepsilon \in (0, \infty)$. The randomization mechanism outputting Z_x on \mathcal{Z} for a given $x \in \mathcal{X}^n$, is called ε -differentially private if for all $x, x' \in \mathcal{X}^n$ with $d_0(x, x') \leq 1$, we have

$$\mathbb{P}(Z_x \in A) \leq e^\varepsilon \mathbb{P}(Z_{x'} \in A), \quad \forall A \subseteq \mathcal{Z} \text{ (measurable)}.$$

We call Z_x an ε -differentially private view of $x \in \mathcal{X}^n$.

DEFINITION OF APPROX. DIFFERENTIAL PRIVACY

For $x, x' \in \mathcal{X}^n$, define the Hamming distance

$$d_0(x, x') := \#\{i : x_i \neq x'_i\}.$$

Definition

Fix $\varepsilon \in (0, \infty)$ and $\delta \in [0, 1]$. The randomization mechanism outputting Z_x on \mathcal{Z} for a given $x \in \mathcal{X}^n$, is called (ε, δ) -approximately differentially private if for all $x, x' \in \mathcal{X}^n$ with $d_0(x, x') \leq 1$, we have

$$\mathbb{P}(Z_x \in A) \leq e^\varepsilon \mathbb{P}(Z_{x'} \in A) + \delta, \quad \forall A \subseteq \mathcal{Z} \text{ (measurable)}.$$

We call Z_x an (ε, δ) -approximately differentially private view of $x \in \mathcal{X}^n$.

FAILURE OF ADP

Consider the following mechanism:

- $\mathcal{Z} = \mathcal{X}^n \cup \{\emptyset\}$, $\delta \in [0, 1]$

$$Z_x = \begin{cases} x, & \text{with probability } \delta, \\ \emptyset, & \text{with probability } 1 - \delta. \end{cases}$$

$$q(z|x) = \mathbb{P}(Z_x = z) = \begin{cases} \delta, & \text{if } z = x, \\ 1 - \delta, & \text{if } z = \emptyset, \\ 0, & \text{else} \end{cases} \leq e^\varepsilon q(z|x') + \delta$$

- This is (ε, δ) -ADP for any $\varepsilon \geq 0!!!$

LOCAL SENSITIVITIES REVISITED

- Recall: *local sensitivity* of f at $x \in \mathcal{X}^n$

$$\Delta_f(x) := \sup_{\substack{x' \in \mathcal{X}^n \\ d_0(x, x') \leq 1}} |f(x) - f(x')|.$$

- Note: For many query functions $f : \mathcal{X}^n \rightarrow \mathbb{R}$

$$Z_x = f(x) + \frac{\Delta_f(x)}{\varepsilon} W, \quad \text{with } W \sim \text{Laplace}(1)$$

is (ε, δ) -ADP, if and only if, $\delta = 1$.

LOCAL SENSITIVITIES REVISITED

$$Z_x = f(x) + \frac{\Delta_f(x)}{\varepsilon} W, \quad \text{with } W \sim \text{Laplace}(1)$$

PROPOSE-TEST-RELEASE

Define

$$\begin{aligned} A_f(x, k) &:= \sup_{y: d_0(x, y) \leq k} \Delta_f(y) \\ D_f(x, b) &:= \min\{k \in \mathbb{N}_0 : A_f(x, k) > b\} \\ \min \emptyset &:= \infty \end{aligned}$$

1. The analyst proposes a value $b > 0$.
2. If $D_f(x, b) + \frac{1}{\varepsilon} \text{Laplace}(1) < \frac{\log(2/\delta)}{2\varepsilon}$, output $Z_x = \emptyset$.
3. Otherwise, output

$$Z_x = f(x) + \frac{b}{\varepsilon} W, \quad \text{with } W \sim \text{Laplace}(1).$$

This satisfies (ε, δ) -ADP.

PROPOSE-TEST-RELEASE

Define

$$A_f(x, k) := \sup_{y: d_0(x, y) \leq k} \Delta_f(y)$$

$$D_f(x, b) := \min\{k \in \mathbb{N}_0 : A_f(x, k) > b\}$$

In step 2 we do the test

$$D_f(x, b) + \frac{1}{\varepsilon} \text{Lap}(1) < \frac{\log(2/\delta)}{2\varepsilon}.$$

Note:

$$b_1 \leq b_2 \quad \Rightarrow \quad D_f(x, b_1) \leq D_f(x, b_2)$$

$$b < \Delta_f(x) \quad \Rightarrow \quad D_f(x, b) = 0$$

$$b \geq \Delta_f \quad \Rightarrow \quad D_f(x, b) = \infty.$$

DIFFERENTIAL PRIVACY: SUMMARY

Pros:

- ▶ DP provides a mathematically rigorous definition of privacy protection.
- ▶ Can develop a theory of optimal privacy mechanisms.
- ▶ It protects against worst case adversaries using any kind of auxiliary information.

Cons:

- ▶ Results are always noisy. Too much noise?
- ▶ Especially difficult for high-dimensional and unbounded data.
- ▶ Many alternative definitions are in use (e.g., ADP, etc.).
- ▶ Optimal data release mechanism depends on the query of interest/the statistical estimation problem. No universally optimal synthetic data release.
- ▶ Many open questions remain...