# Mathematical Statistics

Damien Garreau

Université Côte d'Azur

2022–2023

# 1. General introduction

# Who am I?

- maître de conférence (= assistant professor) in LJAD (Laboratoire Jean Dieudonné)
- before that: postdoctoral researcher (Max Planck Institute, Tübingen, Germany)
- even before: PhD in Inria Paris
- teaching ($\approx$ 200 hours per year)
- **Rest of the time?** research!
- **Goal:** think about open problems whose solution could benefit society, solve them, publish papers with the answer
- examples of topics that interest me at the moment:
  - theoretical foundations of interpretability
  - robustness of text embeddings

# Overview

- **Goal of this course:** introduction to mathematical statistics
  - *estimation:* find parameters of a law given i.i.d. sample
  - **Example:** which percentage of people want to vote for candidate X?
  - *test:* decide between hypotheses on the data in a rational way
  - **Example:** is this vaccine effective in preventing disease X?
- **Organization of the course:** lecture (with me) on Monday, then practical sessions on Wednesday and Thursday (with Luc Lehéricy)
- occasional additional help with Marco Corneli
- **Evaluation:** midterm + final exam
- If you have any question: shoot me an email

  `damien.garreau@unice.fr`
- with a proper subject + question

# Resources

- slides will be on Moodle (`lms.univ-cotedazur.fr/`)
- official name: **SMUMA201**
- other resources (old exams, papers, etc.) $\Rightarrow$ check it regularly!
- wikipedia
- google scholar
- books that I recommend:
  - Rivoirard, Stoltz, *Statistique mathématique en action*, 2009 (second edition 2012)
  - Wasserman, *All of statistics: a concise course in inference*, Springer, 2004 (second edition 2013)

# Organization of the course

- **Disclaimer:** this could change
  1. January 16, 10am-12am (today)
  2. January 23, 10am-12am
  3. January 30, 10am-12am
  4. February 6, 10am-12am
  5. February 13, 10am-12am
  6. February 27, 10am-12am = midterm
  7. March 6, 10am-12am
  8. March 13, 10am-12am
  9. March 20, 10am-12am
  10. March 27, 10am-12am
  11. April 3, 10am-12am
  12. April 10, 10am-12am
- week of February 20 = break
- exam in May (last year May 5)

# Plan of the course

1. statistical modeling
2. estimation (moment method, maximum likelihood)
3. hypothesis testing
4. linear model
5. chi squared tests
6. Kolmogorov-Smirnov
7. density estimation
8. an introduction to machine learning
9. Bayesian statistics

# Sujets de mémoire

- UE de 6 ECTS avec mémoire coeff 4
- rapport de 25 pages + soutenance orale de 30 mn fin juin
- **IM:**
  - travail de recherche en ingénieurie mathématique
  - obligatoirement une partie appliquée
  - choix fin mars, début du travail fin avril
- **MF/MPA:**
  - pas de partie appliquée
  - début après les exams de janvier

# Sujet proposé (avec A. Galligo)

- taille des coefficients de changement de base pour les polynômes symétriques
- polynôme symétrique:

$$P(x_{\sigma(1)}, \ldots, x_{\sigma(n)}) = P(x_1, \ldots, x_n)\,.$$

- plusieurs bases disponibles (monômes symétriques élémentaires,...)
- changements de base connus
- **Exemple:**

$$\sum_{i<j} X_i X_j = \frac{1}{2}\left(\sum_i X_i\right)^2 - \frac{1}{2}\sum_i X_i^2\,.$$

- **Question:** taille de ces coefficients?

# 2. A bit of history

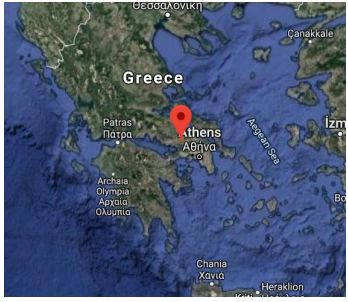
# Etymology

- *Statistics* come from the German *Statistisk* (Gottfried Achenwall, 1749)
- Italian root *statista* (“statesman, politician”)
- New Latin *statisticum* (“of the state”)
- “statistique d’État:” census procedure. How many people? Taxes? Merchandises?

- Mostly *descriptive* statistics, what you did until now
- We focus on *mathematical statistics* (“statistique inférentielle’)

# 2.1. Estimation

# Estimation problem



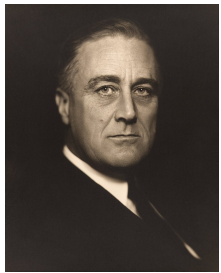

- Thucyclides (c. 460—c. 400 BC), *History of the Peloponnesian War*, battle of Platea (479 BC)
- How to estimate the height of the wall? **First (recorded) statistical methodology**
- Several soldiers count the number of bricks, take the mode, and then multiply by the height of one brick
- Build the ladders accordingly, and win the battle!

# Estimation problem

- United Kingdom, once a year
- **Goal:** make sure that newly minted coin are *conform*
- Same procedure since 1282 AD: take 1 coin every 5000 and have it checked publicly by an expert
- Take 1 every 150 for bimetallic coins
- **Important idea:** too costly to get all the data. Take a small part and *extrapolate*

# Estimation problem

- US, 1936, Franklin Delano Roosevelt *vs* Alfred Landon



- the *Literary Digest* correctly predicted the winner since 1916
- largest poll at the time: **2.4 million** individuals!
- predicts Landon win, Roosevelt loses with 43% of the popular wote
- **Result:** Roosevelt with 62% (19% error!)
- **Why?** *selection bias*: only people in the phone book

# The German tank problem (I)

- Second World War (c. 1943)
- **Problem:** how to assess the production of military equipment for a given month?
- Plan A: James Bond. Plan B: be smart.

- look only at the id numbers of captured tanks

# The German tank problem (II)

- **Goal:** guess $N$, number of tanks produced a given month
- Simplifying a lot, serial numbers $s_1, \ldots, s_n$ are uniformly distributed on $\{1, \ldots, N\}$
- We will see later how to build an *estimator*, the formula is

$$\hat{N} = \max(s_1, \ldots, s_n) \cdot \left(1 + \frac{1}{n}\right) - 1\,.$$

- After the war, who was right? Off by 10% *vs* **double** for Bond.

| Year | Estimated Production | Speer Ministry Statistics | Percentage Error |
|---|---|---|---|
| 1941 | 7,850 | 8,436 | 7% − |
| 1942 | 9,500 | 10,150 | 6% − |
| 1943 | 17,000 | 16,971 | 2% + |
| 1944 | 14,500 | 17,134 | 15% − |
| Total | 48,900 | 52,691 | 7% − |

# 2.2. Testing

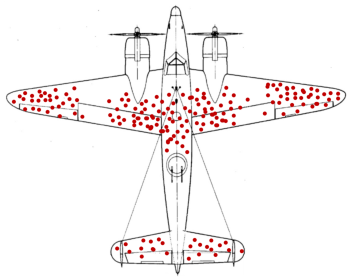
# Testing

- **Example:** pellagra, disease in the poorest regions of Europe and the Americas. unknown vector until 1926
- Blood-sucking fly (*simulium*) has the same geographic range
- People thought it was the vector of the disease

- **But** a bit mysterious: even in the same village, seems to strike only certain houses
- In fact, vitamin $B_3$ (niacin) was the cause
- Now some countries require its addition to grains
- Here a common variable, **poverty**, was the cause for the diet (hence the disease) and the flies

# Bias

- **Question:** where to put armor on aircrafts?
- During WWII, study by the Center for Naval Analyses
- Put armor where the planes are shot the most (red dots)



- **Problem:** only planes which survived
- Abraham Wald (1902-1950): put the armor exactly where there are **no** red dots: these are the critical parts!

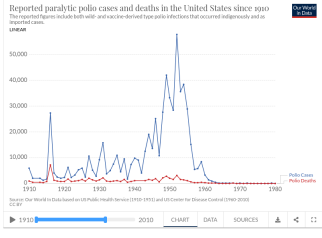
# The Salk vaccine trials (I)

- polio = viral infectious disease
- $R_0 = 5 - 7$, mortality rate $= 5 - 10\%$
- First polio epidemic in the US: 1916
- By the 50s, several vaccines were developed, in particular one by Jonas Salk (1914-1995)

- **Problem:** how can we test the *effectiveness* of the vaccine?
- Naive answer: try on everyone. **This is just not possible, and dangerous!** What if there are some side effects?

# The Salk vaccine trials (II)

- More reasonable: *comparison*
- Give the vaccine to a **treatment group**, compare the response with respect to the **control group**
- the groups should be **random**, control group receives a **placebo**, and the study **double-blind**
- then **hypothesis test** ($H_0$ : vaccine is not effective)
- one can compute $p$-value




# Conclusion: mathematical statistics

Take-away:

- *Descriptive statistics* are about studying the *population*
- *Mathematical statistics* are about making predictions on the population with only a sample
- In simple cases, common sense is enough
- But as soon as we tackle more complex problems, we need maths, in particular **probability theory**

Karl Pearson (1857-1936)

# 3. Probability: a reminder

# Sample space and events

Before speaking of the probability of an event, we need to define what an event is:

- **Sample space:** $\Omega$, the set of all outcomes of a random experiments
- $\omega \in \Omega$ can be thought as a complete description of the situation at the end of the experiment
- **Example:** suppose that the experiment consists in throwing a dice with six faces. What is $\Omega$?
- **Event space:** $\mathcal{F}$ is a set of *subsets* of $\Omega$
- $A \in \mathcal{F}$ can be thought as a *collection of possible outcomes of an experiment*
- **Same example:** what would you propose for $\mathcal{F}$?

*Disclaimer:* $\mathcal{F}$ should satisfy some properties for everything to work, see your courses on probability theory

# Axioms of probability

We call **probability measure** a function $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ that satisfies the following properties:

- $\mathbb{P}(A) \geq 0$ for any $A \in \mathcal{F}$ (a probability is a positive number)
- $\mathbb{P}(\Omega) = 1$ (the probability of something happening is 1)
- if $A_1, A_2, \ldots$ are disjoint events ($A_i \cap A_j = \emptyset$), then

$$\mathbb{P}(A_1 \cup A_2 \cup \cdots) = \sum_{i \geq 1} \mathbb{P}(A_i) \, .$$

(probability of disjoint events is the sum of the probabilities)

In the previous example, if the dice is *fair*, we could use

$$\mathbb{P}(\{1\}) = \cdots = \mathbb{P}(\{6\}) = \frac{1}{6} \, .$$

In our model, the probability of getting a 1 or a 2 would be

$$\mathbb{P}(\{1,2\}) = \mathbb{P}(\{1\} \cup \{2\}) = \mathbb{P}(\{1\}) + \mathbb{P}(\{2\}) = \frac{1}{3} \, .$$

# Properties of probability measures

In statistics, we will rarely be concerned with $\Omega$ and $\mathcal{F}$. But it is important to be familiar with the following properties of $\mathbb{P}$:

- **Monotony:** if $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$
- **Union bound:** for any $A, B \in \mathcal{F}$,

$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B) \ .$$

- **Law of total probability:** let $B_1, B_2, \ldots$ be a partition of $\Omega$ ($B_i \cap B_j = \emptyset$ and $\bigcup_i B_i = \Omega$). Then, for any $A \in \mathcal{F}$,

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A \cap B_i) \ .$$

- **Complement:** for any $A \in \mathcal{F}$,

$$\mathbb{P}(\Omega \setminus A) = 1 - \mathbb{P}(A) \ .$$

# Conditional probability and independence

We will sometime need the notion of *conditional probability*: what is the probability of a certain outcome observing another outcome?

- We denote by $\mathbb{P}(A \mid B)$ the conditional probability of an event $A$ given B
- For any event $B$ with non-zero probability,

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} .$$

- **Question:** what happens if $\mathbb{P}(B) = 0$?
- **Independence:** two random events $A$ and $B$ are called *independent* if

$$\mathbb{P}(A \mid B) = \mathbb{P}(A) .$$

Alternatively, $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$.

# Random variables

Usually, we are not interested in the exact outcome of the experiment $\omega$, but rather by specific numerical properties of the outcome:

- A **random variable** is a mapping $X : \Omega \to \mathbb{R}$
- **Example:** consider an experiment were we flip 10 coins. Then outcomes are binary sequences of length 10:

$$\omega = (H, H, T, H, T, T, T, H, T, H)\,.$$

We could be interested, for instance, in the numbers of heads:

$$X(\omega) = \sum_{i=1}^{10} \mathbb{1}_{\omega_i = H}\,.$$

*Disclaimer:* again, not all functions are ok to work with, see your probability theory course.

# Probabilities and random variables

We can now define the probability that a random variable takes certain values:

- **Discrete case:** let $X$ be a discrete random variable, then we set

$$\mathbb{P}(X = k) = \mathbb{P}(\{\omega \in \Omega \quad \text{such that} \quad X(\omega) = k\}) .$$

- **Example:** assuming that the coin is fair, can you compute $\mathbb{P}(X = 1)$ in the ten coins toss example? What about $\mathbb{P}(X = 2)$?

- **Continuous case:** let $X$ be a continuous variable, then we set

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(\{\omega \in \Omega \quad \text{such that} \quad a \leq X(\omega) \leq b\}) .$$

- **(independence)** if $X$ and $Y$ are independent random variables, then

$$\mathbb{P}(X \leq s \quad \text{and} \quad Y \leq t) = \mathbb{P}(X \leq s)\mathbb{P}(Y \leq t) .$$

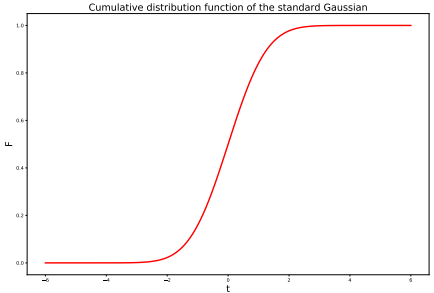
# Cumulative distribution functions

Very often, we prefer to work with the **cumulative distribution function** of a random variable $X$:

- $F_X : \mathbb{R} \to [0, 1]$ is defined by

$$\forall t \in \mathbb{R}, \quad F_X(t) = \mathbb{P}(X \leq t) \ .$$

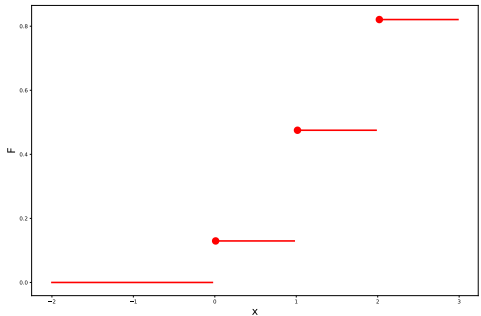
- Knowing $F$ is **equivalent** to knowing the probability
- It usually looks like this:



# Properties of the cumulative distribution function

The following properties come directly from the properties of probability measures:

- **Non-decreasing:** for any $s \leq t$, $F_X(s) \leq F_X(t)$
- **Right continuous:**




- **Takes values in** $[0, 1]$**:** $0 \leq F_X(x) \leq 1$
- $\lim_{t \to -\infty} F_X(t) = 0$
- $\lim_{t \to +\infty} F_X(t) = 1$

# Probability density functions

When the cumulative distribution function is differentiable, we can define

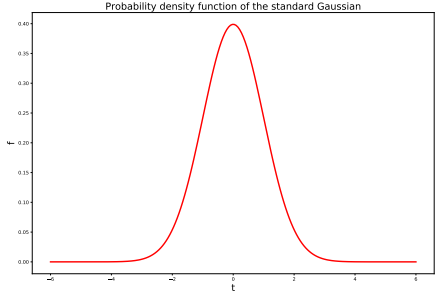
- **Probability density function:** for any $t \in \mathbb{R}$,

$$f_X(t) = \frac{\mathrm{d}F_X(t)}{\mathrm{d}t}\,.$$

- **Intuition:** for a small $\varepsilon$,

$$\mathbb{P}\left(t \leq X \leq t + \Delta t\right) \approx f_X(t)\Delta t\,.$$

- Usually, they look like this:



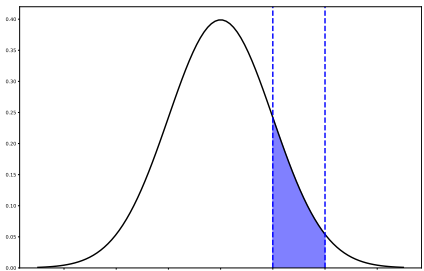
# Properties of probability density functions

We will deal with continuous variables and associated probability functions very often.

- for any $t \in \mathbb{R}$, $f_X(t) \geq 0$
- $\int_{-\infty}^{+\infty} f_X(t)\mathrm{d}t = 1$
- for any event $A$,

$$\mathbb{P}(A) = \mathbb{P}(X \in A) = \int_{t \in A} f_X(t)\mathrm{d}t\,.$$

- **Example:** probability that $1 \leq X \leq 2$ = blue area



# Expectation of a random variable

Informally, the expectation of a random variable $X$ is the **average value** ($=$ the mean) of $X$:

- **Discrete random variable:**

$$\mathbb{E}[X] = \sum_k k \cdot \mathbb{P}(X = k) \, .$$

- **Example:** what is the expected value of a dice roll?
- **Continuous random variable:** if $X$ has a probability density function $f_X$, then

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} t \cdot f_X(t) \mathrm{d}t \, .$$

- If there is a risk of confusion, we specify under which random variable the expectation is taken by writing $\mathbb{E}_X[\cdot]$

# Properties of the expectation

We will use extensively the following properties:

- for any constant $c \in \mathbb{R}$, $\mathbb{E}[c] = c$
- **(homogeneity)** for any $\lambda \in \mathbb{R}$, $\mathbb{E}[\lambda X] = \lambda \mathbb{E}[X]$
- **(additivity)** for any random variables $X$ and $Y$,

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y] .$$

  (we say that the expectation is linear)

- let $A$ be an event, then

$$\mathbb{E}[\mathbb{1}_{X \in A}] = \mathbb{P}(X \in A) .$$

- **(independence)** for any independent random variables $X$ and $Y$,

$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y] .$$

# Variance of a random variable

The **variance** measures how dispersed around its mean a random variable is.

- formally,
$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] .$$
- $\mathrm{Var}(X) \geq 0$, and $\mathrm{Var}(X) = 0$ only if $X = \text{cst}$ a.s.
- $\mathrm{Var}(\lambda X) = \lambda^2 \mathrm{Var}(X)$
- **König-Huygens:** alternative expression given by
$$\mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 .$$

**Exercise:** let $X$ be a random variable such that $f_X(t) = 1$ if $t \in [0, 1]$ and 0 otherwise. Compute the mean and variance of $X$.

# 4. Statistical modeling

# Statistical model

The first step of statistical methodology is to **model the problem at hand**.

**Definition:** A *statistical model* is a pair $(\mathcal{S}, \mathcal{P})$, where $\mathcal{S} = (\Omega, \mathcal{F})$ is a sample space and $\mathcal{P} = \{P \in \mathcal{P}\}$ a family of probability measures.

- **Intuition:** $\mathcal{S}$ is the space of all possible outcomes of the experiment we are modeling, $\mathcal{P}$ is a family of probability distributions that are good models for this experiment
- We say that our model is **parametric** if

$$\mathcal{P} = \{P_\theta, \theta \in \Theta\},$$

  with $\Theta \subseteq \mathbb{R}^D$
- **Inference:** use *observations* to find information on $\theta$

# Observations and sample

**Definition:** we say that a random variable $X$ is an *observation* in the model $(\mathcal{S}, \mathcal{P})$ if it takes values in $\mathcal{S}$ and the law of $X$ belongs to $\mathcal{P}$.

- In the parametric setting, there is a $\theta \in \Theta$ such that, for any event $A$,

$$\mathbb{P}(X \in A) = P_\theta(X \in A)\,.$$

- Often, the observation consists in $n$ i.i.d. random variables $X_1, \ldots, X_n$: we call it a **sample**

- Important: do not confuse the sample, a random variable $(X_1, \ldots, X_n)$, and its **realization**, real numbers $(x_1, \ldots, x_n)$

# Statistical model, example I

- *Literary Digest*: $n$ Americans are interrogated, $X_i$ is the answer of American $i$
- $X_i = 0$ means vote for Landon, $X_i = 1$ means vote for Roosevelt (for example)
- **Model:** $X_i$ are i.i.d. Bernoulli of parameter $\theta \in [0, 1]$, where $\theta$ is the **true proportion** of Roosevelt voters in the population

**Reminder:** $X$ is a Bernoulli of parameter $\theta$ if $X \in \{0, 1\}$ a.s. and

$$\mathbb{P}(X = 1) = \theta = 1 - \mathbb{P}(X = 0) \ .$$

- **Question:** what are the limits of this model?

# Statistical model, example II

- *German tank problem:* $n$ serial numbers are obtained, we model it by a uniform distribution on $\{1, \dots, N\}$

**Reminder:** $X$ follows the uniform distribution on a finite discrete set $\{s_1, \dots, s_k\}$ if $X \in \{s_1, \dots, s_k\}$ a.s. and

$$\mathbb{P}(s_1) = \cdots = \mathbb{P}(s_k) = \frac{1}{k}.$$

- The parameter of the model is $N$, the total number of tanks
- **Question:** what are the limits of this model?

# Statistical model, example III

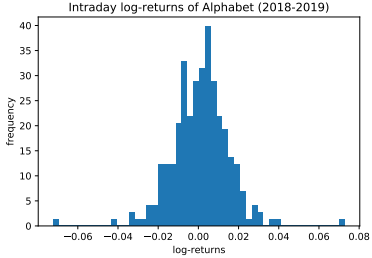
- Consider the value of a stock $V_t$ over time
- The *return* is defined as

$$R_t = \frac{V_{t+1} - V_t}{V_t}.$$

- Not well-behaved, *log*-returns are preferred:

$$L_t = \log\left(\frac{V_{t+1} - V_t}{V_t}\right).$$

- **Example:**



# Statistical model, example III

- Since Bachelier (1900), we model the log-returns by a **Gaussian** distribution

**Reminder:** $X$ follows the Gaussian distribution of parameter $(\mu, \sigma^2)$ if it has density $f_{(\mu,\sigma^2)}$ with respect to the Lebesgues measure on $\mathbb{R}$, where

$$f_{(\mu,\sigma^2)}(u) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \mathrm{e}^{\frac{-(u-\mu)^2}{2\sigma^2}} .$$

In that case we have $\mathbb{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$.

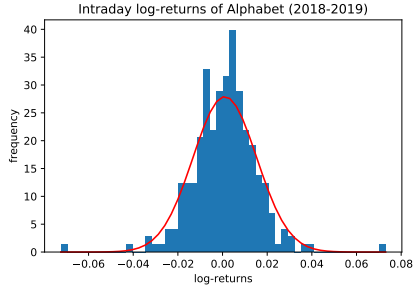
- Our model for each observation is $P_\theta$, with $\theta = (\mu, \sigma^2)$ and $P_\theta$ the Gaussian distribution
- $\Theta = \mathbb{R} \times \mathbb{R}_+$ (no negative standard deviation)

# Statistical model, example III

- **How good is our model?**
- One can compute the sample mean and variance:

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_n)^2 \,.$$

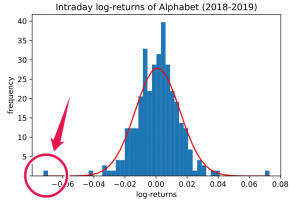
- Then we can visualize the probability density function $f_{(\hat{\mu},\hat{\sigma}^2)}$ on the top of the histogram:



# Statistical model, example III

- **Problem:** extreme events. Especially noted by Mandelbrot (1963).

*"Everybody believes in the [normal approximation], the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact,"* —G. Lippman.



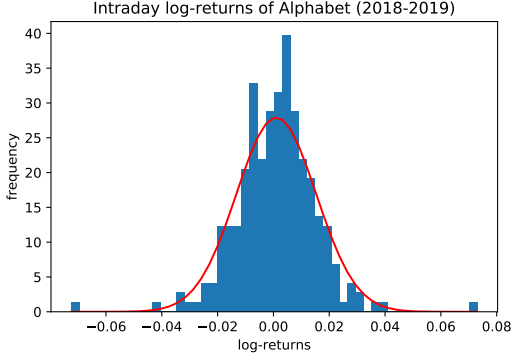

- The probability of this event happening given by the normal distribution is **infinitesimal**...
- ...but these events *do* happen in the real world
- Each year, companies go bankrupt or get incredibly successful.

# Statistical model, example III

- We can compute, in our model:

$$\mathbb{P}_{\hat{\theta}}\left(X \in [-0.07, -0.06]\right) \approx 8.3 \times 10^{-6}\,.$$

- But we typically observe this phenomenon on the course of one year, *i.e.*, 365 observations!



# Non-parametric statistics

- Sometimes we do not make the assumption that $\Theta \subseteq \mathbb{R}^D$: **non-parametric** statistics
- **Examples:**
  - in our model, **every distribution** is possible
  - $\mathcal{P}$ is an infinite-dimensional space, *e.g.*, $X$ has a density on $[0, 1]$
  - finite number of parameters, but increasing with the number of observations
- Important: this is generally a much harder problem
- We focus on *parametric* statistics in this course
- if interested $\rightarrow$ Wasserman, *All of Nonparametric Statistics*, Springer, 2006.

# Identifiability

**Definition:** We say that a statistical model $(\mathcal{S}, \mathcal{P})$ with $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ is *identifiable* if

$$P_{\theta_1} = P_{\theta_2} \quad \Rightarrow \quad \theta_1 = \theta_2 \quad \forall \theta_1, \theta_2 \in \Theta\,.$$

- **Intuition:** different values of the parameter generate different probability distributions
- Thus possible to recover true value of the parameter in the limit

**identifiability $\Rightarrow$ inference is possible**

- **Technical note:** if the model consists of probability density functions, they can differ on a set with measure zero

# 5. Estimation

# Estimation in parametric statistics

- **Goal:** assuming that $X_1, \dots, X_n$ all follow $P_\theta$ for some $\theta \in \Theta$, recover $g(\theta)$ where $g : \Theta \to \mathbb{R}^p$

**Definition:** An *estimator* $\hat{g}$ of $g(\theta)$ is a measurable function of the observations that does not depend on $\theta$. Hence any estimator can be written $\hat{g} = h(X)$.

- not many constraints on $h$ except measurability. Ex: $\hat{g} = 0$.
- it is **very important** that $h$ does not depend on $\theta$, an **unknown** quantity
- strong assumption: existence of $\theta$ such that the observed phenomenon follows $P_\theta$

# Is it doable?

- **Problem:** we are only observing a **finite** number of observations, and each observation is random
- even if we are clever, we could get unlucky
- fundamental example: let $X_1, \ldots, X_n$ be an i.i.d. sample of $X \sim \mathcal{N}(\theta, 1)$
- define

$$\hat{\mu}_n := \frac{1}{n} \sum_{i=1}^{n} X_i$$

(empirical mean = "moyenne empirique")

- the weak law of large numbers tells us that $\hat{\mu}_n \xrightarrow{\mathbb{P}} \mu$
- **asymptotically** (= when we have a lot of observations), there is good hope to recover $g(\theta)$

# Reminder: law of large numbers

**Theorem (Weak Law of Large Numbers):** Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables. Assume that $\mathbb{E}[|X_1|] < +\infty$ and set $\mu := \mathbb{E}[X_1]$. Then

$$\frac{X_1 + \cdots + X_n}{n} \xrightarrow{\mathbb{P}} \mu\,.$$

- **Intuition:** the average of the measurements converges towards the true value
- a stronger statement is true, the *strong* law of large numbers, with almost sure convergence instead of in probability
- multivariate extension: coordinate-wise

# Law of large numbers, in pictures

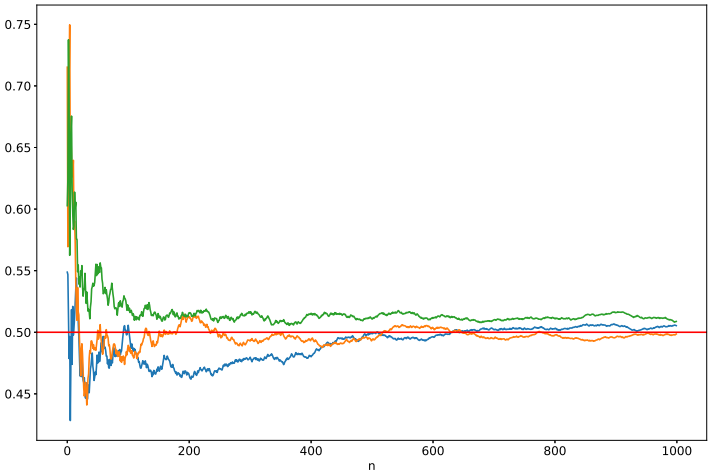



**Figure:** $(\mu_n)_{n\geq 1}$ as a function of $n$ where the $X_i$s are $\mathcal{B}(1/2)$

# Consistent estimators

**Definition:** Let $\hat{g}_n = h_n(X_1, \dots, X_n)$ be an estimator of $g(\theta)$. We say that $\hat{g}_n$ is *consistent* if

$$\forall \theta \in \Theta, \quad \hat{g}_n \xrightarrow[n \to +\infty]{\mathbb{P}} g(\theta)\,.$$

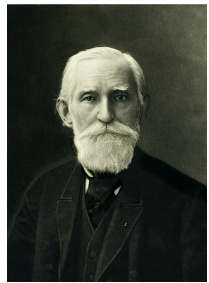
If the convergence is almost sure, we say that $\hat{g}_n$ is *strongly* consistent.

- **Intuition:** the more data, the closer we get from our goal
- **Example:** $\hat{\mu}_n$ is an estimator, and the WLLN shows that it is consistent for $\mu$
- technically speaking, consistency is a statement about the *sequence* of estimators

# 6. The method of moments

# History

- introduced by Chebyshev for his proof of the Central Limit Theorem (1887)
- brought to statistics by Pearson



- Pearson's idea: family of distribution parametrized by the first 4 moments (*Pearson's curves*)
- given data, compute the first 4 empirical moments and pick the distribution with the corresponding population moments

# The method of moments

- **General idea:** given a statistical model $P_\theta$, match the empirical moments with the population moments, and solve for $\theta$
- **Intuition:** WLLN ensures that they will coincide for large $n$
- **Example:** the Gaussian distribution $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$
- $\theta = (\mu, \sigma^2) \in \mathbb{R}^2 \Rightarrow$ compute the first two moments

| | empirical moments | | population moments |
|---|---|---|---|
| $k = 1:$ | $\frac{1}{n}\sum_{i=1}^n X_i =: \hat{\mu}_{1,n}$ | $\Leftrightarrow$ | $\mathbb{E}_\theta[X] = \mu$ |
| $k = 2:$ | $\frac{1}{n}\sum_{i=1}^n X_i^2 =: \hat{\mu}_{2,n}$ | $\Leftrightarrow$ | $\mathbb{E}_\theta\left[X^2\right] = \mu^2 + \sigma^2$ |

- then solve for $\theta$:

$$\hat{\mu}_n = \hat{\mu}_{1,n} \quad \text{and} \quad \hat{s}_n^2 = \hat{\mu}_{2,n} - \hat{\mu}_{1,n}^2 .$$

# Bernoulli distribution

- **Example:** $X_1, \dots, X_n$ i.i.d. $\mathcal{B}(\theta)$ with $\theta \in (0,1)$
- let us follow the recipe: (i) compute the moments
- one parameter $\Rightarrow$ we compute the first moment
- **Recall:** $\mathbb{P}_\theta(X = 1) = \theta$ and $\mathbb{P}_\theta(X = 0) = 1 - \theta$
- therefore

$$\mathbb{E}_\theta[X] = 1 \cdot \theta + 0 \cdot (1 - \theta) = \theta\,.$$

- (ii) match with the empirical moments and solve for $\theta$:

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \mathbb{E}_\theta[X] \Rightarrow \boxed{\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i}\,.$$

# Uniform distribution

- **Reminder:** uniform distribution on $[a, b]$

**Definition:** we say that $X$ has the *uniform distribution* on $[a, b]$ if $X$ has a density with respect to the Lebesgue measure given by

$$\forall t \in \mathbb{R}, \qquad \rho_X(t) = \frac{1}{b-a} \mathbb{1}_{t \in [a,b]} .$$

We write $X \sim \mathcal{U}([a, b])$.

- we can show that, for any $p \geq 0$,

$$\mathbb{E}_{a,b}[X^p] = \frac{1}{p+1} \sum_{k=0}^{p} a^k b^{p-k} .$$

# Uniform distribution, in pictures

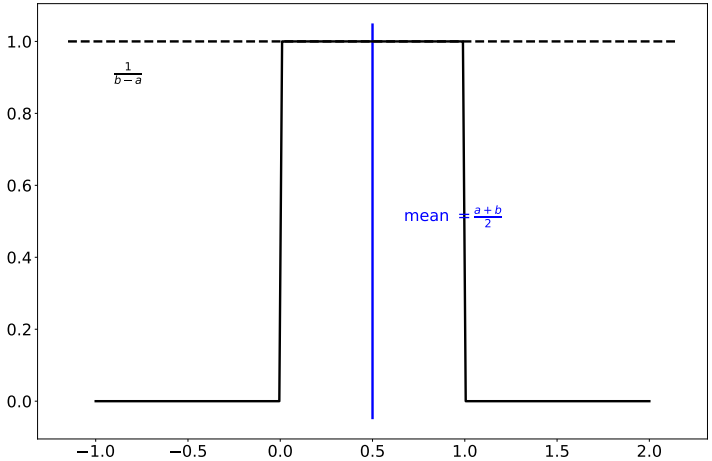



**Figure:** probability density function of the standard uniform distribution

# Uniform distribution (ctd.)

- $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{U}([a,b])$ (German tank problem[1])
- let us use the method of moments to estimate $a$ and $b$
- **Remark:** two parameters to estimated $\Rightarrow$ we need at least two moments
- set $X \sim \mathcal{U}([a,b])$
- recall that $\mathbb{E}_{a,b}[X^p] = \frac{1}{p+1} \sum_{k=0}^{p} a^k b^{p-k}$
- we deduce that

$$\begin{cases} m_1 = \mathbb{E}_{a,b}[X] & = \frac{1}{2}(a+b) \\ m_2 = \mathbb{E}_{a,b}\left[X^2\right] & = \frac{1}{3}(a^2 + ab + b^2)\,. \end{cases}$$

[1] Ruggles and Brody, *An empirical approach to economic intelligence in World War II*, Journal of the American Statistical Association, 1947

# Uniform distribution (ctd.)

- we deduce $b = 2m_1 - a$ from the first equation
- plugging into the second, we see that

$$a^2 - 2am_1 + 4m_1^2 - 3m_2 = 0\,.$$

- solving for $a$ and backtracking, we obtain

$$\begin{cases} a &= m_1 - \sqrt{3(m_2 - m_1^2)} \\ b &= m_1 + \sqrt{3(m_2 - m_1^2)}\,. \end{cases}$$

- **Remark:** Cauchy-Schwarz $\Rightarrow m_2 - m_1^2 \geq 0$
- we obtain the moment estimators:

$$\boxed{\begin{cases} \widehat{a}_n &= \widehat{m}_1 - \sqrt{3(\widehat{m}_2 - \widehat{m}_1^2)} \\ \widehat{b}_n &= \widehat{m}_1 + \sqrt{3(\widehat{m}_2 - \widehat{m}_1^2)}\,. \end{cases}}$$

# Concluding remarks

**Advantages:**

- simple method
- systematic method to produce estimators
- good theoretical properties

**Disadvantages:**

- one need to be able to compute the population moments in **closed-form**. This is **not always the case!**
- not always the "best" estimators

# 7. Maximum likelihood estimation

# Likelihood function

- **Likelihood of an observation:** for a discrete law, probability of observing the outcome under the parameter model ("vraisemblance")
- **Intuition:** how likely is this parametrization of the model given the observations?
- **Definition:**

$$\mathcal{L}(\theta|x) = \mathbb{P}_\theta\left(X = x\right) .$$

- for a continuous distribution, use the density

$$\mathcal{L}(\theta|x) = f_\theta(x) .$$

- if i.i.d. sample, we use the independence property:

$$\begin{aligned}\mathcal{L}(\theta|x_1, \ldots, x_n) &= \mathbb{P}_\theta\left(X_1 = x_1, \ldots, X_n = x_n\right) \\ &= \prod_{i=1}^{n} \mathbb{P}_\theta\left(X_i = x_i\right) = \prod_{i=1}^{n} \mathcal{L}(\theta|x_i) .\end{aligned}$$

# Maximum likelihood estimation

- **Idea:** find the parameter that maximizes the likelihood of the observations in the model
- that is, find the *most likely* parameter given the observations (= "maximum de vraisemblance")
- in other words, pick

$$\hat{\theta}_n^{\mathrm{ML}} \in \arg\max_{\theta \in \Theta} \mathcal{L}(\theta | x_1, \ldots, x_n) \,.$$

- **Trick:** because of the multiplicative nature of the likelihood of an $n$ sample, sometime convenient to maximize the log:

$$\hat{\theta}_n^{\mathrm{ML}} \in \arg\max_{\theta \in \Theta} \log \mathcal{L}(\theta | x_1, \ldots, x_n) \,.$$

- $t \mapsto \log t$ is an increasing function $\Rightarrow$ it does not change the argmax!

