Estadística Matématica – 2022

Emilien Joly – Ayu. —

CIMAT - Guanajuato

Horarios y enlaces

Profesor: Emilien Joly (emilien.joly@cimat.mx)
Ayudante: — (—@cimat.mx)

Fecha de inicio : **Martes 16 de agosto**

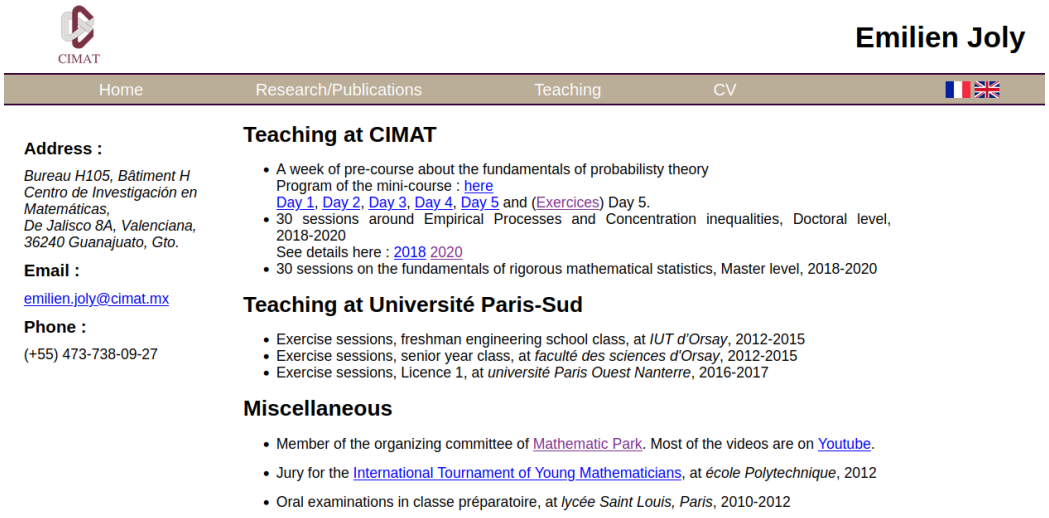
Horarios de clase: martes y jueves, 9:30-10:50
Horarios de prácticas: viernes, 11:00-12:20

Fecha de fin de clases : **viernes 25 de noviembre**

Información útil

La información del curso de EM se encontrará en la pagina web :

- `https://joly415.perso.math.cnrs.fr/EnseignementEG.htm`



CIMAT

Emilien Joly

Home | Research/Publications | Teaching | CV

Address :

Bureau H105, Bâtiment H
Centro de Investigación en Matemáticas,
De Jalisco 8A, Valenciana,
36240 Guanajuato, Gto.

Email :

emilien.joly@cimat.mx

Phone :

(+55) 473-738-09-27

Teaching at CIMAT

- A week of pre-course about the fundamentals of probabilisty theory
  Program of the mini-course : here
  Day 1, Day 2, Day 3, Day 4, Day 5 and (Exercices) Day 5.
- 30 sessions around Empirical Processes and Concentration inequalities, Doctoral level, 2018-2020
  See details here : 2018 2020
- 30 sessions on the fundamentals of rigorous mathematical statistics, Master level, 2018-2020

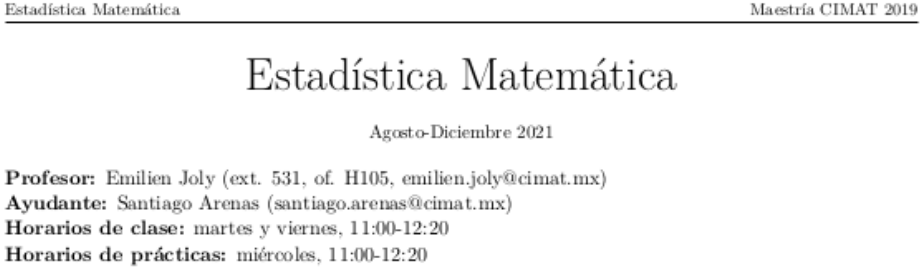
Teaching at Université Paris-Sud

- Exercise sessions, freshman engineering school class, at *IUT d'Orsay*, 2012-2015
- Exercise sessions, senior year class, at *faculté des sciences d'Orsay*, 2012-2015
- Exercise sessions, Licence 1, at *université Paris Ouest Nanterre*, 2016-2017

