Data analysis and stochastic modeling

Lecture 1 – A gentle introduction to probability

Guillaume Gravier

guillaume.gravier@irisa.fr

What are we here for?

1. data from observations
   - see what the data looks like
   - describe the data: distribution, clustering, etc.
   - summarize the data
   - compare data
2. models for decision
   - infer more general properties
   - make a (stochastic) model of the data
   - make decisions: simulation, classification, prediction, etc.

$\Rightarrow$ Provide the elementary tools and techniques

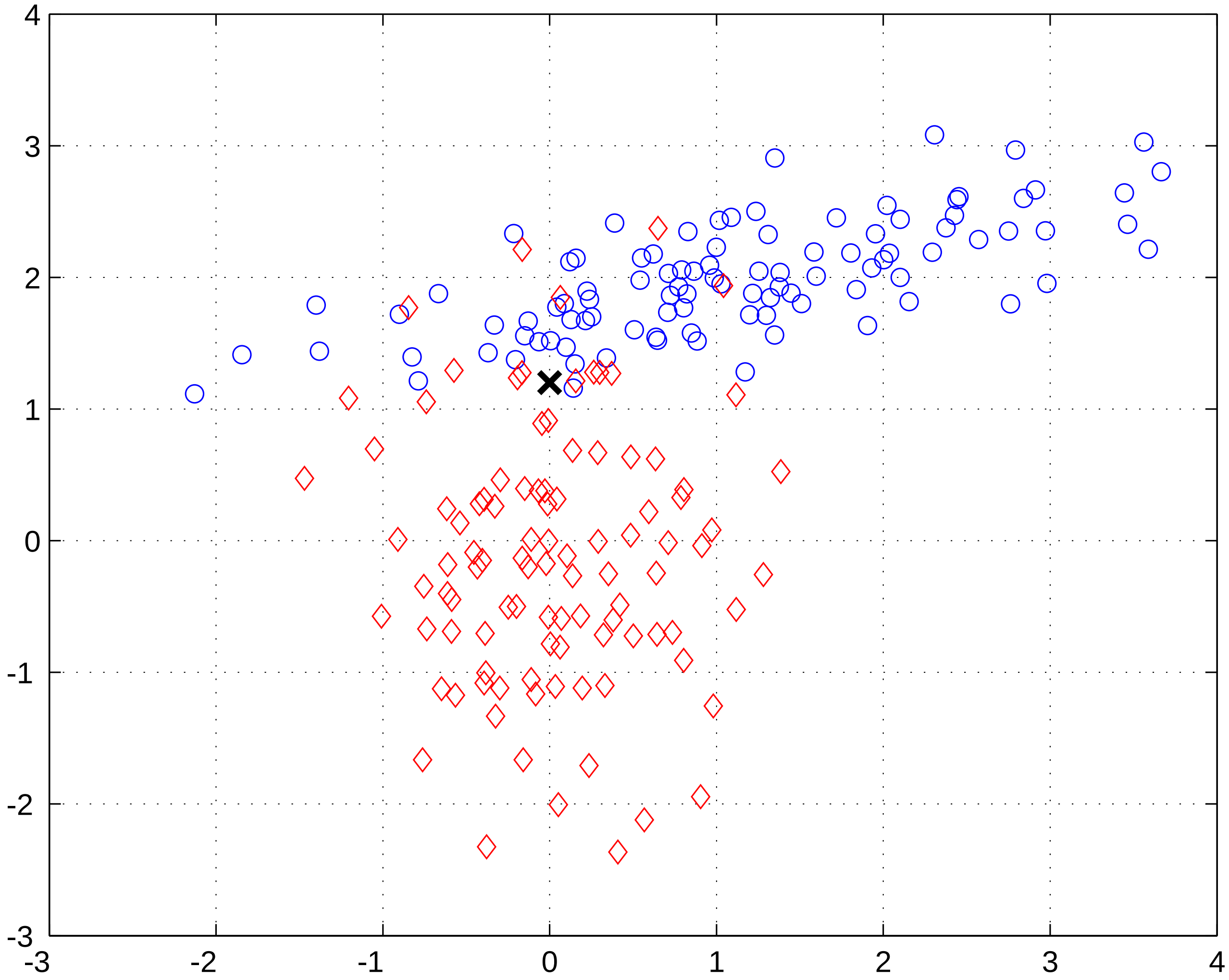
What are we here for?

Tableau 2. — NOMBRE DE MILLIERS DE JOURNÉES DE TRAVAIL PERDUES POUR FAIT DE GRÈVE PAR AN (1954 à 1974, 1968 étant exclu) pour 11 branches d'activité

Années	*Energie*	*Mines*	*Métallurgie*	*Mécanique*	*Bâtiment*	*Chimie*	*Textile*	*Papier*	*Transport*	*Employés de commerce*	*Employés de Banque*
74	102	8	110	879	144	91	130	25	304	39	755
73	909	71	126	508	211	78	105	74	330	0	1
72	151	158	108	938	205	180	101	49	436	20	110
71	148	1	187	1081	225	159	130	75	1235	58	39
70	42	7	63	711	126	157	41	20	247	36	27
69	344	5	59	584	64	89	62	47	496	11	18
67	381	209	693	871	200	185	339	34	655	8	30
66	376	57	114	71	169	93	33	20	547	10	7
65	96	25	31	415	21	50	20	10	209	4	1
64	396	11	165	467	372	90	33	32	666	12	7
63	3925	190	97	548	143	64	62	24	691	7	19
62	549	9	75	290	50	71	28	10	381	0	2
61	295	37	35	576	46	84	36	18	1332	2	24
60	108	111	32	410	39	30	79	10	199	11	0
59	54	60	48	440	151	84	767	8	57	1	21
58	291	44	16	208	88	45	34	12	302	4	0
57	493	56	186	1214	339	103	212	22	705	16	418
56	58	69	82	444	210	58	44	17	118	10	1
55	62	135	173	1409	458	261	61	27	193	4	9
54	119	30	53	528	198	67	66	35	187	4	14

[Source: D. Dacunha & M. Duflo, Probabilités et statistiques, Ed. Masson]

What are we here for?



Why probabilities and statistics will help?

What's the difference between probability and statistics?

- Probability is a field theoretical mathematics

  $\longrightarrow$ rely on axioms and is autonomous from physical reality
- Statistics is the art of collecting, analyzing and interpreting (real) data
  - exploratory statistics $\longrightarrow$ describe, analyze and interpret
  - inferential statistics $\longrightarrow$ generalize, decide and interpret

Probability and statistics will help because most phenomena are hardly considered as deterministic

- because there is an inherent random part (e.g., user behavior)
- because of physical phenomenon too complex to be accurately modeled

What are we gonna talk about?

C01 A gentle introduction to probability (warming up)

C02 Numerical summaries, PCA and the likes

C03 Cluster analysis

C04 Hypothesis testing and variance analysis

C05 From Machine learning to estimation theory and practice

C06 Mixture models and the EM algorithm

C07 Random processes, Markov chains

C08 Hidden Markov chains, continuous time Markov processes

C09 Bayesian networks

C10 Maximum entropy models and conditional random fields

Material available at http://people.irisa.fr/Guillaume.Gravier/ADM

Usefull pillow books

Any introductory book on probabilities and statistics will do the job but I highly recommend the following (some are in French):

- Gilbert Saporta. Probabilités, analyse de données et statistiques. Ed. Technip. 1990, 2006.
- Kishor S. Trivedi. Probability ad Statistics with reliability, queing and computer science applications. Ed Wiley. 2002.

Wikipedia, whether in French or in English, has also great resources regarding statistics (e.g., http://en.wikipedia.org/wiki/Statistics).

