Probability

σ-algebra: Let Ω be a set. A family $\mathcal{F} \subset 2^{\Omega}$ is a σ-algebra if

- $\emptyset, \Omega \in \mathcal{F}$
- $\forall\, A \in \Omega,\ A^c \in \Omega$
- $\forall\, (A_n) \subset \mathcal{F}$ sequence in $\mathcal{F}$, $\bigcup_{i \in \mathbb{N}} A_i \in \mathcal{F}$

Then, the set $(\Omega, \mathcal{F})$ is called a measurable space

Given a topological space (Ω, τ_Ω), the Borel σ-algebra is the smaller σ-algebra that contains τ_Ω

Measurable function: Let $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ be measurable spaces. A function $f: \Omega_1 \longrightarrow \Omega_2$ is called measurable if $f^{-1}(A) \in \mathcal{F}_1,\ \forall\, A \in \mathcal{F}_2$

Probability measure: A probability measure on $(\Omega, \mathcal{F})$ measurable space is a function $P: \mathcal{F} \longrightarrow [0, +\infty]$ such that

- $P[\emptyset] = 0$
- $P[\Omega] = 1$ (without it, P is just a measure on $(\Omega, \mathcal{F})$)
- Given a sequence $(A_n) \subset \mathcal{F}$ two-by-two distinct, then

$$P\left[\bigcup_{n \in \mathbb{N}} A_n\right] = \sum_{n \in \mathbb{N}} P[A_n]$$

← the infinite $+\infty$ for countable union arises from the divergence of this serie

$(\Omega, \mathcal{F}, P)$ is called probability space

↑ Note that $P(A_1) = P(A_2) + P(A_1 - A_2)\ \ \forall\, A_1, A_2 \in \mathcal{F},\ P(A_i) < \infty$

Lemma: Let $(A_n) \subset \mathcal{F}$ a crescent sequence, ie, $A_1 \subseteq A_2 \subseteq \cdots$. Then

$$P\left[\bigcup_{n \in \mathbb{N}} A_n\right] = \lim_{n \to \infty} P[A_n]$$

and, if $(F_n) \subset \mathcal{F}$ is a decrescent sequence and if $P(F_1) < \infty$, then

$$P\left[\bigcap_{n \in \mathbb{N}} F_n\right] = \lim_{n \to \infty} P[F_n]$$

Integrable function: Given $(\Omega, \mathcal{F}, P)$, consider

$$M^+(\Omega, \mathcal{F}) := \{ f : \Omega \longrightarrow \mathbb{R}_{\geq 0} \cup \{+\infty\} \ : \ f \text{ is measurable} \}.$$

A simple function $\varphi : \Omega \longrightarrow \mathbb{R} \cup \{-\infty, +\infty\}$ is a function with a finite non-null values, ie,

$$\varphi = \sum_{i=1}^{n} a_i \chi_{A_i}$$

← standard representation of φ (it's unique)

← characteristic function of $A_i \in \mathcal{F}$

The integral of φ with respect to P is

$$\int_\Omega \varphi \, dP := \sum_{i=1}^{n} a_i P(A_i) \ ; \quad \int_{A_i} \varphi \, dP = a_i P(A_i)$$

If $f \in M^+(\Omega, \mathcal{F})$, the integral of f with respect to P is

$$\int_\Omega f \, dP = \sup \int \varphi \, dP \ ; \quad \int_A f \, dP = \int_\Omega f \chi_A \, dP \quad \forall \ A \in \mathcal{F}$$

Let $f : \Omega \longrightarrow \mathbb{R} \cup \{-\infty, +\infty\}$ be a measurable function, define

$$\underbrace{f^+(x) = \sup\{f(x), 0\}, \quad f^-(x) := \sup\{-f(x), 0\}, \quad |f| := f^+ - f^-}_{\in M^+(\Omega, \mathcal{F})}$$

The integral of f with respect to P is

$$\int_\Omega f\,dP := \int_\Omega f^+\,dP - \int_\Omega f^-\,dP$$

The set of all the functions with finite integral with respect to P is called $L(P)$

Lemma: Every $f \in L(P)$ induces a probability measure given by

$$\tilde{P}(A) = \left| \frac{\int_A f\,dP}{\int_\Omega f\,dP} \right| \quad \forall\, A \in \mathcal{F}$$

Theorem: A measurable function $f \in L(P) \iff |f| \in L(P)$

In special, if $g \in L(P)$ and $|f| \leq |g| \Rightarrow f \in L(P)$ and

$$\left| \int_\Omega f\,dP \right| \leq \int_\Omega |f|\,dP \leq \int_\Omega |g|\,dP$$

Monotone Convergence Theorem: If (f_n) is a monotone increasing sequence of functions in $M^+(\Omega, \mathcal{F})$ which converges to f, then

$$\int_\Omega f\,dP = \lim_{n\to\infty} \int_\Omega f_n\,dP$$

Fatou's Lemma: If $(f_n) \subset M^+(\Omega, \mathcal{F})$, then

$$\int_\Omega (\liminf f_n)\,dP \leq \liminf \left(\int_\Omega f_n\,dP \right)$$

Dominated Convergence Theorem: Let $(f_n) \in L(P)$ which converges almost everywhere to a function f. If $\exists\ g \in L(P)$ such that $|f_n| \leq g\ \ \forall\ n \in \mathbb{N}$, then $f \in L(P)$ and

$$\int_\Omega f\, dP = \lim_{n\to\infty} \int_\Omega f_n\, dP$$

> **Useful bounds**
> - $|e^t - 1| \leq |t| e^{|t|} \quad \forall\ t \in \mathbb{R}$
> - $|t| \leq e^{|t|} \quad \forall\ t \in \mathbb{R}$

Uniform convergence: The sequence (f_n) converges uniformly to f if, for every $\epsilon > 0$, $\exists\ N(\epsilon) > 0$ such that

↑ $N(\epsilon, x) = N(\epsilon)$

$n > N(\epsilon)$ and for $x \in \Omega \Rightarrow |f_n(x) - f(x)| < \epsilon$

Radon-Nikodym Theorem: Let λ and μ be σ-finite measures defined on $\mathcal{F}$ and suppose that $\lambda \ll \mu$. Then, $\exists\ \frac{d\lambda}{d\mu} \in M^+(\Omega, \mathcal{F})$ (not necessarily integrable) function such that

↑ $\lambda \ll \mu \Leftrightarrow \forall\ A \in \mathcal{F}$ such that $\mu(A) = 0$, it implies that $\lambda(A) = 0$

(λ is absolutely continuous with resp. to μ)

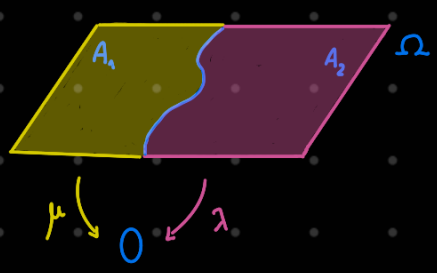
$$\lambda(A) = \int_A \frac{d\lambda}{d\mu}\, d\mu, \quad \forall\ A \in \mathcal{F},$$

and $\frac{d\lambda}{d\mu}$ is unique up to sets of measure μ null

Lebesgue Decomposition Theorem: Let λ and μ be σ-finite measures on $(\Omega, \mathcal{F})$. Then, $\exists$ unique measures $\lambda_1 \ll \mu$, $\lambda_2 \perp \mu$ such that $\lambda = \lambda_1 + \lambda_2$

↑ $\lambda \perp \mu \Leftrightarrow \exists\ A_1, A_2 \in \mathcal{F}$, $A_1 \cap A_2 = \emptyset$, such that
- $A_1 \cup A_2 = \Omega$
- $\mu(A_1) = 0 = \lambda(A_2)$

(λ and μ are mutually singular)



μ-equivalence class: Let $f \in L(\mu)$. Then,

$$[f] := \{ g \in L(\mu) : f(x) = g(x)\ \ \forall\ x \in A \text{ such that } A \in \mathcal{F} \text{ and } \mu(A) \neq 0 \}$$

L^p - spaces: For $1 \leq p < \infty$,

$$L^p(\mu) = \left\{ [f] : f \text{ is } \mu\text{-measurable and } \left[\int_\Omega |f|^p d\mu \right]^{1/p} \right\}$$

are the complet normed vector spaces with the norm

$$\|f\|_p = \left[\int_\Omega |f|^p \, d\mu \right]^{1/p}$$ ← the representant in $[f]$ doesn't matter

