Statistics

Marco Lombardi

Quick announcements 1

Goals

- The main goal of this course is to learn to
  - make a physical model
  - combine it with a statistical model for the measurements
  - infer the parameters of the physical model or, in general, learn as much as you can from the data
  - have fun with physics and statistics!
- All data used are “real”: you will not need to do basic data reduction, but rather high-level modelling.

Quick announcements 2

Programming

- Throughout the course we will use **Python** as a reference programming language
- I will teach only the basics of Python during the first lectures
- For students that are not familiar with Python, there will be specific hands-on lectures taught by your tutor
- You should install asap the **Python 3 Anaconda distribution** on your laptop
- You are encouraged to start playing with Python if you are not proficient with this language

Quick announcements 3

Practicalities

- This year the course will be thought both **remotely** (synchronously) using Zoom and **in presence**
- Each lab experience will last 3-4 lectures; you will work in **groups** of 3-4 people
- All experiences will involve a lot of statistics; as a first consequence the groups will change randomly 😉
- The group will work together to understand the problem, make a physical model, and perform statistical inference using the data provided.

Quick announcements 4

Reports

- Each group **can** submit a report within 2 weeks from the end of the experience.
- The report will be evaluated
- Each student must ensure that **at least 3 reports** have been submitted.
- For each student, we will take into account the three top marks: their average will be used as a “starting mark” for the final exam

Quick announcements 5

Huston we have a problem...

- The success of the course is largely in your hands!
- It is very important for me to have your **feedback** on the lessons during the course
- Stop me if I have been unclear or if you have any doubt: other students will easily have the same doubt.
- Similarly, contact me if you think there is a general problem with the organisation of the course: the earlier you do, the earlier I can take the necessary actions.

Welcome to Lecture 1

The main reference for this course is
THE BOOK

“Information Theory, Inference, and Learning Algorithms”
by David J.C. MacKay
http://www.inference.org.uk/mackay/itila/

See also: “Statistical Rethinking” by Richard McElreath, CRC Press

Definition of Probability

An **ensemble** X is a triplet $(x, \mathscr{A}_X, \mathscr{P}_X)$, where

- x is the **outcome** of a random variable;
- $\mathscr{A}_X = \{a_1, a_2, \ldots, a_I\}$ is a set ("alphabet") that indicates all possible values for x;
- $\mathscr{P}_X = \{p_1, p_2, \ldots, p_I\}$ are the probabilities associated to each value: $P(x = a_i) = p_i$

Probabilities are positive, $p_i \geq 0 \forall i$, and normalised to unity, so that $\sum_{i=1}^{I} p_i = 1$.

Probability of a subset

If $\mathscr{B} \subset \mathscr{A}_X$, we will call the probability of $\mathscr{B}$

$$P(\mathscr{B}) = P(x \in \mathscr{B}) = \sum_{a_i \in \mathscr{B}} P(x = a_i)$$

For discrete events, such as the ones we consider now, we can talk about $P(x = a_i)$.

For continuous variables it makes more sense to talk about probability of a subset, since $P(x = a_i)$ would generally vanish.

Trivial facts

(based on set theory)

- An event can occurs always, such as $P(\mathscr{A}_X) = 1$
- An event can never occur, such as $P(\varnothing) = 0$
- All probabilities are in the unit range, $0 \leq P(x) \leq 1$
- If $\mathscr{B} \cap \mathscr{C} = \varnothing$ are two mutually esclusive events, than $P(\mathscr{B} \cup \mathscr{C}) = P(\mathscr{B}) + P(\mathscr{C})$
- More generally, the inclusion-exclusion rule holds: $P(\mathscr{B} \cup \mathscr{C}) = P(\mathscr{B}) + P(\mathscr{C}) - P(\mathscr{B} \cap \mathscr{C})$
- For the complement event $P(\mathscr{A}_X \backslash \mathscr{B}) = 1 - P(\mathscr{B})$ holds.

Notation

- Lowercase letters (x) are used for **single outcomes**; capital letters (X) for **ensembles**.
- From a single outcome x one cannot derive any statistical indication (average, median, confidence intervals...). The ensemble is required for that!
- I will generally write $P(a_i)$ or $P(x)$, meaning really $P(x = a_i)$; the context should make everything clear.

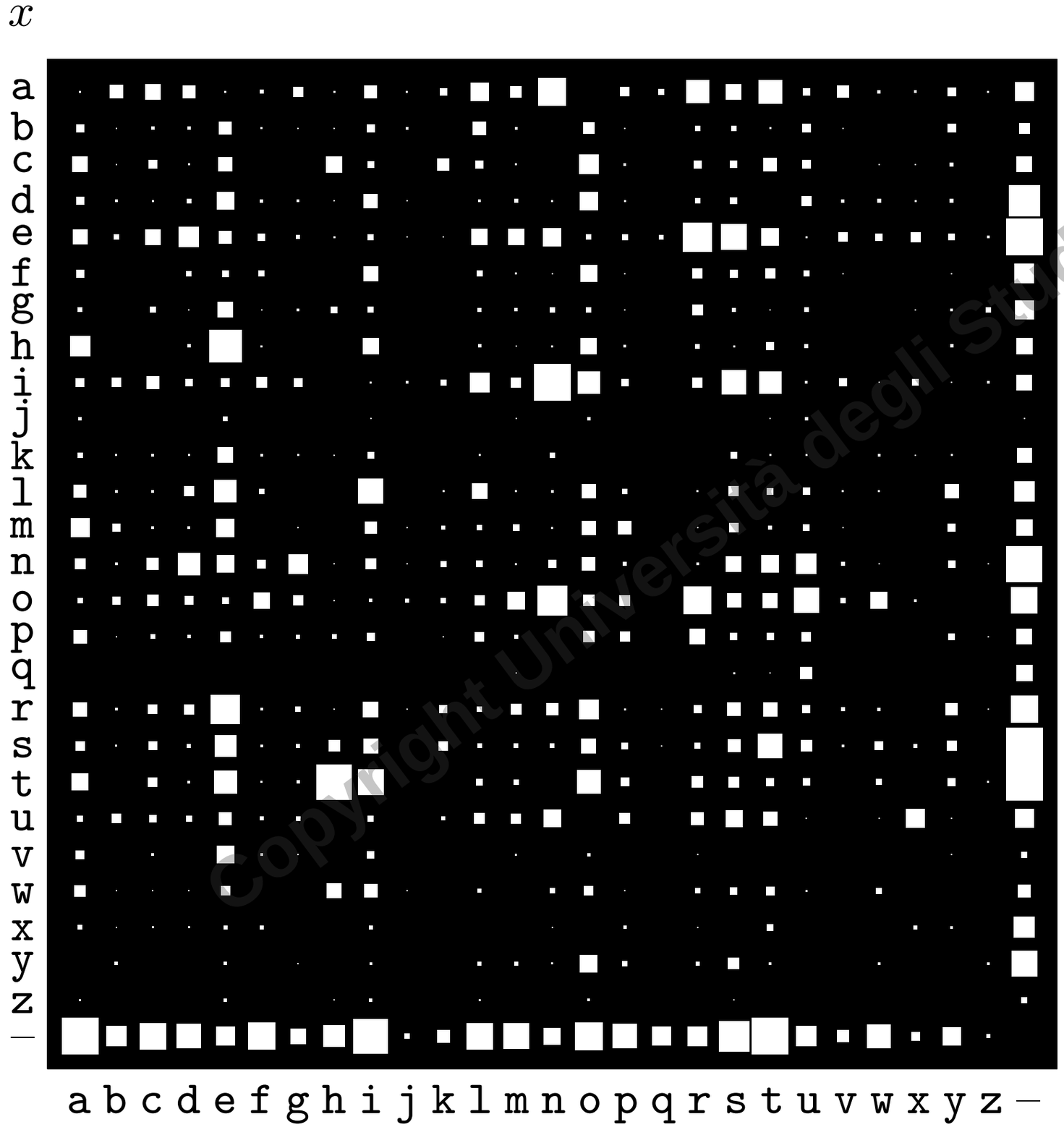
Joint Ensamble

Ensembles have a natural extension for the **Cartesian product**. We call XY an ensemble for the ordered couple (x, y), with

- $x \in \mathcal{A}_X$ and $y \in \mathcal{A}_Y$;
- $P(x, y)$ the probability that the joint event (x, y) is happens.

Note that $P(x, y)$ does not need to be $P_X(x)P_Y(y)$; that is, in general the two events x and y do not need to be **independent**.

Example: couples of letters



Probabilities of all couples of letters from the English language document *"The Frequently Asked Questions Manual for Linux"*.

Marginal and conditional probability

For a joint ensemble XY we can compute the **marginal probability** for the outcome y as

$$P(y) = \sum_x P(x, y) \qquad \text{(and similarly for } x \leftrightarrow y\text{)}$$

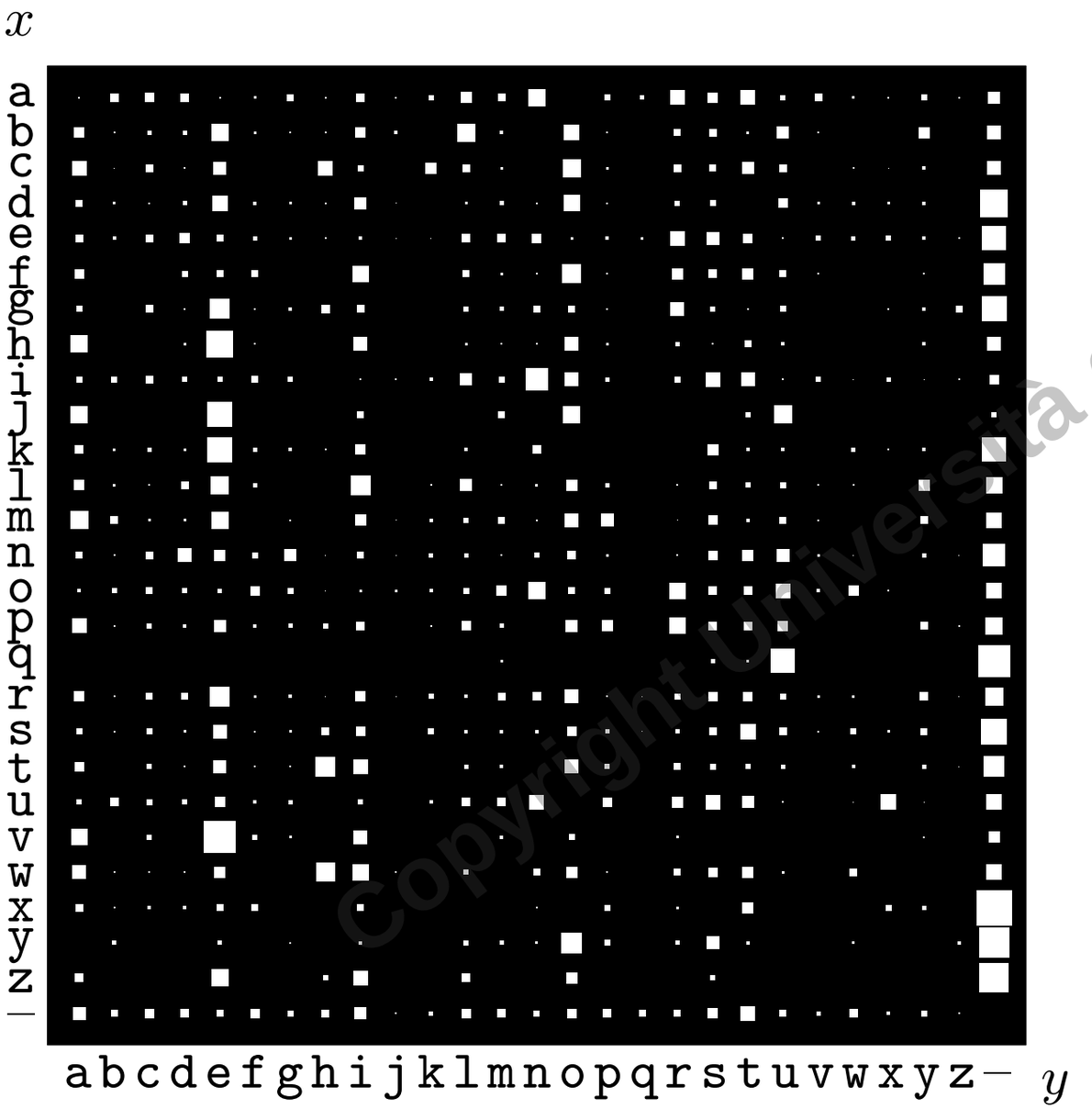
This ignores the outcome of x. The **conditional probability** is

$$P(x \mid y) = \frac{P(x, y)}{P(y)}$$

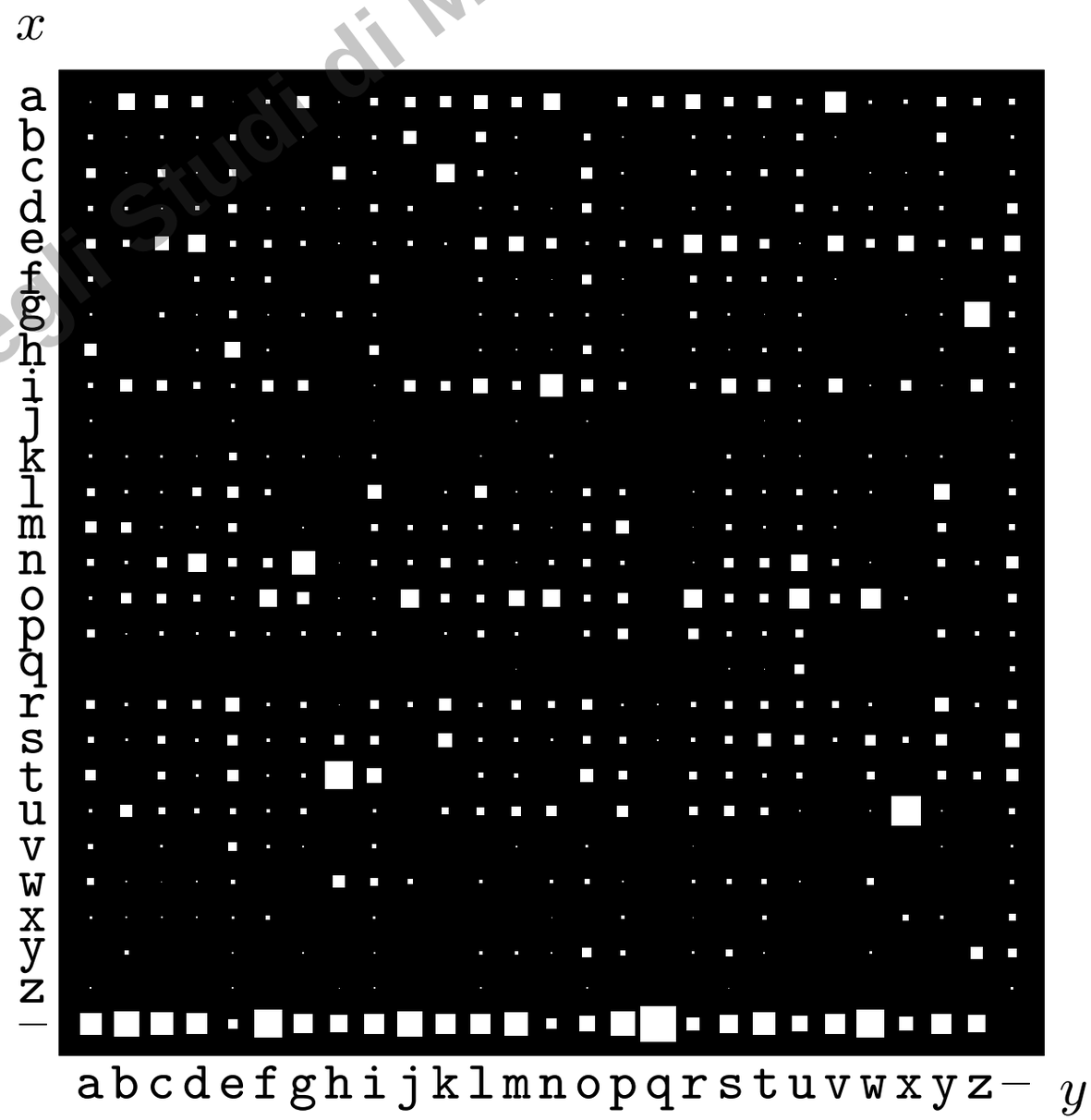
The conditional probability is the probability that the event x happens, given the fact that the event y has happened.

Warning: the conditional probability is not symmetric!

