Discrete and Continuous random variables

EE-209 - Eléments de Statistiques pour les Data Sciences

Outline

Notations

- A capital letter like X denotes a random variable.
- A lower case letter like x denotes a possible observed value. It is a fixed value.
- We can consider some *events*. For example if X and Y are the values obtained by casting two independent dice, we could have

$$\{X = 1\}, \qquad \{X \geq 4\}, \quad \{Y \text{ is even}\}, \quad \{X = Y\}, \quad \{X + Y = 8\}.$$

  Events are formally sets in the theory of probability.
- $\mathbb{P}$ is the general *probability operator*. The probabilities of the events above are simply

$$\mathbb{P}(X = 1), \qquad \mathbb{P}(X \geq 4), \quad \mathbb{P}(Y \text{ is even}), \quad \mathbb{P}(X = Y), \quad \mathbb{P}(X + Y = 8).$$

  The specification of which pmf should be used to calculate the probability is indicated by the random variable itself.

Axioms of probability theory

1. For any events $\mathcal{E}$, $\quad 0 \leq \mathbb{P}(\mathcal{E}) \leq 1$.
2. For any **disjoint** (i.e., incompatible) events $\mathcal{E}_1$ and $\mathcal{E}_2$,

$$\mathbb{P}(\mathcal{E}_1 \cup \mathcal{E}_2) = \mathbb{P}(\mathcal{E}_1) + \mathbb{P}(\mathcal{E}_2).$$

   Example: $\mathbb{P}(|X| \geq 1) = \mathbb{P}(X \geq 1) + \mathbb{P}(X \leq -1)$.

   The previous property generalizes to countable numbers of disjoint events.
3. If $\mathcal{E}$ and $\mathcal{E}^c$ are **complementary events** (i.e., one is the negation of the other), then

$$\mathbb{P}(\mathcal{E} \cup \mathcal{E}^c) = \mathbb{P}(\mathcal{E}) + \mathbb{P}(\mathcal{E}^c) = 1$$

   Example: $\mathbb{P}(X \leq 0) = 1 - \mathbb{P}(X > 0)$.

- A consequence of the second point is that if $\mathcal{E} \subset \mathcal{E}'$ then $\mathbb{P}(\mathcal{E}) \leq \mathbb{P}(\mathcal{E}')$.

  Example: $\mathbb{P}(X \leq 1) \leq 1 - \mathbb{P}(X \leq 2)$.

Discrete random variables (variable aléatoire discrète)

The probability distribution of a discrete random variable X taking values in a countable set $\mathcal{X}$ is given by it *probability mass function* (pmf, *fonction de masse*.)

$$P_X(x) := \mathbb{P}(X = x)$$

Bernoulli random variable. X takes values in $\{0, 1\}$

$$P_X(1) = \mathbb{P}(X = 1) = \theta \quad \text{and} \quad P_X(0) = \mathbb{P}(X = 0) = 1 - \theta.$$

Six faced die. X takes values in $\{1, 2, 3, 4, 5, 6\}$

$$P_X(1) = P_X(2) = P_X(3) = P_X(4) = P_X(5) = P_X(6) = \frac{1}{6}.$$

General discrete distribution on $\{1, \ldots, K\}$.

$$P_X(k) = \mathbb{P}(X = k) = \pi_k \geq 0, \qquad \text{with} \quad \pi_1 + \pi_2 + \ldots + \pi_K = 1.$$

Some classical discrete random variables

Binomial random variable X takes values in $\mathcal{X} = \{1, \dots, n\}$

$$P_X(k) = \mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad \text{for some} \quad p \in [0, 1],$$

with $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ the number of combination of k elements among n.

Geometric random variable X takes values in $\mathcal{X} = \mathbb{N}$

$$P_X(k) = \mathbb{P}(X = k) = p\,(1-p)^k$$

Poisson distribution X takes values in $\mathcal{X} = \mathbb{N}$

$$P_X(k) = \mathbb{P}(X = k) = \frac{\theta^k}{k!} e^{-\theta} \qquad \text{for some} \quad \theta > 0.$$

Joint, marginal, and conditional pmfs

For a pair of discrete variables (X, Y) taking values in $\mathcal{X} \times \mathcal{Y}$, we can define

The joint pmf $P_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y) = \mathbb{P}\big((X = x) \,\&\, (Y = y)\big)$.

The marginal pmfs $P_X(x) = \mathbb{P}(X = x)$ and $P_Y(y) = \mathbb{P}(Y = y)$.

Law of total probability (loi des probabilités totales)

$$P_X(x) = \sum_{y \in \mathcal{Y}} P_{X,Y}(x, y)$$

Conditional distribution

The conditional pmf of X given Y is defined as the pmf

$$P_{X|Y}(x|y) = \frac{P_{X,Y}(x, y)}{P_Y(y)} \quad \text{if} \quad P_Y(y) > 0.$$

Note that if $P_Y(y) = 0$, we can define $P_{X|Y}(\cdot|y)$ to be any pmf: it does not matter...

Expectation and conditional expectation

Expectation (Espérance)

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x \, P_X(x) = \sum_{x \in \mathcal{X}} x \, \mathbb{P}(X = x)$$

Expectation of a function f of X

$$\mathbb{E}[f(X)] = \sum_{x \in \mathcal{X}} f(x) \, P_X(x) = \sum_{x \in \mathcal{X}} f(x) \, \mathbb{P}(X = x)$$

Conditional expectation (Espérance conditionelle)

$$\mathbb{E}[X|Y = y] = \sum_{x \in \mathcal{X}} x \, P_{X|Y}(x|y) = \sum_{x \in \mathcal{X}} x \, \mathbb{P}(X = x|Y = y).$$

Variance and covariance

Variance

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \sum_{x \in \mathcal{X}} (x - \mathbb{E}[X])^2 \, P_X(x).$$

Standard deviation (écart type)

$$\mathrm{std}(X) = \sqrt{\mathrm{Var}(X)}.$$

Covariance

$$\mathrm{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} (x - \mathbb{E}[X])(y - \mathbb{E}[Y]) \, P_{X,Y}(x, y).$$

Indicator functions and indicator variables

Indicator variables are variables that are associated with an event.

Indicator function

For example for the event $\{x \in A\}$, where A is a fixed set, then, we can first define the indicator function

$$x \mapsto 1_{\{x \in A\}} = \begin{cases} 1, & \text{if } x \in A, \\ 0, & \text{else.} \end{cases}$$

When you put a value of x in this function it returns 1 when the statement "$x \in A$" is true and 0 if it is false.

Indicator variable

Then we can apply the indicator function to a random variable X, for example

$$1_{\{X \in A\}}$$

is the random variable equal to 1 when $X \in A$ and 0 else.

Indicator variable: example 1

Assume that X is a random variable over $\{1, \ldots, 6\}$.

$$\mathbb{P}(X \text{ is even}) = P_X(2) + P_X(4) + P_X(6) = \sum_{x=1}^{6} 1_{\{x \text{ is even}\}} P_X(x) = \mathbb{E}[1_{\{X \text{ is even}\}}]$$

More generally

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

In particular, it shows that computing a probability is a particular case of an expectation computation.

Indicator variable: example 2

Assume that X and Y are (possibly dependent) random variables over $\{1, \ldots, 6\}$.

$$\mathbb{E}[1_{\{X=Y\}}] = \sum_{x=1}^{6} \sum_{y=1}^{6} 1_{\{x=y\}} P_{(X,Y)}(x, y) = \sum_{x=1}^{6} P_{(X,Y)}(x, x) = \mathbb{P}(X = Y).$$

Linearity of the expectation

If X is a discrete r.v. then

$$\mathbb{E}[aX + b] = a\,\mathbb{E}[X] + b$$

$$\mathsf{Var}(\,aX + b) = a^2\,\mathsf{Var}(X)$$

$$\mathsf{std}(\,aX + b) = |a|\,\mathsf{std}(X)$$

proofs

Independent random variables

Two random variables X et Y are independent if one of the three equivalent properties hold

1. $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}$,
$$P_{X,Y}(x, y) = P_X(x)\, P_Y(y).$$

2. $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}$ such that $P_Y(y) > 0$, we have
$$P_{X|Y}(x|y) = P_X(x).$$

3. For any functions f and g,
$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\, \mathbb{E}[g(Y)].$$

Proof of $1 \Leftrightarrow 2$.

Proof of $1 \Leftrightarrow 3$.

Distribution of the sum of two non-negative discrete random variables

Assume that X and Y are independent r.v.s taking values in $\mathbb{N}$ and let $Z = X + Y$. What is the pmf of Z?

By the law of total probability

$$\begin{aligned}
\mathbb{P}(Z = n) &= \sum_{k=0}^{n} \mathbb{P}(Z = n \mid X = k)\, \mathbb{P}(X = k) \\
&= \sum_{k=0}^{n} \mathbb{P}(k + Y = n \mid X = k)\, \mathbb{P}(X = k) \\
&= \sum_{k=0}^{n} \mathbb{P}(Y = n - k \mid X = k)\, \mathbb{P}(X = k) \\
&= \sum_{k=0}^{n} \mathbb{P}(Y = n - k)\, \mathbb{P}(X = k) = \sum_{k=0}^{n} P_Y(n - k)\, P_X(k).
\end{aligned}$$

Distribution of the sum of two discrete random variables

Assume that X and Y are independent r.v.s taking values in $\mathbb{Z}$ and let $Z = X + Y$. What is the pmf of Z?

By the law of total probability

$$\begin{aligned}
\mathbb{P}(Z = n) &= \sum_{k=-\infty}^{+\infty} \mathbb{P}(Z = n \mid X = k)\, \mathbb{P}(X = k) \\
&= \sum_{k=-\infty}^{+\infty} \mathbb{P}(k + Y = n \mid X = k)\, \mathbb{P}(X = k) \\
&= \sum_{k=-\infty}^{+\infty} \mathbb{P}(Y = n - k \mid X = k)\, \mathbb{P}(X = k) \\
&= \sum_{k=-\infty}^{+\infty} \mathbb{P}(Y = n - k)\, \mathbb{P}(X = k) = \sum_{k=-\infty}^{+\infty} P_Y(n - k)\, P_X(k).
\end{aligned}$$

Pmf of the sum of two independent discrete random variables

We have proven the following result:

The pmf of the sum of two independent discrete r.v.s is the convolution of the pmfs

- Let X and Y two independent r.v.s with pmfs P_X and P_Y
- Let $Z = X + Y$

Then Z has a probability mass function p_Z given by:

$$P_Z(z) = (P_X * P_Y)(z) := \sum_{y=-\infty}^{+\infty} P_X(z-y)\, P_Y(y) = \sum_{y=-\infty}^{+\infty} P_X(x)\, P_Y(z-x).$$

We say that

- $P_X * P_Y$ is the convolution of P_X and P_Y
- P_X is convolved with P_Y. (P_X est convoluée avec P_Y.)

Application: Sum of two Poisson r.v.s

Let X and Y be two Poisson r.v.s. with $X \sim \text{Pois}(\theta)$ and $Y \sim \text{Pois}(\eta)$. Let $Z = X + Y$. The pmf of Z is

$$
\begin{aligned}
P_Z(n) &= \sum_{k=0}^{n} P_X(k)\, P_Y(n-k) = \sum_{k=0}^{n} \frac{\theta^k}{k!}\, e^{-\theta}\, \frac{\eta^{n-k}}{(n-k)!} e^{-\eta} \\
&= e^{-(\theta+\eta)}\, \frac{1}{n!} \sum_{k=0}^{n} \frac{n!}{k!(n-k)!}\, \theta^k\, \eta^{n-k} \\
&= e^{-(\theta+\eta)}\, \frac{1}{n!}\, (\theta+\eta)^n \sum_{k=0}^{n} \binom{n}{k} \frac{\theta^k}{(\theta+\eta)^k}\, \frac{\eta^{n-k}}{(\theta+\eta)^{n-k}} \\
&= e^{-(\theta+\eta)}\, \frac{1}{n!}\, (\theta+\eta)^n \sum_{k=0}^{n} \binom{n}{k} p^k\, (1-p)^{n-k} \quad \text{with} \quad p := \frac{\theta}{\theta+\eta} \\
&= \frac{(\theta+\eta)^n}{n!}\, e^{-(\theta+\eta)}.
\end{aligned}
$$

A sum of Poisson r.v.s is a Poisson r.v.

We have proven the following result:

Proposition

If X and Y are two Poisson r.v.s. with $X \sim \text{Pois}(\theta)$ and $Y \sim \text{Pois}(\eta)$,

Then $Z = X + Y$ is a Poisson r.v. with $Z \sim \text{Pois}(\theta + \eta)$.

- The sum of two independent r.v. from a certain family of distribution is not necessarily from the same family:
  - The sum of two uniforms r.v.s is not uniform...
  - The sum of two geometric r.v.s is not geometric...
- The pmf of the sum of two random variable is not obtained as the sum or the mean of the pmf, but instead as the **convolution**: $P_X * P_Y$.

Outline

Continuous random variables

Definition: continuous random variable

We say that a random variable is *continuous* if

- it takes values in a subset of $\mathbb{R}$ or $\mathbb{R}^d$
- it can take an uncountable number of different values.

Example: Uniform random variable on $[0, 1]$.

We use the notation $U \sim \mathcal{U}([0, 1])$ for the random variable such that

$$\forall x, x' \quad s.t. \quad 0 \leq x \leq x' \leq 1, \qquad \mathbb{P}(U \in [x, x'] = x' - x).$$

Cumulative density function (Fonction de répartition)

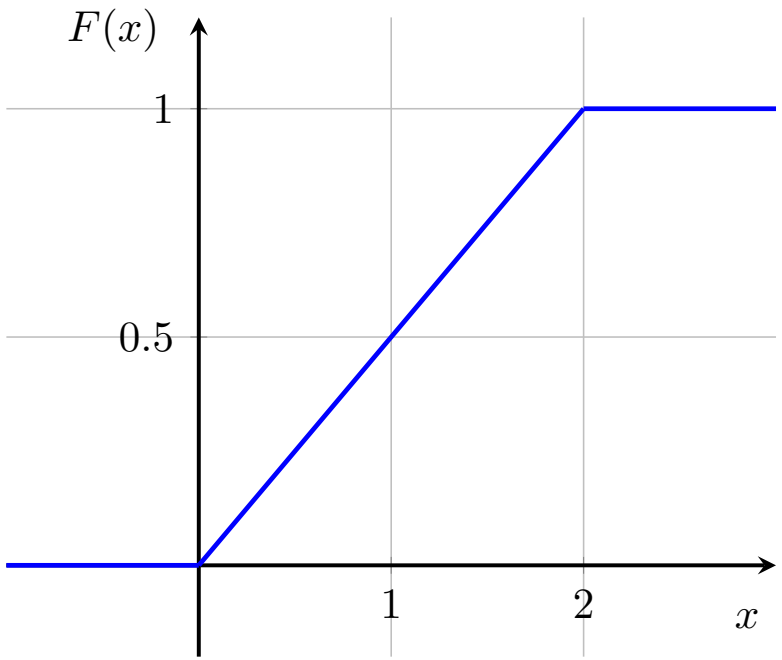
The *cumulative density function* (*fonction de répartition*), or *c.d.f.*, is the function F_X defined by

$$F_X(x) = \mathbb{P}(X \leq x).$$

It is the simplest way to specify the probability distribution of a real-valued r.v.

Example: for a uniform r.v. on $[0, 2]$.

$$F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} 0\,, & \text{if } x \leq 0, \\ \frac{x}{2}\,, & \text{if } 0 \leq x \leq 2, \\ 1\,, & \text{if } x \geq 2. \end{cases}$$

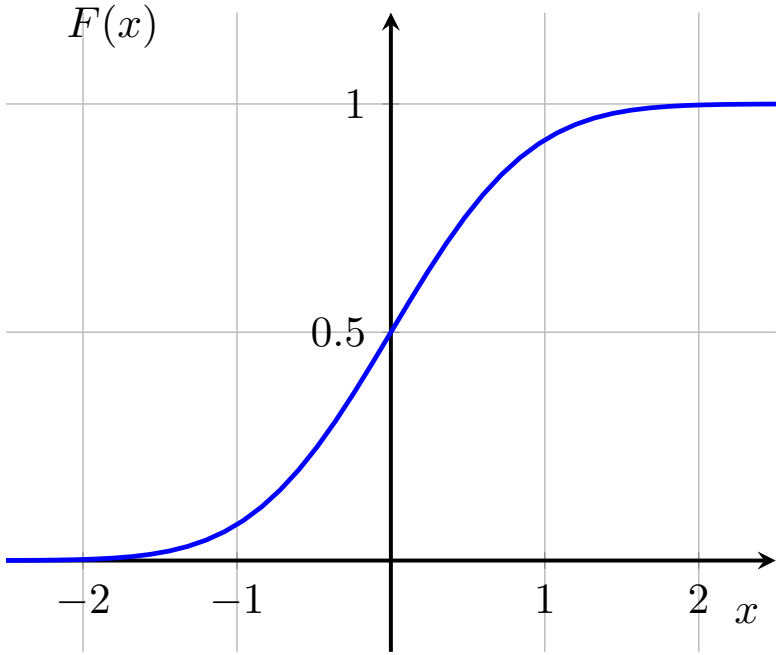


Remark: a cdf must be a non-decreasing function by the second axiom of probability theory, and from 0 to 1 by the third axiom.

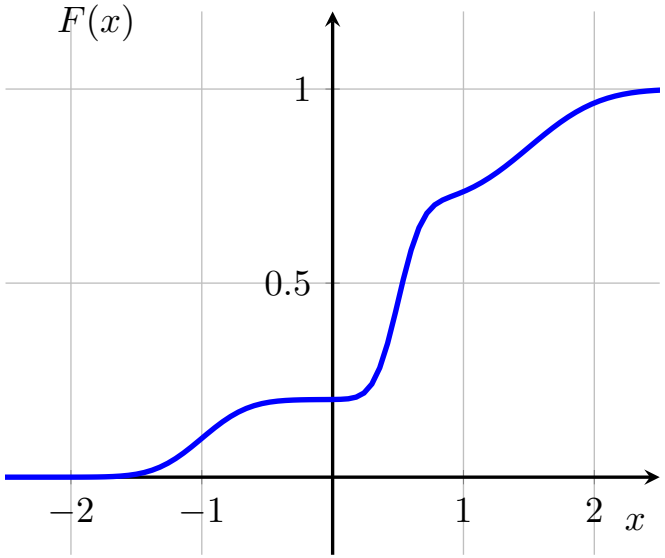
Cumulative density function of the standard normal distribution

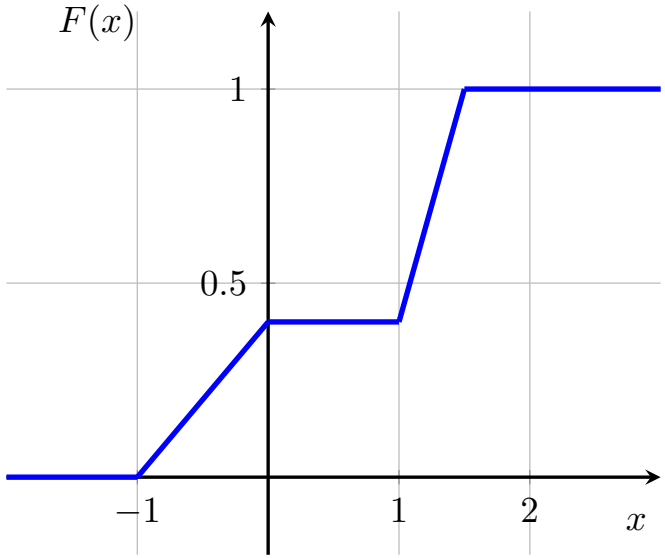
One of the most important distribution in probability and statistics is the *standard normal* or *standard Gaussian* distribution (*gaussienne standard*, ou *normale centrée réduite*). It can be defined from its c.d.f.

$$\mathbb{P}(X \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt$$



More examples of cumulative density functions





In this course, we will consider only continuous cdfs.

It is however possible for a cdf to be discontinuous[1].

[1] One notable example is the *empirical cumulative density function* which is defined for a sample.

Probability density function (Densité de probabilité)

Let F be the c.d.f. of a r.v. X. If there exist a function $f \geq 0$ such that

$$F(x) = \int_{-\infty}^{x} f(t)dt,$$

then f is the *probability density function* of X.

We then have

$$\mathbb{P}(X = x) = \int_{x}^{x} p_X(t)\, dt = 0.$$

$$\mathbb{P}(X \in [a, b]) = F(a) - F(b) = \int_{a}^{b} f(t)\, dt.$$

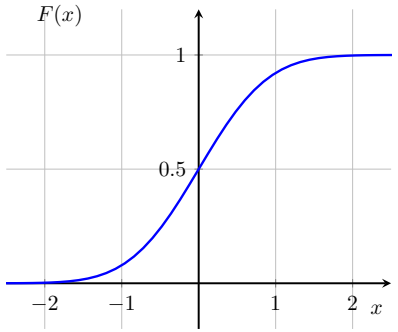
Relations between f and F:

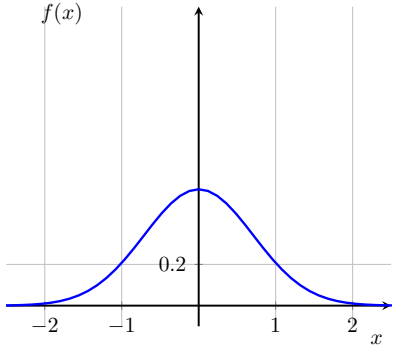
- If F is differentiable and F' is continuous, we have $f(x) = F'(x)$.
- Sometimes F is only piecewise differentiable with F' is continuous on each piece. In that case $f(x) = F'(x)$ on each segment where F' is defined.

Cdfs *vs* pdfs

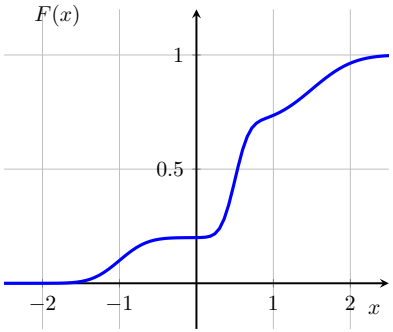
For three cdfs shown in the top row, the corresponding cdfs are shown below.

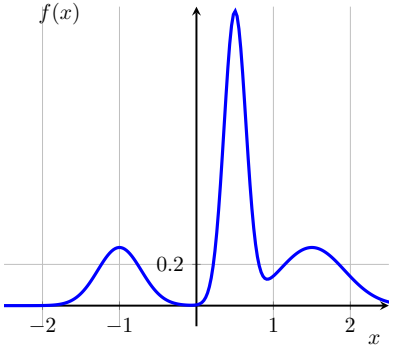
Gaussian



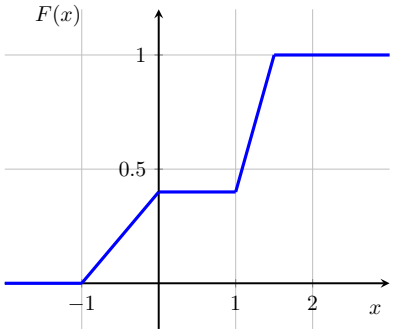


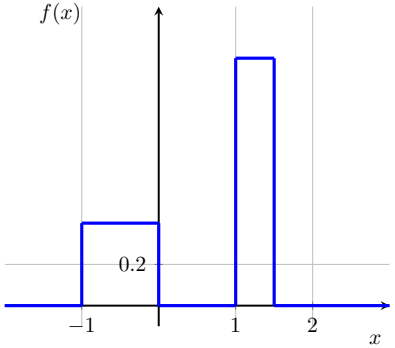
Mixture of Gaussians





Mixture of uniforms





Uniform distribution

Uniform pdf

We say that U taking values in $[a, b]$ follows a uniform distribution on $[a, b]$ and we write $U \sim \mathcal{U}([a, b])$ if its pdf is

$$p_U(u; a, b) = \frac{1}{b-a} 1_{\{a \leq u \leq b\}}.$$

Uniform cdf

For $U \sim \mathcal{U}([a, b])$, its cdf is $\mathbb{P}(U \leq u; a, b) = \begin{cases} 0, & \text{if } u < a, \\ \frac{u-a}{b-a}, & \text{if } a \leq u \leq b, \\ 1, & \text{if } u > b. \end{cases}$

Location & scale: It turns out that if $U \sim \mathcal{U}([0, 1])$ then $U' = a + (b - a)U$ satisfies $U' \sim \mathcal{U}([a, b])$. (proved later)Therefore a is sometimes called a *location* parameter, and $b - a$ a *scale* parameter.

Gaussian distribution

Interactive Normal webpage

The standard normal pdf

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

The Gaussian with mean μ and variance σ^2

We say X follows a Gaussian distribution with expectation μ and variance σ^2, and write $X \sim \mathcal{N}(\mu, \sigma^2)$ if

$$p(x; \mu, \sigma) = \mathcal{N}(x; \mu, \sigma^2) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \Big)$$

Remarks:

- Saying that X follows a *standard normal distribution* is equivalent to $X \sim \mathcal{N}(0, 1)$
- If $X \sim \mathcal{N}(0, 1)$, then for $Y = \mu + \sigma X$ we have $Y \sim \mathcal{N}(\mu, \sigma^2)$.

Exponential distribution

Exponential pdf

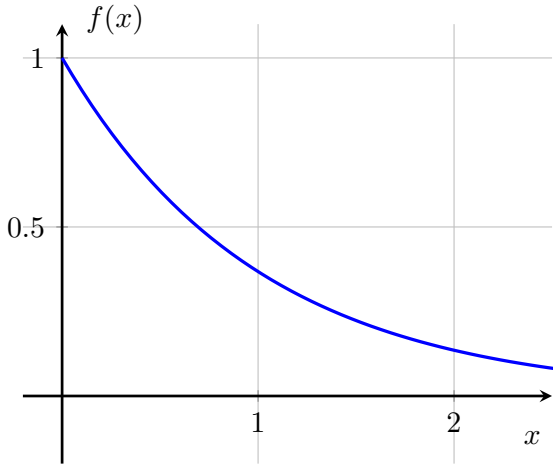
We say that X taking values in $\mathbb{R}_+$ follows an exponential distribution and we write $X \sim \mathcal{E}(\lambda)$ if its pdf is

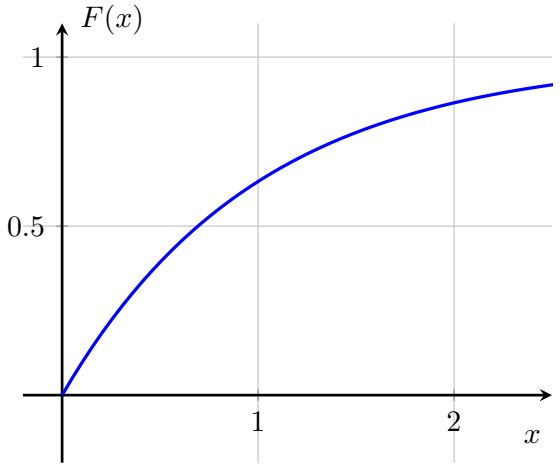
$$p(x;\lambda) = \lambda\, e^{-\lambda x}\, 1_{\{x\geq 0\}}.$$

Exponential cdf

For $X \sim \mathcal{E}(\lambda)$ its cdf is

$$\mathbb{P}(X \leq x;\lambda) = \left(1 - e^{-\lambda x}\right) 1_{\{x\geq 0\}}.$$





Scale: We will prove later in the course that if $X \sim \mathcal{E}(1)$ then $X' = \frac{X}{\lambda} \sim \mathcal{E}(\lambda)$. λ is therefore an *inverse scale* parameter.

Gamma distribution

Gamma pdf

We say that X taking values in $\mathbb{R}_+$ follows a Gamma distribution with *shape* parameter $k > 0$ and *inverse scale* parameter $\lambda > 0$ and we write $X \sim \Gamma(k, \lambda)$ if its pdf is

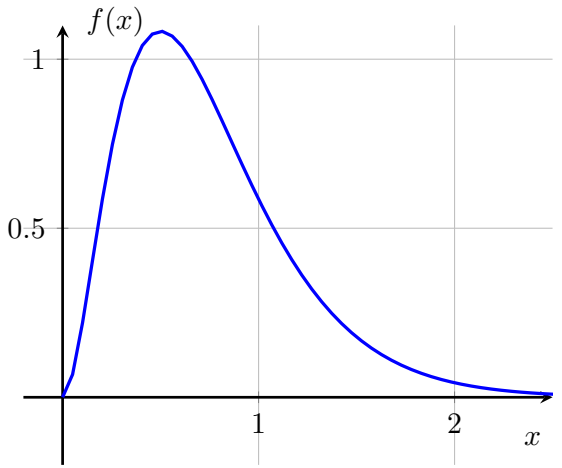
$$p(x; k, \lambda) = \frac{\lambda^k}{\Gamma(k)} x^{k-1} e^{-\lambda x} 1_{\{x \geq 0\}}, \qquad \text{with} \quad \Gamma(k) = \int_0^\infty x^{k-1} e^{-x} \, dx.$$

Remarks:

- $r \mapsto \Gamma(r)$ is the gamma *function*, which satisfies:

$$\forall \alpha > 0, \quad \Gamma(\alpha+1) = \alpha \Gamma(\alpha) \quad \text{and} \quad \forall n \in \mathbb{N}, \quad \Gamma(n+1) = n!$$

- When $k = 1$, we recover the exponential distribution: $\Gamma(1, \lambda) \equiv \mathcal{E}(\lambda)$.



χ^2_n distribution

χ^2_n pdf

We say that Z taking values in $\mathbb{R}_+$ follows a χ^2 distribution with n *degrees of freedom* if it follows a Gamma distribution $Z \sim \Gamma(\frac{n}{2}, \frac{1}{2}) \equiv \chi^2_n$, i.e., if its pdf is

$$p(z;n) = \frac{1}{2^{\frac{n}{2}}\Gamma(1/2)} z^{n/2-1} e^{-\frac{1}{2}z} 1_{\{z\geq 0\}}, \qquad \text{with} \quad \Gamma(r) = \int_0^\infty x^{r-1}e^{-x}\,dx.$$

Remarks:

- We will see later that if $X \sim \mathcal{N}(0,1)$ then $X^2 \sim \chi^2_1$, and that sums of n i.i.d. squared standard normals follow a χ^2_n distribution.

Interactive chi-square webpage

Student's t-distribution

Theorem

Let $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_n \equiv \Gamma(\frac{n}{2}, \frac{1}{2})$ be independent.