Miscellaneous

- Member of the organizing committee of Mathematic Park. Most of the videos are on Youtube.
- Jury for the International Tournament of Young Mathematicians, at *école Polytechnique*, 2012
- Oral examinations in classe préparatoire, at *lycée Saint Louis, Paris*, 2010-2012

- EL temario del curso : aqui



Estadística Matemática

Agosto-Diciembre 2021

Profesor: Emilien Joly (ext. 531, of. H105, emilien.joly@cimat.mx)
Ayudante: Santiago Arenas (santiago.arenas@cimat.mx)
Horarios de clase: martes y viernes, 11:00-12:20
Horarios de prácticas: miércoles, 11:00-12:20

Algunas referencias


Gut, A. (2009) *An intermediate course in probability*.Springer Science & Business Media. [link]


Christensen, R. (2011). *Plane answers to complex questions: the theory of linear models*. Springer Science & Business Media. [link]


Serfling, R. J. (2009). *Approximation theorems of mathematical statistics*. John Wiley & Sons. [link]


Van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge university press. [link]


Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media. [link]


Williams, D. (1991). *Probability with martingales*. Cambridge university press. [link]


Feller, W. (1971) An introduction to Probability Theory and its Applications. John Wiley & Sons. [link]


Knuth, D. (1978) The art of computer programming. Vol. 1: Fundamental algorithms. Atmospheric Chemistry & Physics. [link]

Tareas y exámenes

Fechas de Tareas: (se entregará en sesiones de ayudantía)

- Tarea 1: Inicio 5 septiembre $\rightarrow$ Entrega 16 septiembre
- Tarea 2: Inicio 19 septiembre $\rightarrow$ Entrega 30 septiembre
- Tarea 3: Inicio 3 octubre $\rightarrow$ Entrega 14 octubre
- Tarea 4: Inicio 17 octubre $\rightarrow$ Entrega 28 octubre
- Tarea 5: Inicio 31 octubre $\rightarrow$ Entrega 11 noviembre

Cuenta por **2/5** de la calificación final.

Fechas de examenes:

- Examen 1: El viernes 30 de septiembre (3h) (co. 1/5)
- Examen final: El 2 de diciembre (4h) (co. 2/5)

Cuenta por **3/5** de la calificación final.

Objectivos

1. Familiarizarse con conceptos de teoría de probabilidad de relevancia directa para la formulación de modelos esdadísticos y el manejo de propiedades elementales de métodos estadísticos.
2. Estudiar los llamados métodos de estadística descriptiva (numéricos y gráficos) a la luz de dichos conceptos probabilísticos.
3. Adiestrarse en el entendimiento y manipulación de herramientas matemáticas para inferencia.
4. Estudiar propiedades matemáticas primordiales de estimadores clásicos.
5. Conocer algunas clases de modelos estadísticos de uso general.

1.Preliminares

1.1.Repasos de probabilidad

1.1.1. Espacio de probabilidad

Medidas

- Es una función $\mu : \Sigma \subset \mathcal{P}(\Omega) \to [0; +\infty]$.

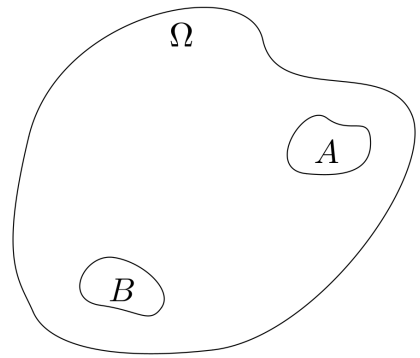
~~Aditiva: Si $A \cap B = \emptyset$, entonces $\mu(A \cup B) = \mu(A) + \mu(B)$~~

- Numerablemente aditiva: Si $(A_n)_{n\in\mathbb{N}}$ son conjuntos disjuntos tq $A = \bigcup_n A_n$, tenemos

$$\mu(A) = \sum_{n\in\mathbb{N}} \mu(A_n)$$

Eso implica $\mu(\emptyset) = 0$

Se tiene que dar un peso no negativo a A (por ejemplo).