FUNDAMENTALS OF PROBABILITY

Basic vocabulary

Study equivalent *objects* on which we observe *variables*:

- object = coins; variable = side on which it falls
- object = manufactured good; variables = dimensions, weight
- object = classification experiment; variable = error rate

Population: a group of equivalent objects
Individual: an object within the group
Sample: a subset of the entire population (given or chosen)
Variable: characteristics describing an individual

Different types of variables:

- numerical: discrete or continuous
- categorical: nominal or ordinal

Basic vocabulary (cont'd)

- *Sample space* or the *universe of possibilities* (Ω)
  - not defined by the experiment but rather by its usage (e.g., throwing a dice)
- *Event*
  - logical assertion with respect to the experiment
    e.g., the result will be greater than 10
  - can often be seen as a subset of Ω
  - an *elementary event* is a subset containing one element
- Some definitions
  - Two events A and B are *incompatible* if the occurence of one exclude the occurence of the other (i.e., if $A \cap B = \emptyset$).
  - A set of events $\{A_1, \ldots, A_n\}$ is said to be *complete* if $\cup_i A_i = \Omega$ and $A_i \cap A_j = \emptyset \quad \forall i \neq j$.

Axioms and properties

The probability theory associates to each a event a number $\in [0, 1]$, the *probability of the event*, satisfying the following **Kolmogorov's axioms**:

- $P[\Omega] = 1$
- each finite set of incompatible events, E_i, satisfies $P[\cup E_i] = \sum P[E_i]$

Consequences:

$$P[\emptyset] = 0 \qquad P[A \cup B] = P[A] + P[B] - P[A \cap B]$$

$$P[\overline{E}] = 1 - P[E] \qquad P[\cup A_i] \leq \sum P[A_i]$$

$$P[A] \leq P[B] \text{ if } A \subset B \qquad \lim_{A_i \to \emptyset} P[A_i] = 0$$

$P[A] = 0$ (resp. $P[A] = 1$) does not imply that A never (resp. always) occurs!

Theorem of total probabilities

Theorem of total probabilities

If B_i *is a complete system of events then,* $\forall$ *event* A*,*
$P[A] = \sum_i P[A \cap B_i]$

Example

Consider a wireless cell with 5 channels, where each channel is in one of two states: busy (0) or available (1). We are interested in the probability that a conf call is not blocked ($X = 1$), knowing that at least 3 channels are required.

Step 1. Define a sample space $\rightarrow$ 5-tuples of 0s and 1s

Step 2. Assign probabilities $\rightarrow$ assume equal probability for each event

Step 3. Identify the events of interest E $\rightarrow$ the event is that "three or more channels are available", represented by the set of 5-tuples that have at least three 1s (16 / 32)

Step 4. Compute the desired probability $\rightarrow$ E is a union of mutually exclusive elementary events E_i with probability 1/32 each, and hence

$$P[X = 1] = \sum_i P[E_i] = \sum_i \frac{1}{32} = \frac{16}{32}$$

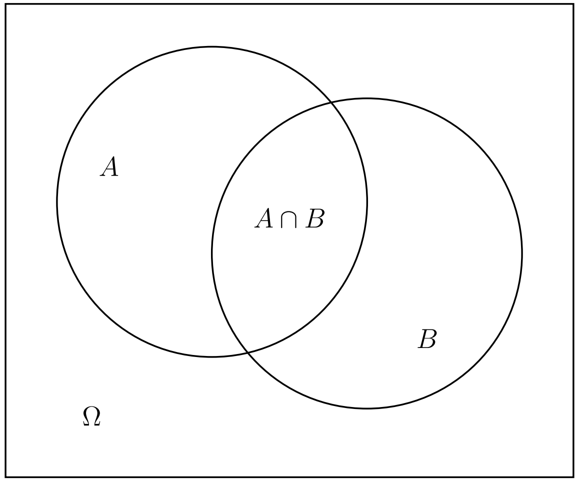
[Trivedi 2002, p. 18]

Conditional probabilities

The **conditional probability** is defined as

$$P[A|B] = \frac{P[A \cap B]}{P[B]}$$

if $P[B] \neq 0$ and undefined otherwise.



This definition leads to the multiplication rule

$$P[A \cap B] = \begin{cases} P[A|B]P[B] & \text{if } P[B] \neq 0 \\ P[B|A]P[A] & \text{if } P[A] \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

Independence

Definition

Two events A and B are independent if and only if $P[A|B] = P[A]$.

The following definition of independence is equivalent:

Two events A and B are independent if $P[A \cap B] = P[A]P[B]$.

Some points worth noting:

- A and B independent $\Rightarrow$ $\overline{A}$ and B independent, A and $\overline{B}$ independent, $\overline{A}$ and $\overline{B}$ independent
- A and B independent and B and C independent does not guarantee that A and C independent

Independence (cont'd)

Three events A, B and C are **mutually independent** if

1. $P[A \cap B \cap C] = P[A]P[B]P[C]$
2. $P[A \cap B] = P[A]P[B]$
3. $P[A \cap C] = P[A]P[C]$
4. $P[B \cap C] = P[B]P[C]$

This can be extended:

n events $A_1, \ldots, A_n$ are mutually independent if and only if for each set of $k \in [2, n]$ distinct indices $i_1, \ldots, i_k$ ($i_j \in [1, n] \forall j$)

$$P[A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}] = P[A_{i_1}]P[A_{i_3}] \ldots P[A_{i_k}]$$

Independence: Tricky examples

Let's throw 2 (fair) dices. The sample space is $\Omega = \{(i,j)\ ;\ 1 \le i,j \le 6\}$

Example 1.

- A first dice is 1, 2 or 3 $\to P[A] = \frac{3}{6}$
- B first dice is 3, 4 or 5 $\to P[B] = \frac{3}{6}$
- C the sum is 9 $\to P[C] = \frac{4}{36}$

We clearly have

$$P[A \cap B \cap C] = \frac{1}{36} = P[A]P[B]P[C]$$

but

$$(\frac{6}{36} =) P[A \cap B] \neq P[A]P[B] (= \frac{9}{36})\ .$$

The inequality also holds for $A \cap C$ and $B \cap C$.

Independence: Tricky examples (cont'd)

Example 2.

- A first dice is 1, 2 or 3 $\rightarrow P[A] = \frac{3}{6}$
- B second dice is 4, 5 or 6 $\rightarrow P[B] = \frac{3}{6}$
- C the sum is 7 $\rightarrow P[C] = \frac{6}{36}$

It is easy to verify that events A, B, C are pairwise independent but

$$(\frac{1}{12} =) P[A \cap B \cap C] \neq P[A]P[B]P[C] (= \frac{1}{24}) \ .$$

Independence: Application to system reliability

Consider a system with several connected components; event A_i = "component i is functioning properly"; define the *reliability* of a component i as $R_i = P[A_i]$, the probability that the component is functioning properly.

Assumption: failure events of components are mutually independent

- Series system: the system fails if any one of its components fails

  $$\rightarrow \quad R_s = P[\cap_i A_i] = \prod_i R_i$$

- Parallel systems: the system fails if all of its components fail

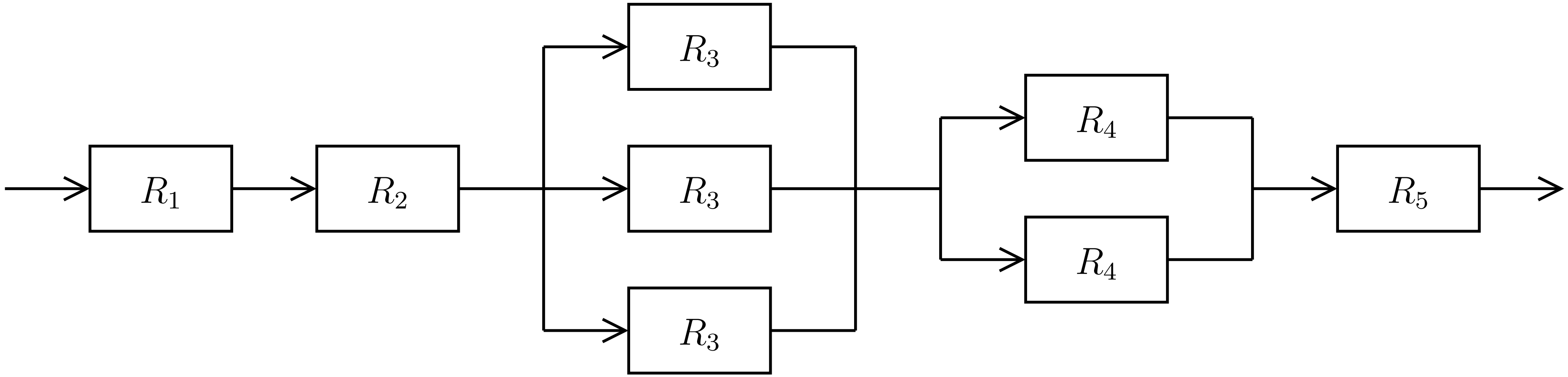
  $$\rightarrow \quad R_s = 1 - P[\cap_i \overline{A}_i] = 1 - \prod_i (1 - R_i)$$

  or, using unreliability instead,

  $$\rightarrow \quad F_s = \prod_i F_i$$

Application to system reliability (cont'd)

