Data Literacy
Lecture 02
Collecting Data

– lecture starts at 10:15 sharp, please hold –

Philipp Hennig

25 October 2021

Eberhard Karls Universität Tübingen
Faculty of Science
Department of Computer Science
Chair for the Methods of Machine Learning

Rough course outline

Data Collection How do we *get* data? — Lectures 2 & 3

- Avoiding *bias* in the collection of data — Experimental Design
- Making sure data is as *informative* as possible — Information

Estimating from Data - What does the data (not) tell us? — Lectures 4–9

- Deriving *estimates* of related quantities from observations — Statistical Concepts
- Can we *trust* the estimate? — Variance of Estimators
- Do data/estimates tell us anything noteworthy? — Tests of Significance
- Did we cheat on ourselves? — Properties and Problems of Testing
- Making Predictions from Data — Regression
- Finding Structure in Data — Dimensionality Reduction and Clustering

Fairness and other societal aspects — Lectures 10 & 11

- Does our analysis put certain people at an disadvantage?
- Can/should we do anything about this?

Style and Techniques — Lectures 12-14

- Collecting, documenting and storing data big and small
- Designing good code in the presence of data
- Visualizing and presenting data, analysis and results

Collecting Data

and why it matters

Collecting data is perhaps the most important part of learning from it. It's also the least formal part.

- how your data is collected can (and will!) affect the result of your analysis
- even if you are not active involved in the collection, you should try to know as much as possible about how data was collected

Today: How to design data collection in *observational* studies to reduce selection bias

Definitions

treatment observational, active, interventional, controlled

Consider data consisting of pairs $(x_i, y_i)_{i=1,\dots,n}$ with a presumed functional relationship $y = f(x)$.

Names of Variables

In statistics x may be called ***independent***, ***controlled***, *or explanatory variable, regressor, treatment, etc.* While y is the ***dependent***, ***response***, *or explained variable, regressand, outcome, etc.* The words *cause* and *effect* should only be used if the causal relationship is known. The index i comes from a **sampling distribution** over some population.

Types of Experiments

In an **observational study**, the experimenter has no control over the values of x. In an **active** or **controlled** study, x can be chosen by the experimenter. The word **interventional** study is more restrictive and often refers to experiments designed to establishing causality, where the mechanism f itself is changed. A **meta-study** is a study that works with data collected by other (observational, controlled, etc.) studies.

An example

which variables, what kind of experiments?

Hypothesis: *Vaccine prevents infection in $c\%$ of people.*

- Independent variable? — vaccine treatment
- Dependent variable? — covid19 infection
- an intervention? — selecting vaccination/placebo
- Sampling distribution? — 'entire population'? In USA/ARG/BRA/GER?

Even observational studies involve selection!

Another example

old news

A local medical laboratory is offering "walk-in" tests for antibodies to SARS-CoV-19 (which signify previous infection and recovery from Covid19) to people who personally come to the lab and pay a fee of 25 EUR. In a press conference on 27 May 2020, the company revealed that, in the first 9 days since the service opened, **1774 persons were tested, of which 184 (10.4%) had a positive test result**. On the same day, official numbers of the health authorities in Tübingen (which are based on PCR tests to directly detect the virus), suggest that just **0.6% of the general population in Tübingen** has so far been infected with SARS-CoV19. Based on the discrepancy between these two numbers, the lab concluded that the **actual percentage of infected people in the general population** may be "**up to 17 times higher** than the official numbers" (presumably because $0.6\% \cdot 17 \approx 10.4\%$).

(Source (rephrased): *Schwäbisches Tagblatt, 27 May 2020*)

- Type of study? — observational
- Sampling distribution? — people who take the test!

Reminder from last week

basic concepts of probability

Notation: The probability for two variables X and Y to take the values $x \in \mathbb{X}$ and $y \in \mathbb{Y}$ is the **joint** probability $P(x, y) = P(X = x \wedge Y = y)$

Sum Rule: The *marginal* probability is then $P(x) \equiv P(X = x) = \sum_{y \in \mathbb{Y}} P(x, y)$

Product Rule: The *conditional* probability for $X = x$ given that $Y = y$ is defined for $P(y) > 0$ through

$$P(x \mid y)P(y) = P(x, y)$$

Bayes' Theorem: The *posterior* probability for (the hypothesis) $X = x$ given (the data) $Y = y$ is

$$\underbrace{P(x \mid y)}_{\text{posterior on } X} = \frac{\overbrace{P(y \mid x)}^{\text{likelihood for } x} \cdot \overbrace{P(x)}^{\text{prior for } x}}{\underbrace{\sum_{x \in \mathbb{X}} P(y \mid x)P(x)}_{\text{evidence for } \mathbb{X}}}$$

Independence Two variables X and Y are **independent**, if and only if their joint distributions factorizes into so-called marginal distributions, i.e. $P(X, Y) = P(X)\ P(Y)$. In that case $P(X|Y) = P(X)$.