L^∞ - spaces: $L^\infty(\mu)$ is the set of all μ-equivalent classes which are bounded up to sets of measure zero, i.e,

$$L^\infty(\mu) := \left\{ [f] : f \text{ is measurable and } \forall A \in \mathcal{F} \text{ such that } \mu(A) \neq 0, \ \sup \{ |f(x)| : x \in A \} < \infty \right\}$$

This space is a normed vector space. In fact, for $A \in \mathcal{F}$ such that $\mu(A) = 0$, for $f \in L^\infty(\mu)$ define

$$S(A) := \sup \{ |f(x)| : x \notin A \}$$
$$\|f\|_\infty := \inf \{ S(A) : A \in \mathcal{F}, \mu(A) \neq 0 \}$$ ← lower number that limits all sets with measure non-null

Bounded linear function: A linear functional on $L^p(\mu)$, for $1 \leq p \leq \infty$, is a function $G : L^p(\mu) \longrightarrow \mathbb{R}$ such that

$$G(af + bg) = a\,G(f) + b\,G(g) \quad \forall\, a, b \in \mathbb{R}, \forall\, f, g \in L^p(\mu).$$

G is bounded if $\exists$ M constant such that $|G(f)| \leq M \|f\|_p$. In this case, the norm or bound of G is

$$\|G\| := \sup \{ |G(f)| : f \in L^p(\mu) \text{ such that } \|f\|_p \leq 1 \}$$

Riesz Representation Theorem: If G is a bounded linear function on $L^p(\mu)$ for $1<p<\infty$, $\exists\, g \in L^q(\mu)$, where $\frac{1}{p}+\frac{1}{q}=1$, such that

$$G(f) = \int_\Omega f g \, d\mu \quad \forall\, f \in L^p(\mu)$$

← We represent G as the integral of a function

and $\|G\| = \|g\|_q$.

For $p \in \{1, \infty\}$, if μ is σ-finite measure and G is a bounded linear functional on $L(\mu)$, then $\exists\, g \in L^\infty(\mu)$ such that

$$G(f) = \int_\Omega f g \, d\mu \quad \forall\, f \in L(\mu)$$

Moreover, $\|G\| = \|g\|_\infty$ and $g \geq 0$ if G is positive, i.e, $\mathrm{Im}(G) \subseteq \mathbb{R}_{\geq 0}$.

Algebra: A family $\tilde{\mathcal{F}} \subset 2^\Omega$ is an algebra of subsets of Ω if

- $\phi, \Omega \in \tilde{\mathcal{F}}$
- $A \in \tilde{\mathcal{F}} \Rightarrow A^c \in \tilde{\mathcal{F}}$
- $A_1, \dots, A_n \in \tilde{\mathcal{F}} \Rightarrow \bigcup_{i=1}^n A_i \in \tilde{\mathcal{F}}$ ← It distinguishes an algebra of a σ-algebra, because here the union is finite

Then, a measure on $\tilde{\mathcal{F}}$ is a function $\mu : \tilde{\mathcal{F}} \longrightarrow [0, \infty]$ such that

- $\mu(\phi) = 0$
- If $(A_n)_{n\in\mathbb{N}} \subset \tilde{\mathcal{F}}$ is a sequence of two-by-two disjoint sets such that $\bigcup_{i\in\mathbb{N}} A_i \in \tilde{\mathcal{F}}$, then

$$\mu\left(\bigcup_{i\in\mathbb{N}} A_i\right) = \sum_{i\in\mathbb{N}} \mu(A_i)$$

Outer measure: Let $H \in 2^\Omega$, we define the outer measure

$$\mu^*(H) = \inf\left\{ \sum_{i\in\mathbb{N}} \mu(A_i) : (A_i)_{i\in\mathbb{N}} \subset \tilde{\mathcal{F}} \text{ such that } H \subset \bigcup_{i\in\mathbb{N}} A_i \right\}$$

↳ It's subadditive: $(H_n)_{n\in\mathbb{N}} \subset 2^\Omega$, $\mu^*\left(\bigcup_{i\in\mathbb{N}} H_i\right) \leq \sum_{i\in\mathbb{N}} \mu^*(H_i)$

μ^*-measurable set: A set $H \subset \Omega$ is called μ^*-measurable if

$$\mu^*(A) = \mu^*(A \cap H) + \mu^*(A \setminus H) \quad \forall\, A \subset \Omega$$

The set of μ^*-measurable sets is called $\mathcal{F}^*$

Carathéodory Extension Theorem: A collection $\mathcal{F}^*$ of all μ^* measurable sets is a σ-algebra containing $\tilde{\mathcal{F}}$. Moreover, if $(A_n)_{n \in \mathbb{N}} \subset \mathcal{F}^*$ is a sequence of two-by-two disjoint sets, then

$$\mu^*\left(\bigcup_{i \in \mathbb{N}} A_i\right) = \sum_{i \in \mathbb{N}} \mu^*(A_i)$$

Hahn Extension Theorem: Suppose that μ is a σ-finite measure on an algebra $\tilde{\mathcal{F}}$. Then, there exists a unique extension of μ to a measure on $\mathcal{F}^*$

Set-up for probability theory:

$(\Omega, \mathcal{F}, P)$

- $\mathcal{F}$: σ-algebra (space of events)
- Ω: abstract space (sample space)

$P: \mathcal{F} \longrightarrow \mathbb{R}_+ \cup \{\infty\}$

- $P(\emptyset) = 0$
- $\{A_j : j \geq 1\}$, $A_j \cap A_k = \emptyset$ $\quad P\left[\bigcup_j A_j\right] = \sum_j A_j$
- $P(\Omega) = 1$ (without it, P is a measure)

Random variable: $X : \Omega \longrightarrow \mathbb{R}$

With $(\mathbb{R}, \mathcal{B})$, X is measurable : $X^{-1}(A) \in \mathcal{F} \quad \forall\, A \in \mathcal{B}$

Probability distribution measure: X has a probability distribution measure μ (in the Borel σ-algebra of $\mathbb{R}$) if

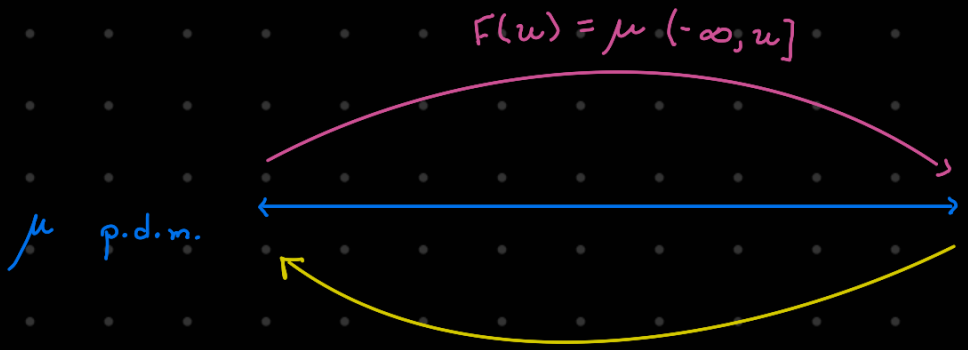
$$\boxed{\mu = X_* P} \quad \text{ie} \quad \mu(A) = P[X^{-1}(A)] \overset{\text{notation}}{=} P[X \in A] \quad \forall\, A \in \mathcal{B}$$

Distribution function: The distribution function associated to X is

$$F_X : \mathbb{R} \longrightarrow [0,1]$$
$$x \longmapsto P[X^{-1}((-\infty, x])] = P[X \leq x]$$

↳ Properties:
- F_X is monotone: $x \leq y \Rightarrow F_X(x) \leq F_X(y)$
- F_X is right-continuous, left-limit exists
- $\lim_{x \to \infty} F_X(x) = 1$, $\lim_{x \to -\infty} F_X(x) = 0$

There is an injection between the probability distribution measures and the distribution functions



$S = \{(a,b] : -\infty \leq a < b \leq +\infty\}$ is semi-algebra over $\mathbb{R}$
- $\emptyset, \mathbb{R} \in S$
- $\forall A, B \in S \Rightarrow A \cap B \in S$
- $\forall A, B \in S \Rightarrow \exists \{C_i\}_{i=1}^{N}$ pairwise-disjoint such that $A \setminus B = \bigcup_{i=1}^{N} C_i$

μ is σ-finite: $\exists \{C_i\}_{i \in \mathbb{N}} \subset \mathcal{B}$ such that
- $\mu(C_i) < \infty$
- $\sum_i C_i = \mathbb{R}$

Discrete random variables:

X is a discrete random variable $\iff \exists\ A \subset \mathbb{R}$ countable such that $X_* P(A) = 1$

Discrete distribution function:

F is a discrete d.f. $\iff \exists\ A = \{x_j\}_{j\geq 1} \subset \mathbb{R}$ and $\exists\ \{p_j\} \subset [0,1]$ such that

$$\sum_{j\geq 1} p_j = 1, \quad F(x) = \sum_{\substack{j\geq 1 \\ x_j \leq x}} p_j$$

summing of j such that $x_j \leq x$

Absolute continuous density function: F is absolutely continuous if $\exists\ f: \mathbb{R} \longrightarrow \mathbb{R}_+$ measurable such that $F(b) - F(a) = \int_a^b f(t)\,dt \quad \forall\ a < b \in \mathbb{R}$

Remark: F is diferenticable almost everywhere and $F' = f$

Singular continuous density function: F is singular $\iff$ $\begin{cases} F \text{ is continuous} \\ F' = 0 \text{ almost everywhere} \end{cases}$

Theorem: Let F be a distribution function. Then, F can be written as a convex combination of a discrete, an absolute continuous an a singular density function.