Let's consider the following system combining series and parallel components



The system's reliability is given by

$$R_s = R_1 R_2 (1 - (1 - R_3)^3)(1 - (1 - R_4)^2) R_5$$

Baye's rule

$$P[B|A] = \frac{P[A|B]P[B]}{P[A]}$$

Thomas Bayes

(c. 1702–1761)

Other form of Bayes' rule assuming a mutually exclusive and collectively exhaustive set of events $B_1, \ldots, B_n$:

$$\begin{aligned} P[B_j|A] &= \frac{P[B_j \cap A]}{P[A]} \\ &= \frac{P[A|B_j]P[B_j]}{\sum_i P[A|B_i]} \end{aligned}$$

Baye’s rule: Application

Three machines M1, M2 and M3 produce bolts. M1 produces on average 0.3 % of faulty bolts, M2 0.8 % and M3 1 %. We mix 1000 bolts in a bag, 500 from M1, 350 from M2 and 150 from M3. We randomly pick one bolt from the bag: It is faulty! What is the probability that it was produced by M1?

- P[M1] = 0.5 P[M2] = 0.35 P[M3] = 0.15
- P[D|M1] = 0.003 P[D|M2] = 0.008 P[D|M3] = 0.15

$$P[M1|D] = \frac{P[D|M1]P[M1]}{P[D|M1]P[M1] + P[D|M2]P[M2] + P[D|M3]P[M3]}$$

Important conditional probabilities equations

- Joint probability

$$P[A \cap B] = P[A, B] = P[A|B]P[B] = P[B|A]P[A]$$

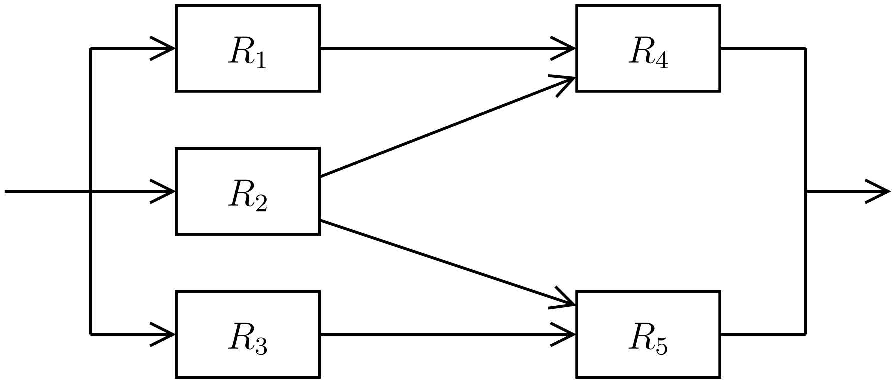
- Marginal probability

$$P[A] = \sum_i P[A|B_i]$$

if the B_i's are mutually exclusive and collectively exhaustive events.

$$P[A|B] = \frac{P[A|B]P[B]}{P[B]} = \frac{P[A, B]}{P[B]}$$

Bayes' rule for non parallel/series systems



$$A = (A_1 \cap A_4) \cup (A_2 \cap A_4) \cup (A_2 \cap A_5) \cup (A3 \cap A_5)$$

From what we have learnt on conditional probabilities, we have

$$P[A] = P[A|A_2]P[A_2] + P[A|\overline{A}_2]P[\overline{A}_2]$$

Two distinct cases

$A_2 \rightarrow A_1$ and A_3 are irrelevant and $P[A|A_2]$ simplifies to a parallel system

$\overline{A}_2 \rightarrow$ two series system in parallel

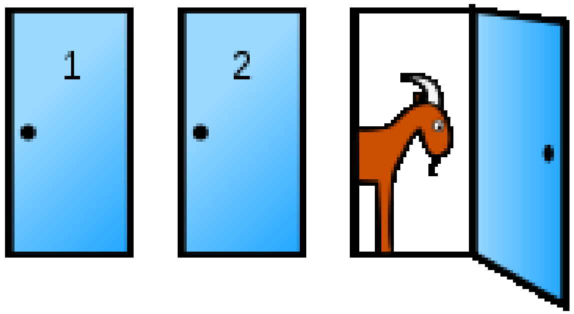
$$R_s = [1 - (1 - R_4)(1 - R_5)]R_2 + [1 - (1 - R_1 R_4)(1 - R_3 R_5)](1 - R_2)$$

[Trivedi 2002, p. 42]

On the limit of combinatorial approaches

Combinatorial approaches can sometimes be tricky:

- Monty Hall paradox



Check http://en.wikipedia.org/wiki/Monty_Hall_problem for a (lengthy) discussion

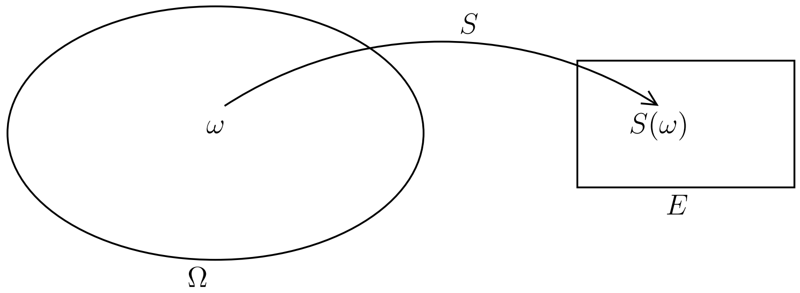
- Two children paradox

  Mr. Smith has two children. At least one of them is a boy. What is the probability that both children are boys?

- Bertrand’s box paradox (same as Monty Hall)
- etc.

Random variables

Definition. *A random variable X on a sample space Ω is a funcion $X : \Omega \to \mathbb{R}$ that assigns a real number $X(\omega)$ to each sample point $\omega \in \Omega$.*



In other words, a random variable is a function mapping the sample space to some values which can be either discrete or continuous.

Examples:

- randomly selecting between 0 and 1 three times and observing the number of 1s in the result

$\omega \in \Omega$	111	110	101	100	011	010	001	000
$P(\omega)$	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125
$X(\omega)$	3	2	2	1	2	1	1	0

- lifetime of a component

  $\longrightarrow$ hard to imagine the sample space!

Random variables: why?

The **inverse image** of a random variable is defined as $A_x = \{\omega \in \Omega$ such that $X(\omega) = x\}$ with the following properties

- $A_x \cap A_y = \emptyset$ if $x \neq y$
- $\cup_{x \in \mathbb{R}} A_x = \Omega$

It is often most convenient to work in the event space defined by the collection of events A_x if our interest is solely in the resulting experimental value of random variable X.