# Example: Bernoulli distribution

- the likelihood of an observation is

$$\mathcal{L}(\theta|x) = \mathbb{P}_\theta(X = x) = \begin{cases} \theta & \text{if } x = 1 \\ 1 - \theta & \text{otherwise} \end{cases}$$

- set $N = \sum_{i=1}^{n} x_i$, then

$$\mathcal{L}(\theta|x_1, \ldots, x_n) = \prod_{i=1}^{n} \theta^{\mathbb{1}_{x_i=1}} (1-\theta)^{\mathbb{1}_{x_i=0}} = \theta^N (1-\theta)^{n-N}$$

- standard procedure: take the log and differentiate

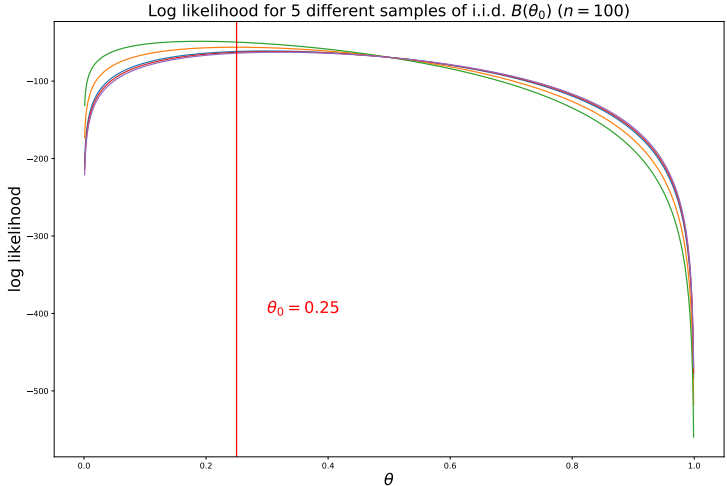
$$\nabla \log \mathcal{L}(\theta|x) = \frac{\mathrm{d}}{\mathrm{d}\theta}(N \log\theta + (n-N)\log(1-\theta)) = \frac{N}{\theta} - \frac{n-N}{1-\theta}\,.$$

- solve for $\theta$:

$$\boxed{\hat{\theta}_n^{\mathrm{ML}} = \frac{N}{n}\,.}$$

(same as the moment estimator)

# Example: Bernoulli distribution



# Example: uniform distribution

- $X_1, \dots, X_n$ i.i.d. $X \sim \mathcal{U}([0, \theta])$ for an unknown $\theta \in \mathbb{R}_+$
- the likelihood of an observation in this model is

$$\mathcal{L}(\theta|x) = f_\theta(x) = \frac{1}{\theta} \mathbb{1}_{x \in [0,\theta]} .$$

- for $n$ i.i.d.observations,

$$\mathcal{L}(\theta|x_1, \dots, x_n) = \frac{1}{\theta^n} \mathbb{1}_{x_1,\dots,x_n \in [0,\theta]} = \frac{1}{\theta^n} \mathbb{1}_{\theta \geq \max_{1 \leq i \leq n} x_i} .$$

- we deduce that

$$\boxed{\hat{\theta}^{\mathsf{ML}} = \max_{1 \leq i \leq n} x_i .}$$

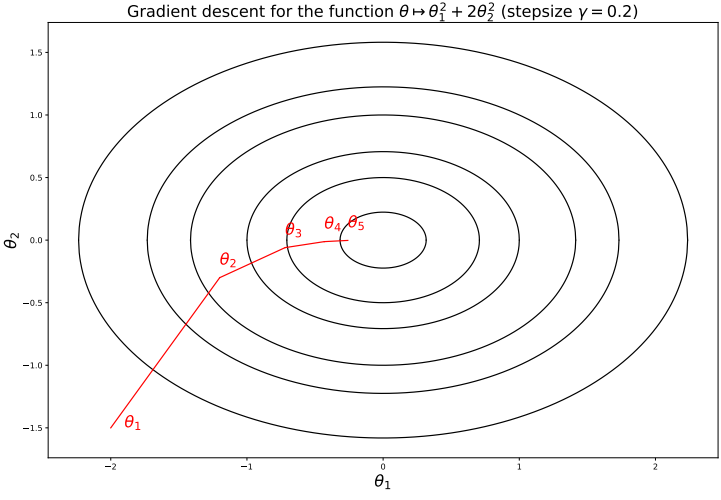
- in particular, it is **different** from

$$\hat{\theta}^{\mathsf{MOM}} = 2\hat{\mu}_{1,n} = \frac{2}{n} \sum_{i=1}^{n} X_i .$$

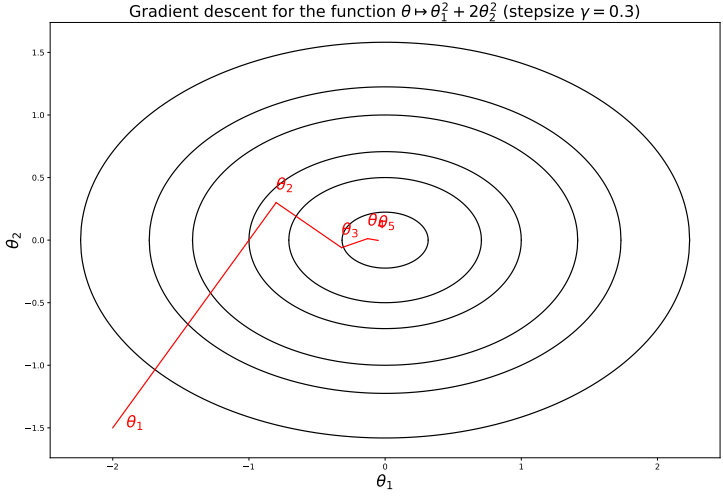
# Maximum likelihood and optimization

- **Important remark:** it is not always possible to obtain the maximum in closed form!
- the expression of
$$\mathcal{L}(\theta|x_1, \ldots, x_n) =: g(\theta)$$
can be *very* complicated
- in that case, we resort to numerical optimization techniques
- the building brick of optimization is **gradient descent:**
$$\begin{cases} \theta^{(0)} & \in \Theta \\ \theta^{(t+1)} & = \theta^{(t)} - \gamma \nabla g(\theta^{(t)}) \,. \end{cases}$$
- good properties if the function to optimize is **convex**

# Gradient descent



# Gradient descent



# Concluding remarks

**Advantages:**

- simple method
- systematic method to obtain estimators
- often good properties (optimal in a sense)

**Disadvantages:**

- need likelihood function
- often hard to solve, need optimization

# 8. Bias variance decomposition

# Bias of an estimator

- how can we judge the quality of an estimator?
- we start with a simple notion, the bias ("biais")

**Definition:** Let $\hat{g} = h(X)$ be an estimator of $g(\theta)$. We call *bias* of $\hat{g}$ the function $b : \Theta \to \mathbb{R}^p$ defined by

$$\text{bias}(\theta) := \mathbb{E}_\theta[\hat{g}] - g(\theta)\,.$$

- **Intuition:** systematic error of $\hat{g}$
- we like **unbiased estimators:** $\forall \theta \in \Theta, \quad b(\theta) = 0$
- or at least *asymptotically unbiased* estimators, that is,

$$\text{bias}(\theta) \xrightarrow[n \to +\infty]{} 0\,.$$

# Example: Bernoulli distribution

- $X_1, \ldots, X_n$ i.i.d. $\mathcal{B}(\theta)$ with unknown $\theta \in [0, 1]$
- we have computed the moment estimator:

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \,.$$

- **Question:** what is the bias of this estimator?
- first compute

$$\mathbb{E}_\theta[\hat{\theta}_n] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\theta[X_i] = \theta \,.$$

- we deduce that

$$\text{bias}(\theta) = \mathbb{E}_\theta[\hat{\theta}_n] - \theta = 0 \,.$$

- the moment estimator for Bernoulli random variables is **unbiased**

# Debiasing estimators

- **Example:** recall that estimating the variance with the method of moments yields

$$\hat{s}_n^2 = \hat{\mu}_2 - \hat{\mu}_1^2 \,.$$

- let us compute

$$\begin{aligned}
\mathbb{E}_\theta\big[\hat{s}_n^2\big] &= \mathbb{E}_\theta\big[\hat{\mu}_2 - \hat{\mu}_1^2\big] && \text{(definition)} \\
&= \mathbb{E}_\theta\left[\frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2\right] && \text{(def. of } \hat{\mu}_{1,n} \text{ and } \hat{\mu}_{2,n}\text{)} \\
&= \mu^2 + \sigma^2 - \frac{1}{n^2}\left(n(\mu^2+\sigma^2) + \frac{n(n-1)}{2}\cdot 2\mu^2\right) && \text{(linearity)} \\
\mathbb{E}_\theta\big[\hat{s}_n^2\big] &= \frac{n-1}{n}\sigma^2 \quad \Rightarrow \quad b(\theta) = \frac{-1}{n}\sigma^2 \neq 0 \,.
\end{aligned}$$

- for this reason, we prefer to use $\hat{\sigma}_n^2 = \frac{n}{n-1}\hat{s}_n^2$.

# Debiasing estimators, ctd.

- back to estimating uniform distribution on $[0, \theta]$
- notice that

$$\mathbb{P}_\theta\left(\max_i X_i \le t\right) = \mathbb{P}_\theta\left(X_1 \le t\right) \cdots \mathbb{P}_\theta\left(X_n \le t\right) = \frac{t^n}{\theta^n} \mathbb{1}_{t \in [0,\theta]}$$

- we get the density $f_\theta(t) = n t^{n-1} \theta^{-n} \mathbb{1}_{t \in [0,\theta]}$
- we can compute

$$\mathbb{E}_\theta[\hat{\theta}^{\mathsf{MLE}}] = \frac{n}{n+1}\theta$$

- thus we prefer to use the *debiased* estimator

$$\hat{\theta}_n = \frac{n+1}{n} \max_{1 \le i \le n} X_i \,.$$

# Mean squared error

- we now turn to another measure of quality of an estimator, the mean squared error ("le risque quadratique")

**Definition:** Let $\hat{g}$ be an estimator of $g(\theta)$. We call *mean squared error* of $\hat{g}$ the quantity

$$\mathrm{MSE}(\hat{g}) := \mathbb{E}_\theta\left[(\hat{g} - g(\theta))^2\right] .$$

- **Intuition:** average error of the estimator from the point of view of $x \mapsto x^2$
- **lower is better**
- can be extended to other loss functions than $(x, y) \mapsto (x - y)^2$
- measure of performance: gives a score that can be estimated, and compared

# Bias-variance decomposition

**Theorem:** Let $\hat{g}$ be an estimator of $g(\theta)$. Then

$$\mathrm{MSE}(\hat{g}) = \mathrm{Var}_\theta(\hat{g}) + b(\theta)^2 .$$

- **Intuition:** decomposition of the MSE in two parts: one depending one the randomness, one depending on the systematic error
- most of the time, difficult to be good on both sides
- we often speak of **bias-variance trade-off**
- one case is clear, though: between two unbiased estimators, pick the one with **lower variance**

# Proof of bias-variance decomposition

The proof is a direct computation:

$$
\begin{aligned}
\mathrm{MSE}(\hat{g}) &= \mathbb{E}_\theta\left[(\hat{g} - g(\theta))^2\right] && \text{(def. of MSE)} \\
&= \mathbb{E}_\theta\left[(\hat{g} - \mathbb{E}_\theta[\hat{g}] + \mathbb{E}_\theta[\hat{g}] - g(\theta))^2\right] && \text{(introduce } \mathbb{E}_\theta[\hat{g}]\text{)} \\
&= \mathbb{E}_\theta\left[(\hat{g} - \mathbb{E}_\theta[\hat{g}])^2 + 2(\hat{g} - \mathbb{E}_\theta[\hat{g}])(\mathbb{E}_\theta[\hat{g}] - g(\theta)) + (\mathbb{E}_\theta[\hat{g}] - g(\theta))^2\right] && \text{(develop)} \\
&= \mathrm{Var}_\theta(\hat{g}) + 2(\mathbb{E}_\theta[\hat{g}] - g(\theta))\mathbb{E}_\theta[\hat{g} - \mathbb{E}_\theta[\hat{g}]] + \mathrm{bias}(\theta)^2 && \text{(definition of } b\text{)} \\
\mathrm{MSE}(\hat{g}) &= \mathrm{Var}_\theta(\hat{g}) + \mathrm{bias}(\hat{g})^2 \,.
\end{aligned}
$$

□

# Example: MSE of the empirical mean

- suppose $X_1, \ldots, X_n$ i.i.d. from a distribution $X$ with mean $\mu$ and (finite) variance $\sigma^2$
- **Question:** what is the mean squared error of $\overline{X}_n$?
- we write

$$\begin{aligned} \mathrm{MSE}(\overline{X}_n) &= \mathbb{E}\left[(\overline{X}_n - \mu)^2\right] \\ &= \mathrm{Var}\left(\overline{X}_n\right) \\ \mathrm{MSE}(\overline{X}_n) &= \frac{\sigma^2}{n} \end{aligned}$$

- back to the i.i.d. $\mathcal{U}([0, \theta])$ case: we have

$$\mathrm{MSE}(\hat{\theta}_n^{MM}) = 4 \cdot \frac{\theta^2}{12} \cdot \frac{1}{n} = \frac{\theta^2}{3n}.$$

# Example, ctd.

- **Question:** is this the best we can do?
- recall that we propose another estimator,

$$\hat{\theta}_n^{MLE} = \frac{n+1}{n} \max_{1 \le i \le n} X_i \,,$$

  also unbiased
- we derived the distribution of $Y_n := \max_{1 \le i \le n} X_i$ and found

$$f(t) = \frac{n}{\theta^n} t^{n-1} \mathbb{1}_{t \in [0,\theta]}$$

  as density
- for the variance, we find

$$\mathrm{Var}\,(Y_n) = \int_0^\theta t^2 \frac{n}{\theta^n} t^{n-1} \mathrm{d}t - \left( \int_0^\theta t \frac{n}{\theta^n} t^{n-1} \mathrm{d}t \right)^2 = \frac{n\theta^2}{(n+1)^2(n+2)} \,.$$

# Example, ctd.

- we deduce that

$$\mathrm{MSE}(\hat{\theta}_n^{\mathsf{MLE}}) = \mathrm{Var}(\hat{\theta}_n^{\mathsf{MLE}}) = \frac{\theta^2}{n(n+2)}\,.$$

- recall that

$$\mathrm{MSE}(\hat{\theta}_n^{\mathsf{MM}}) = \frac{\theta^2}{3n}\,.$$

- the MSE of $\hat{\theta}_n^{\mathsf{MLE}}$ is always smaller
- we would prefer the (debiased) MLE estimator in practice

# 9. Confidence intervals

# Confidence intervals

- **Idea:** give a set that contains the true value of the parameter with high probability (Neyman, 1937)

**Definition:** For any $\alpha \in [0,1]$, a *confidence set* ("région de confiance") for $g(\theta)$ of level $1-\alpha$ is a set $\hat{C}$ measurably constructed with respect to $X$ and such that, for any $\theta \in \Theta$,

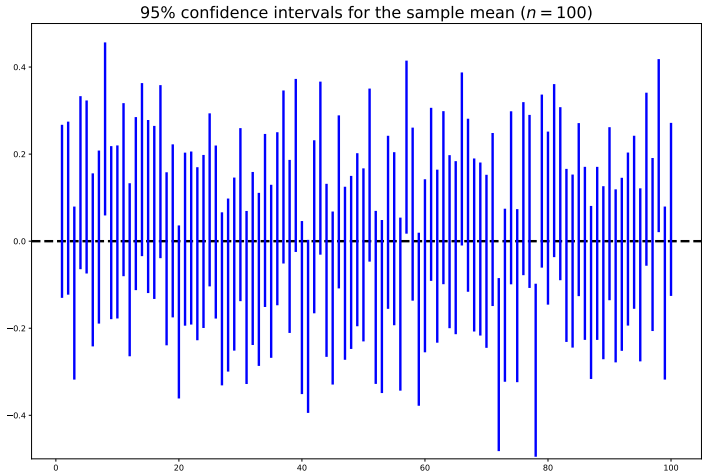
$$\mathbb{P}_\theta \left( g(\theta) \in \hat{C} \right) \geq 1 - \alpha \,.$$

- very often an interval ("intervalle de confiance")
- asymptotic version:

$$\liminf_{n \to +\infty} \mathbb{P}_\theta \left( g(\theta) \in \hat{C}_n \right) \geq 1 - \alpha \,.$$

# Confidence intervals are **random**

- not the value that they contain: $g(\theta)$ is **fixed!**



# Confidence interval, basic example

- let $X_1, \dots, X_n$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$ with **known variance**
- estimator for the mean: $\hat{g}_n = \hat{\mu}_{1,n}$ (empirical mean)
- let us build a **two-sided** confidence interval with level $1 - \alpha$ for $\mu$: we want $\delta$ such that

$$\mathbb{P}\left(\mu \in [\hat{g}_n - \delta, \hat{g}_n + \delta]\right) \geq 1 - \alpha\,.$$

- $\hat{g}_n \sim \mathcal{N}(\mu, \sigma^2/n)$, so we can compute

$$\begin{aligned}
\mathbb{P}\left(\mu \in [\hat{g}_n - \delta, \hat{g}_n + \delta]\right) &= \mathbb{P}\left(-\delta \leq \mu - \hat{g}_n \leq \delta\right) \\
&= \mathbb{P}\left(-\delta \leq \mathcal{N}\left(0, \sigma^2/n\right) \leq \delta\right) \\
&= \mathbb{P}\left(\mathcal{N}(0,1) \in [-\delta\sqrt{n}/\sigma, \delta\sqrt{n}/\sigma]\right) \\
&= \Phi\left(\delta\sqrt{n}/\sigma\right) - \Phi\left(-\delta\sqrt{n}/\sigma\right) \\
&= 2\Phi\left(\delta\sqrt{n}/\sigma\right) - 1 \qquad \text{(symmetry)}
\end{aligned}$$

# Confidence interval, basic example, ctd.

- now we must solve $2\Phi\left(\delta\sqrt{n}/\sigma\right) - 1 \geq 1 - \alpha$, that is,

$$\Phi\left(\delta\sqrt{n}/\sigma\right) \geq 1 - \frac{\alpha}{2}\,.$$

- we obtain

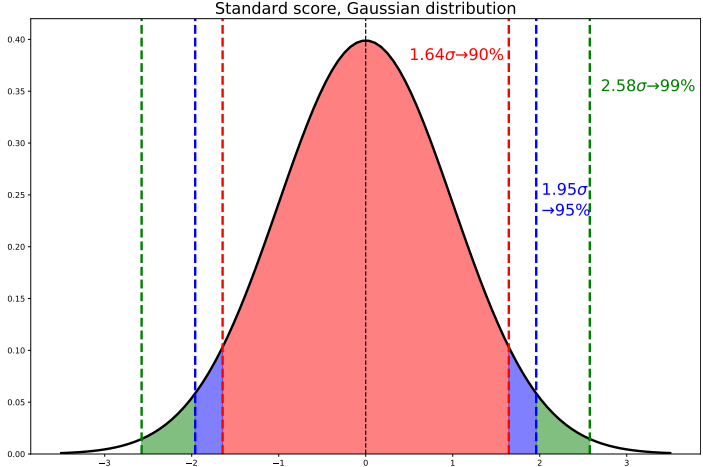
$$\delta = \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\,.$$

- usually, we set $z_p := \Phi^{-1}(p)$, and the confidence interval is

$$\hat{C}_{1-\alpha} = \left[\mu - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \mu + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

# Standard scores of the Gaussian distribution

- typical values for $\alpha$ are $1\%, 5\%$, and $10\%$. associated $z^\star$ are $1.64, 1.95$, and $2.58$, respectively



# Student's $t$-distribution

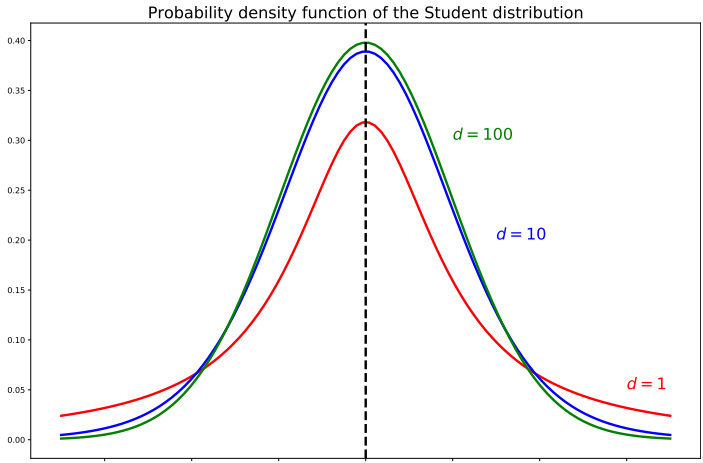
- Student = William Gosset (1876–1937)

- $X_1, \ldots, X_n$ i.i.d. $\mathcal{N}\left(\mu, \sigma^2\right)$
- define $\hat{\sigma}_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \hat{\mu}_{1,n})^2$ the sample variance
- then the random variable

$$T := \frac{\hat{\mu}_{1,n} - \mu}{\hat{\sigma}_n / \sqrt{n}}$$

follows the **Student distribution with $n-1$ degrees of freedom**

# Student's $t$-distribution, in pictures



# Confidence interval, unknown variance

- when situation more complicated, use a **pivotal** quantity
- **Example:** let $X_1, \dots, X_n$ i.i.d. $\mathcal{N}\left(\mu, \sigma^2\right)$ with **unknown variance**
- confidence interval for the empirical mean:

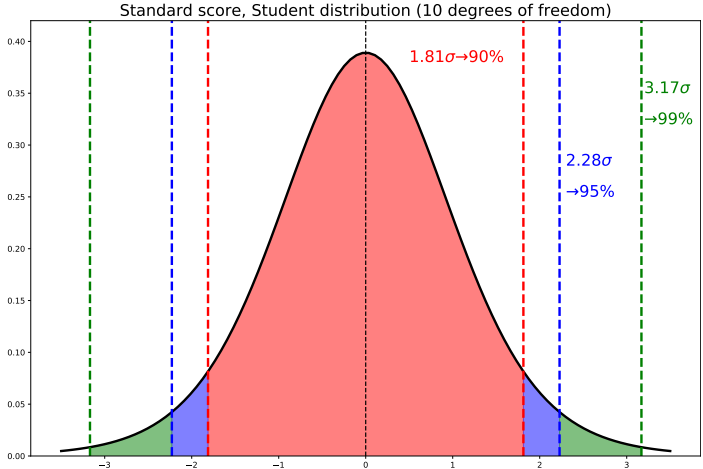
$$\begin{aligned} 1-\alpha &\leq \mathbb{P}\left(-z_{\alpha/2,n-1} \leq T_{n-1} \leq z_{\alpha/2,n-1}\right) \\ &= \mathbb{P}\left(-z_{\alpha/2,n-1} \leq \frac{\hat{\mu}_{1,n}-\mu}{\hat{\sigma}/\sqrt{n}} \leq z_{\alpha/2,n-1}\right) \\ &= \mathbb{P}\left(-z_{\alpha/2,n-1}\frac{\hat{\sigma}_n}{\sqrt{n}} \leq \hat{\mu}_{1,n}-\mu \leq z_{\alpha/2,n-1}\frac{\hat{\sigma}_n}{\sqrt{n}}\right). \end{aligned}$$

- we deduce the confidence interval of level $1-\alpha$

$$\hat{C}_{1-\alpha} = \left[\hat{\mu}_{1,n} - z_{\alpha/2,n-1}\frac{\hat{\sigma}_n}{\sqrt{n}}, \hat{\mu}_{1,n} + z_{\alpha/2,n-1}\frac{\hat{\sigma}_n}{\sqrt{n}}\right]$$

# Standard scores of the Student's $t$-distribution

- typical values for $\alpha$ are $1\%, 5\%$, and $10\%$. associated $z^\star$ are $1.64, 1.96$, and $2.58$, respectively



# 10. Confidence intervals with concentration inequalities

# Chebyshev's inequality

- Bienaymé-Chebyshev inequality (1853-1867)

**Proposition:** Let $Y$ be a random variable such that $\mathbb{E}\left[Y^2\right] < +\infty$. Then, for any $\varepsilon > 0$,

$$\mathbb{P}\left(|Y - \mathbb{E}[Y]| \geq \varepsilon\right) \leq \frac{\mathrm{Var}(Y)}{\varepsilon^2}.$$

- **Intuition:** the values of $Y$ are "concentrated" around its mean
- 75% of values lie within $2\sigma$ of the mean, 89% within $3\sigma$
- **weaker** than Gaussian concentration, which gives 95% and 99.7%
- **But** we can use concentration inequalities to obtain confidence intervals when we do not know the law

# Concentration with Chebyshev

- $X_1, \dots, X_n$ i.i.d. $\mathcal{B}(\theta)$ with $\theta \in (0,1)$
- what is the variance of $X_1$?

$$\mathrm{Var}_\theta(X_1) = \theta(1-\theta).$$

- in particular, **uniform bound** on the variance:

$$\forall \theta \in (0,1), \quad \mathrm{Var}_\theta(X_1) \leq 1/4.$$

- what is the variance of $\hat{\mu}_{1,n}$?

$$\mathrm{Var}_\theta(\hat{\mu}_{1,n}) = \frac{\mathrm{Var}(X_1)}{n} \leq \frac{1}{4n}.$$

- Chebyshev: $\mathbb{P}_\theta(|\hat{\mu}_{1,n} - \theta| \geq \varepsilon) \leq 1/(4n\varepsilon^2)$
- pick $\varepsilon = 1/(2\sqrt{n\alpha})$, confidence interval of level $1-\alpha$ for $\theta$:

$$\hat{C}_{1-\alpha} = \left[\hat{\mu}_{1,n} - \frac{1}{2\sqrt{n\alpha}}, \hat{\mu}_{1,n} + \frac{1}{2\sqrt{n\alpha}}\right]$$

# Hoeffding's inequality

**Proposition (Hoeffding's inequality):**[2] Let $Y_1, \dots, Y_n$ be independent *centered* random variables, and $(a_1, \dots, a_n)$ and $(b_1, \dots, b_n)$ *deterministic* real numbers such that for any $i \in \{1, \dots, n\}$, $Y_i \in [a_i, b_i]$ a.s. Then, for any $t \geq 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n} Y_i \geq t\right) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

- in particular, if $X_1, \dots, X_n$ are i.i.d., and if $a_1 = \dots = a_n = a$ and $b_1 = \dots = b_n = b$,

$$\mathbb{P}_\theta\left(|\hat{\mu}_{1,n} - \mathbb{E}_\theta[X_1]| > t\right) \leq 2\exp\left(\frac{-2nt^2}{(b-a)^2}\right).$$

[2] Hoeffding, *Probability inequalities for sum of bounded random variables*, Journal of the American Statistical Association, 1963

# Confidence intervals with Hoeffding

- same example: $X_1, \dots, X_n$ i.i.d. $\mathcal{B}(\theta)$ with $\theta \in (0,1)$
- we can take $a = 0$ and $b = 1$ since $X_1 \in \{0,1\}$ a.s.
- Hoeffding inequality yields:

$$\mathbb{P}_\theta \left( |\hat{\mu}_{1,n} - \theta| \geq \varepsilon \right) \leq 2\exp\left(-2n\varepsilon^2\right) .$$

- set $\varepsilon = \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}}$, we have obtained the following confidence interval of level $1-\alpha$ for $\theta$:

$$\hat{C}_{1-\alpha} = \left[ \hat{\mu}_{1,n} - \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}}, \hat{\mu}_{1,n} + \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}} \right]$$

- Hoeffding gives a better interval (smaller): **stronger assumptions**

# 11. Convergence of random variables

# Almost sure convergence

**Definition:** We say that the sequence $(X_n)_{n\geq 1}$ converges *almost surely* towards $X$ ("presque sûrement") if

$$\mathbb{P}\left(\lim_{n\to+\infty} X_n = X\right) = 1\,.$$

We write $X_n \xrightarrow{\text{a.s.}} X$.

- **Intuition:** stabilization of all the trajectories
- strongest notion of convergence for random variables
- **Example:** $X_i$ is the amount of energy emitted by a star at time $i$. One day the star will die: $X_i \xrightarrow{\text{a.s.}} 0$

# Convergence in probability

**Definition:** We say that the sequence $(X_n)_{n\geq 1}$ converges *in probability* towards $X$ ("en probabilité") if

$$\forall \varepsilon > 0, \quad \lim_{n\to+\infty} \mathbb{P}\left(|X_n - X| \geq \varepsilon\right) = 0\,.$$

We write $X_n \xrightarrow{\mathbb{P}} X$.

- **Intuition:** measurement errors converge to zero
- almost sure convergence implies convergence in probability
- the converse is **not true!**
- **Example:** law of large numbers

# Convergence in distribution

**Definition:** Let $F_n$ be the cumulative distribution function of $X_n$ and $F$ that of $X$. We say that the sequence $(X_n)_{n\geq 1}$ converges *in distribution* towards $X$ ("en loi") if

$$\lim_{n\to+\infty} F_n(x) = F(x)\,,$$

for all $x$ that is a point of continuity of $F$. We write $X_n \xrightarrow{\mathcal{L}} X$.

- **Intuition:** law of the errors becomes more and more adequately modelized
- weakest form of convergence
- $X$ does not really exist in this definition: we can write $X_n \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$
- **Example:** central limit theorem

# Reminder: central limit theorem

**Theorem (univariate version):** Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables. Assume that $\mathbb{E}\left[\left|X_1^2\right|\right] < +\infty$. Define $\mu := \mathbb{E}[X_1]$ and $\sigma^2 := \mathrm{Var}(X_1)$. Then

$$\sqrt{n}\left(\frac{X_1 + \cdots + X_n}{n} - \mu\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \sigma^2\right) .$$

- **What it means:** the *rescaled* errors between the mean of measurements and the true value are *normaly distributed*.
- multivariate version ($X_1 \in \mathbb{R}^d$):

$$\sqrt{n}\left(\frac{X_1 + \cdots + X_n}{n} - \mu\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \Sigma\right) ,$$

where $\Sigma \in \mathbb{R}^{d \times d}$ is the **covariance matrix** of $X_1$

# Central limit theorem, in pictures

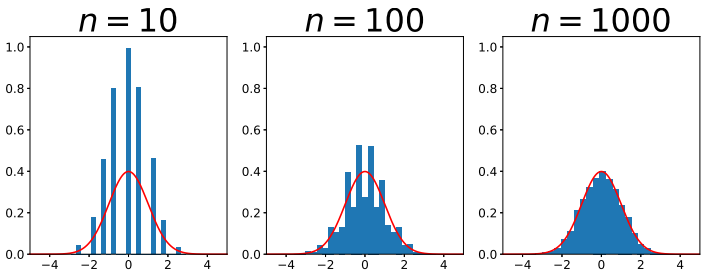



**Figure:** Bernoulli random variables, $10^4$ repetitions, normal distribution density in red

# Convergence in $L^p$

**Definition:** Let $p \geq 1$. We say that the sequence $(X_n)_{n\geq 1}$ *converges in* $L^p$ towards $X$ if the $p$-th absolute moments $\mathbb{E}\left[|X_n|^p\right]$ and $\mathbb{E}\left[|X|^p\right]$ exist and if

$$\lim_{n\to+\infty} \mathbb{E}\left[|X_n - X|^p\right] = 0\,.$$

We write $X_n \xrightarrow{L^p} X$. When $p = 1$ we say that $X_n$ converges in *mean* towards $X$ ("converge en moyenne"), and when $p = 2$ we say that $X_n$ converges in *quadratic mean* towards $X$ ("en moyenne quadratique").

- **Intuition:** controling the moments of the sequence $(X_n)_{n\geq 1}$
- implies convergence in probability
- sometime it is more convenient to compute the moments

# Mapping theorem

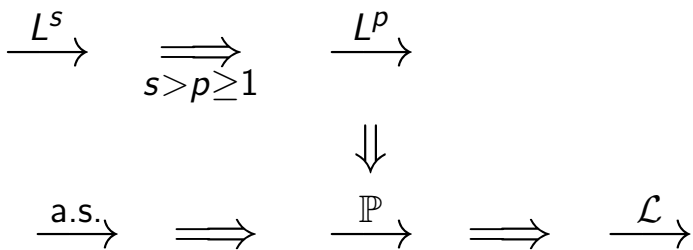
**Theorem:** Assume that $g$ is a continuous mapping on a set that has probability 1. Then

- if $X_n \xrightarrow{\text{a.s.}} X$, then $g(X_n) \xrightarrow{\text{a.s.}} g(X)$;
- if $X_n \xrightarrow{\mathbb{P}} X$, then $g(X_n) \xrightarrow{\mathbb{P}} g(X)$;
- if $X_n \xrightarrow{\mathcal{L}} X$, then $g(X_n) \xrightarrow{\mathcal{L}} g(X)$.

- **Intuition:** we can apply a continuous map
- useful to build upon basic results (LLN, TCL)

# Summary

- we have the following relationships between the different notions of convergence:



# 12. Asymptotic confidence intervals

# Asymptotic normality of estimators

- one way to obtain asymptotic confidence intervals:

**Definition:** We say that an estimator $\hat{g}_n$ of $g(\theta)$ is *asymptotically normal* if it satisfies a central limit theorem statement. That is,

$$\sqrt{n}(\hat{g}_n - g(\theta)) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \sigma^2\right) .$$

- **Intuition:** speed of the convergence of $\hat{g}_n - g(\theta)$ towards zero
- **stronger property** than consistency
- **Problem:** very often, $\sigma^2 = \sigma^2(\theta)$
- we cannot construct a confidence interval that depends on $\theta$!

# Slutsky's theorem

- one solution: replace $\sigma^2$ by an estimate of the variance
- Slutsky's theorem: still possible to get the convergence

**Theorem (Slutsky, 1925):** Let $(X_n)_{n\geq 1}$ and $(Y_n)_{n\geq 1}$ be two sequences of random variables. Suppose that $X_n \xrightarrow{\mathcal{L}} X$ and $Y_n \xrightarrow{\mathbb{P}} c$, where $c$ is a constant. Then

$$X_n + Y_n \xrightarrow{\mathcal{L}} X + c$$
$$X_n Y_n \xrightarrow{\mathcal{L}} cX$$
$$X_n / Y_n \xrightarrow{\mathcal{L}} X/c \text{ if } c \neq 0$$

# Application of Slutsky's theorem

- suppose that $\hat{\sigma}_n^2 \xrightarrow{\mathbb{P}} \sigma^2$, with $\sigma^2 \neq 0$
- suppose also that

$$\sqrt{n}(\hat{\mu}_{1,n} - \mu(\theta)) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \sigma^2\right) ,$$

  where $\mu(\theta) = \mathbb{E}_\theta[X_1]$.
- then, Slutsky's theorem yields

$$\sqrt{\frac{n}{\hat{\sigma}_n^2}}(\hat{\mu}_{1,n} - \mu(\theta)) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, 1\right) .$$

- we obtain the asymptotic confidence interval of level $1 - \alpha$

$$\hat{C}_{1-\alpha} = \left[\hat{\mu}_{1,n} - z_{1-\alpha/2}\frac{\hat{\sigma}_n}{\sqrt{n}}, \hat{\mu}_{1,n} + z_{1-\alpha/2}\frac{\hat{\sigma}_n}{\sqrt{n}}\right]$$

# 13. Delta method

# The delta method

- another possibility: variance-stabilizing transformations

**Theorem:** Let $(U_n)_{n\geq 1}$ be a sequence of random vectors in $\mathbb{R}^m$, $(a_n)_{n\geq 1}$ be a deterministic sequence of real numbers and $\ell : \mathbb{R}^m \to \mathbb{R}^p$ such that

- $a_n \to +\infty$
- there exist a deterministic $U \in \mathbb{R}^m$ and $\Sigma \in \mathbb{R}^{m\times m}$ such that

$$a_n(U_n - U) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma)$$

- $\ell$ is differentiable at $U$ ($\nabla\ell(U) \in \mathbb{R}^{p\times m}$)

Then

$$a_n(\ell(U_n) - \ell(U)) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \nabla\ell(U)\Sigma\nabla\ell(U)^\top\right) .$$

- **Intuition:** Taylor expansion for random variables

# Stabilizing the variance with the delta method

- suppose that we have built an estimator $\hat{g}$ that satisfies

$$\sqrt{n}(\hat{g}_n - g(\theta)) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \sigma^2(\theta)\right) \quad ,$$

  and that we want to find an asymptotic confidence interval for $g(\theta)$
- the delta method tells us that

$$\sqrt{n}(\phi(\hat{g}) - \phi(g(\theta))) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \sigma^2(\theta)\phi'(g(\theta))^2\right) .$$

- if we find $\phi$ increasing such that $\sigma^2(\theta)\phi'(g(\theta))^2 = 1$, then

$$\hat{C}_{1-\alpha} = \left[\phi^{-1}\left(\phi(\hat{g}_n) - \frac{z_{1-\alpha/2}}{\sqrt{n}}\right), \phi^{-1}\left(\phi(\hat{g}_n) + \frac{z_{1-\alpha/2}}{\sqrt{n}}\right)\right]$$

  is a confidence interval of level $1-\alpha$ for $g(\theta)$

# Example: Bernoulli distribution

- $X_1, \ldots, X_n$ i.i.d. $\mathcal{B}(\theta)$
- method of moments:

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \,.$$

- **What is the variance?** central limit theorem:

$$\sigma^2(\theta) = \mathrm{Var}\,(X_1) = \theta(1-\theta)$$

- variance stabilization: we have to solve

$$\theta(1-\theta)\phi'(\theta)^2 = 1 \quad \Rightarrow \quad \phi(\theta) = 2\arcsin\sqrt{\theta}\,.$$

# 14. Statistical testing

# Statistical test

- **Example:** secondary sex ratio (= number of men / number of women *at birth*)
- intuitively, this should be equal to $\frac{1}{2}$: random symmetric assignment
- **Empirically:** $\approx 107/100$
- Arbuthnot (1710): examined birth records for 82 years in London

- used simple procedure, deduced that probability of observed outcome due to chance less than $0.5^{82}$

# Statistical hypotheses

- first informal definition: **decide whether the observations agree with our model**
- initial research hypothesis, truth unknown, *e.g.*, sex ratio $= \frac{1}{2}$
- other examples: efficiency of a drug, better way of teaching...
- formally, we work in a statistical model

$$\mathcal{P} = \{P_\theta \text{ s.t. } \theta \in \Theta\},$$

  and **split** $\Theta$ in two *disjoint* subsets $\Theta_0$ and $\Theta_1$
- **Important:** we do not require $\Theta_0 \cup \Theta_1 = \Theta$
- we define
  - $H_0 : \theta \in \Theta_0$ the null hypothesis ("hypothèse nulle")
  - and $H_1 : \theta \in \Theta_1$ the alternative hypothesis ("hypothèse alternative")
- given realization of $X \sim P_\theta$, **we want to decide whether $H_0$ or $H_1$ holds**

# Statistical testing

**Definition:** we call *test* of $H_0$ versus $H_1$ any function $\phi$ with values in $\{0,1\}$, where $\phi$ is $X$-measurable and can depend on $\Theta_0$ and $\Theta_1$. When $\phi(X)=0$, we conserve $H_0$, when $\phi(X)=1$ we *reject* $H_0$.

- **Remark:** any test can be written $\phi(X)=\mathbb{1}_{h(X)\in R}$, where $h$ is $X$-measurable
- we call $h$ the *test statistic* ("statistique de test") and $R$ the *critical region* ("région de rejet")
- **Important:** *presumed innocent until proven guilty:* reject the null only if enough evidence is collected
- we have to be conservative in choosing $H_0$

# Type I and type II error

- type I error = wrongly rejecting the null = **false positive** ("faux positif")
- type II error = not rejecting a false null hypothesis = **false negative** ("faux négatif")

**Definition:** We define the functions

$$\begin{aligned} \underline{\alpha} : \Theta_0 &\longrightarrow [0,1] \\ \theta &\longmapsto \mathbb{P}_\theta\left(\phi(X) = 1\right) \end{aligned}$$

("risque de première espèce")

and

$$\begin{aligned} \underline{\beta} : \Theta_1 &\longrightarrow [0,1] \\ \theta &\longmapsto \mathbb{P}_\theta\left(\phi(X) = 0\right) \end{aligned}$$

("risque de seconde espèce")

We call *power* of a test the function $1 - \underline{\beta}$ ("fonction puissance").

# Type I and II errors, ctd.

Executive summary:

| Error types | | Null hypothesis is | |
|---|---|---|---|
| Decision | | True | False |
| about | don't reject | correct inference<br>= true negative | type II error<br>= false negative |
| $H_0$ | reject | type I error<br>= false positive | correct inference<br>= true positive |

- think about testing for a disease:
  - **positive** means sick
  - **negative** means healthy
- **Important:** the situation is not symmetric! and generally we want to control the type II error

# Bonus vocabulary

- in the medical world, different vocabulary
- **Sensitivity:** ("sensibilité")

$$\frac{\text{true positive}}{\text{true positive} + \text{false negative}}.$$

- probability of a positive test on positive patients
- **Specificity:** ("spécificité")

$$\frac{\text{true negative}}{\text{true negative} + \text{false positive}}.$$

- probability of a negative test on negative patients

# 15. Size of a test

# Size of a test

**Definition:** We call *size* of a test ("taille") the number

$$\alpha^\star := \sup_{\theta \in \Theta_0} \underline{\alpha}(\theta) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta \left(\phi(X) = 1\right) .$$

We say that the test is *of level* $\alpha$ ("de niveau $\alpha$") if $\alpha^\star \leq \alpha$.

