MSc Applied Information and Data Science

Lucerne University of Applied Sciences and Arts

Data Analytics for Data Scientists

Design of Experiments (DoE)

Lecture 10: Large data quantities

2025

Prof. Dr. Jürg Schwarz

Program: 16:15 until 17:55

16:15	**Begin of the lesson**
	Lecture: Jürg Schwarz ◦ Which study would you choose? ◦ What are large data quantities? ◦ Examples of critical properties of large data quantities Bias & Spurious Correlation ◦ Preview of Lecture 11
	Tutorial: Students / Jürg Schwarz / Assistants ◦ Working on the exercise ◦ Support by Jürg Schwarz / Assistants
17:55	**End of the lesson**

Detailed table of contents → Slide 20

Which study would you choose?

One research question – two studies

Will Donald Trump win the 2016 presidential election?

The population includes 230,000,000 elective voters in the US.

Two studies are being conducted, which differ in terms of data collection:

Study A – includes a data set from a survey with a random sample

Sample size → 400 << 1% of the defined population

Study B – includes an existing data set *(administrative dataset)*

Data set size → 2,300,000 1% of the defined population

Which study would you choose to answer the research question?

Meng (2018)

A first answer – and questions …

It depends …

How can two data sets with different quantities and qualities be compared?

Study A

Which sampling frame is used for sampling?

What controls the response rate?

How to assess the fact that it is a small data set?

Does the topic around the research question influence the response behavior?

Study B

How was the existing data set created?

How to assess the fact that it is an *administrative dataset*?

How to assess the fact that it is a large data set?

Does the topic around the research question influence the (self-)selection?

What are large data quantities?

One of the V dimensions of "Big Data" → ***volume*** (also called*: size, cardinality, ...)*

- Relation of the number of variables vs. sample size
- Relation of size of sample vs. population / size of *administrative dataset* vs. population
- Large data quantities, measured in bytes
- Special storage techniques

Data set from a sample vs. *administrative data set* (German "Verwaltungsdaten")

Samples are taken as part of a study.
Primary goal: To answer research questions

Administrative data are collected for various reasons.
Primary goal: To serve documentary and administrative purposes

Grey area Data from full surveys (census), from social media and from "representative" surveys lie somewhat between data from a sample and administrative data.

This terminology has become established in many fields of research

Made data → Data is generated by researchers ("made").

Found data → Data are obtained administratively and technically ("found")

Mohammadi & Karami (2020) / van Altena et al. (2016) / Mauro et al. (2016)

Research question → "Administration"

Made Data Experimental	**Made Data** Observational (e.g. Social Survey)	**Found Data** Administrative Data	**Found Data** Other Types of Big Data
• Data are collected to investigate a fixed hypothesis.	• Data may be used to address multiple research questions.	• Data are not collected for research purposes.	• Data are not collected for research purposes.
• Usually relatively small in size relatively uncomplex.	• Data may be very large and complex (but usually smaller than big data).	• May be large and complex.	• May be very large and very complex.
• Highly systematic.	• Highly systematic.	• Semi-systematic.	• Some sources will be very unsystematic (e.g. data from social media posts).
• Known sample / population.	Known sample / population.	• Usually a known sample / population.	• Sample / population usually unknown.
		• Multidimensional (i.e. may involve multiple fragments of data which have to be brought together through data linkage).	• Multidimensional (i.e. may involve multiple fragments of data which have too be brought together through data linkage).
		• May be messy (i.e. may involve extensive data management too clean and organize the data).	• Very messy / chaotic.