Then $T := \frac{Z}{\sqrt{V/n}}$ follows the *Student distribution* with n degrees of freedom,with pdf[a]

$$f_T(t) = \frac{1}{\sqrt{n}\, B(\frac{1}{2}, \frac{n}{2})(1 + \frac{t^2}{n})^{\frac{n+1}{2}}} \qquad \text{with} \qquad B(\tfrac{1}{2}, \tfrac{n}{2}) = \frac{\Gamma(\frac{1}{2})\Gamma(\frac{n}{2})}{\Gamma(\frac{n+1}{2})}.$$

[a] The form of the pdf is beyond scope.

Interactive t-distribution webpage

The relationship between the pdf and the cdf can be visualized for classical distributions on the Seeing theory website (Chap. 3)

Expectation (aka Population Mean)

If X is a continuous random variable with a probability density function $p_X(x)$ then the expectation of a function h of X is defined as

$$\mathbb{E}[h(X)] = \int h(x)\, p_X(x)\, dx,$$

... provided the integral exists !.

How do we know if the integral exists?

- if $\forall x,\ h(x) \geq 0$, then $\mathbb{E}[h(X)]$ always exists(sometimes we can have $\mathbb{E}[h(X)] = +\infty$).
- if $\mathbb{E}[|h(X)|] < \infty$ then $\mathbb{E}[h(X)]$ exists and $|\mathbb{E}[h(X)]| < \infty$.
- as a consequence of the previous point, if a r.v. is bounded then $\mathbb{E}[h(X)]$ exists.

Variance. As for discrete variables, $\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

Linearity of the expectation, etc.

If X is an continuous r.v. then

$$\mathbb{E}[aX + b] = a\,\mathbb{E}[X] + b$$

$$\mathsf{Var}(aX + b) = a^2\,\mathsf{Var}(X)$$

$$\mathsf{std}(aX + b) = |a|\,\mathsf{std}(X)$$

proofs

Example 1: Expectation and Variance of $U \sim \mathcal{U}([2,3])$

We consider $U \sim \mathcal{U}([2,3])$

$$\mathbb{E}[U] =: \int_2^3 u\, p_U(u)\, du = \int_2^3 u\, du = \left[\tfrac{1}{2}u^2\right]_2^3 = \tfrac{1}{2}(9-4) = 2.5.$$

$$\begin{aligned}\mathsf{Var}(U) =: \int_2^3 (u-2.5)^2\, p_U(u)\, du &= \int_2^3 (u-2.5)^2\, du \\ &= \int_{-0.5}^{0.5} t^2\, dt = \left[\tfrac{1}{3}t^3\right]_{-0.5}^{0.5} = \tfrac{1}{3}(0.5^3 - (-0.5)^3) = \tfrac{1}{3}\cdot 2 \cdot \tfrac{1}{8} = \frac{1}{12}.\end{aligned}$$

Example 2: Expectation and Variance of $X \sim \Gamma(k, \lambda)$

Given that the pdf is

$$p(x; k, \lambda) = \frac{\lambda^k}{\Gamma(k)} x^{k-1} e^{-\lambda x} 1_{\{x \geq 0\}}, \qquad \text{with} \quad \Gamma(k) = \int_0^\infty \lambda^k x^{k-1} e^{-\lambda x} \, dx,$$

we have $$\mathbb{E}[X] = \int_0^\infty x \, \frac{\lambda^k}{\Gamma(k)} x^{k-1} e^{-\lambda x} = \frac{\Gamma(k+1)}{\Gamma(k)\lambda} \int_0^\infty \frac{\lambda^{k+1}}{\Gamma(k+1)} x^k e^{-\lambda x} = \frac{k}{\lambda},$$

since $\Gamma(k+1) = k\Gamma(k)$.

and $$\mathbb{E}[X^2] = \int_0^\infty x^2 \, \frac{\lambda^k}{\Gamma(k)} x^{k-1} e^{-\lambda x} = \frac{\Gamma(k+2)}{\Gamma(k)\lambda^2} \int_0^\infty \frac{\lambda^{k+2}}{\Gamma(k+2)} x^{k+1} e^{-\lambda x} = \frac{k(k+1)}{\lambda^2}.$$

$$\boxed{\mathbb{E}[X] = \frac{k}{\lambda} \qquad \text{Var}(X) = \frac{k}{\lambda^2}}.$$

Example 3: Expectation and Variance of $X \sim \mathcal{N}(\mu, \sigma^2)$

Expectation

The distribution is symmetric around μ, so by symmetry, we must have $\boxed{\mathbb{E}[X] = \mu}$.

Variance

- For $\mathrm{Var}(X)$, if $X \sim \mathcal{N}(\mu, \sigma^2)$, we can write $X = \mu + \sigma\tilde{X}$ for $\tilde{X} \sim \mathcal{N}(0, 1)$.
- $\mathrm{Var}(X) = \sigma^2 \mathrm{Var}(\tilde{X})$, so we just need to compute $\mathrm{Var}(\tilde{X}) = \mathbb{E}[\tilde{X}^2]$, since $\mathbb{E}[\tilde{X}] = 0$,
- We will prove later in this course that $Y := \tilde{X}^2$ follows a $\chi^2_1 \equiv \Gamma(\frac{1}{2}, \frac{1}{2})$ distribution.
- But we proved that if $Y \sim \Gamma(k, \lambda)$ then $\mathbb{E}[X] = \frac{k}{\lambda}$. So

$$\mathrm{Var}(\tilde{X}) = \mathbb{E}[\tilde{X}^2] = \mathbb{E}[Y] = \frac{1/2}{1/2} = 1.$$

Finally, we proved $\boxed{\mathrm{Var}(X) = \sigma^2}$.

The Cauchy distribution: a distribution without expectation

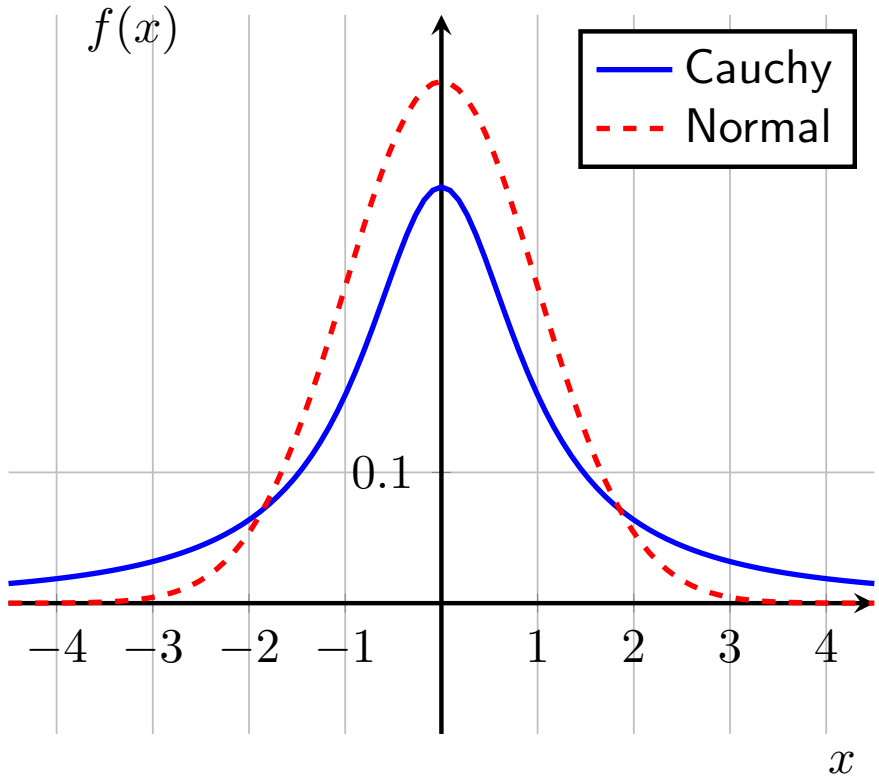
$$p(t) = \frac{1}{\pi(1+t^2)}$$

$$\begin{aligned}\int_0^x t\,p(t)\,dt &= \tfrac{1}{\pi}\int_0^x \frac{t}{1+t^2}\,dt \\ &= \tfrac{1}{2\pi}\left[\log(1+t^2)\right]_0^x \\ &= \tfrac{1}{2\pi}\log(1+x^2) \underset{x\to+\infty}{\longrightarrow} +\infty.\end{aligned}$$

So, we have

$$\begin{cases}\int_0^{+\infty} t\,p(t)\,dt = +\infty \\ \int_{-\infty}^0 t\,p(t)\,dt = -\infty\end{cases}$$

which means that $\mathbb{E}[X]$ does not exist.



For some distributions, $\mathbb{E}[X]$ exists but $\mathbb{E}[X^k]$ does not exist...

Probability of a set

For a random variable that has a continuous cdf F, we have

$$\mathbb{P}(X \in [a,b]) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) = F_X(b) - F_X(a).$$

If X has a pdf p_X, we have $\mathbb{P}(X \in [a,b]) = F_X(b) - F_X(a) = \int_a^b p_X(x)dx.$

For a $A \subset \mathbb{R}$, its indicator function $x \mapsto 1_{\{x \in A\}}$ is defined as

$$1_{\{x \in A\}} = \begin{cases} 1, & \text{if } x \in A \\ 0, & \text{if not.} \end{cases}$$

With this notation, $\mathbb{P}(X \in [a,b]) = \int_a^b p_X(x)dx = \int 1_{\{x \in [a,b]\}}\, p_X(x)dx = \mathbb{E}\big[1_{\{X \in [a,b]\}}\big].$

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

Summary

- Continuous r.v.s take an uncountable number of different values (in $\mathbb{R}$ or $\mathbb{R}^d$).
- Continuous r.v.s have a cumulative density function (cdf) F with $F(x) := \mathbb{P}(X \leq x)$.
- When X has a pdf, $\mathbb{P}(X = x) = 0$ for all x.
- When F is differentiable then $p_X := F'$ is the probability density function (pdf)
- The expectation of $f(X)$ for X with a pdf is $\mathbb{E}[f(X)] = \int f(x)\, p_X(x)\, dx.$
- The probability of event is also an expectation

$$\mathbb{P}(X \in [a, b]) = F(a) - F(b) = \int_a^b p_X(x)\, dx = \mathbb{E}[1_{\{x \in [a,b]\}}].$$

Outline

Support and Range of a scalar random variable

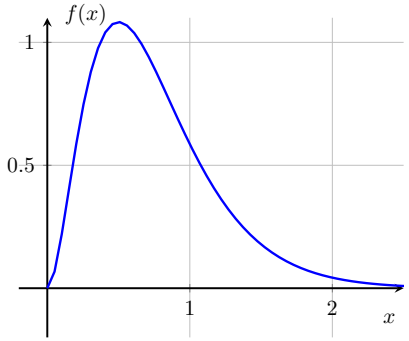
Support:

If X has a pdf p_X then the support of the distribution of X is[2] $\text{Supp}(X) = \overline{\{x \mid p_X(x) > 0\}}$.

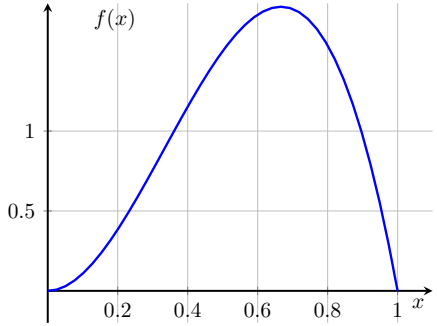
Range:

The range of the distribution is the smallest closed interval containing the support.

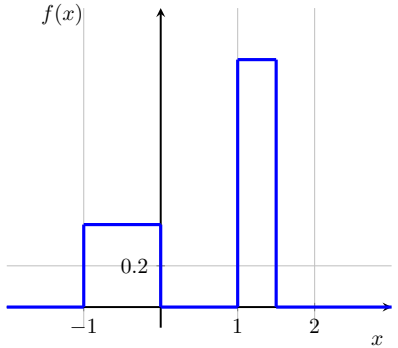
Examples:



Support=Range=$[0, +\infty)$



Support=Range=$[0, 1]$



Support $= [-1, 0] \cup [1, 1.5]$,
Range $= [-1, 1.5]$

[2] If $A \subset \mathbb{R}$ is a set, we denote by $\overline{A}$ the smallest closed set containing A.

Invertible cumulative density function

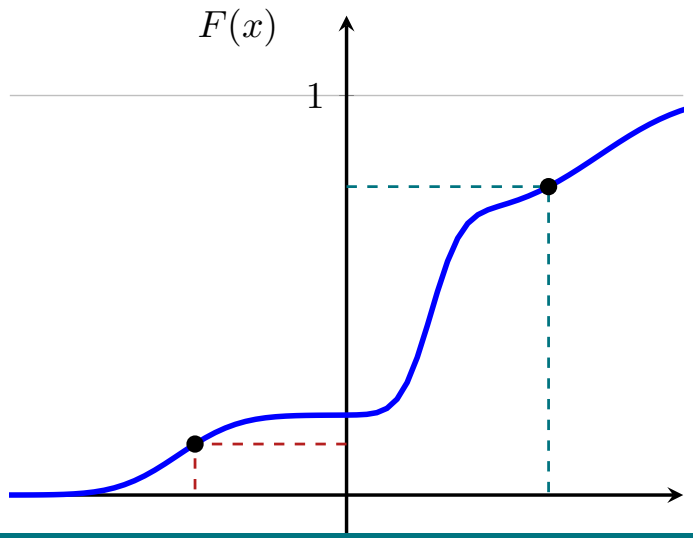
We say that a cdf F is "*invertible on the support of the distribution*" or just "*invertible*" if

$$\forall \alpha \in (0,1), \quad \exists \text{ a unique } x \in \mathbb{R} \quad \text{such that} \quad F(x) = \alpha.$$

- In that case we can define a function: $F^{-1} : (0,1) \to \mathbb{R}$ such that

$$\forall \alpha \in (0,1), \quad F^{-1}(\alpha) \text{ is the unique value } x_\alpha \in \mathbb{R} \text{ such that } F(x_\alpha) = \alpha.$$

- We call F^{-1} the *inverse cdf*.
- All the classical cdfs are invertible: uniform, Gaussian, Gamma, Beta, etc.



Quantiles of a probability distribution

Given the pdf or cdf of a r.v. X it is often useful to be able answer questions of the form:

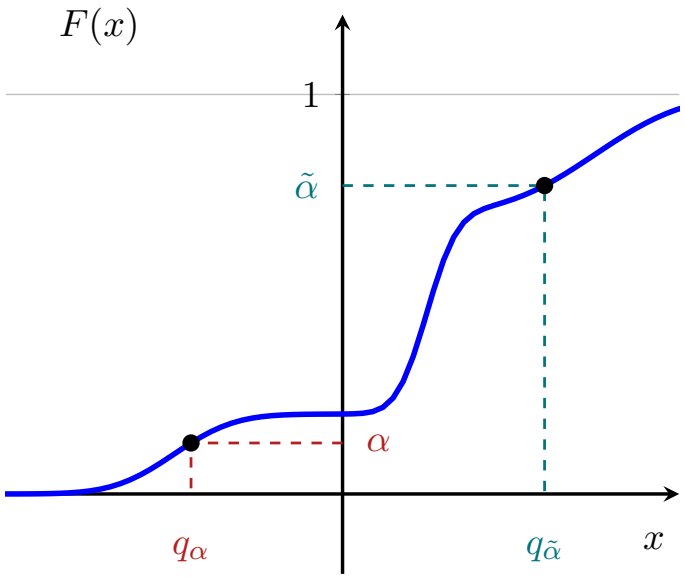
"What is the value of q such that $X < q$ with probability 0.95?"

If the cdf F is invertible it is easy to find this value using F^{-1} :

Quantile of level α of a r.v. X with an invertible cdf.

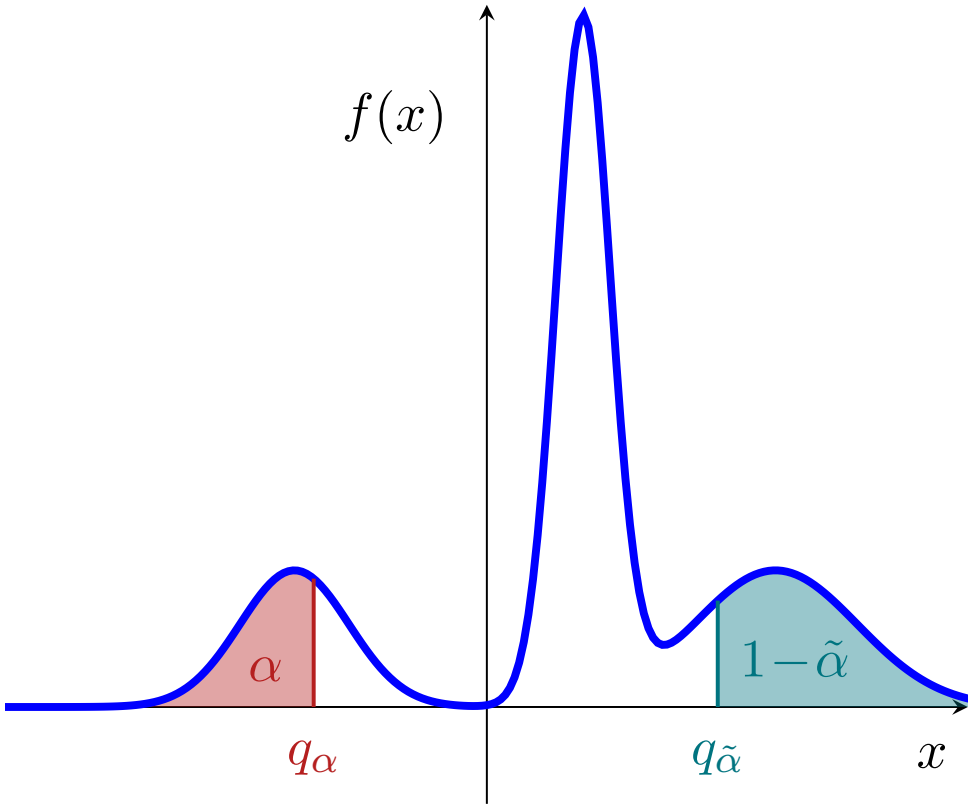
The quantile of level α of X is the unique value q_α such that

$$\mathbb{P}(X \leq q_\alpha) = \alpha \quad \text{or equivalently} \quad q_\alpha = F^{-1}(\alpha).$$



Interpretation of quantiles on the pdf

$$\mathbb{P}(X \leq q_\alpha) = \alpha$$
$$\mathbb{P}(X > q_{\tilde{\alpha}}) = 1 - \tilde{\alpha}$$



- The area under the pdf to the left of q_α is exactly equal to α.
- The area under the pdf to the right of $q_{\tilde{\alpha}}$ is exactly equal to $1 - \tilde{\alpha}$.
- The area under the pdf to the right of $q_{1-\alpha}$ is exactly equal to α.

Quantiles will be key to construct *confidence intervals* and *rejection regions* for hypothesis tests.

Median, quartiles and percentiles

Median

The quantile of level $\alpha = 0.5$ of a distribution is called the median: $m := q_{0.5}$
We therefore have $\mathbb{P}(X \leq m) = 0.5 = \mathbb{P}(X > m)$.
The median is the point such that half of the "probability mass" is on either side of m.

Quartiles

The quartiles are $q_{0.25}$ and $q_{0.75}$. The interquartile is the interval $[q_{0.25}, q_{0.75}]$.

Percentile

If α is in % then we call it *percentile*, e.g., $q_{0.90}$ is the 90th percentile of the distribution.

Empirical quantile

Quantiles can also be defined for a sample. The empirical quantile $\hat{q}_\alpha$ of level α is the value such that a fraction $\frac{\lceil \alpha n \rceil}{n}$ of the data is smaller than $\hat{q}_\alpha$.

Mode

- A mode of a pdf is a local maximum of the pdf (or a point where the pdf becomes infinite).
- If a pdf has a single mode, we say that it is *unimodal*, and we can talk about *"the" mode* of the distribution.
- If it has several isolated modes we say that it is *multimodal*.

Examples:

- the Normal, Exponential, Gamma, and Student distributions are unimodal.
- For Beta distributions, they are sometimes unimodal, sometimes unimodal, depending on the parameters.
- A way to obtain multimodal distribution is to use "mixtures" of distributions.

Sampling from a continuous random variable

Let F be an invertible cdf and F^{-1} its inverse.

Sampling from a r.v. with cdf F from a standard uniform

If

- $U \sim \mathcal{U}[0,1]$
- $X := F^{-1}(U)$

then X is a random variable with cdf equal to F:

$$\mathbb{P}(X \leq x) = F(x).$$

Proof: By definition $\mathbb{P}(U \leq u) = u$. As a consequence;

$$\mathbb{P}(X \leq x) = \mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x).$$

Example: Sampling from an exponential distribution

We consider an exponential r.v. with pdf

$$p(x;\lambda) = \lambda e^{-\lambda x}\, 1_{\{x\geq 0\}}.$$

Then the cdf is $F(x) = \displaystyle\int_0^x \lambda e^{-\lambda t} = \big[-e^{-\lambda t}\big]_0^x = 1 - e^{-\lambda x}$.

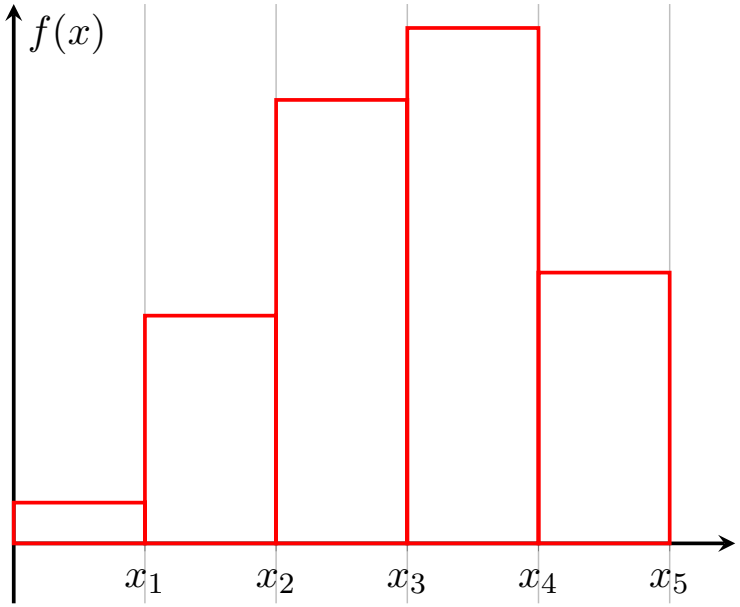
So we can compute $F^{-1}(u) = -\frac{1}{\lambda}\log(1-u)$. This means that if $U \sim \mathcal{U}[0,1]$, then

$$X := -\frac{1}{\lambda}\log(1-U) \sim \mathcal{E}(\lambda).$$

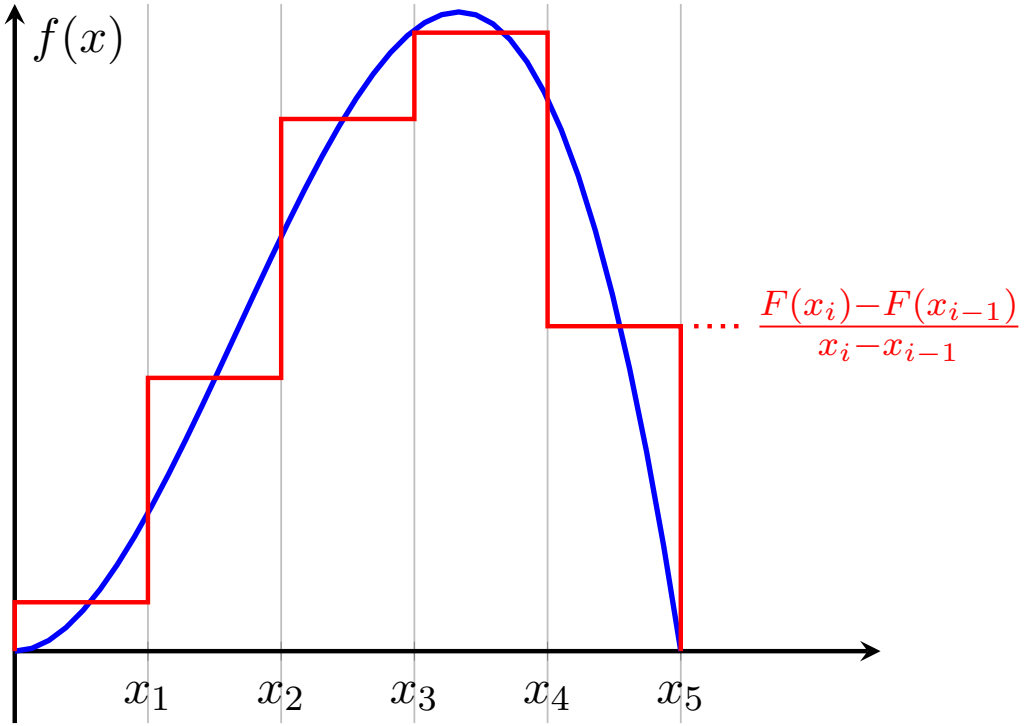
Histograms

- A histogram is typically built from a sample to approximate a density.
- A partition $x_0 < \ldots < x_k$ specifies bins $[x_{k-1}, x_k]$
- The fraction of datapoints which fall in $[x_{k-1}, x_k]$ estimates $\mathbb{P}(X \in [x_{k-1}, x_k])$

What is the connection with probability densities?



Histogram p.d.f., the connection between pdfs and histograms...



- Partition $x_0 < x_1 < \ldots < x_5$
- Original density p_X with cdf F
- Histogram p.d.f. p_h on the partition.

If $X \sim p_X$ and $Y \sim p_h$

$$
\begin{aligned}
&= \mathbb{P}(Y \in [x_{i-1}, x_i]) \\
&= \frac{F(x_i) - F(x_{i-1})}{x_i - x_{i-1}} \cdot (x_i - x_{i-1}) \\
&= F(x_i) - F(x_{i-1}) \\
&= \mathbb{P}(X \in [x_{i-1}, x_i])
\end{aligned}
$$

p_h is a piecewise constant approximation of p_X which assigns the same probability as p_X to each interval $[x_{i-1}, x_i]$.

p_h can therefore be estimated directly from data.

Summary

- The *support* S of a distribution with pdf p_X is $\overline{\{x \mid p_X(x) > 0\}}$
- If the cdf F is invertible on S, we define the quantile function $\alpha \mapsto F^{-1}(\alpha) = q_\alpha$.
- The main property of quantiles is that $\mathbb{P}(X \leq q_\alpha) = \alpha$.
- The median $q_{0.5}$, quartiles $q_{0.25}, q_{0.75}$, and percentiles are particular quantiles.
- Modes are local maxima of the density. Unimodal distributions have a single mode.
- If $U \sim \mathcal{U}([0,1])$, then $X := F^{-1}(U)$ is a r.v. with cdf equal to F.
- The previous property gives a way to sample from any cdf from a uniform sampler.
- A histogram pdf is a piecewise constant approximation of a density on a partition, equal to the mean value of the pdf in each interval of the partition.

Outline

Joint cdfs and densities

We can define the joint cdf for a pair of r.v. (X, Y) by

$$F(x, y) := \mathbb{P}(X \leq x, Y \leq y) := \mathbb{P}\big((X \leq x) \& (Y \leq y)\big).$$

Any function $(x, y) \mapsto f(x, y)$ such that

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(\xi, u)\, d\xi\, du$$

is a joint probability density for the pair (X, Y).

If F is piecewise $\mathcal{C}_2$, a joint probability density can be defined as

$$p_{X,Y}(x, y) = \frac{\partial}{\partial x} \frac{\partial}{\partial y} F(x, y).$$

Conditional density

If $p_{X,Y}(x,y)$ is a joint probability density for the pair of r.v. $(X,Y) \in \mathbb{R}^2$

- We can recover the marginal densities

$$p_X(x) = \int_{\mathbb{R}} p_{X,Y}(x,y)\,dy \quad \text{and} \quad p_Y(y) = \int_{\mathbb{R}} p_{X,Y}(x,y)\,dx.$$

- We can define the conditional density of Y given $X = x$, as follows:

$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)} \quad \text{and} \quad p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}.$$

- As a consequence, we have Bayes' rule:

$$p_{Y|X}(y|x) = \frac{p_{X|Y}(x|y)\,p_Y(y)}{p_X(x)}.$$

Conditional expectation

If $p_{X|Y}(x|y)$ is the conditional probability density of X given Y, then

$$\mathbb{E}[X|Y=y] = \int x\, p_{X|Y}(x|y)\, dx$$

Joint distribution with both discrete and continuous variables

It is perfectly possible to define a joint distribution between a discrete variable and a continuous variable. For example, we can define the pair (Z, X) as follows

- Z is a discrete variable taking value in $\{1, \dots, K\}$ with probability

$$\mathbb{P}(Z = k) = P_Z(k) = \pi_k$$

- given $Z = k$, then X follows a Gaussian distribution $\mathcal{N}(\mu_k, 1)$, that is that

$$p_{X|Z}(x|k) = \mathcal{N}(x; \mu_k, 1) := \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu_k)^2}.$$

Then the joint distribution is specified by $p_{X,Z}(x, k) = p_{X|Z}(x|k) P_X(k) = \mathcal{N}(x; \mu_k, 1)\, \pi_k$.

And we have $p_X(x) = \sum_{k=1}^{K} \mathcal{N}(x; \mu_k, 1)\, \pi_k$ and $P_{Z|X}(k|x) = \frac{\mathcal{N}(x; \mu_k, 1)\, \pi_k}{p_X(x)}$.

Pair of independent continuous random variables

Two random variables X and Y with a joint pdf are independent if one of the three equivalent properties hold

1. $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}$,
$$p_{X,Y}(x, y) = p_X(x)\, p_Y(y).$$
2. $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}$ such that $p_Y(y) > 0$, we have
$$p_{X|Y}(x|y) = p_X(x).$$
3. For any functions f and g,
$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\,\mathbb{E}[g(Y)].$$

The proofs are essentially the same as for discrete variables but replacing sums by integrals.