Example: couples of letters

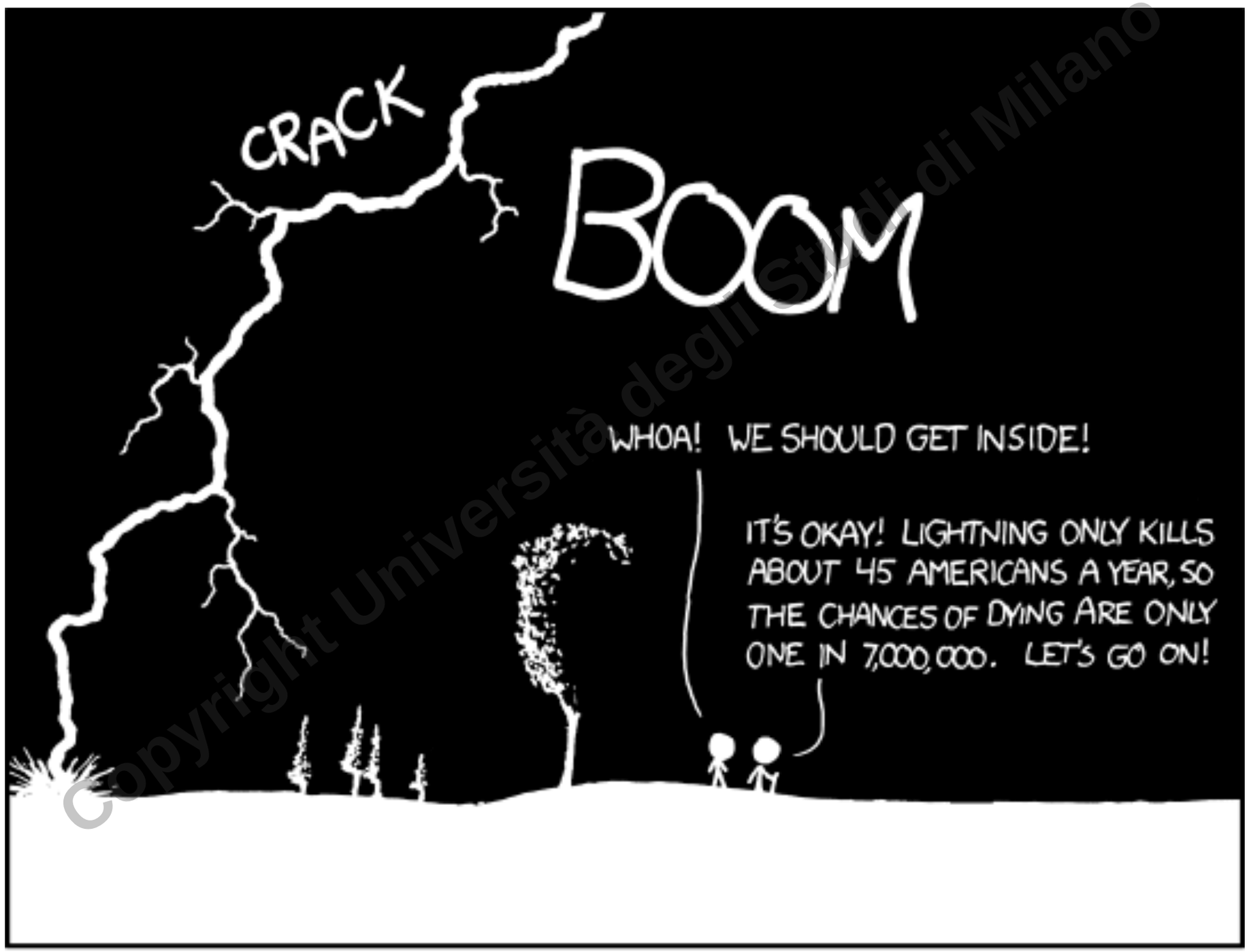


(a) $P(y \mid x)$



(b) $P(x \mid y)$

Conditional Risk



THE ANNUAL DEATH RATE AMONG PEOPLE WHO KNOW THAT STATISTIC IS ONE IN SIX.

Problems...

Problem 1. Charlie has two brothers, Alf and Bob. (a) What is the probability that Charlie is older than Bob? (b) Charlie tells you he is older than Alf. Now, what is the probability that Charlie is older than Bob?

Problem 2. The inhabitants of an island tell the truth 1/3 of the time (they lie with probability 2/3). After one of them makes a statement, you ask another one if that statement was true; the answer you get is “yes”. What is the probability the statement was true?

Product and sum rules

From the definition of the conditional probability one immediately finds the **product rule**

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

This is also called the **chain rule** because it can be applied to multiple variables

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

The marginal probability can then be expressed using the **sum rule**

$$P(x) = \sum_{y} P(x \mid y)P(y)$$

Total probability

Suppose the events $\mathscr{H}_1, \mathscr{H}_2, \ldots, \mathscr{H}_n$ form a **partition** of the entire sample space, i.e.

- They are **mutually exclusive** $\mathscr{H}_i \cap \mathscr{H}_j = \varnothing$
- Their union $\bigcup_i \mathscr{H}_i = \mathscr{A}$ is the sample space

Then $P(x) = P(x \mid \mathscr{H}_1)P(\mathscr{H}_1) + \cdots + P(x \mid \mathscr{H}_n)P(\mathscr{H}_n)$.

Problem. Out of 100 coins one has heads on both sides. One coin is chosen at random and flipped two times. What is the probability to get two heads?

Bayes' Theorem

(something brand new)

A trivial consequence of the previous definitions and propositions is Bayes' theorem

$$P(y\,|\,x) = \frac{P(x\,|\,y)P(y)}{P(x)} = \frac{P(x\,|\,y)P(y)}{\sum_{y'} P(x\,|\,y')P(y')}$$

- $P(y\,|\,x)$ is a probability in y $\Rightarrow$ has to be normalized for y
- The normalisation is granted by the denominator.

Use of Bayes' theorem

Bayes' theorem is used to swap the role of random variables, i.e. to transform $P(x \mid y) \leftrightarrow P(y \mid x)$

Example. Mr. Rossi tests positive for COVID. The test has a reliability of 95% (that is, it is positive for a healthy person in 5% of cases, and is negative for a COVID-infected person in 5% of cases). We also know that on average 1% of people have COVID in Mr. Rossi's age-group. What is the probability that Mr. Rossi has COVID?

TYPE I ERROR: FALSE POSITIVE

TYPE II ERROR: FALSE NEGATIVE

TYPE III ERROR: TRUE POSITIVE FOR INCORRECT REASONS

TYPE IV ERROR: TRUE NEGATIVE FOR INCORRECT REASONS

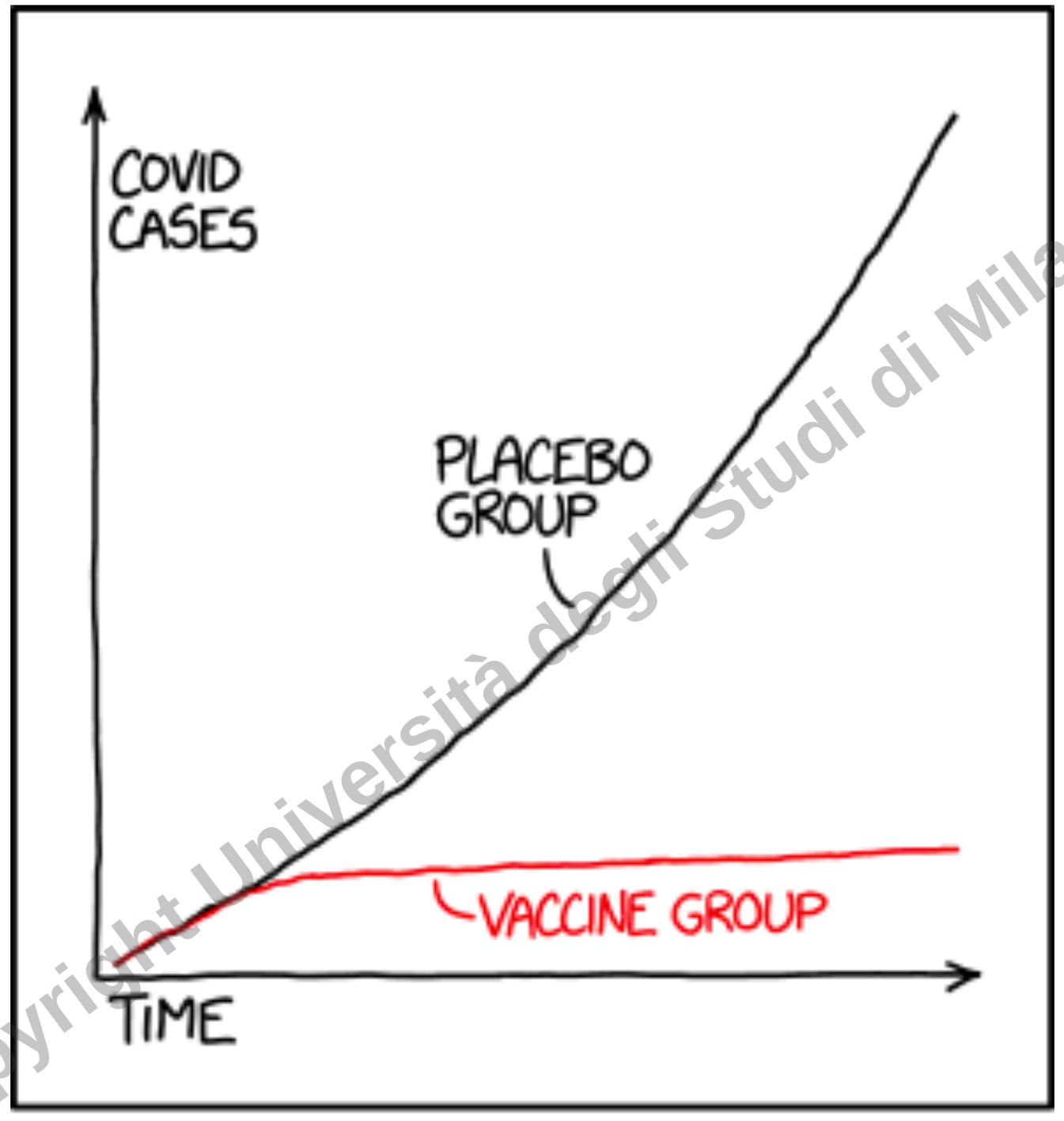
TYPE V ERROR: INCORRECT RESULT WHICH LEADS YOU TO A CORRECT CONCLUSION DUE TO UNRELATED ERRORS

TYPE VI ERROR: CORRECT RESULT WHICH YOU INTERPRET WRONG

TYPE VII ERROR: INCORRECT RESULT WHICH PRODUCES A COOL GRAPH

TYPE VIII ERROR: INCORRECT RESULT WHICH SPARKS FURTHER RESEARCH AND THE DEVELOPMENT OF NEW TOOLS WHICH REVEAL THE FLAW IN THE ORIGINAL RESULT WHILE PRODUCING NOVEL CORRECT RESULTS

TYPE IX ERROR: THE RISE OF SKYWALKER



STATISTICS TIP: ALWAYS TRY TO GET DATA THAT'S GOOD ENOUGH THAT YOU DON'T NEED TO DO STATISTICS ON IT

Definition of Probability

Again!

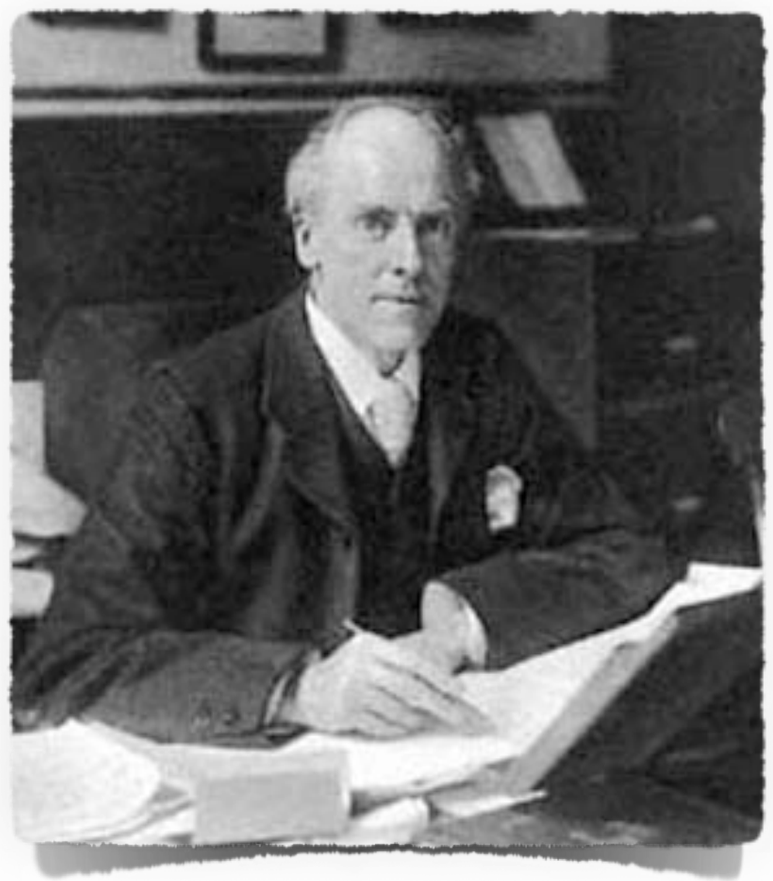
- “Probability describes frequencies of outcomes in random experiment”
  - What is “random”? Hard to define without using the word “probability”
- I will assume we all know what are we talking about, so I will skip formalities…
- [But we will make a digression on the formal, mathematical definition of probability later on]

Coin Tosses

The most obvious random number generator!

Still non that obvious: Pearson had to toss a coin 24,000 times to figure out how it worked (he obtained $f = 0.5005$).

Also, what if the coin **cannot** be tossed? (1913 Liberty Head nickel, worth 3.7M$...)



Interpretation of probability

(Frequentist) ...the ratio of times that we expect an event to occur (#successes / #experiments).

(Classical) ...the ratio of number of favourable and possible outcomes in a (symmetric) experiment.

(Subjectivist) ...an individual's degree of belief in the occurrence of an event.

Interpretation of probability

(Frequentist) ...the ratio of times that we expect an event to occur (#successes / #experiments).

(Classical) ...the ratio of number of favourable and possible outcomes in a (symmetric) experiment.

(Subjectivist) ...an individual's degree of belief in the occurrence of an event.

The men of the street has not troubles with the last interpretation; many statisticians do! [This interpretation is really due to Laplace's 1814 essay]

Cox axioms

(1946, 1961)

Call $B(x)$ the "degree of belief" (dob) in the proposition x.

1. dob's can be ordered: if $B(x) \prec B(y)$ and $B(y) \prec B(z)$, then $B(x) \prec B(z)$.
2. dob of a $\bar{x}$ (negated proposition x) can be computed from dob of x: $B(\bar{x}) = f[B(x)]$.
3. dob of (x, y) can be computed in terms of conditional dob: $B(x, y) = g[B(x \mid y), B(y)]$

Bayesian statistics

- Probabilities can be used to describe **assumptions** (beliefs).
- They can also used to derive **inferences** from these assumptions and from some data; Bayes' theorem is the main tool here.
- People starting with the same assumptions and using the same data will make identical inferences.
- Bayesian statistics is based on **subjective** assumptions. This can be taken as a problem, or rather as a big advantage.
- **You cannot do statistics without making explicit assumptions!**

Nomenclature of Bayes' theorem

$$P(\theta \mid D) = \frac{P(D \mid \theta)\, P(\theta)}{P(D)}$$

- **Likelihood**: a description of the (physical) problem and of the associated measurements
- **Prior**: our belief on the parameters' values before we do any measurement
- **Evidence**: for now just a normalisation term; later on a very important factor!
- **Posterior**: our updated belief on the parameters' values.

Forward probability

The simplest probability problems involve a forward probability calculation: evaluate the probability of one event.

Problem 1. An urn contains K balls, of which B are black and $W = K - B$ are white. Fred draw a ball at random from the urn with replacement N times. What is the probability distribution for the number of black balls drawn?

To solve these problems it is useful to know a few probability distributions.

Simple probability distributions

Discrete ones

Uniform distribution. Typical example: coin, dice

$$P(r \mid N) = \frac{1}{N}$$

Bernoulli distribution. Typical example: single bent coint

$$P(r \mid p) = \begin{cases} p & \text{if } r = 1 \\ 1 - p & \text{if } r = 2 \end{cases}$$

Simple probability distributions

Discrete ones

Binomial distribution. Typical example: number of heads in a bent coin tossed several times

$$P(r\,|\,f,N) = \binom{N}{r} f^r (1-f)^{N-r}$$

Poisson distribution. Rare events on a large population, such as photon counting

$$P(n\,|\,\mu) = \mathrm{e}^{-\mu}\frac{\mu^n}{n!}$$

Inverse probability

In these problems one computes the *conditional probability* of some quantity (parameters) *given* the observations.