Another example

old news

A local medical laboratory is offering "walk-in" tests for antibodies to SARS-CoV-19 (which signify previous infection and recovery from Covid19) to people who personally come to the lab and pay a fee of 25 EUR. In a press conference on 27 May 2020, the company revealed that, in the first 9 days since the service opened, **1774 persons were tested, of which 184 (10.4%) had a positive test result**. On the same day, official numbers of the health authorities in Tübingen (which are based on PCR tests to directly detect the virus), suggest that just **0.6% of the general population in Tübingen** has so far been infected with SARS-CoV19. Based on the discreprancy between these two numbers, the lab concluded that the **actual percentage of infected people in the general population** may be "**up to 17 times higher** than the official numbers" (presumably because $0.6\% \cdot 17 \approx 10.4\%$).

(Source (rephrased): *Schwäbisches Tagblatt, 27 May 2020*)

What's the problem?

$$p(\text{infected}, \text{get test}) = p(\text{infected} \mid \text{get test})p(\text{get test}) > p(\text{infected}) \cdot p(\text{get test}).$$

This study sampled from $p(\text{infected} \mid \text{get test})$, not from $p(\text{infected})$ **sampling bias**

To avoid **sampling bias** estimating $p(\text{test quantity} \mid \text{experiment})$, experiments should satisfy

$$p(\text{test quantity} \mid \text{in experiment}) = p(\text{test quantity} \mid \text{in population})$$

The ideal way to achieve this is if the frequency of the quantity of interest in the sampling population is equal to that frequency in the experiment $p(\text{features in experiment}) = p(\text{features in population})$, (i.e. to **sample from the population**).

A little bit of math

iid. samples allow unbiased estimators

More precisely, the ideal *sample* $\{x_i\}_{i=1,\dots,N}$ should be drawn from **independently, and identically** from *the population* p.

- independently: $p(x_i, x_j \mid \text{in experiment}) = p(x_i \mid \text{in experiment}) \cdot p(x_j \mid \text{in experiment}) \; \forall i, j$
- identically: $p(x_i = x \mid \text{in experiment}) = p(x_j = x \mid \text{in experiment}) \; \forall i, j$
- from the population p: $p(x \mid \text{in experiment}) = p(x \mid \text{in population})$

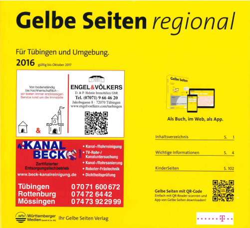
This is important because iid. samples allow the construction of **unbiased estimators**, which are random numbers that are "correct on average" and "converge to the right value with the optimal rate".

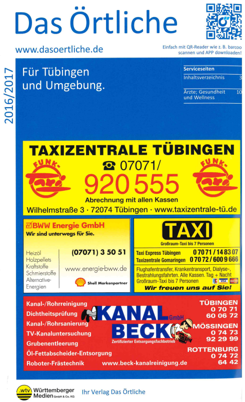
But it can be hard to achieve this…

Live Experiment!

Sample from the Population of Tübingen

- Form groups
- Collect phone books and dice
- Zoom call: Use `dasoertliche.pdf` and `gelbeseiten.pdf`
- **Your task**: Design an algorithm to **draw 10 people uniformly at random** from the phone book
- Timeframe
  - Design the algorithm ($\approx$ 10 mins)
  - Actually run your algorithm ($\approx$ 5 mins)

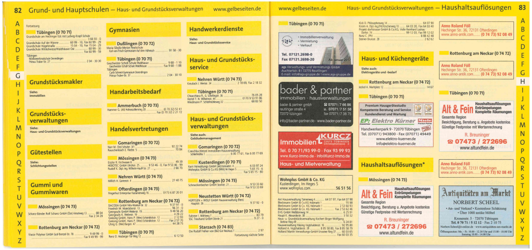


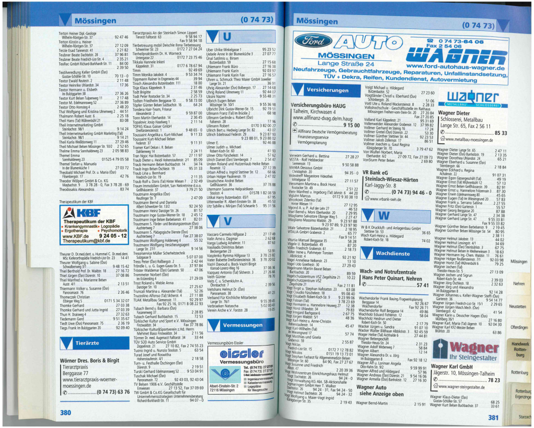


Evaluating the Experiment

What have you learned?