May be merged with found data,
e.g. to control for confounders

Connelly et al. (2016)

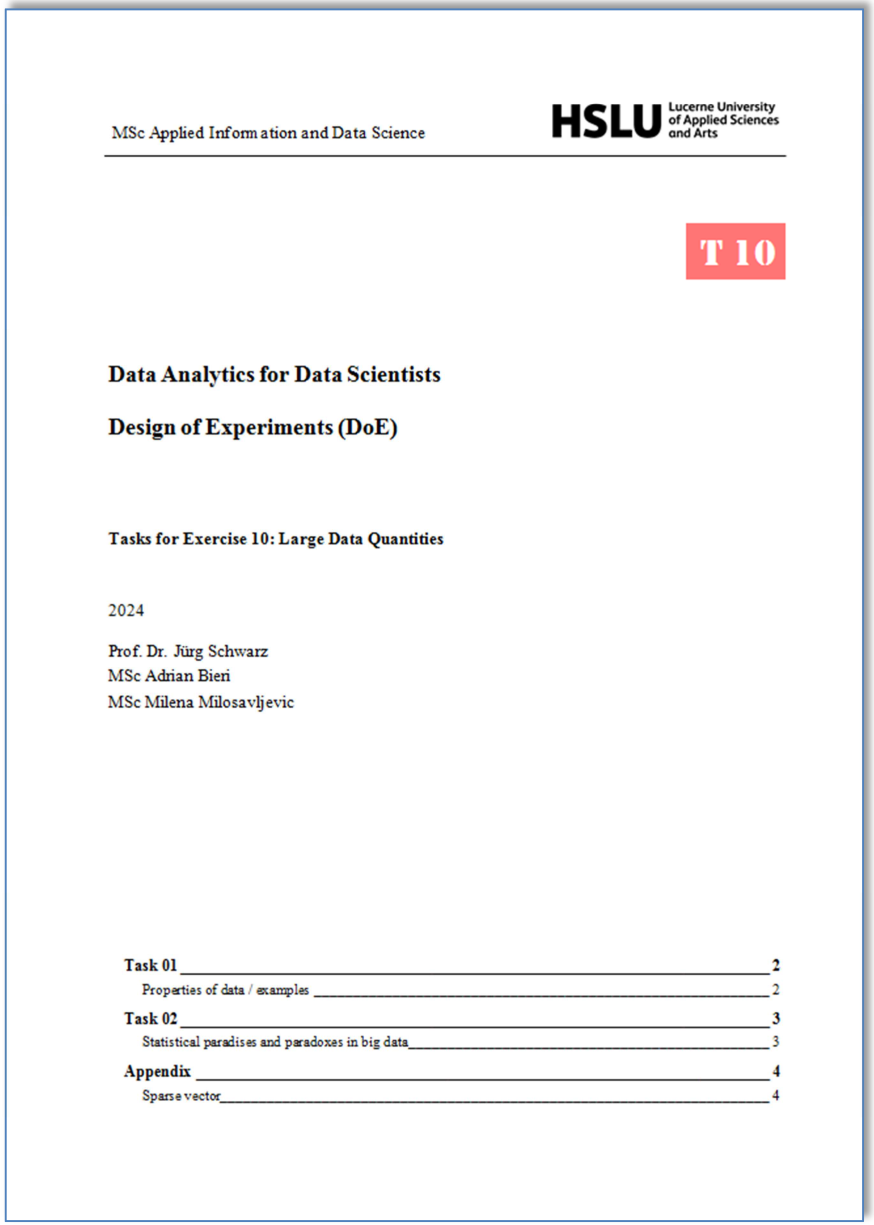
Examples

Made data

…

Found data

…


MSc Applied Information and Data Science

HSLU Lucerne University of Applied Sciences and Arts

T 10

Data Analytics for Data Scientists

Design of Experiments (DoE)

Tasks for Exercise 10: Large Data Quantities

2024

Prof. Dr. Jürg Schwarz
MSc Adrian Bieri
MSc Milena Milosavljevic

Two examples of critical properties of large data quantities

Bias (*Statistical Paradises and Paradoxes*) → Summary starting on Slide 9

Statisticians are increasingly posed with thought-provoking and even paradoxical questions, challenging our qualifications for entering the statistical paradises created by Big Data.

By developing measures for data quality, a framework is suggested to address such a question: Which one should I trust more, a…

- 1% survey with 60% response rate or
- self-reported administrative dataset covering 80% of the population?