Problem 2. There are 11 urns, each containing 10 balls. Urn u contains u black balls and 10 – u white ones. Fred selects one urn at random, give it to Bill, and Bill draws 10 balls with replacement from that urn. If 3 balls are black and 7 white, what is the probability that the urn selected is the u = 3?

These problems invariably requires the use of Bayes' theorem.

More problems

Problem 2 continued. In the urn problem, what is the probability that the next ball drawn by Bill from the same box is black? [This is an example of posterior predictive distribution]

Problem 3. Bill tosses a *bent* coin N times, obtaining a sequence of head and tails. Assume that the coins has a probability f of coming up head. What is the distribution for f if there are n heads in N tosses?

The beta integral will be useful for the solution:

$$\int_0^1 x^n (1-x)^m \, \mathrm{d}x = \frac{n!m!}{(n+m+1)!}$$

Even more problems

Problem 4. Two ordinary dice are thrown. What is the probability distribution for the *sum* of their values? What is the probability distribution for the *difference*?

Problem 5. Consider a unit sphere in an $\mathbb{R}^N$. Show that the fraction of volume of the sphere that is in the surface shell lying at values of radius between $1 - \mathrm{d}R$ and 1 is $f = 1 - (1 - \mathrm{d}R)^N$

The Royal Statistical Society

SALFORD ADVERTISER

THURSDAY JUNE 13, 2002

SALFORD UNI MAN SAYS SALLY CLARK CONVICTION MAY BE WRONG

Maths professor challenges double baby murder case

A SALFORD University Maths professor will challenge evidence used to convict a solicitor of murdering her two baby sons at a conference on cot-deaths next week.

Prof Ray Hill, from Eccles, head of the university's Applied and Discrete Mathematics Research Unit said statistical evidence used to convict Sally Clark, from Wilmslow, in October 2000, was not only quoted out of context and unfairly used to imply guilt, but was actually wrong.

Watching the trial on the TV he became furious and told us: "I shouted at the screen 'that figure's wrong!' They took an estimated figure for the likelihood of one cot death and then just squared it to get this one-in-73 million chance. That's not allowed unless you're sure the events are independent. A bookie wouldn't give you those odds."

He has now studied the Confidential Enquiry into Stillbirths and Deaths in Infancy (CESDI) report, which gives detailed figures on the number of deaths from 1993-1996.

He said: "It seems the chances of two cot deaths in the same family are much higher than the prosecution led the jury to believe."

Prof Hill has written to several national newspapers and is working with Sally Clark's defence team on the campaign to free her.

He will present his full criticism of the evidence at a Developmental Physiology Conference on cot deaths organised by Leicester University on June 28.

The Criminal Cases Review Commission has been looking at the case and is expected to report within the next few weeks. With their report imminent, Sally Clark's defence team and family do not feel it is appropriate to comment.

For more information on the Sally Clark campaign visit www.sallyclark.org.uk

Evidence challenge: Prof Ray Hill (2553-S 02)

The ly Clark case,

"In the recent highly-publicised ca published studies to obtain a figure for the frequency of sudden t death") in families having some of the characteristics of the defendant's family. He went e this figure to obtain a value of 1 in 73 million for the frequency of two cases of SIDS in such a family.

“Prosecutor’s fallacy”

Never confuse $P(x\,|\,y)$ with $P(y\,|\,x)$, especially when rare events are involved!

Real-life example. Sally Clark was accused in 1998 of having killed her first child at 11 weeks of age, then conceived another child and killed it at 8 weeks of age. The prosecutor argument was based on Roy Meadow testimony, where he evaluated the probability of 2 children dying from sudden infant death syndrome to be 1 in 73 millions. He did not evaluate the probability that a mother generates two children and kills both of them.

Problems

Problem 1. A burglary has been committed and 10,000 have their fingerprints compared to a sample from the crime. One man has a matching fingerprint (probability of this happening by chance: 1 in 20,000). What is the probability to find at least one false positive in the city? What is the probability that the man is the burglar?

Problem 2. A box has 1,000,000 coins, and one of these has heads on both sides ("guilty coin"). A coin is extracted at random and it is tossed 15 times, giving 15 heads. The "prosecutor" claims the coin is "guilty".

The Likelihood principle

Example. Urn A contains 5 balls: 1 black, 2 white, 1 green and 1 pink. Urn B contains 500 balls: 200 black, 100 white, 50 yellow, 40 cyan, 30 sienna, 25 green, 25 silver, 20 gold, 10 purple. One urn is selected at random and one ball is drawn. The ball is black. What is the probability that the urn selected is urn A?

Likelihood principle. Given a generative model for data d given parameters θ, $P(d \mid \theta)$, and having observed a particular outcome d_0, all inferences and predictions should depend only on the function $P(d_0 \mid \theta)$.

Non-countable alphabets

- When the alphabet $\mathscr{A}$ is a **non-countable** set the previous definitions cannot be used
    - We cannot have non-countable probabilities p_i such that $\sum_i p_i = 1$.
- In this case we need to drop the request to assign a probability to each possible outcome
    - We will instead assign probabilities to **sets of outcome**
- For that we need a new definition of **probability space**

Probability space

A **probability space** is a triplet $(\mathscr{A}, \mathscr{F}, P)$, where

- $\mathscr{A}$ is the **sample space** or **alphabet**, i.e. the set of possible outcomes of a random experiment
- $\mathscr{F} \subseteq 2^{\mathscr{A}}$ is the **event space**, a set of subsets of $\mathscr{A}$ with a structure called σ-algebra:
  - $\mathscr{A} \in \mathscr{F}$;
  - $\mathscr{F}$ is closed for complement and countable unions (and thus for countable intersections)
- $P\colon \mathscr{F} \to [0,1]$ is the **probability**, a function on $\mathscr{F}$ which is countable additive and such that $P(\mathscr{A}) = 1$

Example 1: fair coin

To describe a fair coin we could make the following choice

- Sample space $\mathscr{A} = \{\mathrm{H}, \mathrm{T}\}$
- Event space given by $\mathscr{F} = \left\{\varnothing, \{\mathrm{H}\}, \{\mathrm{T}\}, \{\mathrm{H}, \mathrm{T}\}\right\}$
- $P(\varnothing) = 0$, $P(\{\mathrm{H}\}) = P(\{\mathrm{T}\}) = 1/2$, and $P(\mathscr{A}) = 1$

N.B. For discrete spaces it is natural to use the **power set** $2^{\mathscr{A}}$ for $\mathscr{F}$; also, probabilities can be defined on individual events ("**atoms**").

Example 2: random number

To describe the extraction of a random number in the range $[0,1]$ we need a little more care.

- Sample space $\mathscr{A} = [0,1] \subset \mathbb{R}$
- Event space $\mathscr{F}$: can be built using all open interval of $[0,1]$, their complement, numerable intersections and unions... that is, the **Borel sets** on $\mathscr{A}$.
- The associated probability is the **Lebesgue measure** on $\mathscr{A}$.

N.B. We cannot assign a probability to an outcome (i.e., a real number) nor to all possible subsets of the unit interval.

Random variables

For many purposes, it can be useful to "translate" the outcome of a random experiment into a different space

Given a probability space $(\mathscr{A}, \mathscr{F}, P)$, a **random variable** is a function $X\colon \mathscr{A} \to A$.

- We assign probabilities to any measurable subset $B \subset A$ using the natural rule $P(B) = P(X^{-1}(B))$.
- This require X to be a **measurable function**, i.e. such that the preimage of a measurable set is also a measurable set (cf. with the definition of a continuous function).
- Often one takes $A = \mathbb{R}$ or $A = \mathbb{R}^n$: this allows one to compute quantities such as the mean and the variance of a random variable.

Random variables: simplifications

In most cases, one can just consider a random variable as **a variable whose value depends randomly on an experiment**.

Example. Consider a fair dice roll.

- The outcome space could be the set
  $\mathscr{A} = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$
    - However, using only this space we could not compute any average!
- We could then map all dice symbols to the values $A = \{1,2,3,4,5,6\}$. We can then compute $E[X] \equiv \bar{X} = 7/2$.
    - Alternatively, just think at X as a random number taken from the set A, thus skipping entirely the event space $\mathscr{F}$.

Probability densities

For the common case of a real random variable X, i.e. when $A = \mathbb{R}$, we can assign a probability density f to X through the definition

$$P(a < X < b) = \int_a^b f(x)\,\mathrm{d}x$$

In order to P to be a valid probability, we require that

- f be Lebesgue integrable
- $f(x) \geq 0 \quad \forall x \in \mathbb{R}$
- f normalised to unity, i.e. $\int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x = 1$

Simple probability distributions

Continuous ones

Normal distribution. Justified by central limit

$$f(x\,|\,\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-(x-\mu)^2/2\sigma^2} \;,\; x \in \mathbb{R}$$

Exponential distribution. Common in decays or other Poisson point processes (with constant rate)

$$f(x\,|\,\lambda) = \lambda\mathrm{e}^{-\lambda x} \;,\; x \geq 0$$

Beta distribution. Conjugate prior of the Binomial:

$$f(x\,|\,\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1} \;,\; x \in [0,1]$$

Simple probability distributions

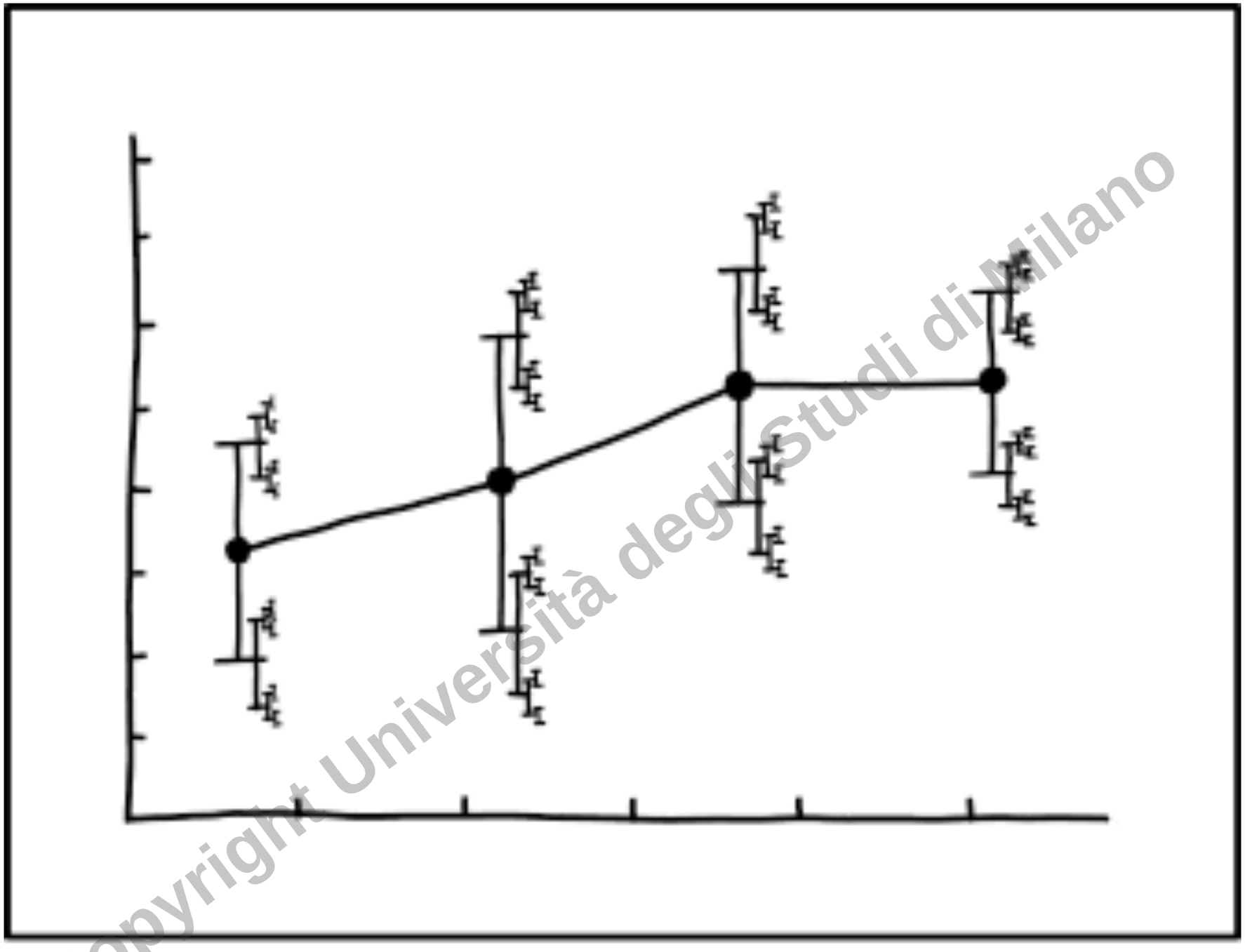
Continuous ones

Log-normal distribution. Central-limit theorem for quantities that multiply

$$f(x\,|\,\mu,\sigma^2) = \frac{1}{x\sqrt{2\pi\sigma^2}}\mathrm{e}^{-(\log x-\mu)^2/2\sigma^2} \quad , \; x > 0$$

Gamma distribution. Generalizes the exponential distribution

$$f(x\,|\,\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}\mathrm{e}^{-\beta x} \quad , \; x \geq 0$$



I DON'T KNOW HOW TO PROPAGATE ERROR CORRECTLY, SO I JUST PUT ERROR BARS ON ALL MY ERROR BARS.

Average, (co)variance...

- **Average**

$$E[X] \equiv \langle X \rangle = \sum_i p_i x_i = \int f(x)\, x \, \mathrm{d}x$$

- **Variance**

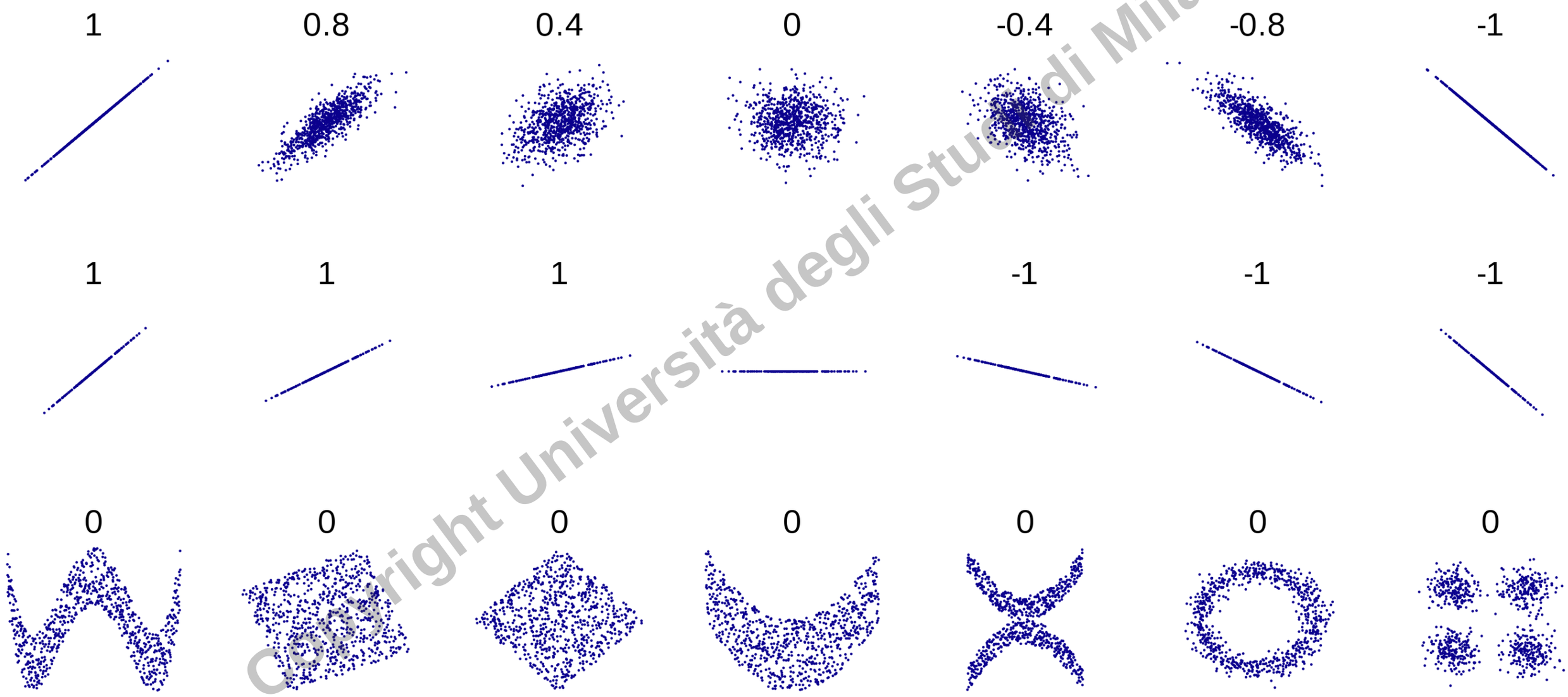
$$\mathrm{Var}[X] = E\left[(X - E[X])^2\right] = E[X^2] - E[X]^2$$