Independent continuous random variables

A collection of random variables $X_1, \ldots, X_n$ are independent if one of the three equivalent properties hold

1. $\forall (x_1, \ldots, x_n) \in \mathcal{X}_1 \times \ldots \times \mathcal{X}_n$,

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = p_{X_1}(x_1) \ldots p_{X_n}(x_n).$$

2. $\forall (x_1, \ldots, x_n) \in \mathcal{X}_1 \times \ldots \times \mathcal{X}_n$ such that $p_{X_{-i}}(x_{-i}) > 0$, we have for all i,

$$p_{X_i|X_{-i}}(x_i|x_{-i}) = p_{X_i}(x_i) \qquad \text{where} \quad X_{-i} := (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n).$$

3. For any functions $f_1, \ldots, f_n$,

$$\mathbb{E}[f_1(X_1) \ldots f_n(X_n)] = \prod_{i=1}^{n} \mathbb{E}[f_i(X_i)].$$

Variance, covariance and correlation

For real valued r.v. X and Y,

$$\mathsf{Var}[X] = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big],$$

$$\mathsf{cov}(X, Y) = \mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y].$$

$$\mathsf{corr}(X, Y) = \frac{\mathsf{cov}(X, Y)}{\sigma_X \sigma_Y} \qquad \text{with} \qquad \sigma_X^2 = \mathsf{Var}[X],\ \sigma_Y^2 = \mathsf{Var}[Y].$$

Now assuming that X is taking values in $\mathbb{R}^d$,

$$\mathsf{Cov}(X) = \mathbb{E}\big[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top\big].$$

Properties of the Variance

The following properties can be verified immediately:

- $\mathrm{Var}(aX + b) = a^2 \,\mathrm{Var}(X)$.
- $\mathrm{cov}(aX + b, cY + d) = ac\,\mathrm{cov}(X, Y)$.
- $\mathrm{corr}(aX + b, cY + d) = \mathrm{corr}(X, Y)$.
- If X and Y are independent, $\mathrm{cov}(X, Y) = 0$.
- In general $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + 2\,\mathrm{cov}(X, Y) + \mathrm{Var}(Y)$
- If X and Y are independent, $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Finally, we also have $|\mathrm{corr}(X, Y)| \leq 1$. The proof is beyond scope.

Sampling from a pair of random variables

Let (X, Y) be a pair of r.v. with joint density $p_{(X,Y)}(x, y)$.

It is usually difficult to sample directly the pair.

However it is possible to

1. Sample $X \sim p_X$ to obtain x
2. Sample $Y \sim p_{(Y|X)(\cdot|x)}$ to obtain y

Note that each step is sampling a scalar random variable.

Pmfs vs pdfs

A number of formulas and results take the same form for pmfs and pdfs by simply replacing sums by integrals.
However it is important to always keep in mind that

- the pmf $P_X(x)$ is the probability of the set $\{x\}$, i.e. $\boxed{P_X(x) = \mathbb{P}(X = x)}$
- while for a pdf $p_X(x)$, we have $\boxed{p_X(x) \neq \mathbb{P}(X = x)}$
- instead $$\boxed{p_X(x) = F'_X(x) = \lim_{h \downarrow 0} \frac{F(x+h) - F(x)}{h} = \lim_{h \downarrow 0} \frac{\mathbb{P}(x \leq X \leq x+h)}{h}}$$

Summary for Joint distribution over r.v.s

- For a pair of r.v.s, $F(x,y) := \mathbb{P}(X \leq x, y \leq y)$ is the joint pdf.
- The joint pdf is $p_{X,Y}(x,y) = \frac{\partial}{\partial x}\frac{\partial}{\partial y}F(x,y)$.
- The marginal density is $p_X(x) = \int_{-\infty}^{\infty} p_{X,Y}(x,y)\,dy$.
- The conditional density is $p_{Y|X}(y|x) = \dfrac{p_{X,Y}(x,y)}{p_X(x)}$.
- Bayes's rule relates both conditionals and marginals: $p_{Y|X}(y|x) = \dfrac{p_{X|Y}(x|y)p_Y(y)}{p_X(x)}$.
- We saw 3 equivalent properties of independence, which generalize to n independent variables.
- We saw several properties of the variance and covariance
- ⚠ For pdfs, $p_X(x) \neq \mathbb{P}(X = x)$.

Proofs and extra material

(beyond the scope of the course)

Linearity of the expectation for a discrete variable: proofs

If X is a discrete r.v. with *probability mass function* P_X with $P_X(x) := \mathbb{P}(X = x)$, then

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$$

Proof:

$$\mathbb{E}[aX + b] = \sum_{x \in \mathcal{X}} (ax + b) P_X(x) = a \sum_{x \in \mathcal{X}} x\, P_X(x) + b \sum_{x \in \mathcal{X}} P_X(x) = a\, \mathbb{E}[X] + b. \quad \square$$

$$\mathsf{Var}(aX + b) = a^2\, \mathsf{Var}(X)$$

Proof:

$$\mathsf{Var}(aX+b) = \mathbb{E}\big[\big(aX+b-(a\mathbb{E}[X]+b)\big)^2\big] = \mathbb{E}\big[\big(a(X-\mathbb{E}[X])\big)^2\big] = a^2\, \mathbb{E}\big[(X-\mathbb{E}[X])^2\big]. \quad \square$$

back

Linearity of the expectation for a discrete variable: proofs

If X is a continuous r.v. with pdf p_X, then

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$$

Proof:

$$\mathbb{E}[aX+b] = \int_{-\infty}^{\infty} (ax+b)p_X(x)\,dx = a\int_{-\infty}^{\infty} x\,P_X(x)\,dx + b\int_{-\infty}^{\infty} p_X(x)\,dx = a\,\mathbb{E}[X]+b. \quad \square$$

$$\mathsf{Var}(aX + b) = a^2\,\mathsf{Var}(X)$$

The proof is the same as for the discrete case *Proof*:

$$\mathsf{Var}(aX+b) = \mathbb{E}\big[\big(aX+b-(a\mathbb{E}[X]+b)\big)^2\big] = \mathbb{E}\big[\big(a(X-\mathbb{E}[X])\big)^2\big] = a^2\,\mathbb{E}\big[(X-\mathbb{E}[X])^2\big]. \quad \square$$

back

Independence: proof of the equivalence of 1 and 2

Proof of 1 $\Rightarrow$ 2:

If $P_Y(y) > 0$ then $P_{X|Y}(x|y) = \dfrac{P_{X,Y}(x,y)}{P_Y(y)} = \dfrac{P_X(x)\,P_Y(y)}{P_Y(y)} = P_X(x).$

Proof of 2 $\Rightarrow$ 1:

If $P_Y(y) = 0$ then no matter what $P_{X|Y}(x|y)$ is $P_{X,Y}(x,y) = P_{X|Y}(x|y)P_Y(y) = 0$ and so the equality $P_{X,Y}(x,y) = P_X(x)P_Y(y)$ holds trivially. Otherwise

$$P_{X,Y}(x,y) = P_{X|Y}(x|y)P_Y(y) = P_X(x)P_Y(y),$$

which proves the result.

back

Independence: proof of the equivalence of 1 and 3

Proof of 1 $\Rightarrow$ 3:

$$\mathbb{E}[f(X)g(Y)] = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} f(x)g(y)P_{X,Y}(x,y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} f(x)\,g(y)P_X(x)P_Y(y)$$
$$= \Big(\sum_{x\in\mathcal{X}} f(x)\mathbb{P}(X=x)\Big)\Big(\sum_{y\in\mathcal{Y}} g(y)P_X(x)P_Y(y)\Big) = \mathbb{E}[f(X)]\,\mathbb{E}[g(Y)].$$

Proof of 3 $\Rightarrow$ 1:

If we take $f(x) = 1_{\{x=x_0\}}$ and $g(y) = 1_{\{y=y_0\}}$, then

On the one hand, $\mathbb{E}[f(X)g(Y)] = \mathbb{E}[1_{\{X=x_0\}}1_{\{Y=y_0\}}] = \mathbb{P}(X=x_0, Y=y_0) = P_{X,Y}(x_0,y_0)$.

On the other,
$\mathbb{E}[f(X)] = \mathbb{E}[1_{\{X=x_0\}}] = \mathbb{P}(X=x_0) = P_X(x_0)$, and similarly $\mathbb{E}[g(Y)] = P_Y(y_0)$.

Combining the two, we get $P_{X,Y}(x_0,y_0) = P_X(x_0)\,P_Y(y_0)$.

Since this is true for any (x_0,y_0) this proves that property 1 holds.

The Cauchy-Schwarz inequality

For $\mathbf{x}, \boldsymbol{y} \in \mathbb{R}^d$ the Cauchy-Schwartz inequality says that $\langle \mathbf{x}, \boldsymbol{y} \rangle| \leq \|\mathbf{x}\| \, \|\boldsymbol{y}\|$.

Cauchy-Schwarz for random variables

Let X and Y be real-valued random variables.

$$|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2]\,\mathbb{E}[Y^2]}$$

Proof: If either $\mathbb{E}[X^2]$ or $\mathbb{E}[Y^2]$ is infinite, the inequality holds. Otherwise We have $2|XY| \leq \frac{1}{c}X^2 + cY^2$ which proves

$$2\mathbb{E}[XY] \leq 2\mathbb{E}[|XY|] \leq \tfrac{1}{c}\,\mathbb{E}[X^2] + c\,\mathbb{E}[Y^2]$$

By setting $c = \sqrt{\frac{\mathbb{E}[X^2]}{\mathbb{E}[Y^2]}}$ and considering both the case of X and $-X$, we get the result.

The covariance inequality

Let X and Y two v.a. such that $\mathbb{E}[|X|^2] < \infty$ and $\mathbb{E}[|Y|^2] < \infty$.

By applying the Cauchy-Schwartz inequality to $\breve{X} = X - \mathbb{E}[X]$ and $\breve{Y} = Y - \mathbb{E}[Y]$, we have

$$|\mathsf{cov}(X,Y)| = |\mathbb{E}[\breve{X}\breve{Y}]| \leq \sqrt{\mathbb{E}[\breve{X}^2]\,\mathbb{E}[\breve{Y}^2]} = \sqrt{\mathsf{Var}(X)\,\mathsf{Var}(Y)}.$$

so that $$|\mathsf{corr}(X,Y)| \leq 1.$$

One hot encoding and the Multinomial

EE-209 - Eléments de Statistiques pour les Data Sciences

One hot encoding

When working with a *nominal* or *ordinal variable* X taking values in $\{1, \ldots, K\}$, it is always more convenient in statistics and machine learning to work with its **indicator vector** representation often called **one hot encoding**:

$$Z = (Z_1, \ldots, Z_K) \quad \text{with} \quad Z_1 = 1_{\{X=1\}}, \quad \ldots, \quad Z_K = 1_{\{X=K\}}.$$

This called **one hot encoding** because Z_k takes values in $\{0, 1\}$ and $Z_1 + \ldots + Z_K = 1$.

Example: For X taking values in $\{1, \ldots, 4\}$ with $P_X(k) = \pi_k$.

$P_X(x)$	x	z
π_1	1	$(1, 0, 0, 0)$
π_2	2	$(0, 1, 0, 0)$
π_3	3	$(0, 0, 1, 0)$
π_4	4	$(0, 0, 0, 1)$

Note that we have:

$$\mathbb{E}[Z_k] = \mathbb{P}(Z_k = 1) = \mathbb{P}(X = k) = \pi_k$$

Continuing with the same example on the next slide...

Counts from sampling a discrete r.v. and the Multinomial

Sampling $n = 17$ independent values from X:

x_i	$P_X(x_i)$	z_{i1}	z_{i2}	z_{i3}	z_{i4}
1	π_1	1	0	0	0
4	π_4	0	0	0	1
1	π_1	1	0	0	0
2	π_2	0	1	0	0
4	π_4	0	0	0	1
4	π_4	0	0	0	1
1	π_1	1	0	0	0
3	π_3	0	0	1	0
4	π_4	0	0	0	1
2	π_2	0	1	0	0
4	π_4	0	0	0	1
3	π_3	0	0	1	0
3	π_3	0	0	1	0
4	π_4	0	0	0	1
3	π_3	0	0	1	0
3	π_3	0	0	1	0
1	π_1	1	0	0	0

	n_1	n_2	n_3	n_4
Counts	4	2	5	6

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} P_X(x_i)$$
$$= \pi_1^{n_1}\, \pi_2^{n_2}\, \pi_3^{n_3}\, \pi_4^{n_4}$$

The probability of the observed sequence depends only on the counts, but...

$$\mathbb{P}(N_1 = n_1, \ldots, N_4 = n_4) = \binom{n}{n_1, n_2, n_3, n_4} \pi_1^{n_1}\, \pi_2^{n_2}\, \pi_3^{n_3}\, \pi_4^{n_4}$$

with the *multinomial coefficient*

$$\binom{n}{n_1, n_2, n_3, n_4} := \frac{n!}{n_1!\, n_2!\, n_3!\, n_4!},$$

which counts the number of ways that a sequence of n numbers in $\{1, 2, 3, 4\}$ in which 1,2,3 and 4 appears respectively exactly n_1, n_2, n_3, n_4 times.

Multinomial random variable

Multinomial pmf

A vector of discrete r.v. $(N_1, \ldots, N_K)$ is said to follow jointly a multinomial distribution with parameters n and $(\pi_1, \ldots, \pi_K)$ and we write $(N_1, \ldots, N_K) \sim \mathcal{M}\big(n, (\pi_1, \ldots, \pi_k)\big)$ if $N_1 + \ldots + N_K = n$ and

$$\mathbb{P}(N_1 = n_1, \ldots, N_K = n_K) = \binom{n}{n_1, \ldots, n_K} \pi_1^{n_1} \ \ldots \pi_K^{n_K},$$

with $\binom{n}{n_1, \ldots, n_K} := \frac{n!}{n_1! \ldots n_K!}$, the *multinomial coefficient*.

Remark:

$$(N_1, N_2) \sim \mathcal{M}\big(n, (\pi_1, 1 - \pi_1)\big) \quad \Leftrightarrow \quad N_1 \sim \mathsf{Bin}(n, \pi_1).$$

The "Multinoulli"?

What happens for $(N_1, \ldots, N_K) \sim \mathcal{M}(1, (\pi_1, \ldots, \pi_K))$?

- $(N_1, \ldots, N_K)$ is a vector of counts such that $N_1 + \ldots + N_k = 1$ so it has to be an **indicator vector** !
- Moreover $1! = 0! = 1$ so $\binom{n}{n_1, \ldots, n_K} = 1$ for all possible values of $n_1, \ldots, n_K$.
- So if $(Z_1, \ldots, Z_K) \sim \mathcal{M}(1, (\pi_1, \ldots, \pi_K))$, then it is an indicator vector and

$$\boxed{\mathbb{P}(Z_1 = z_1, \ldots, Z_K = z_k) = \pi_1^{z_1} \ldots \pi_K^{z_K}}$$

- The distribution of the indicator vector is therefore the counterpart of the Bernoulli and becomes the Bernoulli for $K = 2$.

The Bernoulli, the Binomial, the "Multinoulli" and the Multinomial

$Z \sim \mathsf{Ber}(\pi)$	$(Z_1, \ldots, Z_K) \sim \mathcal{M}(1, \pi_1, \ldots, \pi_K)$
$P_Z(z) = \pi^z (1-\pi)^{1-z}$	$P_{\boldsymbol{Z}}(\boldsymbol{z}) = \pi_1^{z_1} \ldots \pi_K^{z_K}$
$N_1 \sim \mathsf{Bin}(n, \pi)$	$(N_1, \ldots, N_K) \sim \mathcal{M}(n, \pi_1, \ldots, \pi_K)$
$P_{N_1}(n_1) = \binom{n}{n_1} \pi^{n_1} (1-\pi)^{n-n_1}$	$P_{\boldsymbol{N}}(\mathbf{n}) = \binom{n}{n_1, \ldots, n_K} \pi_1^{n_1} \ldots \pi_K^{n_K}$

with

$$\binom{n}{i} = \frac{n!}{(n-i)!i!} \quad \text{and} \quad \binom{n}{n_1, \ldots, n_K} = \frac{n!}{n_1! \ldots n_K!}$$

Continuous random variables: part B

EE-209 - Eléments de Statistiques pour les Data Sciences

Outline

Joint cdfs and densities

We can define the joint cdf for a pair of r.v. (X, Y) by

$$F(x, y) := \mathbb{P}(X \leq x, Y \leq y) := \mathbb{P}\big((X \leq x) \& (Y \leq y)\big).$$

Any function $(x, y) \mapsto f(x, y)$ such that

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(\xi, u)\, d\xi\, du$$

is a joint probability density for the pair (X, Y).

If F is piecewise $\mathcal{C}_2$, a joint probability density can be defined as

$$p_{X,Y}(x, y) = \frac{\partial}{\partial x} \frac{\partial}{\partial y} F(x, y).$$

Conditional density

If $p_{X,Y}(x,y)$ is a joint probability density for the pair of r.v. $(X,Y) \in \mathbb{R}^2$

- We can recover the marginal densities

$$p_X(x) = \int_{\mathbb{R}} p_{X,Y}(x,y)\,dy \quad \text{and} \quad p_Y(y) = \int_{\mathbb{R}} p_{X,Y}(x,y)\,dx.$$

- We can define the conditional density of Y given $X = x$, as follows:

$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)} \quad \text{and} \quad p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}.$$

- As a consequence, we have Bayes' rule:

$$p_{Y|X}(y|x) = \frac{p_{X|Y}(x|y)\,p_Y(y)}{p_X(x)}.$$

Conditional expectation

If $p_{X|Y}(x|y)$ is the conditional probability density of X given Y, then

$$\mathbb{E}[X|Y=y] = \int x\, p_{X|Y}(x|y)\, dx$$

$$\mathbb{E}[X|Y] = \int x\, p_{X|Y}(x|Y)\, dx$$

Law of Total Expectation

$$\mathbb{E}\big[\,\mathbb{E}[X|Y]\,\big] = \mathbb{E}[X]$$

Pair of independent continuous random variables

Two random variables X and Y with a joint pdf are independent if one of the three equivalent properties hold

1. $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}$,
$$p_{X,Y}(x, y) = p_X(x)\, p_Y(y).$$
2. $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}$ such that $p_Y(y) > 0$, we have
$$p_{X|Y}(x|y) = p_X(x).$$
3. For any functions f and g,
$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\, \mathbb{E}[g(Y)].$$

The proofs are essentially the same as for discrete variables but replacing sums by integrals.

Independent continuous random variables

A collection of random variables $X_1, \ldots, X_n$ are independent if one of the three equivalent properties hold

1. $\forall (x_1, \ldots, x_n) \in \mathcal{X}_1 \times \ldots \times \mathcal{X}_n$,

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = p_{X_1}(x_1) \ \ldots \ p_{X_n}(x_n).$$

2. $\forall (x_1, \ldots, x_n) \in \mathcal{X}_1 \times \ldots \times \mathcal{X}_n$ such that $p_{X_{-i}}(x_{-i}) > 0$, we have for all i,

$$p_{X_i|X_{-i}}(x_i|x_{-i}) = p_{X_i}(x_i) \qquad \text{where} \quad X_{-i} := (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n).$$

3. For any functions $f_1, \ldots, f_n$,

$$\mathbb{E}[f_1(X_1) \ldots f_n(X_n)] = \prod_{i=1}^{n} \mathbb{E}[f_i(X_i)].$$

Variance, covariance and correlation

For real valued r.v. X and Y,

$$\mathsf{Var}[X] = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big],$$

$$\mathsf{cov}(X, Y) = \mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y].$$

$$\mathsf{corr}(X, Y) = \frac{\mathsf{cov}(X, Y)}{\sigma_X \sigma_Y} \qquad \text{with} \qquad \sigma_X^2 = \mathsf{Var}[X],\ \sigma_Y^2 = \mathsf{Var}[Y].$$

Now assuming that X is taking values in $\mathbb{R}^d$,

$$\mathsf{Cov}(X) = \mathbb{E}\big[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top\big].$$

Properties of the Variance

The following properties can be verified immediately:

- $\mathrm{Var}(aX + b) = a^2 \,\mathrm{Var}(X)$.
- $\mathrm{cov}(aX + b, cY + d) = ac\,\mathrm{cov}(X, Y)$.
- $\mathrm{corr}(aX + b, cY + d) = \mathrm{corr}(X, Y)$.
- If X and Y are independent, $\mathrm{cov}(X, Y) = 0$.
- In general $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + 2\,\mathrm{cov}(X, Y) + \mathrm{Var}(Y)$
- If X and Y are independent, $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Finally, we also have $|\mathrm{corr}(X, Y)| \leq 1$. The proof is beyond scope.

Joint distribution between a discrete and a continuous variable

It is obvious possible to define the joint distribution between a discrete r.v. C and a continuous r.v. X.

$$p_{X|C}(x|c)\,P_C(c) \qquad \text{or} \qquad P_{C|X}(c|x)\,p_X(x)$$

The joint distribution is neither a pdf nor a pmf as defined in this course but is usually denoted like a pdf $p_{X,C}(x,c)$. (In a more abstract theory of probability the pmf is actually viewed as a "discrete" pdf.)

The **marginal distributions** can be computed using the **law of total probability** takes the forms:

$$p_X(x) = \sum_{c\in\mathcal{C}} p_{X|C}(x|c)\,P_C(c) \qquad \text{and} \qquad P_C(c) = \int P_{C|X}(c|x)\,p_X(x)\,dx.$$

The finite mixture of Gaussians

Finite mixtures of distributions are obtained as the marginal distribution $p_X(x)$ associated with a joint distribution between a discrete variable C and a continuous variables X.

The mixture of Gaussians is the most classical example:

- Let C take values in $\{1, \ldots, 3\}$ with pmf $P_C(k) = \pi_k$
- We can define the conditional density of X given C to be

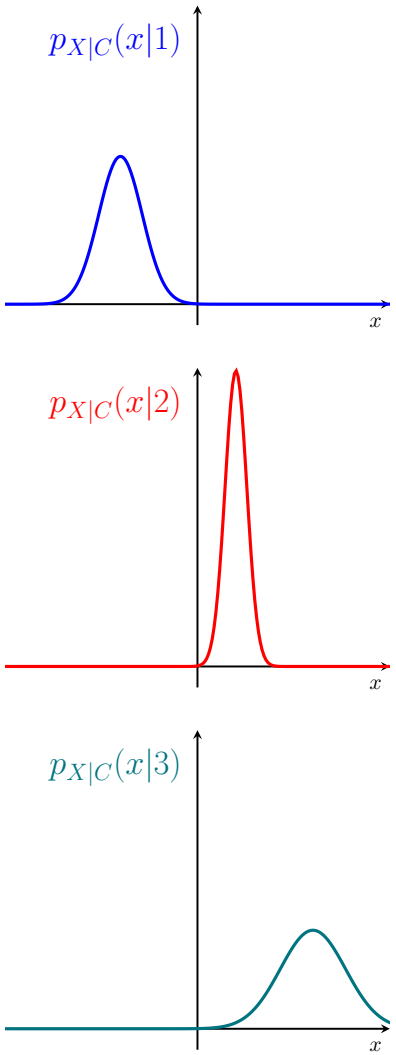
$$p_{X|C}(x|k) = \mathcal{N}(x; \mu_k, \sigma_k^2) := \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(x-\mu_k)^2}{2\,\sigma_k^2}}.$$

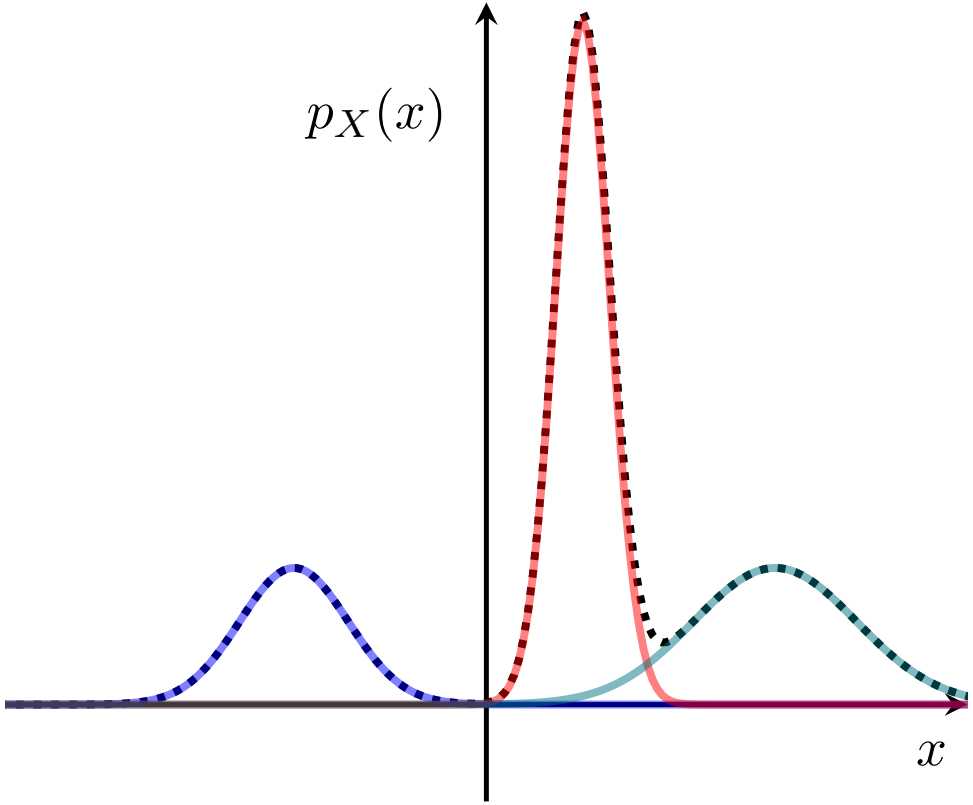
- The marginal is then

$$p_X(x) = \mathcal{N}(x; \mu_1, \sigma_1^2)\,\pi_1 + \mathcal{N}(x; \mu_2, \sigma_2^2)\,\pi_2 + \mathcal{N}(x; \mu_3, \sigma_3^2)\,\pi_3$$

- And by Bayes' rule, we have

$$P_C(k|x) = \frac{p_{X|C}(x|k)P_C(k)}{p_X(x)} = \frac{\mathcal{N}(x; \mu_k, \sigma_k^2)\,\pi_k}{\sum_{j=1}^{3} \mathcal{N}(x; \mu_j, \sigma_j^2)\,\pi_j}.$$





$$p_X(x) = p_{X|C}(x|1)\, P_C(1) + p_{X|C}(x|2)\, P_C(2) + p_{X|C}(x|3)\, P_C(3).$$

$$p_X(x) = \mathcal{N}(x; \mu_1, \sigma_1^2)\, \pi_1 + \mathcal{N}(x; \mu_2, \sigma_2^2)\, \pi_2 + \mathcal{N}(x; \mu_3, \sigma_3^2)\, \pi_3$$

Mixture model and weighted means of densities

We consider the heights distributions of two populations: men and women. Each can be modelled as Gaussian with the following parameters:

	μ	σ
women	161cm	7.1cm
men	175cm	8.5cm

- What is the distribution of height of a "mixed" population which has

$$\pi_1 = 40\% \text{ women and } \pi_0 = 60\% \text{ men?}$$

Let S be the sex variable, with $S = 1$ coding for female and $S = 0$ coding for male. We have

$$\begin{aligned} p(x) &= p(x|S=1)\,p(S=1) + p(x|S=0)\,p(S=0) \\ &= \pi_1\,\mathcal{N}(x;\mu_1,\sigma_1^2) + \pi_0\,\mathcal{N}(x;\mu_0,\sigma_0^2) \\ &= 0.4\,\mathcal{N}(x;161,7.1^2) + 0.6\,\mathcal{N}(x;175,8.5^2) \end{aligned}$$

Sampling from a pair of random variables

Let (X, Y) be a pair of r.v. with joint density $p_{(X,Y)}(x, y)$.

It is usually difficult to sample directly the pair.

However it is possible to

1. Sample $X \sim p_X$ to obtain x
2. Sample $Y \sim p_{(Y|X)(\cdot|x)}$ to obtain y

Note that each step is sampling a scalar random variable.

Pmfs vs pdfs

A number of formulas and results take the same form for pmfs and pdfs by simply replacing sums by integrals.
However it is important to always keep in mind that

- the pmf $P_X(x)$ is the probability of the set $\{x\}$, i.e. $\boxed{P_X(x) = \mathbb{P}(X = x)}$
- while for a pdf $p_X(x)$, we have $\boxed{p_X(x) \neq \mathbb{P}(X = x)}$
- instead $$\boxed{p_X(x) = F'_X(x) = \lim_{h \downarrow 0} \frac{F(x+h) - F(x)}{h} = \lim_{h \downarrow 0} \frac{\mathbb{P}(x \leq X \leq x+h)}{h}}$$

Summary for Joint distribution over r.v.s

- For a pair of r.v.s, $F(x,y) := \mathbb{P}(X \leq x, y \leq y)$ is the joint pdf.
- The joint pdf is $p_{X,Y}(x,y) = \frac{\partial}{\partial x}\frac{\partial}{\partial y}F(x,y)$.
- The marginal density is $p_X(x) = \int_{-\infty}^{\infty} p_{X,Y}(x,y)\,dy$.
- The conditional density is $p_{Y|X}(y|x) = \dfrac{p_{X,Y}(x,y)}{p_X(x)}$.
- Bayes's rule relates both conditionals and marginals: $p_{Y|X}(y|x) = \dfrac{p_{X|Y}(x|y)p_Y(y)}{p_X(x)}$.
- We saw 3 equivalent properties of independence, which generalize to n independent variables.
- We saw several properties of the variance and covariance
- ⚠ For pdfs, $p_X(x) \neq \mathbb{P}(X = x)$.

Outline

Pdf of the sum of two independent continuous random variables

The pdf of the sum of two independent continuous random variable is the convolution of the pdfs

- Let X and Y two independent r.v.s with pdfs p_X and p_Y
- Let $Z = X + Y$

Then Z has a probability density p_Z given by:

$$p_Z(z) = (p_X * p_Y)(z) := \int_{-\infty}^{+\infty} p_X(z-y)\, p_Y(y)\, dy = \int_{-\infty}^{+\infty} p_X(x)\, p_Y(z-x)\, dx.$$

We say that

- $p_X * p_Y$ is the convolution of p_X and p_Y
- p_X is convolved with p_Y.

proof

Application: pdf of the sum of two independent $\mathcal{U}[0,1]$

If U and V are independent uniform distributions on $[0,1]$, what is the distribution of $Y = U + V$?

We have $p_U(u) = p_V(u) = 1_{\{0 \le t \le 1\}}$, and by the previous theorem

$$\begin{aligned} f(y) &= \int_{-\infty}^{\infty} p_V(y-u)\, p_U(u)\, du \\ &= \int_{-\infty}^{\infty} 1_{\{0 \le y-u \le 1\}}\, 1_{\{0 \le u \le 1\}}\, du \\ &= 1_{\{0 \le y \le 1\}} \int_0^y du + 1_{\{1 \le y \le 2\}} \int_{y-1}^1 du \\ &= y\, 1_{\{0 \le y \le 1\}} + (2-y)\, 1_{\{1 \le y \le 2\}}. \end{aligned}$$

Pdf of a scaled version of a random variable $Y = aX$

If X is a continuous r.v. with pdf p_X what is the pdf of $Y = aX$ for $a > 0$?