- How did you select randomly among the pages, using the dice?
- How did you pick a position on the page?
- Do you think your method is **unbiased** among the entries of the phone book?
- Can you think of a group of people from *the population* that is **under-represented** in your sample?





Interlude: More Definitions

Expectations

Definition (Expected Values)

*Consider a random variable X taking values $x \in \mathbb{R}^d$ with density $p(x)$ and a real function $f : \mathbb{R}^d \to \mathbb{R}$. The **expected value** (or expectation) of f is given by*

$$\mathbb{E}_p[f(X)] := \int f(x)p(x)\,dx.$$

Definition (Moments)

*The k-th (non-central) **moment** of the distribution with density p is given by the expectation $\mathbb{E}_p[x^k]$ of the function $f(x) = x^k$. The k-th **central moments** of p are given by $\mathbb{E}_p[(x - \mathbb{E}_p[x])^k]$. In particular, they are*

- ***mean*** $\mathbb{E}_p[x]$
- ***variance*** $\mathrm{var}_p(x) = \mathbb{E}_p[(x - \mathbb{E}_p[x])^2] = \mathbb{E}_p[x^2] - \mathbb{E}_p[x]^2$ *for* $d = 1$. ***standard deviation:*** $\sqrt{\mathrm{var}_p(x)}$
- ***co-variance*** $\mathbb{E}_p[xx^\intercal] - \mathbb{E}[x]\mathbb{E}[x]^\intercal$ *for* $d > 1$.

Sampling ("Monte Carlo") Estimators

Definition

An **estimator** is a random variable defined to approximate some property of a (population) distribution.

Let p be the density of that distribubtion, and consider the "property" ϕ defined by the expectation

$$\phi := \int f(x)p(x)\,dx = \mathbb{E}_p\big(f(x)\big).$$

Let $x_i \sim p, i = 1, \dots, n$ iid. The **sampling ("Monte Carlo") estimator** is

$$\hat{\phi} := \frac{1}{n}\sum_{i=1}^{n} f(x_i).$$

Properties of Monte Carlo Estimator: $\mathbb{E}_p(\hat{\phi})$

Unbiasedness

The MC estimator $\hat{\phi}$ is a random number. What are its properties?

$$\mathbb{E}_p(\hat{\phi}) \overset{\text{Def}}{=} \mathbb{E}_p\left(\frac{1}{n}\sum_{i=1}^{n} f(x_i)\right) \overset{(\star)}{=} \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_p(f(x_i)) \overset{x_i \sim p}{=} \frac{1}{n}\sum_{i=1}^{n}\phi = \frac{1}{n} n\phi = \phi$$

$\rightsquigarrow$ **The MC estimator is unbiased.**

$(\star)$: $\mathbb{E}(\,\cdot\,)$ is linear, i.e. for random variables X, Y and constant α

- $\mathbb{E}(\alpha X) = \alpha\mathbb{E}(X)$
- $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$

Bias is a *technical concept*, a property of *estimators*.

Properties of Monte Carlo Estimator: $\mathrm{var}_p(\hat{\phi})$

Next, let's consider the estimator's variance. It holds

$$\mathrm{var}_p(\hat{\phi}) \overset{Def}{=} \mathbb{E}_p\big((\hat{\phi} - \underbrace{\mathbb{E}_p(\hat{\phi})}_{=\phi})^2\big) = \mathbb{E}_p\left(\left(\frac{1}{n}\sum_{i=1}^{n} f(x_i) - \phi\right)^2\right) = \mathbb{E}_p\left(\left(\frac{1}{n}\sum_{i=1}^{n}\left[f(x_i) - \phi\right]\right)^2\right).$$

Expanding the sum yields

$$\begin{aligned}
\ldots &= \mathbb{E}_p\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[f(x_i) - \phi\right]\left[f(x_j) - \phi\right]\right) \\
&\overset{(\star)}{=} \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Big[\mathbb{E}_p\big(f(x_i)\,f(x_j)\big) - \underbrace{\mathbb{E}_p\big(f(x_i)\,\phi\big)}_{=\phi^2} - \underbrace{\mathbb{E}_p\big(\phi\,f(x_j)\big)}_{=\phi^2} + \underbrace{\mathbb{E}_p\big(\phi^2\big)}_{=\phi^2}\Big] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Big[\mathbb{E}_p\big(f(x_i)\,f(x_j)\big) - \phi^2\Big].
\end{aligned}$$