In the previous example

$$A_0 = \{(0,0,0)\}; A_1 = \{(1,0,0),(0,1,0),(0,0,1)\}$$
$$A_2 = \{(1,1,0),(1,0,1),(0,1,1)\}; A_3 = \{(1,1,1)\}$$

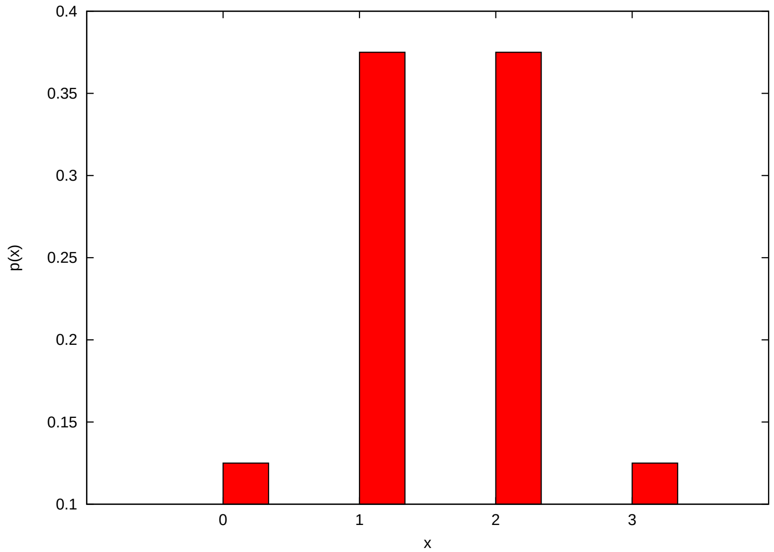
which reduces the sample space of dimension 8 to an event space of dimension 4.

$\rightarrow$ with n trials: 2^n sample points to $n+1$ events!

Probability mass and distribution functions

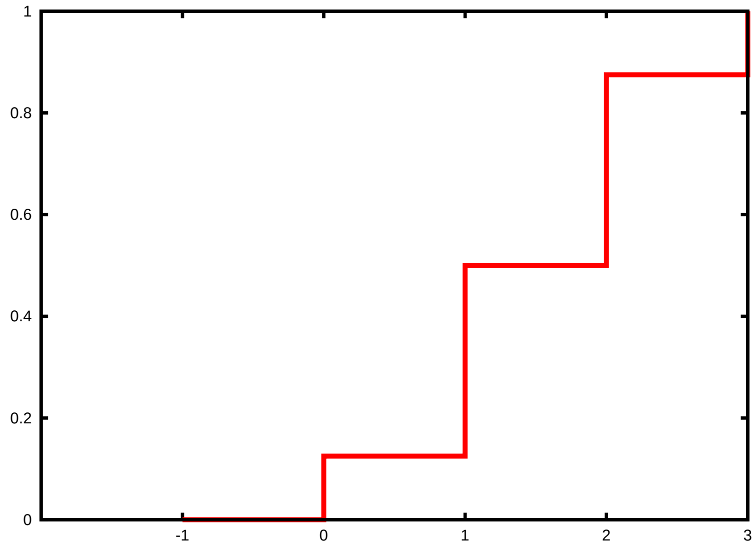
The **probability mass function** (pmf) is defined a function defined as

$$p_X(x) = P[X = x] = P[A_x] = P[\{\omega \in \Omega \text{ such that } X(\omega) = x\}]$$
$$= \sum_{X(\omega)=x} P[\omega]$$



probability mass function (pmf)

$$P[X = x]$$



cumulative distribution function (cdf)

$$P[-\infty < X < x]$$

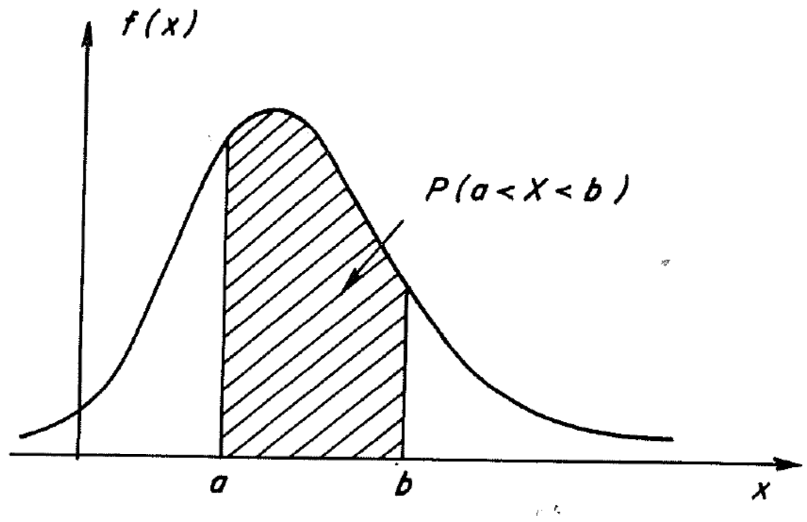
Classical discrete distributions

- **Uniform**: a priori distribution when nothing is known
- **Bernoulli**: $X = 1$ with probability p and 0 with probability $1 - p$
- **Binomial law** : probability that an event that has an occurence probability of p appears k times over n independent trials
- **Poisson**: probability of observing n events occuring at a rate c during a lapse of time T
- **Geometric**: probability of the number of Bernoulli trials before the first "success"
- **Negative binomial**: probability of the number of Bernoulli trials before the r^{th} "success"
- **Hypergeometric**: probability of choosing k defective components among m samples chosen without replacement from a total of n of which d are defective
- etc.

Continuous random variables

- for a continuous variable, $P[X = x] = 0$!!!!!!!
- probability defined over an interval $P(x < X < x + dx) = f(x)dx$
- $f(x)$ is the **probability density function** (pdf)
- $F(x) = \int_{-\infty}^{x} f(x)dx$ is the **cumulative distribution function** (cdf)

$$\begin{aligned} P(a < X < b) &= \int_a^b f(x)dx \\ &= F(b) - F(a) \end{aligned}$$



[Source: G. Sapporta (1990), p. 22]

Classical continuous distributions

- **Uniform**: a priori distribution when nothing is known
- **Exponential**: lifetime of a component or a service
- **Gamma**: somehow equivalent to Poisson in the continous case
- **Normal** or **Gaussian**: almost everything
- **Weibull**: reliability
- etc.

Function of a random variable

Let X be a continuous random variable of density f_X and ϕ a differentiable monotone funcion. $Y = \phi(X)$ is a continuous random variable with density f_Y given by

$$f_Y(y) = \frac{f_X(\phi^{-1}(y))}{|\phi'(\phi^{-1}(y)|} \quad .$$

Example. For $y = \exp(x)$, we have

$$f_Y(y) = \frac{f_X(x)}{\exp(x)} = \frac{f_X(\ln(y))}{y}$$

$\Rightarrow$ not as simple for non monotonous functions

Two random variables

The joint distribution of two random variables X and Y is defined as

$$F_{X,Y}(x, y) = P(X \leq x \cap Y \leq y) \ .$$

If both variables are coutinuous, there often exist a function such that

$$F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v) du dv$$

and, if partial derivative exists,

$$f_{X,Y}(x, y) = \frac{\partial^2 F_{X,Y}(x, y)}{\partial x \partial y} \ .$$

Independence $\Rightarrow F_{X,Y}(x, y) = F_X(x) F_Y(y)$.