Spurious correlation → Summary starting on Slide 15

Big Data are characterized by high dimensionality and large sample size.
These two features raise some unique challenges:

- High dimensionality brings noise accumulation, spurious correlations, and …
- High dimensionality combined with large sample size creates issues such as heavy computational cost and algorithmic instability.

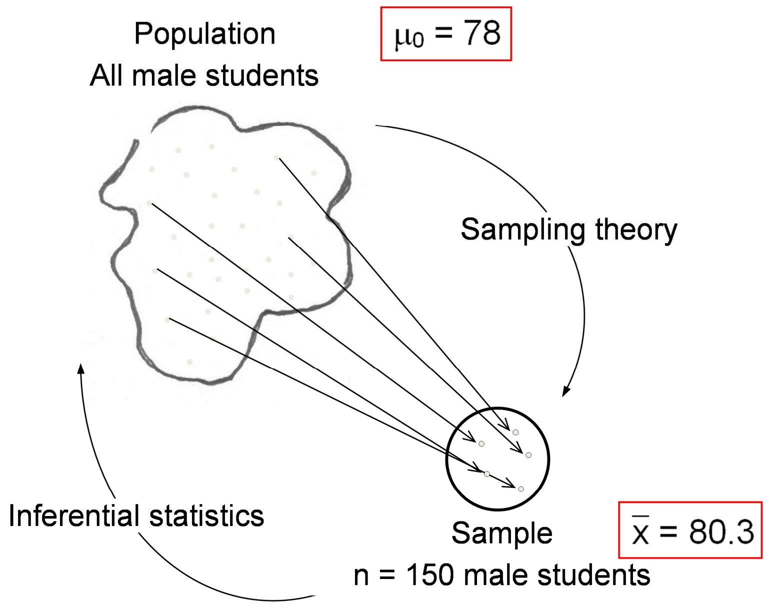
Comment: Spurious correlation here does not mean "storks and babies"!

Meng (2018) / Fan et al. (2017)

Bias (*Statistical Paradises and Paradoxes*)

What is bias?

Deviation between mean μ_0 in the population and sample mean $\bar{x}$



How large is the deviation (bias) between $\bar{x}$ and μ_0?

Three elements determine the bias

1. Data quality measure
2. Data quantity measure
3. Problem difficulty measure

G is a characteristic in the population – e.g., body weight

How can bias be quantified?

Equation to describe the bias between $\bar{x}$ and μ_0

$$\text{Bias} = \bar{x} - \mu_0 = \underbrace{\rho_{R,G}}_{\text{Data Quality}} \times \underbrace{\sqrt{\frac{1-f}{f}}}_{\text{Data Quantity}} \times \underbrace{\sigma_G}_{\text{Problem Difficulty}}$$

$\rho_{R,G}$ *Data defect correlation* → Relationship of characteristic G to the survey method

f *Fraction* → Proportion of the sample or *administrative dataset* to the population

Census $f = 1$ → Data quantity = 0 → $\bar{x} - \mu_0 = 0$ → no bias

No data $f \to 0$ → Data quantity → ∞ → $\bar{x} - \mu_0 \to \infty$ → bias → ∞

σ_G *Variance* → Variation of characteristic G in the population

Example: Let G be constant → $\sigma_G = 0$ → $\bar{x} - \mu_0 = 0$ → no bias

↔ n = 1 is sufficient

G is a characteristic in the population – e.g., body weight

Data defect correlation $\rho_{R,G}$

In $\rho_{R,G}$ the R is a function that shows how data is obtained from the population.

In simple random sampling, the R function generates a randomly generated sequence of elements drawn from the population.
Because of the random process, the selection of an element is independent of G

$\rightarrow \rho_{R,G} = 0$

$$\text{Bias} = \bar{x} - \mu_0 = 0 \times \sqrt{\frac{1-f}{f}} \times \sigma_G = 0$$

In the case of data obtained through administrative technical means, there may be a connection between the (self-)selection of an element and its characteristic G.