- **Covariance**

$$\mathrm{Cov}[X, Y] = E\left[(X - E[X])(Y - E[Y])\right] = E[XY] - E[X]\,E[Y]$$

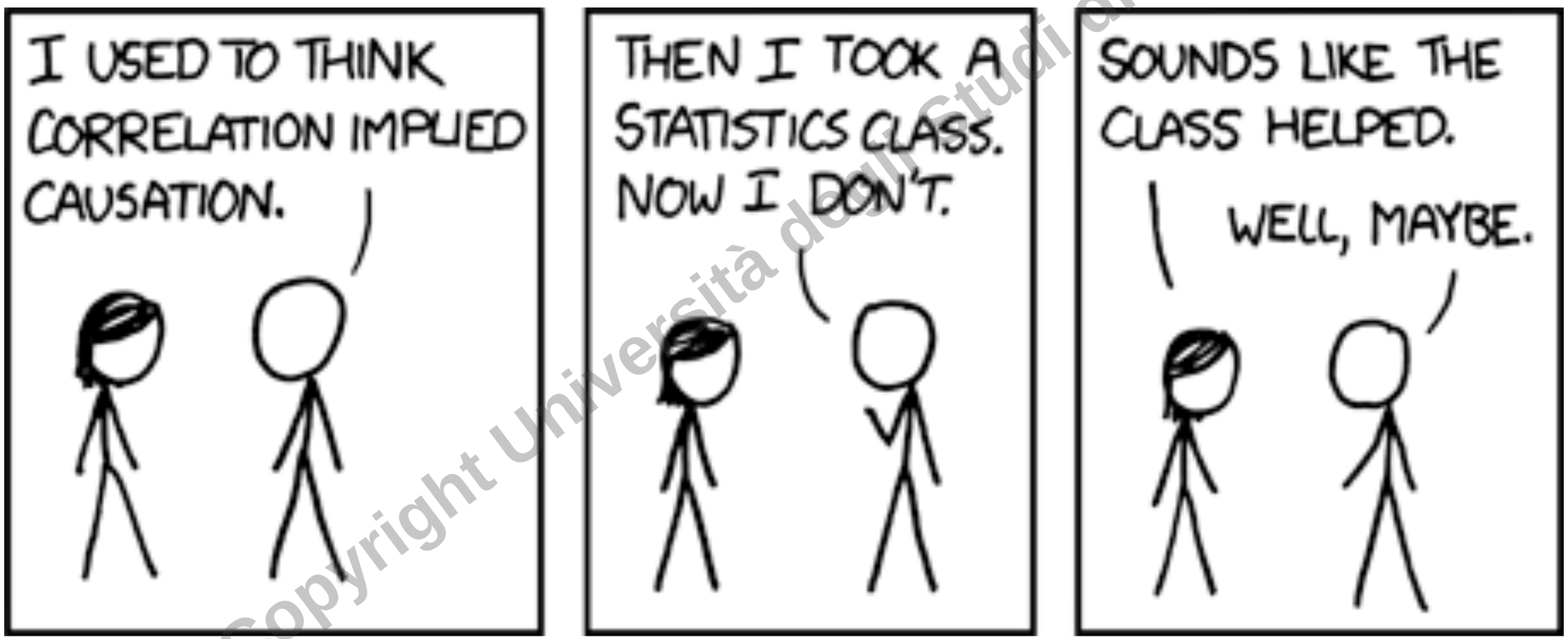
- The error of a measurement is the square root of the variance of its distribution, or its **standard deviation**; this is fully justified only for normal distributions.

Correlation and independence



Different sets with various degrees of correlation

Correlation and independence



Change of random variable

- Suppose X is a random variable with pdf f_X.
- Define Y as a random variable that is a (deterministic) function of X through g: that is, $y = g(x)$
- Than the pdf f_Y of Y can be computed from

$$f_Y(y) = \int \delta\big(y - g(x)\big) f_X(x)\,\mathrm{d}x$$

- This equation holds also for the more general case where $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$.
- If $n = m = 1$, one can also use the cumulative distributions of x and y: corresponding events must have the same probability.

Galaxies!

Problem. Astronomers measure the luminosity (flux) L of stars and galaxies a using a logarithmic scale called *magnitude* m:

$$m = -2.5 \log_{10} \frac{L}{L_{\text{ref}}}$$

Suppose that all galaxies have the same luminosity and are randomly distributed in space with uniform density (a uniform Poisson point process). What is the probability distribution for m?

Simple error propagation

Problem 1. Suppose Y is a random variable that depends on X through a non-linear equation $y = g(x)$. Using a Taylor expansion, write $\mathrm{Var}[Y]$ in terms of $\mathrm{Var}[X]$.

Problem 2. Repeat the problem for the case where X and Y are vector random variables.

Problem 3. Repeat the problem for the case where the relation between x and y is expressed implicitly through $g(x, y) = 0$.

Simple inference

Problem. We "measure" a set of N values generated by a Gaussian distribution with unknown average μ and standard deviation σ. What can we infer about μ and σ?

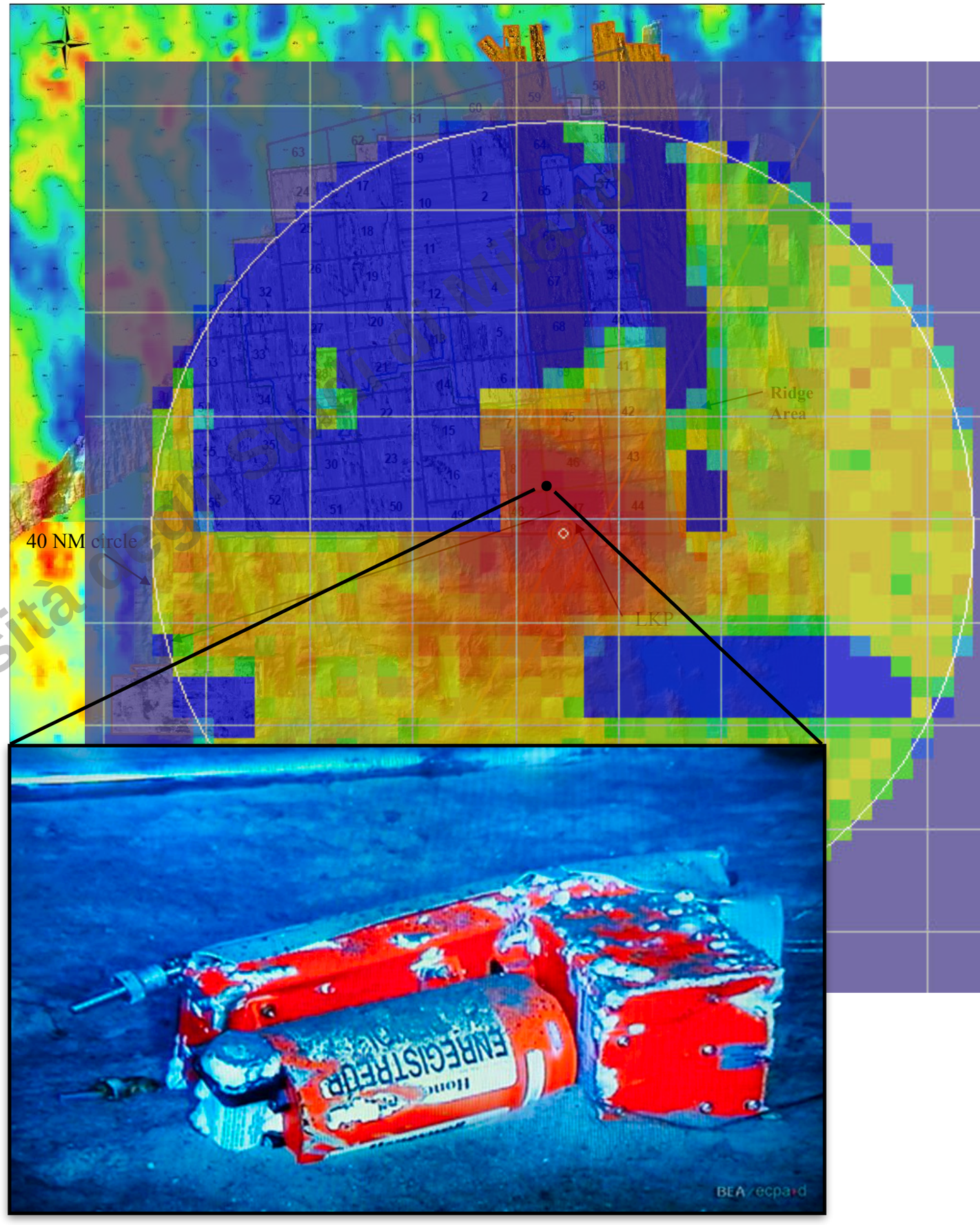
Problem continued. What are the most likely values for μ and σ? What is the most likely value of μ independently of the value of σ? What is the most likely value for σ independently of the value of μ?

Bayesian inference

A real case

- Flight AF447 disappeared in the middle of the Atlantic on June 1, 2009, killing 228 people
- The search for the aircraft has been unsuccessful for two years
- ...until Bayesian techniques were used
- The wreck was then found within one week of underwater search

[Statistical Science 29(1): 69–80, doi: 10.1214/13-STS420]

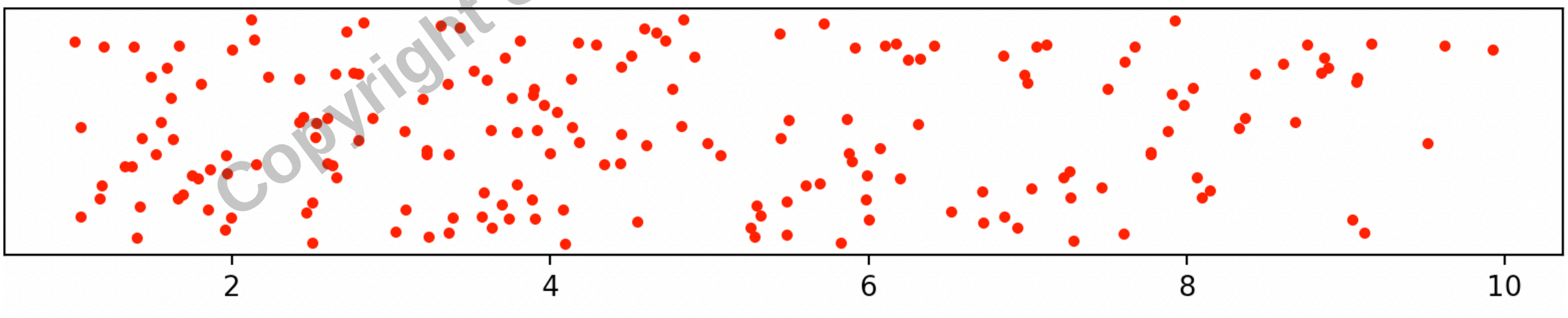


Priors

- Bayesian inference is based on a somewhat arbitrary choice of priors.
- It is not a bug, it is a feature!
- Forces us to state our assumption clearly
- Makes inference trivial and reproducible
- We can then perform “higher” inference on the assumptions
- We can marginalize over the assumptions

“Difficult” inference

Problem. Unstable particles are emitted from a source and decay at a distance x, a real number that follows an exponential distribution with length λ. We can measure x only within a window between $x = 1$ cm and $x = 10$ cm. What can we infer about λ?



Galaxies!

Problem. Astronomers measure the luminosity (flux) L of stars and galaxies a using a logarithmic scale called *magnitude* m:

$$m = -2.5 \log_{10} \frac{L}{L_{\text{ref}}}$$

Suppose that all galaxies have the same luminosity and are randomly distributed in space with uniform density (a uniform Poisson point process). What is the probability distribution for m?

“Difficult” inference

Problem. Unstable particles are emitted from a source and decay at a distance x, a real number that follows an exponential distribution with length λ. We can measure x only within a window between $x = 1$ cm and $x = 10$ cm. What can we infer about λ?

Problem’. Astronomers can measure magnitudes of galaxies only on a given range (say $m \in [6,15]$). Suppose we have a list of measured galaxy magnitudes $\{m_i\}$: what can we say about the parameters of the magnitude distribution?

0° < Glon < 360°
|Glat| > 25°
Total Number of Galaxies: 1094532
Total Area:
J−band: 23758 deg^2
H−band: 23764 deg^2
K−band: 23778 deg^2
Ave Density: 2.78
J−band (1.2 μm)
H−band (1.6 μm)
Ks−band (2.2 μm)
data/model from Glazebrook et al (1994)
Gardner et al (1997)

log(number mag^{-1} deg^{-2})

J, H, or K_s mag

Bayesian model inference

So far we have been using Bayes' theorem to infer simple parameters. This inference relies on assumed assumptions (likelihood form, prior):

$$P(\theta \mid D, H) = \frac{P(D \mid \theta, H) P(\theta \mid H)}{P(D \mid H)}$$

What if we want to test the assumption H? Apply Bayes' theorem again!

$$P(H \mid D) = \frac{P(D \mid H) P(H)}{P(D)}$$

The evidence enters as a new likelihood in this second-order inference!

Bayesian model inference

The trouble with this approach is that often we do not know how to compute the new evidence:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} = \frac{P(D|H)P(H)}{\sum_{H'} P(D|H')P(H')}$$

For many problem, the set of possible hypotheses $\{H\}$ (or theories) is not known and is arbitrarily large!

Again, this shows a significant feature of Bayesian reasoning: it is a “Popperian” approach, where a hypothesis can never be fully accepted!

Evidence ratio

- In a typical situation one has a few (two?) possible models to check with the data.
- The ratio of the posterior distribution for the models is

$$\frac{P(H_1 \mid D)}{P(H_2 \mid D)} = \frac{P(D \mid H_1)}{P(D \mid H_2)} \frac{P(H_1)}{P(H_2)}$$

- The first term is proportional to the ratio of the evidences (how well does the model fit the data?)
- The second term is proportional to the ratio of the priors (what do we think about the models?)
- What would be the frequentist approach here?

Problems

Problem 1. Suppose that you toss a *possibly bent* coin 12 times and you get 3 heads. What can you say about the coin? Is it bent or just a normal fair coin? [Hint: compare two models, a fair coin vs. a bent coin with parameter f]

Problem 2. In the previous problem show that the logarithm of the evidence ratio can grow at most as linear if the coin is bent, but as $\log N_{\text{data}}$ if the coin is fair.

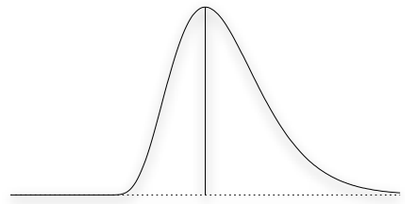
Legal evidence

Problem. Two people have left traces of their own blood at the scene of a crime: the traces are of type O (a common type in the local population, with frequency 60%) and of type AB (rare, frequency 1%). A suspect is tested, and his blood is of type O. Do all these data give evidence in favour of the proposition that the suspect was one of the two people present in at the crime?