Expectation of a random variable

- The **expectation** of a random variable X is defined as
  - discrete case

$$E[X] = \sum_{x \in X(\Omega)} x P[X = x]$$

  - continous case

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx \text{ (if the integral converges)}$$

- The expectation measures the "gravity center" of the distrubtion
- Properties

$$E[a] = a \qquad E[X + a] = E[X] + a$$

$$E[aX] = aE[X] \qquad E[X_1 + X_2] = E[X_1] + E[X_2]$$

Expectation of a random variable (cont'd)

- Expectation of a function of X

$$E[\phi(X)] = \int \phi(x) f(x) dx$$

- Expectation of the product XY

$$E[XY] = \int\int xy f(x,y) dx dy$$

  - X and Y independent $\Rightarrow E[XY] = E[X]E[Y]$
  - but the inverse is not true

Variance and standard deviation

- The **variance** of a random variable X is defined as

$$V[X](=\sigma^2) = E[(X - E[X])^2] = E[X^2] - E[X]^2$$

- The square roots of the variance, σ, is known as the **standard deviation**
- Properties
  - $E[(X-a)^2] = V[X] + (E[X] - a)^2$
  - $V[X-a] = V[X]$
  - $V[aX] = a^2 V[X]$
  - Inequality of Bienaymé-Tchebyshev: $P[|X - E[X]| > k\sigma] \leq \frac{1}{k^2}$

Variance and covariance

- Variance of a sum of variables

$$V[X+Y] = V[X] + V[Y] + 2(\underbrace{E[XY] - E[X]E[Y]}_{COV(X,Y)})$$

  - X and Y independent $\Rightarrow V[X+Y] = V[X] + V[Y]$
  - but the inverse is not true

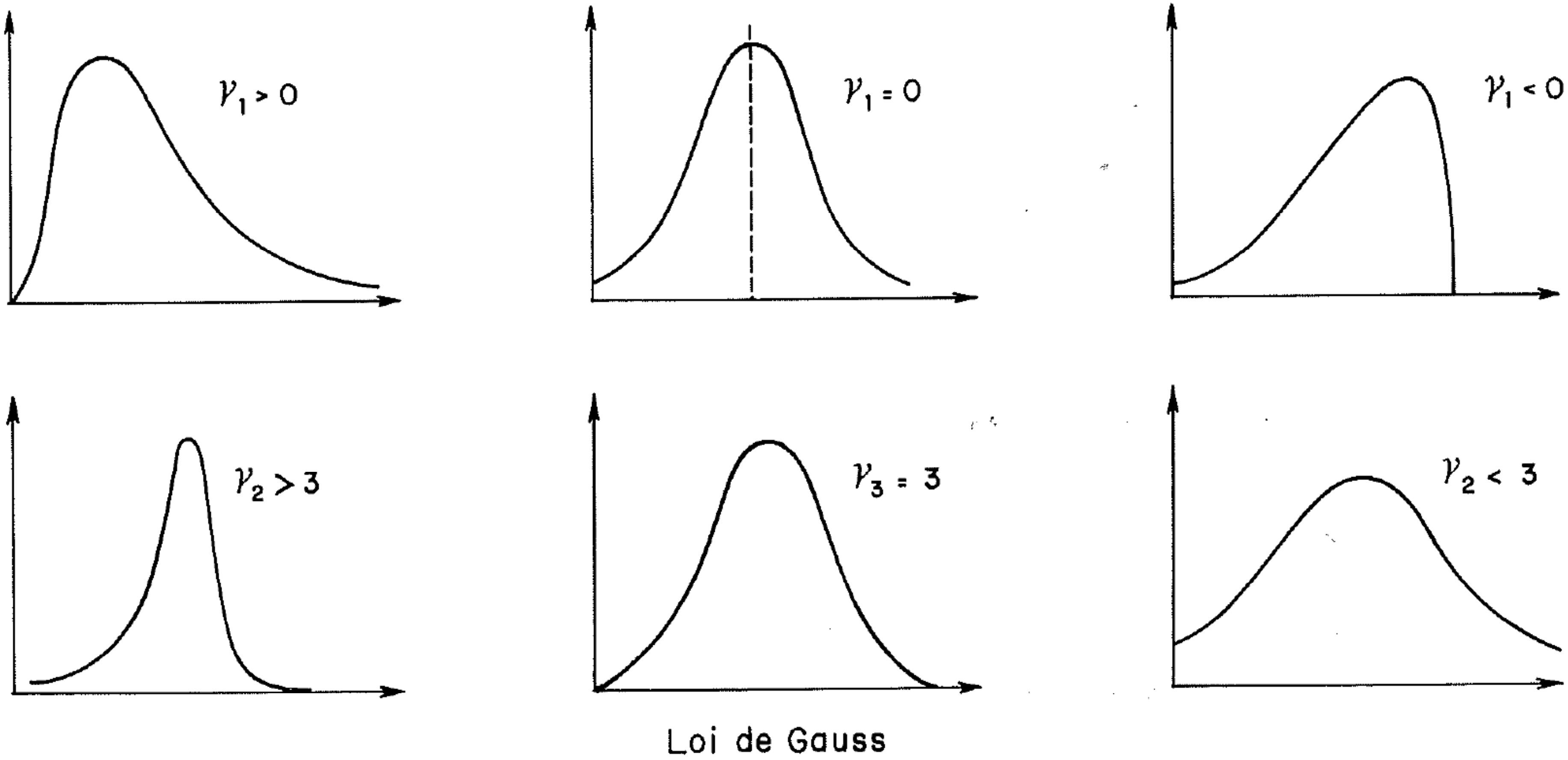
Moments

- the moment of order k is defined as

$$\mu_k = E[(X - E[X])^k]$$

- moments are eventually normalized by σ^k
- typical normalized moments
  - skewness $\gamma_1 = \mu_3/\sigma^3$
    - → symmetric distributions $\Rightarrow \gamma_1 = 0$
    - → heavier tail on the left (resp. right) $\Rightarrow \gamma_1 < 0$ (resp. $\gamma_1 > 0$)
  - kurtosis $\gamma_2 = \mu_4/\sigma^4$
    - → tall and skinny *vs.*. short and squat
    - → importance of the tails in the distribution

Moments (cont'd)



[Source: Saporta 1990, p. 29]

Discrete distributions: Bernoulli and binomial

Jacob Bernoulli

1654–1705

Bernoulli

- Law of a binary random variable X, taking value 1 with probability p and 0 with probability $1 - p$
- Typically used as an indicator function for an event occuring with a probability p
  - $P[X = k] = p^k(1 - p)^{1-k}$
  - $E[X] = p$ and $V[X] = p(1 - p)$

Binomial

- Probability that an event that has an occurence probability of p appears k times over n independent (Bernoulli) trials
- $X \rightsquigarrow \mathcal{B}(n, p)$ if X is the sum of n independent Bernoulli variables

$$P[X = k] = C_n^k p^k (1 - p)^{n-k}$$

$$E[X] = np \text{ and } V[X] = np(1 - p)$$

Classical discrete distributions: Poisson

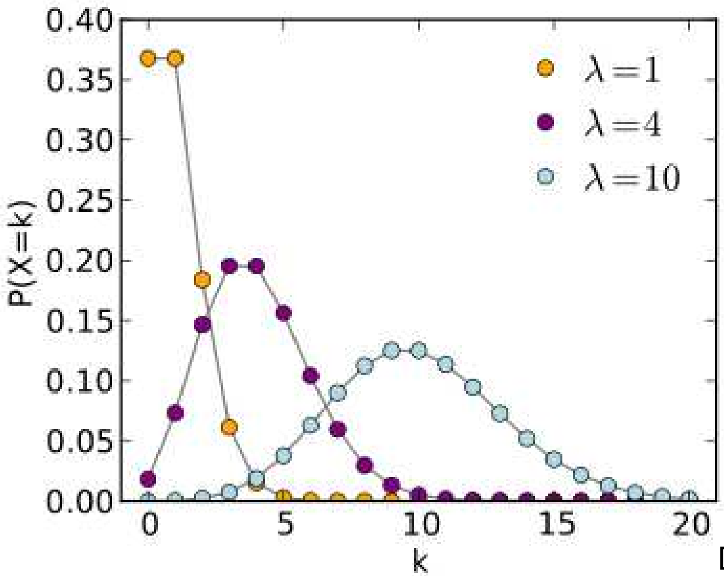
- A Poisson $\mathcal{P}(\alpha)$ is defined as