$\mu(A) > 0$.

Que falta definir?

El conjunto Σ!

1.1.1. Espacio de probabilidad

El conjunto Σ debe de ser una σ-algebra.

- $\Omega \in \Sigma$,
- $A \in \Sigma \implies A^c \in \Sigma$,
- $A, B \in \Sigma \implies A \cup B \in \Sigma$,
- $(A_n)_n \in \Sigma \implies \bigcup_n A_n \in \Sigma$

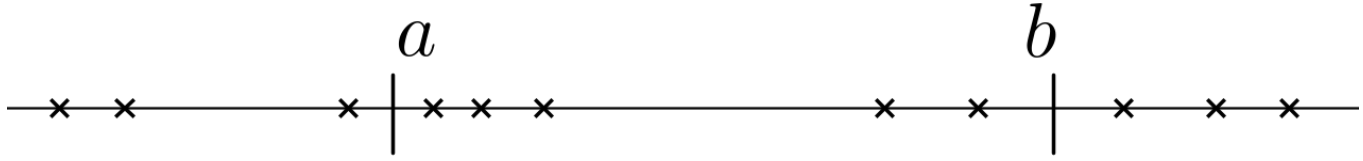
Vocabulario:

- μ es finita si $\mu(\Omega) < +\infty$,
- μ es σ-finita si $\exists (S_n)_n \in \Sigma$ tq $\forall n, \mu(S_n) < +\infty$ y $\bigcup_n S_n = \Omega$,
- μ es una medida de probabilidad si $\mu(\Omega) = 1$.

1.1.1. Espacio de probabilidad

Ejemplos:

- En $\Omega = \mathbb{R}$, podemos definir $\mu([a, b]) = b - a$.
  Ejercicio: Probar que μ es una medida. Que es Σ en este caso?
- Si uno pone puntos sobre la linea real.



Podemos definir
$\mu([a, b]) = \#$número de puntos en el intervalo $[a, b]$.
Ejercicio: Probar que μ es una medida.

1.1.1. Espacio de probabilidad

Variables aleatorias

Una variable aleatoria a valores reales es una función $X : \Omega \to \mathbb{R}$ *medible* entre los espacios de probabilidad

$$(\Omega, \Sigma, \mu) \longrightarrow (\mathbb{R}, \mathcal{B}, P)$$

- Medible significa : $\forall B \in \mathcal{B}, X^{-1}(B) \in \Sigma$
- El conjunto $\mathcal{B}$ son los borelianos (que son?) de $\mathbb{R}$.

Para $\omega \in \Omega$, $X(\omega)$ se llama *realización* de la variable aleatoria X.

1.1.1. Espacio de probabilidad

Ejemplo:
Definimos la función

$$\begin{array}{rccc} X & : & [0,1] & \to & \{0,1\} \\ & & \omega & \mapsto & \mathbb{1}_{\omega > \frac{1}{2}} \end{array}$$

- Entre los espacios $([0,1], \mathcal{B}, \text{Leb})$ y $(\{0,1\}, \mathcal{B}, P)$ tal que $P(\{0\}) = P(\{1\}) = 1/2$ tiene el nombre de una variables de Bernoulli denotado $\mathcal{B}(1/2)$.
- Entre los espacios $([0,1], \mathcal{B}, Q)$ y $(\{0,1\}, \mathcal{B}, P)$ tal que $Q([a,b]) = 2(b \wedge \frac{1}{2} - a \wedge \frac{1}{2})$ tiene el nombre de la variable constante igual a 0.

Todos los elementos de la definición son importantes!

1.1.1. Espacio de probabilidad

Definimos la notación muy commun $\{X \in A\}$.
Lo llamamos *evento* de la variable aleatoria X. Por definición es $\{\omega \in \Omega : X(\omega) \in A\}$ o igualmente $X^{-1}(A)$. Es un elemento de Σ.

Podemos calcular su probabilidad!