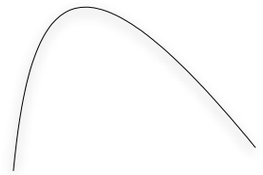
Bayesianism vs. sampling theory

- One of the reasons advocated to prefer a classical (frequentist) analysis over a Bayesian one is computer time
    - Bayesian inference can be computationally challenging because (in principle) one need to compute the evidence
    - Integration in high-dimensions is very hard (remember the problem with the sphere in $\mathbb{R}^n$)
- A possible way out is to use **conjugate priors**, i.e. priors that when multiplied with a likelihood produce posteriors of the same family of distributions
    - This is not always easy or possible to do, especially with complex likelihoods
- An alternative approach is to use an approximation called **Laplace's method**

Laplace's method in 1D



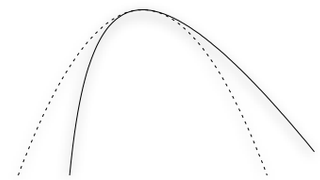
- Suppose we have an un-normalized probability density P^* and we want to find its normalisation constant $Z_P \equiv \int P^*(x)\,\mathrm{d}x$



- If P^* has a single peak at x_0, we can Taylor-expand its log

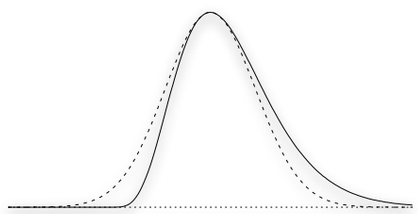
$$\log P^*(x) \simeq \log P^*(x_0) - \frac{c}{2}(x - x_0)^2 + \cdots$$

$$\text{where } c = -\left.\frac{\partial^2}{\partial x^2} \log P^*(x)\right|_{x_0}$$



- Truncating the expansion is equivalent to a Gaussian approximation

$$P^*(x) \simeq Q^*(x) \equiv P^*(x_0)\,\exp\left[-c(x - x_0)^2/2\right]$$



- This gives an approximate estimate for the normalization of P^*

$$Z_P \simeq P^*(x_0)\sqrt{2\pi/c}$$

Laplace's method KD

- The method can be generalised to a K dimensional space.
- We need a multi-dimensional Taylor expansion. Call

$$A_{ij} = -\frac{\partial^2}{\partial x_i \partial x_j} \log P^*(\boldsymbol{x}) \Big|_{\boldsymbol{x}=\boldsymbol{x}_0}$$

- Using the same argument above, we find a multivariate normal distribution for $Q(\boldsymbol{x})$ and the normalising constant is found to be

$$Z_P \simeq Z_Q = P^*(\boldsymbol{x}_0) \sqrt{\frac{(2\pi)^K}{\det A}}$$

Laplace's method: exercise

Problem. A photon counter measure the photon emission rate λ from a source. Suppose that in one minute the counter detects n photons, infer the posterior for λ? Try with a flat prior in λ and in $\log \lambda$. [Hint: use a Poisson distribution for n. Also, compare the final result with Stirling's approximation to the factorial.]

Observation. Laplace's method is not basis-independent: the answer changes depending on the choice for the parametrization. This can be taken as advantage: a suitable choice can make the approximation more accurate.

Conjugate priors

- A conjugate prior is a prior which, when used in a Bayesian inference with a given likelihood, produces a posterior from the same family as the prior.
- This makes Bayesian inference particularly simple: the normalisation of the posterior (evidence) poses no issues, because it is the normalisation of a known distribution.
- The choice of a conjugate prior is strictly dependent on the likelihood; not all likelihoods admit a conjugate prior.

Problem. Suppose we have independent measurements $\{x_n\}$ of an unknown quantity μ, with each measurement $x_n \sim N(\mu, \sigma^2)$. If we choose for $P(\mu)$ a normal distribution, then show that $P(\mu \mid \{x_n\})$ will also be normal.

Exponential family (1D)

- A very large of probability distributions that admit simple conjugate priors is the exponential family.
- This family, in the 1D case, is made of all probability (density) functions that can be factorized as

$$P(x \mid w) = h(x)\ \exp[\eta(w) \cdot T(x) - A(w)]$$

- It is also required that the support of $P(x \mid w)$ does not depend on w.
- Example: the Poisson distribution belongs to this family:

$$h(x) = 1/x!\,, \quad \eta(w) = \ln w\,, \quad T(x) = x\,, \quad A(w) = w\,.$$

- If $\eta(w) = w$ the distribution is said to be in *canonical or natural form*.

Problem. Show that the binomial distribution is part of the exponential family.

Exponential family (*n*D)

- The definition can be simply generalized to vector parameters and probability densities defined in $\mathbb{R}^n$:
  $P(\boldsymbol{x} \mid \boldsymbol{w}) = h(\boldsymbol{x})\ \exp[\boldsymbol{\eta}(\boldsymbol{w}) \cdot \boldsymbol{T}(\boldsymbol{x}) - A(\boldsymbol{w})]$
- In this equation $\boldsymbol{\eta}$: $\mathbb{R}^k \to \mathbb{R}^s$, $\boldsymbol{T}$: $\mathbb{R}^n \to \mathbb{R}^s$ and the product is a scalar product; all other functions are real valued
  - $\boldsymbol{T}(\boldsymbol{x})$ is a **sufficient statistics** for the distribution: it has all the information $\boldsymbol{x}$ provides on the parameter $\boldsymbol{w}$;
  - $\boldsymbol{\eta}(\boldsymbol{w})$ is called the **natural parameter**;
  - $A(\boldsymbol{w})$ is called the **log-partition function** and is linked to the normalization of the distribution.

Exponential family

Conjugate priors

- Suppose a likelihood is written as products of terms like
  $P(\boldsymbol{x}_i \,|\, \boldsymbol{w}) = h(\boldsymbol{x}_i)\ \exp[\boldsymbol{\eta}(\boldsymbol{w}) \cdot \boldsymbol{T}(\boldsymbol{x}_i) - A(\boldsymbol{w})]$
- Define the prior as $P_\pi(\boldsymbol{w} \,|\, \boldsymbol{\alpha}, \beta) = g(\boldsymbol{\alpha}, \beta)\ \exp[\boldsymbol{\eta}(\boldsymbol{w}) \cdot \boldsymbol{\alpha} - \beta A(\boldsymbol{w})]$,
  where $\boldsymbol{\alpha} \in \mathbb{R}^s$ and $\beta > 0$ are two parameters

Problem 1. Show that $P_\pi(\boldsymbol{w} \,|\, \boldsymbol{\alpha}, \beta)$ belongs to the exponential family.

Problem 2. Show that a Bayesian inference produces a posterior of the same form as the prior, with

$$\boldsymbol{\alpha}' = \boldsymbol{\alpha} + \sum_{i=1}^{I} \boldsymbol{T}(\boldsymbol{x}_i) \quad \text{and} \quad \beta' = \beta + I.$$

Conjugate priors

A few examples

Distribution	Parameters	Conjugate prior	Prior iperparameters
Binomial	probability	Beta	2 shape parameters
Poisson	rate	Gamma	shape and rate
Normal	mean	Normal	mean and variance
Normal	precision	Gamma	shape and rate
Normal	mean and precision	Normal-gamma	4 parameters
Exponential	rate	Gamma	shape and rate

Occam's razor

Pluritas non est ponenda sine necessitate

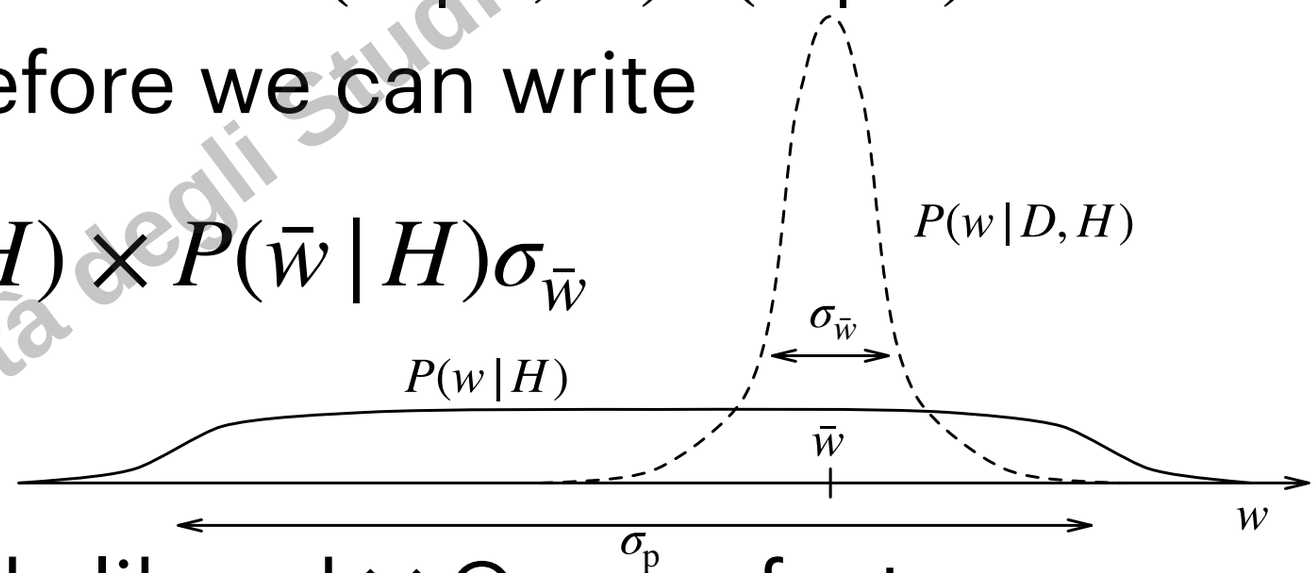
- Bayesian model inference includes "for free" a formulation of Occam's razor.
- This is a consequence of the normalization of the prior:
    - A prior that covers a large volume in the parameter space will have small values
    - This will have a direct consequence in the evidence associated to that prior
- This perfectly works for models with different dimensions in their parameters space
    - No need to talk about degrees of freedom!

Occam factor

The evidence is $P(D\,|\,H) = \int P(D\,|\,w, H)P(w\,|\,H)\,\mathrm{d}w$.

For many problems the posterior $P(D\,|\,w, H)P(w\,|\,H)$ has a narrow peak at $\bar{w}$ and therefore we can write

$$P(D\,|\,H) \simeq P(D\,|\,\bar{w}, H) \times P(\bar{w}\,|\,H)\sigma_{\bar{w}}$$



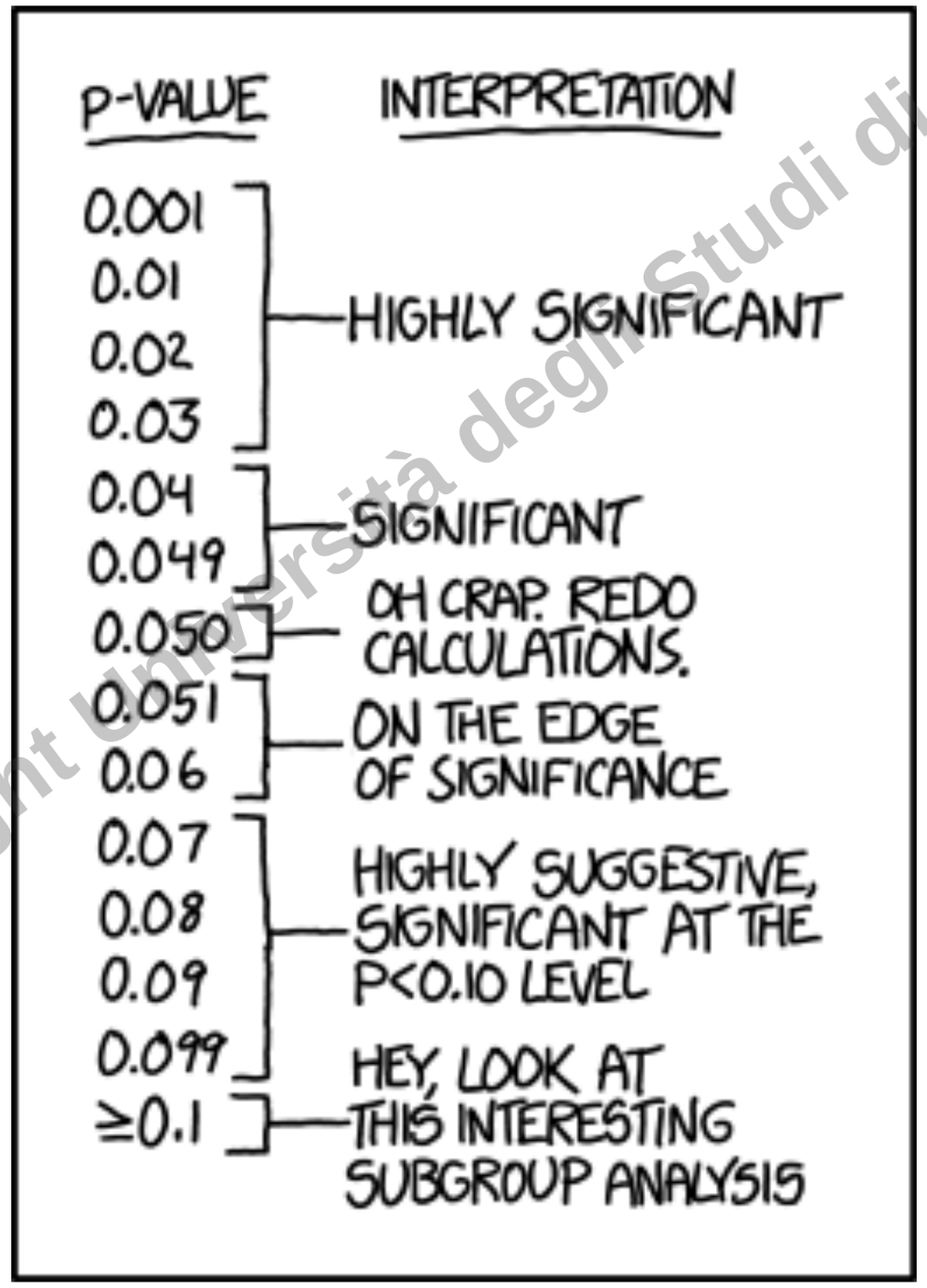
or equivalently

Evidence $\simeq$ Best fit likelihood $\times$ Occam factor

- $\sigma_{\bar{w}}$ is the posterior uncertainty around $\bar{w}$
- If the prior is approximately uniform on a some large interval σ_{p} then Occam's factor is $\sigma_{\bar{w}}/\sigma_{\mathrm{p}} \ll 1$

p-values

If all else fails, use "significant at a *p* > 0.05 level" and hope no one notices.



Vaccine efficacy

Problem. A new vaccine against a disease is being tested. The vaccine is given to N_A people, while N_B people get a placebo. Both groups are monitored: after a year, we find that s_A people and s_B people from the two groups got sick. Is the vaccine working? What is an estimate of its efficacy?

A real case. For the "ChAdOx1 nCoV-19" AstraZeneca SARS-Cov-2 vaccine the main study reports the values in the table. [Lancet, 2021; 397: 99–111]

	A	B
N	4440	4455
s	27	71

p-values & conf. intervals

Beware! Evil things are coming...

Problem 1. In an expensive laboratory Dr. Jack tosses the coin 12 times; the outcome is HHHTHHHHHTHHT. Is the coin biased in favour of H?

Problem 2. In an even more expensive laboratory, we make this experiment. Dr. Jack extracts a number between 1 and 100. Then it tosses a fair coin twice, and each time it reveals the number if the coin is head, or the number + 1 if the coin is tail. Bill, a student of Dr. Jack, has to provide the confidence intervals for the original number extracted. How should Bill proceed?

Monte Carlo methods

Monte Carlo methods

In high-dimensionality it is very difficult to obtain the posterior probability, and is often uninteresting!

Sampling the distribution is often a better and easier choice.

Problem. Suppose that you know the posterior probability for your n-dimensional parameter $\boldsymbol{x}$, $P(\boldsymbol{x})$. Suppose you want to have the probability distribution for a related m-dimensional variable, $\boldsymbol{y} = g(\boldsymbol{x})$. How can you obtain it? Suppose instead you have a *sampling* of $\boldsymbol{x}$, how can you obtain a *sampling* of $\boldsymbol{y}$?

Sampling a distribution

Problem. Suppose we want to compute $\Phi \equiv \langle \phi(\boldsymbol{x}) \rangle = \int \phi(\boldsymbol{x}) P(\boldsymbol{x}) \, \mathrm{d}^S x$. If we have a set of random samples $\{\boldsymbol{x}_n\}$, with each $\boldsymbol{x}_n \sim P$, than an estimate of Φ is given by

$$\hat{\Phi} = \frac{1}{N} \sum_{n=1}^{N} \phi(\boldsymbol{x}_n) \, .$$

What is the expected "error" associated with $\hat{\Phi}$?

Sampling a high-dimensional distribution is still a very challenging task!

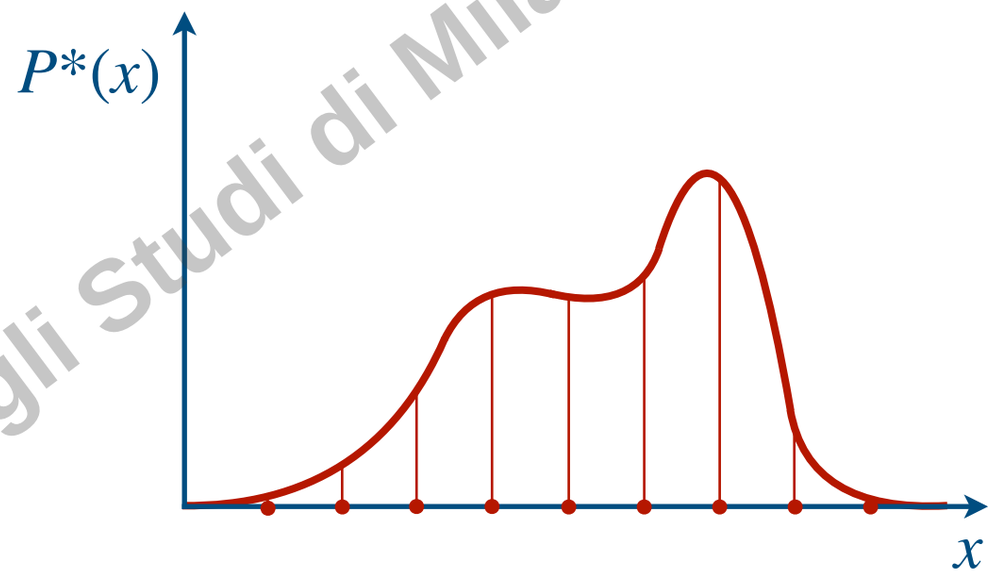
- We need to know the normalizing constant (i.e. the evidence in for the posterior)
- We need to know where $P(\boldsymbol{x})$ is big (i.e., check the entire parameter space!)