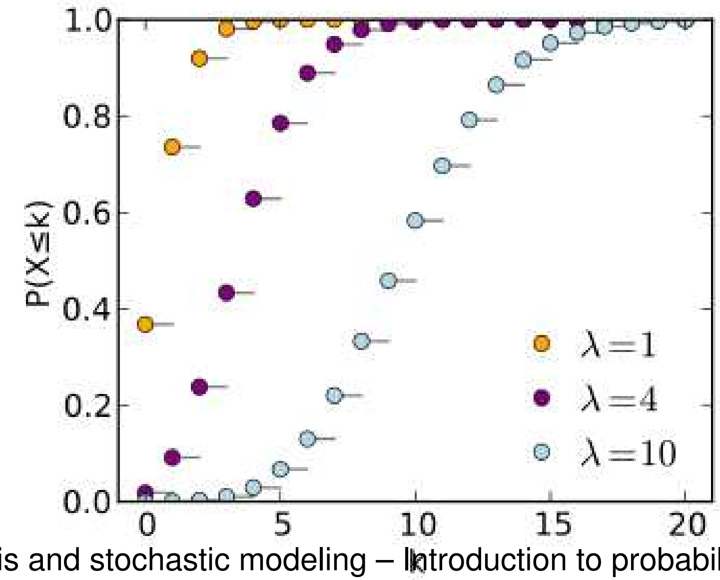
$$P[X = k] = \frac{\alpha^k e^{-\alpha}}{k!}$$

- Probability of k events occuring at a rate λ over an interval of duration t ($\alpha = \lambda t$)
- Approximation of the binomial for small p and large n $(n \geq 20, p \leq 0.05)$
- $E[X] = V[X] = \alpha$

Siméon D. Poisson

1781–1840



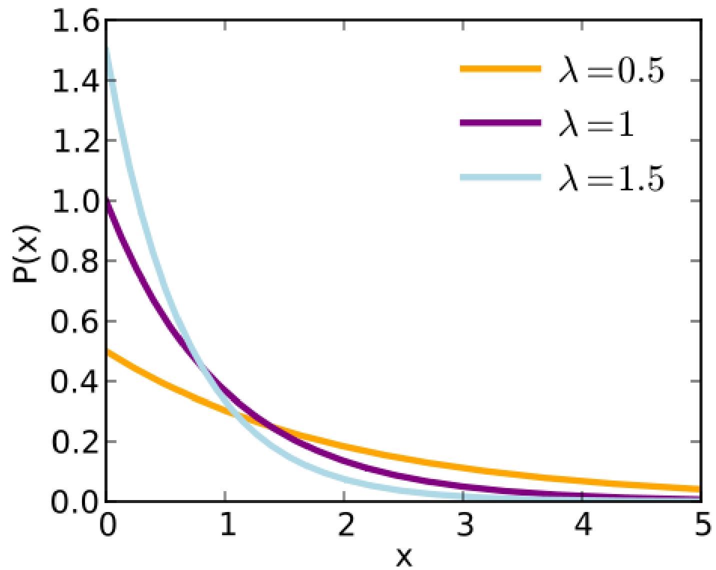


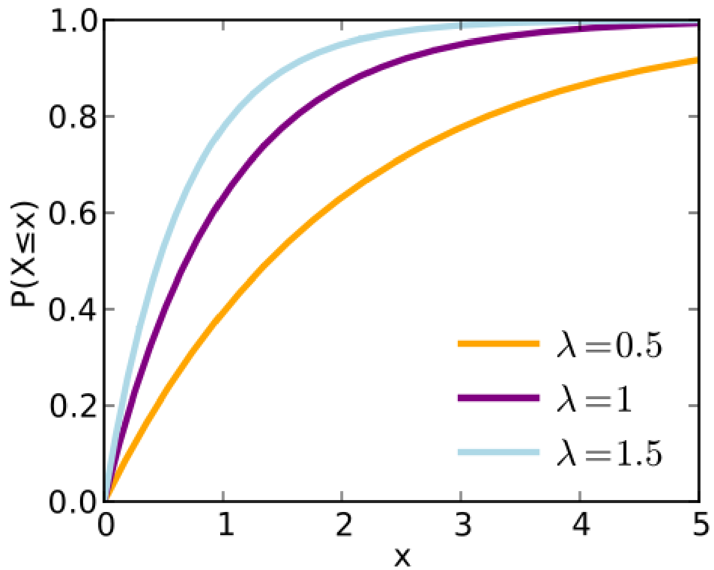
Continuous laws: the exponential law

The exponential law models lifetimes such as

- time between two successive job arrivals
- service time at a server in queing network
- time to failure of a component

$$F(x) = 1 - e^{-\lambda x} \qquad f(x) = \lambda e^{-\lambda x} \qquad E[X] = \lambda^{-1} \qquad V[X] = \lambda^{-2}$$





[Illustration: Skbkekas (from wikipedia.org)]

Memoryless property of the exponential law

- In reliability, $R(t)$ is the probability that an individual lives past t, i.e., $R(t) = P[X > t]$
- From Baye's rule, the probability that an individual dies between t_1 and t_2 knowing that he is alive at t_1 is given by

$$P[t_1 \leq X < t_2 | X > t_1] = \frac{R(t_1) - R(t_2)}{R(t_1)}$$

- Using the exponential law $P[X > t] = e^{-ct}$

$$P[t_1 \leq X < t_2 | X > t_1] = 1 - e^{-c(t_2 - t_1)} = P[X < t_2 - t_1]$$

- No aging of the individual!

Failure rate and mean time to failure

Some definitions in the domain of reliability

- reliability of a component: $R(t) = P[X > t] = 1 - F(t)$
- $f(t)\Delta t$ is the (unconditional) probability of failure in $[t, t + \Delta t[$
- $h(t) = f(t)/R(t)$ is the conditional probability of failure in $[t, t + \Delta t[$, knowing that the component has survived until t
- mean time to failure (MTTF): $E[X] = \int_0^\infty R(t)dt$

For an exponentially distributed lifetime, $R(t) = e^{-ct}$

- $h(t) = c$
- MTTF = c^{-1}
- for a series system
  - $R(t) = \exp\left(-\left(\sum_i \lambda_i\right) t\right)$
  - MTTF = $\left(\sum_i \lambda_i\right)^{-1}$ ($\leq min_i E[X_i]$)

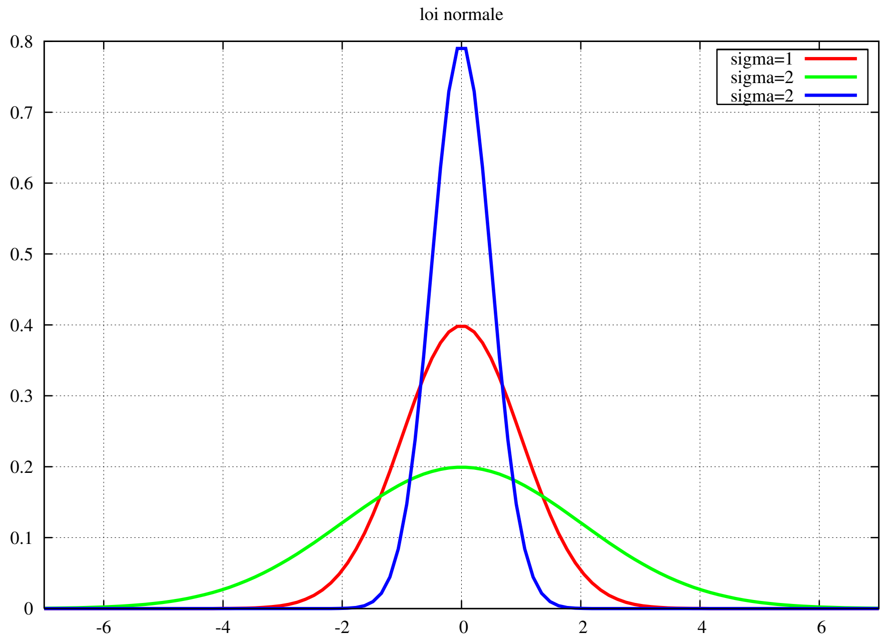
Gaussian law

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

$E[X] = \mu$ and $V[X] = \sigma^2$

Carl Friedrich Gauss

1777–1855



$$P[m - 1.64\sigma < X < m + 1.64\sigma] = 0.90$$
$$P[m - 1.96\sigma < X < m + 1.96\sigma] = 0.95$$
$$P[m - 3.09\sigma < X < m + 3.09\sigma] = 0.998$$

Gaussian law: advantages

Entirely defined by the two parameters μ and σ.