$\exists\ \alpha, \beta \geq 0$, $\alpha + \beta \leq 1$, and $\exists\ F_d, F_{ac}, F_s$ such that

$$F = \alpha F_d + \beta F_{ac} + (1 - \alpha - \beta) F_s$$

Expectation: Let $X: \Omega \longrightarrow \mathbb{R}$ a random variable such that

$$\int_\Omega |X|\,dP < \infty, \longleftarrow \text{ie, } X \text{ is integrable}$$

The same than suppose $X \in L(\Omega)$

then, we define

$$\mathbb{E}[X] = \int_{\Omega} X \, dP$$

Proposition: If $X \geq 0$ (i.e., $X(\Omega) \subset \mathbb{R}_{\geq 0}$), then

$$\sum_{n \in \mathbb{N}} P[X \geq n] \leq \mathbb{E}[X] \leq 1 + \sum_{n \in \mathbb{N}} P[X \geq n]$$

$$\sum_{n \in \mathbb{N}} P\left[X^{-1}[n, +\infty)\right] \leq \mathbb{E}[X] \leq 1 + \sum_{n \in \mathbb{N}} P\left[X^{-1}[n, +\infty)\right]$$

Jensen inequality: Let $\varphi : \mathbb{R} \longrightarrow \mathbb{R}$ be a convex function, ie,

$$\varphi(\theta x + (1-\theta) y) \leq \theta \varphi(x) + (1-\theta) \varphi(y) \quad \forall \ \theta \in [0,1], \ \forall \ x, y \in \mathbb{R}.$$

Let X be a measurable random variable, ie, $\mathbb{E}[|X|] < \infty$. Assume that $\mathbb{E}[|\varphi(X)|] < \infty$. Then,

$$\varphi\left(\mathbb{E}[X]\right) \leq \mathbb{E}[\varphi(X)]$$

Chebyshev inequality: Let $X \geq 0$ be a random variable. Let $f : \mathbb{R}_{\geq 0} \longrightarrow \mathbb{R}_{>0}$ a monotonic non-decreasing function ($x \leq y \Rightarrow f(x) \leq f(y)$). Then,

$$P[X \geq a] \leq \frac{1}{f(a)} \mathbb{E}[f(X)] \qquad P\left[X^{-1}[a, +\infty)\right] \leq \frac{1}{f(a)} \mathbb{E}[f(X)]$$

↑ can be ∞

Independent events: Let $A \in \mathcal{F}$. A is called an *event*.

Let $A_1, A_2, \ldots, A_N \in \mathcal{F}$. They are called independent if, for all $\{n_1, n_2, \ldots, n_r\} \subseteq \{1, \ldots, N\}$, $n_j \neq n_k$ for $j \neq k$, holds

$$P\left[\bigcap_{k=1}^{r} A_{n_k}\right] = \prod_{k=1}^{r} P[A_{n_k}]$$

Independent random variables: Let $X_1, \dots, X_N$ be random variables. They are called independent if, $\forall\ A_1, \dots, A_N \in \mathcal{B}$, holds

$$P\left[\bigcap_{j=1}^{N} \{X_j \in A_j\}\right] = \prod_{j=1}^{N} P[X_j \in A_j]$$

$$P\left[\bigcap_{j=1}^{N} X_j^{-1}(A_j)\right] = \prod_{j=1}^{N} P[X_j^{-1}(A_j)]$$

Independent family of random variables: Let $\{X_\alpha : \alpha \in I\}$ a family of random variables. We say this family is independent if, $\forall\ N \in \mathbb{N}$ and $\forall\ \{\alpha_1, \dots, \alpha_N\} \subset I$ with $\alpha_j \neq \alpha_k$ for $j \neq k$ (ie, for all finite subfamily of $\{X_\alpha\}$), $X_{\alpha_1}, \dots, X_{\alpha_N}$ are independent.

Remark: Any subfamily of an independent family is independent

Random vector: Consider $\mathbb{R}^n$ with its σ-algebra $\mathcal{B}^n$. A n-dimensional random vector is a measurable function $X: \Omega \longrightarrow \mathbb{R}^n$, ie, $\forall\ A \in \mathcal{B}^n$, $X^{-1}(A) \in \mathcal{I}$
Basically, it's a vector of random variables: $X = (X_1, \dots, X_n)$
Its distribution function is

↑ this entrance takes the first coordinate of A

$$F_X : \mathbb{R}^n \longrightarrow [0,1]$$
$$(x_1, \dots, x_n) \longmapsto P[X_1 \leq x_1, \dots, X_n \leq x_n] = P\left[\bigcap_{j=1}^{n} \{X_j \leq x_j\}\right]$$

And its probability distribution measure is

$$\mu_X = X_* P \text{, ie, } \mu(A) = P[X^{-1}(A)] = P[X \in A] \quad \forall\ A \in \mathcal{B}^n$$

Lemma: Let $X=(X_1,\dots,X_n)$ a random vector. Each X_j is associated to a F_{X_j}

$$X_1,\dots,X_n \text{ are independent} \iff F(x_1,\dots,x_n)=\prod_{j=1}^{n} F_{X_j}(x_j) \quad \forall\, x=(x_1,\dots,x_n)\in\mathbb{R}^n$$

$\Updownarrow$ By the identification between F and μ

Lemma: Let $X=(X_1,\dots,X_n)$ a random vector. Each X_j is associated to a μ_{X_j}

$$X_1,\dots,X_n \text{ are independent} \iff \mu_X(A_1\times\cdots\times A_n)=\prod_{j=1}^{n}\mu_{X_j}(A_j) \quad \forall\, A_1\times\cdots\times A_n\in\mathcal{B}^n$$

Theorem: Independence is preserved by measurable functions. Given $X_1,\dots,X_n$ random variables and $f_1,\dots,f_n:\mathbb{R}\to\mathbb{R}$ measurable functions, then $f_1(X_1),\dots,f_n(X_n)$ are independent

↑ we can take functions from more than one random variable and it still holds

Basically, independence is preserved by measurable functions

Theorem: Let X,Y independent random variables such that

- $\mathbb{E}[|X|],\ \mathbb{E}[|Y|]<\infty$

Then,

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

Identically distributed random variables: Let $X_1,\dots,X_n$ be random variables. They are identically distributed if

$$\mu_{X_1}=\mu_{X_2}=\cdots=\mu_{X_n}=(X_1)_*P$$

Weak law of large numbers: Let $X_1, \dots, X_n$ iid random variables. Suppose that

- $\mathbb{E}[X_1^2] < \infty$ ← as they are identically distributed, it's true for the others

Then,

$$\lim_{n\to\infty} \mathbb{E}\left[\left(\frac{X_1+\dots+X_n}{n} - \mathbb{E}[X_1]\right)^2\right] = 0$$

← The empirical average of $\{X_i\}_{i=1}^n$ converges in L^2 to $\mathbb{E}[X_1]$ for n large enough

More general, for a sequence $\{X_i\}_{i=1}^n$ of iid random variables, then

$$\lim_{n\to\infty} X_n \longrightarrow 0 \Longrightarrow \lim_{n\to\infty} \mathbb{E}[|X_n|^\alpha] = 0 \quad \forall \ \alpha > 0$$

(0: function zero)