$\rightarrow \rho_{R,G} \neq 0$

$$\text{Bias} = \bar{x} - \mu_0 = \rho_{R,G} \times \sqrt{\frac{1-f}{f}} \times \sigma_G \neq 0$$

Meng (2018) mentions that "… the data defect correlation $\rho_{R,G}$ is not a quantity that has been well studied, partly because it is not directly estimable."
In the case of administrative technical obtained data, empirical studies are used.

Meng (2021) mentions a current study of Isakov & Kuriwaki (2020) with new estimations.

Meng (2018) / Meng (2021) / Isakov & Kuriwaki (2020)

Statistical paradises and paradoxes **in relation to administrative data sets**

Measure for the size of the bias → Mean-squared error (MSE)

The MSE measures the deviation (bias) of the estimator $\bar{x}$ from the mean μ_0 in the population.

After a few mathematical steps, the result is as follows:

→ $\mathrm{MSE}[\bar{x}] \sim \frac{1-f}{f}$

→ The bias goes to 0 if f goes to 1

→ $\mathrm{MSE}[\bar{x}] \to 0$ if $f \to 1$, that means if $n \to N$.

$f = \frac{n}{N}$ Proportion of sample size of the administrative dataset in the population

<u>Summary</u>

The bias goes to 0 only, if the size n of the *administrative dataset* goes against N (n → N),

The absolute size n of the *administrative dataset* is meaningless without specifying N.

Although n = 2,300,000 is "big", the proportion remains small with $f = \frac{n}{N} = \frac{2{,}300{,}000}{230{,}000{,}000} = 1\%$

→ *The more the data, the surer we fool ourselves.*

Estimates from the data of Trump's election in 2016

Which study would you choose?

Study A – includes a data set from a survey with a random sample

Sample size → 400 << 1% of the defined population

Study B – includes an existing data set *(administrative dataset)*

Data set size → 2,300,000 1% of the defined population

$\rho_{R,G}$ *Data defect correlation* → relationship of characteristic G to the survey method

If an estimate of $\rho_{R,G} = -0.00005$ is used, based on the data of Trump's presidential election in 2016, calculations can be made to answer the question of which study to choose.

