universität wien

Statistics for Data Science, Winter 2023

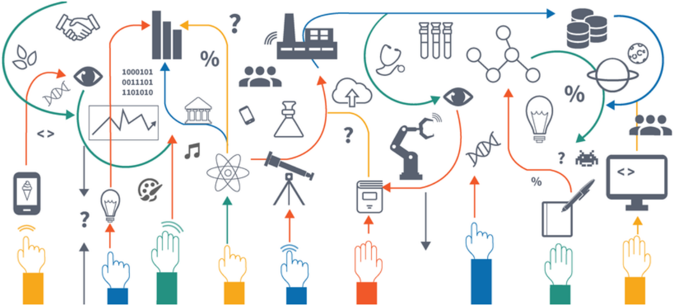
1. Introduction:

Data and Models

Overview

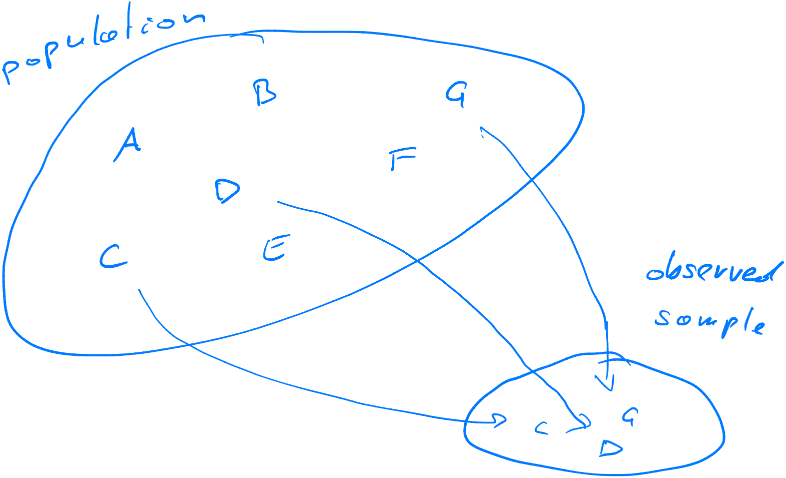
- Introduction (statistical perspective on data)
- Recap: Probability Theory
- Formalism of statistical modeling
- Estimators, tests and confidence intervals

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!
- $\Rightarrow$ ‘assumptions’, statistical model, data generating process

THE STATISTICAL PERSPECTIVE ON DATA: SAMPLING FROM A POPULATION

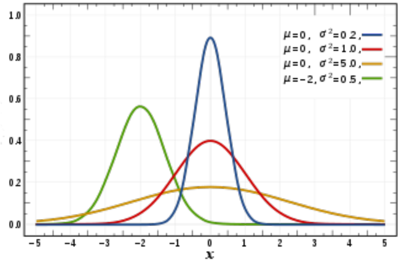


STATISTICAL MODEL VS. ML MODEL

Statistical Model

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

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

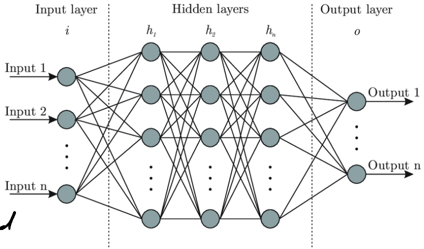


ML Model

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

ML... machine learning

ML ≠ 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:

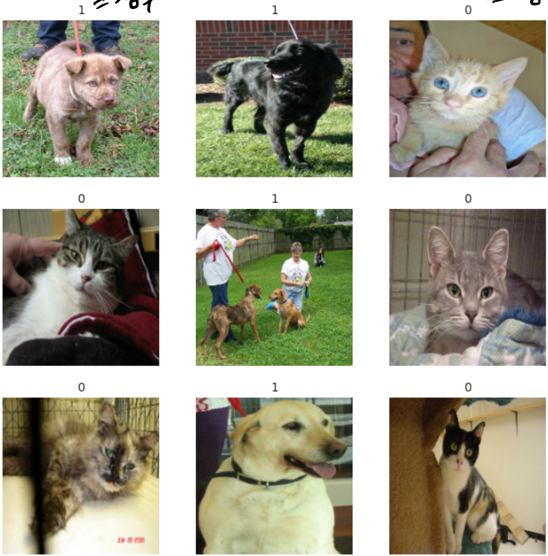
id	age	sex	party
1	37	m	A
2	59	f	B
$\vdots$			$\vdots$
500	25	m	B