Consequently, $\forall\ \varepsilon > 0$, $\lim_{n\to\infty} P[|X_n| > \varepsilon] = 0$

and, specially,

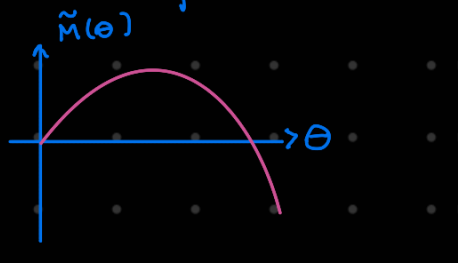
$$\lim_{n\to\infty} P\left[\left|\frac{X_1+\dots+X_n}{n} - \mathbb{E}[X_1]\right| > \varepsilon\right] = 0$$

Upper bound for large deviations: Let $X_1, \dots, X_n$ be iid random variables and $\theta \in \mathbb{R}$ such that $\mathbb{E}[e^{\theta X_1}] < \infty$. Define $M(\theta) = \log \mathbb{E}[e^{\theta X_1}]$. Then, $\forall\ a > 0$,

$$P\left[\frac{X_1+\dots+X_n}{n} - m \geq a\right] \leq e^{n[-(a+\mathbb{E}[X_1])\theta + M(\theta)]}$$

$$\leq e^{-n I(a)},$$

where $I(a) = \sup_{\theta > 0} \{a\theta - \tilde{M}(\theta)\}$ is the Legendre transformation for the convex function $\tilde{M}(\theta) = M(\theta) - \mathbb{E}[X_1]\theta$



$\mathbb{E}_\theta[X]$: When someone writes $\mathbb{E}_\theta$, he is parametrizing $\mathbb{E}$ changing the measure P by a positive and normalized factor e^θ, $\theta \in \mathbb{R}$:

$$\mathbb{E}_\theta[f(x)] = \frac{\mathbb{E}[f(x)e^{\theta x}]}{\mathbb{E}[e^{\theta x}]} = \int_\Omega \frac{f(x)e^{\theta x}\,dP}{\int_\Omega e^{\theta x}\,dP},$$

for f measurable, continuous and bounded

If we have a family of densities $\{p_\theta : \theta \in \Theta \subset \mathbb{R}^n\}$, essa notação pode significar $E_\theta[X] = \int_\Omega p_\theta(x) \log p_\theta(x)\,dx$

$$\mathbb{E}[e^x] < \infty \Rightarrow \mathbb{E}[x^2] < \infty \Rightarrow \mathbb{E}[|x|] < \infty$$

Lower bound for large deviations: Given $X_1, \dots, X_n$ iid random variables, for all open $G \subset \mathbb{R}$,

$$\liminf_{n \to \infty} \left(\frac{1}{n} \log P\left[\frac{X_1 + \cdots + X_n}{n} \in G \right] \right) \geq -\inf_{\alpha \in G} I(\alpha),$$

where $I(\alpha) = \sup_{\lambda \in \mathbb{R}} \{\lambda\alpha - M(\lambda)\}$

Almost sure convergence: A sequence of random variables (X_n) converges almost surely to $X \Leftrightarrow \exists\, A \in \mathcal{F}$, $P[A] = 1$, such that, $\forall\, \omega \in A$, $X_n(\omega) \longrightarrow X(\omega)$, ie, $\forall\, \omega \in A$,

$$X_n(\omega) \longrightarrow X(\omega) \Leftrightarrow \forall\, K > 1,\ \exists\, n \geq 1 \text{ such that } \forall\, m \geq n,\ |X_m(\omega) - X(\omega)| \leq 1/K$$

It's weaker than $\forall\, \omega \in \Omega,\ X_n(\omega) \longrightarrow X(\omega)$

Convergence in probability: A sequence of random variables (X_n) converges in probability to X, ie,

$$X_n \xrightarrow{P} X \Leftrightarrow \forall\, \varepsilon > 0,\ \lim_{n \to \infty} P[|X_n - X| > \varepsilon] = 0$$

Lemma: $X_n \xrightarrow{\text{almost sure}} X \Rightarrow X_n \xrightarrow{P} X$

Convergence almost surely is stronger than convergence in probability

Lemma: $X_n \xrightarrow{P} X \Rightarrow \exists$ subsequence $(X_{n_k}) \subset (X_n)$ such that $X_{n_k} \xrightarrow{\text{almost sure}} X$

Convergence in L^p: For $0 < p < \infty$, given a sequence (X_n) of random variables,

$$X_n \xrightarrow{L^p} X \iff \lim_{n \to \infty} \mathbb{E}[|X_n - X|^p] = 0$$

Lemma: $X_n \xrightarrow{L^p} X$ for $0 < p < \infty \Rightarrow X_n \xrightarrow{P} X$

Theorem: Let (X_n) be a sequence of random variables. Given $0 < p < \infty$, suppose that

- $X_n \xrightarrow{P} X$
- $\exists$ Y random variable bounding X_n: $|X_n| \leq Y \ \forall \ n \in \mathbb{N}$
- $\mathbb{E}[Y^p] < \infty$

Then, $X_n \xrightarrow{L^p} X$

Weak convergence: Given a sequence (μ_n) of probability distribution measures, it converges weakly to μ, i.e,

$$\mu_n \xrightarrow{d} \mu \iff \forall \ (a,b] \text{ such that } \mu(\{a\}) = 0 = \mu(\{b\}),$$
$$\mu_n((a,b]) \longrightarrow \mu((a,b])$$

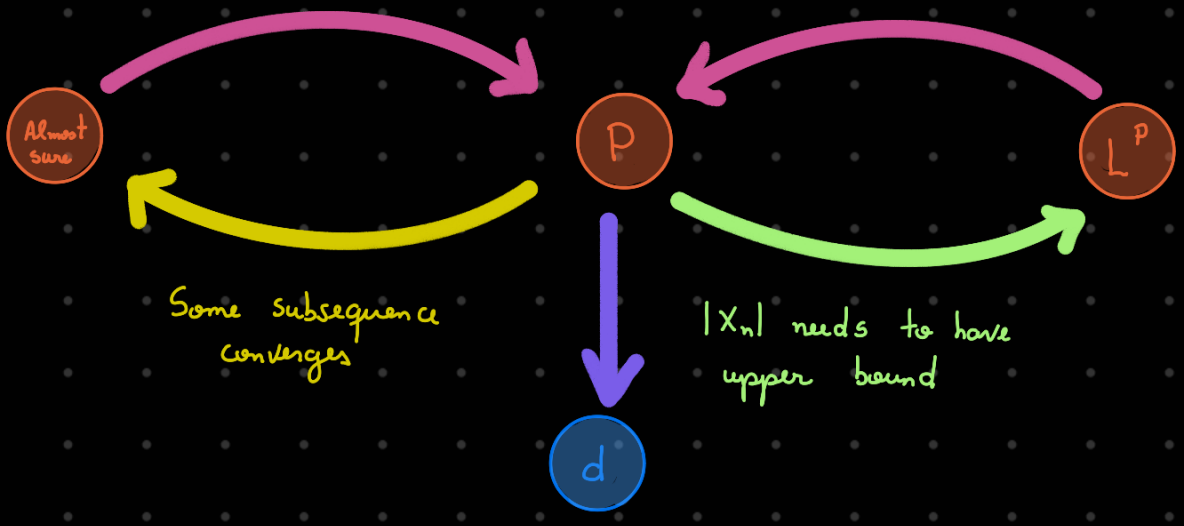
$$\iff F_n(x) \longrightarrow F(x) \ \forall \ x \in \mathbb{R}, \text{ where } F_n, F \text{ are the distribution functions for } \mu_n \text{ and } \mu, \text{ resp.}$$

the definitions are equivalents by the identification $\mu \leftrightarrow F$

this hypothesis is to deal with the δ distributions, because without this, each δ distribution would converge to each others

Convergence in distribution: (X_n) sequence of random variables converges in distribution to X if its sequence (F_{X_n}) of distribution functions converges to F_X

$$X_n \xrightarrow{d} X \iff F_{X_n} \longrightarrow F_X$$



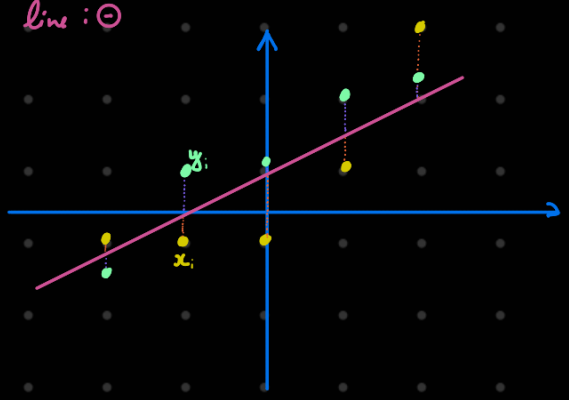
Least squared method: Given two random variables $X, Y : \Omega \longrightarrow \mathbb{R}$, we estimate the intercept β_0 and the slope β_1 parameters as

$$(\beta_0, \beta_1) := \arg\min \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

and the error term of this estimative is defined as

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i \leftarrow \text{with sign} \begin{cases} \text{upper the line} : \oplus \\ \text{under the line} : \ominus \end{cases}$$

This method minimizes the sum of the areas of the squares which side lengths is given by the errors ε_i. The error of each point is given by the length of the vertical line conecting it and the line



It's not a good method when we have different degrees of numerical readings. For example, for 4 observations, $y_1 \in [4,6)$, $y_2 \in [3,4)$, $y_3 \in [5,\infty)$, $y_4 \in [-\infty, 3)$ obtained from a normal process which μ we want to estimate

Y has distribution measure $N(\mu,\sigma)$

1º problem: Interval data: With intervals we can not compute ϵ_i because y_i is not known

2º problem: There is no independent variable X in this problem to estimate μ.