This tells us that …

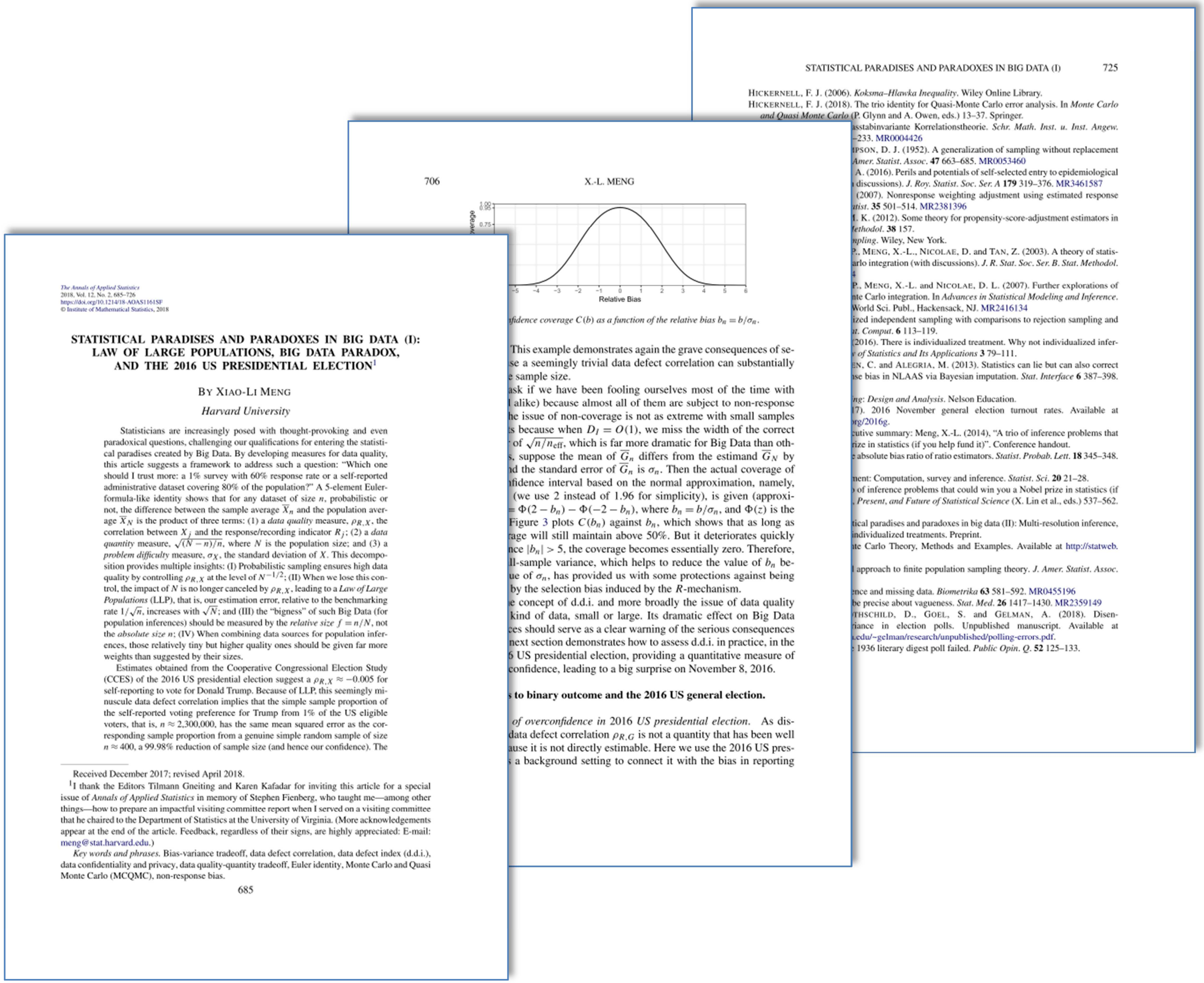
an *administrative dataset* with n = 2,300,000 (f = 1% of US voters)

has the same mean squared error (MSE)

as a simple random sample with n = 400 (f_s << 1% of US voters)

The two studies A and B are equivalent in terms of accuracy as measured by the MSE.

Meng, Xiao-Li (2018): Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. In: The Annals of Applied Statistics 12 (2), S. 685–726.



The Annals of Applied Statistics
2018, Vol. 12, No. 2, 685–726
https://doi.org/10.1214/18-AOAS1161SF
© Institute of Mathematical Statistics, 2018

STATISTICAL PARADISES AND PARADOXES IN BIG DATA (I): LAW OF LARGE POPULATIONS, BIG DATA PARADOX, AND THE 2016 US PRESIDENTIAL ELECTION[1]

BY XIAO-LI MENG

Harvard University

Statisticians are increasingly posed with thought-provoking and even paradoxical questions, challenging our qualifications for entering the statistical paradises created by Big Data. By developing measures for data quality, this article suggests a framework to address such a question: "Which one should I trust more: a 1% survey with 60% response rate or a self-reported administrative dataset covering 80% of the population?" A 5-element Euler-formula-like identity shows that for any dataset of size n, probabilistic or not, the difference between the sample average $\overline{X}_n$ and the population average $\overline{X}_N$ is the product of three terms: (1) a *data quality* measure, $\rho_{R,X}$, the correlation between X_j and the response/recording indicator R_j; (2) a *data quantity* measure, $\sqrt{(N-n)/n}$, where N is the population size; and (3) a *problem difficulty* measure, σ_X, the standard deviation of X. This decomposition provides multiple insights: (I) Probabilistic sampling ensures high data quality by controlling $\rho_{R,X}$ at the level of $N^{-1/2}$; (II) When we lose this control, the impact of N is no longer canceled by $\rho_{R,X}$, leading to a *Law of Large Populations* (LLP), that is, our estimation error, relative to the benchmarking rate $1/\sqrt{n}$, increases with $\sqrt{N}$; and (III) the "bigness" of such Big Data (for population inferences) should be measured by the *relative size* $f = n/N$, not the *absolute size* n; (IV) When combining data sources for population inferences, those relatively tiny but higher quality ones should be given far more weights than suggested by their sizes.