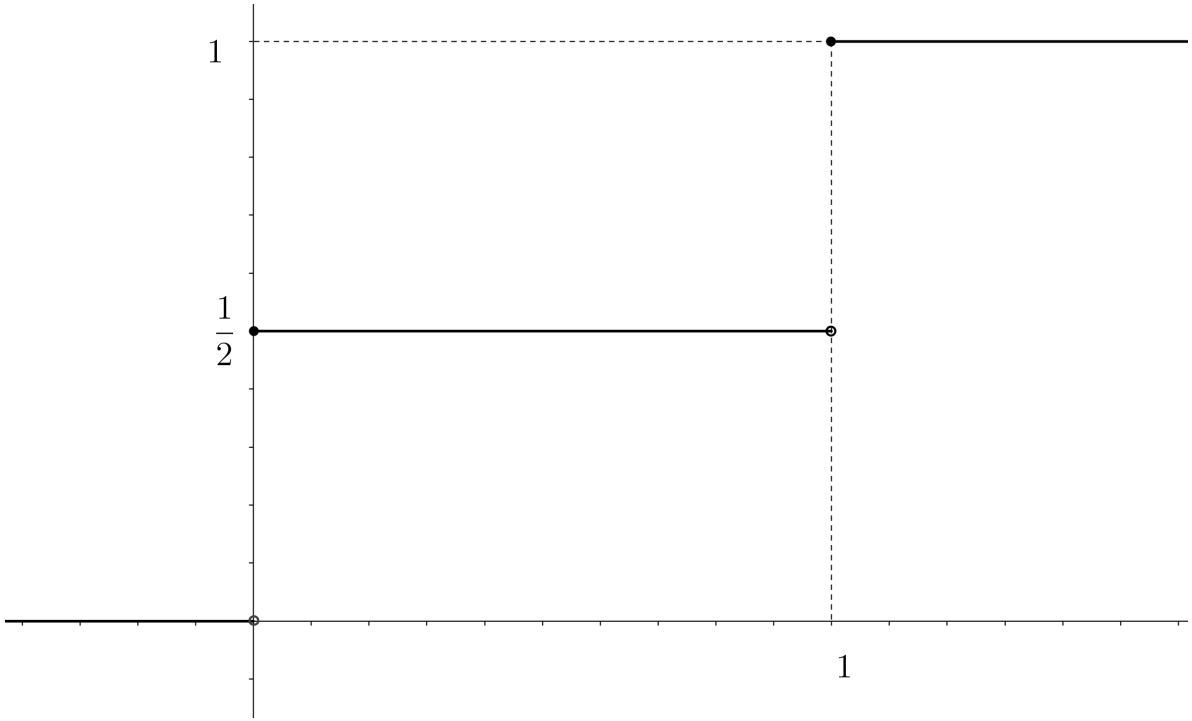
Entonces la escritura $\mathbb{P}(X \in A)$ significa $\mu(X^{-1}(A))$

Igualmente, podemos dar un sentido a las escrituras
$\mathbb{P}(a < X \leq b)$, $\mathbb{P}(X \leq t)$, $\mathbb{P}(X \notin A)$

1.1.1. Espacio de probabilidad

Función de distribución

Sea X una v.a. Para todo $t \in \mathbb{R}$, $(-\infty, t]$ es un intervalo y entonces es un elemento medible. Denotamos $F_X(t) = \mathbb{P}(X \leq t)$ la *función de distribución* de X.



La función de distribución una Bernoulli(1/2).

1.1.1. Espacio de probabilidad

La función de distribución satisface las propiedades siguientes

Proposición

Para una variable aleatoria real X,

1. *la función F_X es no-decreciente.*
2. *la función F_X tiene limites*

$$\lim_{t\to-\infty} F_X(t) = 0 \qquad \lim_{t\to+\infty} F_X(t) = 1$$

3. *la función F_X es continua a la derecha.*

1.1.1. Espacio de probabilidad

Densidad

Se dice que una variable X tiene una densidad f_X si $\forall a, b \in \mathbb{R}$,

$$\mathbb{P}(X \in [a, b]) = \int_a^b f_X(x)dx$$

Cuidado! Hay variables aleatorias que no tienen una función densidad.

Ejemplo: $\mathcal{B}(1/2)$ no tiene densidad!