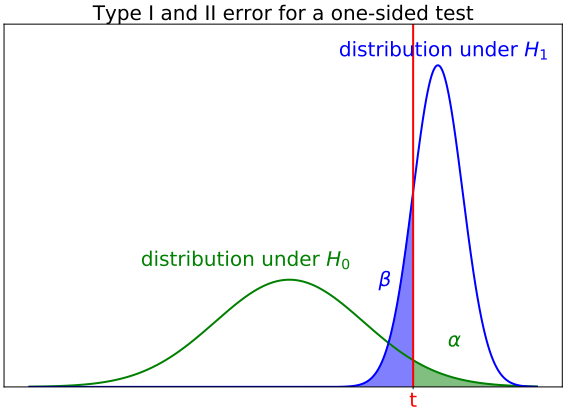
- **Intuition:** maximum probability of wrongly rejecting the null
- asymptotic version:

$$\limsup_{n \to +\infty} \alpha_n^\star \leq \alpha \, .$$

- in a perfect world, we would like to have $\alpha = 0$ and $\underline{\beta} = 0$
- unfortunately, this is often **impossible**: fix a level and try to maximize the power $(1 - \underline{\beta})$

# Link between $\alpha$ and $\beta$

- suppose that $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$ ("hypothèses simples")



# 16. First example

# A first example

- $X_1, \dots, X_n$ i.i.d. $\mathcal{N}(\theta, 1)$, $\theta \in \mathbb{R}$ unknown
- we want to test $H_0 : \theta \geq 1$ vs $H_1 : \theta < 1$
- **Can you think of a test?**
- we consider

$$\phi(X) = \mathbb{1}_{h(X) \in (-\infty, 1)} \quad \text{with} \quad h(X) = \overline{x}_n := \frac{1}{n} \sum_{i=1}^{n} X_i \,.$$

- let $\theta \geq 1$:

$$\begin{aligned} \underline{\alpha}(\theta) &= \mathbb{P}_\theta \left( \phi(X) = 1 \right) \\ &= \mathbb{P}_\theta \left( \overline{x}_n < 1 \right) \\ &= \mathbb{P} \left( \mathcal{N} \left( \theta, \frac{1}{n} \right) < 1 \right) \\ &= \mathbb{P} \left( \mathcal{N}(0,1) < \sqrt{n}(1-\theta) \right) \\ \underline{\alpha}(\theta) &= \Phi(\sqrt{n}(1-\theta)) \,. \end{aligned}$$

# First example, ctd.

- **What is the size of this test?**
- according to the definition:

$$\begin{aligned}\alpha^\star &= \sup_{\theta\in\Theta_0} \underline{\alpha}(\theta) \\ &= \sup_{\theta\geq 1} \Phi(\sqrt{n}(1-\theta))\end{aligned}$$

- $\Phi$ is an increasing function, thus $\Phi(\sqrt{n}(1-\cdot))$ is decreasing:

$$\alpha^\star = \Phi(0) = \frac{1}{2}\,.$$

- not very good...
- **Can we prescribe $\alpha^\star$?**

# First example, improved

- **Idea:** change $R$ such that our test has level $\alpha$
- we propose $R_\alpha = ]-\infty, k_\alpha[$ instead of $]-\infty, 1[$
- let $\theta \geq 1$:

$$\begin{aligned}
\alpha^\star &= \sup_{\theta \in \Theta_0} \mathbb{P}_\theta \left(\phi(X) = 1\right) \\
&= \sup_{\theta \geq 1} \mathbb{P}\left(\mathcal{N}\left(\theta, \frac{1}{n}\right) < k_\alpha\right) \\
&= \sup_{\theta \geq 1} \Phi(\sqrt{n}(k_\alpha - \theta)) \\
\alpha^\star &= \Phi(\sqrt{n}(k_\alpha - 1)) \,.
\end{aligned}$$

- for a given $\alpha$, we solve $\alpha^\star = \alpha$:

$$\boxed{k_\alpha = 1 + \frac{z_\alpha}{\sqrt{n}}}$$

where $z_\alpha$ is, as before, the quantile of order $\alpha$ of the $\mathcal{N}(0,1)$

# Numerical application

- **Reminder:** quantiles of the Gaussian:

$$z_{0.01} = -2.326, \quad z_{0.05} = -1.645, \quad z_{0.1} = -1.28\,.$$

- let us fix $\alpha = 0.1$
- suppose that $n = 5$, then $k_\alpha = 1 - 1.28/\sqrt{5} \approx 0.43$
- we observe

$$x_1 = 0.9, \quad x_2 = -0.1, \quad x_3 = 0.5, \quad x_4 = 1.5, \quad x_5 = -0.8\,.$$

- then $\overline{x}_n = 0.4$, and we would reject $H_0$

# 17. Building tests

# Direct construction

- $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ i.i.d. Gaussian, unknown means and variances $\sigma_X^2$ and $\sigma_Y^2$
- we want to test

$$H_0 : \sigma_X^2 = \sigma_Y^2 \quad \text{vs} \quad H_1 : \sigma_X^2 \neq \sigma_Y^2 \,.$$

- define $\overline{x}_n := \frac{1}{n} \sum_{i=1}^n X_i$ and $\overline{y}_m := \frac{1}{m} \sum_{i=1}^m Y_i$ the **sample means**
- we define the **sample variances**

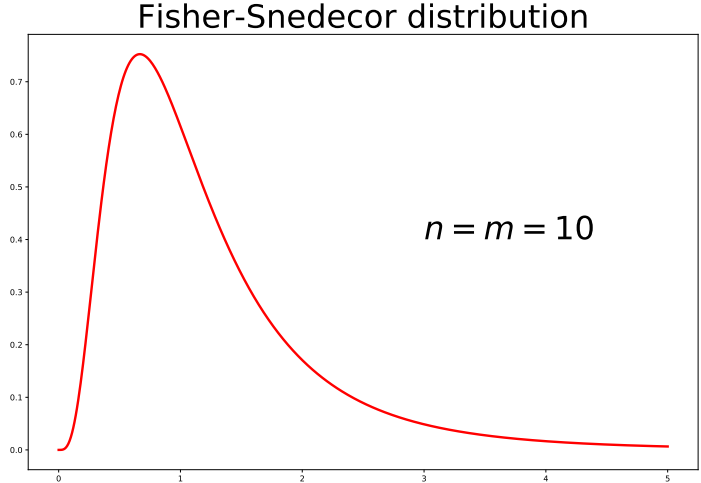
$$S_X^2 := \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{x}_n)^2 \quad \text{and} \quad S_Y^2 := \frac{1}{m-1} \sum_{i=1}^m (Y_i - \overline{y}_m)^2 \,.$$

- then, under the null, the law of

$$F := \frac{S_X^2}{S_Y^2}$$

is known and depend only on $n$ and $m$: $F \sim \mathcal{F}_{n-1,m-1}$

# $F$-distribution




- mean $= \frac{m}{m-2}$, complicated variance

# Direct construction, ctd.

- **Idea:** determine the law of the test statistic under the null and use the quantiles to build $R$
- here, for a given $\alpha$, we find $f_{\alpha/2,n-1,m-1}$ and $f_{1-\alpha/2,n-1,m-1}$ such that

$$\mathbb{P}\left(f_{\alpha/2,n-1,m-1} \leq F \leq f_{1-\alpha/2,n-1,m-1}\right) = 1-\alpha\,.$$

- then we define the $F$-**test** for equality of variances

$$\phi(X,Y) := \mathbb{1}_{\frac{S_X^2}{S_Y^2} \notin [f_{\alpha/2,n-1,m-1}, f_{1-\alpha/2,n-1,m-1}]}\,.$$

- **Example:** for $n = m = 10$ and $\alpha = 0.05$,

$$f_{\alpha/2,n-1,m-1} \approx 0.31 \quad \text{and} \quad f_{1-\alpha/2,n-1,m-1} \approx 3.18\,.$$

# Building tests from confidence intervals

- **Idea:** from any confidence interval, we can build a test of fit
- suppose that $\hat{C}$ is a $1-\alpha$ level confidence interval for $\theta$
- then in order to test

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta \neq \theta_0$$

we can use the test

$$\phi(X) = \mathbb{1}_{\theta_0 \notin \hat{C}}\,.$$

- **What is the level of that test?**
- let $\theta \in \Theta_0$. By definition

$$\begin{aligned}\alpha^\star &= \mathbb{P}_{\theta_0}\left(\phi(X) = 1\right) \\ &= \mathbb{P}_{\theta_0}\left(\theta_0 \notin \hat{C}\right) \\ \alpha^\star &\leq \alpha\end{aligned}$$

# One sample Student $t$-test

- $X_1, \dots, X_n$ i.i.d. $\mathcal{N}\left(\mu, \sigma^2\right)$, $\mu$ and $\sigma$ unknown
- we want to test

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0 \,.$$

- we have seen that

$$T = \frac{\overline{x}_n - \mu}{\hat{\sigma}_n / \sqrt{n}} \sim \mathcal{T}_{n-1} \,,$$

  where $\mathcal{T}_{n-1}$ is the **Student's law** with $n-1$ degrees of freedom
- for any given $\alpha \in (0,1)$, we obtained the $1-\alpha$ level confidence interval for $\mu$

$$\hat{C}_{1-\alpha} = \left[ \hat{\mu}_{1,n} - z_{\alpha/2,n-1} \frac{\hat{\sigma}_n}{\sqrt{n}}, \hat{\mu}_{1,n} + z_{\alpha/2,n-1} \frac{\hat{\sigma}_n}{\sqrt{n}} \right]$$

- $\Rightarrow$ the test $\phi(X) = \mathbb{1}_{\mu_0 \notin \hat{C}_{1-\alpha}}$ has level $\alpha$

# Building tests from likelihood ratio

- **Recall:** likelihood function

$$\mathcal{L}(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f_\theta(x_i)\,.$$

- **Intuition:** $\mathcal{L}(\theta)$ high if $\theta$ in accordance with the observations, low otherwise
- **Idea:** we want to test

$$H_0 : \theta \in \Theta_0 \quad \text{vs} \quad H_1 : \theta \in \Theta_1\,.$$

Let us compare $\mathcal{L}(\theta)$ for $\theta \in \Theta_0$ and $\theta \in \Theta_1$ *via* their ratio

- we set

$$h(X) := \frac{\sup_{\theta\in\Theta_1} \mathcal{L}(\theta; X)}{\sup_{\theta\in\Theta_0} \mathcal{L}(\theta; X)}\,,$$

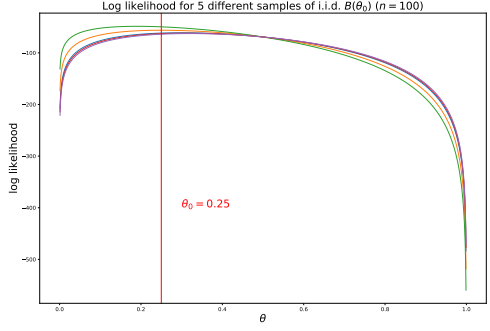
and define the *likelihood ratio* test ("test de rapport de vraisemblance") $\phi(X) = \mathbb{1}_{h(X)>k_\alpha}$

# Gaussian likelihood ratio test

- $X_1, \ldots, X_n$ i.i.d. $\mathcal{N}(\theta, 1)$, $\theta$ unknown
- recall that, in this case,

$$\mathcal{L}(\theta; X) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(\frac{-1}{2}\sum_{i=1}^{n}(X_i - \theta)^2\right).$$

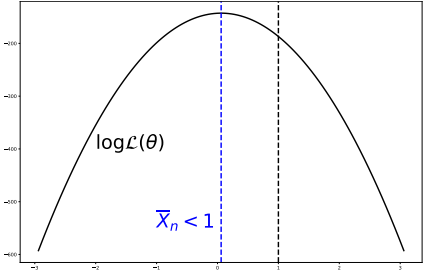
- $\mathcal{L}(\theta; X)$ is increasing on $]-\infty, \overline{x}_n]$ and decreasing on $[\overline{x}_n, +\infty[$

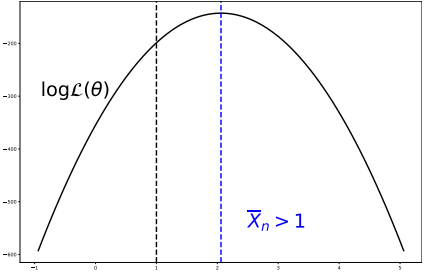


# Gaussian likelihood ratio test, ctd.

- say we want to test $H_0 : \theta \geq 1$ vs $H_1 : \theta < 1$
- thus

$$h(X) = \begin{cases} \mathcal{L}(\overline{x}_n)/\mathcal{L}(1) & \text{if } \overline{x}_n < 1 \\ \mathcal{L}(1)/\mathcal{L}(\overline{x}_n) & \text{if } \overline{x}_n \geq 1 \end{cases}$$



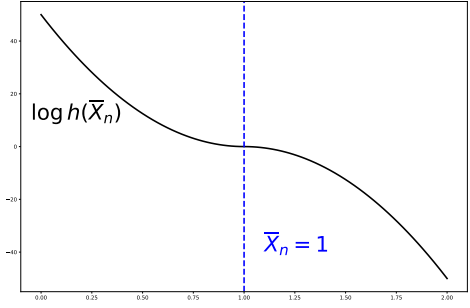
# Gaussian likelihood ratio test, ctd.

- some algebra shows that, for any $\theta_0$, $\theta_1$,

$$\frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_0)} = \exp\left(\frac{n}{2}(\theta_0^2 - \theta_1^2 + 2\overline{x}_n(\theta_1 - \theta_0))\right) .$$

- we deduce

$$h(X) = \exp\left(\frac{n}{2}(\overline{x}_n - 1)^2(2\mathbb{1}_{\overline{x}_n<1} - 1)\right)$$




- since $h$ is decreasing, $\phi(X) = \mathbb{1}_{\overline{x}_n<k_\alpha}$ **as before**

# 18. Properties of a test

# Unbiased test

**Definition:** a test $\phi$ is said to be *unbiased* ("non biaisé") if the power of the test is greater than the level of the test. Namely,

$$\forall \theta \in \Theta_1, \quad 1 - \underline{\beta}(\theta) > \alpha \,,$$

where $\alpha$ is the level of the test.

- **Intuition:** probability that the test rejects is always higher when the alternative is true than under the null
- **Example:** back to first example, one can show that $\underline{\beta}(\theta) = 1 - \Phi(\sqrt{n}(k_\alpha - \theta))$, thus

$$\forall \theta < 1, \quad 1 - \underline{\beta}(\theta) = \Phi(\sqrt{n}(k_\alpha - \theta)) > \alpha$$

by definition of $k_\alpha$

# Consistency

**Definition:** Let $(\phi_n)_{n\in\mathbb{N}}$ be a sequence of tests, each of level $\alpha$. We say that the sequence $(\phi_n)_n$ is *consistent* ("consistant") if the type II errors converge to zero. Namely,

$$\forall \theta \in \Theta_1, \quad \underline{\beta_n}(\theta) \underset{n\to+\infty}{\longrightarrow} 0\,.$$

- **Intuition:** fix the size, power goes to one
- **Same example:**

$$\forall \theta < 1, \quad \underline{\beta_n}(\theta) = 1 - \Phi(\sqrt{n}(k_\alpha - \theta)) \underset{n\to+\infty}{\longrightarrow} 0\,,$$

since $\theta < 1$.

# Uniformly more powerful

**Definition:** Let $\phi(X)$ and $\phi'(X)$ be two tests of level $\alpha$. We say that $\phi(X)$ is *uniformly more powerful* than $\phi'(X)$ ("uniformément plus puissant") if $1 - \underline{\beta} \geq 1 - \underline{\beta}'$. That is,

$$\forall \theta \in \Theta_1, \quad \mathbb{P}_\theta \left(\phi(X) = 1\right) \geq \mathbb{P}_\theta \left(\phi'(X) = 1\right) .$$

- **Intuition:** lower probability of false negative
- we say that $\phi(X)$ is UMP($\alpha$) if it is uniformly more powerful than any test of level $\alpha$ (for the same set of hypotheses)
- UMP($\alpha$) do not always exist!

# 19. $p$-values

# $p$-values

The previous approach is simple but binary (reject or not).

**Definition:** Suppose that we have a family of tests $(\phi_\alpha(X))_{\alpha\in[0,1]}$, each of level $\alpha$. We call *p-value* ("$p$ valeur") associated to observation $X$ the number

$$\hat{\alpha}(X) := \sup\{\alpha \in [0,1] \text{ s.t. } \phi_\alpha(X) = 0\}\,.$$

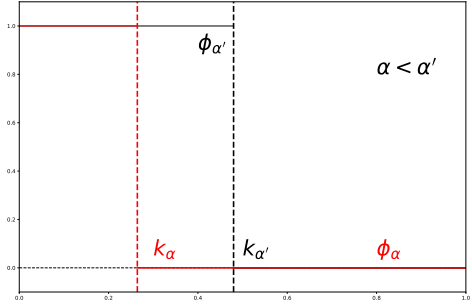
- **Intuition:** highest level such that we do not reject $H_0$
- gives credibility to rejecting $H_0$: if $\hat{\alpha}(X)$ is very small, we are very confident that we should reject $H_0$
- converse not true: we cannot conlude anything from a high $p$-value

# $p$-values, alternative definition

- if $(\phi_\alpha(X))$ is almost surely **increasing** in $\alpha$, then

$$\hat{\alpha}(X) := \inf\{\alpha \in [0,1] \text{ s.t. } \phi_\alpha(X) = 1\}\,.$$

- smallest level such that we reject the null
- **Example:** Gaussian likelihood ratio test. $\alpha < \alpha' \Rightarrow k_\alpha < k_{\alpha'}$




- limit case: $k_{\hat{\alpha}(x)} = \overline{x}_n$, which yields

$$\hat{\alpha}(x) = \Phi(\sqrt{n}(k_{\hat{\alpha}(x)} - 1)) = \Phi(\sqrt{n}(\overline{x}_n - 1))$$

# $p$-value, another definition

**Theorem:** Suppose that for any $\alpha$ we have constructed a test $\phi_\alpha$ of size $\alpha$ that can be written

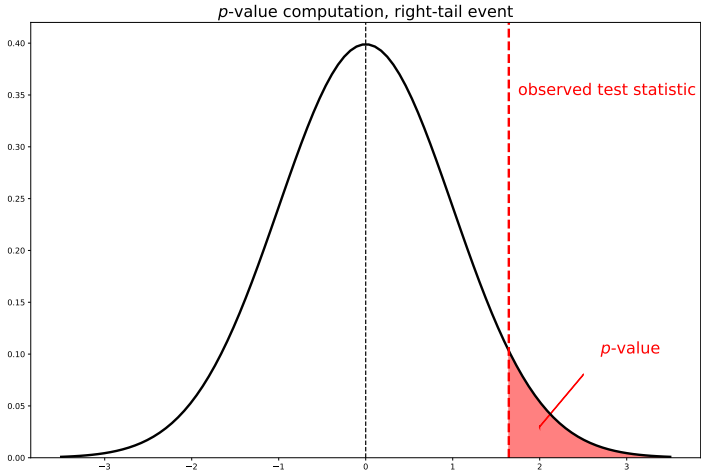
$$\phi_\alpha(X) = \mathbb{1}_{h(X) \geq k_\alpha} .$$

In this case,

$$\hat{\alpha}(x) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta \left( h(X) \geq h(x) \right) .$$

- **Intuition:** probability to observe more extreme test statistic under the null
- straightforward adaptation if $R_\alpha$ is not one-sided
- *size* is important here, cannot be replaced by *level*

# $p$-value, another definition, ctd.



# Disclaimer on $p$-values

- the p-value does not provide the probability that either hypothesis is correct
- $p$-values have become **standard practice** in applied research nowadays, and the criterion in statistical testing has *de facto* become "$p$ small enough"
- usually findings are considered *significant* if $p < 0.05$ (difficult to publish in biology if $p > 0.05$)
- recent phenomenon (Regina Nuzzo, 2014): $p$-**hacking**
- one possible way: collect data until the $p$-value drops under 0.05, or remove data arbitrarily and redo the test
- this may be one of the cause for **reproducibility crisis**

# Disclaimer on *p*-values, ctd.

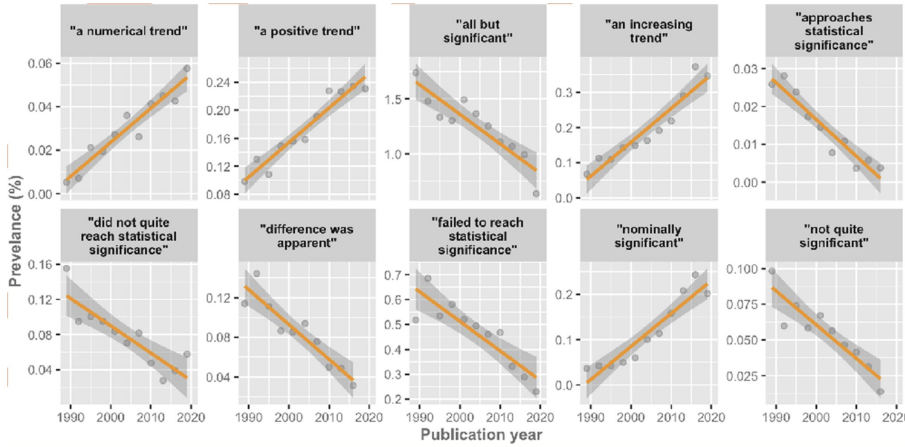



**Figure:** how papers discuss non-significant results (Otte et al, *Analysis of 567,758 randomized controlled trials published over 30 years reveals trends in phrases used to discuss results that do not reach statistical significance*, PLOS biology, 2022)

# Statistical testing methodology

Usually, we like to follow a recipe like this one:

1. set up **null hypothesis** and **alternative hypothesis**
2. make assumptions about the data: **independence, distribution,...**
3. decide which test to use, state the relevant **test statistic** $T$
4. derive the distribution of $T$ under the null
5. select a **significance level** $\alpha$
6. compute corresponding **critical region**
7. compute $t$, measured value of the test statistic
8. reject the null if $t$ belongs to the critical region
9. compute the associated $p$-value and report it along with the test results

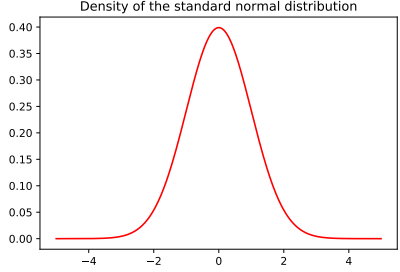
# 20. Gaussian vectors

# The Gaussian distribution

- also called the *normal* distribution (hence the notation $\mathcal{N}$)
- already encountered in the previous lectures: $\mathcal{N}\left(\mu, \sigma^2\right)$ has a density given by

$$f_{(\mu,\sigma^2)}(t) := \frac{1}{\sigma\sqrt{2\pi}} \cdot \mathrm{e}^{\frac{-(t-\mu)^2}{2\sigma^2}} .$$

- ubiquitous model for random errors, in particular because of the TCL




- generalization to higher dimensions: **what happens to $\sigma$?**

# Positive definite matrices

- analogous of positive numbers for matrices

**Definition:** We say that a $d \times d$ symmetric real matrix $M$ is *positive definite* if $x^\top Mx > 0$ for any $x \in \mathbb{R}^d \setminus \{0\}$. We say that it is positive *semi-definite* if $x^\top Mx \geq 0$ for any $x \in \mathbb{R}^d \setminus \{0\}$.

- **Example:** the identity matrix $\mathsf{I}_d$ is positive definite. Indeed,

$$(x, y) \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} = (x, y) \cdot \begin{pmatrix} x \\ y \end{pmatrix} = x^2 + y^2 \,.$$

- we define $S_d^{++}$ (resp. $S_d^{+}$) the set of real positive definite (resp. positive semi-definite) matrices
- **covariance matrices are always positive semi-definite**; positive definite unless one of the variables is a linear function of the others

# The *multivariate* Gaussian distribution

- generalization to the multivariate case ($\mathbb{R}^d$)

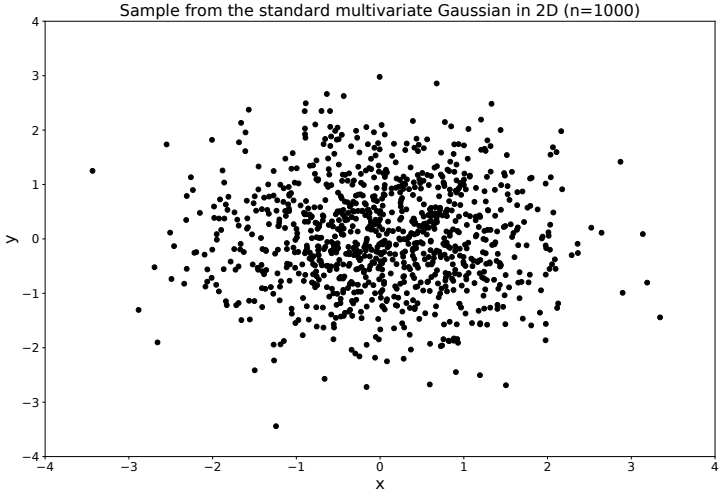
**Definition:** We say that the random vector $X \in \mathbb{R}^d$ follows the multivariate Gaussian distribution with parameters $\mu \in \mathbb{R}^d$ and $\Sigma \in S_d^{++}$ if $X$ has a density given by

$$f_{(\mu,\Sigma)}(x_1, \ldots, x_d) := \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \cdot \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right).$$

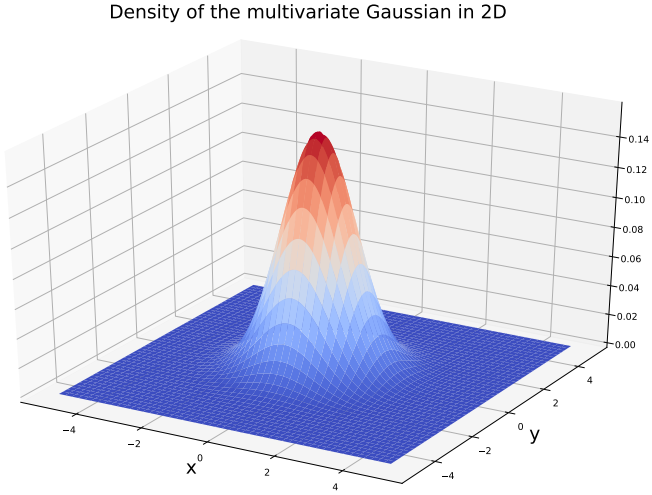
We also say that $X$ is a *Gaussian vector*.

- as expected:
  - $\forall 1 \leq j \leq d$, $\mathbb{E}[X_j] = \mu_j$,
  - and $\forall 1 \leq i, j \leq d$, $\mathrm{Cov}(X_i, X_j) = \Sigma_{ij}$.
- **Remark:** if $\Sigma$ is not invertible, then $X$ does not have a density.

# The multivariate Gaussian in pictures (I)



# The multivariate Gaussian in picture



# Fundamental property

- let $X \sim \mathcal{N}(\mu, \Sigma)$ with $\mu \in \mathbb{R}^d$ and $\Sigma \in S_d^{++}$
- take $A \in \mathbb{R}^{p \times d}$ and $b \in \mathbb{R}^p$ **deterministic**

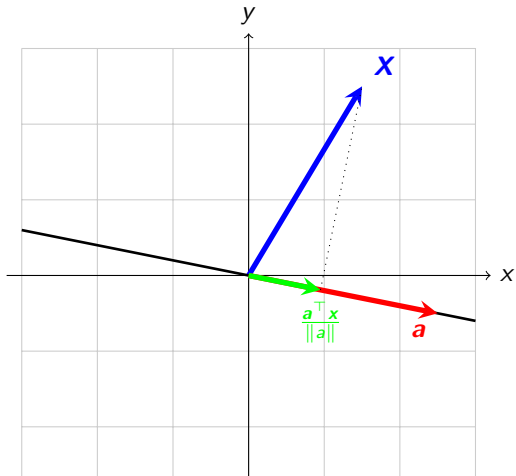
**Proposition:** If $A\Sigma A^\top$ is positive definite, then $AX + b$ is a Gaussian vector of dimension $p$. More precisely,

$$AX + b \sim \mathcal{N}\left(A\mu + b, A\Sigma A^\top\right).$$

- in general, $A\Sigma A^\top$ is positive semi-definite. It is positive definite if full rank ($\text{rank}\left(A\Sigma A^\top\right) = p$)
- **Example:** if $Z \sim \mathcal{N}(0, \mathrm{I}_d)$ and $A \in \mathcal{O}_d$ is an orthogonal matrix, then $AZ \sim \mathcal{N}(0, \mathrm{I}_d)$ (*invariance by rotation*).

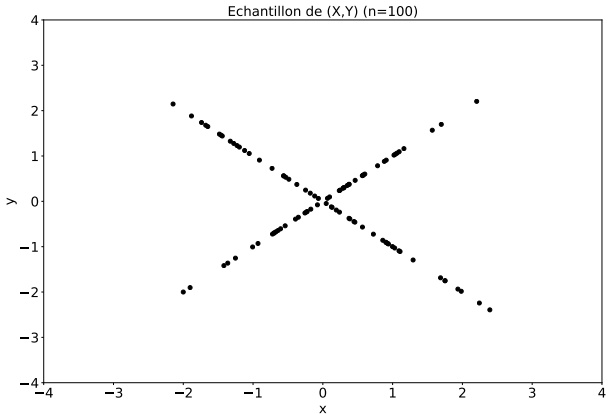
# Alternative definition

- **Alternative definition:** $X = (X_1, \ldots, X_d)^\top$ has multivariate Gaussian distribution if $a^\top X$ is (univariate) Gaussian $\forall a \in \mathbb{R}^d$.



# Gaussian vectors and independence (I)

- if $X$ and $Y$ are multivariate Gaussian and $X \sqcup Y$, then $(X, Y)$ is also a multivariate Gaussian
- this **fails** if the independence condition does not hold!
- counterexample: $X \sim \mathcal{N}(0, 1)$, $W = 2\mathcal{B}(1/2) - 1$ independent from $X$, and $Y = WX$.



# Gaussian vectors and independence (II)

- $(X_1, \dots, X_d)^\top$ Gaussian vector and $\mathrm{Cov}(X_i, X_j) = 0$ implies $X_i \sqcup X_j$.
- but **normally distributed and uncorrelated does not imply independent!**

*Proof:* same example, $X \sim \mathcal{N}(0,1)$, $W = 2\mathcal{B}(1/2) - 1$ independent from $X$, and $Y = WX$. In particular, $X$ **is not independent from** $Y$. But:

$$\begin{aligned}
\mathrm{Cov}(X, Y) &= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] && \text{(definition)} \\
&= \mathbb{E}\left[WX^2\right] - \mathbb{E}[X]^2\,\mathbb{E}[W] && \text{(definition of } Y\text{)} \\
&= \mathbb{E}[W]\,\mathbb{E}\left[X^2\right] - 0 && \text{(independence } + \mathbb{E}[W] = 0\text{)} \\
\mathrm{Cov}(X, Y) &= 0\,. && (\mathbb{E}[W] = 0)
\end{aligned}$$

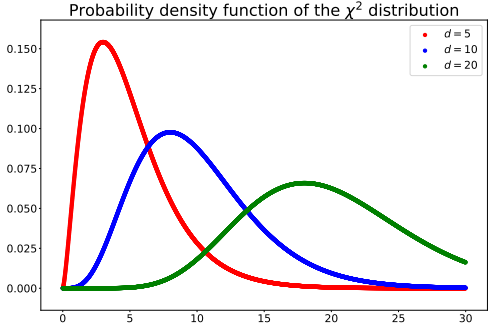
□

# The $\chi^2$ distribution

- let $p \geq 1$, $Z_1, \ldots, Z_p$ i.i.d. $\mathcal{N}(0,1)$, define

$$X := Z_1^2 + \cdots + Z_p^2 \,.$$

- **Notation:** $X \sim \chi^2_p$
- **Consequence:** $\mathbb{E}[X] = p$ and $\mathrm{Var}(X) = 2p$



# Cochran's theorem

**Theorem (Cochran, 1934):** Let $X \sim \mathcal{N}(0, I_d)$ be a Gaussian vector. Let $F$ be a linear subspace of $\mathbb{R}^d$ with dimension $p$. We denote by $P_F$ (resp. $P_{F^\perp}$) the orthogonal projection on $F$ (resp. $F^\perp$). Then

- $P_F X$ and $P_{F^\perp} X$ are independent,

$$P_F X \sim \mathcal{N}(0, P_F), \quad \text{and} \quad P_{F^\perp} X \sim \mathcal{N}(0, P_{F^\perp});$$

- $\|P_F X\|^2$ and $\|P_{F^\perp} X\|^2$ are independent,

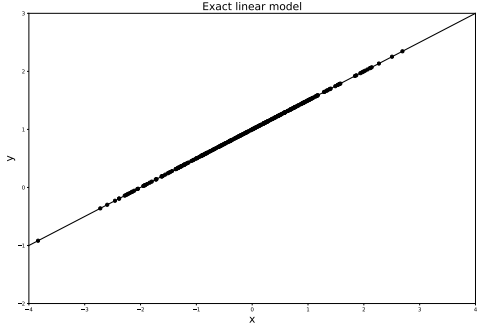
$$\|P_F X\|^2 \sim \chi^2_p, \quad \text{and} \quad \|P_{F^\perp} X\|^2 \sim \chi^2_{d-p}.$$

- equivalent to the Pythagorean theorem "in law"
- an operator $P$ is an orthogonal projection if and only if $P^\top = P = P^2$ ; it is an orthogonal projection on $\mathrm{Im}(P)$, which has dimension $\mathrm{Rank}(P) = \mathrm{Tr}(P)$. Moreover, $P_{F^\perp} = I - P_F$.
- **Example:** sample mean and sample variance (see TD)

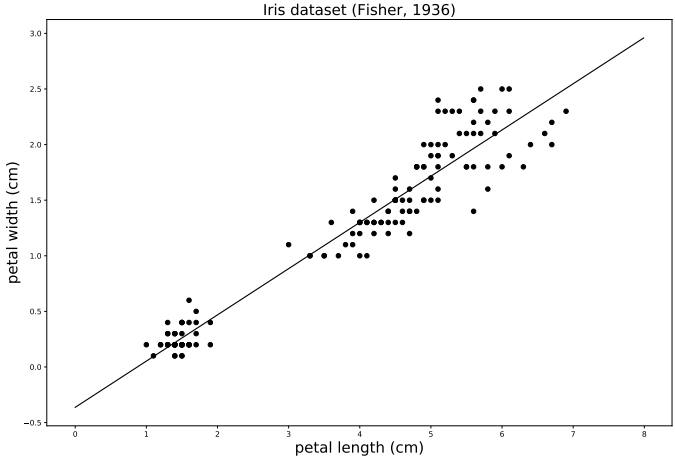
# 21. The linear model

# A statistician's dream

- setting in this chapter: **explicative** variables $X_1, \ldots, X_d \in \mathbb{R}$, **response** variable $Y$ (or *dependent* variable).
- observed $n$ times $(X_i = (X_{i1}, \ldots, X_{id})^\top)$
- assume that $Y = \Phi(X_1, \ldots, X_d)$, can we approximate $\Phi$ and **predict** $Y$ for a new value $x_1, \ldots, x_d$?
- statistician's dream: $\Phi$ is **linear** in the $X_i$s.



# The sad (but interesting!) reality




- **Problem:** even if linear relationship, there is **noise** in the data.

# The linear model

**Linear model:** there exist constant (but unknown) $\beta_0^\star, \beta_1^\star, \dots, \beta_d^\star$ and (random) noise $\varepsilon_i$ such that, for each observation $1 \leq i \leq n$,

$$Y_i = \beta_0^\star + \beta_1^\star X_{i1} + \cdots + \beta_d^\star X_{id} + \varepsilon_i \,.$$

- usually, we set $X_i = (1, X_{i1}, \dots, X_{id})^\top \in \mathbb{R}^{d+1}$ the observations
- **why** this strange 1 in the first position? *(phantom coordinate)*
- set $\beta^\star = (\beta_0^\star, \beta_1^\star, \dots, \beta_d^\star)^\top \in \mathbb{R}^{d+1}$. Then we can write, $\forall 1 \leq i \leq n$,

$$Y_i = X_i^\top \beta^\star + \varepsilon_i \,.$$

- it is **more compact**, but can be confusing. Beware!
- **Linear regression** = find $\beta^\star$

# Matrix notation

- we like the notation to be even more compact
- first set $Y = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top \in \mathbb{R}^n$.
- then regroup the observations (sometime called the *design* matrix)

$$X = \begin{bmatrix} X_1^\top \\ \vdots \\ X_n^\top \end{bmatrix} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1d} \\ 1 & X_{21} & \cdots & X_{2d} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \cdots & X_{nd} \end{pmatrix} \in \mathbb{R}^{n \times (d+1)}.$$

(each *line* is an observation).

- then

$$\boxed{Y = X\beta^\star + \varepsilon}$$

# The *Gaussian* linear model

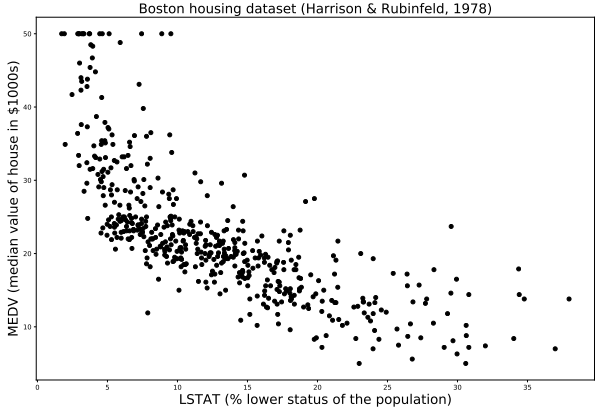
- in some cases, we can model the noise by a multivariate Gaussian:

$$\varepsilon \sim \mathcal{N}(\mu, \Sigma) \ .$$

- thus **Gaussian** linear model (often abridged linear model)
- further assume that the observations are independent, thus $\Sigma$ is **diagonal**
- if we assume the same variance for the $\varepsilon_i$, we say that there is *homoscedasticity* (then $\Sigma = \sigma^2 \, \mathrm{I}$), *heteroscedasticity* otherwise (then $\Sigma = \mathrm{diag}\left(\sigma_1^2, \ldots, \sigma_n^2\right)$)

**We do not need these assumption to do linear regression.**

# Nonlinear features (I)




- in certain cases, we would like to **transform** the features to get a linear fit (*e.g.*, LSTAT$\leftarrow \sqrt{\text{LSTAT}}$)

# Nonlinear features (II)

- turns out, **we can do this**!
- find "smart" transformations $\phi_1, \ldots, \phi_d$, where each

$$\phi_j : \mathbb{R} \to \mathbb{R}$$

  is a **known, deterministic** function.
- the model becomes

$$Y_i = \beta_0^\star + \beta_1^\star \phi_1(X_{i1}) + \cdots + \beta_d^\star \phi_d(X_{id}) + \varepsilon_i \,.$$

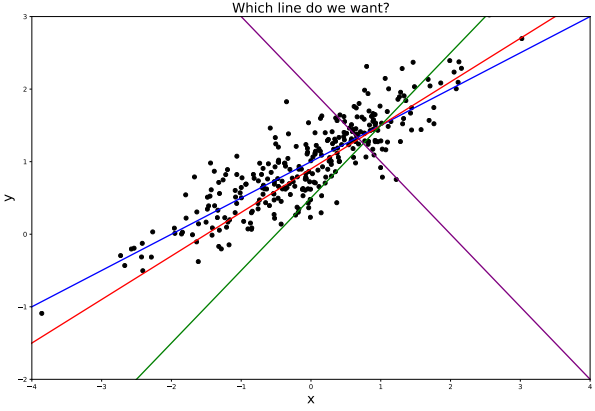
- this is still called linear model! Because **still linear in** $\beta$.
- **Example:** polynomial regression

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon \,.$$

# 22. Linear regression

# Linear regression

- **regression = estimation in presence of a dependent variable.**
- linear regression: find the "best" hyperplane in the linear model
- **Goal:** find a good estimator $\hat{\beta}$ for $\beta^\star$.



# Least-squares criterion

- **Idea:** define a **criterion** for each $\beta \in \mathbb{R}^{d+1}$
- **Intuition:** lower is better (criterion $= 0$ would be perfect fit).
- then find $\hat{\beta}$ that minimizes this criterion
- most common criterion: **least-squares**.

$$\text{crit}\,(\beta) := \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - X_i^\top \beta\right)^2 .$$

*(we sum the squares of the errors made in our linear model)*

- do not forget that this criterion depends on $X_1, \ldots, X_n, Y_1, \ldots, Y_n$.

**Important to understand:** we do not sum the $\varepsilon_i^2$, *a priori* $\beta \neq \beta^\star$!

# Ordinary Least Squares

- the estimator given by the ordinary least squares method is

$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^{d+1}}{\arg\min}\, \mathrm{crit}\,(\beta) = \underset{\beta \in \mathbb{R}^{d+1}}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left( Y_i - X_i^\top \beta \right)^2 \right\}.$$

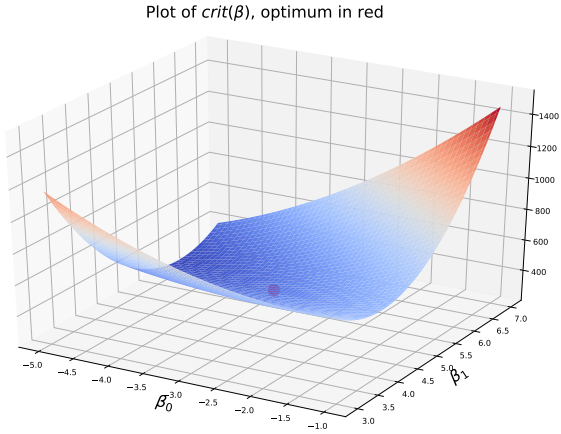
- **History:** Gauss (claimed 1795, published 1809), Legendre (1805). Gauss predicts the future position of Ceres, observed only 40 days by Piazzi in 1801.

- **Question:** how do we find $\hat{\beta}$?