We can again use the cdf

$$F_Y(y) := \mathbb{P}(Y \leq y) = \mathbb{P}(aX \leq y) = \mathbb{P}\Big(X \leq \frac{y}{a}\Big) = F_X\Big(\frac{y}{a}\Big)$$

$$p_Y(y) = F_Y'(y) = \frac{\partial F_X\left(\frac{y}{a}\right)}{\partial y} = \frac{\partial\left(\frac{y}{a}\right)}{\partial y} F_X'\left(\frac{y}{a}\right) = \frac{1}{a} p_X\left(\frac{y}{a}\right)$$

Pdf of $Y = aX$ when $a > 0$

$$p_Y(y) = \tfrac{1}{a} p_X\left(\tfrac{y}{a}\right)$$

Application:

Let X be an exponential r.v. with pdf $f_X(x) = e^{-x}$, what is the density of $Y = \frac{1}{\lambda}X$?

By the previous result, $f_Y(y) = \lambda e^{-\lambda y}$. This is the generic form of an exponential r.v.. We see that the different exponential r.v. are simply scaled versions of one another.

The square of a $\mathcal{N}(0,1)$ is a $\chi^2(1)$

Proposition

If $X \sim \mathcal{N}(0,1)$ and $Y = X^2$ then $Y \sim \chi^2(1) \equiv \Gamma(\frac{1}{2}, \frac{1}{2})$, and in particular

$$p_Y(y) = \frac{1}{\sqrt{2\pi y}} e^{-\frac{1}{2}y}$$

Proof that $X \sim \mathcal{N}(0,1) \;\Rightarrow\; X^2 \sim \chi^2(1)$

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(X^2 \le y) = \mathbb{P}(-\sqrt{y} \le X \le \sqrt{y}) = 2\,\mathbb{P}(0 \le X \le \sqrt{y})$$
$$= 2\big[F_X(\sqrt{y}) - F_X(0)\big]$$

The function F_Y is differentiable for any $y > 0$ and so for any $y > 0$

$$p_Y(y) = F_Y'(y) = 2\frac{d}{dy}\left[F_X(\sqrt{y}) - F_X(0)\right] = 2\frac{1}{2\sqrt{y}}\,F_X'(\sqrt{y})$$
$$= \frac{1}{\sqrt{y}}\,p_X(\sqrt{y}) = \frac{1}{\sqrt{2\pi y}}\,e^{-\frac{1}{2}y}.$$

Summary

In this lecture we have seen

- $f_{X+Y} = f_X * f_Y$
- $f_{aX} = \frac{1}{a} f_X(\frac{\cdot}{a})$
- if $Z \sim \mathcal{N}(0,1)$ then $Z \sim \chi^2_1$.

Outline

I.i.d. sample

At the heart of statistic there is the idea of having a **sample** $\{X_1, \ldots, X_n\}$ of **observations** from a **population**. A "nice" sample is a sample which is i.i.d.

Identically distributed

All observations come from the same population, and so have same distribution/pmf/pdf:

$$\mathbb{P}(X_1 \leq t) = \mathbb{P}(X_i \leq t) \qquad \text{or equivalently} \qquad p_{X_1} = p_{X_i} \quad \text{with} \quad X_i \sim p_{X_i}.$$

Independent

Each observation is drawn purely at random without any dependence to the others, so

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = p_{X_1}(x_1) \ldots p_{X_n}(x_n)$$

I.i.d. = Independent + Identically distributed

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = p(x_1) \ldots p(x_n), \qquad \text{with} \quad p = p_{X_1}.$$

Sample and samples

Statistics and CS/signal processing have different naming conventions.

	Statistics	CS
x_i	observation	sample
$\{x_1, \ldots, x_n\}$	sample	dataset

In this course, I will use the statistics terminology.

Properties of sums and means of i.i.d. r.v.s

- Let $X_1, \ldots, X_n$ be independent, then

$$\mathsf{Var}(X_1 + \ldots + X_n) = \sum_{i=1}^{n} \mathsf{Var}(X_i).$$

- If the r.v. $X_1, \ldots, X_n$ are i.i.d., then

$$\mathsf{Var}(\bar{X}) = \tfrac{1}{n}\mathsf{Var}(X_1).$$

The family of Gaussian distributions is stable

A family $\mathcal{F}$ of random variables/distributions such that

"if X and Y belong to $\mathcal{F}$, then $\alpha X + \beta Y$ also belongs to $\mathcal{F}$" is called a "stable" family of distributions.

Distribution of a linear combination of Gaussians

Let $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ two *independent* r.v.s, then

$$a_1 X_1 + a_2 X_2 \sim \mathcal{N}(a_1\mu_1 + a_2\mu_2,\ a_1^2\sigma_1^2 + a_2^2\sigma_2^2).$$

This shows that the family of Gaussian distributions is stable.

There are other stable families but most of them are complicated. The other one that we have encountered is the family of Cauchy distributions (not obvious).

Distributions of some sums of continuous r.v.s

Sum of $\Gamma(r_i, \lambda)$ random variables

- If $X_1, \ldots, X_n$ are independent with $X_i \sim \Gamma(r_i, \lambda)$ and $Z = X_1 + \ldots + X_n$ then

$$Z \sim \Gamma(r_1 + \ldots + r_n, \lambda).$$

In particular, if we set $r_i = 1$ in the previous result, we get

Sum of i.i.d. Exponential random variables

- If $X_1, \ldots, X_n$ be i.i.d. with $X_i \sim \mathcal{E}(\lambda)$ and $Z = X_1 + \ldots + X_n$ then $Z \sim \Gamma(n, \lambda)$.

And if we consider the case $r_i = \lambda = \frac{1}{2}$, we get

Sum of squares of i.i.d. standard normal random variables

- If $X_1, \ldots, X_n$ be i.i.d. with $X_i \sim \mathcal{N}(0, 1)$ and $Z = X_1^2 + \ldots + X_n^2$ then $X_i^2 \sim \chi^2(1)$ and

$$\forall i, \quad X_i^2 \sim \chi^2(1) \equiv \Gamma(\tfrac{1}{2}, \tfrac{1}{2}), \quad \text{and} \quad Z \sim \chi^2(n) \equiv \Gamma(\tfrac{n}{2}, \tfrac{1}{2}).$$

Summary

In this lecture we have seen

- The concepts of sample and observation from a population
- The concepts of independent observations and of identically distributed observations
- The definition of **i.i.d.** random variables

For some random variables the distribution of the sum of a sample has a known form:

- If $X_i \sim \Gamma(r_i, \lambda)$ independent and $Y = X_1 + \ldots + X_n$, then $Y \sim \Gamma(r_1 + \ldots + r_n, \lambda)$.
- If $E_i \overset{\text{i.i.d.}}{\sim} \mathcal{E}(\lambda)$, then $E_1 + \ldots + E_n \sim \Gamma(n, \lambda)$.
- If $Z_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$, then $Z_1 + \ldots + Z_n \sim \chi^2_n$.

Finally, any linear combination of independent Gaussians is Gaussian.

Proofs and extra material

(beyond the scope of the course)

Sum of two indep. r.v.s: Proof

$$
\begin{aligned}
F_Z(t) = \mathbb{P}(Z \le t) &= \mathbb{P}(X + Y \le t) = \mathbb{P}(X \le t - Y) \\
&= \int_{-\infty}^{\infty} \mathbb{P}(X \le t - Y \mid Y = y)\, p_Y(y)\, dy && \text{law of total probability} \\
&= \int_{-\infty}^{\infty} \mathbb{P}(X \le t - y)\, p_Y(y)\, dy && \text{by independence of } X \text{ and } Y \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{t-y} p_X(x)\, dx\, p_Y(y)\, dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{t} p_X(z-y)\, dz\, p_Y(y)\, dy && \text{change of var.: } x = z - y \\
&= \int_{-\infty}^{t} \int_{-\infty}^{\infty} p_X(z-y)\, p_Y(y)\, dy\, dz && \text{exchanging integration order} \\
&= \int_{-\infty}^{t} f(z)\, dz && \text{with } f(z) := \int_{-\infty}^{\infty} p_X(z-y)\, p_Y(y)\, dy.
\end{aligned}
$$

This shows that f is actually a valid pdf for Z.

Law of Large Numbers
&
Central Limit Theorem

EE-209 - Eléments de Statistiques pour les Data Sciences

Convergence of random variables

Convergence in Probability (convergence en probabilités)

We say that a sequence of r.v.s X_n converges in probability to X and write $X_n \xrightarrow{\mathbb{P}} X$ if

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \varepsilon) = 0, \qquad \text{for any } \varepsilon > 0.$$

Almost sure convergence (convergence presque sûre)

We say that a sequence of r.v.s X_n converges almost surely to X and write $X_n \xrightarrow{\text{a.s.}} X$ if

$$\mathbb{P}\Big(\lim_{n\to\infty} X_n = X\Big) = 1.$$

Theorem

$$\big(X_n \xrightarrow{\text{a.s.}} X\big) \quad \underset{\not\Leftarrow}{\Rightarrow} \quad \big(X_n \xrightarrow{\mathbb{P}} X\big)$$

Strong Law of Large Numbers (SLLN)

If $X_1, \ldots, X_n$ are i.i.d. with $\mathbb{E}[|f(X_1)|] < \infty$, then, as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \xrightarrow{\text{a.s.}} \mathbb{E}[f(X_1)].$$

- In particular,
  the **sample mean** $\bar{X}$ converges to the **population mean** or **expectation** $\mathbb{E}[X_1]$.
- If for some identically distributed r.v.s we have $\frac{1}{n}\sum_{i=1}^{n} f(X_i) \xrightarrow{\mathbb{P}} \mathbb{E}[f(X_1)]$, then we say that there is
  a *weak* law of large numbers.
- The strong law of large numbers implies the weak law of large numbers.

LLN for the empirical mean of i.i.d. Bernoullis $Ber(p)$

We consider

- a sample $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Ber}(p)$.
- $\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ the fraction of the throws where heads was observed.

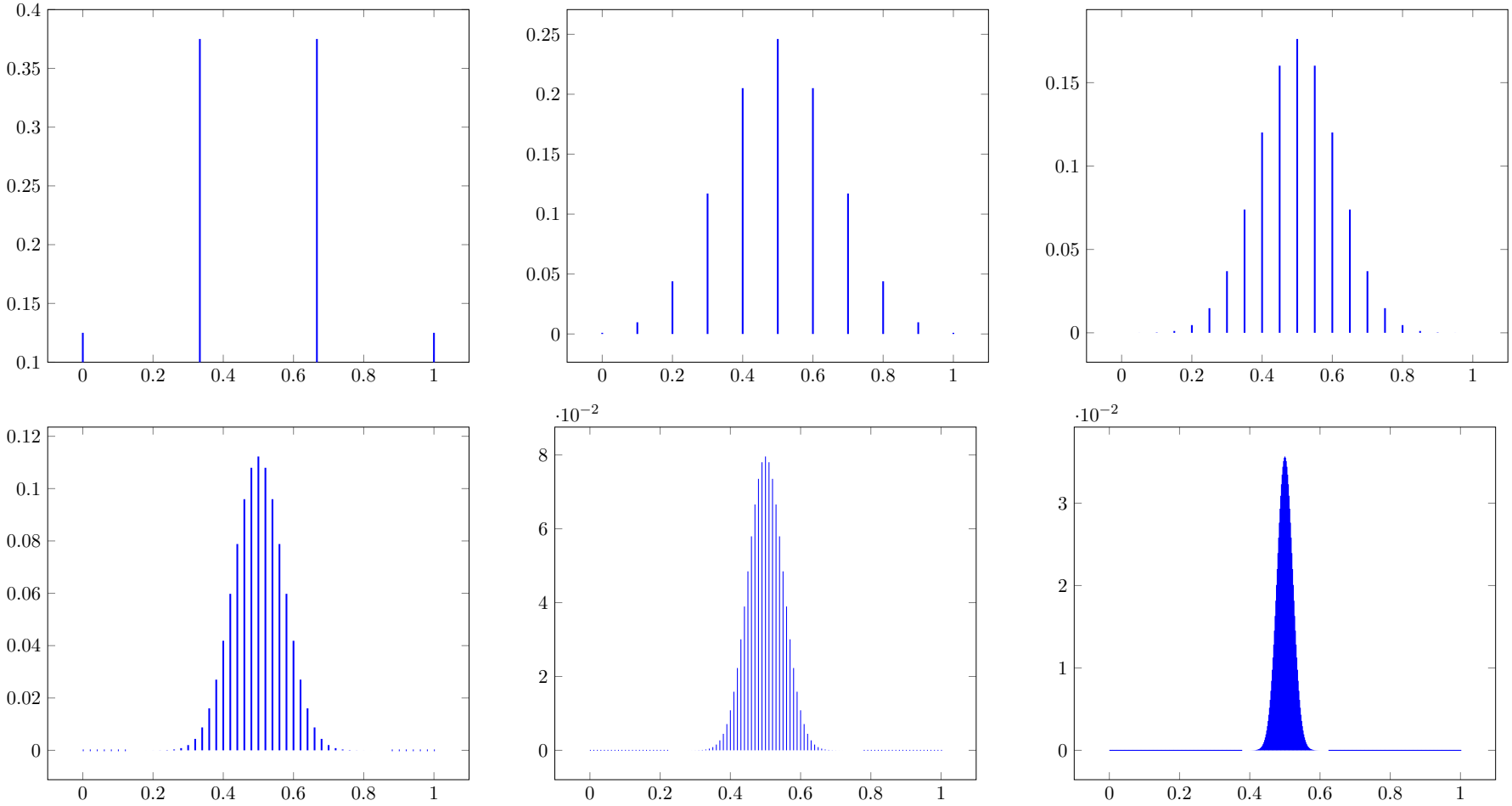
Since $\mathbb{E}[|X_1|] = p < \infty$, by the SLLN, we have

$$\bar{X} \xrightarrow{\text{a.s.}} \mathbb{E}[X_1] = p.$$

Remark:

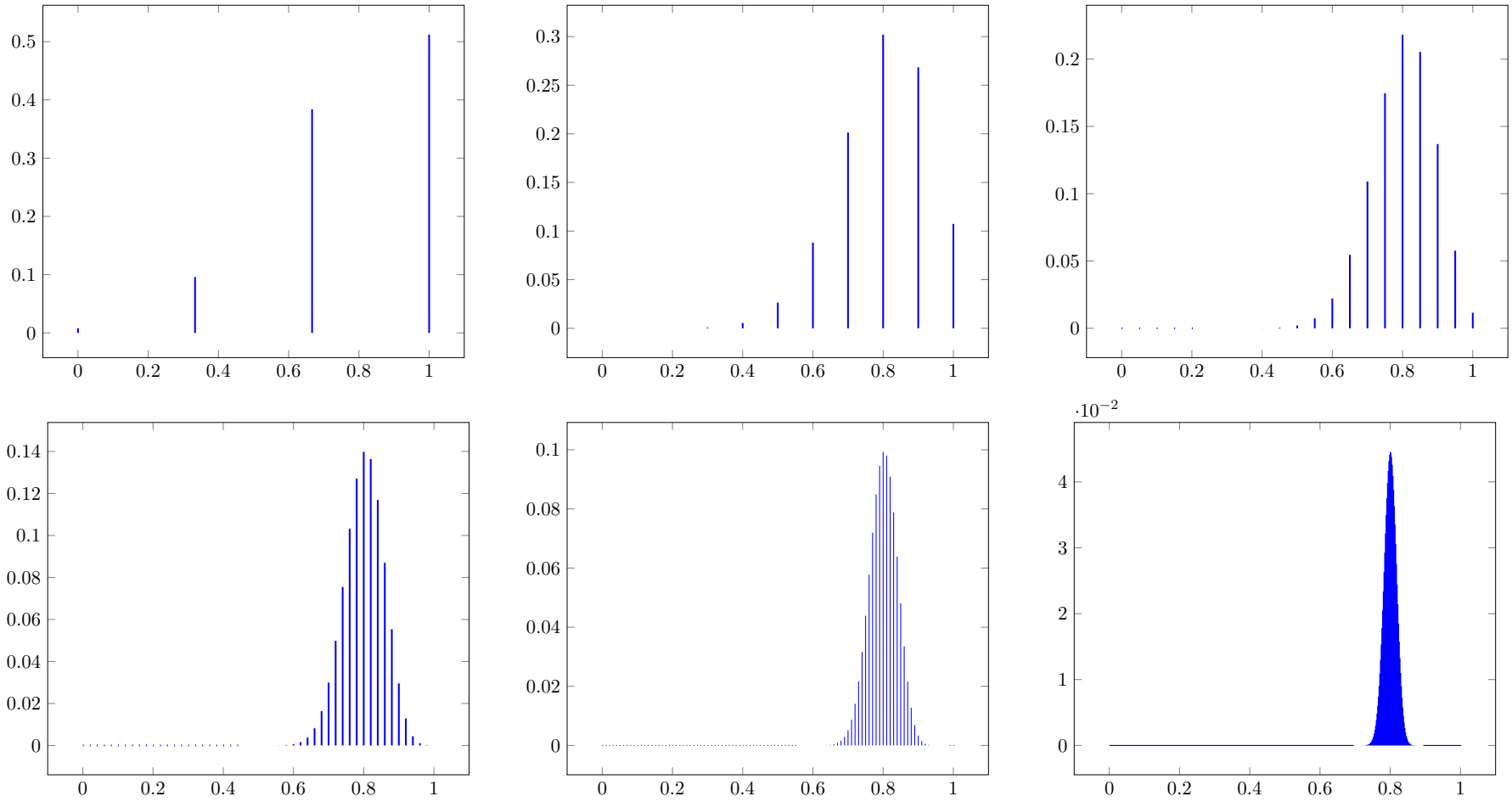
- $N := \sum_{i=1}^{n} X_i = n\bar{X}_n \sim \text{Bin}(n, p)$
- so $\bar{X}_n = \frac{N}{n}$ is a scaled Binomial r.v.

LLN for the Ber(0.5): pmf of $\bar{X}_n$ for $n \in \{3, 10, 20, 50, 100, 500\}$



We see that the distribution of $\bar{X}_n$ concentrates around $p = 0.5$.

LLN for the Ber(0.8): pmf of $\bar{X}_n$ for $n \in \{3, 10, 20, 50, 100, 500\}$



We see that the distribution of $\bar{X}_n$ concentrates around $p = 0.8$.

Convergence in distribution

Definition

Let $(X_n)_{n\geq 0}$ be a sequence of random variables,

- we say that $(X_n)_{n\geq 0}$ converges in distribution to X
- and we write $X_n \xrightarrow{(d)} X$

if, for each point $x \in \mathbb{R}$ where F_X is continuous, $\quad F_{X_n}(x) \underset{n\to\infty}{\longrightarrow} F_X(x), \qquad \forall x \in \mathbb{R}.$

Equivalent definition

For any finite partition $a_0 = -\infty < a_1 < \ldots < a_K = \infty$,

"the histograms of X_n converge to the histograms of X"

in the sense that for any a_{k-1}, a_k where F_X is continuous,

$$\mathbb{P}(X_n \in [a_{k-1}, a_k]) \underset{n\to\infty}{\longrightarrow} \mathbb{P}(X \in [a_{k-1}, a_k]).$$

Central Limit Theorem (CLT)

Theorem

If $X_1, \ldots, X_n$ are i.i.d. with $\mathbb{E}[f(X_1)^2] < \infty$, then,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big(f(X_i) - \mu_f\big) \xrightarrow{(d)} \mathcal{N}(0, \sigma_f^2) \quad \text{with} \quad \mu_f = \mathbb{E}[f(X_1)],\ \sigma_f^2 = \mathsf{Var}(f(X_1)).$$

where $\xrightarrow{(d)}$ is the *convergence in distribution.*

In particular,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{(d)} \mathcal{N}(0, 1) \quad \text{with} \quad \mu = \mathbb{E}[X_1],\ \sigma^2 = \mathsf{Var}(X_1).$$

CLT for the empirical mean of i.i.d. Bernoullis $Ber(p)$

We consider again

- a sample $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Ber}(p)$.
- $\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ the fraction of the throws where heads was observed.

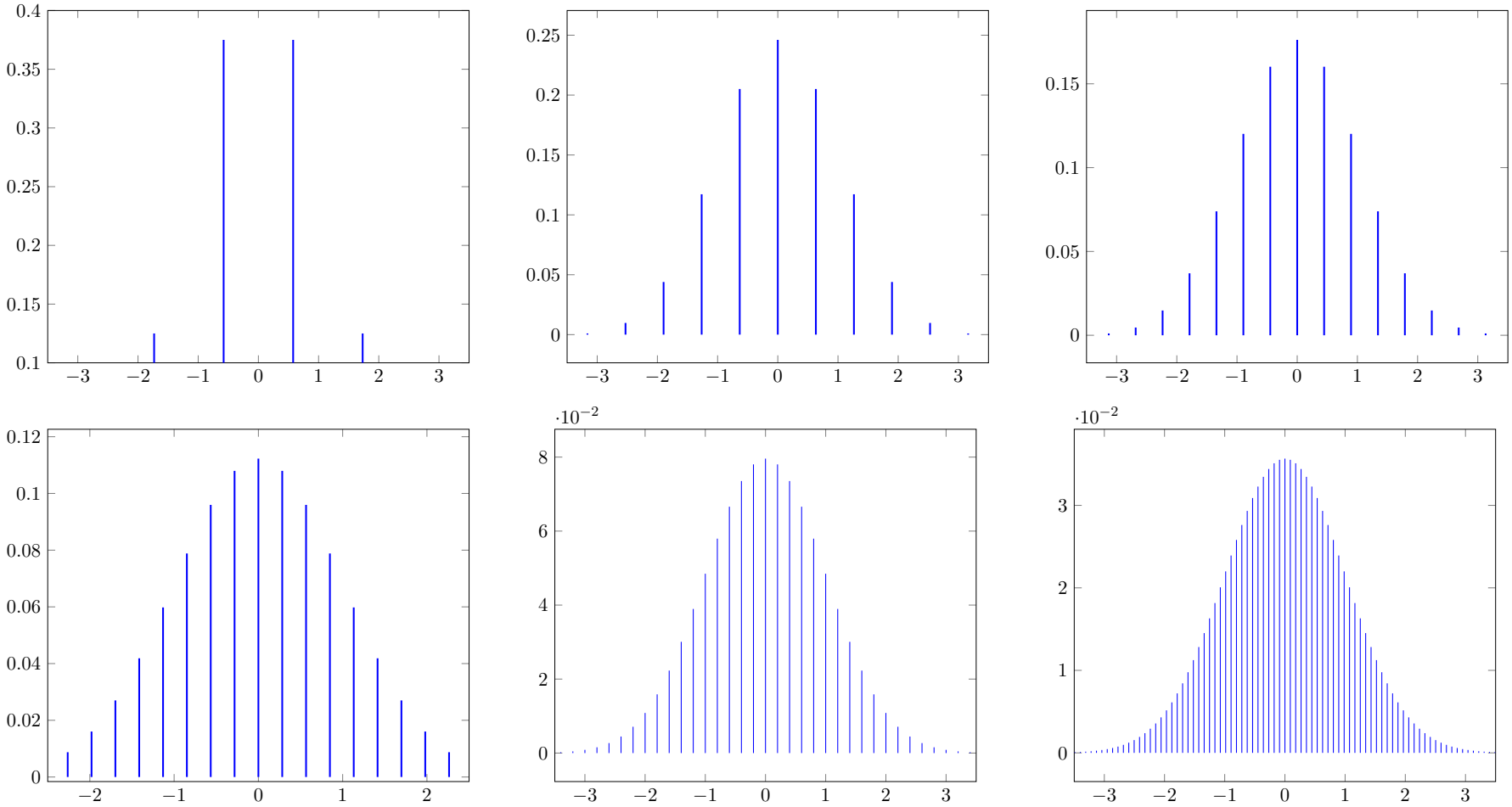
Since $\mathbb{E}[X_1^2] = p < \infty$, by the CLT, we have

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{(d)} \mathcal{N}(0,1) \quad \text{with} \quad \mu = \mathbb{E}[X_1] = p, \quad \sigma^2 = \text{Var}(X_1) = p(1-p).$$

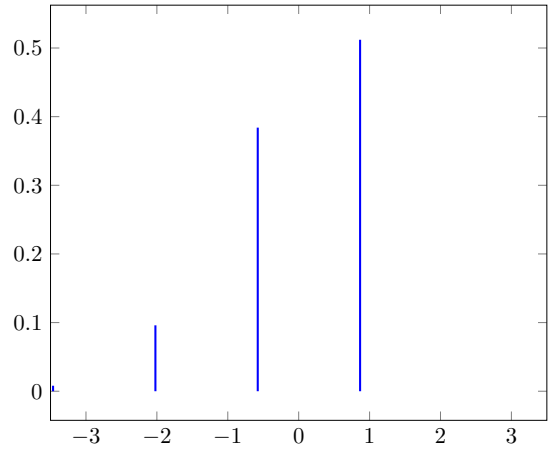
in other words

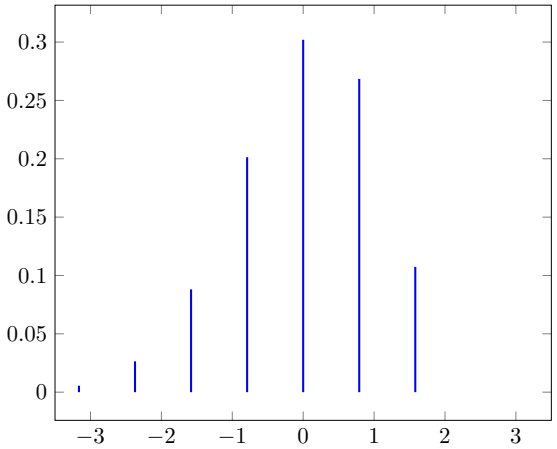
$$\frac{\sqrt{n}\,(\bar{X}_n - p)}{\sqrt{p(1-p)}} \xrightarrow{(d)} \mathcal{N}(0,1)$$

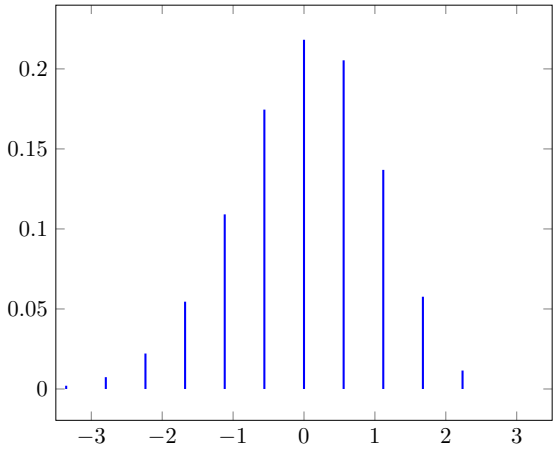
CLT for the Ber(0.5): pmf of $\sqrt{\frac{n}{p(1-p)}}(\bar{X}_n - p)$ for $N \in \{3, 10, 20, 50, 100, 500\}$

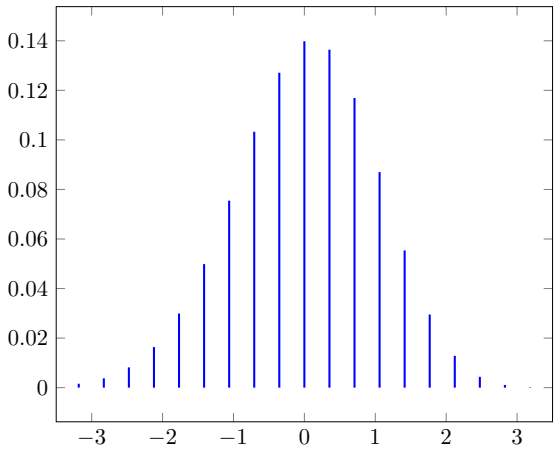


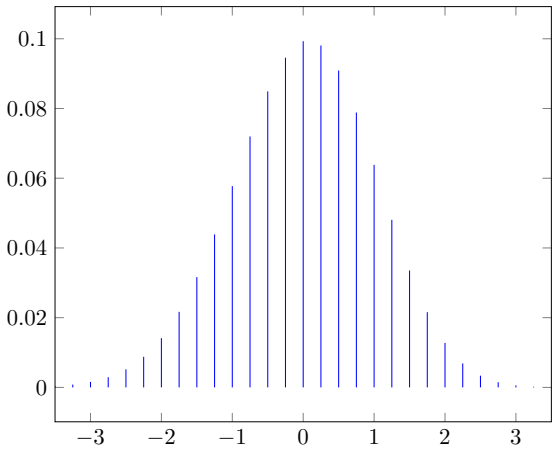
CLT for the Ber(0.8): pmf of $\sqrt{\frac{n}{p(1-p)}}(\bar{X}_n - p)$ for $n \in \{3, 10, 20, 50, 100, 500\}$

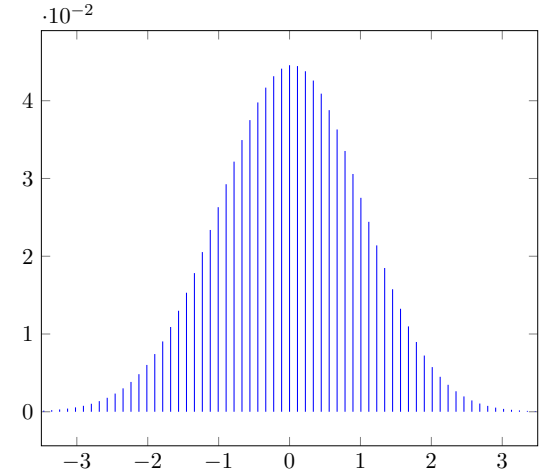












Example 2: LLN and CLT for the empirical mean of i.i.d. $\mathcal{U}[0,1]$ r.v.s.

We consider

- a sample $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{U}[0,1]$.
- $\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ the empirical mean.
- we have $\mathbb{E}[|X_1|] = \mathbb{E}[X_1] = \frac{1}{2}$ and $\text{Var}(X_1) = \frac{1}{12} < \infty$.