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

Example 1.2: Image Classification

Data:

$= y_1$ $= y_3$



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

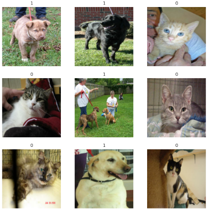
$x_i \in \mathbb{R}^{3p}$, RGB-values

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

p ... # pixels

EXAMPLE 1.2: IMAGE CLASSIFICATION

Data:



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

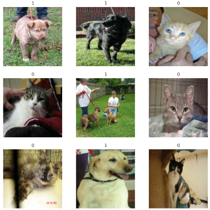
$x_i \in \mathbb{R}^{3p}$, RGB-values

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

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

Example 1.2: Image Classification

Data:



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

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

$y_i \in \{0, 1\}$, 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 $i = 1, \ldots, n$.

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:

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

description	inference
summarize and visualize data	learn about population
describe	generalize/estimate
only data, no models	statistical modeling
no assumptions	idealizations/assumptions
	data generating process?
all data sets are different/unique	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: $p = \dfrac{318}{500} = 0.636$

Example 1.1: electoral survey

Data:

id	party	X_i
1	A	1
2	B	0
$\vdots$	$\vdots$	$\vdots$
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
$\vdots$	$\vdots$	$\vdots$
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
$\vdots$	$\vdots$	$\vdots$
500	B	0

Model:

- $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$
- 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:

- $X_1, \ldots, X_n \overset{iid}{\sim}$ Bernoulli(θ)
- 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:

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

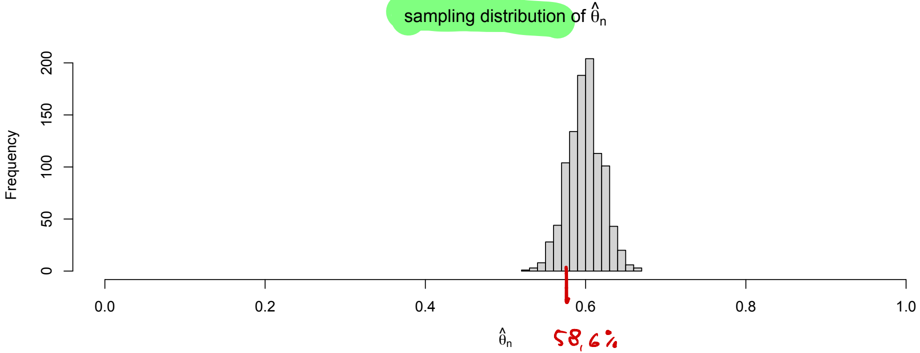
$$\mathbb{E}_\theta[\hat{\theta}_n] = \frac{1}{n} \sum_{i=1}^{n} \underbrace{\mathbb{E}_\theta(X_i)}_{\substack{1 \cdot \theta + 0 \cdot (1-p) \\ = \theta}} = \frac{1}{n} \sum_{i=1}^{n} \theta = \theta$$

"unbiased estimator"

Example 1.1: electoral survey

Model:

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



EXAMPLE 1.1: ELECTORAL SURVEY

Model:

- $X_1, \dots, X_n \overset{iid}{\sim}$ Bernoulli(θ)
- 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)

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

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

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

$P(0{,}544 \leq \theta \leq 0{,}678)$

"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)

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

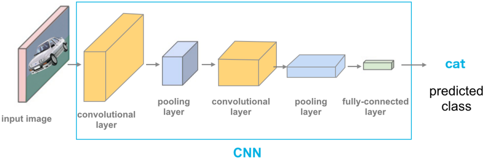
estimation/classification:

Estimate g by a CNN with SGD

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

classification:

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



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

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

validation/error quantification:

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

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

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

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

"quantifies uncertainty of classification"