Luckily, we have a set of different sampling techniques suitable for our purposes!

Converting densities to discrete distributions

(Poisson spatial process)

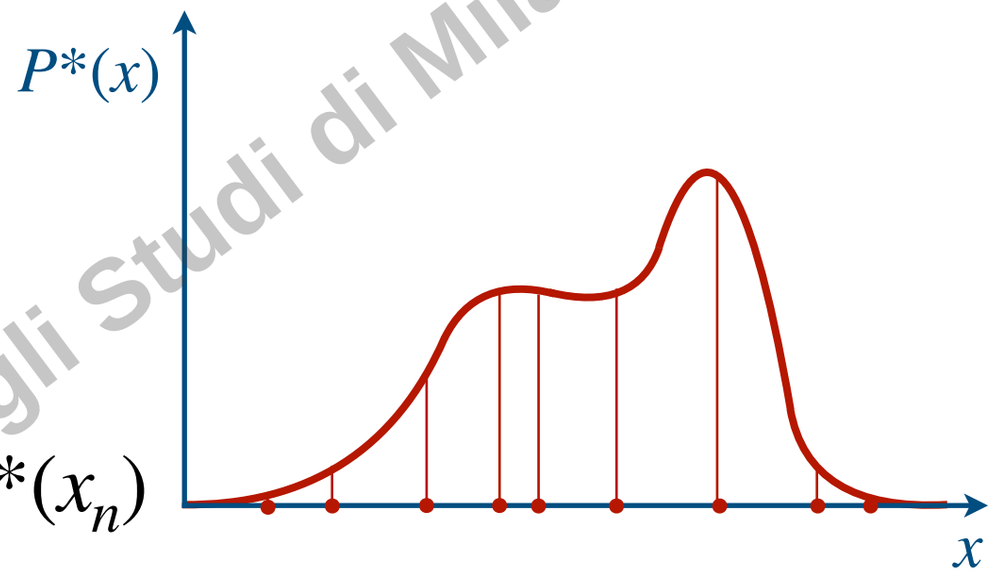
We can easily generate points according to $P^*(x)$ if we discretize this function.



- Divide the interval into equally spaced points $\{x_n\}$.
- Compute $p_n^* = P^*(x_n)$ for each point.
- **Problem.** Can you find a way to generate a x_n according to the p_n?
  - Compute $P_c^* = \sum_{n=1}^{c} p_n^*$, then generate a point uniformly from 0 to P_N^*
  - Take x_c, where c is chosen to be the largest integer such that P_c^* is smaller than the point generated.
- Can this method be generalised to multi-dimensional distributions?

Uniform sampling

- Generate a set of points $\{x_n\}$ uniformly on the support of $P^*(x)$.
- Compute $p_n^* = P^*(x_n)$ for each point.
- Evaluate the normalisation $Z = \sum_n P^*(x_n)$



- Given a function $\phi(x)$, estimate $\Phi = \langle \phi(x) \rangle$ as

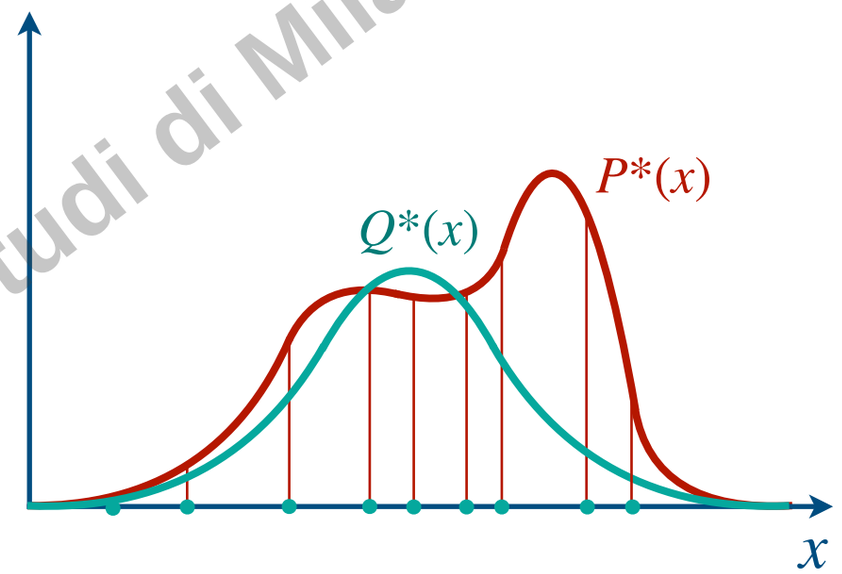
$$\hat{\Phi} = \frac{1}{Z} \sum_n P^*(x_n)\, \phi(x_n)$$

Problem. Can you estimate the variance associated with $\hat{\Phi}$? How does this quantity depend on the dimensionality of the problem when you generalise the algorithm to multi-dimensional distribution densities?

Importance sampling

- Suppose we have a probability density Q^* which is not too dissimilar from P^* and such that $\Omega_{P^*} \subset \Omega_{Q^*}$.
- Suppose also we can easily generate a set of points $\{x_n\}$ according to Q^*.
- Compute $w_n = P^*(x_n)/Q^*(x_n)$ for each point.
- Given a function $\phi(x)$, estimate $\Phi = \langle \phi(x) \rangle$ as

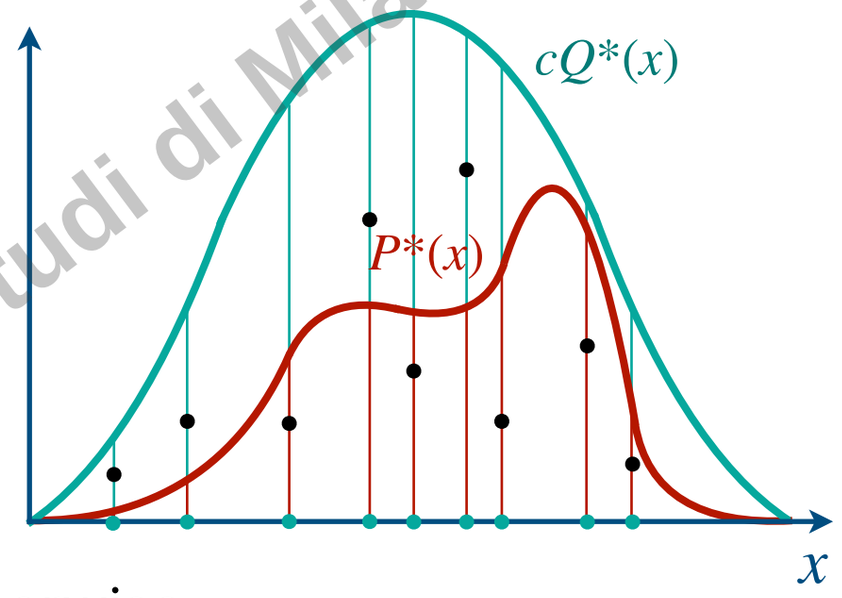
$$\hat{\Phi} = \sum_n w_n \, \phi(x_n) \Big/ \sum_n w_n$$



Problem. Can you estimate the variance associated with $\hat{\Phi}$? How does this quantity depend on the dimensionality of the problem when you generalise the algorithm to multi-dimensional distribution densities?

Rejection sampling

- Suppose we have a probability density Q^* such that $\exists c \,|\, cQ^*(x) \geq P^*(x)\, \forall x.$
- Suppose also we can easily generate a set of points $\{x_n\}$ according to Q^*.
- Generate u_n uniformly in the range $[0,\, cQ^*(x_n)]$
- Accept x_n if $P^*(x_n) \geq u_n$ and discard the point otherwise.
- The set of accepted points $\{x_n\}$ is distributed according to $P(x)$; therefore, given a function $\phi(x)$, we can estimate $\Phi = \langle \phi(x) \rangle$ simply as $\hat{\Phi} = \sum_n \phi(x_n) \big/ N.$



Problem. Can you estimate the variance associated with $\hat{\Phi}$? How does this quantity depend on the dimensionality of the problem when you generalise the algorithm to multi-dimensional distribution densities? How efficient is this algorithm?

Problems

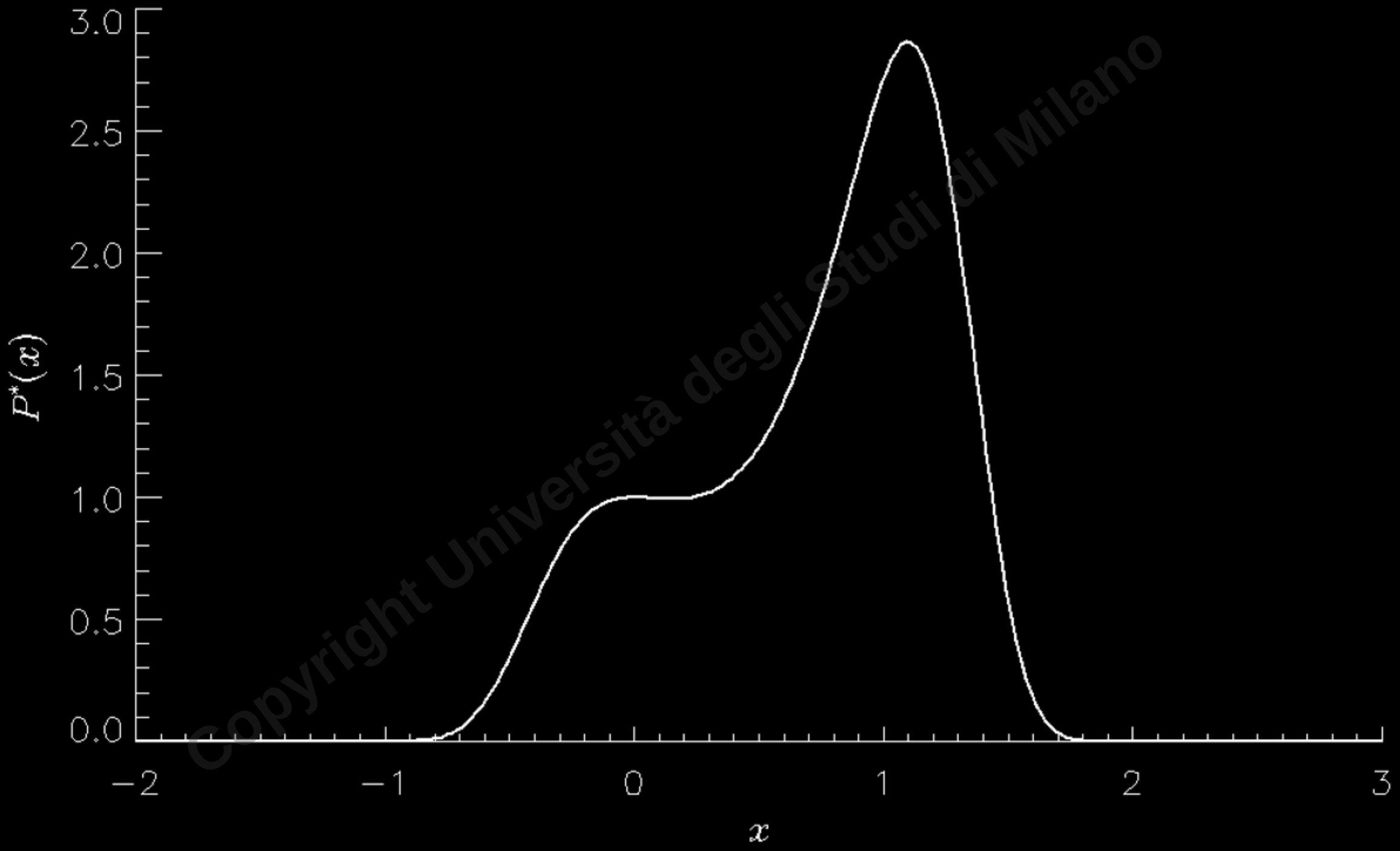
Problem 1. Generate $N = 1000$ points according to the following two probability densities

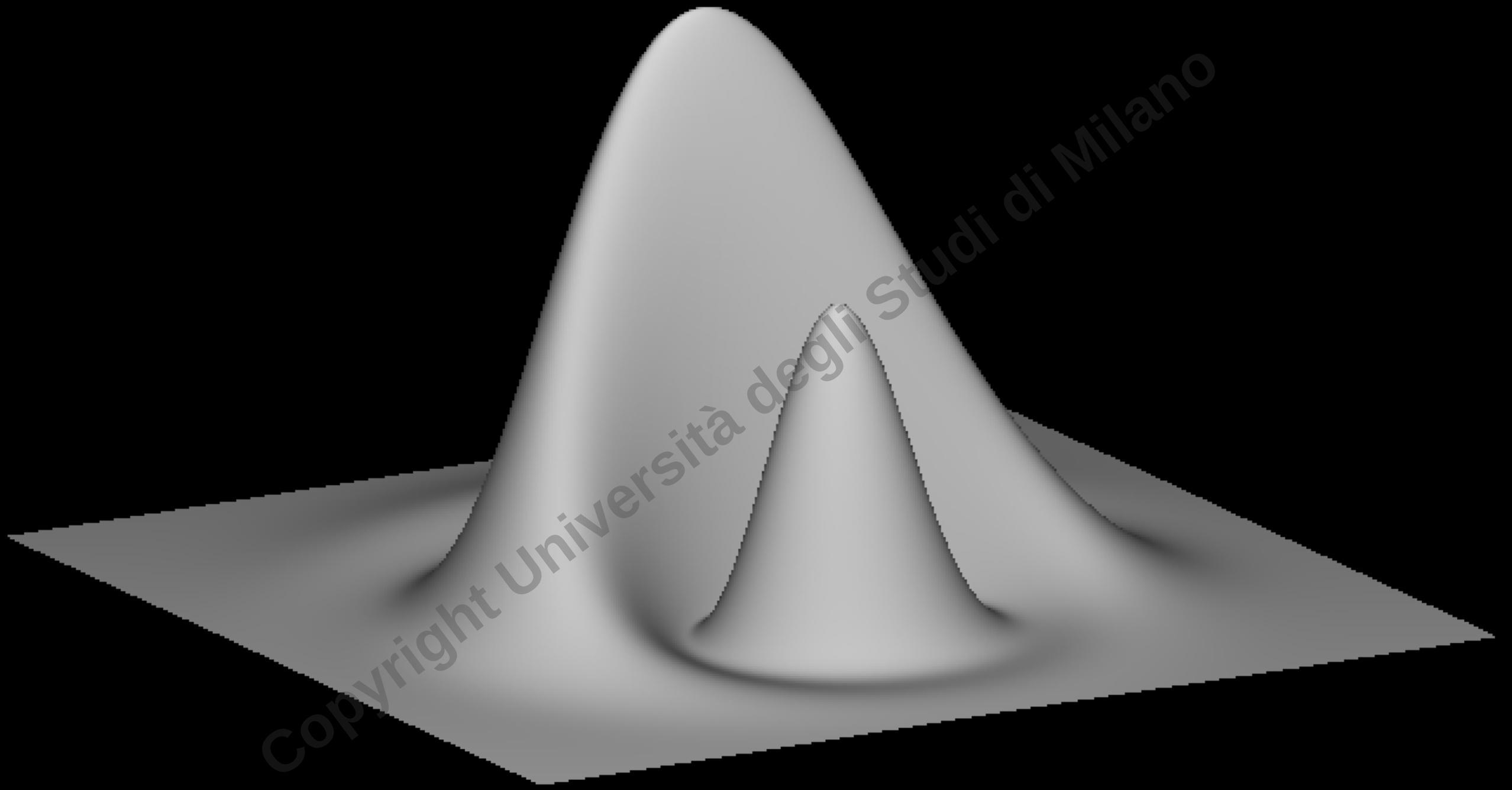
$$P_1(x) \propto \begin{cases} \mathrm{e}^{-x/2} & \text{for } x \in [1,10] \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad P_2(x) \propto \mathrm{e}^{-3x^4+5x^3-x^2}$$

Evaluate the average of $\phi(x) = x^3$ when x is distributed according to these two probability densities.