**Idea:** crit is a convex, smooth function of $\beta \Rightarrow$ differentiate and set derivatives to 0.

# Least squares in picture

- now we have to work a bit because crit is a function of $d + 1$ variables:



# Calculus aparte

- **Reminder:** let $f : \mathbb{R}^N \to \mathbb{R}^M$, then the *gradient* of $f$ is defined as

$$\nabla f = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \dots & \frac{\partial f_M}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \dots & \frac{\partial f_M}{\partial x_2} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_1}{\partial x_N} & \frac{\partial f_2}{\partial x_N} & \dots & \frac{\partial f_M}{\partial x_N} \end{pmatrix} \in \mathbb{R}^{N \times M}$$

- **Example:** when $f$ is real-valued ($M = 1$), $\nabla f$ is a vector, thus a column

# Calculus aparte, ctd.

- let us consider first the function $f : x \mapsto Ax$, with $x \in \mathbb{R}^N$ and $A \in \mathbb{R}^{M\times N}$ a fixed matrix
- let $j \in \{1, \dots, M\}$, then we know that

$$(Ax)_j = A_{j,1}x_1 + A_{j,2}x_2 + \cdots + A_{j,N}x_N \,.$$

- let $i \in \{1, \dots, N\}$, then

$$\frac{\partial}{\partial x_i}(Ax)_j = A_{j,i} \,.$$

- we deduce from this computation that

$$\boxed{\forall A \in \mathbb{R}^{M\times N}, \qquad \nabla(Ax) = A^\top}$$

# Calculus aparte, ctd.

- more complicated: let $B \in \mathbb{R}^{N\times N}$ and define $f : x \mapsto x^\top Bx$
- set $1 \in \{1, \dots, N\}$, then

$$(Bx)_j = B_{j,1}x_1 + B_{j,2}x_2 + \cdots + B_{j,N}x_N\,.$$

- we deduce that

$$x^\top Bx = \sum_{j,k=1}^{n} B_{j,k}x_j x_k\,.$$

- therefore,

$$\frac{\partial}{\partial x_i}(x^\top Bx) = \sum_{j=1}^{n}(B_{i,j} + B_{j,i})x_j\,.$$

- in a concise form:

$$\boxed{\forall B \in \mathbb{R}^{N\times N}, \qquad \nabla(x^\top Bx) = (B + B^\top)x}$$

# Normal equation

- now we have all the tools required to obtain the normal equation
- recall that

$$\text{crit}\,(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - X_i^\top \beta\right)^2 .$$

- let us develop this expression:

$$\begin{aligned}(Y_i - X_i^\top\beta)^2 &= (Y_i - X_i^\top\beta)^\top(Y_i - X_i^\top\beta)\\ &= (Y_i - \beta^\top X_i)(Y_i - X_i^\top\beta) \qquad (Y_i \in \mathbb{R})\\ &= Y_i^2 - Y_i\beta^\top X_i - Y_i X_i^\top\beta + \beta^\top X_i X_i^\top\beta\\ (Y_i - X_i^\top\beta)^2 &= \beta^\top X_i X_i^\top\beta - 2Y_i X_i^\top\beta + Y_i^2 .\end{aligned}$$

- thus: $\text{crit}\,(\beta) = \frac{1}{n}\sum_{i=1}^{n}\beta^\top X_i X_i^\top\beta - \frac{2}{n}\sum_{i=1}^{n} Y_i X_i^\top\beta + \sum_{i=1}^{n} Y_i^2$

$$\Rightarrow \qquad \text{crit}\,(\beta) = \beta^\top\left(\tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i X_i^\top\right)\beta - 2\left(\tfrac{1}{n}\textstyle\sum_{i=1}^{n} Y_i X_i^\top\right)\beta + \text{cst.}$$

# Normal equation, ctd.

- we notice that:

$$X_i X_i^\top = \begin{pmatrix} 1 \\ X_{i,1} \\ \vdots \\ X_{i,d} \end{pmatrix} \begin{pmatrix} 1 & X_{i,1} & \cdots & X_{i,d} \end{pmatrix} = \begin{pmatrix} 1 & X_{i,1} & \cdots & X_{i,d} \\ X_{i,1} & X_{i,1}^2 & \cdots & X_{i,1}X_{i,d} \\ \vdots & \vdots & \ddots & \vdots \\ X_{i,d} & X_{i,1}X_{i,d} & \cdots & X_{i,d}^2 \end{pmatrix}$$

- thus

$$\left( \sum_{i=1}^n X_i X_i^\top \right)_{j,k} = \sum_{i=1}^n X_{i,j} X_{i,k} = (X^\top X)_{j,k} \,.$$

- in the same spirit,

$$\left( \sum_{i=1}^n Y_i X_i^\top \right)_{1,j} = \sum_{i=1}^n Y_{i,1} X_{i,j} = (Y^\top X)_{1,j} \,.$$

# Normal equation, ctd.

- therefore we can write crit in a more concise way:

$$\mathrm{crit}\,(\beta) = \frac{1}{n}\beta^\top X^\top X\beta - \frac{2}{n}Y^\top X\beta + \mathrm{cst}\,.$$

- it is a **convex function**, let us differentiate with respect to $\beta$ using the calculus results:

$$\begin{aligned}\nabla\mathrm{crit}\,(\beta) &= \nabla\left(\frac{1}{n}\beta^\top X^\top X\beta - \frac{2}{n}Y^\top X\beta\right)\\ &= \frac{2}{n}X^\top X\beta - \frac{2}{n}X^\top Y\,.\end{aligned}$$

- then we set this equation to zero:

$$\boxed{X^\top X\beta = X^\top Y\,.}$$

- this is known as the **normal equation**

# Inverting a matrix

- we have a linear system of equations to solve (in $\beta$):

$$X^\top X\beta = X^\top Y\,.$$

- if $X^\top X$ is invertible, no problem: we can define

$$\hat{\beta}_n = (X^\top X)^{-1} X^\top Y\,.$$

- in some cases ($n < d+1$), we have to work a bit more
- we define a *pseudo* inverse
- for this, we need the notion of singular value decomposition

# Singular Value Decomposition

**Definition-Theorem (singular value decomposition):** Let $A \in \mathbb{R}^{M \times N}$. Then there exist (i) $U \in \mathbb{R}^{M \times M}$ orthogonal, (ii) $V \in \mathbb{R}^{N \times N}$ orthogonal, and (iii) $\Sigma \in \mathbb{R}^{M \times N}$ diagonal with positive entries such that

$$A = U \Sigma V^\top .$$

The matrix $\Sigma$ is unique up to ordering of its diagonal elements.

- we call $\sigma_i := \Sigma_{ii}$ the **singular values** of $A$
- they are the square root of the eigenvalues of $A^\top A$
- only $\operatorname{rank}(A)$ of them are non-zero
- the columns of $U$ (resp. $V$) are the eigenvectors of $AA^\top$ (resp. $A^\top A$)

# Generalized inverse

- pseudo-inverse of a diagonal matrix:

$$\begin{pmatrix} d_1 & 0 & \cdots & 0 & 0 & \\ 0 & \ddots & \ddots & \vdots & \vdots & \cdots \\ \vdots & \ddots & \ddots & 0 & 0 & \cdots \\ 0 & \cdots & 0 & d_p & 0 & \end{pmatrix} \mapsto \begin{pmatrix} d_1^\dagger & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & d_p^\dagger \\ 0 & \cdots & 0 & 0 \\ & \vdots & \vdots & \end{pmatrix}$$

  where $x^\dagger = x^{-1}$ is $x \neq 0$ and 0 otherwise
- the **Moore-Penrose pseudo-inverse** of $M$ is then defined as

$$M^\dagger = V \Sigma^\dagger U^\top .$$

  We always have $M^\dagger M M^\dagger = M^\dagger$ and $M M^\dagger M = M$.
- **Example:** if $M$ is invertible, then $M^{-1} = M^\dagger$.
- from now on, we set $(X^\top X)^{-1} = (X^\top X)^\dagger$

# Conclusion on least squares

- now we can look at the solutions:

**Theorem (James, 1978):** The complete set of solutions of $Ax = b$ is given by

$$z = A^\dagger b + (\mathrm{I}_{d+1} - A^\dagger A)w\,,$$

for $w \in \mathbb{R}^{d+1}$.

- $A^\dagger A$ is an orthogonal projection since it is symmetric and $(A^\dagger A)^2 = (A^\dagger A A^\dagger)A = A^\dagger A$, so $I_{d+1} - A^\dagger A$ is the orthogonal projection on $\mathrm{Im}(A^\dagger A)^\perp$ and

$$\begin{aligned}\|A^\dagger b + (\mathrm{I}_{d+1} - A^\dagger A)w\|^2 &= \|(A^\dagger A)A^\dagger b + (\mathrm{I}_{d+1} - A^\dagger A)w\|^2 \\ &= \|A^\dagger b\|^2 + \|(\mathrm{I}_{d+1} - A^\dagger A)w\|^2\,.\end{aligned}$$

- taking the Moore-Penrose pseudo-inverse guarantees that **we take the solution with smallest Euclidean norm**.

# 23. Least squares as maximum likelihood estimation

# Maximum likelihood (I)

Let us make the following assumption (homoscedastic Gaussian linear model):

**Assumption:** the $\varepsilon_i$ are i.i.d. Gaussian random variable $\mathcal{N}\left(0, \sigma^2\right)$, *i.e.*, $\varepsilon \sim \mathcal{N}\left(0, \sigma^2 \mathsf{I}_n\right)$

- likelihood of the $i$-th observation (knowing $x_i$):

$$\mathcal{L}(y; \beta, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(\frac{-(y - x_i^\top \beta)^2}{2\sigma^2}\right).$$

- independence:

$$\begin{aligned}\mathcal{L}(y_1, \ldots, y_n; \beta, \sigma) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(\frac{-(y_i - x_i^\top \beta)^2}{2\sigma^2}\right) \\ &= \frac{1}{\sigma^n (2\pi)^{n/2}} \cdot \exp\left(\frac{-1}{2\sigma^2} \sum_{i=1}^{n} (y_i - x_i^\top \beta)^2\right).\end{aligned}$$

# Maximum likelihood (II)

- **Recall:** we want to maximize $\mathcal{L}(y_1, \ldots, y_n; \beta, \sigma)$ for parameters $(\beta, \sigma) \in \mathbb{R}^{d+1} \times \mathbb{R}_+$
- fix $\sigma > 0$, find $\beta$ that maximizes

$$\exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^\top \beta)^2\right).$$

- equivalent to **minimizing**

$$\sum_{i=1}^{n}(y_i - x_i^\top \beta)^2.$$

- $\Rightarrow$ we recover the same solution as before for $\beta$!
- namely,

$$\hat{\beta}^{\mathsf{MLE}} = (X^\top X)^{-1} X^\top Y.$$

# Maximum likelihood (III)

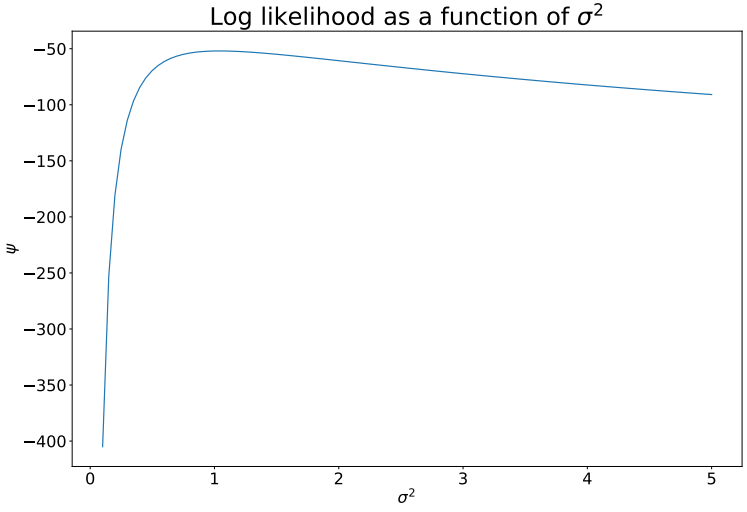
- ***Quid* of the variance?**
- fix $\hat{\beta}$ and study the function

$$\psi : \sigma^2 \mapsto \log \mathcal{L}(y_1, \ldots, y_n; \hat{\beta}, \sigma^2)\,.$$

- according to previous computations:

$$\begin{aligned}
\psi(\sigma^2) &= \log\left(\frac{1}{(\sigma^2)^{n/2}(2\pi)^{n/2}} \cdot \exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^\top \beta)^2\right)\right) \\
&= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^\top \hat{\beta})^2 - \frac{n}{2}\log 2\pi \\
&= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^\top \hat{\beta})^2 + \text{cst}\,.
\end{aligned}$$

# Maximum likelihood (IV)



# Maximum likelihood (V)

- smooth function, let us differentiate:

$$\psi'(\sigma^2) = \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - x_i^\top \hat{\beta})^2$$

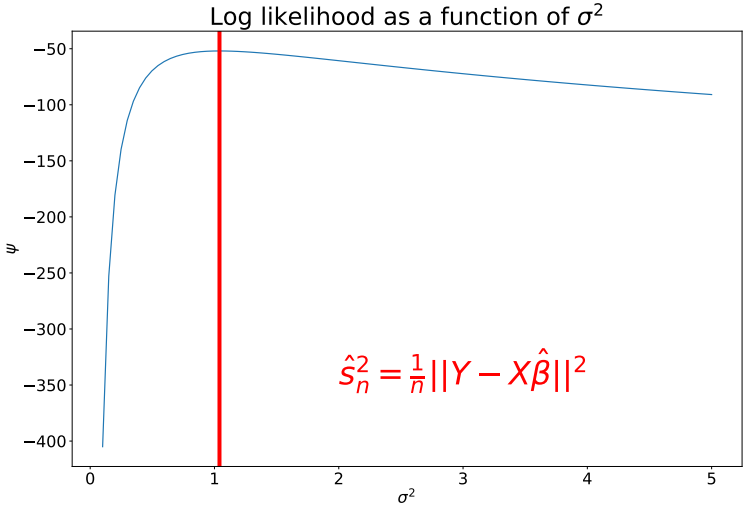
- point of maximum attained if $\psi'(\sigma^2) = 0$, that is

$$\frac{-n}{\sigma^2} + \frac{1}{\sigma^4}\sum_{i=1}^{n}(y_i - x_i^\top \hat{\beta})^2 = 0$$

- we deduce

$$\hat{s}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^\top \hat{\beta})^2 .$$

# Maximum likelihood (VI)



# Maximum likelihood (VII)

**Theorem:** Let us assume the Gaussian linear model

$$Y = X\beta + \varepsilon \quad \text{where} \quad \varepsilon \sim \mathcal{N}\left(0, \sigma^2 \mathsf{I}_n\right) ,$$

and assume that $X$ has full rank $(d+1)$. Then the maximum likelihood estimator of $\theta = (\beta, \sigma^2)$ is given by $\hat{\theta}_n = (\hat{\beta}_n, \hat{s}_n^2)$, with

$$\hat{\beta}_n = (X^\top X)^{-1} X^\top Y \quad \text{and} \quad \hat{s}_n^2 = \frac{1}{n} \left\| Y - X\hat{\beta}_n \right\|^2 .$$

Moreover, $\hat{\beta}_n$ and $\hat{s}_n^2$ are *independent*, and we have

$$\hat{\beta}_n \sim \mathcal{N}\left(\beta, \sigma^2 (X^\top X)^{-1}\right) \quad \text{and} \quad \frac{n}{\sigma^2} \hat{s}_n^2 \sim \chi^2(n-d-1) .$$

# Maximum likelihood (VIII)

- in particular, $\hat{\beta}_n$ is an **unbiased** estimator of $\beta$
- recall that a chi-squared distribution with $p$ degrees of freedom has expectation $p$
- thus

$$\mathbb{E}\left[\frac{n}{\sigma^2}\hat{\mathsf{s}}_n^2\right] = n - d - 1\,,$$

  and we deduce the following **unbiased** estimator of $\sigma^2$:

$$\hat{\sigma}_n^2 = \frac{1}{n-d-1}\left\|Y - X\hat{\beta}_n\right\|^2 .$$

- **Conclusion:** least squares and maximum likelihood yield the same estimator for $\beta$

# Proof of the theorem (I)

- set $\Pi = X(X^\top X)^{-1}X^\top \in \mathbb{R}^{n\times n}$
- $\Pi$ is an **orthogonal projection:**

$$\Pi^2 = \Pi \quad \text{and} \quad \Pi^\top = \Pi\,.$$

- under our assumptions, $Y - X\beta = \varepsilon \sim \mathcal{N}\left(0, \sigma^2 \mathsf{I}_n\right) \Rightarrow$ we may apply Cochran's theorem to $\frac{\varepsilon}{\sigma}$
- first consequence:

$$\Pi\varepsilon = \Pi(Y - X\beta) = X\hat{\beta} - X\beta \sim \mathcal{N}\left(0, \sigma^2\Pi\right)\,.$$

- $Y \sim \mathcal{N}\left(X\beta, \sigma^2 \mathsf{I}_n\right)$, and $X$ full rank $\Rightarrow X^\top X$ invertible, therefore

$$\begin{aligned}\hat{\beta} = (X^\top X)^{-1}X^\top Y &\sim \mathcal{N}\left((X^\top X)^{-1}X^\top X\beta, \sigma^2(X^\top X)^{-1}X^\top((X^\top X)^{-1}X^\top)^\top\right)\\ &= \mathcal{N}\left(\beta, \sigma^2(X^\top X)^{-1}\right)\,.\end{aligned}$$

# Proof of the theorem (II)

- second consequence:

$$(I - \Pi)\varepsilon = Y - X\hat{\beta} \sim \mathcal{N}\left(0, \sigma^2(I - \Pi)\right)$$

  is an **independent** random variable from $\hat{\beta}$
- moreover, we know the law of the squared norm:

$$\frac{1}{\sigma^2}\left\|Y - X\hat{\beta}\right\|^2 \sim \chi^2_p$$

- with our notation,

$$\frac{1}{\sigma^2}\left\|Y - X\hat{\beta}\right\|^2 = \frac{n}{\sigma^2}\hat{s}^2_n\,.$$

- what is $p$? In fact, $\Pi$ projection on $\mathrm{Im}\,(X)$, thus $p = n - d - 1$ as expected. $\square$

# Gauss-Markov theorem

**Theorem (Gauss-Markov, 1821-1912):** In the linear model, assume that the errors satisfy

1. $\mathbb{E}[\varepsilon_i] = 0$ *(centered)*;
2. $\mathrm{Var}(\varepsilon_i) = \sigma^2$ *(homoscedastic)*;
3. $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ *(uncorrelated)*.

Let $\tilde{\beta}_n$ be another unbiased estimator of $\beta$ such that $\tilde{\beta} = CY$. Then

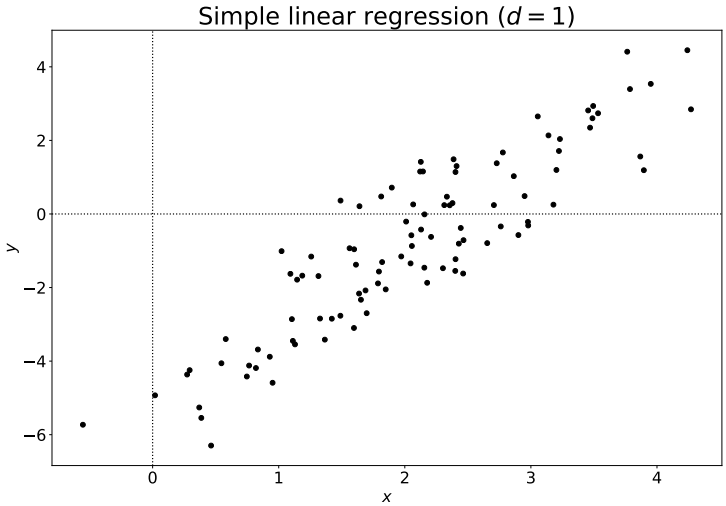
$$\mathrm{Cov}(\hat{\beta}_n) \lesssim \mathrm{Cov}(\tilde{\beta}_n)$$

- **Intuition:** provided it exists, the estimator given by the ordinary least squares is the **best linear unbiased estimator**
- **Note:** Gauss proved the result with Gaussian assumption on the data, Markov got rid of it

# 24. Simple linear regression

# Simple linear regression (I)

The general case can be a bit mysterious, let us explicit everything in a simple case: $d = 1$



# Simple linear regression (II)

- we observe $x_1, \ldots, x_n \in \mathbb{R}$, thus

$$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \in \mathbb{R}^{n \times 2} .$$

- straightforward computation gives

$$X^\top X = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} .$$

- invertible if the $x_i$s **are not constant**:

$$\det\left(X^\top X\right) = n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2 = n^2 \mathrm{Var}\left(x\right) > 0 .$$

# Simple linear regression (II)

- let us introduce the following notation:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i\,,$$

$$\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i\,, \quad \text{and} \quad \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

- with these notation, $\mathrm{Var}\,(x) = \overline{x^2} - \overline{x}^2$ and $\mathrm{Cov}\,(x, y) = \overline{xy} - \overline{x} \cdot \overline{y}$
- *Nota bene:*

$$\boxed{\overline{xy} \neq \overline{x} \cdot \overline{y} \quad \text{and} \quad \overline{x^2} \neq \overline{x}^2}$$

# Simple linear regression (III)

- with these notation:

$$X^\top X = n \begin{pmatrix} 1 & \overline{x} \\ \overline{x} & \overline{x^2} \end{pmatrix} \quad \text{and} \quad \det\left(X^\top X\right) = n^2(\overline{x^2} - \overline{x}^2)\,.$$

- we deduce that

$$(X^\top X)^{-1} = \frac{1}{n(\overline{x^2} - \overline{x}^2)} \begin{pmatrix} \overline{x^2} & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix}.$$

- on the other side,

$$X^\top Y = n \begin{pmatrix} \overline{y} \\ \overline{xy} \end{pmatrix}$$

- thus

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y = \frac{1}{\overline{x^2} - \overline{x}^2} \begin{pmatrix} \overline{x^2} \cdot \overline{y} - \overline{x} \cdot \overline{xy} \\ \overline{xy} - \overline{x} \cdot \overline{y} \end{pmatrix}$$

# Simple linear regression (III)

- we first write

$$\hat{\beta}_1 = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2}.$$

- then we notice that

$$\begin{aligned}
\hat{\beta}_0 &= \frac{\overline{x^2} \cdot \overline{y} - \overline{x} \cdot \overline{xy}}{\overline{x^2} - \overline{x}^2} \\
&= \frac{(\overline{x^2} - \overline{x}^2) \cdot \overline{y} + \overline{x}^2 \cdot \overline{y} - \overline{x} \cdot \overline{xy}}{\overline{x^2} - \overline{x}^2} \\
&= \overline{y} + \overline{x} \cdot \frac{\overline{x} \cdot \overline{y} - \overline{xy}}{\overline{x^2} - \overline{x}^2} \\
\hat{\beta}_0 &= \overline{y} - \hat{\beta}_1 \overline{x}
\end{aligned}$$

- we call $\hat{\beta}_0$ the **intercept** and $\hat{\beta}_1$ the **slope**

# Pearson correlation coefficient

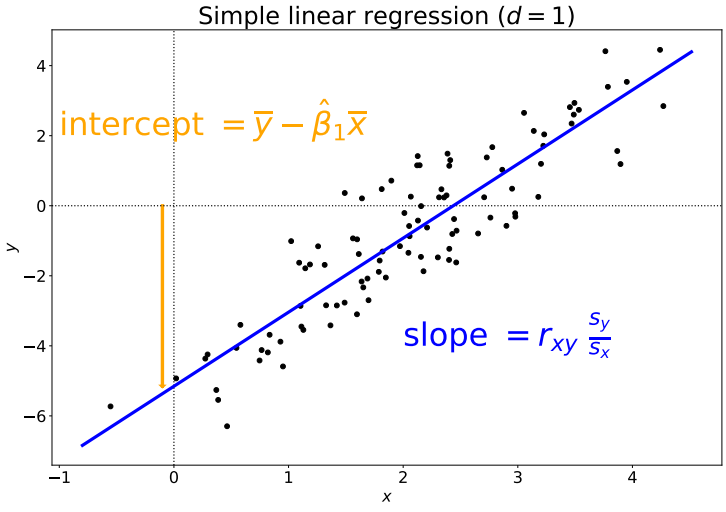
- define the **empirical variances** $s_x^2 = \overline{x^2} - \overline{x}^2$ and $s_y^2 = \overline{y^2} - \overline{y}^2$
- **Definition:** we set $r_{xy}$ the **Pearson correlation coefficient** between $x$ and $y$:

$$r_{xy} := \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{s_x s_y} .$$

- $r_{xy}$ is **invariant** by rescaling
- the **slope** can be written

$$\hat{\beta}_1 = r_{xy} \frac{s_y}{s_x} .$$

# Simple linear regression, example



# Coefficient of determination (I)

How good is the fit?

**Definition:** recall the Pearson correlation coefficient

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}\,.$$

We define the **coefficient of determination** as

$$R^2 = r_{y\hat{y}}^2 \in [0, 1]\,.$$

- **Intuition:** proportion of the variance in the dependent variable that is explained from the independent variables
- $R^2 = 1$ if perfect linear relation, $R^2 = 0$ if uncorrelated

$\Rightarrow$ *higher is better*

# Coefficient of determination (II)

Define the following quantities:

- the **total sum of squares**

$$SS_{\text{tot}} = \sum_{i=1}^{n} (y_i - \overline{y})^2 \,,$$

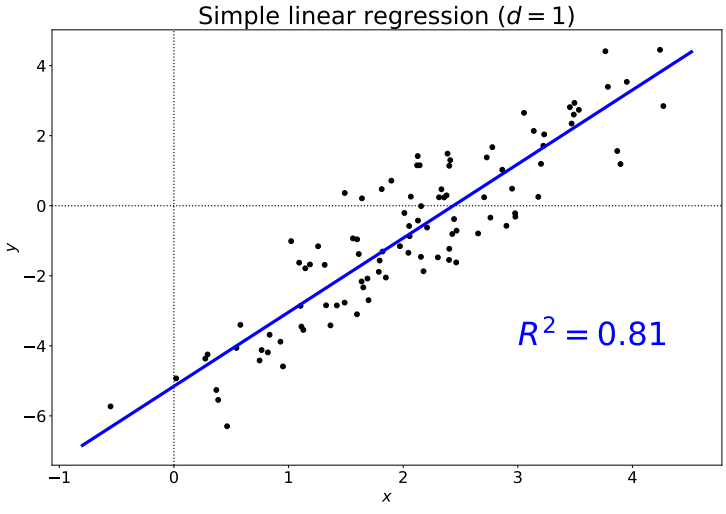
- the **sum of squares of residuals**

$$SS_{\text{res}} = \sum_{i=1}^{n} (y_i - X_i^\top \hat{\beta})^2 \,.$$

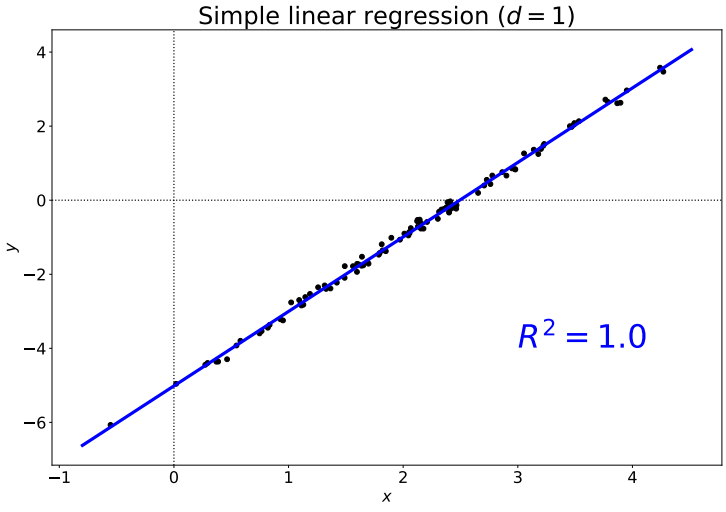
Then the following expression (more general) holds:

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} \,.$$

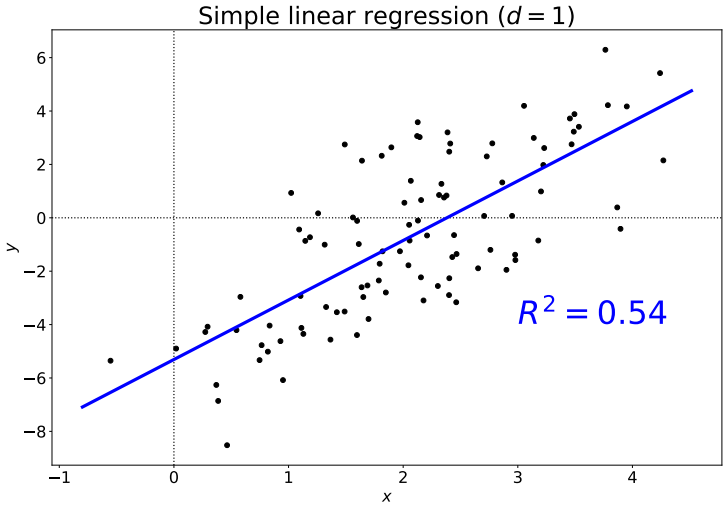
# Coefficient of determination (III)



# Coefficient of determination (IV)



# Coefficient of determination (V)



# Back to the Student $t$-distribution

- more general definition:

**Definition:** let $Z \sim \mathcal{N}(0,1)$ be a standard Gaussian and $V \sim \chi^2_d$ an independent chi-squared random variable. Then

$$T = \frac{Z}{\sqrt{V/d}}$$

follows the Student $t$-distribution with $d$ degrees of freedom.

# Testing for significance

- in the Gaussian linear model ($d = 1$), suppose that we want to test

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0 \,.$$

- **Intuition:** test whether the independent variable helps to predict the value of the dependent variable
- test statistic

$$T_n = \frac{\hat{\beta}_{1,n}}{\hat{\sigma}_n} \sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

- Student law with $n - 2$ degrees of freedom under the null
- leads to the test

$$\phi(Y) = \mathbb{1}_{|T_n| > t_{n-2,1-\alpha/2}} \,,$$

where $t_{n-2,1-\alpha/2}$ quantile of order $1 - \alpha/2$ of the $\mathcal{T}_{n-2}$ distribution

# Confidence interval for a new observation

- natural estimator for a **new observation** $x$:

$$\hat{y} = \hat{\beta}_{0,n} + \hat{\beta}_{1,n} x \,.$$

- notice that $\hat{y} = A\hat{\beta}_n$ with $A = \begin{pmatrix} 1 & x \end{pmatrix}$
- recall that

$$\hat{\beta}_n \sim \mathcal{N}\left(\beta, \sigma^2 (X^\top X)^{-1}\right) .$$

- we deduce that

$$\hat{y} \sim \mathcal{N}\left(A\beta, \sigma^2 A (X^\top X)^{-1} A^\top\right) .$$

- a quick computation yields

$$\hat{y} \sim \mathcal{N}\left(\beta_0 + \beta_1 x, \frac{\sigma^2}{n} \cdot \frac{\overline{x^2} - 2x \cdot \overline{x} + x^2}{\overline{x^2} - \overline{x}^2}\right)$$

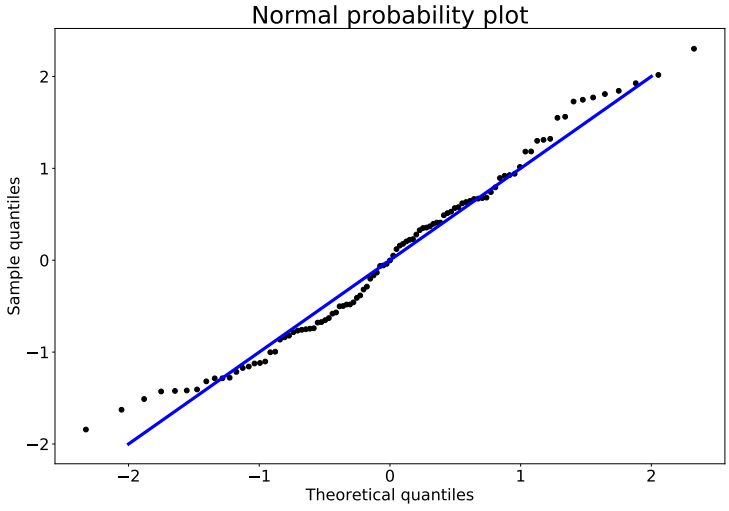
# Residuals analysis

**Definition:** we call residuals *("résidus")* the numbers

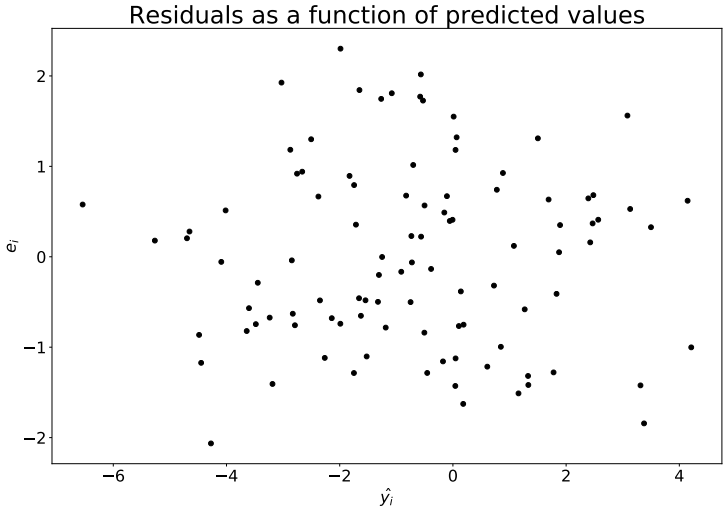
$$e_i = y_i - \hat{y}_i \, .$$

- **Intuition:** error in the prediction on the data
- approximation of the $\varepsilon_i$, to which **we do not have access!**
- we can check our assumptions *("analyse des résidus")*:
  - normality $\Rightarrow$ the $e_i$ should be approximately Gaussian
  - homoscedasticity $\Rightarrow$ equal variance
  - independence $\Rightarrow$ structure of the $e_i$

# Testing residuals for normality



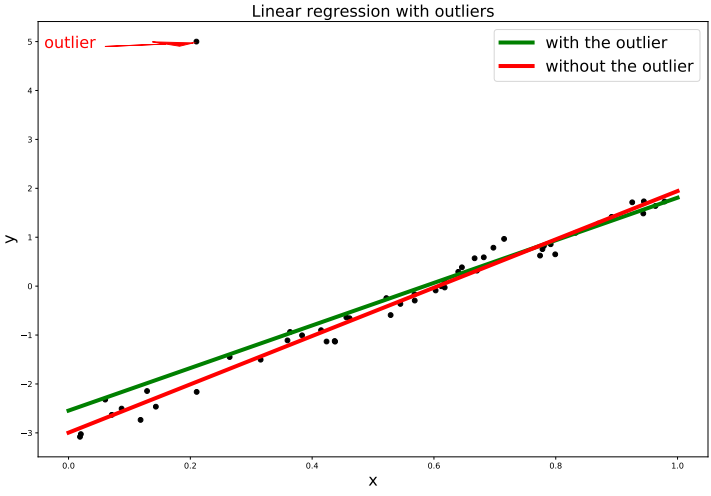
# Testing residuals for homoscedasticity



# 25. Weighted least squares

# Outliers

- problem with OLS: **sensitive to outliers**



# Weighted least squares (I)

- one possible solution: put **more weight** on certain observations
- define $W \in \mathbb{R}^{n \times n}$ a **weight matrix**, often *diagonal*
- weighted least squares solves

$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^{d+1}}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} W_{ii} (Y_i - X_i^\top \beta)^2 \right\}.$$

- intuitively, weights should be inversely proportional to the uncertainty (no noise $\Rightarrow$ infinite weight).

- the normal equations in this case are

$$\boxed{X^\top W X \beta = X^\top W Y .}$$

# Weighted least squares (II)

- gives rise to the estimator

$$\hat{\beta}^{\mathrm{WLS}} = (X^\top W X)^{-1} X^\top W Y \,.$$

- when the errors are **uncorrelated**, $\hat{\beta}^{\mathrm{WLS}}$ is the **best linear unbiased estimator** if $W_{ii} = \frac{1}{\sigma_i}$.

# 26. Estimating the parameters of a discrete distribution

# Multinomial distribution

**Definition:** let $n$ and $k$ be finite integers and $p \in [0,1]^k$ such that

$$p_1 + \cdots + p_k = 1\,.$$

Let $A_1, \ldots, A_k$ a set of **distinct** outcomes. For $1 \leq i \leq n$, we define the independent random variables $E_i$, each taking the value $A_j$ with probability $p_j$. Then we say that the vector $N = (N_1, \ldots, N_k)$ where

$$N_j = \sum_{i=1}^{n} \mathbb{1}_{E_i = A_j}$$

has the *multinomial distribution ("loi multinomiale")* with parameters $n, k, p$.

- **Intuition:** $N_j$ counts the number of times the experiments yielded $A_j$
- **Important:** while the experiments are independent, the $N_i$ are not (in particular, $\sum_i N_i = n$)

# Example: the Bernoulli distribution

- let $p \in (0,1)$
- set $n = 1$, $k = 2$, and
$$p_1 = p \quad \text{and} \quad p_2 = 1 - p\,.$$
- one experiment $E_1$ with two outcomes $\{A_1, A_2\}$
- then $N_1$ has the **Bernoulli** distribution $\mathcal{B}(p)$:
$$\begin{aligned} \mathbb{P}(N_1 = 0) &= \mathbb{P}(E_1 = A_2) \\ &= p_2 \\ \mathbb{P}(N_1 = 0) &= 1 - p \end{aligned}$$
and $\mathbb{P}(N_1 = 1) = 1 - \mathbb{P}(N_1 = 0) = p$.
- same reasoning for $N_2 \sim \mathcal{B}(1-p)$

# Example: the binomial distribution

- let $p \in (0,1)$
- set $n \geq 1$, $k = 2$,

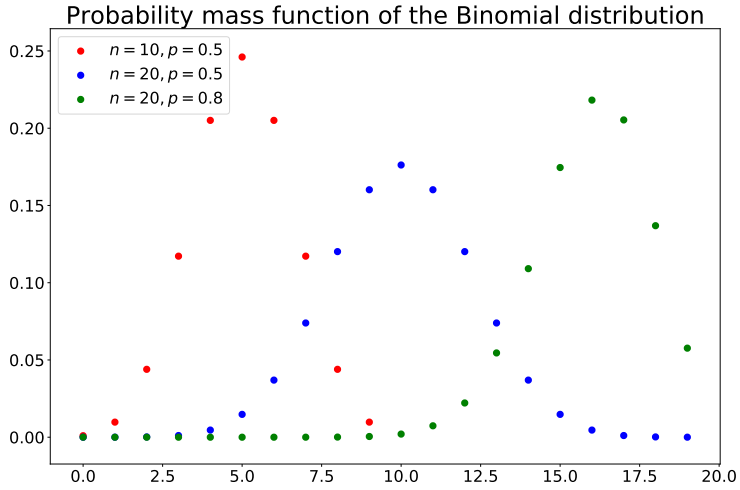
$$p_1 = p \quad \text{and} \quad p_2 = 1 - p\,.$$

- $n$ independent experiments $E_j$ with two outcomes $\{A_1, A_2\}$
- then $N_1$ has the **binomial distribution** $\mathcal{B}(n,p)$: for any $m \in \{0,\dots,n\}$,

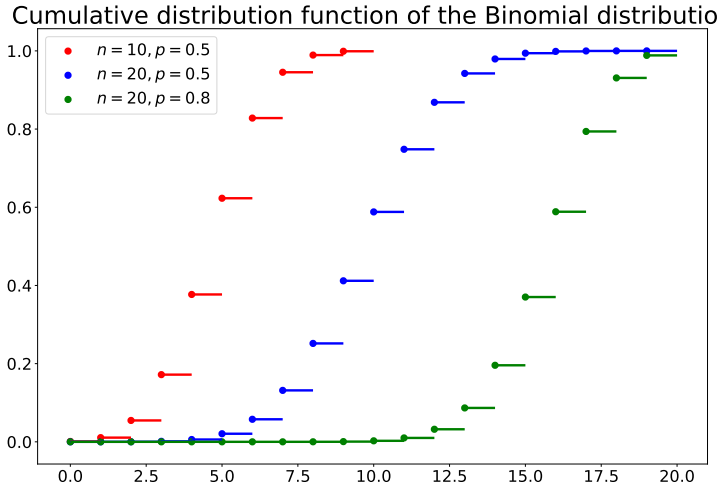
$$\begin{aligned}\mathbb{P}(N_1 = m) &= \mathbb{P}(\exists S \subseteq \{1,\dots,n\} \text{ s.t. } |S| = m \text{ and} \\ &\qquad \forall i \in S, E_i = A_1, \forall i \notin S, E_i = A_2) \\ &= \binom{n}{m} p_1^m p_2^{n-m}\end{aligned}$$

$$\mathbb{P}(N_1 = m) = \binom{n}{m} p^m (1-p)^{n-m}$$

# Binomial distribution in pictures (I)



# Binomial distribution in pictures (II)



# Properties of the multinomial distribution (I)

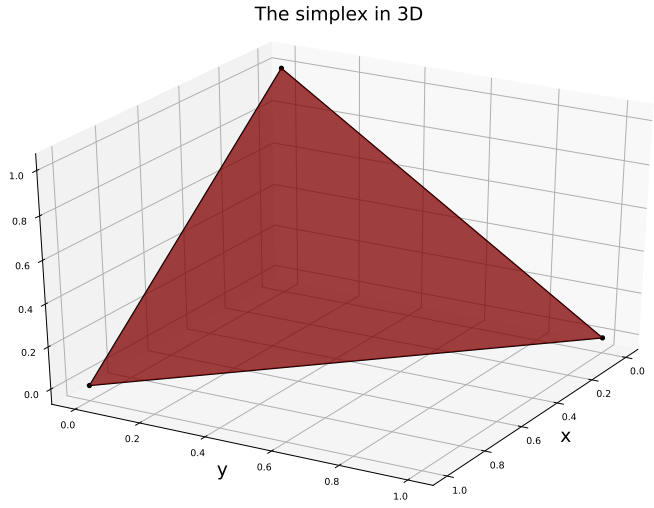
- set $n, k \geq 1$ integers and $p = (p_1, \ldots, p_k)^\top$ such that $p_j \geq 0$ and $\sum_j p_j = 1$
- let us consider $N \sim \mathcal{M}(n, p)$

**Property:** $N/n$ takes values almost surely in the **simplex**:

$$\Delta_k = \left\{ t = (t_1, \ldots, t_k) \in \mathbb{R}_+^k \text{ s.t. } \sum_j t_j = 1 \right\}.$$

*Proof:* $\sum_j N_j = \sum_j \sum_i \mathbb{1}_{E_i = A_j} = n$ almost surely. $\square$

# The simplex


The simplex in 3D
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
y
0.0
0.2
0.4
0.6
0.8
1.0
x

# Properties of the multinomial distribution (II)

**Property:** the probability mass function of $X \sim \mathcal{M}(n, p)$ is given by

$$\mathbb{P}(N_1 = n_1, \dots, N_k = n_k) = \frac{n!}{n_1! \cdots n_k!} p_1^{n_1} \cdots p_k^{n_k}.$$

*Proof:* same idea than for the binomial. □

**Property:** it holds that $\mathbb{E}[N_j] = np_j$, $\mathrm{Var}(N_j) = np_j(1-p_j)$, and $\mathrm{Cov}(N_j, N_{j'}) = -np_j p_{j'}$.