Joint probability density: Let $X: \Omega_1 \to \mathbb{R}$, $Y: \Omega_2 \to \mathbb{R}$ be random variables. Having said that, we obtain a new measure space $(\Omega_1 \times \Omega_2, \Sigma^*, P^*)$, where Σ^* is the product σ-algebra and $P^* := P_1 \otimes P_2$, ie, $P^*(x,y) = P_1(x) P_2(y)$.

The joint probability density is the function $\rho_{x,y} : \Omega_1 \times \Omega_2 \longrightarrow \mathbb{R}^2$

$$P^*((x,y) \in A) = \int_A \rho_{x,y}(x,y)\, dx\, dy \quad \forall\ A \in \mathcal{B}_{\mathbb{R}^2},$$

↳ It implies $\rho_{x,y} \geq 0$ and $\int_{\mathbb{R}^2} \rho_{x,y}\, dx\, dy$

ie,

$$(X \otimes Y)^*(P)(A) = P((X \otimes Y)^{-1}(A)) = \int_A \rho_{x,y}(x,y)\, dx\, dy \quad \forall\ A \in \mathcal{B}_{\mathbb{R}^2}$$

↳ If X, Y are independent, $\rho_{x,y} = \rho_x \cdot \rho_y$, where the probability density ρ_x and ρ_y are defined analogously

Note it's different from probability distribution measures. In that case, we define a measure on $(\mathbb{R}, \mathcal{B})$ in terms of a probability measure P. Here, we have two measures (P on Σ and $\mu := dx\, dy$ on $\mathcal{B}_{\mathbb{R}^2}$) and we are defining a density function

↳ It's just the Radon-Nykodin derivative

Marginal density: Given $\rho_{x,y}$ joint distribution density, we define the marginal density with respect to X and Y, respectively, as

$$\rho_x(x) = \int_{\mathbb{R}} \rho_{x,y}(x,y)\, dy, \quad \rho_y(y) = \int_{\mathbb{R}} \rho_{x,y}(x)\, dx$$

Likelihood function: Let $X_1, \dots, X_n$ be random variables. Let Θ a set called the parameter space. Fixed $\theta \in \Theta$, let $\rho_{X_1,\dots,X_n;\theta} = \rho(x_1,\dots,x_n|\theta)$ be the joint probability density of $X_1,\dots,X_n$. Given $X_1 = x_1, \dots, X_n = x_n$, the likelihood function is defined as

(↘ we start with some model which gives the dependance in θ.)

$$L : \Theta \longrightarrow \mathbb{R}$$

$$\theta \longmapsto L(\theta|x_1,\dots,x_n) := \rho(x_1,\dots,x_n|\theta)$$

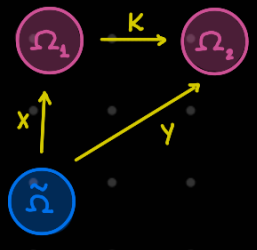
If, for a sample $x_1,\dots,x_n$, $L(\theta_1|x_1,\dots,x_n) > L(\theta_2|x_1,\dots,x_n)$, then $x_1,\dots,x_n$ is more likely if $\theta=\theta_1$ than if $\theta=\theta_2$

Given $X_1,\dots,X_n$ random variables, $x_1,\dots,x_n$ are ↗ Points such that $x_i \in \mathrm{Im}(X_i)$, $1 \le i \le n$

Likelihood Principle: If $x_1,\dots,x_n$ and $y_1,\dots,y_m$ are two sample points such that $L(\theta|x_1,\dots,x_n) \propto L(\theta|y_1,\dots,y_m)$ $\forall \theta \in \Theta$, then $\{x_1,\dots,x_n\} = \{y_1,\dots,y_m\}$

Infinitely divisible random variable: A random variable X is infinitely divisible if, $\forall n \in \mathbb{N}$, $\exists$ $X_1,\dots,X_n$ iid random variables such that $X = \sum_{i=1}^{n} X_i$

Sufficient statistics: Let Ω_1, Ω_2 be finite non-empty sets and let $K: \Omega_1 \longrightarrow \Omega_2$ a statistic which transforms the random variable X to $Y = K(X)$. Then, given a density $p(x|\theta)$ for X, it induces a density for Y: $q(y|\theta) = \sum_{x \in K^{-1}(y)} p(x|\theta)$



If the quantity below doesn't depend on θ, we say K is a suffient statistics for X

$$r(x|\theta) = \frac{p(x|\theta)}{q(K(x)|\theta)}$$

In this case: $p(x|\theta) = r(x)\, q(K(x)|\theta)$ and K contains all de dependense of p in θ and it preserves all the informations about θ.

Minimal sufficient statistics: A sufficient statistical K for X is minimal if any other sufficient statistics for X is a function of K, ie, if $\bar{K}$ is a suffient statistics for X, then $\exists$ $f: \Omega_2 \to \Omega_2$ measurable function such that $\bar{K} = f \circ K$ a.e. with respect to the underlied measure

K is minimal $\Leftrightarrow$ $\left(\dfrac{p(x,\theta)}{p(y,\theta)} \text{ is independent of } \theta \iff K(x) = K(y) \quad \forall\, x,y \in \Omega_1 \right)$

Neyman-Fisher factorization criterion: Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a family of probability measures on a measurable space $(\Omega, \mathcal{F})$ absolutely continuous with respect to a σ-finite measure μ. Let T a statistic between $(\Omega, \mathcal{F}, P_\theta)$ and another measurable space $(\tilde{\Omega}, \tilde{\mathcal{F}})$

T is sufficient with respect to $\mathcal{P}$ $\iff$ $\forall\, \theta \in \Theta$, $\forall\, x \in \Omega$, $\exists\, g_\theta : \tilde{\Omega} \to \mathbb{R}_{\geq 0}$

$\forall\, x \in \Omega$, $\exists\, h : \Omega \to \mathbb{R}_{\geq}$, $h(x) \neq 0$ almost everywhere

(both measurable function and in $L(\Omega, \mu)$)

such that

$$\frac{dP_\theta}{d\mu}(x) = p(x|\theta) = g_\theta(T(x)) \cdot h(x) \quad \text{almost everywhere}$$

Random matrix: Let $\{X_i : \Omega \longrightarrow \mathbb{R}^n : 1 \leq i \leq p\}$ be a family of random vectors. A random matrix of dimension $n \times p$ is the matrix

$$M := [X_1\ X_2 \cdots X_p] = \begin{bmatrix} x_1^1 & x_2^1 & \cdots & x_p^1 \\ x_1^2 & x_2^2 & \cdots & x_p^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^n & x_2^n & \cdots & x_p^n \end{bmatrix}$$

Multivariate normal distribution: Let X be a n-dimensional random vector $X = (X_1, \ldots, X_n)$. The n-variate normal distribution is the probability measure $\mathcal{N}_n(\mu, \Sigma)$ such that the probability density is given by

$$p_X(\underbrace{x_1, \cdots, x_n}_{\vec{x}}) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2} (\vec{x} - \mu)^T \Sigma^{-1} (\vec{x} - \mu) \right),$$

omitted the parameters. Precisely, it should be $p(x_1, \ldots, x_n | \mu, \Sigma)$

Notice it's well defined because $\mathcal{N}_n(\mu,\Sigma)(X^{-1}(A)) = \int_A \rho_x(x)\,dx \quad \forall A \in \mathcal{B}_{\mathbb{R}^n}$

