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

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

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

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 $\{1, \dots, K\}$**.**

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

Some classical discrete random variables

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

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

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

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

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

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

Standard deviation (écart type)

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

Covariance

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

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 $\{1, \dots, 6\}$.

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

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)

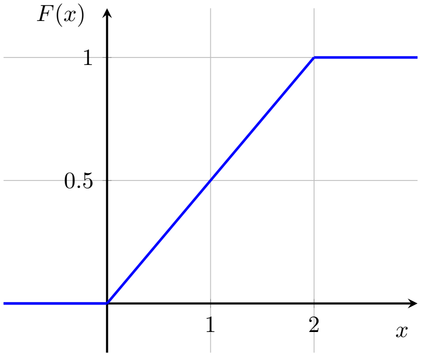
The *cumulative density function* (*fonction de répartition*), or *c.d.f.*, is the function F_X defined by

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

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

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

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

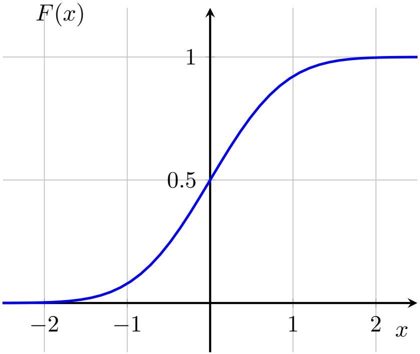


Remark: a cdf must be a non-decreasing function by the second axiom of probability theory, and from 0 to 1 by the third axiom.

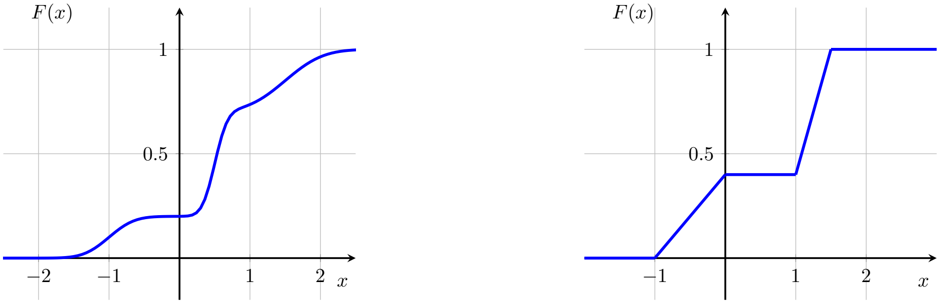
Cumulative density function of the standard normal distribution

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

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

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

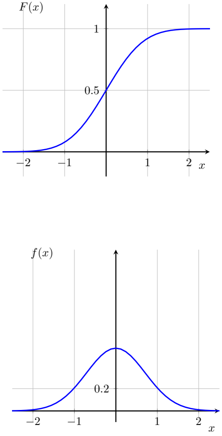
Relations between f and F:

- 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.

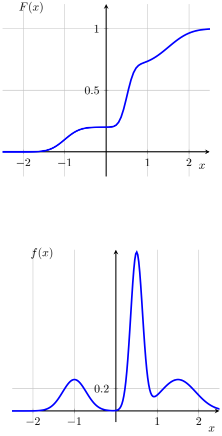
Cdfs *vs* pdfs

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

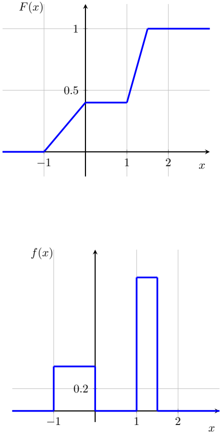
Gaussian



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

$$\mathbb{P}(U \leq u; a, b) = \begin{cases} 0, & \text{if } \quad u < a, \\ \frac{u-a}{b-a}, & \text{if } \quad a \leq u \leq b, \\ 1, & \text{if } \quad u > b. \end{cases}$$

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

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

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

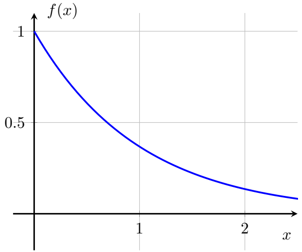
$$p(x;\lambda) = \lambda\, e^{-\lambda x}\, 1_{\{x \geq 0\}}.$$

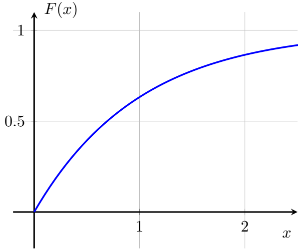
Exponential cdf

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

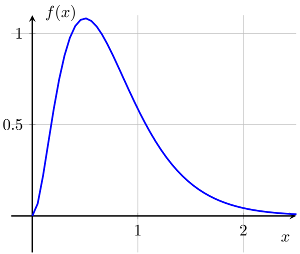
$$p(x; k, \lambda) = \frac{\lambda^k}{\Gamma(k)} x^{k-1} e^{-\lambda x} 1_{\{x \geq 0\}}, \qquad \text{with} \quad \Gamma(k) = \int_0^\infty x^{k-1} e^{-x}\, dx.$$

Remarks:

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

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

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

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

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

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 Var(X), if $X \sim \mathcal{N}(\mu, \sigma^2)$, we can write $X = \mu + \sigma\tilde{X}$ for $\tilde{X} \sim \mathcal{N}(0, 1)$.
- Var$(X) = \sigma^2$ Var$(\tilde{X})$, so we just need to compute 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 $\chi_1^2 \equiv \Gamma(\frac{1}{2}, \frac{1}{2})$ distribution.
- But we proved that if $Y \sim \Gamma(k, \lambda)$ then $\mathbb{E}[X] = \frac{k}{\lambda}$. So