Since $\mathbb{E}[|X_1|] < \infty$, by the SLLN, we have

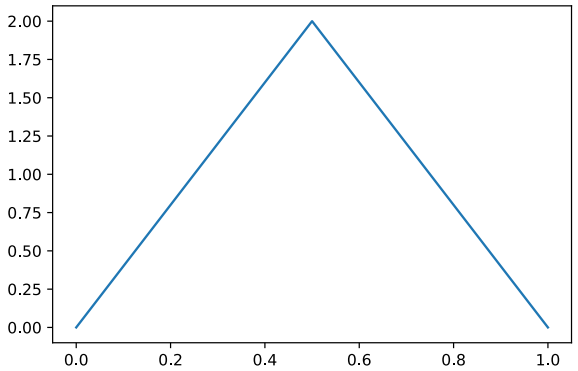
$$\bar{X} \xrightarrow{\text{a.s.}} \mathbb{E}[X_1] = \frac{1}{2}.$$

Since the mean and variance are finite, we have $\mathbb{E}[X_1^2] < \infty$, and by the CLT, we have

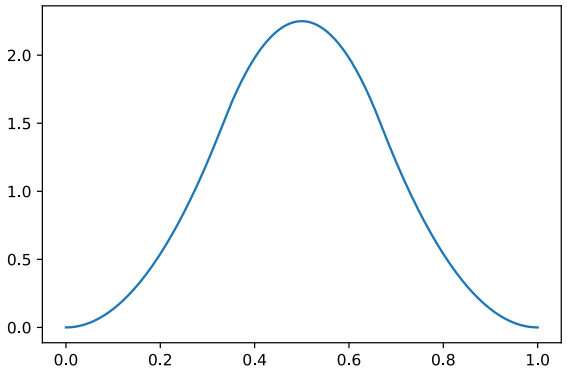
$$\frac{\sqrt{n}\,(\bar{X}_n - 0.5)}{\sqrt{1/12}} \xrightarrow{(d)} \mathcal{N}(0,1).$$

LLN for means of Uniforms $\bar{X}_n$ with $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{U}[0,1]$

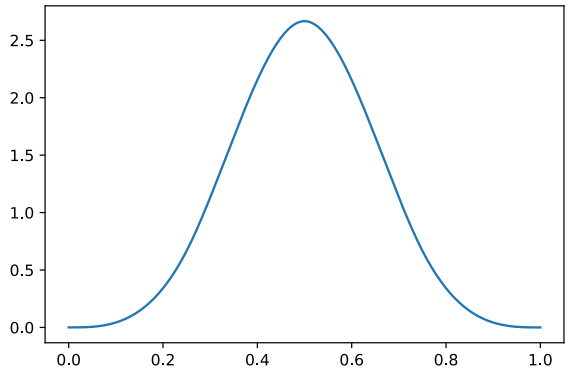
Probability density functions $p_{\bar{X}_n}$ of $\bar{X}_n$:



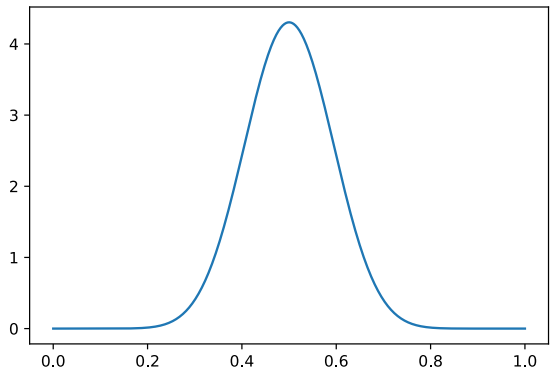
$n = 2$



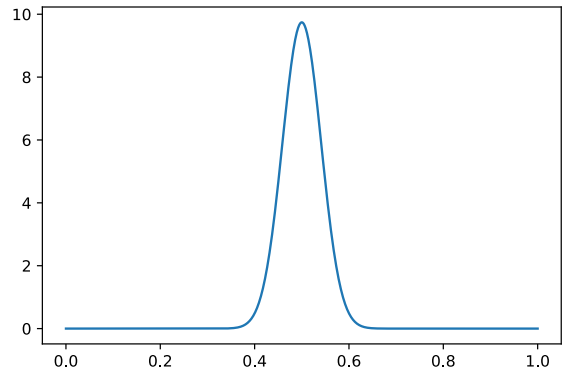
$n = 3$



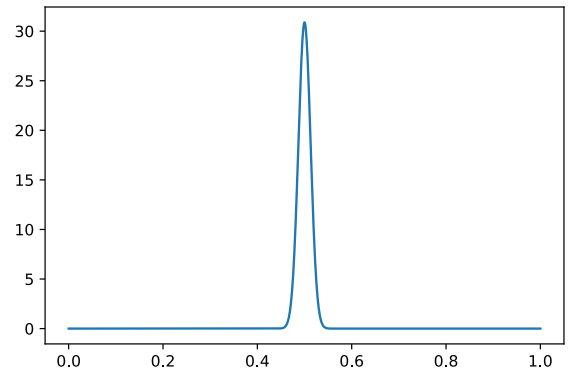
$n = 4$



$n = 10$



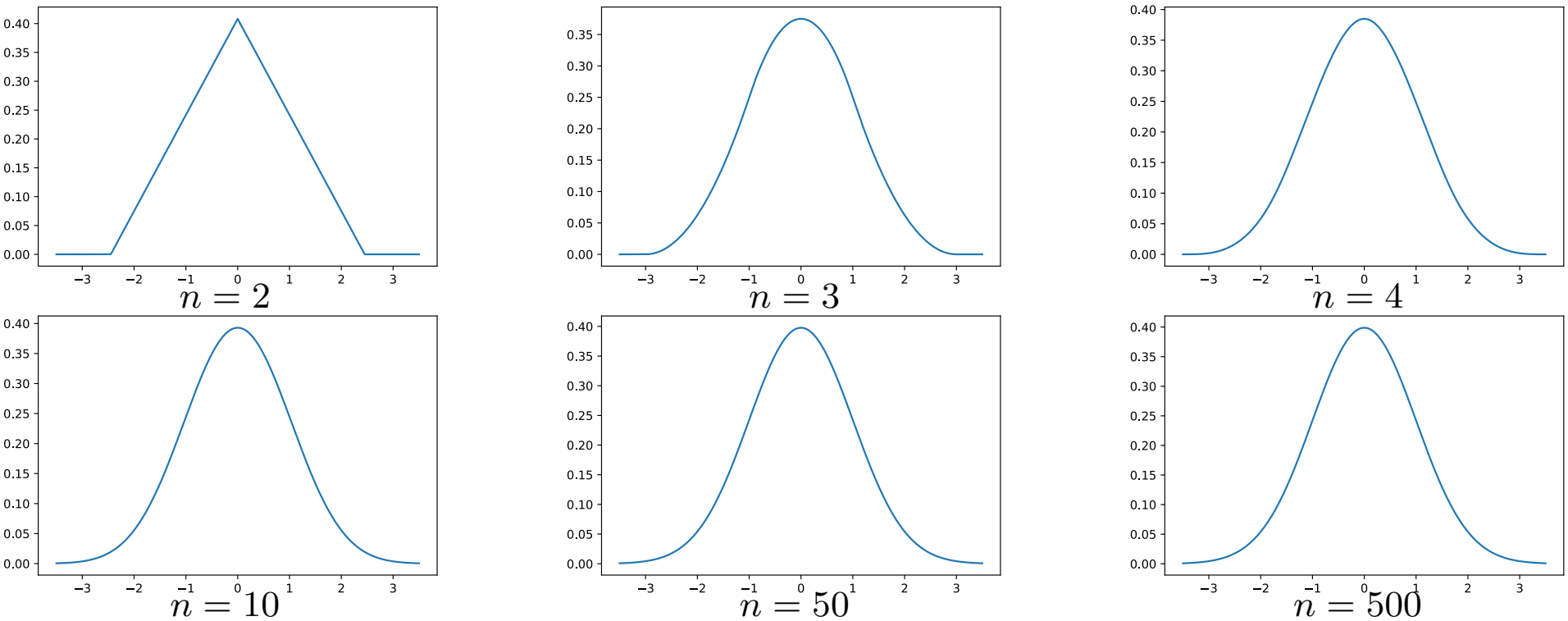
$n = 50$



$n = 500$

CLT$^+$ for *standardized* means $\sqrt{12n}(\bar{X}_n - 0.5)$ with $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{U}[0,1]$

Probability density functions $p_{\sqrt{12n}(\bar{X}_n - 0.5)}$ of $\sqrt{12n}(\bar{X}_n - 0.5)$:



Actually, the result seen here is stronger than the CLT because the *pdfs* of $\sqrt{12n}(\bar{X}_n - 0.5)$ become Gaussian (and not only the cdfs).

CLT combined with Slutsky's lemma for the case $\hat{\sigma} \xrightarrow{\mathbb{P}} \sigma$.

We will often use the CLT to know how close $\bar{X}_n$ is from $\mu := \mathbb{E}[X_1]$, but this depends on σ which is typically unknown...

Fortunately, the CLT is still valid if we have an estimate : $\hat{\sigma}$ of σ which converges to it.

Theorem

If $\hat{\sigma} \xrightarrow{\mathbb{P}} \sigma$, then $\dfrac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} \xrightarrow{(d)} \mathcal{N}(0,1)$ with $\mu = \mathbb{E}[X_1]$, $\sigma^2 = \mathrm{Var}(X_1)$.

This is guaranteed by a theoretical result called Slutsky's lemma beyond the scope of the course.

Relationship between the TCL and the LLN

- We always have $\mathbb{E}[X^2] \geq \mathbb{E}[|X|]^2$ so that if $\mathbb{E}[X^2] < \infty$ then $\mathbb{E}[|X|] < \infty$ as well.
- So if the conditions to apply the TCL are met then the SLLN applies as well.

Central Limit Theorem: multivariate version

We consider now r.v. $X_i = (X_{i1}, \ldots, X_{id})^\top$ taking values in $\mathbb{R}^d$.

Theorem

If $X_1, \ldots, X_n$ are i.i.d. with $\mathbb{E}[\|X_1\|^2] < \infty$, then,

$$\sqrt{n}(\bar{X} - \boldsymbol{\mu}) \xrightarrow{(d)} \mathcal{N}(0, \boldsymbol{\Sigma}),$$

with $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^\top$, and $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ the covariance matrix of X_1 with entry

$$\boldsymbol{\Sigma}_{jk} = \mathrm{cov}(X_{1j}, X_{1k}) = \mathbb{E}[(X_{1j} - \mu_j)(X_{1k} - \mu_k).]$$

Continuous mapping theorem

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}^k$ is a *continuous function*

$$\text{if} \quad Y_n \xrightarrow{\text{a.s.}} Y \quad \text{then} \quad f(Y_n) \xrightarrow{\text{a.s.}} f(Y)$$

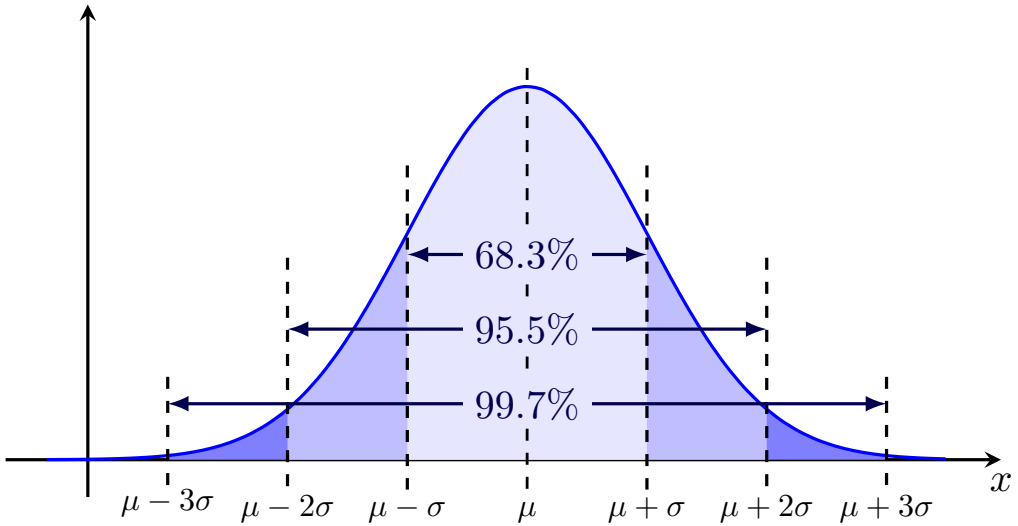
$$\text{if} \quad Y_n \xrightarrow{\mathbb{P}} Y \quad \text{then} \quad f(Y_n) \xrightarrow{\mathbb{P}} f(Y)$$

$$\text{if} \quad Y_n \xrightarrow{(d)} Y \quad \text{then} \quad f(Y_n) \xrightarrow{(d)} f(Y).$$

Confidence Intervals

EE-209 - Eléments de Statistiques pour les Data Sciences

The 68.3 - 95.5 - 99.7 rule

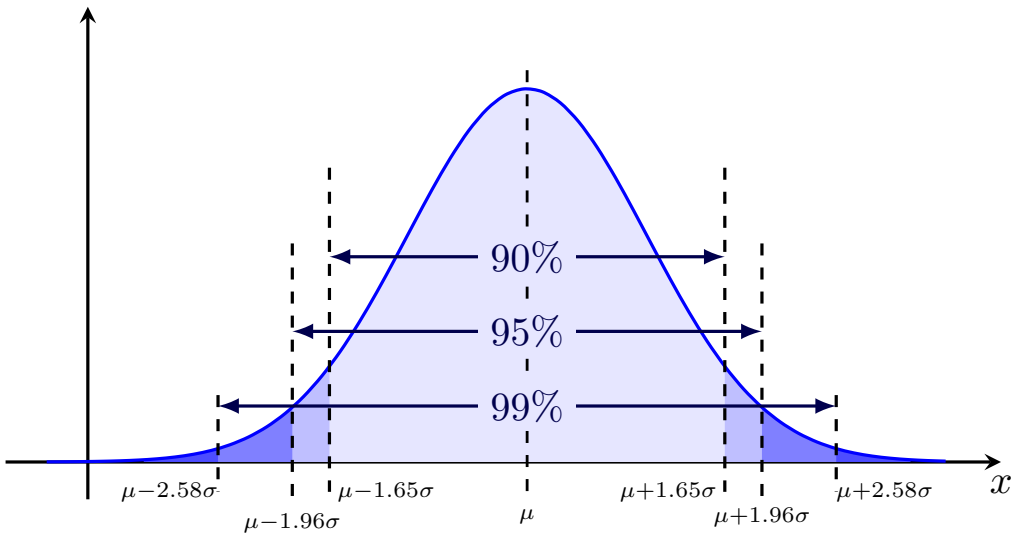


For $X \sim \mathcal{N}(\mu, \sigma^2)$

$$\mathbb{P}(X \in [\mu - \sigma, \mu + \sigma]) \approx 0.683$$
$$\mathbb{P}(X \in [\mu - 2\sigma, \mu + 2\sigma]) \approx 0.955$$
$$\mathbb{P}(X \in [\mu - 3\sigma, \mu + 3\sigma]) \approx 0.997$$

We see that the probability that a Gaussian random variable takes a value which is further away from the expectation μ than 3σ (even 2 σ) is fairly small.

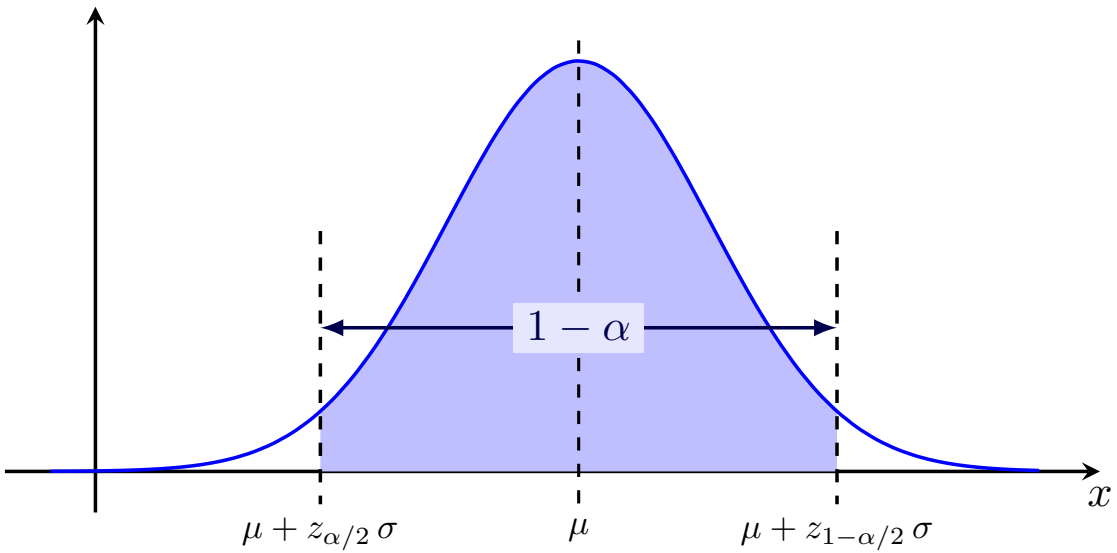
Intervals with guarantees at 90%, 95% and 99%



For $X \sim \mathcal{N}(\mu, \sigma^2)$

$$\mathbb{P}(X \in [\mu - 1.645\,\sigma, \mu + 1.645\,\sigma]) \approx 0.90$$
$$\mathbb{P}(X \in [\mu - 1.960\,\sigma, \mu + 1.960\,\sigma]) \approx 0.95$$
$$\mathbb{P}(X \in [\mu - 2.576\,\sigma, \mu + 2.576\,\sigma]) \approx 0.99$$
$$\mathbb{P}(X \in [\mu - 3.291\,\sigma, \mu + 3.291\,\sigma]) \approx 0.999$$

High probability interval for a single Gaussian observation $X \sim \mathcal{N}(\mu, \sigma^2)$



$$\begin{aligned} & \mathbb{P}(\quad X \quad \in [\mu - q\,\sigma, \mu + q\,\sigma]) \\ = \; & \mathbb{P}(X - \mu \in \quad [\,-q\,\sigma,\; q\,\sigma\,] \quad) \\ = \; & \mathbb{P}\Big(\tfrac{X-\mu}{\sigma} \in \quad [\,-q,\; q\,] \quad\Big) \\ = \; & \mathbb{P}(\quad Z \quad \in \quad [\,-q,\; q\,] \quad), \end{aligned}$$

where $Z \sim \mathcal{N}(0,1)$.

And $\mathbb{P}(Z \in [-q, q]) = \Phi(q) - \Phi(-q) = 1 - 2\Phi(-q)$ where Φ is the standard Gaussian cdf.

So $\mathbb{P}(Z \in [-q, q]) = 1-\alpha \quad \Leftrightarrow \quad -q = z_{\alpha/2} \quad \Leftrightarrow \quad q = z_{1-\alpha/2} = |z_{\alpha/2}|$.

We have $\quad \mathbb{P}\big(X \in [\mu - z_{1-\alpha/2}\,\sigma, \mu + z_{1-\alpha/2}\,\sigma]\big) = 1-\alpha$

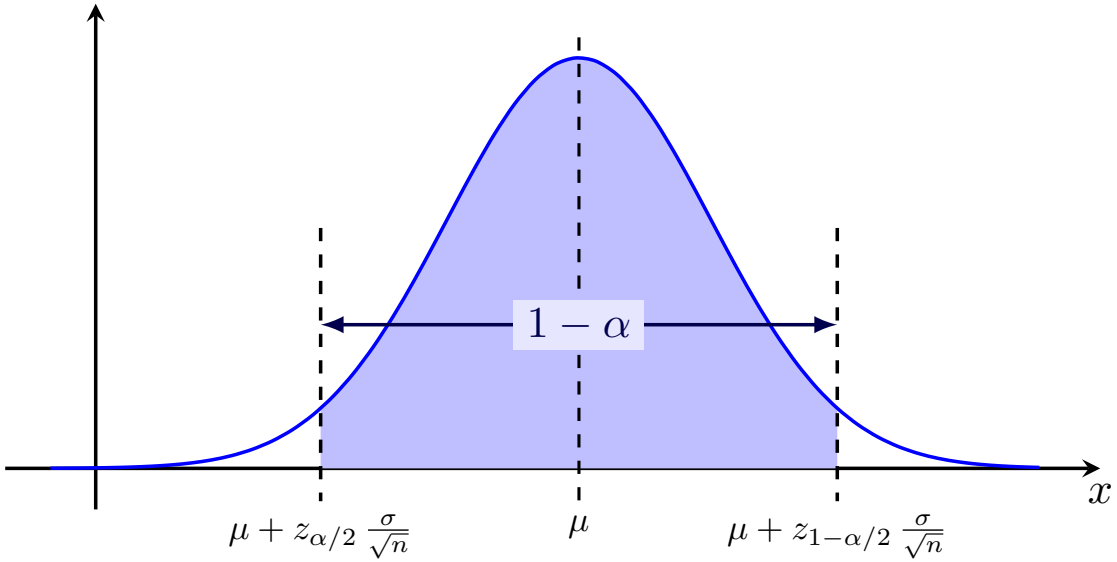
or equivalently $\quad \mathbb{P}\big(X \in [\mu - |z_{\alpha/2}|\,\sigma, \mu + |z_{\alpha/2}|\,\sigma]\big) = 1-\alpha$

High probability interval for the empirical mean of i.i.d. Gaussian data

$$\begin{aligned} \text{If} \quad X_1, \ldots, X_n &\overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2) \\ \text{then} \quad X_1 + \ldots + X_n &\sim \mathcal{N}(n\mu, n\sigma^2) \\ \text{so that} \quad \bar{X} &\sim \mathcal{N}(\mu, \tfrac{\sigma^2}{n}) \end{aligned}$$

where $\bar{X} := \frac{1}{n}(X_1 + \ldots + X_n)$.

$\text{std}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ is called the *standard error*.



$$\begin{aligned} \text{So we have} \quad & \mathbb{P}\big(\bar{X} \in \big[\mu - z_{1-\alpha/2}\,\tfrac{\sigma}{\sqrt{n}}, \mu + z_{1-\alpha/2}\,\tfrac{\sigma}{\sqrt{n}}\big]\big) = 1-\alpha \\ \text{or equivalently} \quad & \mathbb{P}\big(\bar{X} \in \big[\mu - |z_{\alpha/2}|\,\tfrac{\sigma}{\sqrt{n}}, \mu + |z_{\alpha/2}|\,\tfrac{\sigma}{\sqrt{n}}\big]\big) = 1-\alpha \\ \text{e.g.} \quad & \mathbb{P}\big(\bar{X} \in \big[\mu - 1.96\,\tfrac{\sigma}{\sqrt{n}}, \mu + 1.96\,\tfrac{\sigma}{\sqrt{n}}\big]\big) = 0.95 \end{aligned}$$

Confidence interval: key idea

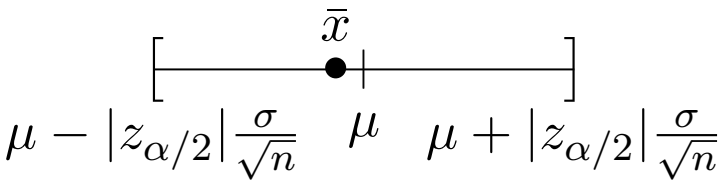
$$\mathbb{P}(\bar{X} \in [\mu - c, \mu + c]) = \mathbb{P}(|\bar{X} - \mu| \leq c) = \mathbb{P}(\mu \in [\bar{X} - c, \bar{X} + c])$$

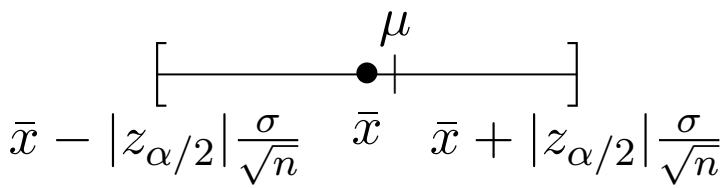
Indeed $\bar{X} \leq \mu + c \quad \Leftrightarrow \quad \mu \geq \bar{X} - c.$ And $\mu - c \leq \bar{X} \quad \Leftrightarrow \quad \mu \leq \bar{X} + c.$

So we have $\mathbb{P}\left(\mu \in \left[\bar{X} - z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}}\right]\right) = 1 - \alpha$

or equivalently $\mathbb{P}\left(\mu \in \left[\bar{X} - |z_{\alpha/2}|\,\frac{\sigma}{\sqrt{n}}, \bar{X} + |z_{\alpha/2}|\,\frac{\sigma}{\sqrt{n}}\right]\right) = 1 - \alpha$

e.g. $\mathbb{P}\left(\mu \in \left[\bar{X} - 1.96\,\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\,\frac{\sigma}{\sqrt{n}}\right]\right) = 0.95$





- $\left[\bar{X} - |z_{\alpha/2}|\,\frac{\sigma}{\sqrt{n}}, \bar{X} + |z_{\alpha/2}|\,\frac{\sigma}{\sqrt{n}}\right]$ is a $1 - \alpha$ level *confidence interval.*
- $\left[\bar{X} - 1.96\,\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\,\frac{\sigma}{\sqrt{n}}\right]$ is a 95% level *confidence interval.*

It is the interval which is random, not μ !

Example: estimating a temperature

Let's assume that we are trying to measure a temperature (e.g. below a glacier) with a device which is fairly unstable. We assume that the standard deviation of the measurement error is known and equal to $\sigma = 0.6$.

We collect the following list of measured values:

$$[-0.1,\ 0.3, -0.8,\ 0.0,\ 0.1, -0.8,\ 0.7, -1.9, -0.9, -0.3]$$

We have $n = 10$ and $\bar{x} = -0.37$.
And we get the following 95% confidence interval:

$$\left[\bar{x} - 1.96\ \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\ \frac{\sigma}{\sqrt{n}}\right] = \left[-0.37 - 1.96 \cdot \frac{0.6}{\sqrt{10}}, -0.37 + 1.96 \cdot \frac{0.6}{\sqrt{10}}\right] = [-0.74, 0.00].$$

Note that in this example σ was known which is rarely the case

The case of the Gaussian empirical mean $\bar{X}$ when σ is unknown

We assume again that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ so that $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$.

We can estimate σ^2 using the *unbiased variance estimate* $S^2 := \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$.

The way we proceeded before was using the fact that $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$...

But, given that S^2 is a random variable, a priori, we cannot simply replace σ^2 by S^2 in the previous equation.However we have the following result:

Theorem: the appropriately standardized mean follows a Student distribution...

Under the assumption that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$:

(i) $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$, (ii) $S^2/\sigma^2 \sim \frac{1}{n-1} \chi^2_{n-1}$, (iii) $\bar{X}$ and S^2 are independent r.v.s

and $T := \dfrac{\bar{X} - \mu}{S/\sqrt{n}}$ follows Student's t-distribution St_{n-1} with $n-1$ degrees of freedom.

The Student distribution

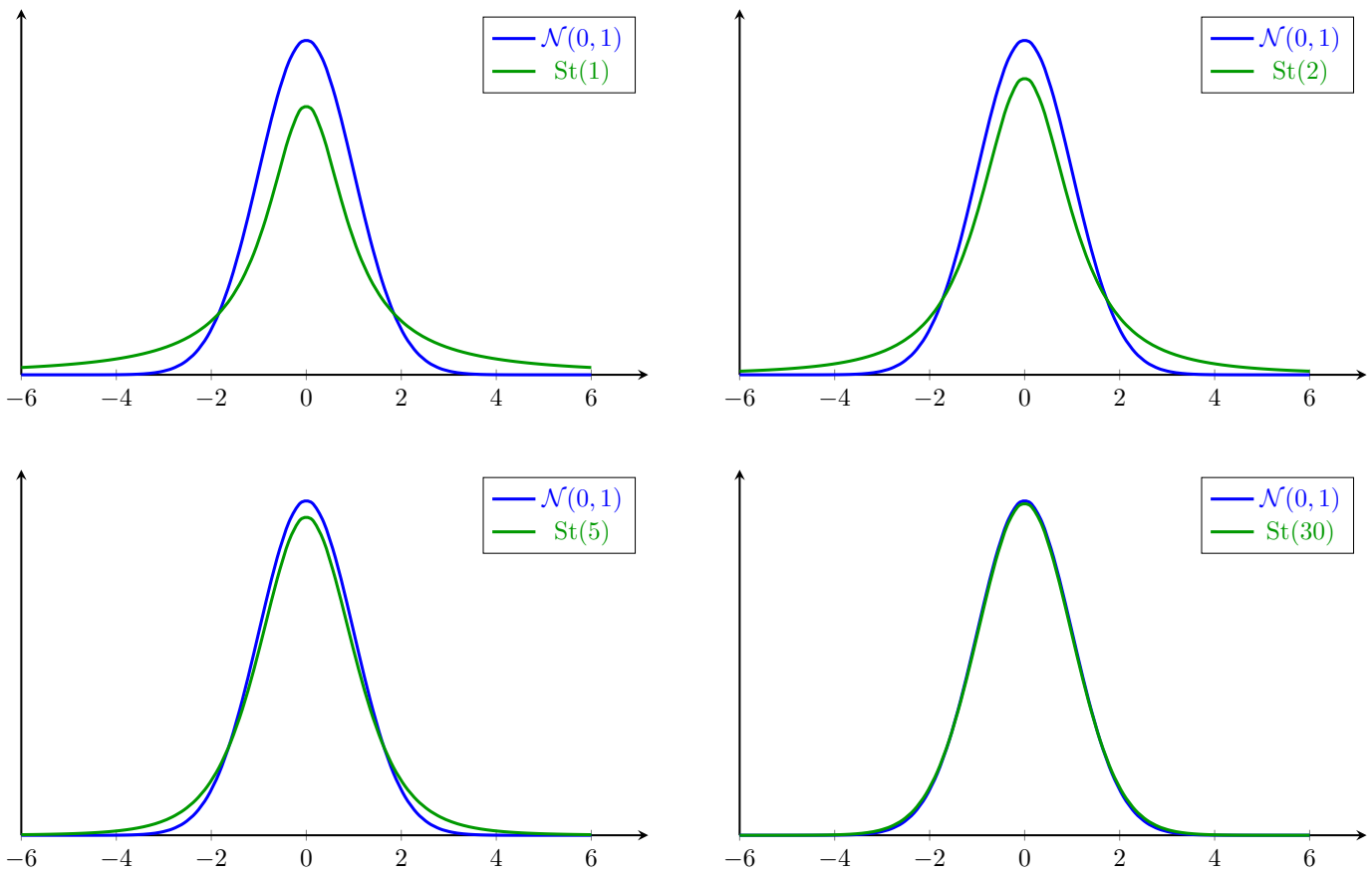
If $T \sim \mathrm{St}_n$ then its pdf is

$$p_T(t) = c_n \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}},$$

where c_n is a normalizing constant.

Away from the mean, the Gaussian density decreases *faster than exponentially*, while the Student t density decreases only *polynomially*.

Intervals containing 95% of the probability mass are thus wider for the Student.



Comparing the Student pdfs with $n = 1, 2, 5, 30$ d.f.s with a $\mathcal{N}(0,1)$.