Estimates obtained from the Cooperative Congressional Election Study (CCES) of the 2016 US presidential election suggest a $\rho_{R,X} \approx -0.005$ for self-reporting to vote for Donald Trump. Because of LLP, this seemingly minuscule data defect correlation implies that the simple sample proportion of the self-reported voting preference for Trump from 1% of the US eligible voters, that is, $n \approx 2{,}300{,}000$, has the same mean squared error as the corresponding sample proportion from a genuine simple random sample of size $n \approx 400$, a 99.98% reduction of sample size (and hence our confidence). The

Received December 2017; revised April 2018.

[1]I thank the Editors Tilmann Gneiting and Karen Kafadar for inviting this article for a special issue of *Annals of Applied Statistics* in memory of Stephen Fienberg, who taught me—among other things—how to prepare an impactful visiting committee report when I served on a visiting committee that he chaired to the Department of Statistics at the University of Virginia. (More acknowledgements appear at the end of the article. Feedback, regardless of their signs, are highly appreciated: E-mail: meng@stat.harvard.edu.)

Key words and phrases. Bias-variance tradeoff, data defect correlation, data defect index (d.d.i.), data confidentiality and privacy, data quality-quantity tradeoff, Euler identity, Monte Carlo and Quasi Monte Carlo (MCQMC), non-response bias.

685

Spurious Correlation: A story in five steps

Overview and example

Step 1: Regression equation in vector notation

General equation of a multiple regression model with d variables

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_d x_d + \varepsilon$$

Inclusion of index i for the data points in the data set: i = 1, 2, ... n (n being sample size)

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_d x_{d,i} + \varepsilon_i$$

→ vector / matrix notation

$$\begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} & \ldots & x_{d,1} \\ 1 & x_{1,2} & x_{2,2} & \ldots & x_{d,1} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{1,n} & x_{2,n} & \ldots & x_{d,n} \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \ldots \\ \beta_d \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \ldots \\ \varepsilon_n \end{bmatrix}$$

$$\mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + \varepsilon$$

$\boldsymbol{\beta}$ = vector of coefficients β_i

Step 2: Sparse vector β – Simplified

Typically found in:
Classic study ↔ Made data

Relation: n larger than d

d small

n large

y	x_1	x_2	x_3
47.8	1	1	7.5
84.0	1	1	2.9
1.7	1	1	1.7
35.3	1	1	8.8
93.5	1	1	1.9
2.6	1	1	1.6
54.4	1	1	6.2
8.8	1	1	1.1
81.2	1	1	2.9
42.2	1	1	0.9
27.8	1	1	9.7
61.9	1	1	0.3
11.2	1	1	6.3
...	...	...	...

Typically found in:
"Big Data" ↔ Found data

Relation: d larger than n

d large

n small

y	x_1	x_2	x_3	...	x_p
90.1	1	1	0.2		0.8
13.8	1	1	7.4		0.8
98.0	1	1	9.9		0.5
26.6	1	2	3.3		2.0
69.6	1	2	8.9		0.6
51.0	1	2	8.1		9.7

Example of research in genetics
n = 38 data points *(chips)*
d = 3,051 variables *(genes)*

If d is (much) larger than n, the vector of the coefficients is "sparsely populated" → *sparse**

*More about "Sparse vector" in the appendix of exercise 10

Step 3: Variable Selection in Stepwise Regression

The selection of variables x_j and the estimation of coefficients β_j in the regression model

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_d x_{d,i} + \varepsilon$$

is made from the data using an algorithm:

Step 1: The x_j with the strongest correlation $y \sim x_i$ is included in the equation.