Central-limit theorem *If X_n is a set of random variables from the same law with mean μ and standard deviation σ, then*

$$\frac{1}{\sqrt{n}}\left(\frac{X_1+X_2+\ldots+X_n-n\mu}{\sigma}\right)\longrightarrow\mathcal{N}(0,1)$$

Theorem of Lindeberg *Let $X_1,\ldots,X_n$ be independent random variables of means μ_i and standard deviations σ_i. Under certain coniditions*

$$\frac{\sum_i X_i-m_i}{\sqrt{\sum_i \sigma_i^2}}\longrightarrow\mathcal{N}(0,1)$$

Interpretation: if a variable results from a large number of small additive causes, its law is a Gausian law.

Properties of the Gaussian law

Changing variable

If $X \longrightarrow \mathcal{N}(\mu, \sigma^2)$, then

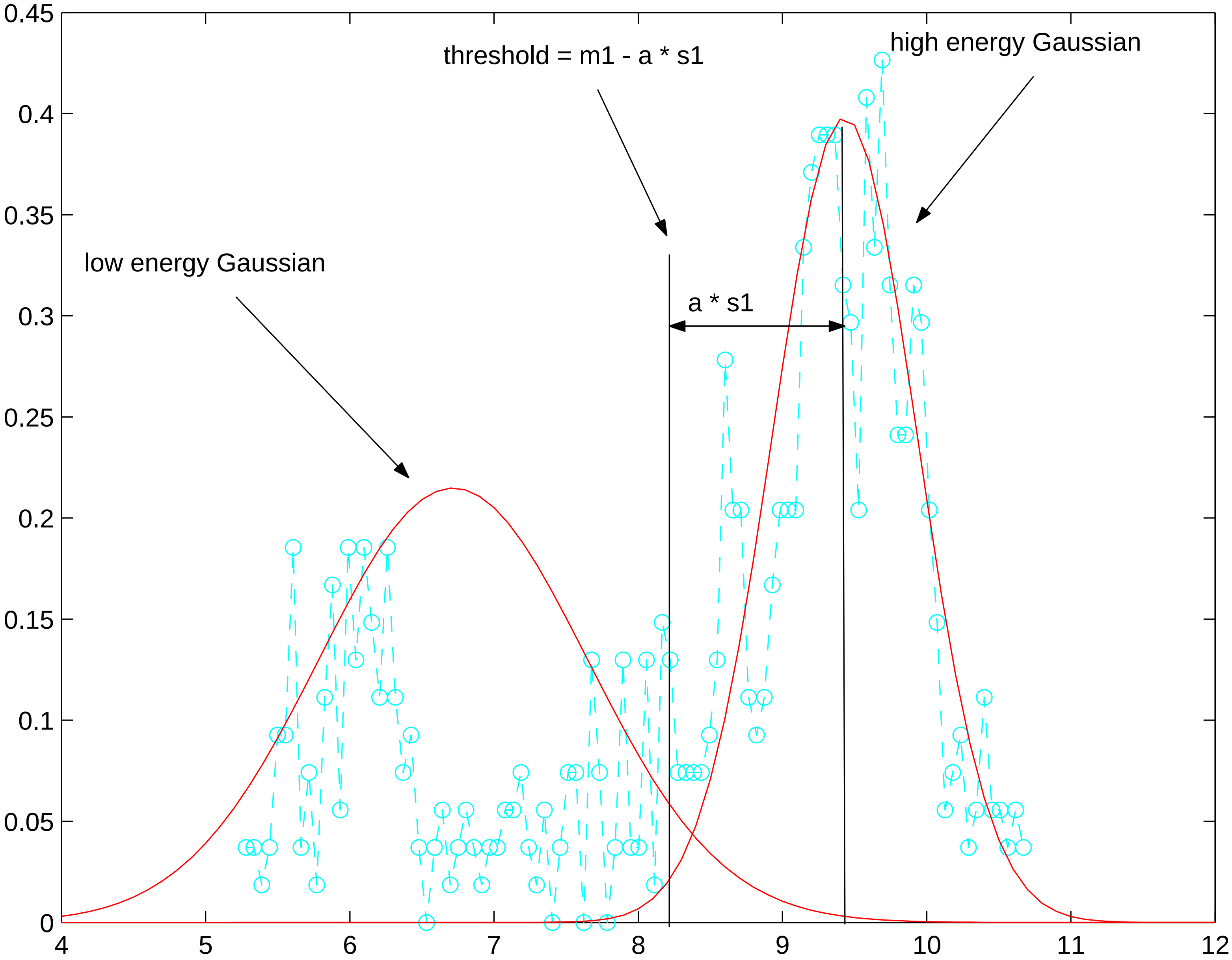
$$U = (X - \mu)/\sigma \rightsquigarrow \mathcal{N}(0, 1)$$

Additivity

If X_1 and X_2 are two independent variables with respective laws $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$, then the law of $X_1 + X_2$ is Gaussian with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$.

Warning: the property of additivity only applies if the variables are independent!

A practical use of the Gaussian law

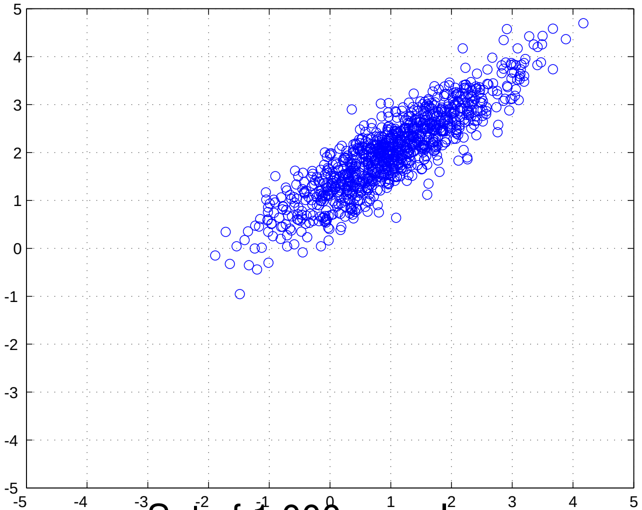


Multivariate Gaussian density

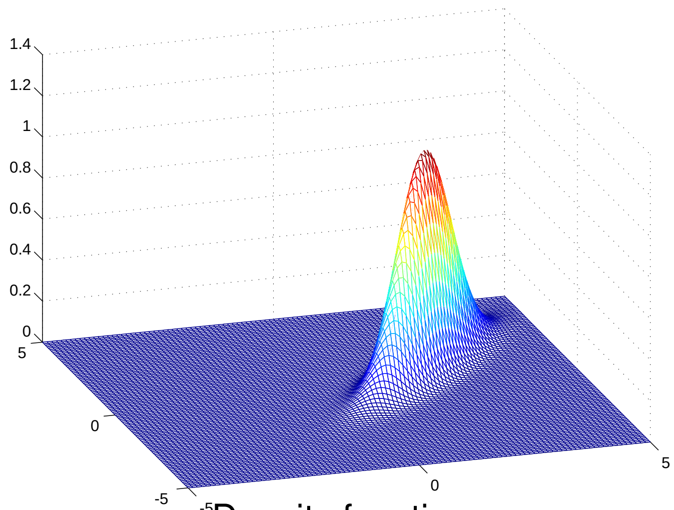
Definition

X is a Gaussian vector of dimension p if any linear combination of its components $a'X$ is a Gaussian in dimension 1.

$$f(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\}$$



Set of 1,000 samples



Density function

Example of a multivariate Gaussian with $m = [21]$ and $\theta = \pi/6$.