Confidence Interval for the Gaussian $\bar{X}$ using the Student t-distribution

Given that $T := \dfrac{\bar{X} - \mu}{S/\sqrt{n}} \sim \mathrm{St}_{n-1}$, if

- $t_{\alpha/2}^{(n-1)}$ is the quantile of level $\frac{\alpha}{2}$ of a St_{n-1},
- $\tau_{\frac{\alpha}{2}} := t_{1-\alpha/2}^{(n-1)} = |t_{\alpha/2}^{(n-1)}|$ is the quantile of level $1 - \frac{\alpha}{2}$ of a St_{n-1},

$$
\begin{aligned}
\text{then} \qquad 1-\alpha &= \mathbb{P}(-\tau_{\frac{\alpha}{2}} \le T \le \tau_{\frac{\alpha}{2}}) \\
&= \mathbb{P}(-\tau_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \le \bar{X} - \mu \le \tau_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}) \\
&= \mathbb{P}(-\tau_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \le \mu - \bar{X} \le \tau_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}) \\
&= \mathbb{P}(\bar{X} - \tau_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \le \mu \le \bar{X} + \tau_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}})
\end{aligned}
$$

So $\left[\bar{X} - \tau_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \bar{X} + \tau_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right]$ is a confidence interval of level $1 - \alpha$

Comparing $t^{(n)}_{1-\alpha/2}$ and $z_{1-\alpha/2}$

	Values of $\tau_{\alpha/2}$ for $n =$								$\lvert z_{\alpha/2}\rvert$
$1-\alpha$	1	2	5	10	20	50	100	∞	
0.90	6.31	2.92	2.02	1.81	1.73	1.67	1.66	1.645	1.645
0.95	12.7	4.30	2.57	2.23	2.09	2.01	1.98	1.960	1.960
0.99	63.6	9.92	4.03	3.17	2.85	2.68	2.62	2.576	2.576

- For $n = 6$, $\left[\bar{X} - 2.57\frac{S}{\sqrt{n}}, \bar{X} + 2.57\frac{S}{\sqrt{n}}\right]$ is a confidence interval of level 95% for μ.
- For $n = 51$, $\left[\bar{X} - 2.01\frac{S}{\sqrt{n}}, \bar{X} + 2.01\frac{S}{\sqrt{n}}\right]$ is a confidence interval of level 95% for μ.

When $n \to \infty$ we have $t^{(n)}_{1-\alpha/2} \longrightarrow z_{1-\alpha/2} = |z_{\alpha/2}|$.

Example: estimating a temperature again but with σ unknown

We are still trying to measure a temperature (e.g. below a glacier) with a device which is fairly unstable, but now σ is unknown.
We have the same list of measured values:

$$[-0.1,\ 0.3, -0.8,\ 0.0,\ 0.1, -0.8,\ 0.7, -1.9, -0.9, -0.3]$$

and we still have $n = 10$ and $\bar{x} = -0.37$.

We compute the sample standard deviation $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = 0.747$

And we get the following 95% Student confidence interval using that $t^{(9)}_{0.975} = 2.26$.

$$\left[\bar{x} - \tau_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \bar{x} + \tau_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right] = \left[-0.37 - 2.26 \cdot \frac{0.747}{\sqrt{10}},\ -0.37 + 2.26 \cdot \frac{0.747}{\sqrt{10}}\right] = [-0.90,\ 0.16].$$

The confidence interval is larger because $s > \sigma$ and because $t_{1-\alpha/2} > z_{1-\alpha/2}$.

Asymptotic Confidence Intervals

- What if $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P$ with P unknown?
- Can we still determine a confidence interval for $\mu = \mathbb{E}[X_1]$?
- If we assume that $\mathbb{E}[X_1^2] < \infty$, then by the CLT, $\dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{(d)} \mathcal{N}(0,1)$.
- Even better, if $\mathbb{E}[X_1^2] < \infty$, by the CLT + Slutsky's lemma, $\dfrac{\bar{X} - \mu}{S/\sqrt{n}} \xrightarrow{(d)} \mathcal{N}(0,1)$.

then if n is large

$$
\begin{aligned}
1-\alpha \quad &\approx \quad \mathbb{P}\left(-|z_{\alpha/2}| \leq \tfrac{\bar{X}-\mu}{S/\sqrt{n}} \leq |z_{\alpha/2}|\right) \\
&= \mathbb{P}\left(-|z_{\alpha/2}|\tfrac{S}{\sqrt{n}} \leq \bar{X} - \mu \leq |z_{\alpha/2}|\tfrac{S}{\sqrt{n}}\right) \\
&= \mathbb{P}\left(-|z_{\alpha/2}|\tfrac{S}{\sqrt{n}} \leq \mu - \bar{X} \leq |z_{\alpha/2}|\tfrac{S}{\sqrt{n}}\right) \\
&= \mathbb{P}\left(\bar{X} - |z_{\alpha/2}|\tfrac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + |z_{\alpha/2}|\tfrac{S}{\sqrt{n}}\right)
\end{aligned}
$$

So $\left[\bar{X} - \tau_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}, \bar{X} + \tau_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right]$ is a *approximate* confidence interval of level $1-\alpha$ when n is sufficiently large. This is called an asymptotic CI, and $1-\alpha$ is its *nominal probability coverage*. Note that we used S here but any *consistent* estimator of σ could be used.

Asymptotic Confidence Intervals: Application to the Bernoulli

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Ber}(p)$ with p unknown, and we consider the estimator $\hat{p} := \bar{X}$.

- Given that $\mathbb{E}[X_1^2] = \mathbb{E}[X_1] = p < \infty$ and that $\text{Var}(X) = p(1-p)$, by the CLT

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \xrightarrow{(d)} \mathcal{N}(0,1).$$

- $p(1-p)$ is unknown but can be estimated consistently by $\hat{p}(1-\hat{p})$, and by Slutsky

$$\frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})}/\sqrt{n}} \xrightarrow{(d)} \mathcal{N}(0,1).$$

We therefore have the asymptotic CI at level $1-\alpha$ *(nominal probability coverage)*

$$p \in \left[\hat{p} - |z_{\alpha/2}| \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, \hat{p} + |z_{\alpha/2}| \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \right].$$

Wald confidence intervals for the Maximum Likelihood Estimator

By the CLT, if $\hat{\theta} = \hat{\theta}_{\text{MLE}}$ is the maximum likelihood estimator for θ based on an i.i.d. sample of size n, and if $I_1(\theta) > 0$, then

$$\sqrt{nI_1(\theta)}(\hat{\theta} - \theta) = \sqrt{I(\theta)}(\hat{\theta} - \theta) \xrightarrow{(d)} \mathcal{N}(0, 1).$$

With Slutsky's lemma, we also have $\sqrt{I(\hat{\theta})}(\hat{\theta} - \theta) \xrightarrow{(d)} \mathcal{N}(0, 1).$

So we have

$$\begin{aligned} 1 - \alpha &\approx \mathbb{P}(-|z_{\alpha/2}| \leq \sqrt{I(\hat{\theta})}(\hat{\theta} - \theta) \leq |z_{\alpha/2}|) \\ &= \mathbb{P}\left(\hat{\theta} - \frac{|z_{\alpha/2}|}{\sqrt{I(\hat{\theta})}} \leq \theta \leq \hat{\theta} + \frac{|z_{\alpha/2}|}{\sqrt{I(\hat{\theta})}}\right) \end{aligned}$$

Finally, $\left[\hat{\theta} - \frac{|z_{\alpha/2}|}{\sqrt{I(\hat{\theta})}}, \hat{\theta} + \frac{|z_{\alpha/2}|}{\sqrt{I(\hat{\theta})}}\right]$ is an asymptotic confidence interval of level $1 - \alpha$ which is thus valid when n is large.

Wald CI for the MLE of the parameter p in the Bernoulli model

Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Ber}(p_*)$. If $N = \sum_{i=1}^n X_i = n\bar{X}$. then

- the log-likelihood is $\ell(p) = N \log p + (n - N) \log(1 - p)$.
- the score function is $\ell'(p) = \frac{N}{p} - \frac{n-N}{1-p} = \frac{(1-p)N - p(n-N)}{p(1-p)} = \frac{N - pn}{p(1-p)}$.
- the stationary points of ℓ satisfy $\ell'(p) = 0$. The unique solution is $\hat{p} = \frac{N}{n} = \bar{X}$.
- $\ell'(p) > 0$ for $p < \hat{p}$ and $\ell'(p) < 0$ for $p > \hat{p}$ so $\hat{p}$ attains the maximum and is the MLE.
- the Fisher Information is $I(p) = \text{Var}\big(\ell'(p)\big) = \frac{\text{Var}(N)}{p^2(1-p)^2} = \frac{np(1-p)}{p^2(1-p)^2} = \frac{n}{p(1-p)}$.
- It can be estimated by the observed information $I(\hat{p}) = \frac{n}{\hat{p}(1-\hat{p})}$.

Using the definition of the Wald confidence interval we have $p \in \left[\hat{p} - \frac{|z_{\alpha/2}|}{\sqrt{I(\hat{p})}},\ \hat{\theta} + \frac{|z_{\alpha/2}|}{\sqrt{I(\hat{p})}}\right]$

After replacement $p \in \left[\hat{p} - |z_{\alpha/2}| \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + |z_{\alpha/2}| \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$.This is the same asymptotic confidence interval as the one we had obtained from applying the general CLT.

Confidence Interval: general view

In the case of a scalar parameter θ

- instead of looking for a *pointwise estimator* $\hat{\theta}$ which aims at being close to θ
- we try to find a (short) interval $[\hat{\Theta}_l, \hat{\Theta}_u]$ such that

$$\mathbb{P}\big(\theta \in [\hat{\Theta}_l, \hat{\Theta}_u]\big) \geq 1 - \alpha.$$

- This interval is often of the form $[\hat{\theta} - m, \hat{\theta} + m]$ where m is the *margin of error* (MOE).
- Often, $m = q\frac{\sigma}{\sqrt{n}}$ or $m = q\frac{\hat{\sigma}}{\sqrt{n}}$ where q is the quantile of a distribution that does not depend on any (unknown) parameter, where σ is the standard deviation of a single observation and $\frac{\sigma}{\sqrt{n}}$ is called the *standard* error (SE).
- A confidence interval is a way to quantify our *uncertainty* about our estimate, and the MOE and SE are ways to measure it.
- When an approximate confidence interval is built it targets a level $1 - \alpha$, which is called the *nominal probability coverage*.
- It can be different from the actual value of $\mathbb{P}\big(\theta \in [\hat{\Theta}_l, \hat{\Theta}_u]\big)$, which is called the *actual probability coverage* and which is often unknown.

Pivots and how to construct confidence intervals

A **pivot** is a statistic $T(X_1, \ldots, X_n, \theta)$ that depends on the sample and on the parameter of interest in such a way that its distribution does not depend on θ.

For example:

- For $X_i \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, $\quad T := \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ is a pivot.
- For $X_i \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, $\quad T := \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim \text{St}_{n-1}$ is a pivot, for $S^2 = \frac{n}{n-1}(\overline{X^2} - \bar{X}^2)$.

We can also have some asymptotic pivots:

- For $X_i \ldots, X_n$ i.i.d. with $\mathbb{E}[X_1^2] < \infty$, $\quad T := \frac{\bar{X}-\mu}{\hat{\sigma}/\sqrt{n}} \xrightarrow{(d)} \mathcal{N}(0,1)$.
- For $\hat{\theta} = \hat{\theta}_{MLE}$, $\quad T := \sqrt{I(\hat{\theta})}(\hat{\theta} - \theta) \xrightarrow{(d)} \mathcal{N}(0,1)$.

Wilson score CI for the MLE of the parameter p in the Bernoulli model

We can also consider the central limit theorem based on the true variance $p(1-p)$:

$$\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \xrightarrow{(d)} \mathcal{N}(0,1) \quad \text{so that} \quad \frac{n(\hat{p}-p)^2}{p(1-p)} \xrightarrow{(d)} \chi_1^2.$$

Let $z = z_{1-\alpha/2}$ be an $1-\alpha/2$ normal quantile.Then z^2 is an $1-\alpha$ quantile of the χ_1^2.

Therefore

$$\mathbb{P}\Big(\frac{n(\hat{p}-p)^2}{p(1-p)} \leq z^2\Big) \underset{n\to\infty}{\longrightarrow} 1-\alpha.$$

The inequality can be rewritten as

$$np^2 - 2np\hat{p} + n\hat{p}^2 \leq z^2 p(1-p)$$

$$p^2(n+z^2) - 2p(n\hat{p} + \tfrac{1}{2}z^2) + n\hat{p}^2 \leq 0$$

Calculations show that this is equivalent to

$$p \in \Big[\hat{p}_z - \frac{\hat{\sigma}_z}{\sqrt{n}}\, z,\ \hat{p}_z + \frac{\hat{\sigma}_z}{\sqrt{n}}\, z\Big]$$

with $\hat{p}_z := \dfrac{n\hat{p} + z^2\frac{1}{2}}{n+z^2}$,

and $\hat{\sigma}_z := \dfrac{n}{n+z^2}\sqrt{\hat{p}(1-\hat{p}) + \dfrac{z^2}{4n}}$.

Uncertainty, confidence intervals and significant figures

EE-209 Eléments de Statistiques pour les Data Sciences

Significant figures

In data science, continuous quantities are not known to arbitrary precision. On the contrary any measurement has a level of uncertainty. This means that in the decimal expansion of a number, digits beyond the precision, should be dropped and the number should be rounded to the last significant figure.

For example, if the uncertainty is $0.000\underline{1}$ then $x = 0.0\underbrace{\underline{452}}_{\text{significant digits}}71298 \rightarrow x = 0.0453$

For example, if the uncertainly is $\underline{1} \cdot 10^2$ then $x = \underbrace{\underline{50}}_{\text{significant digits}}32 \rightarrow x = 5.0 \cdot 10^3$.

Note that in the examples above the uncertainty is specified with a **single** significant digit, which determines the level of precision.
So we removed digits beyond the precision level and rounded to the closest number at the same precision.

Margin of error and reporting numbers at an appropriate precision level

In statistical estimation the level of uncertainty u is determined rigorously by confidence intervals: for symmetric CIs, u is the *margin of error*.

If a CI is of the form $[\hat{\theta} - u, \hat{\theta} + u]$ then it is usually considered that

1. u should be reported with $k = 1$ or 2 (or at most 3) significant digits[1].
2. the precision of reported values for $\hat{\theta}$, $\hat{\theta} - u$ and $\hat{\theta} + u$ should be the same as for u.
3. the rounding should be done "outwards", i.e. down for $\hat{\theta} + u$, and up for $\hat{\theta} + u$.

For example, if $\hat{\theta} = 1.069871$ and $u = 0.002415$,we can calculate $\hat{\theta} - u = 1.067456$ and $\hat{\theta} + u = 1.072286$;if we keep 2 significant digits, then we report

$$u = 0.00\underline{24}, \quad \hat{\theta} = \underline{1.0699} \quad \text{and} \quad [\hat{\theta} - u, \hat{\theta} + u] = [\underline{1.0674}, \underline{1.0723}].$$

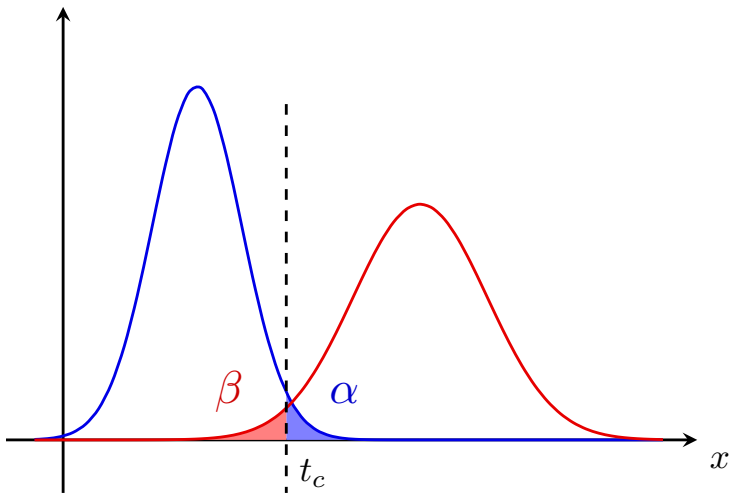
If the first significant digit in u is a $1, 2$ or 3 we usually keep more significant digits than if it is a larger number, because the relative error produced by rounding is larger.

[1] The **G**uide to the Expression of **U**ncertainty in **M**easurement (**GUM**) published by the Joint Committee for Guides in Metrology (JCGM) recommends to report uncertainty *"to at most two significant digits"* (7.2.6).

Hypothesis testing

EE209 - Eléments de Statistiques pour les Data Sciences

Telling apart two distributions based on an observation



Let's assume that X can follow two distributions:

- Under the *null hypothesis* $H_0 : X \sim \mathcal{N}(\mu_0, \sigma_0^2)$
- Under the *alternate hypothesis* $H_1 : X \sim \mathcal{N}(\mu_1, \sigma_1^2)$,

with $\mu_0 < \mu_1$ and σ_0, σ_1 not too large.

Can we try to decide based on an observation x of X which hypothesis is the correct one?

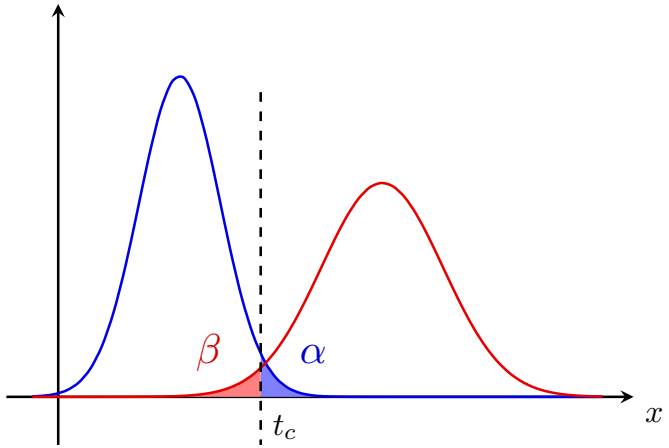
We can choose a *critical value* t_c on the value x, and

- if $x \leq t_c$, decide that H_0 is correct
- if $x > t_c$, decide that H_1 is correct

We can define

- $\alpha = \mathbb{P}_0(X > t_c)$ where $\mathbb{P}_0$ is "the probability **if** H_0 is true"
- $\beta = \mathbb{P}_1(X \leq t_c)$ where $\mathbb{P}_1$ is "the probability **if** H_1 is true"

Types of errors



With the same setting as on previous slide, let's denote by Δ our decision with

- $\Delta = 0$ if we decide that H_0 is correct
- $\Delta = 1$ if we decide that H_1 is correct

We have:

- $\{\Delta = 0\} = \{X \leq t_c\}$ and
- $\{\Delta = 1\} = \{X > t_c\}$

	$\Delta = 0$	$\Delta = 1$
H_0	☺	Type I-error ☹
H_1	Type II-error ☹	☺

$$\mathbb{P}_0(\Delta = 1) = \mathbb{P}_0(X > t_c) = \alpha$$

$$\mathbb{P}_1(\Delta = 0) = \mathbb{P}_0(X \leq t_c) = \beta$$

The probabilities of the configurations are

	$\Delta = 0$	$\Delta = 1$
H_0	$1 - \alpha$	α
H_1	β	$1 - \beta$

Telling apart two distributions based on a sample

Let's assume that $X_1, \ldots, X_n$ are i.i.d. but can follow two distributions:

- Hypothesis $H_0 : X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu_0, \sigma_0^2)$
- Hypothesis $H_1 : X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu_1, \sigma_1^2)$,

Can we try to decide based on a sample $x_1, \ldots, x_n$ which hypothesis is the correct one?

We can for example compute $\bar{x}$ and use the fact that

- Hypothesis $H_0 : \bar{X} \sim \mathcal{N}(\mu_0, \frac{\sigma_0^2}{n})$
- Hypothesis $H_1 : \bar{X} \sim \mathcal{N}(\mu_1, \frac{\sigma_1^2}{n})$,

Since the variance decrease with n, with a well chosen t_c, the probability of error should decrease with n.

Testing an alternative with one hypothesis to privilege by default

When deciding between hypotheses, the situation is very often asymmetric: there is one hypothesis which should be privileged by default.

Ham vs spam.
If a spam filter has to decide between two hypotheses

- This email is valid correspondence ("ham")
- This email is spam

it is much worse to classify ham as spam than the opposite.

By default we would rather consider that a mail is ham. This will be the null hypothesis, H_0.

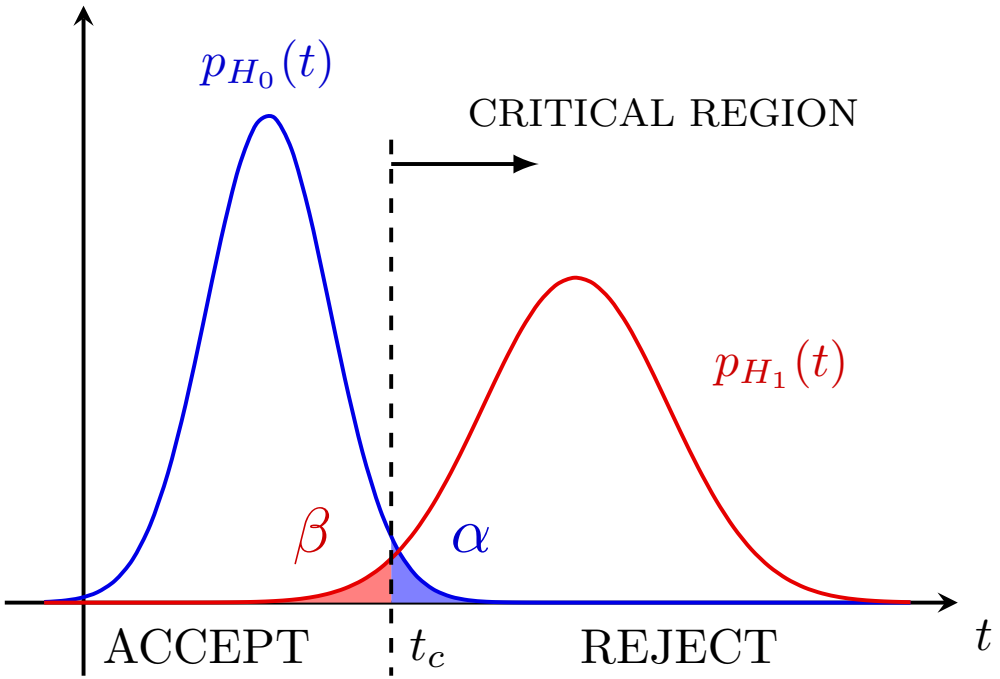
Tumor vs not.
If the result of an analysis based on a radio or a CT-SCAN has to detect the presence of a tumor, it is much worse to fail to detect an existing tumor than to detect something which will turn out later not to be. So here the null hypothesis H_0 will be "there is a tumor."

The Neyman-Pearson hypothesis testing framework

We assume that

- the data follows a distribution $p(\cdot;\theta)$ from a statistical model parameterized by $\theta \in \Theta$.
- Under the null hypothesis $\theta \in H_0 \subset \Theta$, and under the alternate hypothesis, $\theta \in H_1 \subset \Theta$.
- $H_0 \cap H_1 = \varnothing$.
- We assume that there is a *statistic* of the data $T = T(X_1, \ldots, X_n)$ which tends to be small under H_0 and larger under H_1
- The null hypothesis H_0 is privileged by default
- Our priority is to make sure that the Type-I error $\alpha = \mathbb{P}_0(\Delta = 1)$ is low.
- We will thus choose the *critical value* t_c on T to guarantee that α is low.

The Neyman-Pearson testing framework: vocabulary



- α is the significance level (Type-I error)
- $1-\alpha$ is the confidence level
- β is the Type-II error level
- $1-\beta$ is the power

- We will decide that H_1 is correct (i.e. set $\Delta = 1$) typically if $T \in [t_c, +\infty)$ which is called the *critical region* of the test. This set can take other forms.
- if $\Delta = 1$ we say that "we reject the null hypothesis" and that the result is "statistically significant.

One-sided Gaussian test

We assume that $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ where σ is known.

- We consider the simple alternative $H_0 : \mu = \mu_0$ *vs* $H_1 : \mu = \mu_1$ with $\mu_1 > \mu_0$.
- We consider the *test statistic*

$$T(X_1, \dots, X_n) = T := \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}.$$

We have $T \overset{H_0}{\sim} \mathcal{N}(0,1)$, so $\mathbb{P}_0(T > z_{1-\alpha}) = \alpha$ and we can choose $t_c = z_{1-\alpha}$ to control the type-I error.

We will reject the null hypothesis if T falls in the *critical region* $[z_{1-\alpha}, +\infty)$. In that case, we also say that $\bar{X}$ is *significantly larger* than μ_0.

We have $T - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} \overset{H_1}{\sim} \mathcal{N}(0,1)$, so

$$\beta = \mathbb{P}_1(T \leq t_c) = \mathbb{P}_1\left(T - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} \leq t_c - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right) = \Phi\left(-\sqrt{n}\left(\frac{\mu_1 - \mu_0}{\sigma}\right) + t_c\right)$$

Simple hypothesis vs composite hypothesis

A simple hypothesis is a hypothesis $H_k = \{\theta_k\}$ which specifies a single value for θ. A non-simple hypothesis is called a *composite* hypothesis.

Simple alternative

$$H_0 : \theta = \theta_0 \qquad \textit{vs} \qquad H_1 : \theta = \theta_1.$$

Composite alternative

Assuming that $\theta \in \mathbb{R}$,

$$H_0 : \theta = \theta_0 \qquad \textit{vs} \qquad H_1 : \theta > \theta_0.$$

$$H_0 : \theta = \theta_0 \qquad \textit{vs} \qquad H_1 : \theta < \theta_0.$$

$$H_0 : \theta = \theta_0 \qquad \textit{vs} \qquad H_1 : \theta \neq \theta_0.$$

Other alternatives leading to the one-sided Gaussian test

$H_0 : \mu = \mu_0$ *vs* $H_1 : \mu > \mu_0$

Given that t_c is only determined by the distribution under H_0, the *critical region* is again $[z_{1-\alpha}, +\infty)$

$H_0 : \mu = \mu_0$ *vs* $H_1 : \mu = \mu_1$ with $\mu_1 < \mu_0$.

In this case, we can reject if T is lower than an *critical value* such that $\mathbb{P}_0(T < t_c) = \alpha$, which entails $t_c = z_\alpha$. Of course, the *critical region* is now $(-\infty, z_\alpha]$.

$H_0 : \mu = \mu_0$ *vs* $H_1 : \mu < \mu_0$

Given that t_c is only determined by the distribution under H_0, this case is the same as the case just before for the determination of the *critical region*.

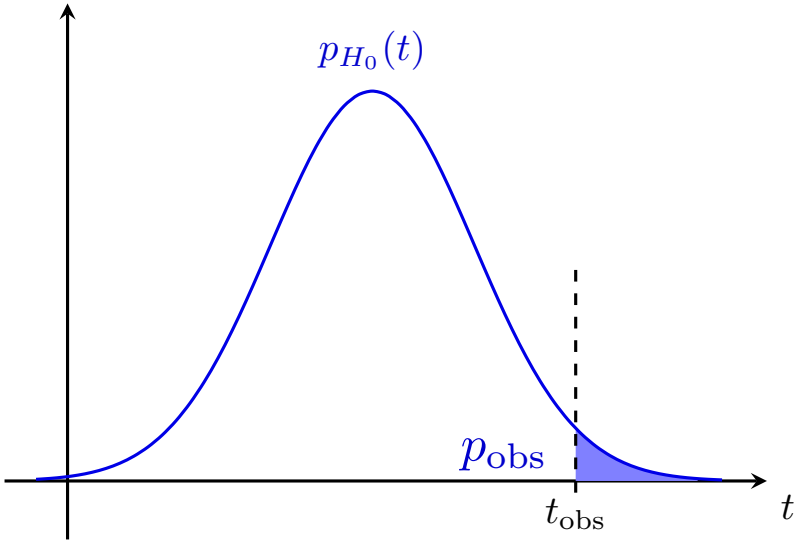
p-value

One limitation of the test methodology that we have to choose a *significance level* α. I could be useful to report a value such that one can easily assess whether the test would be rejected at other levels and which would directly measure the significance of the value t_{obs}.

p-value definition

If t_{obs} is the observed value of the *test statistic* T then the associated p-value is

$$p_{\text{obs}} = \mathbb{P}_0(T \geq t_{\text{obs}}).$$



Interpretations of the p-value

The p-value is

- the probability to observe a more extreme value of T than t_{obs} under H_0.
- the smallest significance level such that the null would be rejected for $T = t_{\text{obs}}$.
- the significance level of the test with $t_c = t_{\text{obs}}$.
- a measure of significance of the test statistic value t_{obs}.

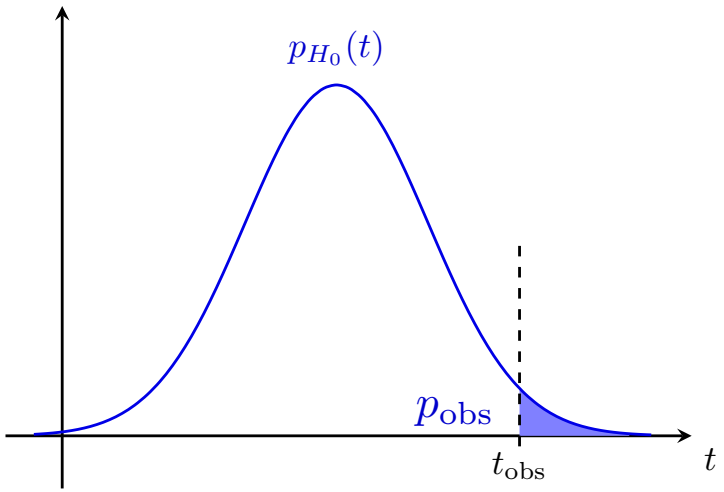
Test decision in terms of the p-value

By definition H_0 is rejected iff $(t_{\text{obs}} > t_c) \Leftrightarrow (p_{\text{obs}} < \alpha)$.

Example: p-value for a one-sided Gaussian test.

We have $T \overset{H_0}{\sim} \mathcal{N}(0,1)$, so

$$p_{\text{obs}} = \mathbb{P}_0(T \geq t_{\text{obs}}) = 1 - \Phi(t_{\text{obs}}).$$



Two-sided Gaussian test

We assume that $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ where σ is known.

- We consider the composite alternative $H_0 : \mu = \mu_0$ *vs* $H_1 : \mu \neq \mu_0$.
- We consider the *test statistic*

$$|T(X_1, \dots, X_n)| = |T| := \left| \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right|.$$

We have $T \overset{H_0}{\sim} \mathcal{N}(0,1)$, so $\mathbb{P}_0\big(|T| > z_{1-\alpha/2}\big) = 1 - \mathbb{P}_0\big(z_{\alpha/2} \leq T \leq z_{1-\alpha/2}\big) = \alpha$ and we can choose $t_c = z_{1-\alpha/2}$ to control the type-I error.

In case of rejection of the null hypothesis, we say that $\bar{X}$ is *significantly different* from μ_0.