Step 2: The next x_j with the strongest correlation $y \sim x_i + x_j$ is included.

Step 2 is repeated until the addition of further x-variables does not significantly increase the R square any further or until all variables are included.

Step 4: Multicollinearity / Correlations of the independent variables

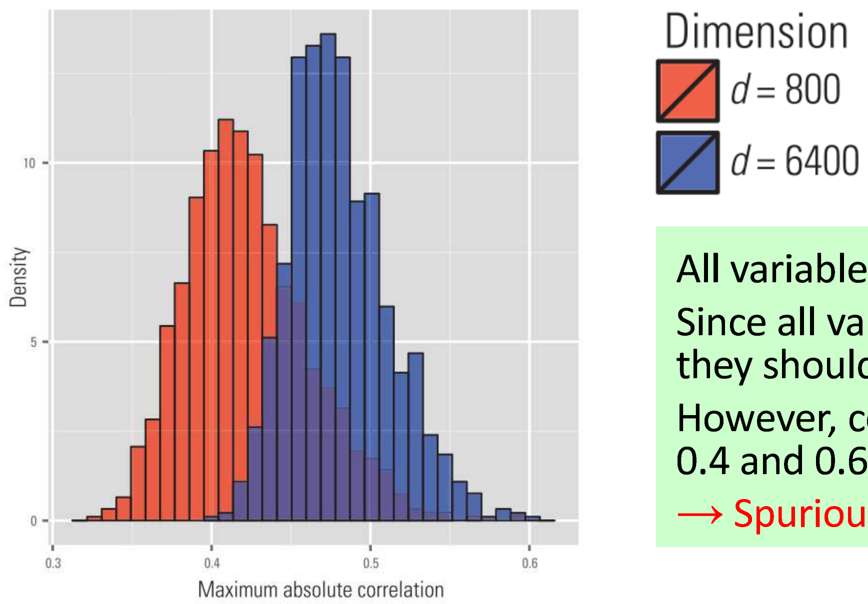
Multicollinearity occurs when the independent variables (x_j) correlate strongly.

Symptoms of multicollinearity

In the case of strong correlation, the standard errors of the coefficients are estimated inaccurately and the tests and confidence intervals therefore become inaccurate.

- The probability increases that a "good" independent variable proves to be not significant.
- Coefficients β_j could occur with the opposite sign than expected.
- In the case of stepwise regression, the values of the estimates for coefficients β_j are inconsistent.

Step 5: Example: Simulation of Fan et al. (2017)



All variables are correlated in pairs.
Since all variables are random variables, they should in principle be uncorrelated.
However, correlations between about 0.4 and 0.6 arise, for example at d = 6,400.
→ Spurious correlation

n = 60 / Two variants of d: d = 800 and d = 6,400

There are few data points n with many variables d ↔ sparse vector β

This creates a large number of strong spurious correlations

→ Stepwise regression is …
- strongly affected by multicollinearity
- less stable

→ Strong bias in parameter estimation

Richman & Roberts (2023)

Assessing Spurious Correlations in Big Search Data

… it also presents vast new risks that scientists or the public will identify meaningless and totally spurious 'relationships' between variables.

This study is the first to quantify the magnitude of the spurious correlation problem for big search data.

Preview of Lecture 11

What has happened so far

The more the data, the surer we fool ourselves is a call for action.

It is important to know which properties large data quantities have.
There are specific differences to "small" data sets.

It’s equally important to know the influence on design of experiments / statistics and to consider the critical properties of large data sets in the design.

What follows in Lecture 11

When designing experiments in social media, it is likely to face many problems, questions and choices in how to proceed:

- Which study design is given?
  Which one should be / can be chosen?
- Population bias
- Samples …
- etc.

Table of contents