If X follows $\mathcal{N}_n(\mu,\Sigma)$, so holds

$$\mu = \mathbb{E}[X]$$
$$\Sigma_{ij} = \mathbb{E}[(X_i-\mu_i)(X_j-\mu_j)] = \mathrm{Cov}(X_i,X_j)$$

The inverse of Σ is called precision matrix: $Q = \Sigma^{-1}$

Theorem (Characterization of normal ($X \sim \mathcal{N}_k(\mu,\Sigma)$) random vector): Let $X = (X_1,\dots,X_n)$ a random vector. Then,

$X \sim \mathcal{N}_n(\mu,\Sigma)$ $\iff$

$\exists\, Z = (Z_1,\dots,Z_k)$ iid random vector
$\exists\, \mu = (\mu_1,\dots,\mu_\ell) \in \mathbb{R}^\ell$
$\exists\, A \in M_{\ell\times k}(\mathbb{R})$
such that
$X = AZ + \mu$, $Z_i \sim \mathcal{N}(0,1)$ $\forall\, 1\le i\le k$, $AA^T = \Sigma$

(X has the standard probability measure up to an affine map)

χ^2-distribution: Let $X_1,\dots,X_n$ be random variables following the standard normal distribution, ie,

$$\mathcal{N}(0,1)(X_i^{-1}(A)) = \mu(A) \quad \forall A \in \mathcal{B}_{\mathbb{R}},\ \mu \text{ Lebesgue measure in } \mathbb{R}$$

Then, consider $Q := \sum_{i=1}^n X_i^2$. This new random variable $Q:\Omega \longrightarrow \mathbb{R}$ will follow a new distribution, called χ^2 with n degrees of freedom, ie,

$$\chi^2_n(Q^{-1}(A)) = \mu(A) \quad \forall A \in \mathcal{B}_{\mathbb{R}}$$

Its probability density is

$$\rho_Q(x \mid n) = \begin{cases} \dfrac{x^{\frac{n}{2}-1}\, e^{-\frac{x}{2}}}{2^{\frac{n}{2}}\,\Gamma\left(\frac{n}{2}\right)}, & x > 0 \\ 0, & \text{otherwise} \end{cases}$$

- $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt$
- $\Gamma(z) = (z-1)!$
- $\Gamma(z+1) = z\,\Gamma(z)$
- $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$

Sample following a distribution: A sample $(x_1, x_2, \dots, x_n)$ which follows a distribution $\mathcal{D}$ is just a point in the image set of some random vector X which follows $\mathcal{D}$. In other, $\exists\ X : \Omega \longrightarrow \mathbb{R}$ random vector and $w \in \Omega$ such that $X(w) = (x_1, \dots, x_n)$ and such that $\mu = X_* \mathcal{D}$, where μ is the Lebesgue measure in $\mathbb{R}^n$:

$$\mathcal{D}(X^{-1}(A)) = \mu(A)$$

Generating according to one distribution: Generate a random variable following a probability distribution $\mathcal{D}$ is the process of seek a random variable $X : \Omega \longrightarrow \mathbb{R}$ such that $\mu = X_* \mathcal{D}$. It's the same for random vectors. Then, generate a sample of points following $\mathcal{D}$, formally, is the process of take a point in the image of a random variable X such that $\mu = X_* \mathcal{D}$

Left/Right invariant measures: Let $(G, \cdot)$ be a locally compact Hausdorff topological group ($\forall\ g \in G,\ \exists$ K compact set and U open set such that $x \in U \subset K$).

Let $g \in G$ and $S \subset G$. Then, define the translations

- Left: $gS = \{gs : s \in S\}$
- Right: $Sg = \{sg : s \in S\}$

→ It preserves Borel sets. As the group operation is continuous, it preserves the topology and, consequently, the σ-algebra $\mathcal{B}$

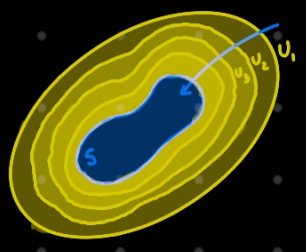
A measure μ on $(G, \mathcal{B})$ is

- Left-translation-invariante if: $\mu(gS) = \mu(S)\ \forall\ g \in G, \forall\ S \in \mathcal{F}$ (Notation: LTI)
- Right-translation-invariante if: $\mu(Sg) = \mu(S)\ \forall\ g \in G, \forall\ S \in \mathcal{F}$ (Notation: RTI)

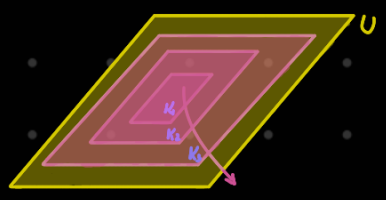
Haar's theorem: Up to a multiplicative constant, $\exists!$ nontrivial measure on $\mathcal{B}$ of G such that

- μ is LTI: $\mu(gS) = \mu(S) \quad \forall g \in G, \forall S \in \mathcal{B}$
- μ is finite on compacts: $\mu(K) < \infty \quad \forall K \subset G$ compact
- μ is outer regular on Borel sets: $\mu(S) = \inf\{\mu(U) : S \subset U, U \in \tau_G\} \quad \forall S \in \mathcal{B}$
- μ is inner regular on open sets: $\mu(U) = \sup\{\mu(K) : K \subset U, K \text{ compact}\} \quad \forall U \in \tau_G$

Such a measure is called the left Haar measure on G



Outer regular



inner regular

This measure is unique with the normalization condition: $\mu(G) = 1$

Left Haar measure → $U \neq \emptyset \Rightarrow \mu(U) > 0$ (there is no trivial set of measure zero)

Left Haar measure → K compact $\Rightarrow \mu(K) < \infty$

↳ It's the usually called Haar measure

→ Considered a generalization of Lebesgue, despite Lebesgue not be always a Haar (case of the cone of covariance matrices in the Wishart distribution)
In $(\mathbb{R}^n, +)$, they coincides. In $GL(n, \mathbb{R})$, they are diffrent by a multiplicative factor which consider the determinant of the matrices

Similarly, $\exists!$ right Haar measure ν (→ They are not necessarily equal) up to a positive cte

Right Haar measure: $\nu(S) = a\mu(S^{-1})$, $a > 0$, $S^{-1} = \{g \in G : g^{-1}g = id \ \forall g \in S\} \quad \forall S \in \mathcal{B}$

Modular function: $\exists \Delta : G \longrightarrow (0, +\infty)$, $g \longmapsto \Delta(g)$ such that, for each $S \in \mathcal{B}$ fixed, $\nu(g^{-1}S) = \Delta(g)\nu(S)$

↳ Modular function

Then, if $\Delta(g) = 1$, ν is right and left invariant and, consequently, $\nu = \mu$

The groups which $\Delta(g)$ are called unimodular groups

Radon measure: Let $(\Omega, \mathcal{B})$ a measurable space with the Borel σ-algebra, where Ω is Hausdorff. A Radon measure in $(\Omega, \mathcal{B})$ is inner regular and locally finite measure.

Every Haar measure is a Radon measure

Dirac measures are Radon

In $(\mathbb{R}^n, +)$: Lebesgue $\equiv$ Haar $\equiv$ Radon

	Lebesgue	Haar	Radon
Domínio de definição	$\mathbb{R}^n$	$(G, \cdot)$ Locally compact Hausdorff topological groups	(Ω, τ) Locally compact Hausdorff topological spaces
Invariance	Translation	Group operation by left	

The cone $M_+(\Omega)$ of finite measures over a (finite) set Ω is a set of Radon measures, specially for the subset $P_+(\Omega)$ of probability measures

Polish space: A Polish space is a topological space (Ω, τ) which is

- Separable: It contains a countable dense subset
- Completely metrizable: It's homeomorphic to a complete metric space

Probability measures on the Borel σ-algebra of a Polish space is a Radon measure

Raw moment: Let $X: \Omega \to \mathbb{R}$ be a random variable with density function (with respect to the underlied σ-algebras and measures) ρ_x. Then, the n-th raw moment is given by

$$\langle X^n \rangle := \int_{\mathbb{R}} x^n \rho_x(x) \, d\mu(x)$$

Central moment: Let $X: \Omega \to \mathbb{R}$ be a random variable with density function (with respect to the underlied σ-algebras and measures) ρ_x. Given $c \in \mathbb{R}$, consider the integral

$$\langle X_c^n \rangle = \int_{\mathbb{R}} (x - c)^n \rho_x(x) \, d\mu(x)$$

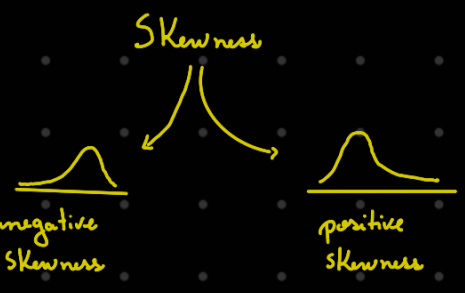
When $c = \langle X \rangle =: \mathbb{E}_x[X]$, $\langle X_c^n \rangle$ is called the n-th central moment of X