Properties of the covariance matrix

The covariance matrix is symmetric, definite positive,

$$\Sigma = V\ D\ V'$$

where

V are the eigen vectors defining the principal axes (orientation of the density) and

D are the eigen values defining the dispertion along the axes.

Theorem

The components of a Gaussian vector are independent if and only if Σ is a diagonal matrix, *i.e.* if the components are not correlated.

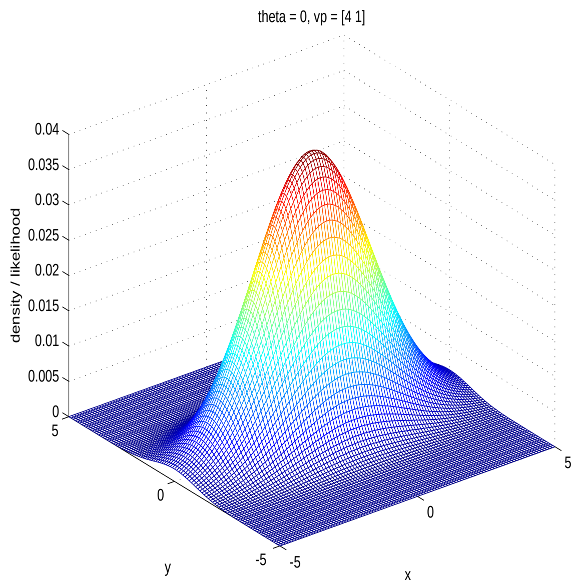
Illustration of 2D Gaussians

From the correlation point of view

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_2\sigma_1 & \sigma_2^2 \end{pmatrix}$$

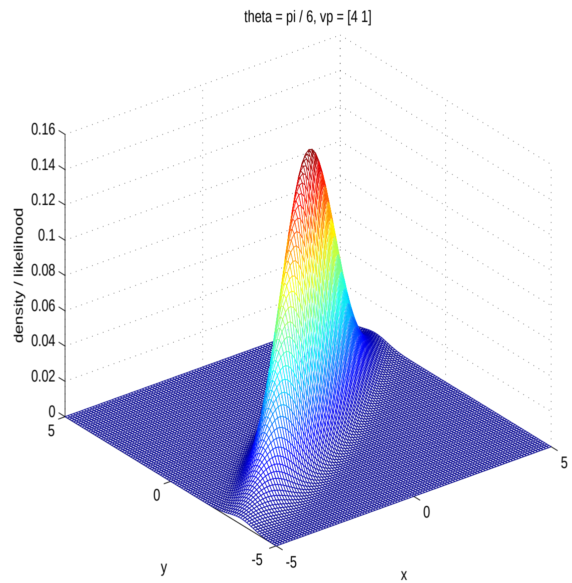
From the geometric point of view

$$V = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}$$



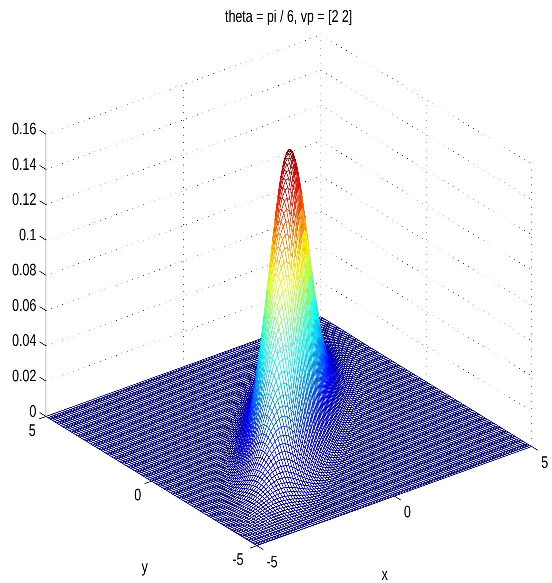
$\theta = 0$

$D = \text{diag}(4 \quad 1)$



$\theta = \pi/6$

$D = \text{diag}(4 \quad 1)$



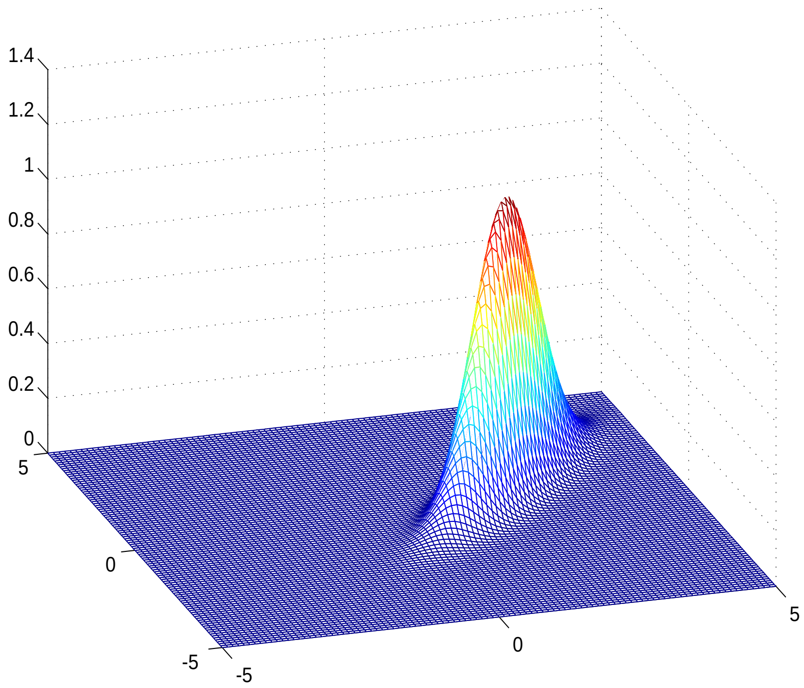
$\theta = \pi/6$

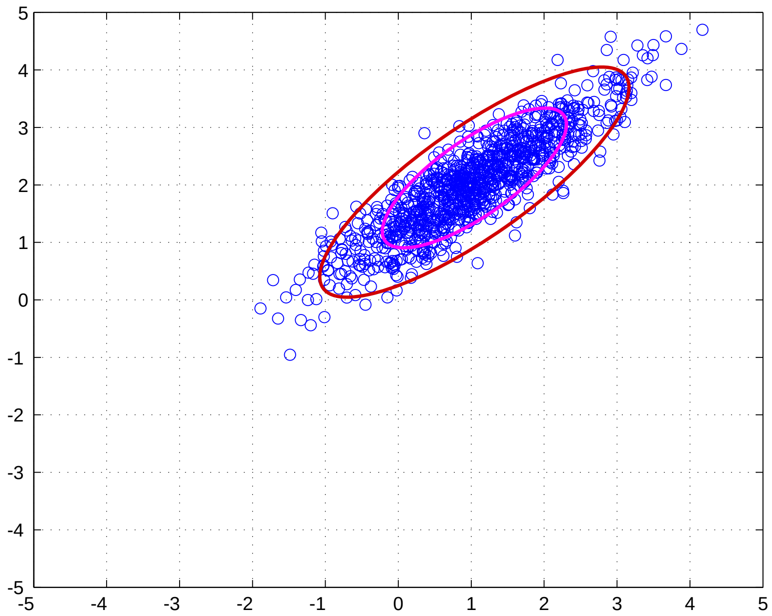
$D = \text{diag}(2 \quad 2)$

Isodensity ellipsoids

Isodensity curves are (hyper)ellipsoids whose equation is given by

$$(x-\mu)'\Sigma^{-1}(x-\mu) = c$$





90% of the samples lie whithin the pink ellipse 99% of the samples lie whithin the red ellipse