*Proof:* $1 = (p_1 + \cdots + p_k)^n$ then $p_i \frac{\partial}{\partial p_i}$ trick □

# Estimating the parameters of a discrete distribution (I)

- consider a **discrete set** $\mathcal{X} = \{x_1, \dots, x_k\}$
- we are given an i.i.d. sample $X = (X_1, \dots, X_n)$ with common law given by $p = (p_1, \dots, p_k)$:

$$\forall i \in \{1, \dots, n\}, \ \forall j \in \{1, \dots, k\}, \quad \mathbb{P}(X_i = x_j) = p_j \, .$$

- **Example:** fair dice ($x_j = j, p_j = 1/6$)
- **How can we estimate $p$?**
- consider the random variables $\mathbb{1}_{X_i = x_j}$. They are i.i.d. Bernoulli with mean $p_j$
- we can use the **method of moments**
- since $\mathbb{E}\left[\mathbb{1}_{X_i = x_j}\right] = p_j$, we set

$$\hat{p}_j = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i = x_j} \, .$$

# Estimating the parameters of a discrete distribution (II)

- for any $1 \leq j \leq k$, define

$$N_j = \sum_{i=1}^{n} \mathbb{1}_{X_i = x_j} .$$

- we can rewrite $\hat{p}_j = N_j / n$
- $N = (N_1, \ldots, N_k)^\top$ has the **multinomial distribution**
- in particular,

$$\mathbb{E}\left[\frac{N_j}{n}\right] = p_j ,$$

  and the $\hat{p}_j$ are unbiased, consistent estimators of the $p_j$
- we can also construct **joint confidence intervals** since we have the joint law

# 27. Chi-squared test for fit of a distribution

# Chi-squared test for fit of a distribution

- **Statistical model:** $X = (X_1, \dots, X_n)$ i.i.d. with support $\mathcal{X} = (x_1, \dots, x_k)$ with $p = (p_1, \dots, p_k)$ such that

$$\forall 1 \leq i \leq n, \ \ \forall 1 \leq j \leq k, \quad \mathbb{P}(X_i = x_j) = p_j$$

- **General idea:** we are given a reference law

$$p^0 = (p_1^0, \dots, p_k^0),$$

  with **full support** on $\mathcal{X}$ (that is, $p_j^0 > 0$ for all $j$)
- we want to **test** whether $p = p^0$ or not *("test d'ajustement à une loi donnée")*
- formally:

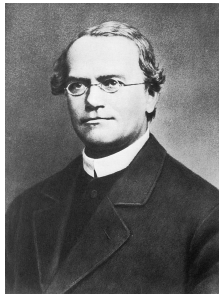
$$H_0 : \quad \forall j \in \{1, \dots, k\}, \quad p_j = p_j^0$$

vs

$$H_1 : \quad \exists j \in \{1, \dots, k\} \text{ such that } p_j \neq p_j^0 .$$

# Example: testing the laws of heredity

- **Example:** genetics (Gregor Johann Mendel, 1822–1884)



- suppose that the color of tulips is red (R), blue (B), or purple
- we have a theory: the color is determined a single gene that has two alleles: R and B
- thus purple parents (*RB*) theoretically yield 3 possibilities (RR, RB, and BB) with probability 0.25, 0.5, and 0.25
- **Can we test our hypothesis?**

# Chi-squared distance

- **Idea:** under $H_0$, $\hat{p}$ is close to $p^0$
- thus we can take as test statistic a **distance** between $\hat{p}$ and $p^0$

**Definition:** we define the *chi-squared distance* (*"distance du chi deux"*) between $\hat{p}_n$ and $p^0$ by

$$D_n^2(\hat{p}_n, p^0) = n \sum_{j=1}^{k} \frac{(\hat{p}_j - p_j^0)^2}{p_j^0} = \sum_{j=1}^{k} \frac{(N_j - np_j^0)^2}{np_j^0}.$$

- **Intuition:** weighted, re-normalized Euclidean distance between $\hat{p}_n$ and $p^0$
- often rewritten as a function of the integer counts

# Chi-squared distance under the null (I)

**Theorem:** the chi-squared distance has the following asymptotic behaviors:

- under $H_0$: $p = p^0$, we have
$$D_n^2(\hat{p}_n, p^0) \xrightarrow{\mathcal{L}} \chi^2_{k-1};$$
- on the other side, under $H_1$: $p \neq p^0$, we have
$$D_n^2(\hat{p}_n, p^0) \longrightarrow +\infty$$
almost surely.

- **Intuition:** we know the law of the test statistic under $H_0$ **and** $H_1$ when $n \to +\infty$
- in particular, it does not depend on $p^0$!

# Chi-squared distance under the null (II)

*Proof:* we start with the behavior under $H_0$.

- for any $i \in \{1, \dots, n\}$, set

$$Z_i = \left( \frac{1}{\sqrt{p_1^0}} (\mathbb{1}_{X_i = x_1} - p_1^0), \dots, \frac{1}{\sqrt{p_k^0}} (\mathbb{1}_{X_i = x_k} - p_k^0) \right)^\top .$$

- the $Z_i$ are i.i.d., centered, and have a covariance matrix $\Gamma$ such that

$$\Gamma_{j,j} = \mathrm{Var} \left( \frac{1}{\sqrt{p_j^0}} (\mathbb{1}_{X_i = x_j} - p_j^0) \right) = \frac{p_j^0 (1 - p_j^0)}{p_j^0} = 1 - p_j^0 ,$$

$$\Gamma_{j,j'} = \mathrm{Cov} \left( \frac{1}{\sqrt{p_j^0}} (\mathbb{1}_{X_i = x_j} - p_j^0), \frac{1}{\sqrt{p_{j'}^0}} (\mathbb{1}_{X_i = x_{j'}} - p_{j'}^0) \right) = -\sqrt{p_j^0 p_{j'}^0} .$$

# Chi-squared distance under the null (III)

- denote by $\sqrt{p^0}$ the vector $(\sqrt{p_1^0}, \ldots, \sqrt{p_k^0})^\top$
- the **central limit theorem** yields

$$\sqrt{n}\left(\frac{1}{\sqrt{p_1^0}}(\hat{p}_1 - p_1^0), \ldots, \frac{1}{\sqrt{p_k^0}}(\hat{p}_k - p_k^0)\right)^\top = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i \xrightarrow{\mathcal{L}} V\,,$$

  where $V \sim \mathcal{N}(0, \Gamma)$
- by the **continuous mapping theorem**,

$$D_n^2(\hat{p}_n, p^0) \xrightarrow{\mathcal{L}} \|V\|^2\,.$$

- we notice that $\Gamma = \mathsf{I}_k - \sqrt{p^0}\sqrt{p^0}^\top$ is the **orthogonal projection** on the subspace orthogonal to $\mathrm{Vec}\left(\sqrt{p^0}\right)$
- by **Cochran's theorem**, $\|V\|^2 \sim \chi^2_{k-1}$

# Chi-squared distance under the null (IV)

We now turn towards the behavior under $H_1$.

- under $H_1$, there exists an integer $j_0$ such that

$$p_{j_0} \neq p^0_{j_0} .$$

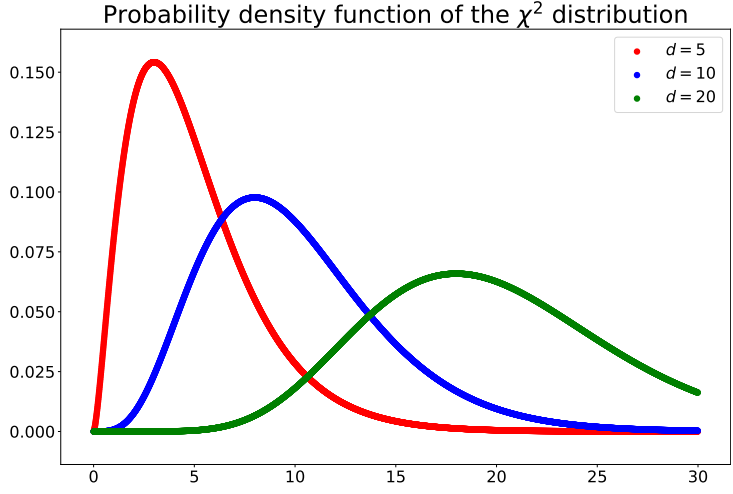
- we write

$$\begin{aligned} D_n^2(\hat{p}_n, p^0) &= n \sum_{j=1}^{k} \frac{(\hat{p}_j - p_j^0)^2}{p_j^0} \\ &\geq n \frac{(\hat{p}_{j_0} - p^0_{j_0})^2}{p^0_{j_0}} \\ &\sim n \frac{(p_{j_0} - p^0_{j_0})^2}{p^0_{j_0}} \end{aligned}$$

$$D_n^2(\hat{p}_n, p^0) \longrightarrow +\infty \quad \text{a.s.} \qquad \square$$

# Reminder: the $\chi^2$ distribution

- let $X \sim \chi^2_d$, then $\mathbb{E}[X] = d$ and $\mathrm{Var}(X) = 2d$



# Quantiles of the $\chi^2$ distribution (I)

- recall **quantile function:** $c_{d,p}$ is a number such that

$$\mathbb{P}\left(\chi^2_d \leq c_{d,p}\right) \geq p\,.$$

- very often given as a table:

| d \ p | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 0.95 | 0.99 |
|---|---|---|---|---|---|---|---|
| 1 | 0.45 | 0.71 | 1.07 | 1.64 | 2.71 | 3.84 | 6.63 |
| 2 | 1.39 | 1.83 | 2.41 | 3.22 | 4.61 | 5.99 | 9.21 |
| 3 | 2.37 | 2.95 | 3.66 | 4.64 | 6.25 | 7.81 | 11.34 |
| 4 | 3.36 | 4.04 | 4.88 | 5.99 | 7.78 | 9.49 | 13.28 |
| 5 | 4.35 | 5.13 | 6.06 | 7.29 | 9.24 | 11.07 | 15.09 |
| 6 | 5.35 | 6.21 | 7.23 | 8.56 | 10.64 | 12.59 | 16.81 |
| 7 | 6.35 | 7.28 | 8.38 | 9.8 | 12.02 | 14.07 | 18.48 |
| 8 | 7.34 | 8.35 | 9.52 | 11.03 | 13.36 | 15.51 | 20.09 |
| 9 | 8.34 | 9.41 | 10.66 | 12.24 | 14.68 | 16.92 | 21.67 |
| 10 | 9.34 | 10.47 | 11.78 | 13.44 | 15.99 | 18.31 | 23.21 |

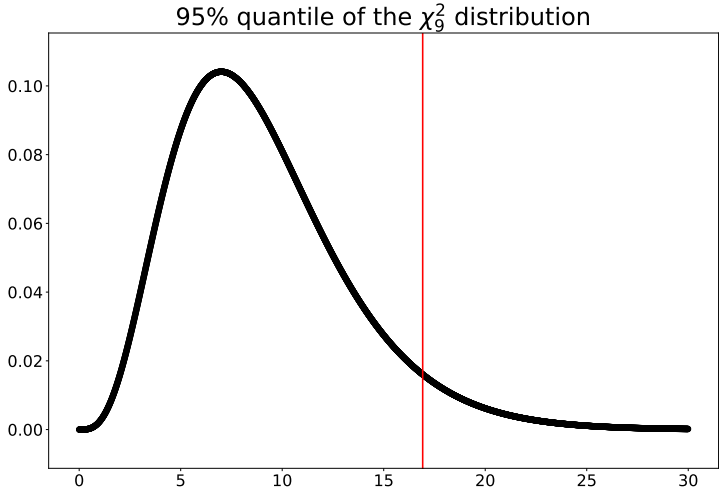
# Quantiles of the $\chi^2$ distribution (II)

- **How to read this table?**
- suppose that we are looking for $c_{9,0.95}$ (a number such that 95% of the observations from a $\chi^2_9$ random variable are smaller):

| d \ p | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | **0.95** | 0.99 |
|---|---|---|---|---|---|---|---|
| 1 | 0.45 | 0.71 | 1.07 | 1.64 | 2.71 | **3.84** | 6.63 |
| 2 | 1.39 | 1.83 | 2.41 | 3.22 | 4.61 | **5.99** | 9.21 |
| 3 | 2.37 | 2.95 | 3.66 | 4.64 | 6.25 | **7.81** | 11.34 |
| 4 | 3.36 | 4.04 | 4.88 | 5.99 | 7.78 | **9.49** | 13.28 |
| 5 | 4.35 | 5.13 | 6.06 | 7.29 | 9.24 | **11.07** | 15.09 |
| 6 | 5.35 | 6.21 | 7.23 | 8.56 | 10.64 | **12.59** | 16.81 |
| 7 | 6.35 | 7.28 | 8.38 | 9.8 | 12.02 | **14.07** | 18.48 |
| 8 | 7.34 | 8.35 | 9.52 | 11.03 | 13.36 | **15.51** | 20.09 |
| **9** | **8.34** | **9.41** | **10.66** | **12.24** | **14.68** | **16.92** | **21.67** |
| 10 | 9.34 | 10.47 | 11.78 | 13.44 | 15.99 | **18.31** | 23.21 |

- thus $c_{9,0.95} = 16.92$

# Quantiles of the $\chi^2$ distribution (III)




- 95% of the observations fall **to the left** of the red line

# Chi-squared test for fit of a distribution

- finally, we propose the following test *("test d'ajustement du chi deux")*:

$$\boxed{\phi(X_1, \dots, X_n) = \mathbb{1}_{D_n^2(\hat{p}_n, p^0) > c_{k-1,1-\alpha}}.}$$

- asymptotically, the test is of size $\alpha$
- the test is also **consistent** (power converges to 1)
- **Remark:** the guarantees only holds for large $n$!
- usually, it is recommended to ensure that $n \geq 30$ and $np_j^0 \geq 5$ for all $j \in \{1, \dots, k\}$
- if it is not the case, **regroup cases**

# Chi-squared test: example

- recall genetics example: $p^0_{RR} = 0.25$, $p^0_{RB} = 0.50$, and $p^0_{BB} = 0.25$
- we want to test

$$H_0 : p = p^0 \quad \text{vs} \quad H_1 : p \neq p^0 .$$

- we repeat **independently** and **in the same conditions** $n = 100$ time the crossing experiment
- we obtain the following data:

| color | red | purple | blue |
|---|---|---|---|
| number | 27 | 55 | 18 |

- sample size is **large enough**
- the $\chi^2$ statistic is given by

$$D_n^2(\hat{p}_n, p^0) = \frac{(27-25)^2}{25} + \frac{(55-50)^2}{50} + \frac{(18-25)^2}{25} = 2.62$$

- the 95% quantile for the $\chi^2_2$ distribution is 5.99, thus **we do not reject the null**

# 28. Extensions

# Chi-squared test for continuous distributions (I)

- possible to adapt the $\chi^2$ statistic to **continuous distribution**
- **Idea:** split the space in $k$ classes $A_1, \dots, A_k$
- from an i.i.d. sample $X_1, \dots, X_n$, we construct

$$\forall i \in \{1, \dots, n\}, \quad Y_i = j \text{ if } X_i \in A_j \,.$$

- then define $\hat{p}$ as

$$\hat{p}_j = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{Y_i = j}$$

- finally, take

$$p_j^0 = \mathbb{P}\left(X \in A_j\right) \,,$$

  and use the previous test
- **Remark:** a rule of thumb is to choose the $A_j$ such that the $p_j^0$ are approximately equal

# Chi-squared test for continuous distributions (II)

- **Example:** suppose that we want to test

$$H_0 : X \sim \mathcal{E}(1) \quad \text{vs} \quad H_1 : X \not\sim \mathcal{E}(1)\,.$$

- split $\mathbb{R}_+$ in five intervals:

$$\mathbb{R}_+ = [0, t_1) \cup [t_1, t_2) \cup [t_2, t_3) \cup [t_3, t_4) \cup [t_4, +\infty)$$

- we write

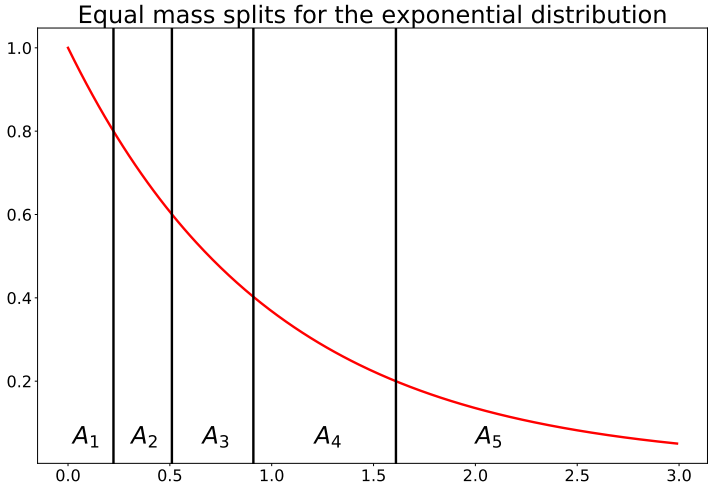
$$\mathbb{P}\left(\mathcal{E}(1) \in [t_j, t_{j+1})\right) = 1 - \mathrm{e}^{-t_{j+1}} - (1 - \mathrm{e}^{-t_j}) = \mathrm{e}^{-t_j} - \mathrm{e}^{-t_{j+1}}\,.$$

- then solve:

$$\begin{cases} \mathrm{e}^{-0} - \mathrm{e}^{-t_1} &= 1/5 \\ \mathrm{e}^{-t_1} - \mathrm{e}^{-t_2} &= 1/5 \\ \mathrm{e}^{-t_2} - \mathrm{e}^{-t_3} &= 1/5 \\ \mathrm{e}^{-t_3} - \mathrm{e}^{-t_4} &= 1/5 \\ \mathrm{e}^{-t_4} - 0 &= 1/5 \end{cases} \Rightarrow \begin{cases} \mathrm{e}^{-t_1} &= 4/5 \\ \mathrm{e}^{-t_2} &= 3/5 \\ \mathrm{e}^{-t_3} &= 2/5 \\ \mathrm{e}^{-t_4} &= 1/5 \end{cases}$$

# Chi-squared test for continuous distributions (III)

- that is, $t_1 = 0.22, t_2 = 0.51, t_3 = 0.91,$ and $t_4 = 1.61$



# Chi-squared test for continuous distributions (IV)

- suppose that we observe 100 realizations of $X$, dispatching as follows:

| class | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ |
|---|---|---|---|---|---|
| number | 20 | 27 | 31 | 15 | 7 |

- we compute the test statistic:

$$D_n^2(\hat{p}_n, p^0) = \frac{(20-20)^2}{20} + \frac{(27-20)^2}{20} + \frac{(31-20)^2}{20} + \frac{(15-20)^2}{20} + \frac{(7-20)^2}{20}$$
$$= 18.2$$

- the 95% quantile for the $\chi_4^2$ distribution is 9.49: **we reject the null**

# Parametric family of distributions (I)

- let us now consider a parametric family of **discrete** distributions

$$\mathcal{F} = \{p(\theta), \theta \in \Theta\},$$

  with $\Theta \subseteq \mathbb{R}^D$
- we want to test

$$H_0 : p \in \mathcal{F} \quad \text{vs} \quad H_1 : p \notin \mathcal{F}$$

  *("test d'ajustement à une famille de lois")*
- **Idea:** pick a reference law in $\mathcal{F}$
- **Solution:** construct $\hat{\theta}_n$ the maximum likelihood estimator of $\theta$
- we can then consider

$$p(\hat{\theta}_n) = \Big(p_1(\hat{\theta}_n), \ldots, p_k(\hat{\theta}_n)\Big)^\top \in \mathbb{R}^k.$$

# Parametric family of distributions (II)

- we then compare the empirical frequencies to $p(\hat{\theta}_n)$ as before, giving rise to the test statistic

$$D_n^2(\hat{p}_n, \mathcal{F}) = n \sum_{j=1}^{k} \frac{(\hat{p}_j - p_j(\hat{\theta}_n))^2}{p_j(\hat{\theta}_n)} .$$

**Theorem:** Suppose that $D < k - 1$. Assume that the mapping

$$p : \theta \mapsto p(\theta) = (p_1(\theta), \ldots, p_k(\theta))$$

is one-to-one, $C^2$, and has no singular point. Then, if $\hat{\theta}_n$ is a consistent estimator of $\theta$,

$$D_n^2(\hat{p}_n, \mathcal{F}) \xrightarrow{\mathcal{L}} \chi^2_{k-D-1}$$

under the null.

# Parametric family of distributions (III)

- under $H_1$, $D_n^2(\hat{p}_n, \mathcal{F}) \to +\infty$ almost surely if the distance from $p$ to $\mathcal{F}$ is $> 0$
- in definitive, we have the following test *("test d'ajustement à une famille paramétrée de lois")*:

$$\phi(X_1, \ldots, X_n) = \mathbb{1}_{D_n^2(\hat{p}_n, \mathcal{F}) > c_{k-D-1,1-\alpha}},$$

  where $c_{\cdot,\cdot}$ denote the $\chi^2$ quantiles as before
- also possible to extend for **continuous** laws as explained previously
- we will see in a future course the **Kolmogorov-Smirnov test** that avoids binning the data

# Chi-squared test for independence (I)

- **Setting:** $n$ observations $(X_1, Y_1), \dots, (X_n, Y_n)$ i.i.d. taking **discrete** values
- denote by $\nu$ the joint law of $(X, Y)$
- we want to test

$$H_0 : X \text{ independent from } Y \quad \text{vs} \quad H_1 : \text{not independent}$$

*("test du chi deux d'indépendance")*

- **Idea:** transform this problem into a *particular case* of the previous test:

$$H_0 : \nu \in \mathcal{F} \quad \text{vs} \quad H_1 : \nu \notin \mathcal{F}$$

- **Question:** which family to consider?

# Chi-squared test for independence (II)

- denote by $\mathcal{X} = \{x_1, \dots, x_r\}$ (resp. $\mathcal{Y} = \{y_1, \dots, y_s\}$) the support of $X$ (resp. $Y$)
- $\mathcal{P}(\mathcal{X})$ (resp. $\mathcal{P}(\mathcal{Y})$) the set of distributions on $\mathcal{X}$ (resp. $\mathcal{Y}$) with **positive mass**
- we set

$$\mathcal{F} = \{p \otimes q, \;\; p \in \mathcal{P}(\mathcal{X}), q \in \mathcal{P}(\mathcal{Y})\}$$

- **Intuition:** set a discrete distributions on $\mathcal{X} \times \mathcal{Y}$ that can be written as a product
- the parameter of this family is

$$\theta = (p_1, \dots, p_{r-1}, q_1, \dots, q_{s-1})^\top \in \mathbb{R}^{r+s-2}$$

# Chi-squared test for independence (III)

- the maximum likelihood estimator is given by $(\hat{p}, \hat{q})$, where

$$\hat{p}_j = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i = x_j} \quad \text{and} \quad \hat{q}_{j'} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{Y_i = y_{j'}} .$$

- on the other side, the empirical frequencies are given by

$$\hat{\nu}_{j,j'} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(X_i, Y_i) = (x_j, y_{j'})} .$$

- the test statistic is

$$D_n^2(\hat{p}_n, \hat{q}_n, \hat{\nu}_n) = n \sum_{j=1}^{r} \sum_{j'=1}^{s} \frac{(\hat{\nu}_{j,j'} - \hat{p}_j \hat{q}_{j'})^2}{\hat{p}_j \hat{q}_{j'}} .$$

# Chi-squared test for independence (IV)

- one can check the assumptions of the theorem, and show that

$$D_n^2(\hat{p}_n, \hat{q}_n, \hat{\nu}_n) \xrightarrow{\mathcal{L}} \chi^2_{k-D-1}$$

  under the null
- **Question:** how many degrees of freedom?
- here $k = rs$ (number of parameters for $\nu$) and $D = r - 1 + s - 1 = r + s - 2$ (number of parameters for $\theta$)
- thus

$$k - D - 1 = rs - (r + s - 2) - 1 = (r - 1)(s - 1)\,.$$

- in definitive, we have the following test *("test du chi deux d'indépendance")*:

$$\boxed{\phi(X_1, Y_1, \ldots, X_n, Y_n) = \mathbb{1}_{D_n^2(\hat{p}_n, \hat{q}_n, \hat{\nu}_n) > c_{(r-1)(s-1),1-\alpha}}\,,}$$

  where $c_{\cdot,\cdot}$ denote the $\chi^2$ quantiles as before

# Chi-squared test for homogeneity (I)

- **particular case** of the chi-squared test for independence
- **discrete** distributions $a$ and $b$ on $\{x_1, \ldots, x_k\}$
- we are given i.i.d. samples $S_1, \ldots, S_n$ and $T_1, \ldots, T_m$ from $a$ and $b$
- we want to test *("test du chi deux d'homogénéité")*:

$$H_0 : a = b \quad \text{vs} \quad H_1 : a \neq b\,.$$

- **Example:** categories of population (men / women), $k$ outcomes
- **Idea:** set

$$((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1}), \ldots, (X_{n+m}, Y_{n+m})) = \\ ((S_1, 1), \ldots, (S_n, 1), (T_1, 2), \ldots, (T_m, 2))$$

- then test for $X$ independent of $Y$!

# Chi-squared test for homogeneity (II)

- more precisely, for any $1 \leq j \leq k$, set

$$N_j = \sum_{i=1}^{n} \mathbb{1}_{S_i = x_j}\,, M_j = \sum_{i=1}^{n} \mathbb{1}_{T_i = x_j}\,, \quad \text{and} \quad \hat{p}_j = \frac{N_j + M_j}{n+m}\,.$$

- then the test statistic is

$$D^2_{n,m}(\hat{p}_{n,m}) = \sum_{j=1}^{k} \left( \frac{(N_j - n\hat{p}_j)^2}{n\hat{p}_j} + \frac{(M_j - m\hat{p}_j)^2}{m\hat{p}_j} \right)\,.$$

- under $H_0$, $D^2_{n,m}(\hat{p}_{n,m}) \xrightarrow{\mathcal{L}} \chi^2_{k-1}$ and $D^2_{n,m}(\hat{p}_{n,m}) \xrightarrow{\text{a.s.}} +\infty$ under $H_1$
- chi-squared test for homogeneity *("test du chi deux d'homogénéité")*:

$$\boxed{\phi(S_1, \ldots, S_n, T_1, \ldots, T_m) = \mathbb{1}_{D^2_{n,m}(\hat{p}_{n,m}) > c_{k-1,1-\alpha}}\,,}$$

# 29. Cumulative distribution function

# Recap on cumulative distribution functions (I)

**Definition:** Let $X$ be a *real-valued* random variable. We call *cumulative distribution function* of $X$ *("fonction de répartition")* the map $F_X : \mathbb{R} \to \mathbb{R}$ defined by

$$\forall x \in \mathbb{R}, \quad F_X(x) = \mathbb{P}(X \leq x) \ .$$

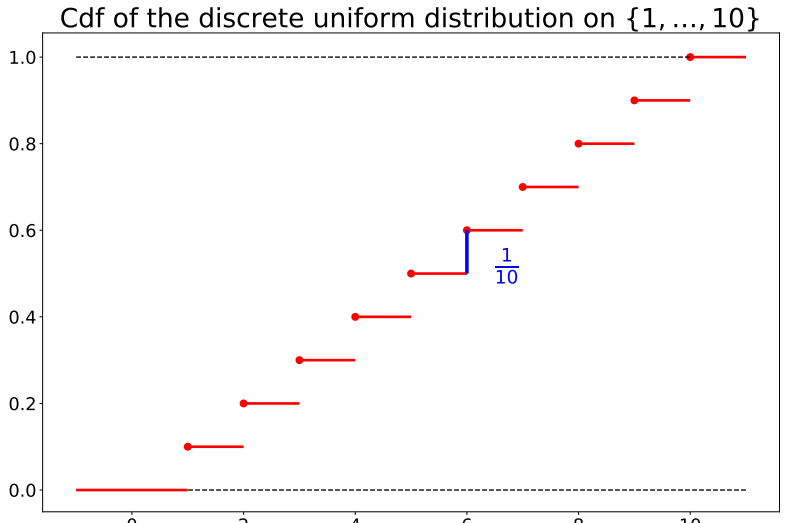
- **Beware:** the $\leq$ is important in the case of *discrete* distributions
- suppose that $X \in \{x_1, \ldots, x_k\}$ a.s., then

$$F_X(x) = \sum_{x_j \leq x} \mathbb{P}(X = x_j) \ .$$

- **Example:** uniform distribution on $\{1, \ldots, k\}$:

$$\mathbb{P}(X = x_j) = \frac{1}{k} \quad \Rightarrow \quad F_X(x) = \frac{1}{k} \left| \{ j \in \{1, \ldots, k\} \text{ s.t. } j \leq x \} \right| \ .$$

# Recap on cumulative distribution functions (II)



# Recap on cumulative distribution functions (III)

- **Continuous distribution:** if there exists a *density*, we can compute an integral since

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_X(t)\mathrm{d}t\,.$$

- **Example:** uniform distribution on $[0,1]$, density given by

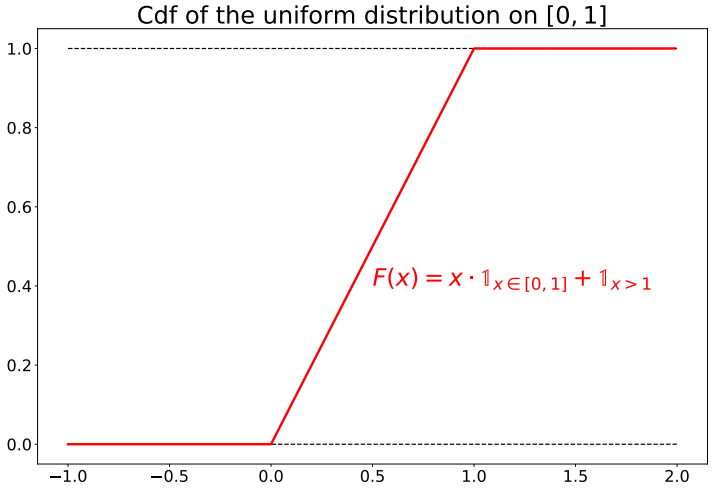
$$f_X(t) = \mathbb{1}_{t\in[0,1]}\,.$$

- we write

$$F_X(x) = \int_{-\infty}^{x} \mathbb{1}_{t\in[0,1]}\mathrm{d}t$$

$$\Rightarrow F_X(x) = \begin{cases} 0 & \text{if} \quad x < 0 \\ x & \text{if} \quad x \in [0,1] \\ 1 & \text{otherwise} \end{cases}$$

# Recap on cumulative distribution functions (IV)



# Recap on cumulative distribution functions (V)

- sometimes the integral is **harder** to compute
- **Example:** standard Gaussian distribution $\mathcal{N}(0,1)$, density given by

$$f_X(t) = \frac{1}{\sqrt{2\pi}} \cdot \mathrm{e}^{\frac{-t^2}{2}} .$$

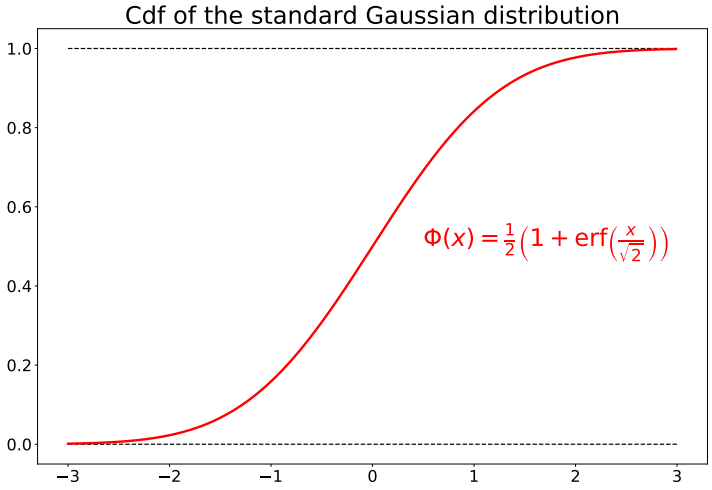
- **Problem:** we can not find a primitive with usual functions
- we define the **error function**

$$\operatorname{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^{x} \mathrm{e}^{-t^2} \mathrm{d}t = \sqrt{\frac{2}{\pi}} \int_{0}^{x} \mathrm{e}^{-t^2} \mathrm{d}t .$$

- the cdf of the Gaussian is usually denoted by $\Phi$ and can be written

$$\Phi(x) = \frac{1}{2} \left( 1 + \operatorname{erf}\left( \frac{x}{\sqrt{2}} \right) \right) .$$

# Recap on cumulative distribution functions (VI)



# Recap on cumulative distribution functions (VII)

**Property:** Let $X$ be a real-valued random variable and $F_X$ its cumulative distribution function. Then

- $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to +\infty} F_X(x) = 1$;
- $F_X$ is a non-decreasing function;
- $F_X$ is right-continuous *("càdlàg")*;
- $F_X$ is discontinuous at the atoms of $\mathbb{P}$.

Moreover, any function satisfying these properties is the cumulative distribution function of a random variable.

*Proof:* elementary properties of $\mathbb{P}$. $\square$

# Recap on cumulative distribution functions (VIII)

- define the **generalized inverse**

$$\forall q \in [0,1], \quad F^{-1}(q) = \inf\{x \in \mathbb{R} : \ F(x) \geq q\}\,.$$

- it holds that

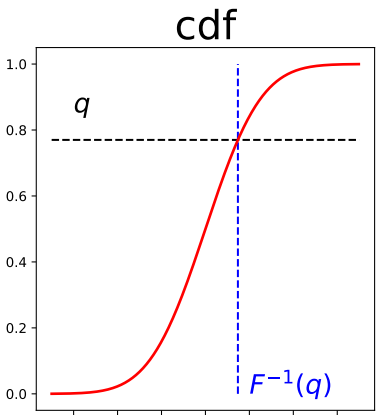
$$q \leq F(x) \quad \Leftrightarrow \quad F^{-1}(q) \leq x\,.$$

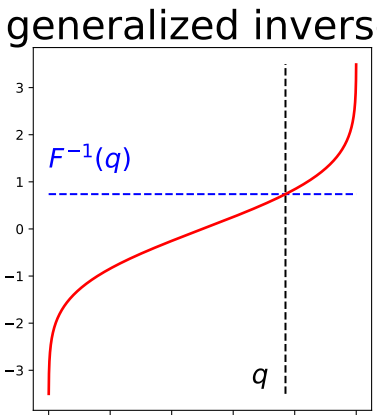
- in particular, $F \circ F^{-1}(q) \geq q$, with equality iff $q \in \mathrm{Im}\,(F)$

**Proposition:** For any random variable with cumulative distribution function $F$, is $U \sim \mathcal{U}([0,1])$, then $F^{-1}(U)$ is a random variable with cdf $F$.

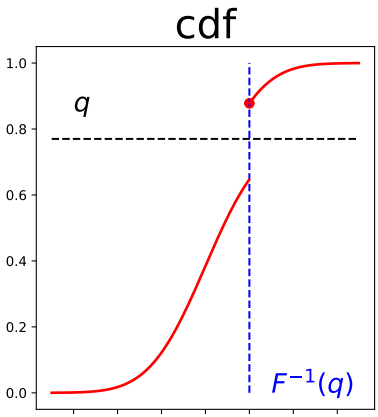
*Proof:* $\mathbb{P}\left(F^{-1}(U) \leq t\right) = \mathbb{P}\left(U \leq F(t)\right) = F(t)$ $\square$

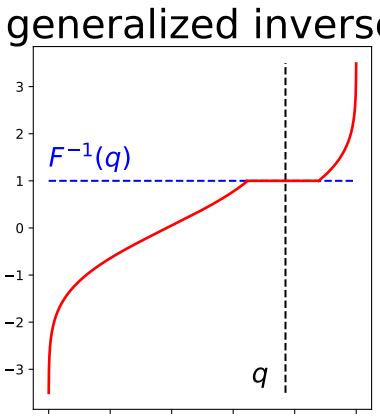
# Recap on cumulative distribution functions (IX)



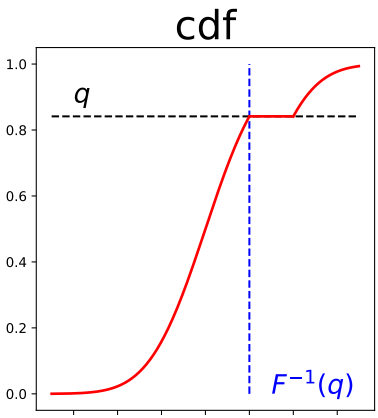



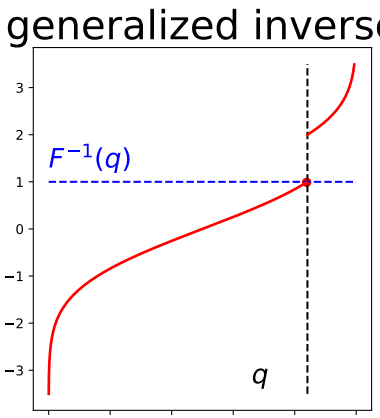
# Recap on cumulative distribution functions (X)



# Recap on cumulative distribution functions (XI)



# 30. Estimating the cumulative distribution function

# Empirical cumulative distribution function (I)

- **Question:** given i.i.d. samples $X_1, \dots, X_n$ of $X$, can we estimate the cumulative distribution function $F_X$?

**Definition:** The *empirical cumulative distribution function* associated to $X_1, \dots, X_n$ *("fontion de répartition empirique")* is the random variable $F_n$ defined by

$$F_n : \mathbb{R} \longrightarrow [0,1]$$
$$x \longmapsto \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq x}$$

- **Important:** $F_n$ is a **random** function!
- we shorten to "ecdf"

# Empirical cumulative distribution function (II)

- **How to build $F_n$?** first rank the $X_i$s (form the *order statistics*):

$$(X_1, \ldots, X_n) \longmapsto (X_{(1)}, \ldots, X_{(n)}),$$

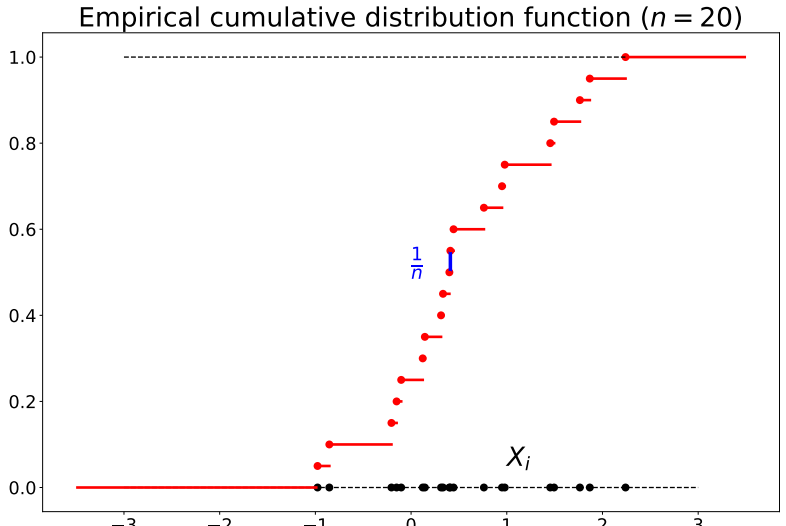
  with $\min\{X_i\} = X_{(1)} \leq \cdots \leq X_{(n)} = \max\{X_i\}$
- $F_n$ is discontinuous only at the $X_{(i)}$, jumping from $1/n$
- $F_n$ takes the value 0 on $(-\infty, X_{(1)})$ and 1 on $[X_{(n)}, +\infty)$
- $F_n$ is constant on each interval $[X_{(i)}, X_{(i+1)})$, more precisely:

$$F_n(X_{(j)}-) = \frac{j-1}{n} \quad \text{and} \quad F_n(X_{(j)}) = \frac{j}{n}.$$

- in particular,

$$F_n(-\infty) = 0 \quad \text{and} \quad F_n(+\infty) = 1 \quad \text{a.s.}$$

# Empirical cumulative distribution function (III)



# Convergence of the ecdf (I)

**Proposition:** Let $x \in \mathbb{R}$. Let $X_1, \dots, X_n$ be an i.i.d. sample from $X$, and let $F$ be the cdf from $X$. Define $F_n$ the ecdf associated to the sample $X_1, \dots, X_n$ as before. Then, when $n \to +\infty$, it holds that

$$\forall x \in \mathbb{R}, \qquad \sqrt{n}(F_n(x) - F(x)) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, F(x)(1 - F(x))\right) .$$

In particular, for any fixed $x \in \mathbb{R}$, $F_n(x)$ is an unbiased consistent estimator of $F(x)$.