Ejemplo: La función $f : \mathbb{R} \to \mathbb{R}$ tq $f(x) = \mathbb{1}_{[0,1]}(x)$ es una función de densidad asociada a la variable denotada $U([0, 1])$ y dada por

$$\begin{array}{ccccc} X & : & [0,1] & \to & [0,1] \\ & & \omega & \mapsto & \omega \end{array}$$

de $([0, 1], \mathcal{B}, \text{Leb})$ hasta $([0, 1], \mathcal{B}, \text{Leb})$.

1.1.1. Espacio de probabilidad

Proposición

Una función de densidad debe de cumplir

- $f_X : \mathbb{R} \to \mathbb{R}_+$.
- $\int_{-\infty}^{+\infty} f_X(x)dx = 1$.

P_X

La medida de probabilidad del espacio $(\mathbb{R}, \mathcal{B}, \underline{P})$ se denota P_X y vale por definición

$$\forall B \in \mathcal{B}, P_X(B) = \mu(X^{-1}(B)).$$

Se llama ley de probabilidad de la v.a. X.

1.1.1. Espacio de probabilidad

Notación:
La determinación de P_X,F_X o f_X es suficiente para caracterizar completamente la variable aleatoria X. Usamos la notación $\sim$ para decir "X tiene la distribución ...". Usaremos indistintamente

$$X \sim P_X$$
$$X \sim F_X$$
$$X \sim f_X$$

o otras nociones no ambiguas como $X \sim \mathcal{B}(p)$.
Deben de conocer:
$\mathcal{B}(p)$, $\mathcal{B}(n,p)$, $\mathcal{P}(\lambda)$, $\mathcal{E}(\lambda)$, $\mathcal{N}(\mu,\sigma^2)$, $\mathcal{U}([a,b])$, Rad(p), $\mathcal{G}(p)$,