Problem 2. Generate $N = 1000$ points according to the 2D density $P(\boldsymbol{x}) = \mathrm{e}^{-|\boldsymbol{x}|^2} \cos^2(2\,|\boldsymbol{x} - \boldsymbol{x}_0|)$ for $\boldsymbol{x}_0 = (0,1)$.

Evaluate the average of $\phi(\boldsymbol{x}) = x_1^2 x_2$.





Markov chains

Definition. A **discrete-time** stochastic process that satisfy the rule

$$P(x_t \,|\, x_{t-1}, x_{t-2}, \ldots, x_{t-k}) = P(x_t \,|\, x_{t-1})$$

In other words, a state only depends on the state at the previous step! This is referred to as the **Markov property**.

Markov chains can have **discrete states** (or **finite states**) if the possible values for each state x_t are discrete (or finite).

Moreover, in our discussion we will consider only **time-homogeneous** Markov chains: for these the transition probability does not depend on the time-step.

Finite state Markov chain

A state i **communicate** to state j if there is a possible path that connects the two both ways: this is an equivalence relation. If any state can communicate with any other state, the chain is said to be **irreducible**.

A state is **recurrent** if the probability I will get back to it is one; otherwise it is **transient**. An irreducible chain has either recurrent or transient states.

The **period** of a state i is defined as

$$T_i = \text{gcm}\{t : P(x_t = i \,|\, x_0 = i) > 0\}$$

A state is **aperiodic** if $T_i = 1$.

Stationary distributions

A finite-state time-homogeneous MC can be represented by a transition matrix $\mathbf{P}$, where $P_{ij} = P(x_t = i \,|\, x_{t-1} = j)$. This allow us to compute the probability at step a given step as

$$\pi_t = \mathbf{P}\pi_{t-1} = \ldots = \mathbf{P}^t \pi_0$$

Definition. A distribution is said to be **stationary** if $\pi = \mathbf{P}\pi$.

The Perron-Frobenious theorem guarantees that an irreducible and aperiodic MC has a unique stationary distribution.

Under fairly the same conditions, the stationary distribution can be obtained from any starting distribution π_0 by takin the limit

$$\pi = \lim_{t \to \infty} P^t \pi_0$$

Random walk

Problem 1. A Markov chain over $\mathbb{R}$ is generated by taking $x_0 = 0$ and setting $x_t \,|\, x_{t-1} \sim \mathrm{N}(0, \sigma^2)$. What is the probability distribution for the t-th point? Write a program to check your result.

Problem 2. A discrete Markov chain over the integers $[0, N-1)$ is generated by takin $x_0 = 0$ and

$$P(x_t \,|\, x_{t-1}) = \begin{cases} 1/2 & \text{if } x_t = x_{t-1}, \\ 1/4 & \text{if } x_t - x_{t-1} \equiv \pm 1 \bmod N \\ 0 & \text{otherwise} \end{cases}$$

Use a vector-matrix notation to find the probability distribution for x_t. Is this chain irreducible? Does it admit recurrent states?

Detailed balance

The definition of a stationary distribution can be easily generalized to continuous MCs.

Definition. A probability density $\pi(x)$ is stationary if
$\int P(x \mid x')\, \pi(x')\, \mathrm{d}x' = \pi(x).$

For both discrete- and continuous-state MC, we can consider a closely related concept:

Definition. A probability density satisfies the detailed balance equation if $\pi(x)P(x' \mid x) = \pi(x')P(x \mid x')$. Essentially, this equation states that there is an identical "flux" of probabilities from state x to state x' and vice-versa.

Problem. Show that if $\pi(x)$ satisfies the detailed balance equation, then it is a stationary distribution.

Metropolis-Hastings algorithm

Idea. Sort of a rejection sampling, with a proposal density that depends on the current point!

1. Start with any point x_0 in the sample space.
2. Generate a new point x' using a chosen proposal distribution $Q(x' \mid x_t)$.
3. Compute
$$a = \frac{P^*(x')}{P^*(x_t)} \frac{Q(x_t \mid x')}{Q(x' \mid x_t)}$$
4. Accept the new point if $a \geq 1$; accept it with probability a if $a < 1$, or duplicate the original.
5. Go back to step 2.

Quick check. Show that this algorithm satisfies the detailed balance.

Metropolis: practical considerations

- The generated data-points, as in any MC, are **not independent**: they are characterised by a *correlation-time*
- The first points need to be discarded: they still have "memory" of x_0
- The algorithm has maximum efficiency when the correlation-time is minimised
- If s is a typical scale associated to $Q(x'|x)$, after T steps one will have covered approximately a length $s\sqrt{T}$
- One needs to run the chain for at least $T \simeq (L/s)^2$, where L is the scale of the posterior.

Efficient Markov chains

- A key aspect of the standard Metropolis algorithm is a **wise choice** of the **proposal distribution** $Q(x' \mid x)$
- small jumps will make the datapoints x and x' highly correlated
- small jumps will also make it difficult to discover isolated peaks in the posterior distribution
- long jumps will almost certainly result in a rejection of x'
- best if $Q(x' \mid x)$ mimics the shape of the posterior distribution, but that is unknown!
- A number of solutions exists: the most interesting one use the concept of affine-invariant steps

Clustering

K-means clustering

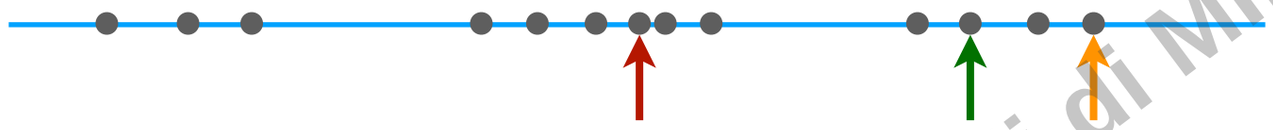
- One of the simplest techniques for **unsupervised learning** (with a funny name: *K* is just the number of clusters!)
- Can be applied to data represented as points in a metric space (typically, $\mathbb{R}^n$)
- Is based on the idea that, each datapoint, has associated a **latent variable**, the cluster to which it belongs
- Idea: try to find out the latent variable from the data themselves...
- ...and classify the data using the latent variables

Latent variables

- Suppose we know which cluster each point belongs
    - The definition of clusters is then trivial: just group the points belonging to each cluster!
- On the other hand, suppose we have a definition of clusters (e.g., as a partition of the point space)
    - The association of each point to its cluster is trivial!
- Idea: use the two steps above in turn, starting from random assignments or cluster definitions💬
- This is an example of **expectation-maximization, or E-M for short, algorithm** (probably the most famous one)

K-means clustering

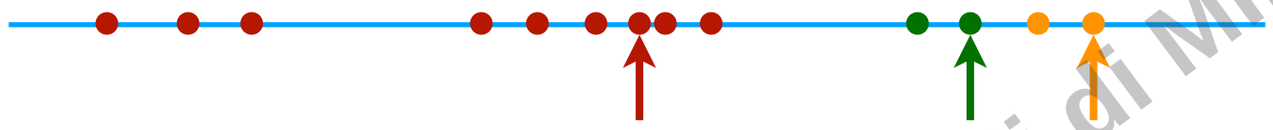
Description of the algorithm



1. Pick K random datapoints as tentative centers for the clusters
2. Assign each datapoint to the closest center (M-step)
3. For each cluster, compute its new center by taking the average of the coordinates of the datapoint associated to it (E-step)
4. Repeat steps 2 and 3 until the algorithm converges (i.e., step 2 does not change the assignment of each point)
5. [Optional] Compute the sum of the variances of each cluster
6. Repeat the entire algorithm from step 1 a few times; take the final result that is associated to the smallest sum of variances.

K-means clustering

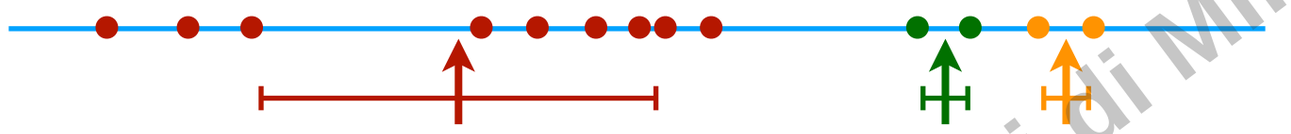
Description of the algorithm



1. Pick K random datapoints as tentative centers for the clusters
2. Assign each datapoint to the closest center (M-step)
3. For each cluster, compute its new center by taking the average of the coordinates of the datapoints associated to it (E-step)
4. Repeat steps 2 and 3 until the algorithm converges (i.e., step 2 does not change the assignment of each point)
5. [Optional] Compute the sum of the variances of each cluster
6. Repeat the entire algorithm from step 1 a few times; take the final result that is associated to the smallest sum of variances.

K-means clustering

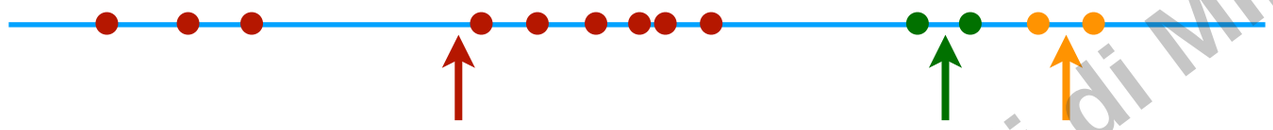
Description of the algorithm



1. Pick K random datapoints as tentative centers for the clusters
2. Assign each datapoint to the closest center (M-step)
3. For each cluster, compute its new center by taking the average of the coordinates of the datapoints associated to it (E-step)
4. Repeat steps 2 and 3 until the algorithm converges (i.e., step 2 does not change the assignment of each point)
5. [Optional] Compute the sum of the variances of each cluster
6. Repeat the entire algorithm from step 1 a few times; take the final result that is associated to the smallest sum of variances.

K-means clustering

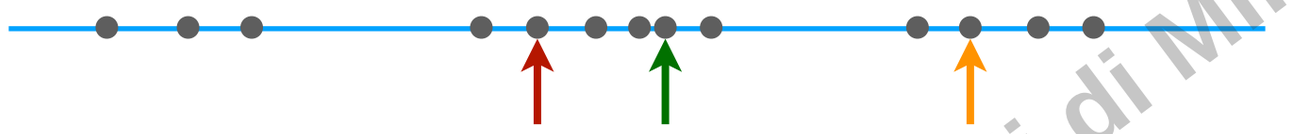
Description of the algorithm



1. Pick K random datapoints as tentative centers for the clusters
2. Assign each datapoint to the closest center (M-step)
3. For each cluster, compute its new center by taking the average of the coordinates of the datapoints associated to it (E-step)
4. Repeat steps 2 and 3 until the algorithm converges (i.e., step 2 does not change the assignment of each point)
5. [Optional] Compute the sum of the variances of each cluster
6. Repeat the entire algorithm from step 1 a few times; take the final result that is associated to the smallest sum of variances.

K-means clustering

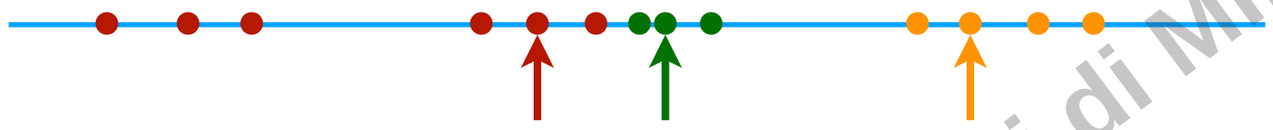
Description of the algorithm



1. Pick K random datapoints as tentative centers for the clusters
2. Assign each datapoint to the closest center (M-step)
3. For each cluster, compute its new center by taking the average of the coordinates of the datapoints associated to it (E-step)
4. Repeat steps 2 and 3 until the algorithm converges (i.e., step 2 does not change the assignment of each point)
5. [Optional] Compute the sum of the variances of each cluster
6. Repeat the entire algorithm from step 1 a few times; take the final result that is associated to the smallest sum of variances.

K-means clustering

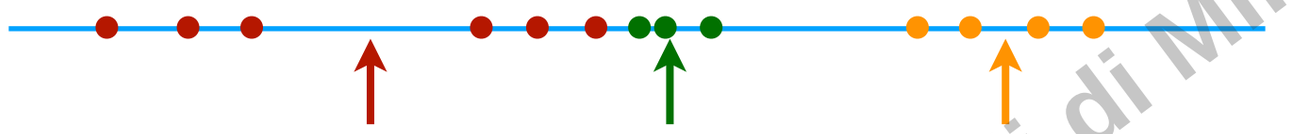
Description of the algorithm



1. Pick K random datapoints as tentative centers for the clusters
2. Assign each datapoint to the closest center (M-step)
3. For each cluster, compute its new center by taking the average of the coordinates of the datapoints associated to it (E-step)
4. Repeat steps 2 and 3 until the algorithm converges (i.e., step 2 does not change the assignment of each point)
5. [Optional] Compute the sum of the variances of each cluster
6. Repeat the entire algorithm from step 1 a few times; take the final result that is associated to the smallest sum of variances.

K-means clustering

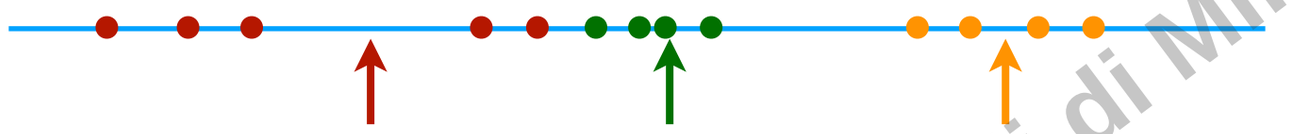
Description of the algorithm



1. Pick K random datapoints as tentative centers for the clusters
2. Assign each datapoint to the closest center (M-step)
3. For each cluster, compute its new center by taking the average of the coordinates of the datapoints associated to it (E-step)
4. Repeat steps 2 and 3 until the algorithm converges (i.e., step 2 does not change the assignment of each point)
5. [Optional] Compute the sum of the variances of each cluster
6. Repeat the entire algorithm from step 1 a few times; take the final result that is associated to the smallest sum of variances.

K-means clustering

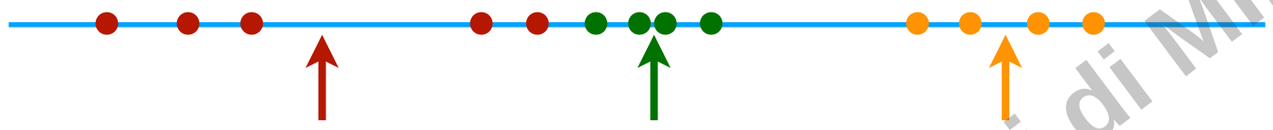
Description of the algorithm



1. Pick K random datapoints as tentative centers for the clusters
2. Assign each datapoint to the closest center (M-step)
3. For each cluster, compute its new center by taking the average of the coordinates of the datapoints associated to it (E-step)
4. Repeat steps 2 and 3 until the algorithm converges (i.e., step 2 does not change the assignment of each point)
5. [Optional] Compute the sum of the variances of each cluster
6. Repeat the entire algorithm from step 1 a few times; take the final result that is associated to the smallest sum of variances.

K-means clustering

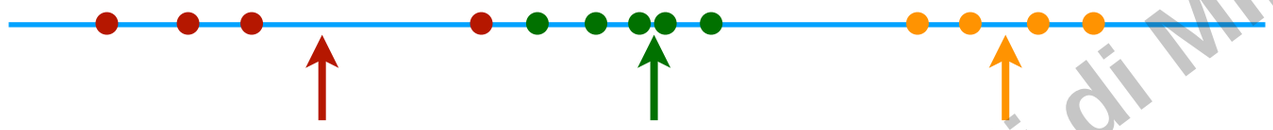
Description of the algorithm



1. Pick K random datapoints as tentative centers for the clusters
2. Assign each datapoint to the closest center (M-step)
3. For each cluster, compute its new center by taking the average of the coordinates of the datapoints associated to it (E-step)
4. Repeat steps 2 and 3 until the algorithm converges (i.e., step 2 does not change the assignment of each point)
5. [Optional] Compute the sum of the variances of each cluster
6. Repeat the entire algorithm from step 1 a few times; take the final result that is associated to the smallest sum of variances.