The p-value is $p_{\text{obs}} = \mathbb{P}_0(|T| \geq |t_{\text{obs}}|) = 2\big(1 - \Phi(|t_{\text{obs}}|)\big)$.

Relationship between Gaussian confidence intervals and Gaussian tests

Two-sided test:

The null hypothesis is **not** rejected iff as $|T| \leq t_c$ but

$$-t_c \leq \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \leq t_c \quad \Leftrightarrow \quad \bar{X} - t_c \frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{X} + t_c \frac{\sigma}{\sqrt{n}}.$$

But $t_c = z_{1-\alpha/2}$, so the null hypothesis is rejected at the level of significance α iff

$$\mu_0 \notin \left[\bar{X} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}\right].$$

In other words:

The hypothesis that $\mu = \mu_0$ is rejected at a level of significance α
if and only if
μ_0 is not inside the (symmetric) Gaussian confidence interval of level $1 - \alpha$.

Relationship between Gaussian confidence intervals and Gaussian tests

One-sided test:

The null hypothesis is **not** rejected iff as $T \leq t_c$ but

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \leq t_c \quad \Leftrightarrow \quad \bar{X} \leq \mu_0 + t_c \frac{\sigma}{\sqrt{n}} \quad \Leftrightarrow \quad \bar{X} - t_c \frac{\sigma}{\sqrt{n}} \leq \mu_0.$$

But $t_c = z_{1-\alpha}$, so the null hypothesis is rejected at the level of significance α iff

$$\mu_0 \notin \left[\bar{X} - z_{1-\alpha} \frac{\sigma}{\sqrt{n}}, \, +\infty\right).$$

In other words:

The hypothesis that $\mu = \mu_0$ is rejected at a level of significance α
if and only if
μ_0 is not inside the semi-infinite upper Gaussian confidence interval of level $1 - \alpha$.

One-sided Student test

We assume that $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ where σ is **unknown**.

- We consider the simple alternative $H_0 : \mu = \mu_0$ *vs* $H_1 : \mu = \mu_1$ with $\mu_1 > \mu_0$.
- We consider the *test statistic*

$$T(X_1, \dots, X_n) = T := \frac{\bar{X} - \mu_0}{S/\sqrt{n}}.$$

We have $T \overset{H_0}{\sim} \mathsf{St}_{n-1}$, so $\mathbb{P}_0(T > t_{1-\alpha}^{(n-1)}) = \alpha$ and we can choose $t_c = t_{1-\alpha}^{(n-1)}$ to control the type-I error.

We have $T - \frac{\mu_1 - \mu_0}{S/\sqrt{n}} \overset{H_1}{\sim} \mathsf{St}_{n-1}$, so

$$\beta = \mathbb{P}_1(T \leq t_c) = \mathbb{P}_1\Big(T - \frac{\mu_1 - \mu_0}{S/\sqrt{n}} \leq t_c - \frac{\mu_1 - \mu_0}{S/\sqrt{n}}\Big) = F_{\mathsf{St}_{n-1}}\big(-\sqrt{n}\big(\tfrac{\mu_1 - \mu_0}{S}\big) + t_c\big).$$

The p-value is $p_{\text{obs}} = \mathbb{P}_0(T \geq t_{\text{obs}}) = \big(1 - F_{\mathsf{St}_{n-1}}(t_{\text{obs}})\big)$.

Two-sided Student test

We assume that $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ where σ is **unknown**.

- We consider the composite alternative $H_0 : \mu = \mu_0$ *vs* $H_1 : \mu \neq \mu_0$.
- We consider the *test statistic*

$$|T(X_1, \dots, X_n)| = |T| := \left| \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \right|.$$

We have $T \overset{H_0}{\sim} \mathsf{St}_{n-2}$, so $\mathbb{P}_0\big(|T| > t_{1-\alpha/2}^{(n-1)}\big) = 1 - \mathbb{P}_0\big(t_{\alpha/2}^{(n-1)} \leq T \leq t_{1-\alpha/2}^{(n-1)}\big) = \alpha$ and we can choose $t_c = t_{1-\alpha/2}^{(n-1)}$ to control the type-I error.

The p-value is $p_{\text{obs}} = \mathbb{P}_0(|T| \geq |t_{\text{obs}}|) = 2\big(1 - F_{\mathsf{St}_{n-1}}(|t_{\text{obs}}|)\big)$

One-sided asymptotic Gaussian test

We assume that $X_1, \dots, X_n \overset{\text{iid}}{\sim} P$ where P is unknown, but we assume that $\mathbb{E}[X_1^2] < \infty$.

- We consider the simple alternative $H_0 : \mu = \mu_0$ *vs* $H_1 : \mu = \mu_1$ with $\mu_1 > \mu_0$.
- We consider the *test statistic*

$$T(X_1, \dots, X_n) = T := \frac{\bar{X} - \mu_0}{\hat{\sigma}/\sqrt{n}},$$

where $\hat{\sigma}$ is a consistent estimator of σ, like S for example.

By the CLT, under H_0, $T \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)$, so $\mathbb{P}_0(T > z_{1-\alpha}) \xrightarrow[n\to\infty]{} \alpha$ and we can choose $t_c = z_{1-\alpha}$ to *asymptotically* control the type-I error.

Symmetrically, under H_1, $T - \frac{\mu_1-\mu_0}{\hat{\sigma}/\sqrt{n}} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)$, so

$$\beta = \mathbb{P}_1(T \le t_c) = \mathbb{P}_1\Big(T - \frac{\mu_1 - \mu_0}{\hat{\sigma}/\sqrt{n}} \le t_c - \frac{\mu_1 - \mu_0}{\hat{\sigma}/\sqrt{n}}\Big) \approx \Phi\big(-\sqrt{n}\big(\tfrac{\mu_1-\mu_0}{\hat{\sigma}}\big) + t_c\big).$$

We can define similarly two-sided asymptotic Gaussian tests.

One-sided χ^2 test for the variance σ^2

We assume that $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ where μ is **unknown**.

- We consider the simple alternative $H_0 : \sigma = \sigma_0$ *vs* $H_1 : \sigma = \sigma_1$ with $\sigma_1 > \sigma_0$.
- We consider the *test statistic*

$$T(X_1, \dots, X_n) = T := (n-1)\frac{S^2}{\sigma_0^2} = \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

We have $T \overset{H_0}{\sim} \chi^2_{n-1}$, so if $\chi^2_{n-1,1-\alpha}$ is the quantile of level $1-\alpha$ of a χ^2_{n-1} distribution, $\mathbb{P}_0(T > \chi^2_{n-1,1-\alpha}) = \alpha$ and we can choose $t_c = \chi^2_{n-1,1-\alpha}$ to control the type-I error.

We have $\frac{\sigma_0^2}{\sigma_1^2} T \overset{H_1}{\sim} \chi^2_{n-1}$, so $\beta = \mathbb{P}_1(T \leq t_c) = \mathbb{P}_1\left(\frac{\sigma_0^2}{\sigma_1^2} T \leq \frac{\sigma_0^2}{\sigma_1^2} t_c\right) = F_{\chi^2_{n-1}}\left(\frac{\sigma_0^2}{\sigma_1^2} t_c\right)$.

The p-value is $p_{\text{obs}} = \mathbb{P}_0(T \geq t_{\text{obs}}) = \left(1 - F_{\chi^2_{n-1}}(t_{\text{obs}})\right)$, with $F_{\chi^2_{n-1}}$ the cdf of a χ^2_{n-1} r.v.
We could define similarly a two-sided χ^2 test.

Two-sided Wald test

We assume that $\hat{\theta}$ is the MLE for the parameter θ based on an i.i.d. sample $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p(\cdot; \eta)$ with $\theta = \psi(\eta)$. We consider the log-likelihood $\ell(\theta)$, the Fisher information matrix $I(\theta)$.

- We consider the composite alternative $H_0 : \theta = \theta_0$ *vs* $H_1 : \theta \neq \theta_0$.
- We consider the *test statistic* $|T|$ with

$$T(X_1, \ldots, X_n) = T := \sqrt{I(\hat{\theta})}(\hat{\theta} - \theta_0).$$

By the CLT with Slutsky, under H_0, $T \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)$, so $\mathbb{P}_0(|T| > z_{1-\alpha/2}) \underset{n\to\infty}{\longrightarrow} 1-\alpha$ and we can choose $t_c = z_{1-\alpha/2}$ to *asymptotically* control the type-I error.

We can define an asymptotic p-value $\mathbb{P}_0(|T| > |t_{\text{obs}}|) = 2(1 - \Phi(|t_{\text{obs}}|))$.

Summary

- In the Neyman-Pearson framework a null hypothesis H_0 is the default hypothesis.
- We can *reject* the null hypothesis in favor of an alternative if the value of a *test statistics* is larger than a *critical value*.
- We focus on controling the Type-I error level α, aka the *significance level*.
- Instead of setting the *critical value* based on a *significance level*.
- The p-value p_{obs} is the probability $\mathbb{P}_0(T \geq t_{\text{obs}})$.
- It is possible to construct one and two-sided Gaussian and Student tests.
- It is possible to construct asymptotic Gaussian tests.
- One form of asymptotic test for $\hat{\theta}_{\text{MLE}}$ is the Wald test.
- The null is rejected at the confidence level α in a two-sided test iff the parameter μ_0 or θ_0 is not in the corresponding (symmetric) confidence interval.
- The same holds for one-sided test, but with one-sided confidence intervals.

Bayesian Statistics

EE209 - Eléments de Statistiques pour les Data Sciences

An estimator that takes the form of a probability distribution...

In the context of estimation, we have seen so far:

Point Estimators: the MLE, the method of moments (MM) produce estimators that take the form of a single number (or vector).

Confidence Intervals: i.e., interval estimators that are specified by two numbers. (They can be generalized in higher dimension with confidence regions which are sets.)

A school of thought in statistics called Bayesian statistic considers that estimation being intrinsically uncertain, we should express our uncertainty about the estimated quantity by providing an estimator which is itself a probability distribution, but this time about the unknown parameter. Bayesians propose a simple methodology to compute these probabilities, which is based on treating unknown parameters as random variables + specifying an initial "a priori" distribution on these parameters and applying... the Bayes rule.

Frequentist statisticians, who represent the other main school of thought have traditionally had reservations about the Bayesian principles and have proposed another way producing an estimator taking the form of a probability distribution, the bootstrap, which we will not cover in this course.

Today, it is well accepted that both frequentist and Bayesian principle have merits.

About notations

In this chapter we will simplify/losen some of the notations for legibility:

We will write

- $p(x)$ for $p_X(x)$ or $P_X(x)$: we will use the same notations for pmf and pdfs and drop the index indicating whose r.v. this is the pdf of. The choice of the letter in the argument will implicitly specify this.
- similarly, we will write $p(x|y)$ for $p_{X|Y}(x|y)$, $p(x,y)$ for $p_{(X,Y)}(x,y)$, and $p(y)$ for $p_Y(y)$. Note that this does not mean that we assume that X and Y have the same distribution... Indices, might reappear in ambiguous cases.
- for reason that will become clear, we will write $p(x|\theta)$ instead of the previous $p_\theta(x)$ or $p(x;\theta)$ to indicate the dependence on a parameter of the statistical model. This is the "Bayesian way" of writing this dependence.

Bayesian estimation for a single observation

Bayesians treat the parameter θ as a **random variable**.

A priori

The Bayesian has to specify an *a priori* distribution $p(\theta)$ for the model parameters θ, which models her/his prior belief of the relative plausibility of different values of the parameter.

A posteriori

The observation contributes through the likelihood: $p(x|\theta)$.
The *a posteriori* distribution on the parameters is then

$$p(\theta|x) = \frac{p(x|\theta)\,p(\theta)}{p(x)} \propto p(x|\theta)\,p(\theta) \qquad \text{with} \quad p(x) = \int p(x \mid \theta)\,p(\theta)\,d\theta$$

$\rightarrow$ The Bayesian estimator is therefore a probability distribution on the parameters.

This estimation procedure is called Bayesian inference.

Bayesian inference for a sample of observations

Assume that we have a sample $\mathcal{D}_n = \{x_1, \ldots, x_n\}$ of observations that are i.i.d. from a distribution with pmf/pdf $p(x \mid \theta)$ in a statistical model $\mathcal{P}_\Theta$ for some value of θ. The data $\mathcal{D}_n$ is often called the *evidence*.

As before, we use an a priori distribution (or prior distribution) $p(\theta)$ over Θ.

Since for a given θ the data is i.i.d., the likelihood now takes the form:

$$p(\mathcal{D}_n \mid \theta) = p(x_1, \ldots, x_n \mid \theta) = p(x_1 \mid \theta) \ldots p(x_n \mid \theta)$$

The a posteriori distribution (or posterior distribution) is obtained again using Bayes' rule

$$p(\theta \mid \mathcal{D}_n) = \frac{p(\mathcal{D}_n \mid \theta)\, p(\theta)}{p(\mathcal{D}_n)} \qquad \text{with} \qquad p(\mathcal{D}_n) = \int p(\mathcal{D}_n \mid \theta)\, p(\theta)\, d\theta.$$

The Beta distribution

A beta random variable $\theta \sim \text{Beta}(\alpha, \beta)$ is a random variable defined on the interval $[0, 1]$ and whose density takes form

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\theta^{\alpha-1}\,(1-\theta)^{\beta-1}\,1_{\{0\leq\theta\leq 1\}}, \quad \text{for} \quad \alpha, \beta > 0.$$

$$\mathbb{E}[\theta] = \frac{\alpha}{\alpha+\beta} \qquad \text{and} \qquad \text{Var}(\theta) = \frac{\alpha}{\alpha+\beta}\,\frac{\beta}{\alpha+\beta}\,\frac{1}{1+\alpha+\beta}.$$

So if we define $\theta_0 := \mathbb{E}[\theta]$ and $n_0 := \alpha + \beta$, we have

$$p(\theta) = \frac{\Gamma(n_0)}{\Gamma\big(n_0\theta_0\big)\,\Gamma\big(n_0(1-\theta_0)\big)}\,\theta^{n_0\theta_0-1}\,(1-\theta)^{n_0(1-\theta_0)-1}\,1_{\{0\leq\theta\leq 1\}}.$$

$$\mathbb{E}[\theta] = \theta_0 \qquad \text{and} \qquad \text{Var}(\theta) = \frac{\theta_0(1-\theta_0)}{1+n_0}$$

Bayesian inference for the Beta-Bernoulli model

Let $\mathcal{D}_n = \{x_1, \ldots, x_n\}$, with $X_i \overset{iid}{\sim} \text{Ber}(\theta)$, so that we have $p(x_i \mid \theta) = \theta^{x_i}(1-\theta)^{1-x_i}$.

For the whole sample, the likelihood is $p(\mathcal{D}_n \mid \theta) = p(x_1 \mid \theta) \ldots p(x_n \mid \theta) = \theta^N (1-\theta)^{n-N}$.

And we use the prior distribution $p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}\, 1_{\{0 \leq \theta \leq 1\}}$.

We can calculate the posterior distribution:

$$
\begin{aligned}
p(\theta \mid \mathcal{D}_n) &\propto p(\mathcal{D}_n \mid \theta)\, p(\theta) \\
&\propto \theta^N (n-\theta)^{n-N}\, \theta^{\alpha-1}(1-\theta)^{\beta-1}\, 1_{\{0 \leq \theta \leq 1\}} \\
&\propto \theta^{N+\alpha-1}\, (1-\theta)^{n-N+\beta-1}\, 1_{\{0 \leq \theta \leq 1\}} \\
&= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(N+\alpha)\Gamma(n-N+\beta)}\, \theta^{N+\alpha-1}\, (1-\theta)^{n-N+\beta-1}\, 1_{\{0 \leq \theta \leq 1\}}
\end{aligned}
$$

So $\theta \mid \mathcal{D}_n \sim \text{Beta}(N+\alpha, n-N+\beta)$.

Posterior mean and posterior variance (in the Beta-Bernoulli model)

Based on the posterior distribution $\theta \mid \mathcal{D}_n \sim \text{Beta}(N + \alpha, n - N + \beta)$,

We can compute:

- the **posterior mean**

$$\theta_{\text{PM}} = \mathbb{E}[\,\theta \mid \mathcal{D}_n\,] = \frac{N + \alpha}{n + \alpha + \beta} = \frac{n}{n + \alpha + \beta}\,\frac{N}{n} + \frac{\alpha + \beta}{n + \alpha + \beta}\,\frac{\alpha}{\alpha + \beta}$$
$$= \frac{n}{n + n_0}\;\hat{\theta}_{\text{MLE}} \;+\; \frac{n_0}{n + n_0}\;\theta_0.$$

  θ_{PM} can be seen as a *point estimator* which, in this case (and others as we will see), is a convex combination of the MLE and the prior mean.

- the **posterior variance**

$$\text{Var}(\,\theta \mid \mathcal{D}_n\,) = \frac{\theta_{\text{PM}}\,(1 - \theta_{\text{PM}})}{1 + n + n_0},$$

  which is way to quantify how the posterior distribution is concentrated.

The posterior mode aka the *maximum a posteriori* (MAP)

Another *point estimator* that can be derived from the whole posterior distribution, is the posterior mode which is usually called the MAP or *maximum a posteriori*.

$$\begin{aligned}\theta_{\text{MAP}} &= \arg\max_{\theta} \log p(\theta \mid \mathcal{D}_n) = \arg\max_{\theta} \log \left(\frac{p(\mathcal{D}_n \mid \theta)}{} p(\theta) p(\mathcal{D}_n) \right) \\ &= \arg\max_{\theta} \log p(\mathcal{D}_n \mid \theta) + \log p(\theta).\end{aligned}$$

Note that θ_{MAP} resembles the MLE because $\log p(\mathcal{D}_n \mid \theta) = \ell(\theta)$ is the log-likelihood.

For Bayesian estimation in the Bernoulli model with a Beta prior we get

$$\log p(\theta \mid \mathcal{D}_n) = (N + \alpha - 1) \log \theta + (n - N + \beta - 1) \log(1 - \theta) + \text{cst}$$

if $\theta \in [0, 1]$ and $-\infty$ if $\theta \notin [0, 1]$.
We thus find

$$\theta_{\text{MAP}} = \frac{N + \alpha - 1}{n + \alpha + \beta - 2} \qquad \text{so that for } \alpha = \beta = 1 \quad \theta_{\text{MAP}} = \hat{\theta}_{\text{MLE}}.$$

Bayesian inference for the mean of a Gaussian, assuming that σ^2 is known

Let $\mathcal{D}_n = \{x_1, \ldots, x_n\}$, with $x_i \,|\, \mu \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, $\quad \mu \sim \mathcal{N}(\mu_0, \tau^2)$.

$$
\begin{aligned}
p(\mu \mid \mathcal{D}_n) &\propto p(\mathcal{D}_n \mid \mu)\, p(\mu) \\
&\propto \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2} \frac{1}{(2\pi\tau^2)^{1/2}} e^{-\frac{1}{2\tau^2}(\mu-\mu_0)^2} \\
&\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (\mu - x_i)^2 - \frac{1}{2\tau^2}(\mu-\mu_0)^2\right) \\
&\propto \exp\left\{-\frac{1}{2}\left(\left[\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right]\mu^2 - 2\left[\frac{n\bar{x}}{\sigma^2}+\frac{\mu_0}{\tau^2}\right]\mu + \ldots\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\frac{n\tau^2+\sigma^2}{\tau^2\sigma^2}\right]\left(\mu^2 - 2\left[\frac{n\tau^2\,\bar{x}+\sigma^2\,\mu_0}{n\tau^2+\sigma^2}\right]\mu + \ldots\right)\right\}
\end{aligned}
$$

$\mu \mid \mathcal{D}_n \sim \mathcal{N}(\mu_{\text{post}}, \tau^2_{\text{post}})$ with

$$
\mu_{\text{post}} = \frac{n\tau^2}{n\tau^2+\sigma^2}\,\bar{x} + \frac{\sigma^2}{n\tau^2+\sigma^2}\,\mu_0, \qquad \tau^2_{\text{post}} = \frac{\sigma^2}{n+\frac{\sigma^2}{\tau^2}}.
$$

Posterior distribution for μ, assuming that σ^2 is known

$$\mu \mid \mathcal{D}_n \sim \mathcal{N}(\mu_{\text{post}}, \tau^2_{\text{post}}) \quad \text{with} \quad \mu_{\text{post}} = \frac{n\tau^2}{n\tau^2 + \sigma^2}\,\bar{x} + \frac{\sigma^2}{n\tau^2 + \sigma^2}\,\mu_0, \qquad \tau^2_{\text{post}} = \frac{\sigma^2}{n + \frac{\sigma^2}{\tau^2}}.$$

Of course, the parameters of the posterior are the **posterior mean** and **posterior variance**:

$$\mu_{\text{PM}} = \mathbb{E}[\mu \mid \mathcal{D}_n] = \mu_{\text{post}} \qquad \text{and} \qquad \text{Var}(\mu \mid \mathcal{D}_n) = \tau^2_{\text{post}}.$$

For the MAP, since the mode of a Gaussian distribution is also its mean we have

$$\theta_{\text{MAP}} = \theta_{\text{PM}}.$$

Bayesian inference for a sample of observations

Given an i.i.d. sample $x_1, \ldots, x_n$, in frequentist statistics, we have

$$p_\theta(x_1, \ldots, x_n) = p_\theta(x_1)\, p_\theta(x_2)\, \ldots p_\theta(x_n)$$

Similarly, in Bayesian terms, we have

$$p(x_1, \ldots, x_n \mid \theta) = p(x_1|\theta)\, p(x_2|\theta)\, \ldots\, p(x_n|\theta). \qquad (*)$$

But since θ is a random variable, the above equation is **not equivalent to**

$$p(x_1, \ldots, x_n) = p(x_1)\, p(x_2)\, \ldots\, p(x_n).$$

Equation $(*)$ is true if

$X_1, \ldots, X_n$ are "independent **given** θ" which is different than $X_1, \ldots, X_n$ are independent.

This makes sense if you think of observations from a biased coin with unknown bias. The concept of **conditional independence** is a complicated concept and we will not explore it further, but it important to remember that the observations are independent only "given θ."

Consecutive updates of the posterior

$$\begin{aligned} p(\theta \mid \mathcal{D}_n)\, p(\mathcal{D}_n) = p(\mathcal{D}_n \mid \theta)\, p(\theta) &= p(x_1 \mid \theta) \ldots p(x_n \mid \theta)\, p(\theta) \\ &= p(x_n \mid \theta)\, p(\mathcal{D}_{n-1} \mid \theta)\, p(\theta) \\ &= p(x_n \mid \theta)\, p(\theta \mid \mathcal{D}_{n-1})\, p(\mathcal{D}_{n-1}). \end{aligned}$$

But $p(\mathcal{D}_n) = p(x_1, \ldots, x_n) = p(x_n \mid \mathcal{D}_{n-1})\, p(\mathcal{D}_{n-1})$. So

$$p(\theta \mid \mathcal{D}_n) = p(x_n \mid \theta)\, p(\theta \mid \mathcal{D}_{n-1})\, \frac{p(\mathcal{D}_{n-1})}{p(\mathcal{D}_n)} = \frac{p(x_n \mid \theta)\, p(\theta \mid \mathcal{D}_{n-1})}{p(x_n \mid \mathcal{D}_{n-1})}.$$

And

$$\int p(x_n \mid \theta)\, p(\theta \mid \mathcal{D}_{n-1})\, d\theta = \int p(x_n \mid \theta)\, \frac{p(\mathcal{D}_{n-1} \mid \theta)\, p(\theta)}{p(\mathcal{D}_{n-1})} d\theta = \frac{p(\mathcal{D}_n)}{p(\mathcal{D}_{n-1})} = p(x_n \mid \mathcal{D}_{n-1}).$$

Finally

$$\boxed{p(\theta \mid \mathcal{D}_n) = \frac{p(x_n \mid \theta)\, p(\theta \mid \mathcal{D}_{n-1})}{p(x_n \mid \mathcal{D}_{n-1})} = \frac{p(x_n \mid \theta)\, p(\theta \mid \mathcal{D}_{n-1})}{\int p(x_n \mid \theta)\, p(\theta \mid \mathcal{D}_{n-1})\, d\theta}}.$$

The posterior is the new prior after having seen data...

If we apply the formula we established

$$p(\theta \mid \mathcal{D}_n) = \frac{p(x_n \mid \theta)\, p(\theta \mid \mathcal{D}_{n-1})}{p(x_n \mid \mathcal{D}_{n-1})}.$$

to data arriving one by one, we get

$$p(\theta \mid x_1) = \frac{p(x_1 \mid \theta)\, p(\theta)}{p(x_1)} \qquad \text{and} \qquad p(\theta \mid x_1, x_2) = \frac{p(x_2 \mid \theta)\, p(\theta \mid x_1)}{p(x_2 \mid x_1)} \qquad \text{etc}$$

So the posterior after seeing x_1 becomes the new prior before seeing x_2 and if we apply Bayes rule, we obtain a new posterior which is the same as if we had observed x_1 and x_2 at the same time !

The posterior acts as a memory of what we have learned, which can be updated when new information arrives.

The Dirichlet distribution

We say that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ follows the Dirichlet distribution and note

$$\boldsymbol{\theta} \sim \text{Dir}(\boldsymbol{\alpha})$$

for $\boldsymbol{\theta}$ in the simplex $\triangle_K = \{\boldsymbol{u} \in \mathbb{R}_+^K \mid \sum_{k=1}^K u_k = 1\}$ and if it has the density

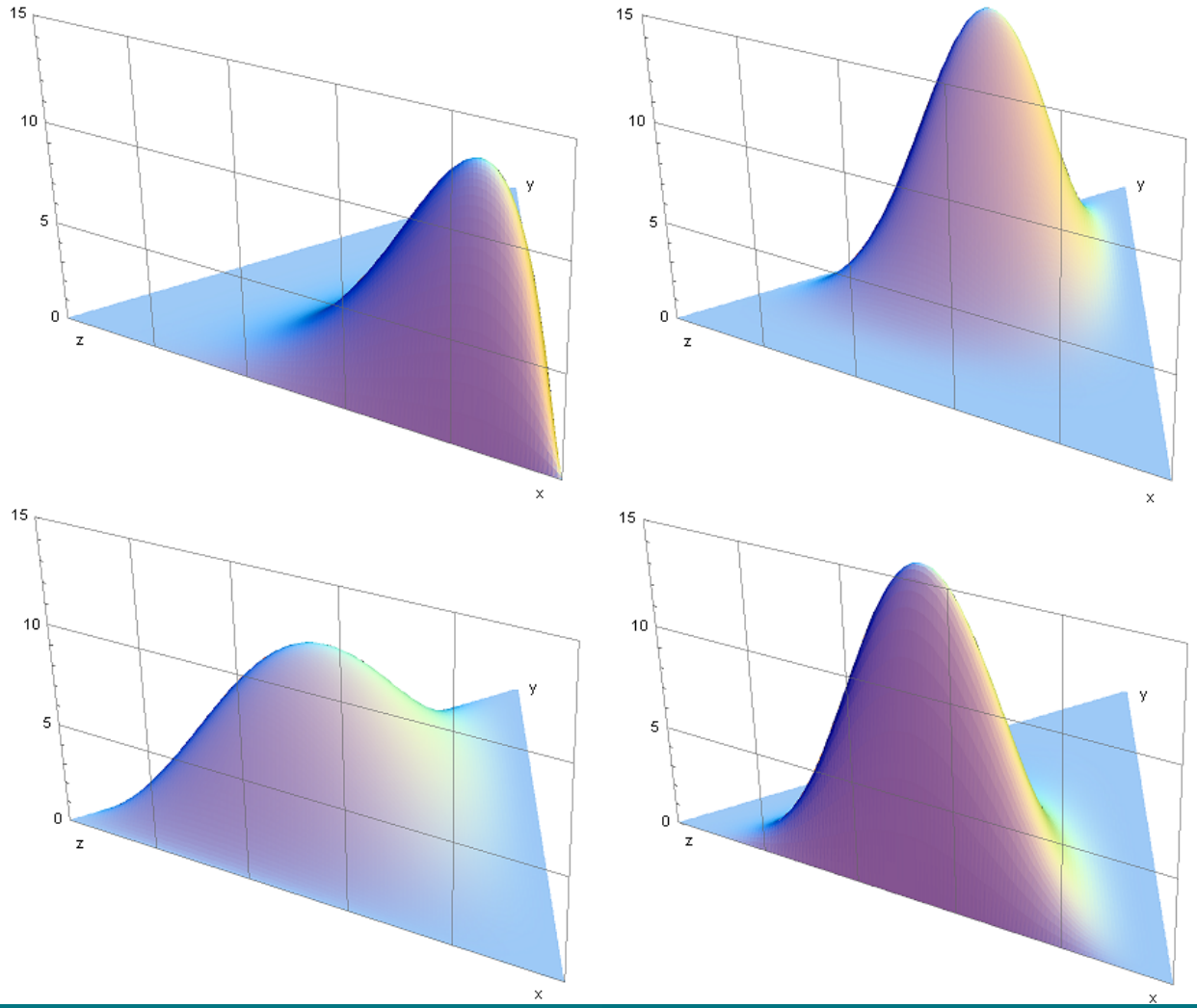
$$p(\boldsymbol{\theta}; \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\prod_{k=1}^K \Gamma(\alpha_k)} \theta_1^{\alpha_1 - 1} \ldots \theta_{K-1}^{\alpha_{K-1} - 1} \left(1 - \sum_{j=1}^{K-1} \theta_j\right)^{\alpha_K - 1} 1_{\{\forall 1 \leq j \leq K-1,\ 0 \leq \theta_j \leq 1\}}.$$

where

$$\alpha_0 = \sum_k \alpha_k \quad \text{and} \quad \Gamma(x) := \int_0^\infty t^{x-1} e^{-t} dt.$$

In fact to be rigorous, we should write $p(\theta_1, \ldots, \theta_{K-1}; \boldsymbol{\alpha})$ but it is more convenient to write $p(\boldsymbol{\theta}; \boldsymbol{\alpha})$ and keep in mind that $\theta_K = 1 - \sum_{j=1}^{K-1} \theta_j$.