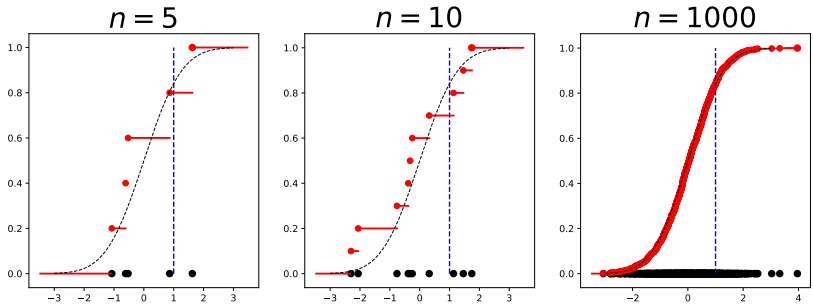
*Proof:* Recall that $F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq x}$. The random variables $\mathbb{1}_{X_i \leq x}$ are i.i.d. Bernoulli random variables of parameter

$$\mathbb{E}\left[\mathbb{1}_{X_i \leq x}\right] = \mathbb{P}\left(X_i \leq x\right) = F(x) .$$

In particular, they have finite variance $F(x)(1 - F(x))$. The result follows from the central limit theorem. $\square$

# Convergence of the ecdf (II)

- **Pointwise convergence:** fix $x \in \mathbb{R}$, $F_n(x) \xrightarrow{\text{a.s.}} F(x)$



# Convergence of the ecdf (III)

- in fact a stronger result holds: recall that, for any $f : \mathbb{R} \to \mathbb{R}$, we define

$$\|f\|_\infty = \sup_{x \in \mathbb{R}} |f(x)| \, .$$

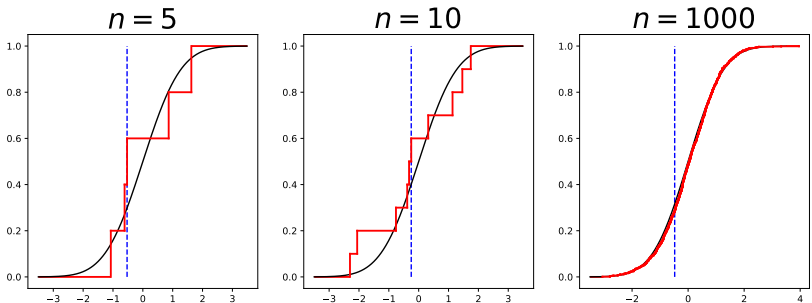
**Theorem (Glivenko-Cantelli, 1933):** Let $X_1, \ldots, X_n$ be an i.i.d. sample from $X$, and let $F$ be the cdf from $X$. Define $F_n$ the ecdf associated to the sample $X_1, \ldots, X_n$ as before. Then, when $n \to +\infty$, it holds that

$$\|F_n - F\|_\infty \xrightarrow{\text{a.s.}} 0 \, .$$

- in other words, $F_n$ is a *strongly consistent estimator* of $F$ in the sense of $\|\cdot\|_\infty$

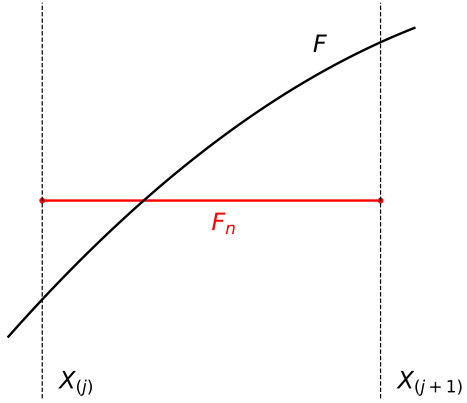
# Convergence of the ecdf (IV)

- **Uniform convergence:** $\|F_n - F\|_\infty \xrightarrow{\text{a.s.}} 0$



# Convergence of the ecdf (V)

- **How to compute $\|F_n - F\|_\infty$?**
- since $F$ is non-decreasing and $F_n$ is piecewise-constant, $\|F_n - F\|_\infty$ attained at an extremity on each segment $[X_{(j)}, X_{(j+1)}]$



# Convergence of the ecdf (VI)

- recall that
$$F_n(X_{(j)}-) = \frac{j-1}{n} \quad \text{and} \quad F_n(X_{(j)}) = \frac{j}{n}\,.$$
- **Idea:** at each end of a segment $X_{(j)}$, we compare to the two possible values
- we have obtained a very simple way to compute
$$\|F_n - F\|_\infty = \max_{1\le i\le n} \left\{ \max \left\{ \left| F(X_{(i)}) - \frac{i}{n} \right|, \left| F(X_{(i)}) - \frac{i-1}{n} \right| \right\} \right\}\,.$$
- **Computational cost:** order the $n$-sample and compute the max: $\mathcal{O}(n \log n)$

# 31. Kolmogorov-Smirnov test

# Kolmogorov-Smirnov test (I)

- **Setting:** we are given a reference distribution $F_{\text{ref}}$, and an i.i.d. sample $X_1, \dots, X_n$ from an unknown distribution $F$
- we want to test

$$H_0 : F = F_{\text{ref}} \quad \text{vs} \quad H_1 : F \neq F_{\text{ref}}$$

*("test d'ajustement à une loi donnée")*

- **Idea:** we know that, for large $n$, $\|F_n - F\|_\infty$ vanishes (according to Glivenko-Cantelli)
- let us consider

$$h_n(X, F_{\text{ref}}) = h_n(X_1, \dots, X_n, F_{\text{ref}}) = \|F_n - F_{\text{ref}}\|_\infty \ .$$

- for large $n$, $F_n \approx F$, thus $h_n$ **is small under the null** and we reject $H_0$ if $h_n$ is too large

# Kolmogorov-Smirnov test (II)

- under $H_0$,

$$
\begin{aligned}
h_n(X, F_{\text{ref}}) &= \sup_{x \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq x} - F_{\text{ref}}(x) \right| \\
&\sim \sup_{x \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{F_{\text{ref}}^{-1}(U_i) \leq x} - F_{\text{ref}}(x) \right| \\
&\sim \sup_{x \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{U_i \leq F_{\text{ref}}(x)} - F_{\text{ref}}(x) \right| \\
h_n(X, F_{\text{ref}}) &\sim \sup_{q \in \text{Im}(F_{\text{ref}})} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{U_i \leq q} - q \right|
\end{aligned}
$$

  where $U_1, \ldots, U_n$ is an i.i.d. $\mathcal{U}([0,1])$ sample
- in particular, if $F_{\text{ref}}$ is continuous *("sans atomes")*, the distribution of $h_n$ under the null does not depend on $F_{\text{ref}}$ *("statistique libre")*

# Kolmogorov-Smirnov test (III)

- **For small** $n$**,** $(n \leq 100)$, we can use this property to compute quantiles from $h_n$ under the null by simulating $h_{n,U} = h_n(U, F_{\mathcal{U}})$
- **For large** $n$**,** $(n > 100)$, we use an asymptotic result:

**Theorem:** Under the null, when $n \to +\infty$,

$$\sqrt{n} h_n(X, F_{\text{ref}}) \xrightarrow{\mathcal{L}} K = \sup_{t \in [0,1]} |B(t)| \ ,$$

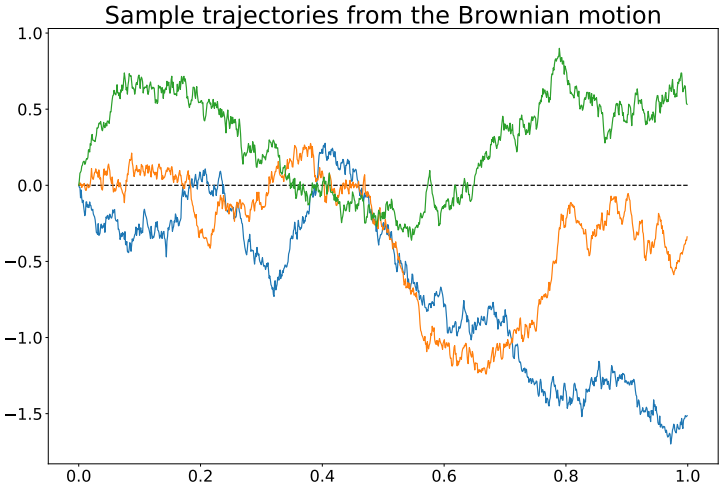
where $B(\cdot)$ is the Brownian bridge.

- sometimes $K$ is called the **Kolmogorov** distribution
- we know the cdf of $K$ and we can invert it numerically to find quantiles:

$$\mathbb{P}(K \leq x) = 1 - 2 \sum_{k=1}^{+\infty} (-1)^{k+1} \mathrm{e}^{-2k^2 x^2} \ .$$

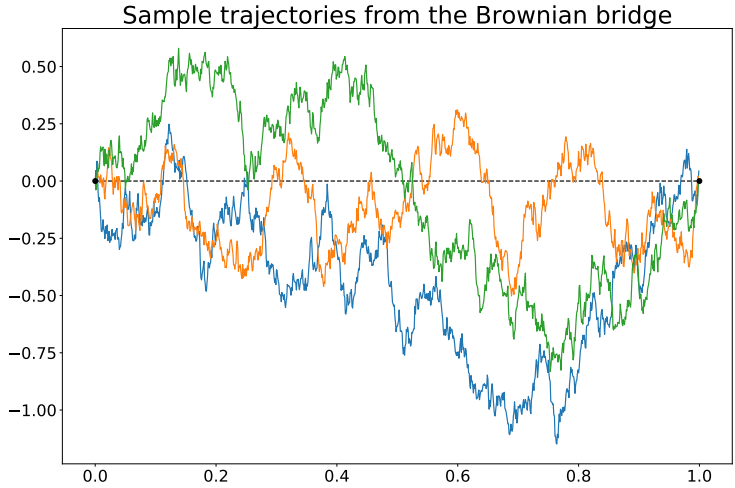
# Reminder: Brownian motion

- $W(0) = 0$ a.s., $W(\cdot)$ a.s. $C^0$, ind. increments, $W(t) - W(s) \sim \mathcal{N}(0, t-s)$



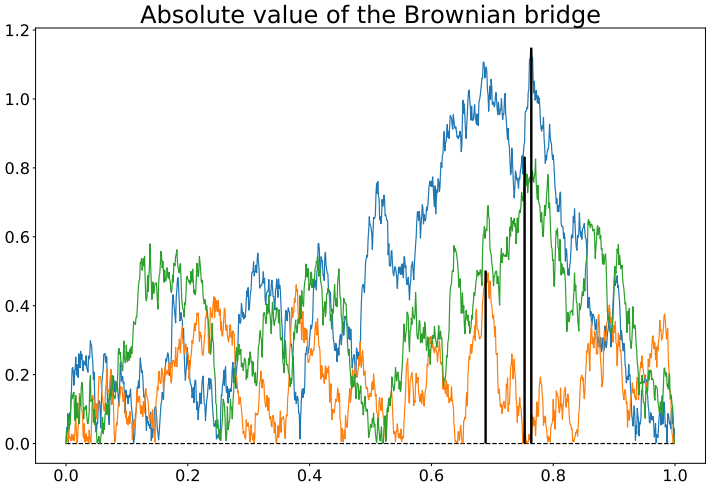
# Brownian bridge

- $B(t) = (W(t)|W(1) = 0)$ *("pont brownien")*

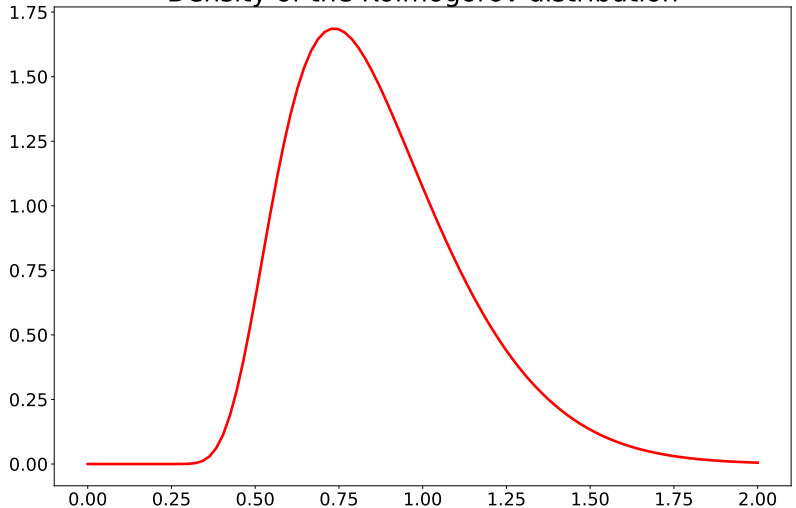


# Kolmogorov distribution (I)

- recall $K = \sup_{t\in[0,1]} |B(t)|$ (in black below)



# Kolmogorov distribution (II)



# Kolmogorov-Smirnov test

- we let $k_{1-\alpha}$ be the quantile of level $1-\alpha$ of $K$:

$$\mathbb{P}\left(\sqrt{n}h_n(X, F_{\text{ref}}) \leq k_{1-\alpha}\right) \approx \mathbb{P}(K \leq k_{1-\alpha}) = 1-\alpha\,.$$

- we deduce the following test *("test de Kolmogorov-Smirnov")*:

$$\boxed{\phi(X, F_{\text{ref}}) = \mathbb{1}_{h_n(X,F_{\text{ref}})>\xi_{n,1-\alpha}}\,,}$$

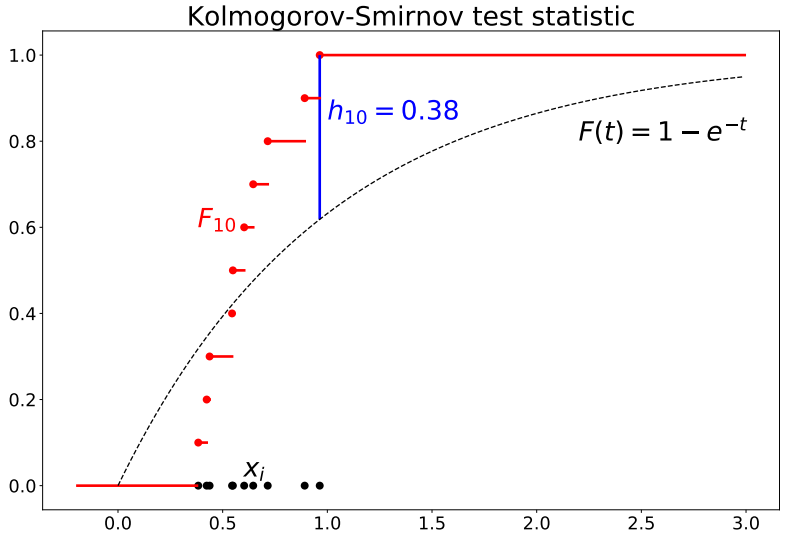
  where $\xi_{n,1-\alpha}$ comes from a table for small $n$, whereas for large $n$ we take $\xi_{n,1-\alpha} = n^{-1/2}k_{1-\alpha}$
- $\phi_\alpha$ is consistent and asymptotically of size $\alpha$
- the confidence region is **a band of half-length** $\xi_{n,1-\alpha}$ **around** $F_n$ containing only cdf:

$$\hat{C}_{1-\alpha} = \{G \text{ cdf}: \ \forall x \in \mathbb{R}, |F_n(x) - G(x)| \leq n^{-1/2}k_{1-\alpha}\}$$

# Kolmogorov-Smirnov test: example (I)

- we want to test
$$H_0 : F = \mathcal{E}(1) \quad \text{vs} \quad H_1 : F \neq \mathcal{E}(1)\,.$$
- we are given 10 observations
$$x = (0.55, 0.72, 0.6, 0.54, 0.42, 0.65, 0.44, 0.89, 0.96, 0.38)$$
- let us recall that
$$h_{10}(x, \mathcal{E}(1)) = \max_{1 \leq i \leq 10} \left\{ \max \left\{ \frac{i}{10} - 1 + \mathrm{e}^{-x_{(i)}}, 1 - \mathrm{e}^{-x_{(i)}} - \frac{i-1}{10} \right\} \right\},$$
since the cdf of $\mathcal{E}(1)$ is $F(t) = 1 - \mathrm{e}^{-t}$
- we obtain
$$h_{10}(x, \mathcal{E}(1)) \approx 0.38$$

# Kolmogorov-Smirnov test: example (II)



# Kolmogorov-Smirnov test: example (III)

- small $n$: we read a quantile table

| n \ p | 0.90 | 0.95 | 0.975 | 0.99 | 0.995 |
|---|---|---|---|---|---|
| 1 | 0.900 | 0.950 | 0.975 | 0.990 | 0.995 |
| 2 | 0.684 | 0.776 | 0.842 | 0.900 | 0.929 |
| 3 | 0.565 | 0.636 | 0.708 | 0.785 | 0.829 |
| 4 | 0.493 | 0.565 | 0.624 | 0.689 | 0.734 |
| 5 | 0.447 | 0.509 | 0.563 | 0.627 | 0.669 |
| 6 | 0.410 | 0.468 | 0.519 | 0.577 | 0.617 |
| 7 | 0.381 | 0.436 | 0.483 | 0.538 | 0.576 |
| 8 | 0.358 | 0.410 | 0.454 | 0.507 | 0.542 |
| 9 | 0.339 | 0.387 | 0.430 | 0.480 | 0.513 |
| 10 | 0.323 | 0.369 | 0.409 | 0.457 | 0.489 |
| 11 | 0.308 | 0.352 | 0.391 | 0.437 | 0.468 |
| 12 | 0.296 | 0.338 | 0.375 | 0.419 | 0.449 |

- $h_{10} > \xi_{10,0.95}$: at level 5% we **reject** the null

# 32. Extensions

# Parametric family of distributions (I)

- let us now consider a parametric family of distributions

$$\mathcal{F} = \{F_\theta : \ \theta \in \Theta\},$$

  with $\Theta \subseteq \mathbb{R}^D$
- we want to test

$$H_0 : F \in \mathcal{F} \quad \text{vs} \quad H_1 : F \notin \mathcal{F}$$

  *("test d'ajustement à une famille de lois")*
- **Idea:** pick a reference in $\mathcal{F}$
- we *could* construct $\hat{\theta}_n$ the maximum likelihood estimator of $\theta$, and then consider

$$h'_n(X, \mathcal{F}) = \left\| F_n - F_{\hat{\theta}_n} \right\|_\infty = \sup_{x \in \mathbb{R}} \left| F_n(x) - F_{\hat{\theta}_n}(x) \right| .$$

- **Problem:** the distribution of $h'_n(X, \mathcal{F})$ depends on $\hat{\theta}_n$!

# Parametric family of distribution (II)

- in some specific cases, we have a way out
- **Example:** exponential distributions $\mathcal{E}(\lambda)$
- in that case

$$F_\lambda(x) = (1 - \mathrm{e}^{-\lambda x})\mathbb{1}_{x \geq 0}\,.$$

- the *maximum likelihood estimator* of $\lambda$ is given by

$$\hat{\lambda}_n = \frac{1}{\overline{X}_n}$$

**Proposition:** Let $X \sim \mathcal{E}(1)$ and $\lambda > 0$. Then $X/\lambda \sim \mathcal{E}(\lambda)$.

*Proof:* $\mathbb{P}(X/\lambda \leq x) = \mathbb{P}(X \leq \lambda x) = 1 - \mathrm{e}^{-\lambda x}$. $\square$

# Parametric family of distribution (III)

- let $(X'_1, \ldots, X'_n)$ be an i.i.d. $\mathcal{E}(1)$ sample. We write

$$
\begin{aligned}
h'(X, \mathcal{E}) &= \sup_{x \in \mathbb{R}_+} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq x} - \left(1 - \mathrm{e}^{-x/\overline{X}_n}\right) \right| \\
&\sim \sup_{x \in \mathbb{R}_+} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X'_i \leq \lambda x} - \left(1 - \mathrm{e}^{-\lambda x/\overline{X'}_n}\right) \right| \\
&\sim \sup_{t \in \mathbb{R}_+} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X'_i \leq t} - \left(1 - \mathrm{e}^{-t/\overline{X'}_n}\right) \right|
\end{aligned}
$$

which **does not depend on $\lambda$.**

- thus we can easily tabulate the law and obtain a test **specific to the exponential distribution**

# Two-sample Kolmogorov-Smirnov test (I)

- we are given two i.i.d. samples $X = (X_1, \dots, X_n)$ and $Y = (Y_1, \dots, Y_m)$
- we denote by $F$ (resp. $G$) the distribution of $X_1$ (resp. $Y_1$), and we want to test

$$H_0 : F = G \quad \text{vs} \quad H_1 : F \neq G$$

*("test d'homogénéité")*

- **Idea:** form the empirical cumulative distributions

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq x} \quad \text{and} \quad G_m(x) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{Y_i \leq x} \,,$$

and define the test statistic

$$h_{n,m}(X, Y) = \|F_n - G_m\|_\infty = \sup_{x \in \mathbb{R}} |F_n(x) - G_m(x)| \,.$$

# Two-sample Kolmogorov-Smirnov test (II)

- **Under** $H_0$**,** $h_{n,m}$ depends on $F$ only through $\mathrm{Im}\,(F)$
- as before, we are reduced to $h_{n,m}(U, V)$, where $U$ (resp. $V$) is a $n$-sample (resp. $m$-sample) of $\mathcal{U}([0,1])$
- we can simulate $h_{n,m}(U, V)$ and compute the quantiles $d_{n,m,\cdot}$:

$$\mathbb{P}\left(h_{n,m}(U, V) \leq d_{n,m,1-\alpha}\right) = 1 - \alpha\,.$$

- we deduce the following test *("test d'homogénéité de Kolmogorov-Smirnov")*:

$$\boxed{\phi(X, Y) = \mathbb{1}_{h_{n,m}(X,Y) > d_{n,m,1-\alpha}}\,,}$$

- $\phi(X, Y)$ is consistent and of size $\alpha$ when the distributions are continuous

# 33. Density estimation

# Density estimation (I)

- **Setting:** $X$ random variable with values in $\mathbb{R}^d$
- we assume that $X$ has **unknown** density $f$ with respect to the Lebesgue measure on $\mathbb{R}^d$
- we are given $(X_1, \ldots, X_n)$ i.i.d. sample from $X$
- **Problem:** how to build an estimator $\hat{f}_n$ of $f$?
- **Important remark:** $f$ is a function, we want to build

$$\hat{f}_n = \hat{f}(X_1, \ldots, X_n) : \mathbb{R}^d \longrightarrow \mathbb{R}$$
$$x \longmapsto \hat{f}_n(x)$$

- **non-parametric** estimation problem
- ubiquitous problem in statistics, examples include data visualization, classification, anomaly detection, etc.

# Density estimation (II)

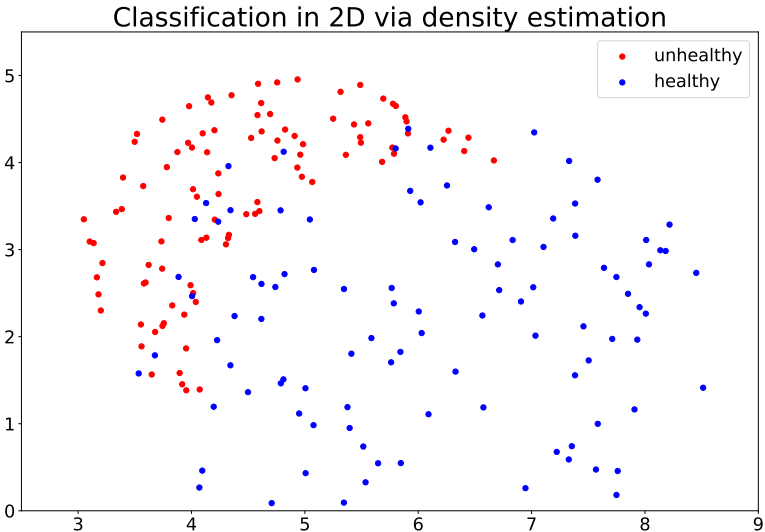
- **Example:** classification: $(X_i, Y_i)_{1 \leq i \leq n}$ i.i.d. distributed according to $(X, Y)$
- $X \in \mathbb{R}^2$ measurement of concentration of 2 proteins in the blood
- $Y = 1 \Rightarrow$ unhealthy, $Y = 0 \Rightarrow$ healthy
- **Goal:** predict as accurately as possible the label $y$ of a new observation $x$
- let us denote by $f$ (resp. $g$) the density of $X$ conditionally to $Y = 1$ (resp. $Y = 0$)
- if $\mathbb{P}(Y = 0) = \mathbb{P}(Y = 1)$, it is possible to show that the best predictor for $y$ is given by

$$y_n^\star(x) = \mathbb{1}_{f(x) > g(x)} \, .$$

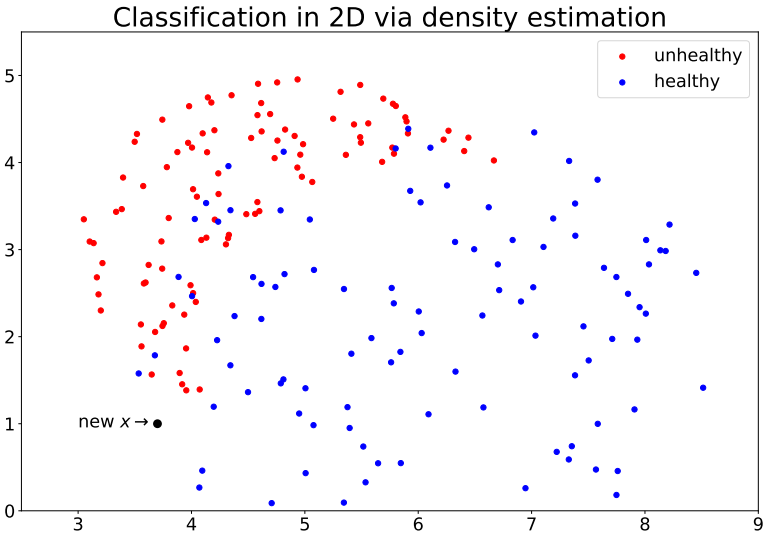
- suppose that we build $\hat{f}_n$ (resp. $\hat{g}_n$) estimators of $f$ (resp. $g$), then we can propose the following estimator of $y$:

$$\hat{y}_n(x) = \mathbb{1}_{\hat{f}_n(x) > \hat{g}_n(x)} \, .$$

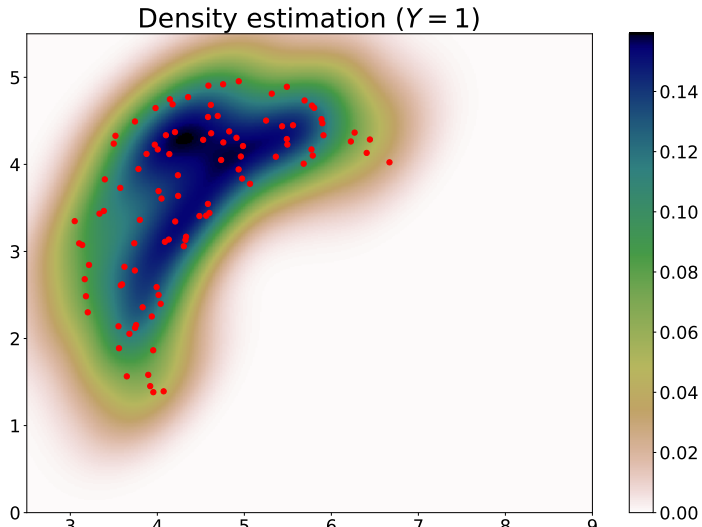
# Density estimation (III)



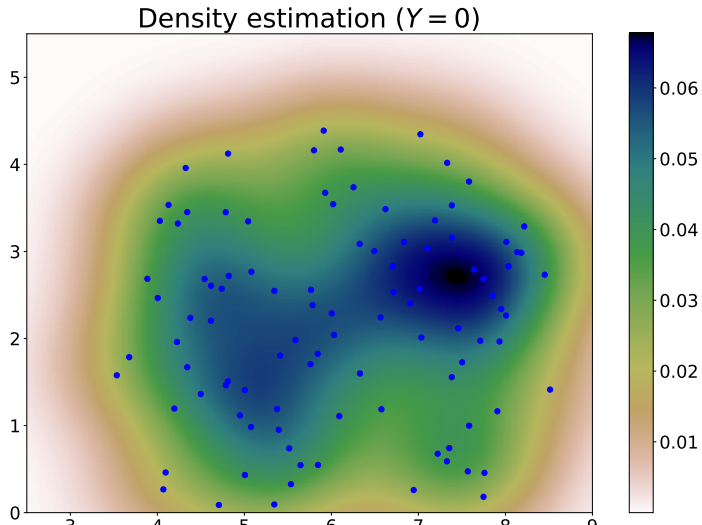
# Density estimation (IV)



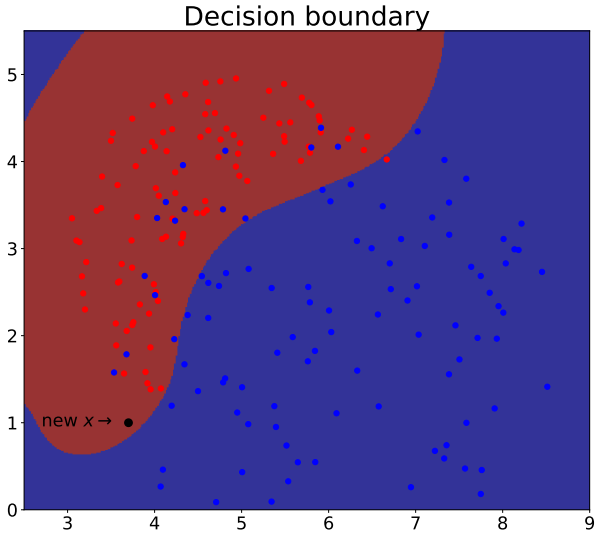
# Density estimation (V)



# Density estimation (VI)



# Density estimation (VII)



# Distance between distributions (I)

- **Question:** what does it mean for $\hat{f}_n$ to be "close" to $f$?
- what we really want: $\hat{\mu}_n$ close to $\mu$, that is,

$$\forall A \text{ Borel set}, \quad \int_A f(x)\mathrm{d}x \approx \int_A \hat{f}_n(x)\mathrm{d}x\,.$$

**Definition:** let $P$ and $Q$ be two probability measures on a sigma algebra $\mathcal{F}$ of the sample space $\Omega$. We define the *total variation distance* between $P$ and $Q$ *("distance en variation totale")* by

$$d_{\mathrm{TV}}(P,Q) = \sup_{A\in\mathcal{F}} |P(A) - Q(A)|\,.$$

- this will be our criterion to judge if we built a suitable $\hat{f}_n$

# Distance between distributions (II)

- **Example:** empirical measure. Given $(x_1, \dots, x_n)$, we set

$$\forall A \text{ Borel}, \quad \mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{x_i \in A} .$$

- **Intuition:** count how many times $X$ falls into $A$
- in other words, $\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$ where $\delta_x$ is the Dirac in $x$
- **Remark:** the cumulative distribution function of $\mu_n$ is the empirical cumulative distribution function $F_n$
- **Problem:** if $\mu$ has a density with respect to the Lebesgue measure, $d_{\mathrm{TV}}(\mu_n, \mu) = 1$ (take $A = (x_1, \dots, x_n)$)

# Distance between distributions (III)

**Lemma (Scheffé, 1950):** Suppose that $P$ and $Q$ have densities $f$ and $g$ with respect to the Lebesgue measure. Then

$$d_{\mathrm{TV}}(P, Q) = \frac{1}{2} \|f - g\|_1 \, .$$

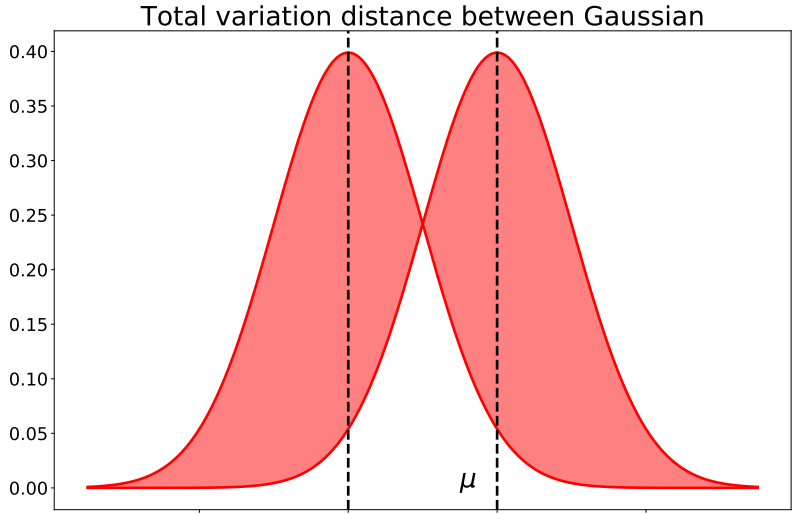
- **Good news:** it suffices to look at the density
- abuse of notation:

$$d_{\mathrm{TV}}(f, g) = d_{\mathrm{TV}}(F, G) \, .$$

- **Proof:** TD. □

# Distance between distributions (III)

- **Example:** $P \sim \mathcal{N}(0,1)$, $Q \sim \mathcal{N}(\mu,1)$

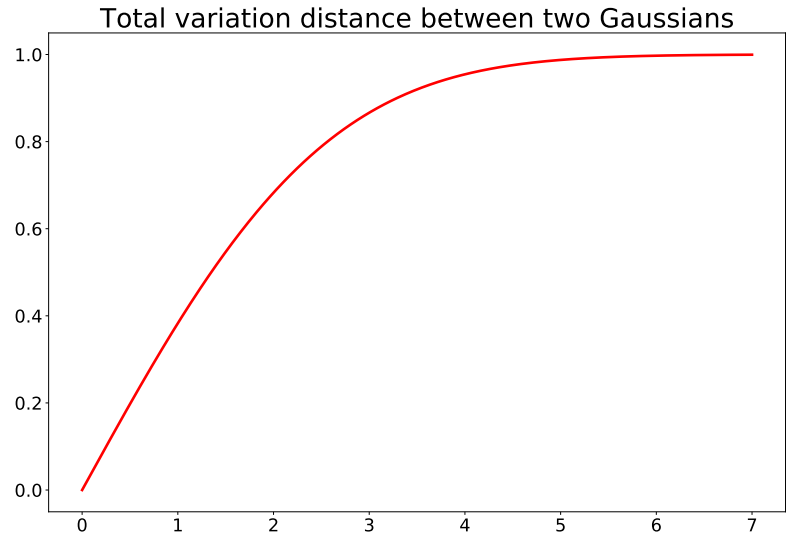


# Distance between distributions (IV)

- we can compute (fix $\mu > 0$):

$$
\begin{aligned}
d_{\mathrm{TV}}(P, Q) &= \frac{1}{2} \int \left| \frac{1}{\sqrt{2\pi}} \mathrm{e}^{\frac{-x^2}{2}} - \frac{1}{\sqrt{2\pi}} \mathrm{e}^{\frac{-(x-\mu)^2}{2}} \right| \mathrm{d}x \\
&= \frac{1}{2\sqrt{2\pi}} \left\{ \int_{-\infty}^{\mu/2} \left( \mathrm{e}^{\frac{-x^2}{2}} - \mathrm{e}^{\frac{-(x-\mu)^2}{2}} \right) \mathrm{d}x - \int_{\mu/2}^{+\infty} \left( \mathrm{e}^{\frac{-x^2}{2}} - \mathrm{e}^{\frac{-(x-\mu)^2}{2}} \right) \mathrm{d}x \right\} \\
&= \Phi\left(\frac{\mu}{2}\right) - \Phi\left(\frac{-\mu}{2}\right) \\
d_{\mathrm{TV}}(P, Q) &= \mathrm{erf}\left(\frac{\mu}{2\sqrt{2}}\right)
\end{aligned}
$$

# Distance between distributions (V)



# 34. Sliding window density estimation

# Sliding window (I)

- **Idea:** we know how to estimate $F$:

$$\forall x \in \mathbb{R}, \quad F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq x} \, .$$

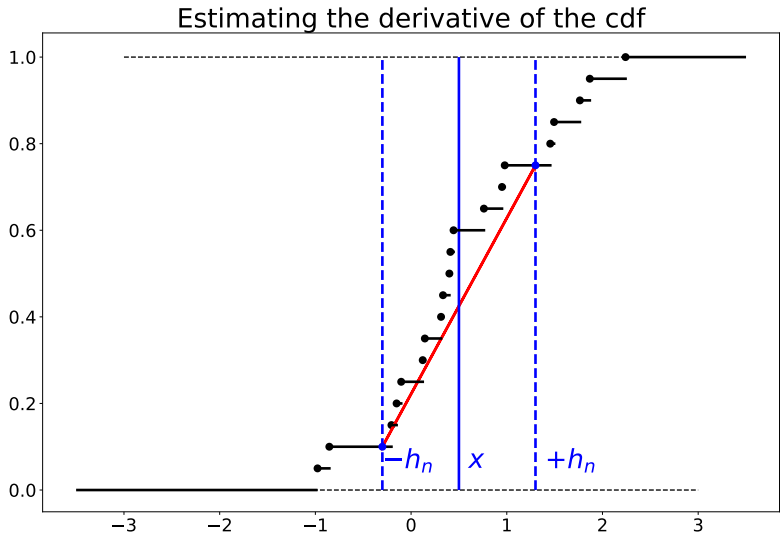
- why not take the derivative of this estimator?
- **Problem:** not differentiable...
- **Solution:** fix a bandwidth $h_n > 0$ and form a symmetric approximation of the derivative:

$$\forall x \in \mathbb{R}, \quad \hat{f}_n(x) = \frac{F_n(x + h_n) - F_n(x - h_n)}{2h_n} \, .$$

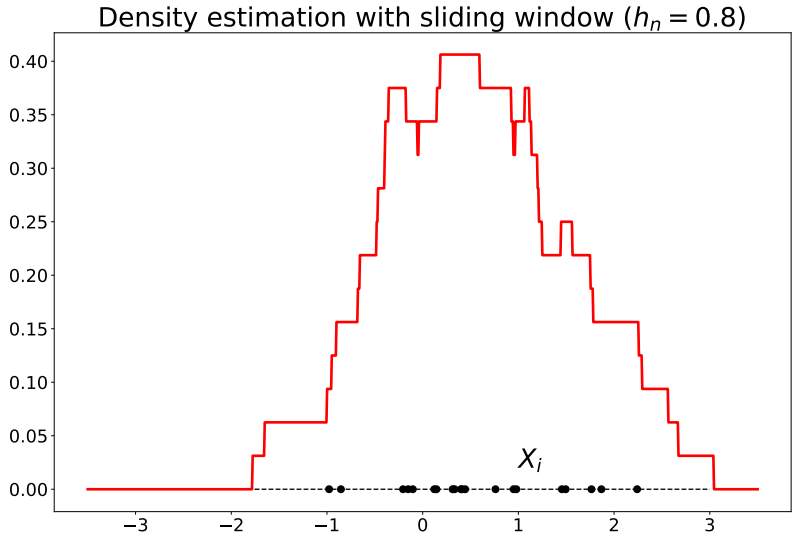
- indeed, for small $h$, Taylor expansion yields

$$\frac{1}{2h}(F(x) + hf(x) - F(x) + hf(x))) \approx f(x) \, .$$

# Sliding window (II)



# Sliding window (III)



# Sliding window (IV)

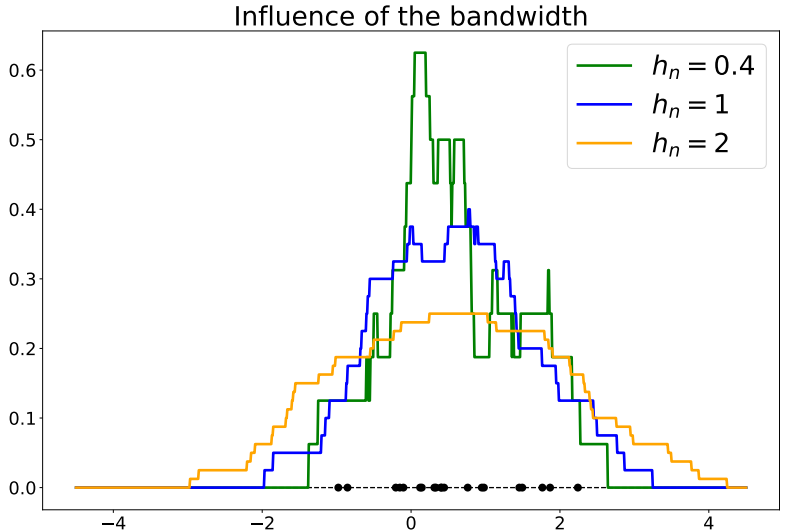
**Property:** For any $h_n > 0$, the estimator $\hat{f}_n$ is a probability density function.