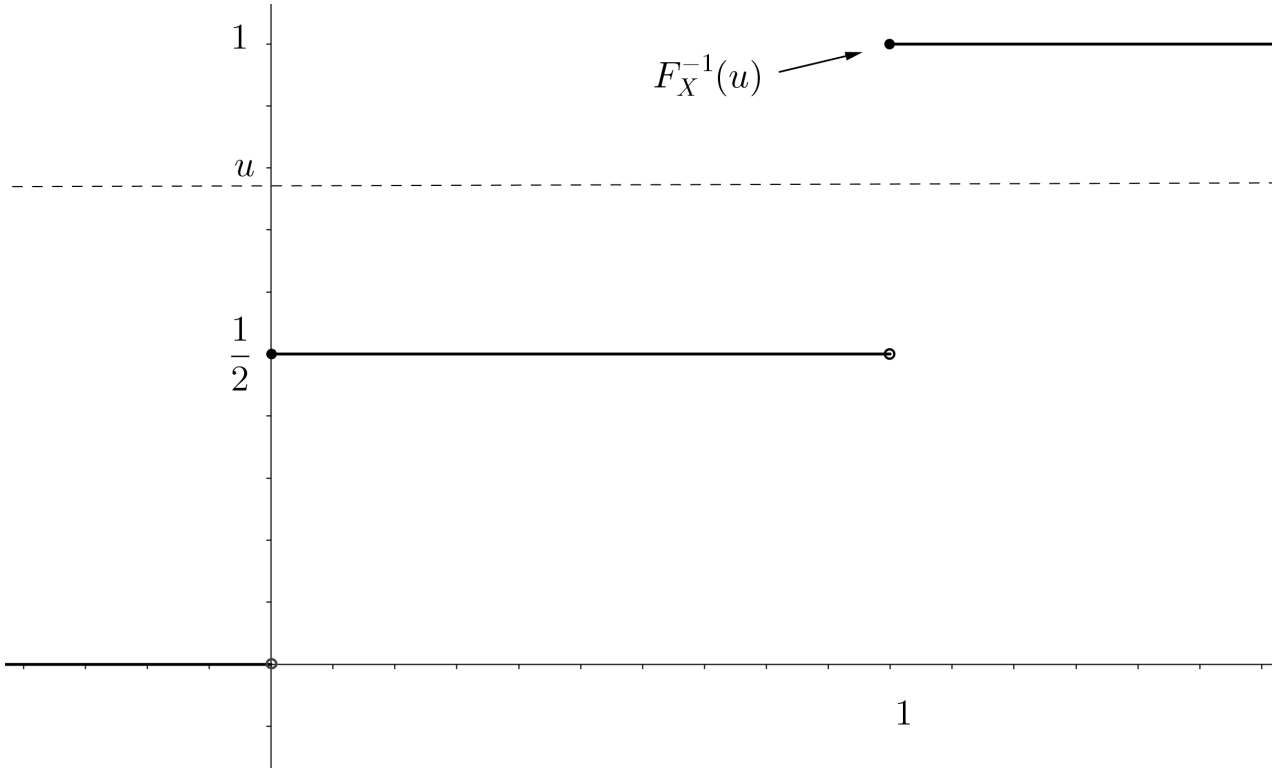
1.1.1. Espacio de probabilidad

Cuantiles

La función F_X es una función no-decreciente lo que permite definir la pseudo-inversa

$$F_X^{-1}(u) = \inf\{t \in \mathbb{R} : F_X(t) \geq u\}$$

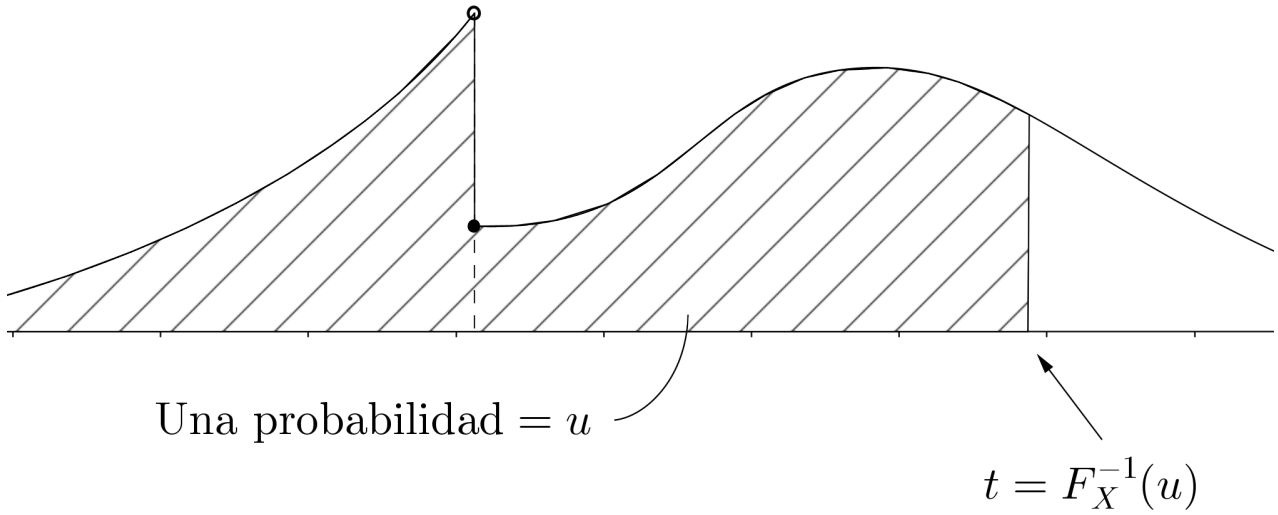
para $u \in [0, 1]$. La cantidad $F_X^{-1}(u)$ se llama *cuantile* de orden u.



1.1.1. Espacio de probabilidad

Interpretación en caso de densidad

El cuantile t de orden u es el punto del eje de abscisa tal que la superficie debajo de la curva de f_X contenga una probabilidad igual a u.



- Cuando $u + \frac{1}{2}$, el cuantile se llama *mediana*
- Cuando $u = \frac{1}{4}$ o $u = \frac{3}{4}$, el cuantile se llama *cuartile*.

1.1.2. Calculo con medidas de probabilidad

Valores esperados

Las medidas de probabilidad son medidas particulares $\rightarrow$ hay una noción de $\int$ y de integrabilidad.