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

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

The Cauchy distribution: a distribution without expectation

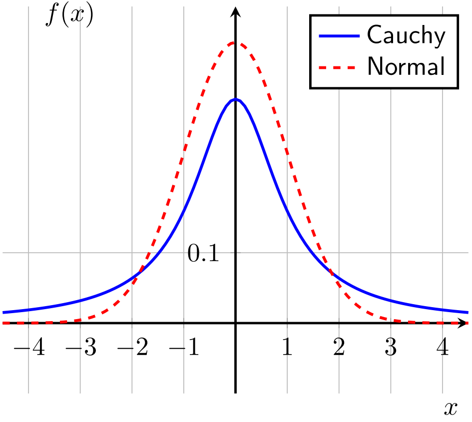
$$p(t) = \frac{1}{\pi(1+t^2)}$$

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

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

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

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 $\quad \mathbb{E}[f(X)] = \int f(x)\, p_X(x)\, dx.$
- 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

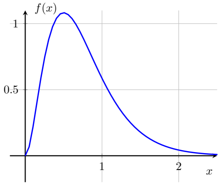
Support:

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

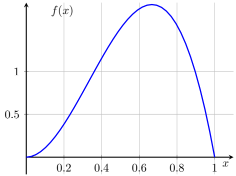
Range:

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

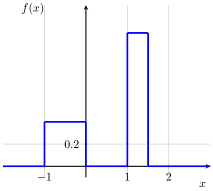
Examples:



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

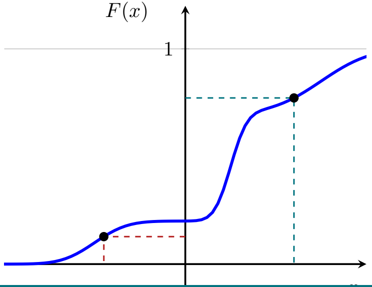
$$\forall \alpha \in (0,1), \quad \exists \text{ a unique } x \in \mathbb{R} \quad \text{such that} \quad F(x) = \alpha.$$

- 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:

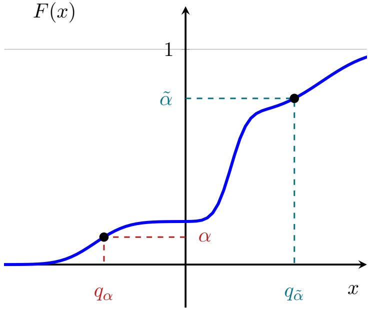
"What is the value of q such that $X < q$ with probability 0.95?"

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.

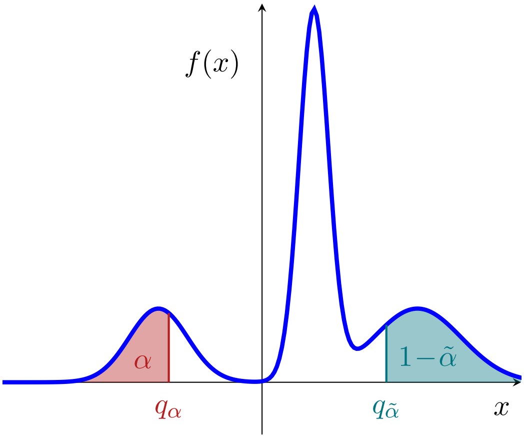
The quantile of level α of X is the unique value q_α such that

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



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 $\quad \mathbb{P}(X \leq m) = 0.5 = \mathbb{P}(X > m)$.
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 $\frac{[\alpha n]}{n}$ 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 $F(x) = \displaystyle\int_0^x \lambda e^{-\lambda t} = \big[-e^{-\lambda t}\big]_0^x = 1 - e^{-\lambda x}$.

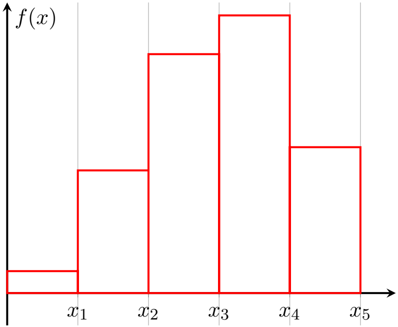
So we can compute $F^{-1}(u) = -\frac{1}{\lambda}\log(1-u)$. This means that if $U \sim \mathcal{U}[0,1]$, then

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

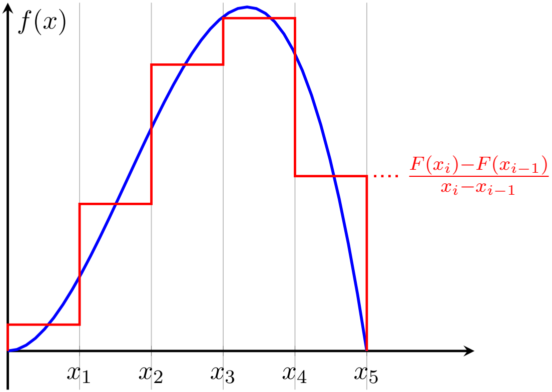
Histograms

- 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 $\{1, \ldots, K\}$ 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 $P_{Z|X}(k|x) = \dfrac{\mathcal{N}(x; \mu_k, 1)\, \pi_k}{p_X(x)}$.

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,

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

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

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

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

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

Properties of the Variance

The following properties can be verified immediately:

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

Finally, we also have $|\mathsf{corr}(X, Y)| \leq 1$. 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:

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

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

Proof:

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

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:

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

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

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

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

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