↳ We introduce this notation to distinguish when we are considering the density ρ_x of X or we are just computing $\int_\Omega x \, dx$

Standarzed moment: Let $X : \Omega \to \mathbb{R}$ be a random variable with the underlied density function ρ_x. The n-th standardized moment of X is the n-th central moment divided by the n-th standard deviation

$$\langle X_s^n \rangle = \frac{\mathbb{E}_x[(X - \mathbb{E}_x[X])^n]}{\mathbb{E}_x[(X - \mathbb{E}_x[X])^2]^{\frac{n}{2}}}$$

#	Raw	Central	Standardized
1	Mean	0	0
2		Variance	1
3			Skewness
4			Kurtosis
5			Hyperskewness
6			Hypertailedness



Mixed moments: Let $X_1, \dots, X_n$ be random variables with the underlied joint density function ρ_x. For any $K_i \in \mathbb{N}$, we call

- $\mathbb{E}_x[X_1^{K_1} \dots X_n^{K_n}]$ the mixed moment of order $K = \sum_{i=1}^n K_i$
- $\mathbb{E}_x[(X_1 - \mathbb{E}_x[X_1])^{K_1} \dots (X_n - \mathbb{E}_x[X_n])^{K_n}]$ the central mixed moment of order $K = \sum_{i=1}^n K_i$

If $K = 2$, there is only one central mixed moment of each 2 random variables, say X_1, X_2.

$$\mathrm{Cov}(X_1, X_2) = \mathbb{E}_x[(X_1 - \mathbb{E}_x[X_1])(X_2 - \mathbb{E}_x[X_2])]$$

So, the covariance of X_1 and X_2 is the second central mixed moment of them (or, equivalently, the second central moment of the random vector $\vec{X} = (X_1, X_2)$)

↳ Obvious, $\mathrm{Cov}(X_1, X_2)$ is unique and depends on the measures in the domain and codomain of $\vec{X}$)

If X_1 and X_2 are random vectors of same size, $\mathrm{Cov}(\vec{X}_1, \vec{X}_2)$ is the matrix of covariance

If $K = 3$, we have the freedom to choose

- 2 random variables X_1 and X_2

  and, consequently, we have

  $\mathbb{E}_x[(X_1 - \mathbb{E}_x[X_1])(X_2 - \mathbb{E}_x[X_2])^2]$

  $\mathbb{E}_x[(X_1 - \mathbb{E}_x[X_1])^2(X_2 - \mathbb{E}_x[X_2])]$

- 3 random variables X_1, X_2 and X_3

  and, consequently, we have

  $\mathbb{E}_x[(X_1 - \mathbb{E}_x[X_1])(X_2 - \mathbb{E}_x[X_2])(X_3 - \mathbb{E}_x[X_3])]$

Consequently, we can define the coSkewness and coKurtosis in the same way as the covariance, but it's not unique for $K > 1$

Cumulant: Let $X : \Omega \longrightarrow \mathbb{R}$ be a random variable. We define the moment generating function as

$M_x : \mathbb{R} \longrightarrow \mathbb{R}$, $t \longmapsto \mathbb{E}_x[e^{tx}]$,

when this expectation is finite. This name is justified because

$$M_x(t) = 1 + t\,\underbrace{\mathbb{E}_x[X]}_{\int x\,\rho_x(x)\,d\mu(x)} + \frac{t^2}{2!}\mathbb{E}_x[X^2] + \frac{t^3}{3!}\mathbb{E}_x[X^3] + \cdots$$

(1st raw, 2nd raw, 3rd raw)

(If $\vec{X}$ is a random vector, the $M_{\vec{x}}$ is defined as $M_{\vec{x}}(\vec{t}) = \mathbb{E}[e^{\langle \vec{t}, \vec{x} \rangle}]$, $t \in \mathbb{R}^n$)

When M_x is defined, we call the cumulant-generating the logarithm of that one

$K : \mathbb{R} \longrightarrow \mathbb{R}$, $t \longmapsto \ln \mathbb{E}[e^{tX}]$,

The n-th cumulant of X is the n-th coefficient of the Maclaurin expansion of K

$$K(t) = \sum_{n \in \mathbb{N}} \frac{t^n}{n!} \kappa_n, \qquad \kappa_n = \frac{d^{(n)}K}{d\,t^{(n)}}(0)$$

- $\kappa_n(X + c) = \begin{cases} \kappa_n(X) + c, & \text{if } n = 1 \\ \kappa_n(X), & \text{if } n > 1 \end{cases}$ (invariant by translation)
- $\kappa_n(cX) = c^n \kappa_n(X)$ (n-homogeneous)
- $\kappa_n(X_1 + \cdots + X_k) = \sum_{i=1}^{k} \kappa_n(X_i)$ (linear on the sum)

- $\kappa_1(X) = \langle X \rangle$
- $\kappa_2(X) = \mathrm{Var}(X)$

} Important: The only probability distribution such that all the cumulants of order ≥ 3 of a random variable which follows this distribution is the normal distribution (trivial δ-distribution if $\kappa_2 = 0$)

when $\{p_\theta\}$ is parametric, we write $\mathrm{Var}_\theta(X)$

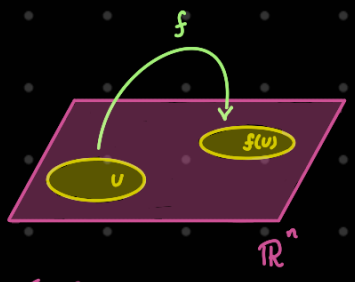
In general, it's not unique. Then, usually, people write $\kappa_{r_1,\dots,r_\ell}$

For random vector, we will have the cumulants obtained by the coefficients of the Maclaurin series for multiple variables. Then, if $M_{\vec{X}}(\vec{t})$ is the moment-generating function and it's finite, then a r-th cumulant of $\vec{X}$ is

$$\kappa_r(\vec{X}) = \frac{\partial^{r_1}}{\partial t_1^{r_1}} \cdots \frac{\partial^{r_\ell}}{\partial t_n^{r_\ell}} K(t)\Big|_{t=0}, \qquad r = \sum_{i=1}^{\ell} r_i$$

Invariance of domain: If U is an open set of $\mathbb{R}^n$ and $f: U \to \mathbb{R}^n$ is a continuous injective map, then $f(U)$ is open in $\mathbb{R}^n$ and f is homeomorphism between U and $f(U)$

↳ Continuous injective functions which preserve the dimension of the space are open



Complete statistics: A statistic T is complete for a family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ of probability measures if, $\forall$ measurable function f,

$$\mathbb{E}_\theta[f \circ T] = c \quad \forall \theta \in \Theta \Rightarrow f \circ T = c \text{ a.e. with respect to all } P_\theta \in \mathcal{P}$$

If the expectation of $f \circ T$ is constant for all θ, then $f \circ T$ need to be equal to this constant (a.e. with respect to the distributions of $\mathcal{P}$)

Note that, for each f, we can have another constant c; c is constant in Θ.

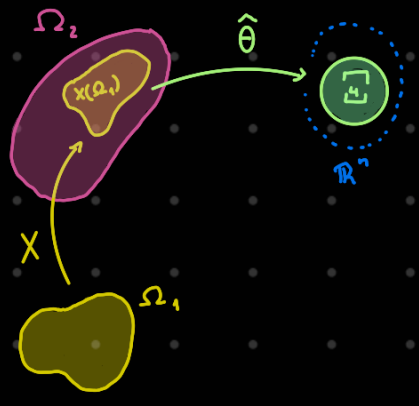
There is non-constant invisible functions of T for $\mathcal{P}$, i.e, with $\mathbb{E}_\theta$ null for all the parameters

Theorem: If T is complete and sufficient, then T is minimal sufficient

	Sufficient	Minimal	Complete
CONCEPT	$T(X)$ contains all the needed information about the parameter present in the "data of X"	It's the "lowest" statistics which contains all the needed information about parameter present in the "data of X"	Guarantees there is no functions non-trivial (cte) of $T(X)$ with constant expectation for all θ
ROLE	Reduction of data without lack of information about the parameter	Best reduction of data without lack of information about the parameter	Provides the uniqueness of unbiased estimator of θ based on $T(X)$ with minimum variance

we can define it on a manifold, but we need to care about the support of the functions

Estimator: Let $(\Omega, \mathcal{F})$ and $(\Theta \subset \mathbb{R}^n, \mathcal{G})$ be measurable spaces. An estimator is a measurable function $\hat{\theta}: \Omega \longrightarrow \Theta$. Ω is called sample space and Θ is called sample estimates space or sometimes just parameter space.