K-means clustering

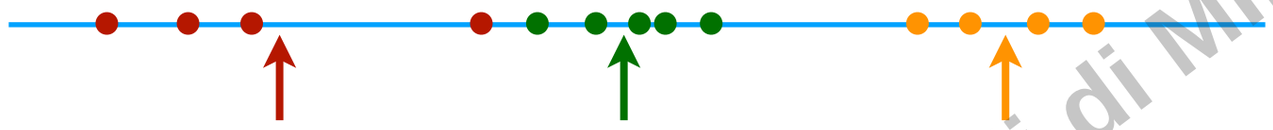
Description of the algorithm



1. Pick K random datapoints as tentative centers for the clusters
2. Assign each datapoint to the closest center (M-step)
3. For each cluster, compute its new center by taking the average of the coordinates of the datapoints associated to it (E-step)
4. Repeat steps 2 and 3 until the algorithm converges (i.e., step 2 does not change the assignment of each point)
5. [Optional] Compute the sum of the variances of each cluster
6. Repeat the entire algorithm from step 1 a few times; take the final result that is associated to the smallest sum of variances.

K-means clustering

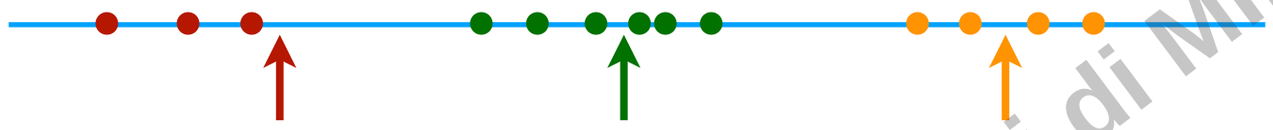
Description of the algorithm



1. Pick K random datapoints as tentative centers for the clusters
2. Assign each datapoint to the closest center (M-step)
3. For each cluster, compute its new center by taking the average of the coordinates of the datapoints associated to it (E-step)
4. Repeat steps 2 and 3 until the algorithm converges (i.e., step 2 does not change the assignment of each point)
5. [Optional] Compute the sum of the variances of each cluster
6. Repeat the entire algorithm from step 1 a few times; take the final result that is associated to the smallest sum of variances.

K-means clustering

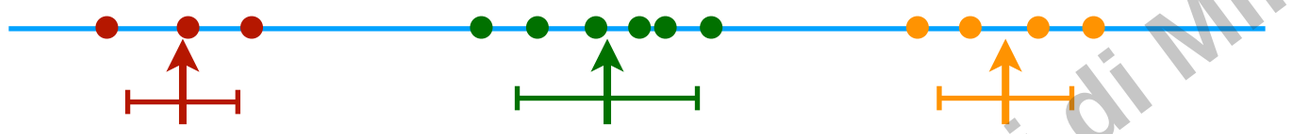
Description of the algorithm



1. Pick K random datapoints as tentative centers for the clusters
2. Assign each datapoint to the closest center (M-step)
3. For each cluster, compute its new center by taking the average of the coordinates of the datapoints associated to it (E-step)
4. Repeat steps 2 and 3 until the algorithm converges (i.e., step 2 does not change the assignment of each point)
5. [Optional] Compute the sum of the variances of each cluster
6. Repeat the entire algorithm from step 1 a few times; take the final result that is associated to the smallest sum of variances.

K-means clustering

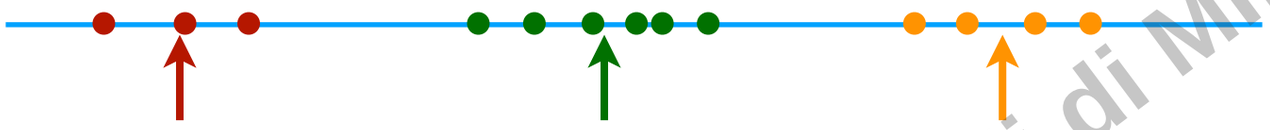
Description of the algorithm



1. Pick K random datapoints as tentative centers for the clusters
2. Assign each datapoint to the closest center (M-step)
3. For each cluster, compute its new center by taking the average of the coordinates of the datapoints associated to it (E-step)
4. Repeat steps 2 and 3 until the algorithm converges (i.e., step 2 does not change the assignment of each point)
5. [Optional] Compute the sum of the variances of each cluster
6. Repeat the entire algorithm from step 1 a few times; take the final result that is associated to the smallest sum of variances.

K-means clustering

Description of the algorithm



1. Pick K random datapoints as tentative centers for the clusters
2. Assign each datapoint to the closest center (M-step)
3. For each cluster, compute its new center by taking the average of the coordinates of the datapoints associated to it (E-step)
4. Repeat steps 2 and 3 until the algorithm converges (i.e., step 2 does not change the assignment of each point)
5. [Optional] Compute the sum of the variances of each cluster
6. Repeat the entire algorithm from step 1 a few times; take the final result that is associated to the smallest sum of variances.

K-means clustering

Details

- The algorithm can be easily written in $\mathbb{R}^n$
- Requires the concept of a **distance** (metric vector space) and thus works nicely only if such a concept makes sense
    - Cannot applied to different parameter spaces if one cannot normalize the various parameters somehow
    - Makes sense only if the clusters are approximately **spherical** and of the **same size**
- Is extremely simple and fast to implement
    - The stopping criterium is clear (no issues with floating point approximations)
    - There are no numerical issues with it

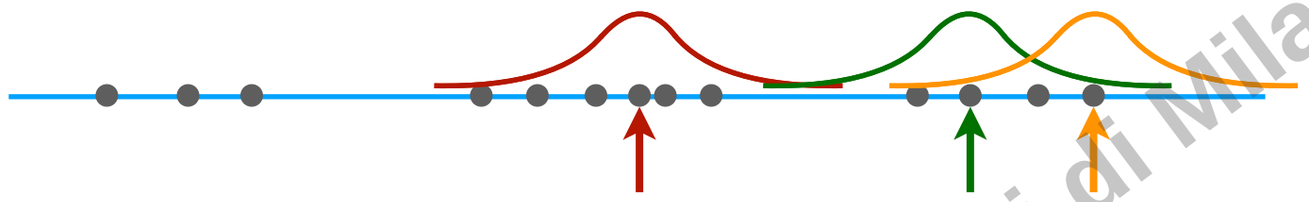
K-means clustering

Soft version

- Idea: use as latent variables non-binary "**responsibilities**," the probabilities that each data-point belongs to each cluster
- This requires a probability density that depends on the distance from the cluster center.
    - Typically, a normal distribution with fixed or variable width
- **E-step**: update the responsibilities by using the current cluster centers and widths (and their associated pdf)
- **M-step**: compute new cluster centers (and optionally cluster widths) by taking into account the responsibilities
- In case the chosen pdf is a multinomial normal distribution, use
    - weighted average for the cluster centers
    - [optional] weighted variances for the cluster widths
- Thus the point distribution is modelled as a (Gaussian) **mixture**

K-means clustering

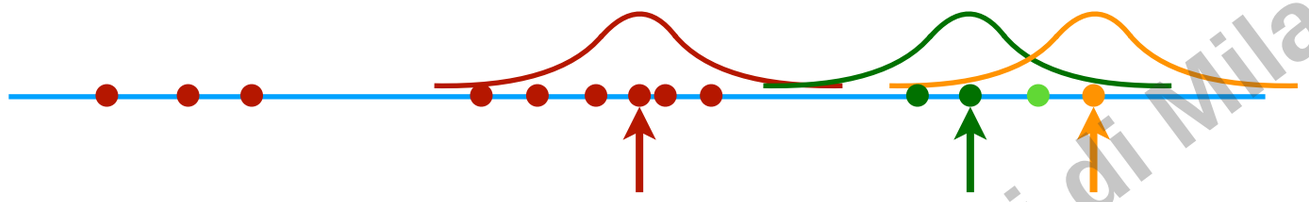
Soft clustering



1. Pick K random datapoints as tentative centers for the clusters; assign fixed width to each cluster distribution
2. Computes the responsibilities of each point (E-step)
3. For each cluster, compute its new center (and width) by taking the weighted average of the coordinates of the associated datapoints (M-step)
4. Repeat steps 2 and 3 until the algorithm converges (i.e., the change in the cluster centers and widths are small enough)
5. [Optional] Compute the sum of the variances of each cluster
6. Repeat the entire algorithm from step 1 a few times; take the final result that is associated to the smallest sum of variances.

K-means clustering

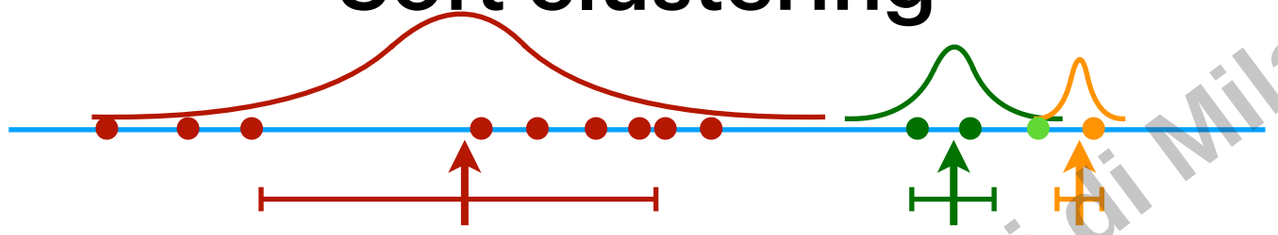
Soft clustering



1. Pick K random datapoints as tentative centers for the clusters; assign fixed width to each cluster distribution
2. Computes the responsibilities of each point (E-step)
3. For each cluster, compute its new center, width, and richness by taking the weighted average of the coordinates of the associated datapoints (M-step)
4. Repeat steps 2 and 3 until the algorithm converges (i.e., the change in the cluster centers and widths are small enough)
5. [Optional] Compute the (weighted) sum of the variances of each cluster
6. Repeat the entire algorithm from step 1 a few times; take the final result that is associated to the smallest sum of variances.

K-means clustering

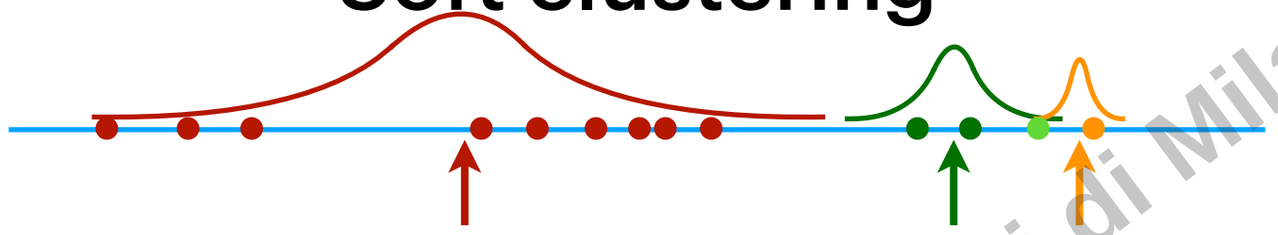
Soft clustering



1. Pick K random datapoints as tentative centers for the clusters; assign fixed width to each cluster distribution
2. Computes the responsibilities of each point (M-step)
3. For each cluster, compute its new center, width, and richness by taking the weighted average of the coordinates of the associated datapoints (E-step)
4. Repeat steps 2 and 3 until the algorithm converges (i.e., the change in the cluster centers and widths are small enough)
5. [Optional] Compute the (weighted) sum of the variances of each cluster
6. Repeat the entire algorithm from step 1 a few times; take the final result that is associated to the smallest sum of variances.

K-means clustering

Soft clustering

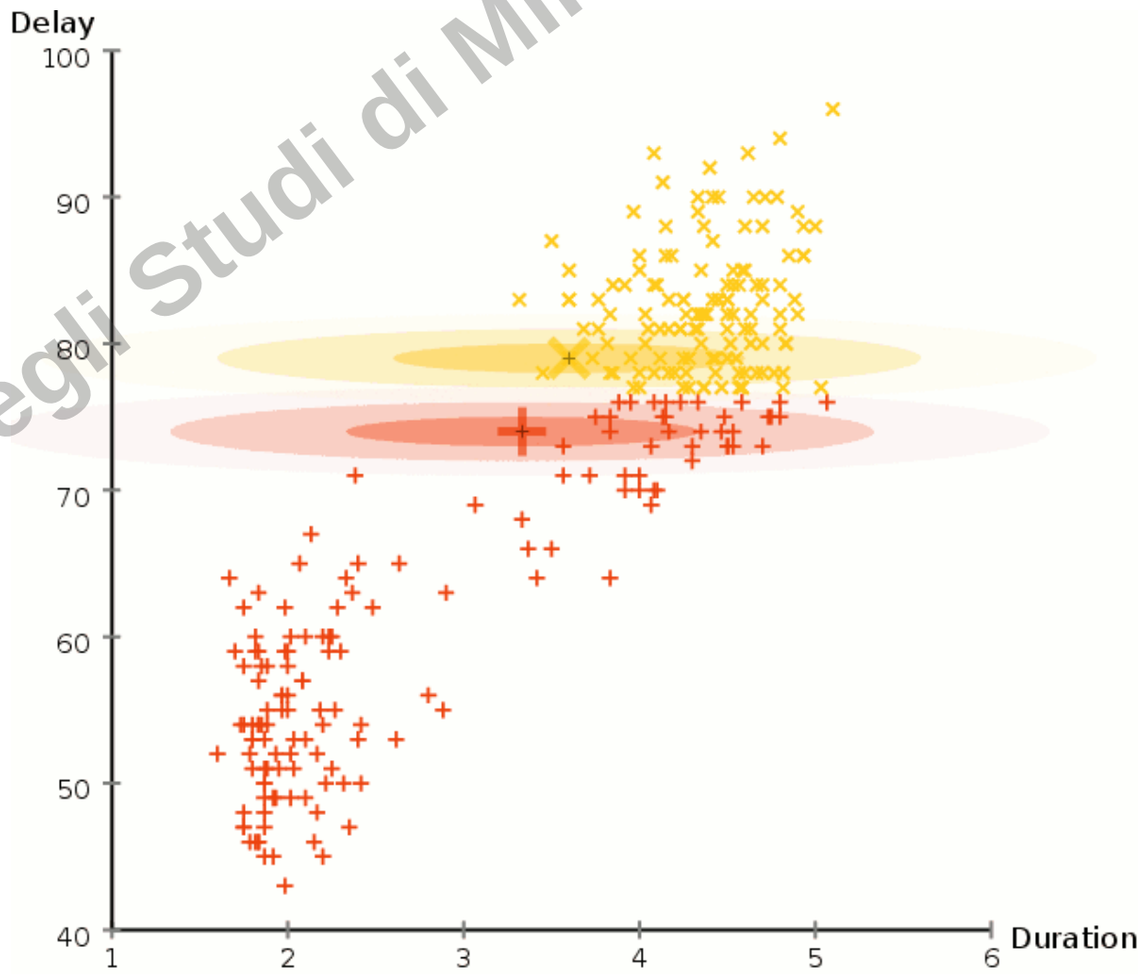


1. Pick K random datapoints as tentative centers for the clusters; assign fixed width to each cluster distribution
2. Computes the responsibilities of each point (M-step)
3. For each cluster, compute its new center, width, and richness by taking the weighted average of the coordinates of the associated datapoints (E-step)
4. Repeat steps 2 and 3 until the algorithm converges (i.e., the change in the cluster centers and widths are small enough)
5. [Optional] Compute the (weighted) sum of the variances of each cluster
6. Repeat the entire algorithm from step 1 a few times; take the final result that is associated to the smallest sum of variances.

Soft K-mean clustering

Variations

- The pdf used to describe the responsibilities is often a multivariate normal distribution (Gaussian Mixture Model, or GMM)
- For multidimensional data, one can, for example
    - Keep all covariances identical and multiples of the identity (“circular” same-size clusters)
    - Keep all covariances multiples of the identity (“circular” different-size clusters)
    - Keep all covariances diagonal (“elliptical” clusters with axes aligned to the coordinates)
    - Put no constraints on the covariances (general “elliptical” clusters)
- The method as been generalised to noisy and incomplete datapoints (so-called *extreme deconvolution*)



Soft K-mean clustering

Issues

- The algorithm is essentially a maximum-likelihood one
- As such, it suffers from a possible severe issue when a cluster can contain a single point
    - In this case the (co)variance goes to zero...
    - ...and a divergence is found!
    - The algorithm fails miserably
- It also fails to converge to the global maximum in case of a bad initial choice of the centers (similarly to the hard version)
- There are some techniques to mitigate these issues (split-and-merge algorithm)

K-mean clustering

Implementation

- Both the hard and the soft K-mean clustering are quite simple to implement (try!)
- Alternatively, one can opt for ready-to-use implementations in ***scikit-learn***
  - `sklearn.cluster.KMeans` for hard K-mean clustering
  - `sklearn.mixture.GaussianMixture` for soft K-mean clustering
- The library contains many other clustering algorithms, including `sklearn.mixture.BayesianGaussianMixture`, a *variational Bayesian* version of the soft K-mean clustering.