By splitting the sum over j, we obtain

$$\operatorname{var}_p(\hat{\phi}) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[\mathbb{E}_p\big(f(x_i)\, f(x_j)\big) - \phi^2\right]$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} \Bigg[\underbrace{\mathbb{E}_p\big(f(x_i)^2\big) - \phi^2}_{j=i} + \sum_{j \neq i} \left[\mathbb{E}_p\big(f(x_i)\, f(x_j)\big)\big) - \phi^2\right]\Bigg].$$

($\star\star$): For independent random variables X, Y holds

$$\mathbb{E}(X\,Y) = \mathbb{E}(X)\,\mathbb{E}(Y)$$

Using ($\star\star$) yields $\mathbb{E}_p\big(f(x_i)\, f(x_j)\big)\big) = \mathbb{E}_p\big(f(x_i)\big)\, \mathbb{E}_p\big(f(x_j)\big) = \phi^2$. It follows

$$\ldots = \frac{1}{n^2} \sum_{i=1}^{n} \left[\mathbb{E}_p\big(f(x_i)^2\big) - \phi^2\right] = \frac{1}{n^2} \sum_{i=1}^{n} \operatorname{var}_p\big(f(x)\big) = \frac{1}{n} \operatorname{var}_p\big(f(x)\big) = \mathcal{O}(n^{-1})$$

$\rightsquigarrow$ **Sampling converges slowly: The expected error** $\sqrt{\operatorname{var}_p(\hat{\phi})}$ **drops as** $\mathcal{O}(n^{-1/2})$**.**

"Controlling for" Bias

Importance Sampling

If we can not produce samples from p, but only (iid.) from q, we can still produce an unbiased estimate, the *importance sampling* (aka. *(re-)weighted sampling*) estimate. The **core insight** is

$$\phi = \mathbb{E}_p(f(x)) \overset{\text{Def}}{=} \int f(x)p(x)dx = \int f(x)\frac{p(x)}{q(x)}q(x)dx \overset{\text{Def}}{=} \mathbb{E}_q\left(f(x)\frac{p(x)}{q(x)}\right).$$

Given n iid. samples $x_i \sim q$, let

$$\tilde{\phi} := \frac{1}{n}\sum_{i=1}^{n} f(x_i)w(x_i) \quad \text{with} \quad w(x) := \frac{p(x)}{q(x)} \quad \text{(sample weight)}$$

- This estimator is **unbiased**, i.e. $\mathbb{E}_q(\tilde{\phi}) = \phi$
- But, it can be very **imprecise**: $\operatorname{var}_q(\tilde{\phi}) = \frac{1}{n}\operatorname{var}_q\left(f(x)\frac{p(x)}{q(x)}\right)$

You can only correct for the sampling bias you are aware of (we need to know $p(x_i)/q(x_i)$)!

- In **observational** studies in which you want to identify *expected values* of observables in the *population p* from a *small number* of samples, the “ideal” strategy is to sample **independent and identically distributed** from the whole population.
- This is “ideal” in the sense that it produces the **unbiased** sampling estimator with known (up to bias) error and “good” convergence rate.
- But samples from the population p are rarely accessible
  - If samples come from a different sampling population q, can “control for” sampling bias by **importance sampling**
  - But can only correct for biases that are *known* (in the sense of knowing $w(x) = p(x)/q(x)$ in the whole support of p)
- Also, sometimes we simply can not observe the quantity of interest directly. Let’s talk about this in two weeks…

Please provide feedback.

An Item for your Calendar?

Research Seminar on the Methods of Machine Learning

- No lecture next week (public holiday)
- Check out the `Methods of Machine Learning Research Seminar` tag at `https://talks.tue.ai/`
- First talk on **3 November** by **Frederik Künstner**, University of British Columbia

An Optimization View of Probabilistic Models: Convergence of Expectation Maximization in KL divergence

Expectation maximization (EM) is the default algorithm for fitting probabilistic models with missing or latent variables, yet we lack a good understanding of its non-asymptotic convergence properties. Previous works show results along the lines of "EM converges at least as fast as gradient descent" by assuming the conditions for the convergence of gradient descent apply to EM. This approach is loose and does not capture that EM makes more progress than a gradient step, but the assumptions fail to hold for textbook examples of EM like Gaussian mixtures. We derive convergence rates in Kullback-Leibler divergence for the common setting of exponential family distributions by making connections between EM and mirror descent. In contrast to previous works, the analysis is invariant to the choice of parametrization as it directly compares the probability distributions, and holds with minimal assumptions.