- Cuando X es una variable de densidad f_X,

$$\mathbb{E}[X] = \int_{\mathbb{R}} x f_X(x) dx.$$

- Si $f(x) = (x - \mathbb{E}[X])^2$,

$$\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \int_{\mathbb{R}} (x - \mathbb{E}[X])^2 f_X(x) dx$$

  se llama *varianza* de X y se denota $\mathrm{Var}(X)$.

Cuidado: No siempre existen $\mathbb{E}[X]$ y $\mathrm{Var}(X)$! En general, hay que probar que esas cantidades existen antes de calcularlas.

1.1.2. Calculo con medidas de probabilidad

Ejemplos:

- Valor esperado de $X \sim \mathcal{U}([2,9])$.
  X tiene la densidad $f_X(x) = \mathbb{1}_{[2,9]}(x) \times \frac{1}{7}$.

$$\mathbb{E}[X] = \int_{\mathbb{R}} x \mathbb{1}_{[2,9]}(x) \times \frac{1}{7} dx = \int_2^9 \frac{x}{7} dx$$
$$= \left[\frac{x^2}{14}\right]_2^9 = \frac{81-4}{14} = \frac{77}{14} = \frac{11}{2} = \mathbf{5.5}$$

- Valor esperado de $X \sim \mathcal{G}(p)$.

$$\mathbb{E}[X] = \int_{\mathbb{R}} x dP_X(x) = \sum_{k=1}^{\infty} q^{k-1} pk$$
$$= p \sum_{k=0}^{\infty} q^k (k+1) = pS$$

1.1.2. Calculo con medidas de probabilidad

Ejemplos:

- Valor esperado de $X \sim \mathcal{G}(p)$.
  Sea $S = \sum_{k=0}^{\infty} q^k(k+1)$ y $h(x) = \sum_{k=0}^{\infty} q^k x^k$ definida sobre $[0, q^{-1})$. Claramente, $h(x) = \frac{1}{1-xq}$ y

$$h'(x) = \sum_{k=1}^{\infty} k x^{k-1} q^k = \sum_{k=0}^{\infty} (k+1) x^k q^{k+1}$$

y $h'(1) = qS$. Por otro lado,

$$h'(x) = -q \frac{-1}{(1-xq)^2} = \frac{q}{(1-xq)^2},$$

lo que implica $h'(1) = \frac{q}{p^2}$ y $S = \frac{1}{p^2}$.
Finalmente, $\mathbb{E}[X] = pS = \frac{p}{p^2} = \boldsymbol{\frac{1}{p}}$.

1.1.2. Calculo con medidas de probabilidad

Algunas propiedades:

El valor esperado $\mathbb{E}[X]$ se define como una integral al respecto de una medida. Lo que permite afirmar sin prueba:

1. El valor esperado es lineal : $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$.
2. Si $X \geq 0$ entonces $\mathbb{E}[X] \geq 0$.
3. Monotonicidad: Si $0 \leq X_n \uparrow X$ entonces $\mathbb{E}[X_n] \uparrow \mathbb{E}[X]$.
4. Fatou: Si $X_n \geq 0$, $\mathbb{E}[\liminf X_n] \leq \liminf \mathbb{E}[X_n]$.
5. Convergencia dominada: Si $\forall n$, $|X_n(\omega)| \leq V(\omega)$ con $\mathbb{E}[V] < \infty$ y $X_n(\omega) \underset{n\to\infty}{\to} X(\omega)$, entonces
$$\mathbb{E}[X_n] \underset{n\to\infty}{\to} \mathbb{E}[X].$$
6. Jensen: Sea $\phi : \mathbb{R} \to \mathbb{R}$ una función convexa tq $\mathbb{E}[\phi(X)] < \infty$, entonces
$$\phi(\mathbb{E}[X]) \leq \mathbb{E}[\phi(X)].$$

1.1.2. Calculo con medidas de probabilidad

Notación $\mathbb{E}[g(X)]$

Según la definición, se puede dar un sentido a la notación $\mathbb{E}[g(X)] := \int_{\mathbb{R}} g(x) dP_X(x)$.

- Si X tiene densidad f_X, $\mathbb{E}[g(X)] = \int_R g(x) f_X(x) dx$.
- Si X es discreta sobre $\mathcal{X}$, $\mathbb{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x) \mathbb{P}(X = x)$.