**Proof:** Since $F_n$ is non-decreasing, $\hat{f}_n \geq 0$ a.s. Moreover,

$$\begin{aligned}\int \hat{f}_n(x)\mathrm{d}x &= \int \frac{F_n(x+h_n) - F_n(x-h_n)}{2h_n}\mathrm{d}x \\ &= \frac{1}{2nh_n}\sum_{i=1}^{n}\int \mathbb{1}_{X_i \in (x-h_n, x+h_n)}\mathrm{d}x \\ &= \frac{1}{2nh_n}\sum_{i=1}^{n} 2h_n \\ \int \hat{f}_n(x)\mathrm{d}x &= 1\end{aligned}$$

□

# Sliding window (V)



# Consistency of sliding window estimation (I)

**Proposition:** Let $(X_1, \dots, X_n)$ be an i.i.d. sample from a real-valued random variable $X$ with density $f$ with respect to the Lebesgue measure. Let $\hat{f}_n$ be the sliding window estimator with bandwidth $h_n > 0$. Suppose that $h_n \to 0$ and $nh_n \to +\infty$ when $n \to +\infty$. Then, for almost every $x \in \mathbb{R}$,

$$\mathbb{E}\left[(\hat{f}_n(x) - f(x))^2\right] \longrightarrow 0\,.$$

In particular, $\hat{f}_n$ is a *consistent* estimator of $f$ for $d_{\mathrm{TV}}$.

- **Intuition:** $h_n$ needs to be small (we want the local information about the derivative)...
- but **not so small**: we want enough points in $[x - h_n, x + h_n]$

# Consistency of sliding window estimation (II)

**Proof:** Bias-variance decomposition: let $x \in \mathbb{R}$,

$$\mathbb{E}\left[(\hat{f}_n(x) - f(x))^2\right] = \left(\mathbb{E}\left[\hat{f}_n(x)\right] - f(x)\right)^2 + \mathrm{Var}\left(\hat{f}_n(x)\right).$$

- bias term:

$$\begin{aligned}\mathbb{E}\left[\hat{f}_n(x)\right] &= \frac{\mathbb{E}\left[F_n(x+h_n) - F_n(x-h_n)\right]}{2h_n} \\ &= \frac{F(x+h_n) - F(x-h_n)}{2h_n} \\ \mathbb{E}\left[\hat{f}_n(x)\right] &= f(x) + o(h_n)\,.\end{aligned}$$

Since $h_n \to 0$,

$$\mathrm{bias}(\hat{f}_n(x)) = \mathbb{E}\left[\hat{f}_n(x)\right] - f(x) \longrightarrow 0\,.$$

# Consistency of sliding window estimation (III)

- variance term: we already noticed that

$$2nh_n\hat{f}_n(x) = \sum_{i=1}^{n} \mathbb{1}_{X_i \in (x-h_n, x+h_n]}\,.$$

- sum of i.i.d. Bernoulli with parameter

$$p = \mathbb{E}\left[\mathbb{1}_{X_i \in (x-h_n, x+h_n]}\right] = F(x+h_n) - F(x-h_n)\,.$$

- thus variance $np(1-p)$, and

$$\begin{aligned}
\mathrm{Var}\left(\hat{f}_n(x)\right) &= \frac{1}{4n^2h_n^2}\mathrm{Var}\left(2nh_n\hat{f}_n(x)\right) \\
&= \frac{np(1-p)}{4n^2h_n^2} \\
\mathrm{Var}\left(\hat{f}_n(x)\right) &\leq \frac{1}{4nh_n} \cdot \frac{F(x+h_n) - F(x-h_n)}{h_n} \longrightarrow 0 \qquad \square
\end{aligned}$$

# 35. Histogram density estimation

# Histogram density estimation (I)

- **Idea:** split $\mathbb{R}$ in bins, and count how many $X_i$ are in each bin
- formally, for each $n$ we define $(A_{k,n})_{k\in\mathbb{Z}}$ a partition of $\mathbb{R}$:

$$\forall n \in \mathbb{N}, \quad \bigcup_{k\in\mathbb{Z}} A_{k,n} = \mathbb{R} \text{ and } A_{k,n} \cap A_{k',n} = \emptyset \quad \forall k \neq k' .$$

- from this partition, we form

$$\forall k \in \mathbb{Z}, \forall x \in A_{k,n}, \quad \hat{f}_n(x) = \frac{|\{i \text{ s.t. } X_i \in A_{k,n}\}|}{n \cdot \lambda(A_{k,n})} ,$$

  where $\lambda$ is the Lebesgue measure
- **Important:** we restrict ourselves to *regular* partitions, that is,

$$\forall k \in \mathbb{Z}, \quad A_{k,n} = (kh_n, (k+1)h_n] ,$$

  where $h_n > 0$

# Histogram density estimation (II)

- in this case, $\lambda(A_{k,n}) = h_n$ for any $k \in \mathbb{Z}$, and

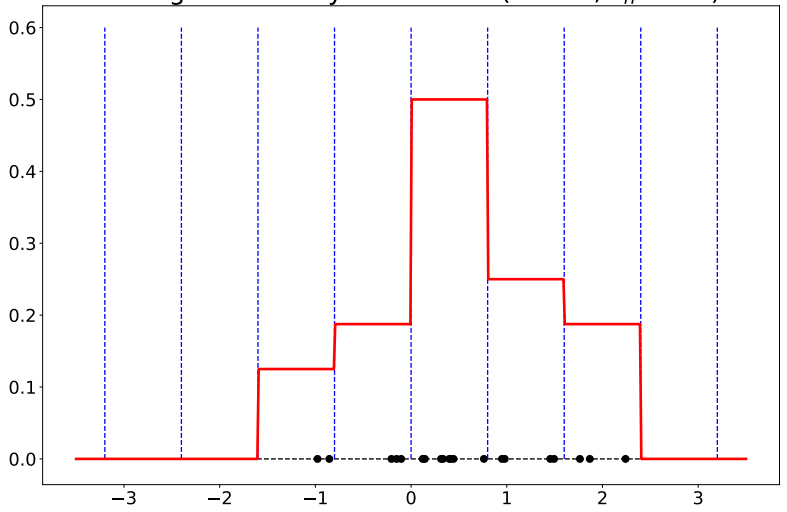
$$\forall k \in \mathbb{Z}, \forall x \in A_{k,n}, \quad \hat{f}_n(x) = \frac{|\{i \text{ s.t. } X_i \in A_{k,n}\}|}{nh_n} .$$

- let us check that $\hat{f}_n$ is a **density**:

$$\begin{aligned}
\int \hat{f}_n(x)\mathrm{d}x &= \sum_{k\in\mathbb{Z}} (\text{value on } A_{k,n}) \times (\text{length of } A_{k,n}) \\
&= \sum_{k\in\mathbb{Z}} \frac{|\{i \text{ s.t. } X_i \in A_{k,n}\}|}{nh_n} \times h_n \\
&= \frac{1}{n} \sum_{k\in\mathbb{Z}} |\{i \text{ s.t. } X_i \in A_{k,n}\}|
\end{aligned}$$

$$\int \hat{f}_n(x)\mathrm{d}x = 1 .$$

# Histogram density estimation (III)



# Histogram density estimation (IV)

**Proposition:** Let $(X_1, \dots, X_n)$ be an i.i.d. sample from a real-valued random variable $X$ with density $f$ with respect to the Lebesgue measure. Let $\hat{f}_n$ be the histogram estimator with bandwidth $h_n > 0$. Suppose that $h_n \to 0$ and $nh_n \to +\infty$ when $n \to +\infty$. Suppose additionally that $|f'(x)| \leq L$ for some $L > 0$. Then, for almost every $x \in \mathbb{R}$,

$$\mathbb{E}\left[(\hat{f}_n(x) - f(x))^2\right] \longrightarrow 0\,.$$

In particular, $\hat{f}_n$ is a consistent estimator of $f$ for $d_{\mathrm{TV}}$.

- **Intuition:** same as before, $h_n$ needs to be small in order to get the local behavior...
- but not too small!

# Histogram density estimation (V)

**Proof:** bias-variance decomposition again, let us look at the bias:

$$\mathbb{E}\left[\hat{f}_n(x)\right] = \mathbb{E}\left[\frac{|\{i \text{ s.t. } X_i \in A_{k,n}\}|}{nh_n}\right]$$
$$= \mathbb{E}\left[\frac{1}{nh_n}\sum_{i=1}^{n} \mathbb{1}_{X_i \in A_{k_x,n}}\right]$$

where $k_x \in \mathbb{Z}$ is such that $x \in A_{k_x,n}$.

$$\mathbb{E}\left[\hat{f}_n(x)\right] = \frac{1}{nh_n} \cdot n\left[F((k_x+1)h_n) - F(k_x h_n)\right]$$
$$= \frac{F((k_x+1)h_n) - F(k_x h_n)}{(k_x+1)h_n - k_x h_n}$$
$$= f(x')$$

for some $x' \in [k_x h_n, (k_x+1)h_n]$ (Rolle theorem).

# Histogram density estimation (VI)

Thus, there exists some $x'' \in [k_x h_n, (k_x + 1)h_n]$ such that

$$\begin{aligned}
\mathbb{E}\left[\hat{f}_n(x)\right] - f(x) &= f(x') - f(x) \\
&= f'(x'') \cdot (x' - x) \\
&\leq |f'(x'')| \cdot |x' - x| \\
\mathbb{E}\left[\hat{f}_n(x)\right] - f(x) &\leq L h_n
\end{aligned}$$

where we used the regularity assumption. We have proved:

$$\text{bias}(\hat{f}_n(x))^2 = \left(\mathbb{E}\left[\hat{f}_n(x)\right] - f(x)\right)^2 \leq L^2 h_n^2 .$$

# Histogram density estimation (VII)

Now let us look at the variance term.

$$\begin{aligned}\operatorname{Var}\left(\hat{f}_n(x)\right) &= \operatorname{Var}\left(\frac{1}{nh_n}\sum_{i=1}^{n}\mathbb{1}_{X_i\in[k_x h_n,(k_x+1)h_n]}\right)\\ &= \frac{1}{n^2h_n^2}\cdot np(1-p)\,,\end{aligned}$$

with

$$\begin{aligned}p &= \mathbb{P}\left(X_i\in[k_x h_n,(k_x+1)h_n]\right)\\ &= F((k_x+1)h_n) - F(k_x h_n)\\ &= h_n f(x')\,,\end{aligned}$$

for some $x'\in[k_x h_n,(k_x+1)h_n]$.

# Histogram density estimation (VIII)

We deduce

$$\begin{aligned}\mathrm{Var}\left(\hat{f}_n(x)\right) &= \frac{1}{n^2 h_n^2} \cdot nh_n f(x')(1 - h_n f(x')) \\ &\leq \frac{nh_n f(x')}{n^2 h_n^2} + \frac{nh_n^2 f(x')^2}{n^2 h_n^2} \\ \mathrm{Var}\left(\hat{f}_n(x)\right) &\leq \frac{f(x')}{nh_n} + \frac{f(x')^2}{n}\,.\end{aligned}$$

Finally,

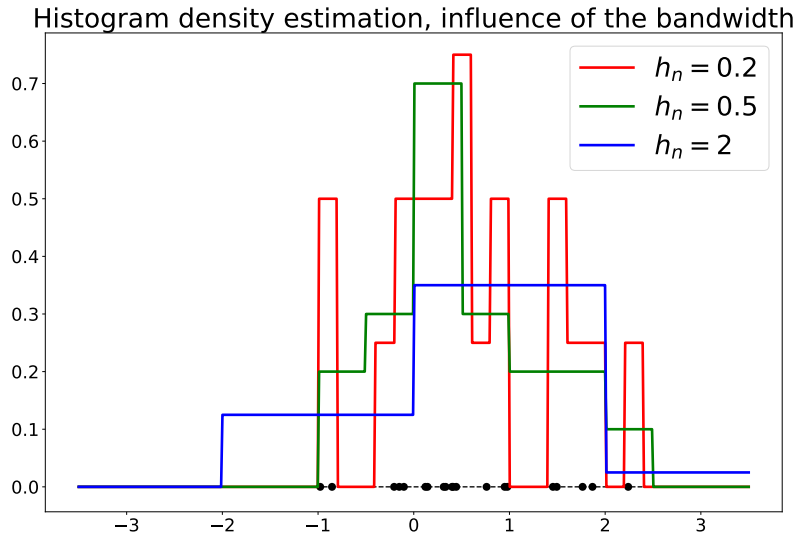
$$\mathbb{E}\left[(\hat{f}_n(x) - f(x))^2\right] \leq L^2 h_n^2 + \frac{f(x')}{nh_n} + \frac{f(x')^2}{n} \longrightarrow 0\,. \qquad \square$$

- **Remark:** last display gives us optimal bandwidth

$$h_n = \left(\frac{f(x')}{2nL^2}\right)^{\frac{1}{3}},$$

but we do not know $f(x')$ and $L$ in practice...

# Histogram density estimation (V)



# 36. Kernel density estimation

# Kernel density estimation (I)

- **Main tool:** kernel function

**Definition:** we say that a function $K : \mathbb{R} \to \mathbb{R}_+$ is a *kernel function ("noyau")* if

$$\int K(x)\mathrm{d}x = 1 \quad \text{and} \quad \int K^2(x)\mathrm{d}x < +\infty\,.$$

- for any $h > 0$, we also define

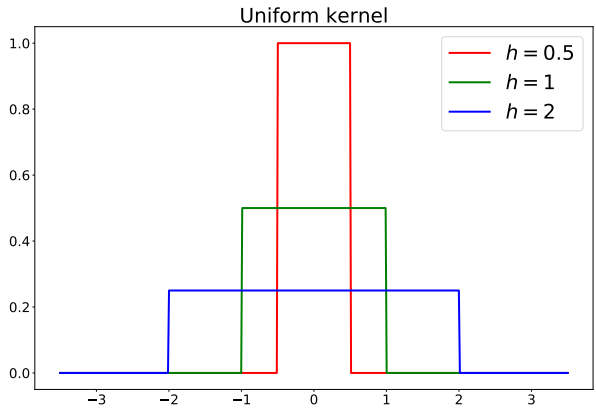
$$\forall x \in \mathbb{R}, \quad K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right)\,.$$

- **Intuition:** $K$ is a bump, $h$ adjusts the width of the bump

# Kernel density estimation (II)

- **Example:** the uniform kernel:

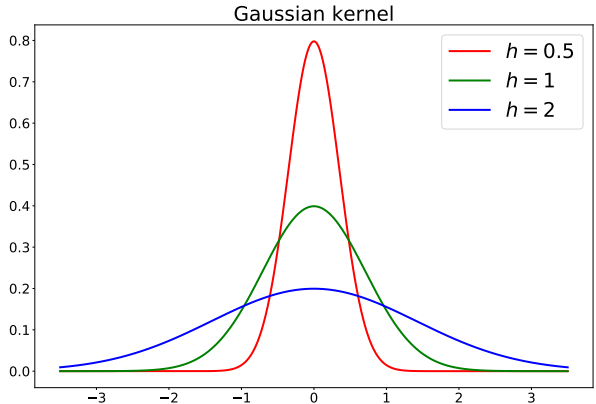
$$\forall x \in \mathbb{R}, \quad K(x) = \frac{1}{2}\mathbb{1}_{x\in[-1,1]}\,.$$



# Kernel density estimation (III)

- **Example:** the Gaussian kernel:

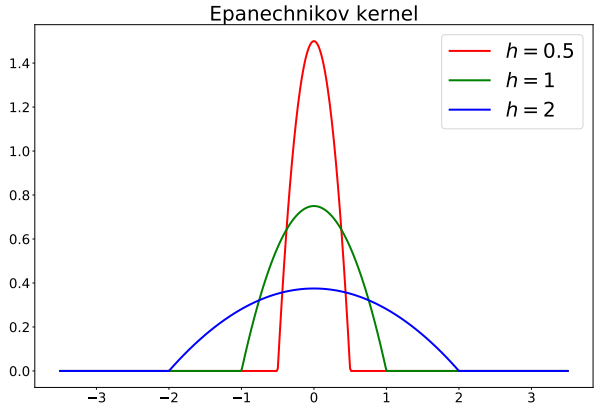
$$\forall x \in \mathbb{R}, \quad K(x) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left(\frac{-x^2}{2}\right).$$



# Kernel density estimation (IV)

- **Example:** Epanechnikov kernel:

$$\forall x \in \mathbb{R}, \quad K(x) = \frac{3}{4}(1 - x^2)_+ .$$



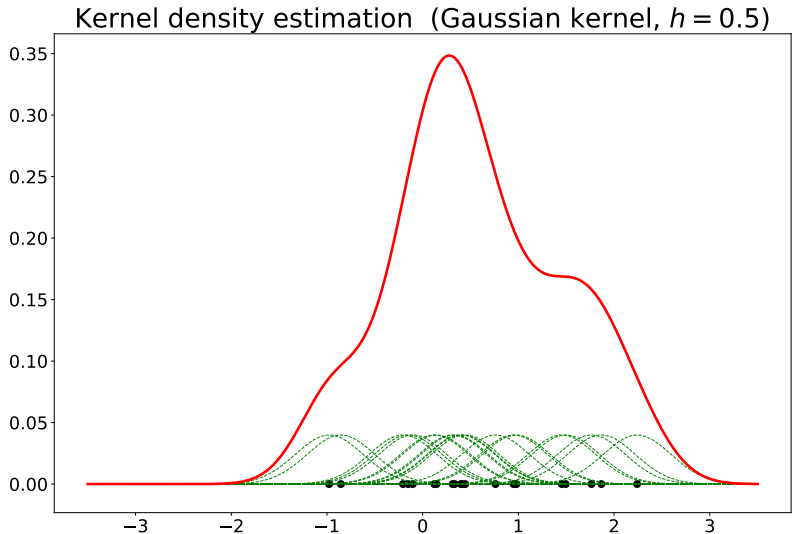
# Kernel density estimation (V)

**Definition:** Let $(X_1, \dots, X_n)$ be an i.i.d. sample from a real-valued random variable $X$. For any $x \in \mathbb{R}$, we define the kernel density estimator $\hat{f}_{n,K}$ by

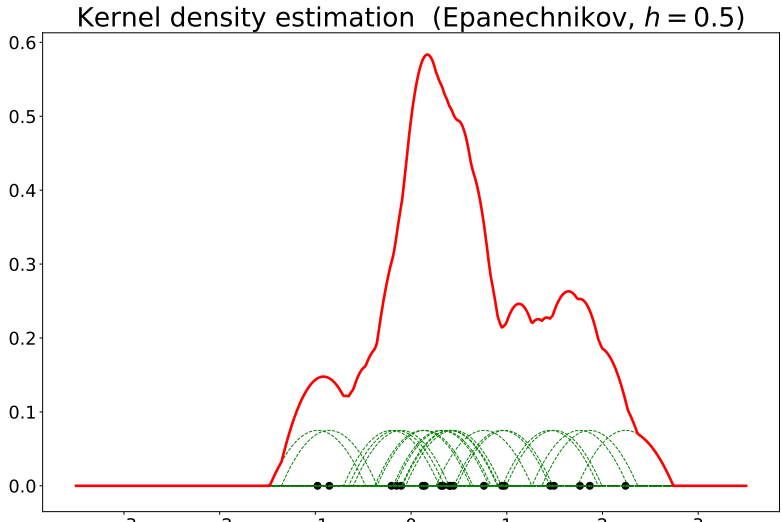
$$\hat{f}_{n,K}(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right).$$

- **Intuition:** replace the Diracs in the empirical measure by something smoother
- $\hat{f}_{n,K}$ is a density ($K_h$ sums to one by change of variable)
- **Remark:** if $K$ is the uniform kernel, we recover the sliding window estimator

# Kernel density estimation (VI)



# Kernel density estimation (VII)



# Kernel density estimation (VIII)

- **Idea:** take a kernel that incorporates the *a priori* information we have on $f$ (regularity, compact support, etc.)
- we also have a consistency result:

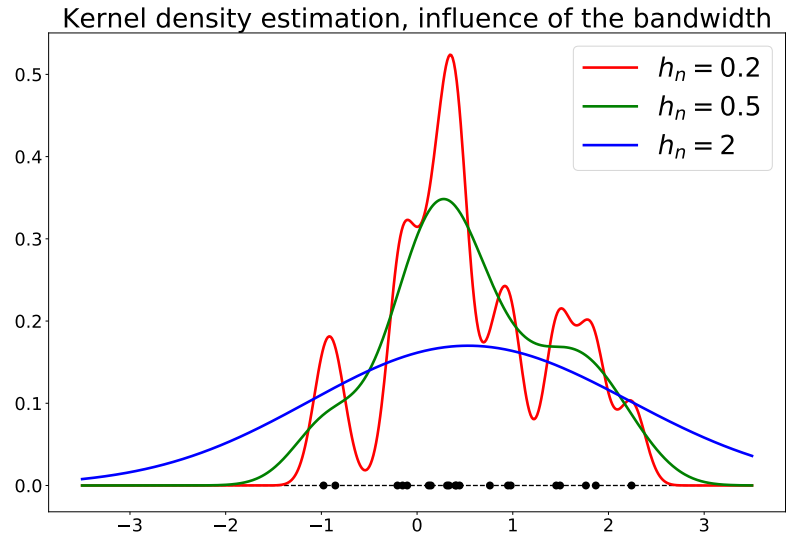
**Proposition:** Let $(X_1, \ldots, X_n)$ be an i.i.d. sample from a real-valued random variable $X$ with density $f$ with respect to the Lebesgue measure. Let $\hat{f}_{n,K}$ be the kernel estimator associated to kernel $K$ and bandwidth $h_n > 0$. Suppose that $h_n \to 0$ and $nh_n \to +\infty$ when $n \to +\infty$. Then

$$d_{\mathrm{TV}}(\hat{f}_{n,K}, f) \xrightarrow{\mathbb{P}} 0$$

when $n \to +\infty$.

- **Proof:** TD. $\square$

# Kernel density estimation (IX)



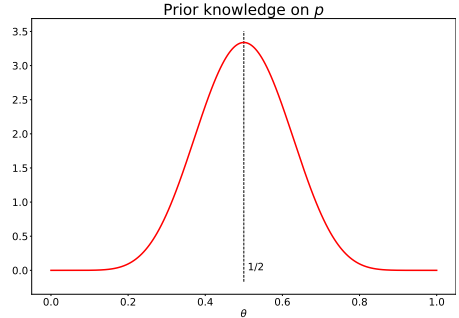
# 37. Introduction to Bayesian statistics

# Bayesian statistics (I)

- so far, we have studied **frequentist** statistics
- road map looked like:
  - collect data
  - make assumptions on the model (i.i.d., Gaussian, etc.)
  - set a statistical model, with an unknown but fixed parameter $\theta$
  - estimate $\theta$ / test for $\theta$ / find a confidence interval for $\theta$ / etc.
- **Example:** $\beta^\star$ in the linear model
- the red part is different in Bayesian statistics
- data generation is two-layer process:
  1. **generate** the parameter $\theta$ according to a distribution $\pi$
  2. sample $X_1, \ldots, X_n$ according to this parameter
- **Intuition:** quantify our *prior belief* on $\theta$

# Bayesian statistics (II)

- **Example:** proportion of females at birth (course #3)
- $X_i = 1$ if female, 0 otherwise
- **frequentist approach:** $X_i$ are i.i.d. $\mathcal{B}(\theta)$, where $\theta$ is a fixed, unknown number $\in [0, 1]$
- **Bayesian approach:** we are *pretty sure* that $\theta$ is going to be close to $1/2$
- *let us incorporate this prior knowledge as a distribution on p*
- intuitively, it is going to look like this:



# Bayesian statistics (III)

- **Question:** which distribution can we use as $\pi$?
- cannot use Gaussian, we have to be in $[0,1]$ a.s.

**Definition:** Let $a, b > 0$. We say that $Z$ has the *Beta distribution* with parameters $a$ and $b$ if $Z$ has density

$$f_{a,b}(t) = \frac{t^{a-1}(1-t)^{b-1}}{B(a,b)} \mathbb{1}_{t\in[0,1]}\,,$$

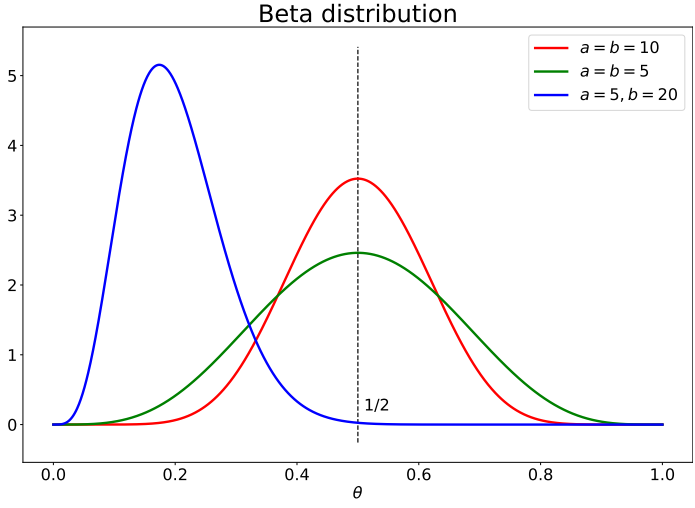
where $B(a,b)$ is a normalizing constant $(= \int_0^1 t^{a-1}(1-t)^{b-1}\mathrm{d}t)$.

- $Z \sim B(a,b)$ has mean $\frac{a}{a+b}$
- $Z \sim B(a,a)$ is centered in $1/2$ and has variance

$$\mathrm{Var}(Z) = \frac{1}{4(2a+1)}\,.$$

# Bayesian statistics (IV)

- high $a \Rightarrow$ high confidence that $\theta$ close to 1



# Bayesian statistics (V)

- now we have fixed a **prior distribution** $\pi = B(a, a)$ on the parameter $\theta$ *("distribution de probabilités a priori")*
- **What next?**
- second important idea of Bayesian inference: update our prior with respect to the observations
- **Intuition:** prior belief contains what you think about the data
- observing data is going to *move* the distribution:
  - maybe the data agrees with our belief: it will *reinforce* it
  - maybe it does not: it will *move away*
- in the end, we will obtain a **posterior distribution**
- **Spoiler alert:** the prior distribution is less and less important as we have more and more data
- **Question:** how do we do this??

# Bayesian statistics (VI)

**Theorem (Bayes):** For any given event $A$ and $B$, such that $\mathbb{P}(B) > 0$, we have

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B \mid A)\,\mathbb{P}(A)}{\mathbb{P}(B)}.$$

Thomas Bayes (1701–1761)

# Bayesian statistics (VII)

- **Main idea:** write distribution of parameter conditionally to data using Bayes theorem

$$\pi(\theta|X_1,\dots,X_n) = \frac{p(X_1,\dots,X_n|\theta)\pi(\theta)}{\int_{\theta\in\Theta} p(X_1,\dots,X_n|\theta)\mathrm{d}\pi(\theta)}$$

- $\pi(\theta|X_1,\dots,X_n)$ is the *posterior distribution* of $\theta$: the distribution of $\theta$ given the observations (this is what we are looking for!)
- $p(X_1,\dots,X_n|\theta)$ is the *likelihood* of the observations: we already know how to compute this
- $\pi(\theta)$ is the *prior distribution*: we decide what it is
- the denominator generally does not matter (it does not depend on $\theta$), and we will use the notation $\propto$ ("proportional to")

# Bayesian statistics (VIII)

- **Back to the example:** recall that we chose

$$\pi(\theta) \propto \theta^{a-1}(1-\theta)^{a-1},$$

for some $a > 0$

- the likelihood of $X_1, \ldots, X_n$ is given by

$$\begin{aligned} p(x_1, \ldots, x_n|\theta) &= \prod_{i=1}^{n} \mathbb{P}_\theta\left(X_i = x_i\right) \\ &= \prod_{i=1}^{n} \theta^{\mathbb{1}_{x_i=1}}(1-\theta)^{\mathbb{1}_{x_i=0}} \\ &= \theta^{\sum_{i=1}^{n} \mathbb{1}_{x_i=1}}(1-\theta)^{\sum_{i=1}^{n} \mathbb{1}_{x_i=0}} \\ p(x_1, \ldots, x_n|\theta) &= \theta^{\sum_{i=1}^{n} \mathbb{1}_{x_i=1}}(1-\theta)^{n-\sum_{i=1}^{n} \mathbb{1}_{x_i=1}} \end{aligned}$$

# Bayesian statistics (IX)

- **In definitive,**

$$\begin{aligned}\pi(\theta|X_1,\dots,X_n) &\propto p(X_1,\dots,X_n|\theta)\cdot\pi(\theta)\\ &\propto \theta^{\sum_{i=1}^n \mathbb{1}_{X_i=1}}(1-\theta)^{n-\sum_{i=1}^n \mathbb{1}_{X_i=1}}\cdot\theta^{a-1}(1-\theta)^{a-1}\\ \pi(\theta|X_1,\dots,X_n) &\propto \theta^{a+\sum_{i=1}^n \mathbb{1}_{X_i=1}-1}(1-\theta)^{a+n-\sum_{i=1}^n \mathbb{1}_{X_i=1}-1}\end{aligned}$$

- in particular, there is no need to compute the proportionality constant: we recognize a Beta distribution

$$\boxed{\theta\sim B\left(a+\sum_{i=1}^n \mathbb{1}_{X_i=1}, a+n-\sum_{i=1}^n \mathbb{1}_{X_i=1}\right)}$$

conditionally to $X_1,\dots,X_n$

- **Remark (i):** we have a *distribution* for the parameter! much more information than in the frequentist setting

# Bayesian statistics (X)

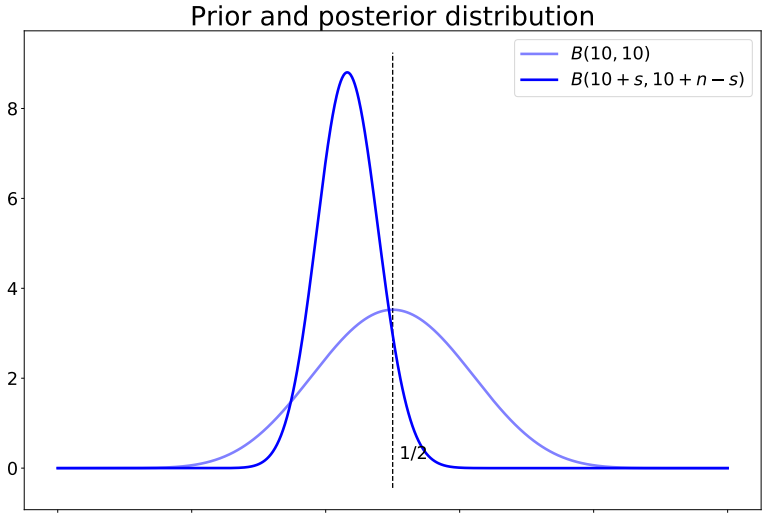
- **Remark (ii):** we can see how the observations are going to *modify* our prior belief
- for instance, many observations $= 1$ means that

$$a + \sum_{i=1}^{n} \mathbb{1}_{X_i=1} \gg a + n - \sum_{i=1}^{n} \mathbb{1}_{X_i=1} \, .$$

- $\Rightarrow$ *skew the distribution to the right*, moving away from our prior belief
- **Vocabulary:** when both the prior distribution and the posterior distribution belong to the same family, we say that they are *conjugate*

# Bayesian statistics (XI)

- example with $n = 100$ and 42 females ($= 1$)



# 38. Bayesian estimation

# Bayes estimator (I)

- we have obtained a **posterior distribution** for the parameter
- we can do many things with that!
- **Why not "estimation"?**
- **Motivation:** a distribution is complicated, practitioners often want a *single number*
- one possibility: return the mean of the posterior distribution
- **Bayes estimator:**

$$\begin{aligned}\hat{\theta}^{\text{Bayes}} &= \mathbb{E}_{\pi(\theta|X_1,\ldots,X_n)}[\theta] \\ &= \int_{\theta\in\Theta} \theta \ \mathrm{d}\pi(\theta|X_1,\ldots,X_n)\end{aligned}$$

# Bayes estimator (II)

- **Example:** back to the Bernoulli-Beta case
- we know that the expectation of a $B(\alpha, \beta)$ distribution is given by

$$\frac{\alpha}{\alpha + \beta}.$$

- recall that the posterior distribution we obtained was

$$\theta \sim B\left(a + \sum_{i=1}^{n} \mathbb{1}_{X_i=1}, a + n - \sum_{i=1}^{n} \mathbb{1}_{X_i=1}\right),$$

  conditionally to $X_1, \ldots, X_n$
- we obtain

$$\hat{\theta}^{\text{Bayes}} = \frac{a + \sum_{i=1}^{n} \mathbb{1}_{X_i=1}}{2a + n}.$$

# Bayes estimator (III)

- **Recall:** we obtained
$$\hat{\theta}^{\text{Bayes}} = \frac{a + \sum_{i=1}^{n} \mathbb{1}_{X_i=1}}{2a + n}.$$
- **What is happening here?**
- when $n$ is very small with respect to $a$, we have
$$\hat{\theta}^{\text{Bayes}} \approx \frac{1}{2},$$
we recover the mean of our prior
- when $n$ is very large with respect to $a$, we have
$$\hat{\theta}^{\text{Bayes}} \approx \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i=1},$$
we end up with the method of moments estimator

# Maximum a posteriori (I)

- another possibility: look at the *most probable value of the parameter*
- **Maximum a posteriori (MAP)** returns the *mode* of the posterior distribution:

$$\begin{aligned}\hat{\theta}^{\text{MAP}} &\in \arg\max_{\theta\in\Theta} \pi(\theta|X_1,\ldots,X_n) \\ &= \arg\max_{\theta\in\Theta} \frac{p(X_1,\ldots,X_n|\theta)\pi(\theta)}{\int_{\theta\in\Theta} p(X_1,\ldots,X_n|\theta)\mathrm{d}\pi(\theta)} \\ &= \arg\max_{\theta\in\Theta} p(X_1,\ldots,X_n|\theta)\pi(\theta)\,.\end{aligned}$$

- once again, the constant does not count!

# Maximum a posteriori (II)

- **Back to our example:** we had obtained

$$\theta \mid X_1, \ldots, X_n \sim B\left(a + \sum_{i=1}^{n} \mathbb{1}_{X_i=1}, a + n - \sum_{i=1}^{n} \mathbb{1}_{X_i=1}\right),$$

that is,

$$\pi(\theta|X_1, \ldots, X_n) \propto \theta^{a+\sum_{i=1}^{n} \mathbb{1}_{X_i=1}-1}(1-\theta)^{a+n-\sum_{i=1}^{n} \mathbb{1}_{X_i=1}-1}.$$

- one can show that the mode of a $B(\alpha, \beta)$ distribution is attained at

$$\frac{\alpha - 1}{\alpha + \beta - 2},$$

whenever $\alpha, \beta > 1$

# Maximum a posteriori (III)

- in our case, we obtain
$$\hat{\theta}^{\text{MAP}} = \frac{a - 1 + \sum_{i=1}^{n} \mathbb{1}_{X_i=1}}{2(a-1) + n} .$$
- **What is happening here?**
- two cases:
  - when $n$ is large, then
$$\hat{\theta}^{\text{MAP}} \approx \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i=1} ,$$
    we end up with the frequentist estimator!
  - when $a$ is large,
$$\hat{\theta}^{\text{MAP}} \approx \frac{1}{2} ,$$
    the mode of our prior

# Maximum a posteriori (IV)

- **What happens with another prior?**
- for instance, let us suppose that we are completely unsure about $\theta$, and set

$$\theta \sim \mathcal{U}([0,1]) \ .$$

- in other words, $\pi(\theta) \propto 1$
- then

$$\pi(\theta|X_1, \ldots, X_n) \propto p(X_1, \ldots, X_n|\theta) \, .$$

- in that case, MAP is just maximum likelihood estimation!
- **Remark:** the uniform prior is often called a *non-informative* prior

# Maximum a posteriori (V)

- **What happens when $\Theta$ is not bounded?**
- there is no uniform distribution on $\mathbb{R}$!
- let us do it anyway and consider $\pi(\theta) \propto 1$ on $\mathbb{R}$
- we sample $X_1, \ldots, X_n$ i.i.d. $\mathcal{N}(\theta, 1)$ conditionally to $\theta$
- let us compute the posterior distribution:

$$\begin{aligned}
\pi(\theta|X_1, \ldots, X_n) &\propto p(X_1, \ldots, X_n|\theta)\pi(\theta) \\
&\propto \prod_{i=1}^{n} \exp\left(\frac{-(X_i - \theta)^2}{2}\right) \\
&\propto \exp\left(\frac{-1}{2}\sum_{i=1}^{n}(X_i - \theta)^2\right) \\
\pi(\theta|X_1, \ldots, X_n) &\propto \exp\left(\frac{-1}{2}\sum_{i=1}^{n}(X_i^2 - 2X_i\theta + \theta^2)\right)
\end{aligned}$$

# Maximum a posteriori (VI)

$$\begin{aligned}
\pi(\theta|X_1,\dots,X_n) &\propto \exp\left(\frac{-1}{2}\sum_{i=1}^{n}(X_i^2 - 2X_i\theta + \theta^2)\right)\\
&\propto \exp\left(\frac{-1}{2}(n\theta^2 - 2\theta n\overline{X}_n)\right)\\
&\propto \exp\left(\frac{-n}{2}(\theta^2 - 2\overline{X}_n\theta)\right)\\
\pi(\theta|X_1,\dots,X_n) &\propto \exp\left(\frac{-n}{2}(\theta - \overline{X}_n)^2\right)
\end{aligned}$$

- we recognize a **Gaussian distribution**:

$$\theta \sim \mathcal{N}\left(\overline{X}_n, \frac{1}{n}\right),$$

conditionally to $X_1,\dots,X_n$

# Maximum a posteriori (VII)

- **What is the MAP in this case?**
- the mode of a Gaussian is also the mean, thus

$$\hat{\theta}^{\text{MAP}} = \overline{X}_n \,.$$

- non-informative prior: MAP coincides with maximum likelihood estimation
- even though $\pi(\theta)$ was not "well-defined"!
- we say that $\pi(\theta)$ is an **improper prior**

# 39. Credible intervals

# Credible intervals (I)

- **Credible intervals** are the analogous of *confidence intervals* for Bayesians
- But very different in spirit!
- **Key idea:** the parameter is random and the data is fixed
- for frequentists, it is the *opposite*

**Definition:** Let $\theta$ be the parameter of interest and $\pi(\theta|X_1, \dots, X_n)$ the posterior distribution of $\theta$ conditionally to the data. Let $\alpha \in [0, 1]$. We say that $I$ is a *credible interval* for $\theta$ at level $1 - \alpha$ if

$$\mathbb{P}(\theta \in I \,|\, X_1, \dots, X_n) \geq 1 - \alpha\,.$$

- **Remark:** credible intervals are not unique

# Credible intervals (II)

- we can choose the *narrowest* interval containing the mode of the distribution
- this is called **highest posterior density interval (HPDI)**
- if $I = [x, y]$, generally we require
  - mode $\in I$
  - $f(x) = f(y)$
  - $F(y) - F(x) = 1 - \alpha$
- difficult to solve in general!
- **Example (i):** Gaussian example, we obtained

$$\theta \sim \mathcal{N}\left(\overline{X}_n, \frac{1}{n}\right),$$

  conditionally to $X_1, \ldots, X_n$
- the condition on the density gives

$$x = \overline{X}_n - z \quad \text{and} \quad y = \overline{X}_n + z\,.$$

# Credible intervals (III)

- then we write

$$\begin{aligned}\mathbb{P}\left(\theta \leq t \,|\, X_1, \ldots, X_n\right) &= \mathbb{P}\left(\mathcal{N}\left(\overline{X}_n, 1/n\right) \leq t \,\middle|\, X_1, \ldots, X_n\right) \\ &= \Phi(\sqrt{n}(t - \overline{X}_n))\end{aligned}$$

- we deduce

$$F(y) - F(x) = \Phi(z\sqrt{n}) - \Phi(-z\sqrt{n}) = 2\Phi(z\sqrt{n}) - 1\,.$$

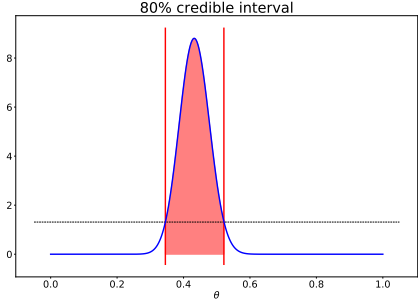
- finally, we recover $z = \frac{z_{\alpha/2}}{\sqrt{n}}$
- we have obtained the same interval than in the frequentist case
- this is not always the case, and sometimes HPDI are **much more difficult to compute**

# Credible intervals (IV)

- **Example (ii):** back to the first example
- we obtained

$$\theta \sim B\left(a + \sum_{i=1}^{n} \mathbb{1}_{X_i=1}, a + n - \sum_{i=1}^{n}\right) .$$

- Beta distributions are *unimodal*
- but even solving $f(x) = f(y)$ is not trivial!
- we resort to *approximate* algorithms



# Quick summary

- the Bayesian framework is very *flexible*
- allows us to incorporate **prior knowledge** about the parameter
- concept such as estimation and confidence intervals are easier to understand...
- but often very hard to compute!
- **Bonus:** when sample size is large, frequentists and Bayesians agree

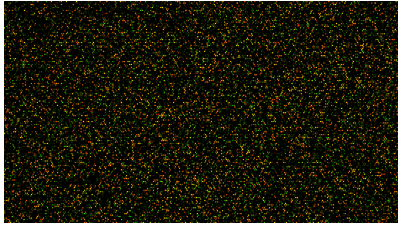
**Theorem (Bernstein-von Mises, 1949):** Suppose that $X_1, \ldots, X_n$ are i.i.d. $P_{\theta_0}$. Suppose that the maximum likelihood estimator converges towards $\theta_0$. If the prior distribution puts positive mass at $\theta_0$, then the asymptotic distribution of the maximum a posteriori does not depend on the prior.