Dirichlet distribution II



Bayesian estimation of a multinomial random variable

Let $\mathcal{D}_n$ be an i.i.d. sample $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ with

- $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iK}) \sim \mathcal{M}(1, \boldsymbol{\theta})$ following a "multinoulli" r.v. variable
- $\boldsymbol{\theta} \sim \text{Dir}(\boldsymbol{\alpha})$, i.e. $\boldsymbol{\theta}$ follows a Dirichlet prior.

$$p(\boldsymbol{\theta}) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \qquad \text{and} \qquad p(\boldsymbol{z}_i | \boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{z_{ik}}$$

We have

$$p(\boldsymbol{\theta} | \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) = \frac{p(\boldsymbol{\theta}) \prod_i p(\boldsymbol{z}_i | \boldsymbol{\theta})}{p(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)} \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \prod_{i=1}^{n} \prod_{k=1}^{K} \theta_k^{z_{ik}} \propto \prod_{k} \theta_k^{\sum_{i=1}^{n} z_{ik} + \alpha_k - 1}$$

So that $(\boldsymbol{\theta} \mid \mathcal{D}_n) \sim \text{Dir}(\alpha_1 + N_1, \ldots, \alpha_K + N_K)$ with $N_k = \sum_{i=1}^{n} z_{ik}$.

Posterior Mean and Posterior Variance in the Dirichlet-Multinomial model

If $\boldsymbol{\theta} \sim \text{Dir}(\alpha_1, \ldots, \alpha_K)$ then

$$\theta_k^{\text{prior}} = \mathbb{E}[\theta_k] = \frac{\alpha_k}{\alpha_0} \quad , \quad \text{Var}(\theta_k) = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)} \quad \text{with} \quad \alpha_0 = \sum_k \alpha_k.$$

We have shown that $(\boldsymbol{\theta} \mid \mathcal{D}_n) \sim \text{Dir}(\alpha_1 + N_1, \ldots, \alpha_K + N_K)$

Posterior mean

$$\theta_{\text{PM},k} := \mathbb{E}[\theta_k|\mathcal{D}_n] = \frac{\alpha_k + N_k}{\alpha_0 + n} = \frac{\alpha_0}{\alpha_0 + n} \frac{\alpha_k}{\alpha_0} + \frac{n}{\alpha_0 + n} \frac{N_k}{n}.$$

Posterior variance

$$\text{Var}(\, \theta_k \mid \mathcal{D}_n \,) = \frac{\theta_{\text{PM},k}\,(1 - \theta_{\text{PM},k})}{\alpha_0 + n + 1}$$

Conjugate priors

A family of prior distribution $\mathcal{P}_A = \{p_\alpha(\theta) \mid \alpha \in A\}$

is said to be **conjugate** to a model $\mathcal{P}_\Theta$, if, for a sample

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p(\cdot \mid \theta) \qquad \text{with} \qquad p(\cdot \mid \theta) \in \mathcal{P}_\Theta,$$

the distribution q defined by $q(\theta) := p(\theta | x_1, \ldots, x_n) = \dfrac{p_\alpha(\theta) \prod_i p(x_i \mid \theta)}{\int p_\alpha(\theta) \prod_i p_\theta(x_i) d\theta}$

is such that $q \in \mathcal{P}_A$.

Note that if we could also have used Bayesian notations $p(\theta \mid \alpha)$ instead of $p_\alpha(\theta)$.

Conjugate families

Likelihood	Conjugate prior
Bernoulli/Binomial	Beta
Multinomial	Dirichlet
Poisson	Gamma
Normal with fixed σ^2	Normal
Normal with fixed μ	Inverse gamma
Normal multivar. Normal with fixed Σ	Normal
multivar. Normal with fixed Σ with fixed $\boldsymbol{\mu}$	Inverse Wishart
Exponential	Gamma

In all the examples that we have considered (Bernoulli, Gaussian and Multinomial model) we have used the conjugate prior for convenience each time. This is not necessary but then the integrals cannot be computed analytically in general...

Posterior expectations and the predictive distribution

The principle of Bayesian estimation is that the prior and posterior distribution model the *uncertainty* that we have in the estimation process. As a consequence, one should always integrate over the uncertainty. So the final estimate for a function $f(\theta)$ is

$$\mathbb{E}[f(\theta) \mid \mathcal{D}_n] = \int f(\theta)\, p(\theta|\mathcal{D}_n)\, d\theta.$$

Of course the posterior mean is a particular example.
In particular the predictive distribution is the pmf/pdf of a new observation x' from the *model* given the *evidence* provided by the data $\mathcal{D}_n = \{\mathbf{x}_1, \ldots, x_n\}$

$$\mathbb{E}\big[p(x'|\theta) \mid \mathcal{D}_n\big] = p(x'|x_1, \ldots, x_n) = \int p(x'|\theta)\, p(\theta|x_1, \ldots, x_n)\, d\theta.$$

We will not discuss the calculations of the predictive distribution further in this course.

Bayesian inference for the precision of a Gaussian

Reminder: $X \sim \Gamma(k, \beta)$ then $p(x) = \frac{\beta^k}{\Gamma(k)} x^{k-1} \exp\big(-\beta x\big)$

By definition the *precision* is $\lambda = \sigma^{-2}$

Let $\mathcal{D}_n = \{x_1, \dots, x_n\}$, with $x_i \,|\, \lambda \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \lambda^{-1})$, $\quad \lambda \sim \Gamma(\frac{n_0}{2}, \frac{n_0 \sigma_0^2}{2})$.

If $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \mu)^2$, then the Gaussian likelihood takes the form

$$p(\mathcal{D}_n|\lambda) = \big(\tfrac{\lambda}{2\pi}\big)^{\frac{n}{2}} \exp\big(-\lambda \tfrac{n\hat{\sigma}^2}{2}\big) \quad \text{and} \quad p(\lambda) = \frac{\big(\frac{n_0\sigma_0^2}{2}\big)^{\frac{n_0}{2}}}{\Gamma\big(\frac{n_0}{2}\big)} \lambda^{\frac{n_0}{2}-1} \exp\big(-\lambda \tfrac{n_0\sigma_0^2}{2}\big)$$

$$p(\lambda|\mathcal{D}_n) \propto \lambda^{\frac{n_0+n}{2}-1} \exp\Big(-\lambda\big[\tfrac{n_0\sigma_0^2}{2} + \tfrac{n\hat{\sigma}^2}{2}\big]\Big)$$

So that $\lambda|\mathcal{D}_n \sim \Gamma\big(\frac{n_0+n}{2}, \frac{n_0\sigma_0^2 + n\hat{\sigma}^2}{2}\big)$

$$\mathbb{E}[\lambda|\mathcal{D}_n] = \frac{n_0+n}{n_0\sigma_0^2 + n\hat{\sigma}^2} = \Big(\frac{n_0}{n_0+n}\sigma_0^2 + \frac{n}{n_0+n}\hat{\sigma}^2\Big)^{-1}, \qquad \text{Var}(\lambda|\mathcal{D}_n) = \frac{n_0+n}{(n_0\sigma_0^2 + n\hat{\sigma}^2)^2}.$$

Improper priors

For the Gaussian model with prior $\mu \sim \mathcal{N}(\mu_0, \tau^2)$ we found $\mu \mid \mathcal{D}_n \sim \mathcal{N}(\mu_{\text{post}}, \tau^2_{\text{post}})$ with

$$\mu_{\text{post}} = \frac{n\tau^2}{n\tau^2 + \sigma^2}\,\bar{x} + \frac{\sigma^2}{n\tau^2 + \sigma^2}\mu_0, \qquad \tau^2_{\text{post}} = \frac{\sigma^2}{n + \frac{\sigma^2}{\tau^2}}.$$

If we don't have a strong prior on μ it is natural to let $\tau^2 \to +\infty$. In that case the prior is not a probability distribution anymore, but the prior on μ is uniform over all real numbers! This is an example of an improper prior. The posterior distribution is still well defined with

$$\mu_{\text{post}} = \bar{x}, \qquad \tau^2_{\text{post}} = \frac{\sigma^2}{n}.$$

In that case μ_{post} is unbiased (and coincidentally $\text{Var}(\theta \mid \mathcal{D}_n) = \text{Var}(\bar{X})$).

For the Beta-Bernoulli model we found $\theta \mid \mathcal{D}_n \sim \text{Beta}(N + \alpha, n - N + \beta)$, and the posterior mean was $\theta_{\text{PM}} = \frac{N+\alpha}{n+\alpha+\beta}$. Letting $\alpha \to 0$ and $\beta \to 0$ corresponds to using an improper prior $\text{Beta}(0, 0)$ and we again obtain an unbiased posterior mean $\theta_{\text{PM}} = \frac{N}{n}$.

Frequentist Probability *vs* Bayesian probability

For *frequentists*, the concept of probability is grounded in the law of large numbers. For them, it only makes sense to talk about the probability of some event, if this event occurs in a random experiment that can (a least theoretically) be repeated indefinitely, so that the probability can be defined as the limiting frequency of occurence of the event.

Examples: Probability that a coin falls on heads, probability of winning at the lottery, probability that an isotope disintegrates in less than 1 sec, probability that my friend is upset on Mondays.

For *Bayesians*, probability distributions can be used as well to express a *belief* about possible outcomes or truth. And this belief can be updated based on a likelihood (consisting of frequentist or Bayesian probabilities) based on Bayes rules.

Examples: Probability that the universe is finite, probability that my friend is upset today, probability that the missing mass in the universe is larger than x.

Some remarks on Bayesian methods

Subjective vs objective priors. A difficultly encountered in practice is how to choose the prior. *Subjective Bayesians* argue that the prior should reflect their prior knowlegde and beliefs. *Objective Bayesians* seeks ways to choose priors that are neutral or optimal in some sense (improper priors, Jeffreys' priors, etc).

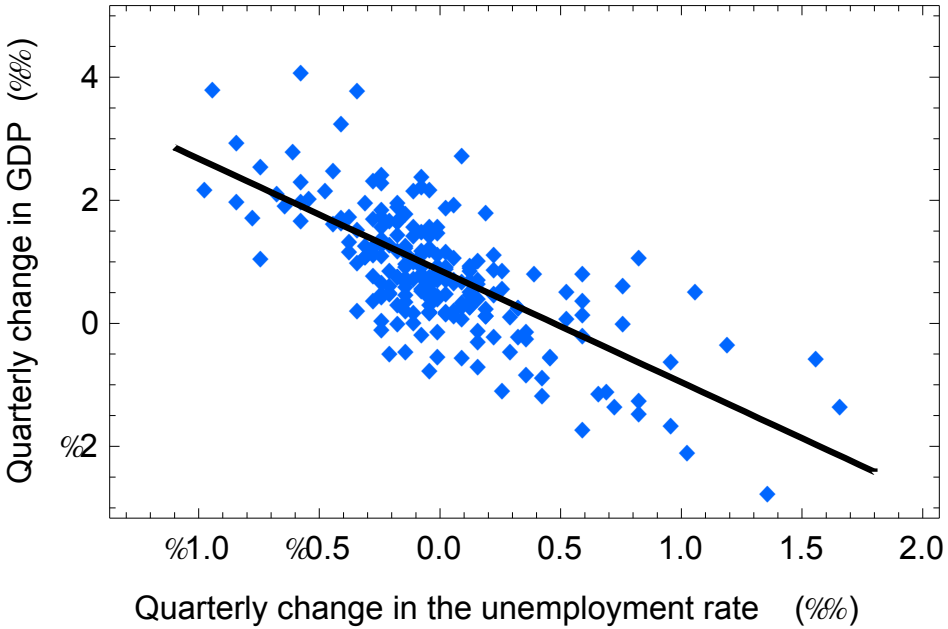
Bayesian uncertainty vs frequentist uncertainty. Bayesians consider data as fixed/given and the uncertainty on θ is obtained by combining "prior belief" encoded in the a priori distribution with the likelihood of the different values of the parameters, whereas for frequentist the uncertainty on θ is the uncertainty of $\hat{\theta}$ which depends on the distribution of the data.

Assuming that the model is correct. The logic of Bayesian inference requires that the data really comes from the likelihood, otherwise we cannot really apply Bayes' rule.

Linear regression

Eléments de Statistiques pour les Data Sciences

Simple linear regression



In economics, Okun's law is an empirical relationship between the increase in unemployment rate x and the increase in GDP y.

In statistics x are y are called

- y the *response* (or *dependent variable*)
- x the *explanatory* (or *independent*) *variable*

- What is the "best" linear function of x, so of the form $ax + b$, to approximate y?
- We will define "the best" as the one which minimizes the *mean squared error* (MSE).

This is the problem of linear regression. We talk about *simple* linear regression when there is a single explanatory variable.

Simple linear regression from a sample : statement

We consider a collection of observations (x_i, y_i) and the question: What is the linear (or more precisely affine) transformation of x that best approximates y in the least square sense?

$$\min_{a,b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} (y_i - a\, x_i - b)^2$$

$$\min_{a,b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} (y_i - a\, x_i)^2 - 2b \frac{1}{n} \sum_{i=1}^{n} (y_i - a\, x_i) + b^2$$

so that $\hat{b} = \bar{y} - a\bar{x}$ for a given value of a. Replacing b by its optimal value we get:

$$\min_{a,b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} (\breve{y}_i - a\breve{x}_i)^2 \qquad \text{with} \quad \breve{x}_i := x_i - \bar{x},\ \breve{y}_i := y_i - \bar{y}.$$

$$\min_{a \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \breve{y}_i^2 - 2a \frac{1}{n} \sum_{i=1}^{n} \breve{x}_i \breve{y}_i + a^2 \frac{1}{n} \sum_{i=1}^{n} \breve{x}_i^2.$$

Simple linear regression from a sample : statement

$$\min_{a\in\mathbb{R}} \tfrac{1}{n-1}\sum_{i=1}^{n} \breve{y}_i^2 - 2a\,\tfrac{1}{n-1}\sum_{i=1}^{n} \breve{x}_i\,\breve{y}_i + a^2\,\tfrac{1}{n-1}\sum_{i=1}^{n} \breve{x}_i^2$$

We consider the sample variances and covariance

$$s_x^2 := \tfrac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2, \quad s_y^2 := \tfrac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2, \quad s_{xy} := \tfrac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

and the correlation coefficient $r := \dfrac{s_{xy}}{s_x\,s_y}$.

We can rewrite the optimization problem as

$$\min_{a\in\mathbb{R}} s_y^2 - 2a\,s_{xy} + a^2 s_x^2, \qquad \text{so that} \quad \hat{a} = \frac{s_{xy}}{s_x^2} = \frac{r s_x s_y}{s_x^2} = r\,\frac{s_y}{s_x}.$$

Simple linear regression in the sample case: solution

We found that the best affine function of X to approximate Y in the least square sense is of the form

$$\hat{y} := \hat{f}(x) := \hat{a}x + \hat{b} \qquad \text{with} \qquad \hat{a} = r\,\frac{s_y}{s_x}, \quad \hat{b} = \bar{y} - \hat{a}\bar{x}.$$

So that for any value x we have the *estimated response* $\hat{y}$.

$$\hat{y} = \hat{f}(x) = \bar{y} + r\,\frac{s_y}{s_x}\,(x - \bar{x}).$$

If we consider the *residual* $e_i := y_i - f^*(x_i)$, then we can write

$$y_i = \hat{y}_i + e_i = \bar{y} + r\,s_y\,\frac{x_i - \bar{x}}{s_x} + e_i.$$

Properties of the residuals $e_i := y_i - \hat{y}_i$

Given that $\hat{b} = \bar{y} - \hat{a}\bar{x}$, we have $\hat{y}_i - \bar{y} = \hat{a}x_i + \hat{b} - \bar{y} = \hat{a}(x_i - \bar{x})$, and thus

$$-\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(\hat{y}_i - y_i) = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}) = \sum_{i=1}^{n} \hat{a}(x_i - \bar{x}) = 0.$$

$$\sum_{i=1}^{n} e_i x_i = \sum_{i=1}^{n} e_i(x_i - \bar{x}) = \sum_{i=1}^{n}(y_i - \hat{y}_i)(x_i - \bar{x}) = \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) - \sum_{i=1}^{n}(\hat{y}_i - \bar{y})(x_i - \bar{x})$$

$$= (n-1)s_{xy} - \sum_{i=1}^{n} \hat{a}(x_i - \bar{x})(x_i - \bar{x}) = (n-1)s_{xy} - (n-1)\hat{a}\, s_x^2 = 0.$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y} + e_i)^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n} e_i(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n} e_i^2,$$

but $$\sum_{i=1}^{n} e_i(\hat{y}_i - \bar{y}) = \hat{a}\sum_{i=1}^{n} e_i(x_i - \bar{x}) = 0.$$

Summary: properties of the residuals $e_i = y_i - \hat{y}_i$

Let $(x_1, y_1), \ldots, (x_n, y_y)$ be a collection of datapoints for linear regression.

Let

- s_x^2 and s_y^2 be the *sample variances*.
- $s_{xy} := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$ the *sample covariance*.
- $r := \frac{s_{xy}}{s_x s_y}$ the *sample correlation* also called *correlation coefficient*.

Let $\hat{a} := r \frac{s_y}{s_x}, \quad \hat{b} := \bar{y} - \hat{a}\bar{x}, \quad \hat{y}_i := \hat{a}x_i + \hat{b}.$

Then the *residuals* $e_i := y_i - \hat{y}_i$ satisfy the following properties:

- They are *centered*: $\bar{e} = \frac{1}{n} \sum_{i=1}^{n} e_i = 0.$
- They are *empirically decorrelated* from the x_i: $\sum_{i=1}^{n} e_i (x_i - \bar{x}) = 0$

Empirical variances of the estimated $\hat{y}_i$ and of the residuals e_i

We have $\bar{\hat{y}} = \bar{y} - \bar{e} = \bar{y}$. So the sample variance of $\hat{y}_i$ is

$$\frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\hat{a}^2(x_i - \bar{x})^2 = r^2\frac{s_y^2}{s_x^2}s_x^2 = r^2 s_y^2.$$

But we have proven that

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \frac{1}{n-1}\sum_{i=1}^{n}e_i^2,$$

so that

$$\frac{1}{n-1}\sum_{i=1}^{n}e_i^2 = s_y^2 - r^2 s_y^2 = s_y^2\,(1 - r^2).$$

Pythagoras and a decomposition between explained and residual variance

$$\overbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}^{(n-1)s_y^2} = \overbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}^{(n-1)r^2\,s_y^2} + \overbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}^{(n-1)(1-r^2)\,s_y^2}$$

$$\text{TSS} = \text{ESS} + \text{RSS}$$

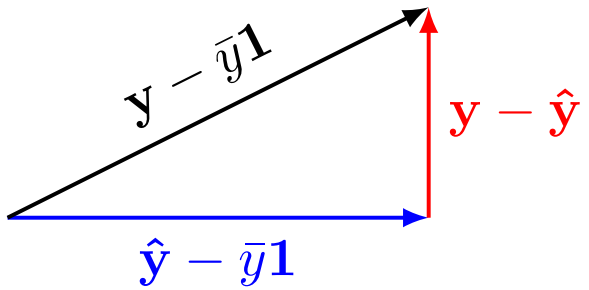
$$\text{Total SoS} = \text{Explained SoS} + \text{Residual SoS}$$

with SoS=sum of squares.

Let

- $\mathbf{y} = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$
- $\hat{\mathbf{y}} = (\hat{y}_1, \dots, \hat{y}_n)^\top \in \mathbb{R}^n$,
- $\mathbf{1} = (1, \dots, 1)^\top \in \mathbb{R}^n$.

The decomposition of the TSS corresponds to the Pythagorean triangle:



Note that

$$\hat{\mathbf{y}} - \bar{y}\mathbf{1} = \hat{a}(\mathbf{x} - \bar{x}\mathbf{1}) = r\frac{s_y}{s_x}(\mathbf{x} - \bar{x}\mathbf{1}).$$

Coefficient of determination

The *coefficient of determination* noted R^2 is defined as the fraction of the variance explained by the *explanatory variable*

$$R^2 := \frac{\mathsf{ESS}}{\mathsf{TSS}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{r^2 s_y^2}{s_y^2} = r^2.$$

So the coefficient of determination is the *square of the correlation coefficient* between x and y in the data.

Linear regression with a vector of explanatory variables

Given a dataset

$$\mathcal{D}_n = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\},$$

with $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$, we consider

- the *design matrix* $\boldsymbol{X}$ and
- the vector of *responses* $\boldsymbol{y}$

defined as

$$\boldsymbol{X} = \begin{bmatrix} — & \mathbf{x}_1^\top & — \\ — & \mathbf{x}_2^\top & — \\ & \vdots & \\ — & \mathbf{x}_n^\top & — \end{bmatrix} \quad \text{and} \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Remark: most of the time it is relevant to

- center the data: $\mathbf{x}_i^{\mathsf{c}} = \mathbf{x}_i - \bar{\mathbf{x}}$
- normalize via e.g. $x_{ij}^{\mathsf{s}} = x_{ij}^{\mathsf{c}} / \widehat{\sigma}_j$ or mapping $\mathbf{x}_{ij}^{\mathsf{c}}$ to $[0, 1]$, etc

Linear regression *aka* ordinary least square regression (OLS)

Given a dataset $\mathcal{D}_n = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ we have

$$\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} — & \mathbf{x}_1^\top & — \\ — & \mathbf{x}_2^\top & — \\ — & \vdots & — \\ — & \mathbf{x}_n^\top & — \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} y_1 - \mathbf{x}_1^\top \boldsymbol{\beta} \\ y_2 - \mathbf{x}_2^\top \boldsymbol{\beta} \\ \vdots \\ y_n - \mathbf{x}_n^\top \boldsymbol{\beta} \end{bmatrix}$$

So that we have

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{\beta}^\top \mathbf{x}_i)^2 = \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$$

with

- the vector of responses $\boldsymbol{y}^\top = (y_1, \ldots, y_n) \in \mathbb{R}^n$
- the design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ whose ith row is equal to $\mathbf{x}_i^\top$.

Solving linear regression

We can rewrite the MSE as $\frac{1}{n}Q(\boldsymbol{\beta})$ with

$$Q(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} - 2\,\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y} + \|\boldsymbol{y}\|^2.$$

A minimum has to be stationary point, i.e., such that $\nabla Q(\boldsymbol{\beta}) = 0$. To compute the gradient, we can use the property that for Q differentiable, we have

$$Q(\boldsymbol{\beta} + \boldsymbol{h}) = Q(\boldsymbol{\beta}) + \nabla Q(\boldsymbol{\beta})^\top \boldsymbol{h} + o(\|\boldsymbol{h}\|),$$

where $o(\|\boldsymbol{h}\|)$ is a higher order term in $\boldsymbol{h}$. In our case, we have

$$Q(\boldsymbol{\beta} + \boldsymbol{h}) = Q(\boldsymbol{\beta}) + \boldsymbol{h}^\top \boldsymbol{X}^\top \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\,\boldsymbol{h} + \boldsymbol{h}^\top \boldsymbol{X}^\top \boldsymbol{X}\,\boldsymbol{h} - 2\,\boldsymbol{h}^\top \boldsymbol{X}^\top \boldsymbol{y}$$

from which we deduce that $\nabla Q(\boldsymbol{\beta}) = 2\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} - 2\,\boldsymbol{X}^\top \boldsymbol{y}$.

Normal equations

We have thus established that the stationary points of Q satisfy the

Normal equations: $\boxed{\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{0}}$

Given that $\boldsymbol{X}^\top \boldsymbol{X}$ is a positive semi-definite matrix, the curvature of the function is non-negative everywhere and so all stationary points are global minima. The normal equation thus characterize exactly the vectors $\boldsymbol{\beta}$ which are solutions to the linear regression problem.

If $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible, then there is a unique solution to the normal equations and and $\widehat{\boldsymbol{\beta}}$ is given by:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}.$$

Remarks:

- $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible iff the columns of $\boldsymbol{X}$ are linearly independent

$\rightarrow$ they are linearly dependent iff one of them is a linear combination of the others.

- $\boldsymbol{X}^\top \boldsymbol{X}$ is never invertible for $p > n$.

Linear or affine regression?

Compare the linear vs affine functions of $\mathbf{x}$

$$f_{\boldsymbol{\beta}}(\mathbf{x}) = \boldsymbol{\beta}^\top \mathbf{x} \qquad \text{vs} \qquad f_{\boldsymbol{\beta},b}(\mathbf{x}) = \boldsymbol{\beta}^\top \mathbf{x} + b = \widetilde{\boldsymbol{\beta}}^\top \widetilde{\mathbf{x}}$$

With a new definition of the variables

$$\widetilde{\boldsymbol{\beta}} = \begin{bmatrix} \boldsymbol{\beta} \\ b \end{bmatrix} \quad \text{and} \quad \widetilde{\mathbf{x}} = \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix}$$

we can rewrite an affine model in dimension p as a linear model in dimension $p+1$, in which the last column of the design matrix is $\mathbf{1} = (1, \ldots, 1)^\top \in \mathbb{R}^n$.

Exercise: What is the value of $\hat{b}$ if the data is centered?

Gaussian conditional model and linear regression

We decide to model the conditional distribution of Y given X by

$$Y \mid X \sim \mathcal{N}(\boldsymbol{\beta}^\top X + b, \sigma^2)$$

or equivalently $Y = \boldsymbol{\beta}^\top X + \varepsilon \quad$ with $\quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Given a dataset $\mathcal{D}_n = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ we can consider the likelihood of $\boldsymbol{\beta}$ in the conditional model of Y given X and estimate $\boldsymbol{\beta}$ using the maximum likelihood principle.

Likelihood for one pair

$$p(y_i \mid \mathbf{x}_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(y_i - \boldsymbol{\beta}^\top \mathbf{x}_i)^2}{\sigma^2} \right)$$

Negative log-likelihood

$$-\ell(\boldsymbol{\beta}, \sigma^2) = -\sum_{i=1}^{n} \log p(y_i | \mathbf{x}_i) = \frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - \boldsymbol{\beta}^\top \mathbf{x}_i)^2}{\sigma^2}.$$

Gaussian conditional model and linear regression

$$\min_{\sigma^2, \boldsymbol{\beta}} \frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - \boldsymbol{\beta}^\top \mathbf{x}_i)^2}{\sigma^2}$$

The minimization problem in $\boldsymbol{\beta}$

$$\min_{\boldsymbol{\beta}} \frac{1}{2\sigma^2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$$

that we recognize as the usual linear regression.
Optimizing over σ^2, we find:

$$\widehat{\sigma}^2_{\mathsf{MLE}} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{\boldsymbol{\beta}}^\top_{\mathsf{MLE}} \mathbf{x}_i)^2$$

Properties if the data is actually Gaussian

Assume that $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ with

Full column rank *fixed* design: $\operatorname{rank}(\boldsymbol{X}) = p$ (which implies $n \geq p$).

I.i.d. centered **Gaussian** noise: $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$

then

- $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\beta}^*, \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1})$
- $S^2 = \frac{1}{n-p} \|\hat{\boldsymbol{y}} - \boldsymbol{y}\|_2^2 \sim \frac{\sigma^2}{n-p} \chi^2_{n-p}$
- $\widehat{\boldsymbol{\beta}}$ and S^2 are independent

All of these are used for

- t-test and to construct confidence intervals
- Only valid if the data is Gaussian (= model is well-specified)

Simpson's paradox

EE-209 Eléments de Statistiques pour les Data Sciences

Simpson's paradox

We consider the kidney stone (k.s.) recovery example from Charig et al. (1986)[1]
Two treatments are compared:

$T = a$: open surgery

$T = b$: percutaneous neurolithotomy (removes k.s. by a small puncture wound)

The study considered the records of 700 patients:

- Out of 350 patients who received $T = a$, 273 recovered fully (78%)
- Out of 350 patients who received $T = b$, 289 recovered fully (83%).

	Patients with		Overall
	small stones	large stones	
$T = a$	$\mathbf{93}\%\ (81/87)$	$\mathbf{73}\%\ (192/263)$	$78\%\ (273/350)$
$T = b$	$87\%\ (234/270)$	$69\%\ (55/80)$	$\mathbf{83}\%\ (289/350)$

Paradox: it seems that b performs better overall but worse in each subcase!!

[1] Charig, C. R., Webb, D. R., Payne, S. R., & Wickham, J. E. (1986). Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. Br Med J (Clin Res Ed), 292(6524), 879-882.

Simpson's paradox: explanation

	Patients with		Overall
	small stones	large stones	
$T = a$	$\mathbf{93}\%\ (81/87)$	$\mathbf{73}\%\ (192/263)$	$78\%\ (273/350)$
$T = b$	$87\%\ (234/270)$	$69\%\ (55/80)$	$\mathbf{83}\%\ (289/350)$

- $T = a$ chosen often for large k.s. (263/343),
- $T = b$ chosen often for small k.s. (270/357).
- Recovery for large stones is cleary lower than for small stones !
- Large k.s. is a more serious condition overall and $T = a$ is used more often for large k.s..

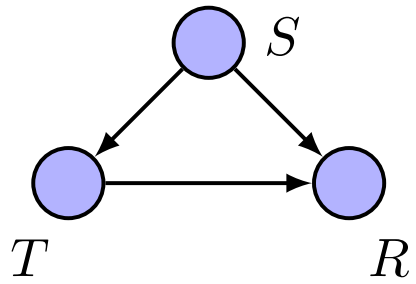
So the overall recovery rate is misleading because the two groups of patients are not comparable, the patients who received $T = a$ tend to be more seriously ill than the patients who received $T = b$...

Simpson's paradox: analysis

S size of the stone (small vs big)

T treatment (a vs b)

R recovery (0 vs 1)



We can write

$$\mathbb{P}(R = r, S = s, T = t) = \mathbb{P}(R = r \mid T = t, S = s)\, \mathbb{P}(T = t \mid S = s)\, \mathbb{P}(S = s).$$

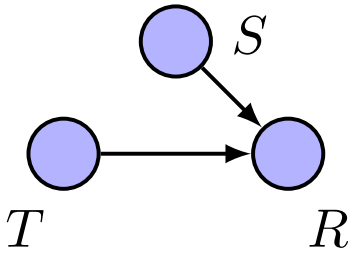
$$\hat{\mathbb{P}}(T = a \mid S = \text{small}) = \frac{87}{357} \approx 0.24, \quad \textit{vs} \quad \hat{\mathbb{P}}(T = a \mid S = \text{large}) = \frac{263}{343} \approx 0.77.$$

The fact that S has an effect on both T and R introduces a confusion between the effect of T on R and the effect of S on R.

S is called a *confounder* or *confounding variable* (*facteur de confusion*/*variable confondante*).

Randomized Trial / Essai randomisé

To avoid confounding it is necessary to avoid that the *treatment* or *intervention* T depends on any other variable that can affect the outcome, i.e., by rendering T independent from these variables.



This is exactly what is done in *randomized clinical trials*:

- Patients are assigned at random to the groups receiving the different treatments.
- If there is a single treatment to test the patient are assigned at random to the treatment or to a *control* group, which will receive a *placebo*. In that case, it is a *randomized controlled trial.*

If this had been done for the study on kidney stone we would have had $\mathbb{P}(T = t \mid S = s) = \mathbb{P}(T = t)$. In other words, the probability of receiving the treatment is independent of whether the patient has small or large kidney stones.

In some cases, using a randomized trial can pose ethical problems and other methods must be applied.

In many cases, only *observational data* is available (economics, political and social sciences).