Given a random variable $X: \Omega_1 \longrightarrow \Omega_2$, an estimator based on X is an estimator $\hat{\theta}: X(\Omega_1) \longrightarrow \Theta$

We define the bias of $\hat{\theta}$ as the function

$B: \Theta \longrightarrow \mathbb{R}^n, \quad \theta \longrightarrow B(\theta) = \mathbb{E}_\theta[\hat{\theta}(X)] - \theta$

An estimator $\hat{\theta}$ is called unbiased for some $\theta \in \Theta$ if $B(\theta) = 0$, ie, if $\mathbb{E}_\theta[\hat{\theta}] = \theta$. In this case, usually is called "the true value of θ".

The error of $\hat{\theta}$ for some $\theta \in \Theta$ is defined as the function

$\mathcal{E}_\theta: \Omega_1 \longrightarrow \mathbb{R}, \quad x \longmapsto \mathcal{E}_\theta(x) = \hat{\theta}(x) - \theta$

Consequently, the mean squared error for some $\theta \in \Theta$ is defined as

$\overline{\mathcal{E}}^2_\theta(\hat{\theta}) = \mathbb{E}_\theta[\mathcal{E}_\theta(x)^2] = \mathbb{E}_\theta[(\hat{\theta}(x) - \theta)^2]$

The variance of the estimator $\hat{\theta}$ for some $\theta \in \Theta$ is

$$\mathrm{Var}_\theta(\hat{\theta}) := \mathbb{E}_\theta[(\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}])^2]$$

For $\theta \in \Theta$, the variance connects the bias of $\hat{\theta}$ with its squared mean error by

$$\mathrm{Var}_\theta(\hat{\theta}) = \bar{E}^2_\theta(\hat{\theta}) - B(\theta)^2$$

↑ variance of an estimator measures how far away the squared mean error is from the bias^2

Conditional expectation: Let $(\Omega_1, \mathcal{F}_1, P)$ be a probability space and let $\alpha : \Omega_1 \to \Omega_2$, $\alpha \in L(P)$, be a random variable with values in $(\Omega_2, \mathcal{F}_2)$. Let $\sigma(\alpha) \subset \mathcal{F}_1$ be the sub-σ-algebra of $\mathcal{F}_1$ generated by α, ie, the smaller sub-σ-algebra of $\mathcal{F}_1$ such that α is still measurable. Consider the function $\sigma(\alpha)$-measurable $g : \Omega_1 \to \mathbb{R}$ such that

$$\int_A \alpha(x)\, dP(x) = \int_A g(x)\, dP(x) \quad \forall\, A \in \sigma(\alpha)$$

g is called the conditional expectation of α. It's denoted by $\mathbb{E}[\alpha \mid \sigma(\alpha)]$ and it's unique up to P-equivalence.

Now, let $T : \Omega_1 \to \Omega_3$ be a random variable between $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_3, \mathcal{F}_3)$. Consider the sub-$\sigma$-algebra $\sigma(T)$ generated by T. The conditional expectation of α given T is the unique (up P-equivalence) $\sigma(T)$-measurable function $h : \Omega_1 \to \mathbb{R}$ such that

$$\int_B \alpha(x)\, dP(x) = \int_B h(x)\, dP(x) \quad \forall \quad B \in \sigma(T),$$

when $h \in L(P)$. We denote that conditional expectation by $\mathbb{E}_x[X \mid T]$. Then, we can write

$$\mathbb{E}_x[X \mid T] := \mathbb{E}_x[\alpha \mid \sigma(T)],$$

when it exists and is finite

Lehmann-Scheffé Theorem: Let $X: \Omega_1 \longrightarrow \Omega_2$ be a random variable and let $T: \Omega_2 \longrightarrow \Omega_3$ be a complete and sufficient statistic for a family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ of probability measures in $(\Omega_2, \mathcal{F}_2)$. Let $\hat{\theta}(X)$ be a unbiased estimator of $\theta \in \Theta$ such that $\hat{\theta} = f \circ T(X)$, for some measurable function $f: \Omega_3 \longrightarrow \Theta$. Then, $\hat{\theta}$ is the unique (up to sets of measure zero) minimum variance unbiased estimator of θ, ie,

$$\mathrm{Var}_\theta(\hat{\theta}(X)) \leq \mathrm{Var}_\theta(\hat{\theta}'(X)),$$

for all unbiased estimator $\hat{\theta}'$ of θ.

Rao-Blackwell Theorem: Let $X: \Omega_1 \longrightarrow \Omega_2$ be a random variable and let $T: \Omega_2 \longrightarrow \Omega_3$ be a sufficient statistic for a family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ of probability measures in $(\Omega_2, \mathcal{F}_2)$. Let $\hat{\theta}$ be a unbiased estimator of $\theta \in \Theta$. Then, the unbiased estimator for θ given by

$$\hat{\theta}^*(X): \Omega_2 \longrightarrow \Theta, \quad \hat{\theta}^*(X) = E[\hat{\theta}(X) \mid T(X)]$$

is such that

$$\mathrm{Var}_\theta(\hat{\theta}(X)) \geq \mathrm{Var}_\theta(\hat{\theta}^*(X))$$

Statistical model: Let $(\Omega, \mathcal{F})$ be a measurable space. A statistical model for Ω is a subset $\mathcal{P}' \subset \mathcal{P}$ of the set $\mathcal{P}$ of probability measures in $(\Omega, \mathcal{F})$. Moreover, let Θ a non-empyt set. A parametric statistical model is a set $\mathcal{P}' = \{P_\theta : \theta \in \Theta\} \subset \mathcal{P}$ indexed by Θ. In this case, Θ is called a space of parameters for $\mathcal{P}'$. When another measure μ in $(\Omega, \mathcal{F})$ is underlied and the Radon-Nikodym derivative of $dP/d\mu$ of P and μ exists for all $P \in \mathcal{P}'$, the set of this functions may be called a statistical model for Ω in some literatures.

Regular statistical model: Let $\mathcal{P}' = \{P_\theta : \theta \in \Theta\}$ be a parametric statistical model on $(\Omega, \mathcal{F})$.

It's called regular if all the conditions below hold:

- Global differentiation: $\Theta \subset \mathbb{R}^n$, $\Theta \in \tau_{\mathbb{R}^n}$
- Differentiation of the Likelyhood function: The function $\theta \longmapsto L(\theta; \cdot)$ is $C^1(\Theta)$ (or suffitially differentiable) μ-almost everywhere in Ω

- Dominance of fixed measure: $\exists\ \mu$ σ-finite measure such that, $\forall\ \theta \in \Theta$, $P_\theta \ll \mu$
- Constant support: $\{x \in \Omega : p_\theta(x) > 0\}$ doesn't depend on θ, up to μ-almost everywhere
- Permutation between derivative and integral: Hold

$$\frac{\partial}{\partial \theta_i} \int_A f(x,\theta)\, d\mu(x) = \int_A \left(\frac{\partial}{\partial \theta_i} f(x,\theta)\right) d\mu(x), \quad \forall\ A \in \mathcal{F},\ i = 1,\dots,n,$$

for all $f(\cdot,\theta) \in L^1(A)$, $f(x,\cdot) \in C^1(\Theta)$, fixed $\theta \in \Theta$ and $x \in \Omega$, respectively.

- Well behaved Fisher information: The Fisher information $I(\theta)$ for $\theta \in \Theta$, given by

$$I(\theta) = \mathbb{E}_\theta\left[\left(\partial_\theta \ln(p_\theta(X))\right)\left(\partial_\theta \ln(p_\theta(X))\right)^T\right]$$

is a non-singular operator, ie, $\forall\ \theta \in \Theta$, $I(\theta)$ exists, is continuous and is strictly positive in Θ

↳ small variations in θ produce detectable variations on the variation

The definition of regular model is not a consense in the literature. Some conditions, sometimes, are removed. When a model is not regular, it's called singular parametric statistical model