- in particular, $\hat{\theta}^{\text{MAP}}$ and $\hat{\theta}^{\text{MLE}}$ have the same asymptotic distribution

# 40. Curse of dimensionality

# High-dimensional data

- **Example:** genomics
- microarray data = expression state of a large number of genes (sometimes the whole genome!)

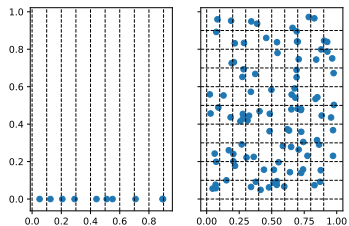


- informally, activity of the genes for each sample
- typically $n \approx 10$ and $d \approx 10^5$

$$d \gg n$$

# Curse of dimensionality (I)

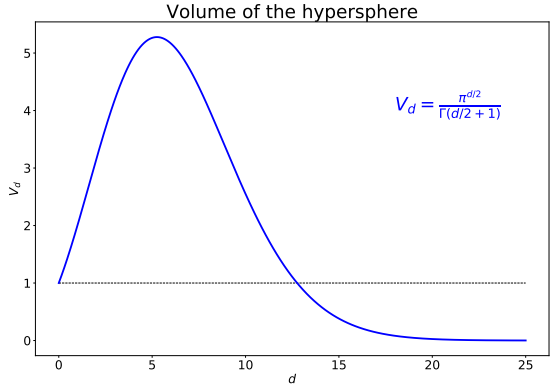
- **Example:** density estimation
- we want to be local enough: sample points that are at distance $\varepsilon$




- in dimension $d$, we need $\left(\frac{1}{\varepsilon}\right)^d$ points
- This is not realistic!

# Curse of dimensionality (II)

- other problem: most of the mass of hypercubes is near the boundary
- **Equivalent statement:** uniformly sampled points are more likely to be close to a face than another point (!)




- Our intuition is wrong!

# Curse of dimensionality (III)

- **From now on I assume that the data are centered**
- **Reminder:** Gaussian linear model

$$Y = X\beta^\star + \varepsilon \in \mathbb{R}^n,$$

  with $X \in \mathbb{R}^{n \times d}$ and $\beta^\star \in \mathbb{R}^d$
- we have seen the least squares estimator (courses 5 and 6):

$$\hat{\beta}^{\mathrm{LS}} = (X^\top X)^{-1} X^\top Y.$$

- **Problems:** $X^\top X \in \mathbb{R}^{d \times d}$ is a very large matrix!
- when $d \gg n$, $X^\top X$ is not invertible! Indeed, $X^\top X$ is not full rank:

$$\mathrm{rank}\left(X^\top X\right) \leq \min(\mathrm{rank}\left(X\right), \mathrm{rank}\left(X^\top\right)) \leq n < d.$$

- we have seen one solution: generalized inverse
- another is **ridge regression** (also called Tikhonov regularization)

# 41. Ridge regression

# Ridge regression (I)

- **Idea:** add a perturbation so that $X^\top X$ is invertible
- more precisely, consider

$$\hat{\beta}^{\text{Ridge}} = (X^\top X + \lambda \, \mathrm{I}_d)^{-1} X^\top Y \,,$$

  for some $\lambda > 0$ (Hoerl and Kennard, Ridge regression: Biased estimation for nonorthogonal problems, 1970)
- add a "ridge" on the diagonal of $X^\top X$
- as seen in TD 6, $\hat{\beta}^{\text{Ridge}}$ is solution of

$$\hat{\beta}^{\text{Ridge}} \in \underset{\beta \in \mathbb{R}^d}{\arg\min} \left\{ \|Y - X\beta\|^2 + \lambda \, \|\beta\|^2 \right\} . \tag{1}$$

- **Bonus:** equivalent formulation:

$$\hat{\beta}^{\text{Ridge}} \in \underset{\beta \in \mathbb{R}^d}{\arg\min} \, \|Y - X\beta\|^2$$
$$\text{subject to } \|\beta\|^2 \leq t \,.$$

# Ridge regression (II)

**Proof of (1):**

- let us set $R(\beta) = \|Y - X\beta\|^2 + \lambda \|\beta\|^2$
- $R$ is a **convex** function
- we compute

$$\begin{aligned}\nabla R(\beta) &= \nabla \left[(Y - X\beta)^\top (Y - X\beta) + \lambda \beta^\top \beta\right] \\ &= \nabla \left[Y^\top Y - 2\beta^\top X^\top Y + \beta^\top X^\top X \beta + \lambda \beta^\top \beta\right] \\ \nabla R(\beta) &= -2X^\top Y + 2X^\top X\beta + 2\lambda\beta\end{aligned}$$

- note that

$$\nabla R(\beta) = 0 \quad \Leftrightarrow \quad (X^\top X + \lambda \mathrm{I}_d)\beta = X^\top Y\,.$$

- since $X^\top X + \lambda \mathrm{I}_d$ is invertible, we can conclude. $\square$

# Ridge regression (III)

- **Question:** why is $X^\top X + \lambda \mathsf{I}_d$ invertible?
- the matrix $X^\top X + \lambda \mathsf{I}_d$ is singular iff

$$\det\left(X^\top X + \lambda \mathsf{I}_d\right) = 0\,.$$

- that is, $\chi_{X^\top X}(-\lambda) = 0$, where $\chi_M$ denotes the **characteristic polynomial of matrix** $M$
- in other words,

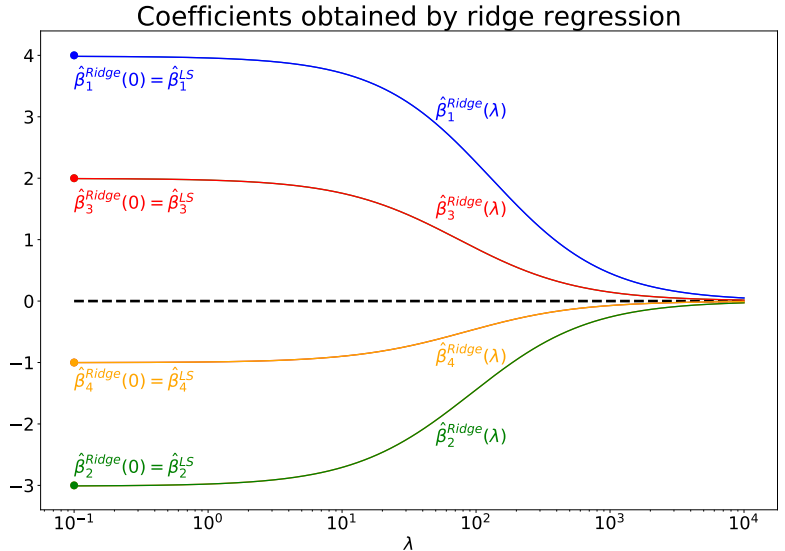
$$\det\left(X^\top X + \lambda \mathsf{I}_d\right) = 0 \quad \Leftrightarrow \quad -\lambda \in \mathrm{Spec}\left(X^\top X\right)\,.$$

- since $X^\top X$ is positive semi-definite,

$$\mathrm{Spec}\left(X^\top X\right) \subseteq \mathbb{R}_+\,.$$

- **Conclusion:** for any $\lambda > 0$, $X^\top X + \lambda \mathsf{I}_d$ is invertible

# Ridge regression (IV)



# Ridge regression (V)

- let us prove the shrinking in the case of **orthonormal inputs**
- we make the assumption that $X^\top X = \mathrm{I}_d$
- then

$$\begin{aligned}\hat{\beta}^{\text{Ridge}} &= (X^\top X + \lambda \mathrm{I}_d)^{-1} X^\top Y \\ &= (\mathrm{I}_d + \lambda \mathrm{I}_d)^{-1} X^\top Y \\ &= \frac{1}{1+\lambda}(X^\top X)^{-1} X^\top Y \\ \hat{\beta}^{\text{Ridge}} &= \frac{1}{1+\lambda}\hat{\beta}^{\text{LS}}\end{aligned}$$

- in particular, since $\lambda > 0$,

$$\boxed{\forall 1 \leq j \leq d, \quad \left|\hat{\beta}_j^{\text{Ridge}}\right| < \left|\hat{\beta}_j^{\text{LS}}\right|}$$

- ridge regression shrinks the coefficients!

# Ridge regression (VI)

- **What happens for non-orthonormal inputs?**
- back to singular value decomposition:

$$X = U\Sigma V^\top \quad \text{with} \quad U \in \mathcal{O}(n),\ \ \Sigma \in \mathrm{Diag}(n,d),\ \text{and}\ V \in \mathcal{O}(d)\,.$$

- that is, $UU^\top = U^\top U = \mathsf{I}_n$, $VV^\top = V^\top V = \mathsf{I}_d$, and $\Sigma$ is diagonal with

$$\Sigma_{jj} = \sigma_j \geq 0\,,$$

  the *singular values*
- **Remark:** the $\sigma_j^2$ are the eigenvalues of $X^\top X$
- there is unicity if we take

$$\sigma_1 \geq \cdots \geq \sigma_d\,.$$

# Ridge regression (VII)

- we can use SVD to obtain the **spectral formulation** of $\hat{\beta}^{\text{LS}}$
- recall that $X = U\Sigma V^\top$ and deduce successively $X^\top = V\Sigma^\top U^\top$,

$$X^\top X = V\Sigma^\top \Sigma V^\top, \quad X^\top X + \lambda \mathrm{I}_d = V(\Sigma^\top \Sigma + \lambda \mathrm{I}_d)V^\top,$$

and $\left(X^\top X + \lambda \mathrm{I}_d\right)^{-1} = V(\Sigma^\top \Sigma + \lambda \mathrm{I}_d)^{-1} V^\top$.

- we write

$$\begin{aligned}
\hat{\beta}^{\text{LS}} &= (X^\top X)^\dagger X^\top Y \\
&= (V\Sigma^\top U^\top U \Sigma V^\top)^\dagger V \Sigma^\top U^\top Y \\
&= (V\Sigma^\top \Sigma V^\top)^\dagger V \Sigma^\top U^\top Y \\
\hat{\beta}^{\text{LS}} &= V(\Sigma^\top \Sigma)^\dagger \Sigma^\top U^\top Y
\end{aligned}$$

- thus $\boxed{\hat{\beta}^{\text{LS}} = VDU^\top Y}$, with $D \in \mathbb{R}^{d \times n}$ diagonal such that

$$D_{jj} = \begin{cases} \frac{1}{\sigma_j} & \text{if} \quad \sigma_j > 0 \\ 0 & \text{otherwise} \end{cases}$$

# Ridge regression (VIII)

- as for the ridge estimator:

$$\begin{aligned}\hat{\beta}^{\text{Ridge}} &= (X^\top X + \lambda \mathrm{I}_d)^{-1} X^\top Y \\ &= (V\Sigma^\top \Sigma V^\top + \lambda V V^\top)^{-1} V \Sigma^\top U^\top Y \\ &= V(\Sigma^\top \Sigma + \lambda \mathrm{I}_d)^{-1} V^\top V \Sigma^\top U^\top Y \\ \hat{\beta}^{\text{Ridge}} &= V(\Sigma^\top \Sigma + \lambda \mathrm{I}_d)^{-1} \Sigma^\top U^\top Y\end{aligned}$$

- thus $\boxed{\hat{\beta}^{\text{Ridge}} = VD'U^\top Y}$, with $D' \in \mathbb{R}^{d \times n}$ diagonal such that

$$D'_{jj} = \frac{\sigma_j}{\sigma_j^2 + \lambda}\,.$$

- since $\lambda > 0$, $D'_{jj} < D_{jj}$: in these new coordinates, the ridge estimator shrinks the coefficients

# Ridge regression (IX)

- **Another bonus of the spectral formulation:** computational cost
- original formulation:
$$\hat{\beta}^{\text{Ridge}} = (X^\top X + \lambda \, \mathrm{I}_d)^{-1} X^\top Y \,,$$
thus we need to invert a $d \times d$ matrix
- this costs $\mathcal{O}\left(d^3\right)$ operations (in truth a bit less): this is not possible if $d \approx 10^5$!
- spectral formulation:
$$\hat{\beta}^{\text{Ridge}} = VD'U^\top Y \,.$$
- we are reduced to the cost of computing the SVD of $X$, that is, $\mathcal{O}\left(nd^2\right)$ (in the regime $n \ll d$)
- since $n \ll d$, $nd^2 \ll d^3$: this is a **huge** computational gain

# Bias of $\hat{\beta}^{\text{Ridge}}$ (I)

**Lemma:** Assume that we are in the Gaussian linear model *(fixed design)* $Y = X\beta^\star + \varepsilon$, with $\varepsilon \sim \mathcal{N}\left(0, \sigma^2 \mathsf{I}_n\right)$. Then the bias of the ridge estimator is given by

$$\mathbb{E}[\hat{\beta}^{\text{Ridge}}] - \beta^\star = -\lambda(X^\top X + \lambda \mathsf{I}_d)^{-1}\beta^\star \,.$$

In particular, if $X = U\Sigma V^\top$ is the SVD of $X$ as before,

$$\left\|\mathbb{E}[\hat{\beta}^{\text{Ridge}}] - \beta^\star\right\|^2 = \lambda^2 \sum_{j=1}^{d} \frac{1}{(\sigma_j^2 + \lambda)^2} \cdot \left(V_j^\top \beta^\star\right)^2 \,,$$

where $V_j$ is the $j$th column of $V$.

- **Intuition:** since for any $\lambda > 0$, $\mathbb{E}[\hat{\beta}^{\text{Ridge}}] \neq \beta^\star$, ridge regression is biased
- **Remark:** when $\lambda \to 0$, the bias goes to 0

# Bias of $\hat{\beta}^{\text{Ridge}}$ (II)

**Proof of the lemma:** recall that $\hat{\beta}^{\text{Ridge}} = (X^\top X + \lambda \mathrm{I}_d)^{-1} X^\top Y$. We are in the Gaussian linear model, thus $Y = X\beta^\star + \varepsilon$. Thus

$$\begin{aligned}
\mathbb{E}[\hat{\beta}^{\text{Ridge}}] - \beta^\star &= \mathbb{E}\left[(X^\top X + \lambda \mathrm{I}_d)^{-1} X^\top Y\right] - \beta^\star \\
&= \mathbb{E}\left[(X^\top X + \lambda \mathrm{I}_d)^{-1} X^\top (X\beta^\star + \varepsilon)\right] - \beta^\star \\
&= (X^\top X + \lambda \mathrm{I}_d)^{-1} X^\top X \beta^\star - \beta^\star \\
\mathbb{E}[\hat{\beta}^{\text{Ridge}}] - \beta^\star &= \left[(X^\top X + \lambda \mathrm{I}_d)^{-1} X^\top X - \mathrm{I}_d\right] \beta^\star .
\end{aligned}$$

- there is a trick:

$$(X^\top X + \lambda \mathrm{I}_d)^{-1} (X^\top X + \lambda \mathrm{I}_d) = \mathrm{I}_d ,$$

thus

$$(X^\top X + \lambda \mathrm{I}_d)^{-1} X^\top X - \mathrm{I}_d = -\lambda (X^\top X + \lambda \mathrm{I}_d)^{-1} .$$

# Bias of $\hat{\beta}^{\text{Ridge}}$ (III)

- thus

$$\begin{aligned}\left\|\mathbb{E}[\hat{\beta}^{\text{Ridge}}] - \beta^\star\right\|^2 &= \left\|-\lambda(X^\top X + \lambda \mathsf{I}_d)\beta^\star\right\|^2 \\ &= \lambda^2 \left\|V(\Sigma^\top\Sigma + \lambda \mathsf{I}_d)^{-1} V^\top \beta^\star\right\|^2 \\ &= \lambda^2 \left\|(\Sigma^\top\Sigma + \lambda \mathsf{I}_d)^{-1} V^\top \beta^\star\right\|^2\end{aligned}$$

since the Euclidean norm is rotation-invariant. We deduce that

$$\left\|\mathbb{E}[\hat{\beta}^{\text{Ridge}}] - \beta^\star\right\|^2 = \lambda^2 \sum_{j=1}^{d} \frac{1}{(\sigma_j^2 + \lambda)^2} \cdot (V_j^\top \beta^\star)^2 \,,$$

where $V_j$ is the $j$th column of $V$. $\square$

# Variance of $\hat{\beta}^{\text{Ridge}}$ (I)

**Lemma:** Set $W = (X^\top X + \lambda \mathrm{I}_d)^{-1} X^\top X$. Then, in the Gaussian linear model, the covariance of the ridge estimator is given by

$$\mathrm{Cov}(\hat{\beta}^{\text{Ridge}}) = \sigma^2 W (X^\top X)^{-1} W^\top .$$

In particular, if $X = U \Sigma V^\top$ is the SVD of $X$, then

$$\mathbb{E}\left[ \left\| \hat{\beta}^{\text{Ridge}} - \mathbb{E}[\hat{\beta}^{\text{Ridge}}] \right\|^2 \right] = \sigma^2 \sum_{j=1}^{d} \frac{\sigma_j^2}{(\sigma_j^2 + \lambda)^2} .$$

- **Intuition:** larger $\lambda$ reduces the variance ($\to 0$ when $\lambda \to +\infty$)
- take $\tilde{\sigma}_j$ the singular values of $\frac{1}{\sqrt{n}} X$ ($\sigma_j^2 = n \tilde{\sigma}_j^2$). Then, when $\lambda \to 0$, the variance reduces to

$$\sigma^2 \sum_{j=1}^{d} \frac{1}{\sigma_j^2} = \sigma^2 \sum_{j=1}^{d} \frac{1}{n \tilde{\sigma}_j^2} \approx \boxed{\frac{\sigma^2 d}{n}} .$$

# Variance of $\hat{\beta}^{\text{Ridge}}$ (II)

**Proof of the lemma:** recall that for any random vector $Z \in \mathbb{R}^d$ and any deterministic matrix $M \in \mathbb{R}^{d \times d}$,

$$\mathrm{Cov}(MZ) = M\,\mathrm{Cov}(Z)\,M^\top .$$

*Hint:* remind yourself the Gaussian vector property

- notice that

$$\begin{aligned} W\hat{\beta}^{\text{LS}} &= (X^\top X + \lambda \mathrm{I}_d)^{-1} X^\top X \cdot (X^\top X)^{-1} X^\top Y \\ &= (X^\top X + \lambda \mathrm{I}_d)^{-1} X^\top Y \\ W\hat{\beta}^{\text{LS}} &= \hat{\beta}^{\text{Ridge}} \end{aligned}$$

- therefore (see course 6 for the covariance of $\hat{\beta}^{\text{LS}}$)

$$\begin{aligned} \mathrm{Cov}(\hat{\beta}^{\text{Ridge}}) &= W\mathrm{Cov}(\hat{\beta}^{\text{LS}})W^\top \\ &= \sigma^2 W(X^\top X)^{-1} W^\top \end{aligned}$$

# Variance of $\hat{\beta}^{\text{Ridge}}$ (III)

- now we notice that

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\beta}^{\text{Ridge}} - \mathbb{E}[\hat{\beta}^{\text{Ridge}}]\right\|^2\right] &= \sum_{j=1}^{d} \mathbb{E}\left[\left(\hat{\beta}_j^{\text{Ridge}} - \mathbb{E}[\hat{\beta}_j^{\text{Ridge}}]\right)^2\right] \\
&= \sum_{j=1}^{d} \text{Cov}(\hat{\beta}^{\text{Ridge}})_{jj} \\
&= \text{trace}\left(\text{Cov}(\hat{\beta}^{\text{Ridge}})\right) \\
\mathbb{E}\left[\left\|\hat{\beta}^{\text{Ridge}} - \mathbb{E}[\hat{\beta}^{\text{Ridge}}]\right\|^2\right] &= \sigma^2 \text{trace}\left(W(X^\top X)^{-1} W^\top\right) .
\end{aligned}$$

- let us simplify this expression, first by seeing that

$$(X^\top X)^{-1} = V(\Sigma^\top \Sigma)^\dagger V^\top$$

# Variance of $\hat{\beta}^{\text{Ridge}}$ (IV)

- then writing

$$\begin{aligned} W &= (X^\top X + \lambda \mathrm{I}_d)^{-1} X^\top X \\ &= V(\Sigma^\top \Sigma + \lambda \mathrm{I}_d)^{-1} \Sigma^\top \Sigma V^\top \end{aligned}$$

- we obtain

$$W(X^\top X)^{-1} W^\top = V(\Sigma^\top \Sigma + \lambda \mathrm{I}_d)^{-2} \Sigma^\top \Sigma V^\top .$$

- finally, we write

$$\begin{aligned} \operatorname{trace}\left(W(X^\top X)^{-1} W^\top\right) &= \operatorname{trace}\left(V(\Sigma^\top \Sigma + \lambda \mathrm{I}_d)^{-2} \Sigma^\top \Sigma V^\top\right) \\ &= \operatorname{trace}\left((\Sigma^\top \Sigma + \lambda \mathrm{I}_d)^{-2} \Sigma^\top \Sigma\right) \\ \operatorname{trace}\left(W(X^\top X)^{-1} W^\top\right) &= \sum_{j=1}^{d} \frac{\sigma_j^2}{(\sigma_j^2 + \lambda)^2} , \end{aligned}$$

which concludes the proof. □

# Mean squared error of $\hat{\beta}^{\text{Ridge}}$ (I)

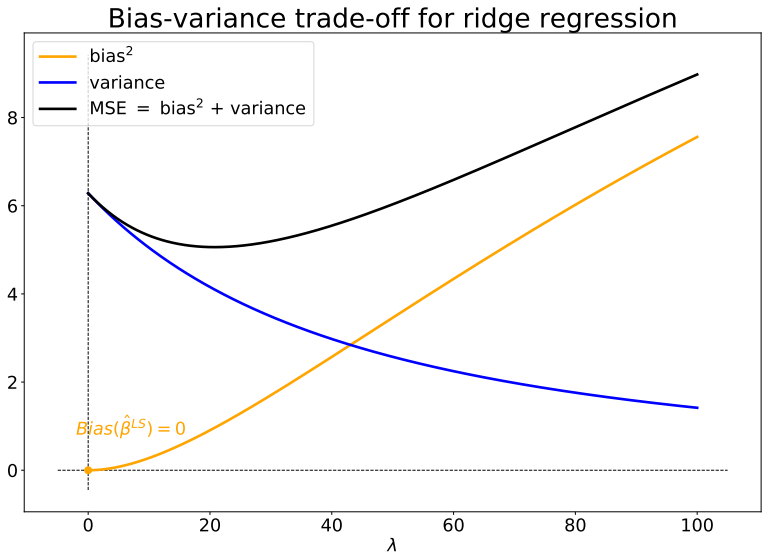
**Theorem:** Assume that we are in the Gaussian linear model *(fixed design)* $Y = X\beta^\star + \varepsilon$, with $\varepsilon \sim \mathcal{N}\left(0, \sigma^2 \mathsf{I}_n\right)$. Let $X = U\Sigma V^\top$ be the SVD of $X$. Then

$$\mathrm{MSE}\left(\hat{\beta}^{\text{Ridge}}\right) = \lambda^2 \sum_{j=1}^{d} \frac{1}{(\sigma_j^2 + \lambda)^2} \cdot \left(V_j^\top \beta^\star\right)^2 + \sigma^2 \sum_{j=1}^{d} \frac{\sigma_j^2}{(\sigma_j^2 + \lambda)^2} \cdot$$

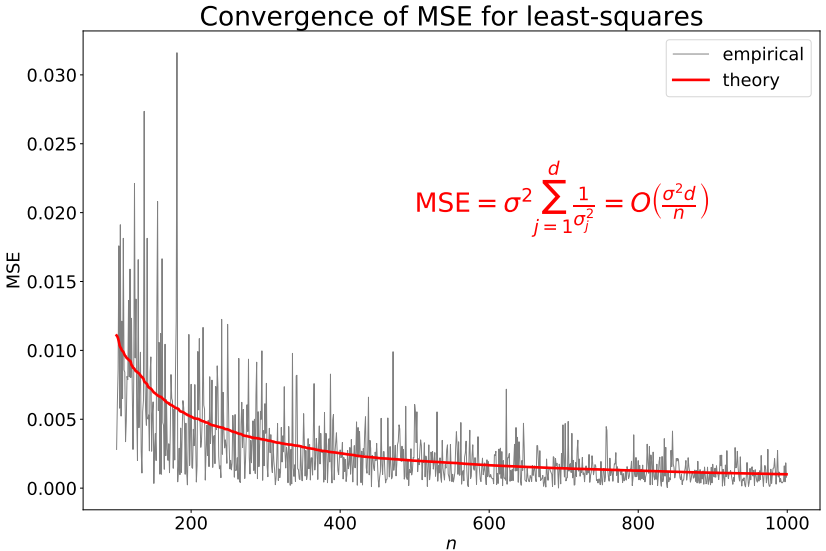
- **Intuition:** we accept some bias to reduce the MSE
- we can choose $\lambda$ to balance the bias and the variance part (*bias-variance trade-off*) and hopefully minimize the MSE
- **Remark:** typically,[3] there exist a range of $\lambda$ such that $\mathrm{MSE}(\lambda)$ is smaller than $\mathrm{MSE}(0)$ (least-squares case)

[3] not always! see Kobak, Lomond, Sanchez, *The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization*, JMLR, 2020

# Mean squared error of $\hat{\beta}^{\text{Ridge}}$ (II)



# Mean squared error of $\hat{\beta}^{\text{Ridge}}$ (III)



# Mean squared error of $\hat{\beta}^{\text{Ridge}}$ (IV)

- **How to choose $\lambda$?**
- back to the orthonormal case ($X^\top X = \mathsf{I}_d$)
- in this case, $\sigma_j = 1$ for any $j$, and

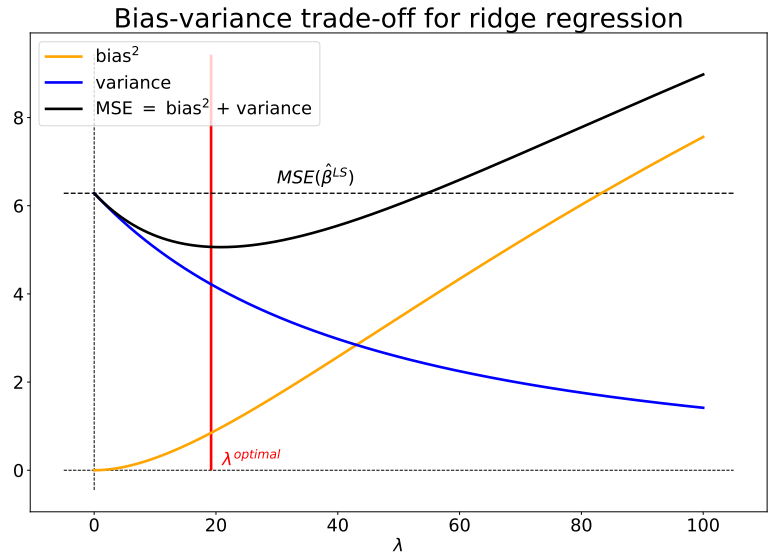
$$\mathrm{MSE}(\lambda) = \frac{\lambda^2}{(1+\lambda)^2} \|\beta^\star\|^2 + \frac{\sigma^2 d}{(1+\lambda)^2} .$$

- we can study the function $\lambda \mapsto \mathrm{MSE}(\lambda)$, and we find a minimum at

$$\lambda = \frac{\sigma^2 d}{\|\beta^\star\|^2} .$$

- **Unfortunately,** we know neither $\sigma^2$ nor $\|\beta^\star\|^2$...
- in practice, we use cross-validation

# Mean squared error of $\hat{\beta}^{\text{Ridge}}$ (V)



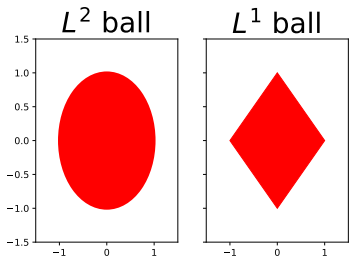
# 42. Least Absolute Shrinkage and Selection Operator

# The Lasso (I)

- **Idea:** Least Absolute Shrinkage (Lasso) replaces the $L^2$ norm by the $L^1$ norm [4]
- namely, for some $\lambda > 0$, find

$$\hat{\beta}^{\text{Lasso}} \in \underset{\beta \in \mathbb{R}^d}{\arg\min} \left\{ \|Y - X\beta\|^2 + \lambda \|\beta\|_1 \right\} .$$

- recall that $\|x\|_1 = \sum_{i=1}^{d} |x_i|$:




[4] Tibshirani, *Regression Shrinkage and Selection via the Lasso*, 1986

# The Lasso (II)

- recall that we are looking for:

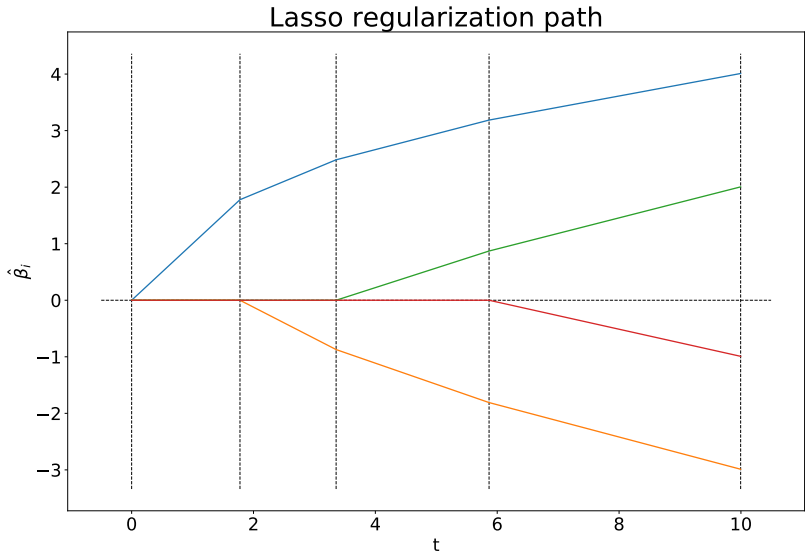
$$\hat{\beta}^{\text{Lasso}} \in \underset{\beta \in \mathbb{R}^d}{\arg\min} \left\{ \|Y - X\beta\|^2 + \lambda \|\beta\|_1 \right\} . \tag{2}$$

- equivalent formulation (see TD):

$$\hat{\beta}^{\text{Lasso}} \in \underset{\beta \in \mathbb{R}^d}{\arg\min} \|Y - X\beta\|^2$$
$$\text{subject to } \|\beta\|_1 \leq t .$$

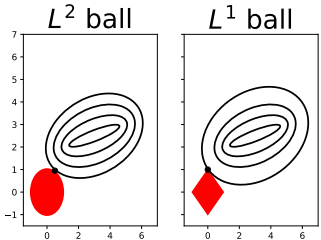
- **Question:** how do we find $\hat{\beta}^{\text{Lasso}}$?
- **Problem:** (2) is convex, but **not differentiable**
- several possibilities:
  - subgradient methods
  - **Least-angle regression:** also gives the regularization path (for the same computational cost)

# The Lasso (III)



# The Lasso (IV)

- **Intuition:** $L^1$ norm promotes sparsity (sparse $=$ *"parcimonieux"*)




- in red, the constraint $\|\beta\|_\cdot \leq t$
- in black the level sets of $\|Y - X\beta\|^2$
- $L^1$ ball has many facets and edges: often the solution has many zero coordinates!

# The Lasso (IV)

- let us define
$$\|\beta\|_0 = |\{i \text{ s.t. } \beta_i \neq 0\}| \,,$$
the number of non-zero coordinates of vector $\beta$
- **Definition:** when
$$\|\beta\|_0 \ll d \,,$$
we say that $\beta$ is *sparse ("parcimonieux")*
- **Bet on sparsity:** assume that the groundtruth is sparse
  - if it is, we do well with the Lasso
  - if not, then no method is going to perform well
- also important to have a variable selection method
- quickly identify which covariates are important

# The Lasso (V)

- let us try to understand why in the **orthonormal case** ($X^\top X = \mathsf{I}_d$)
- we write

$$\begin{aligned} R(\beta) &= \|Y - X\beta\|^2 + \lambda \|\beta\|_1 \\ &= (Y - X\beta)^\top (Y - X\beta) + \lambda \|\beta\|_1 \\ R(\beta) &= Y^\top Y - 2\beta^\top X^\top Y + \beta^\top X^\top X\beta + \lambda \|\beta\|_1 \end{aligned}$$

- recall that $\hat{\beta}^{\mathsf{LS}} = (X^\top X)^{-1} X^\top Y = X^\top Y$:

$$\begin{aligned} R(\beta) &= -2\beta^\top \hat{\beta}^{\mathsf{LS}} + \beta^\top \beta + \lambda \|\beta\|_1 + \mathsf{cst} \\ &= \sum_{j=1}^{d} \left\{ -2\beta_j \hat{\beta}_j^{\mathsf{LS}} + \beta_j^2 + \lambda |\beta_j| \right\} + \mathsf{cst} \end{aligned}$$

# The Lasso (VI)

- that is,

$$R(\beta) = \sum_{j=1}^{d} R_j(\beta_j) + \text{cst}\,,$$

  with $R_j(x) = -2\hat{\beta}_j^{\text{LS}} x + x^2 + \lambda\,|x|$

- let us look at a coordinate such that $\hat{\beta}_j^{\text{Lasso}} \neq 0$
- then we can differentiate:

$$\frac{\partial R(\beta)}{\partial \beta_j} = \frac{\partial R_j(\beta_j)}{\partial \beta_j} = -2\hat{\beta}_j^{\text{LS}} + 2\beta_j + \lambda \text{sign}\,(\beta_j)$$

- $\hat{\beta}_j^{\text{Lasso}}$ solves

$$-2\hat{\beta}_j^{\text{LS}} + 2x + \lambda \text{sign}\,(x) = 0\,.$$

# The Lasso (VII)

- a nonzero solution must satisfy, depending on its sign,

$$-2\hat{\beta}_j^{\mathsf{LS}} + 2x + \lambda = 0 \quad \text{or} \quad -2\hat{\beta}_j^{\mathsf{LS}} + 2x - \lambda = 0\,.$$

- that is,

$$x = \hat{\beta}_j^{\mathsf{LS}} - \frac{\lambda}{2} \text{ if } x > 0 \quad \text{or} \quad x = \hat{\beta}_j^{\mathsf{LS}} + \frac{\lambda}{2} \text{ if } x < 0\,.$$

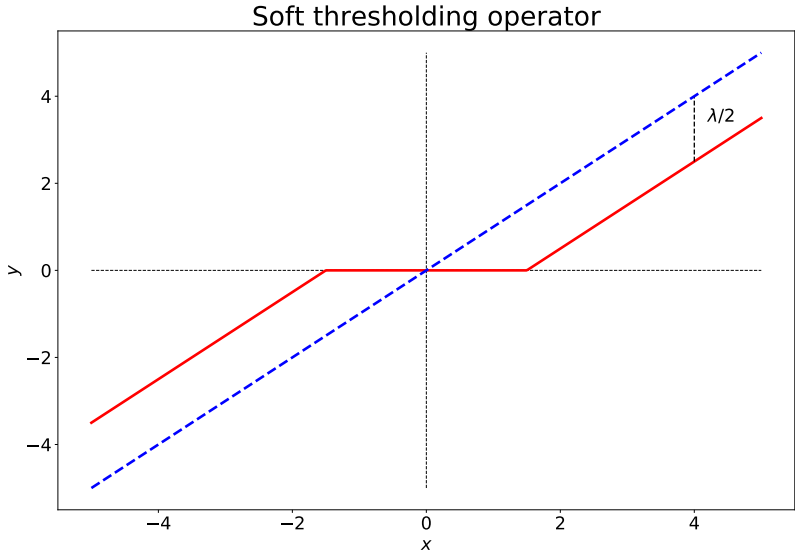
- we deduce the three possibilities:

$$\begin{cases} \hat{\beta}_j^{\mathsf{LS}} - \frac{\lambda}{2} > 0 & \Rightarrow \hat{\beta}_j^{\mathsf{Lasso}} = \hat{\beta}_j^{\mathsf{LS}} - \frac{\lambda}{2} \\ \hat{\beta}_j^{\mathsf{LS}} + \frac{\lambda}{2} < 0 & \Rightarrow \hat{\beta}_j^{\mathsf{Lasso}} = \hat{\beta}_j^{\mathsf{LS}} + \frac{\lambda}{2} \\ \hat{\beta}_j^{\mathsf{LS}} \in \left[\frac{-\lambda}{2}, \frac{\lambda}{2}\right] & \Rightarrow \hat{\beta}_j^{\mathsf{Lasso}} = 0 \end{cases}$$

In short,

$$\hat{\beta}_j^{\mathsf{Lasso}} = \operatorname{sign}\left(\hat{\beta}_j^{\mathsf{LS}}\right) \cdot \left(\left|\hat{\beta}_j^{\mathsf{LS}}\right| - \frac{\lambda}{2}\right)_+$$

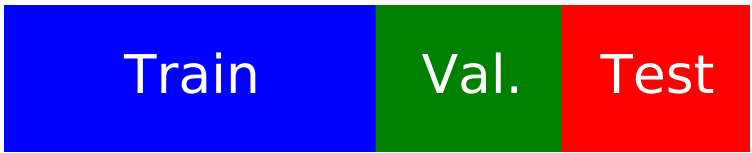
# The Lasso (VIII)



# 43. Cross-validation

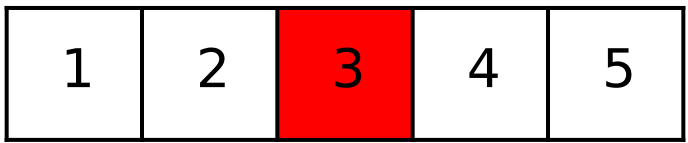
# Cross-validation (I)

- model $\hat{f}$ trained on
$$\mathcal{X}_{\text{train}} = \{(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}, 1 \leq i \leq n\}\,.$$
- we assume that our model $\hat{f}_\lambda$ depends on a parameter $\lambda$
- we want to find $\lambda$ that minimizes some criterion *on new data* (test error)
- **Idea:** estimate it by creating a test set from the train set
- we call this new set the **validation set**
- **Ideally,** with enough data, the picture is the following:



# Cross-validation (II)

- **General idea:** pick the best $\lambda$ on the train/val sets
- **Problem:** in general, data is scarce
- we want to use all the available data
- **Idea:** $K$-fold cross-validation *("validation croisée")*:
  - split the training set in $K$ parts
  - for each $i \in \{1, \dots, K\}$, train on $K-1$ parts
  - and compute the test error on the remaining part
  - aggregate the errors
- typical choice: $K = 5, 10$ or $n$ (*leave-one-out*)
- schematically:



# Cross-validation (III)

- more details: $\kappa : \{1, \ldots, n\} \longrightarrow \{1, \ldots, K\}$ indexing function
- $\kappa(i)$ tells us in which box observation $i$ belongs
- we define $\hat{f}^{-k}$ the model trained on **all observations not in box** $k$
- then the cross-validation estimate of the prediction error is

$$\mathrm{CV}(\hat{f}) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \hat{f}^{-\kappa(i)}(x_i)) \,.$$

- when our model depends on a parameter $\lambda$, we define similarly

$$\mathrm{CV}(\hat{f}, \lambda) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \hat{f}_{\lambda}^{-\kappa(i)}(x_i)) \,.$$

- then $\lambda \mapsto \mathrm{CV}(\hat{f}, \lambda)$ is an estimate of the test error curve, and we can choose $\hat{\lambda}$ minimizing $\mathrm{CV}(\hat{f}, \lambda)$

# Cross-validation (IV)

- **Computational cost of cross-validation:** $K\times$ cost of training our model on $(1-1/K)n$ points $+$ $n\times$ cost of prediction
- this can be **huge** and generally constrains $K$ to small values
- but for small $K$, we are estimating the **averaged test error**

$$\overline{E_{\text{test}}} = \mathbb{E}_{X_1,\dots,X_n}[E_{\text{test}}] \ ,$$

  since the training sets for $\hat{f}^{-k}$ are quite different
- generally we are interested in $E_{\text{test}}$, pushing for larger $K$
- $K = 5$ or 10 is a good in-between[5]
- **Remark:** in certain cases, it is possible to compute CV faster (see TD)

[5] Arlot and Lerasle, *Why $V = 5$ is enough in $V$-fold cross-validation*, preprint, 2012