Cuidado: Es posible que $\mathbb{E}[g(X)]$ no tenga sentido! (Integrabilidad!).

Ejemplos:

- Para $g(x) = \mathbb{1}_A(x)$, tenemos $\mathbb{E}[g(X)] = \mathbb{P}(X \in A)$.
- Si g es acotada por una función G integrable tq $G \times f_X$ es integrable, entonces $\mathbb{E}[g(X)]$ tiene sentido y $\mathbb{E}[g(X)] \leq \mathbb{E}[G(X)]$.
- En particular si $\forall x, g(x) \leq M$, $\mathbb{E}[g(X)]$ existe y $\mathbb{E}[g(X)] \leq M$.

1.1.2. Calculo con medidas de probabilidad

Independencia

- Se dice que dos eventos A y B son *independientes* si $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$.
- Se dice que dos variables aleatorias X y Y son *independientes* si $\forall a, b, c, d \in \mathbb{R}$ con $a \leq b$ y $c \leq d$, los dos eventos $\{X \in [a, b]\}$ y $\{Y \in [c, d]\}$ son independientes.

La intuición es que dos eventos independientes no tienen efecto uno sobre el otro. Como consecuencia, es suficiente de conocer la probabilidad de A y B calcular la probabilidad del evento "A y B". Eso permite de calcular la probabilidad de eventos complejos que involucran a muchas variables aleatorias independientes.

1.1.2. Calculo con medidas de probabilidad

Proposición

Sean X y Y dos variables independientes. Sean f y g dos funciones a valores reales y medibles entonces,

$$\mathbb{E}\left[f(X)g(Y)\right] = \mathbb{E}\left[f(X)\right]\mathbb{E}\left[g(Y)\right].$$

Prueba.

Ver un curso de teoría de la medida. (Tomar aproximaciones de funciones $f_n = \sum \alpha_i \mathbb{1}_{I_i}$). □

Comentario: En efecto es una caracterización de la independencia $\rightarrow$ un "si y solo si".

1.1.2. Calculo con medidas de probabilidad

Desigualdad de Markov

Proposición

Sea X una variable aleatoria real y $g : \mathbb{R} \to \mathbb{R}_+^*$ creciente tq $\mathbb{E}[g(X)] < \infty$. Entonces,

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[g(X)]}{g(t)}.$$

Prueba.

La función g siendo creciente $\mathbb{P}(X \geq t) = \mathbb{P}(g(X) \geq g(t))$ y

$$\begin{aligned}
\mathbb{E}[g(X)] &= \mathbb{E}\left[g(X)\mathbb{1}_{g(X)\geq g(t)}\right] + \mathbb{E}\left[g(X)\mathbb{1}_{g(X)<g(t)}\right] \\
&\geq \mathbb{E}\left[g(t)\mathbb{1}_{g(X)\geq g(t)}\right] \\
&= g(t)\mathbb{P}(g(X) \leq g(t))
\end{aligned}$$

□

1.1.2. Calculo con medidas de probabilidad

Ejemplos:

- Si $g = \mathrm{id}$ y sean X y Y dos variables positivas y independientes, entonces

$$\mathbb{P}(XY \geq t) \leq \frac{\mathbb{E}[XY]}{t} = \frac{\mathbb{E}[X]\,\mathbb{E}[Y]}{t}.$$

- Si $g(x) = x^2$ y que consideramos la variable aleatoria $|X - \mathbb{E}[X]|$,

$$\begin{aligned}\mathbb{P}(|X - \mathbb{E}[X]| \geq t) &= \mathbb{P}\left((X - \mathbb{E}[X])^2 \geq t^2\right) \\ &\leq \frac{\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]}{t^2} \\ &= \frac{\mathrm{Var}(X)}{t^2}.\end{aligned}$$

Esa desigualdad se llama *desigualdad de Chebyshev*.

1.1.2. Calculo con medidas de probabilidad

Variables i.i.d.

Sean $X_1, \ldots, X_n$ una colección de variables aleatorias. Se dice que $X_1, \ldots, X_n$ son *i.i.d.* si $\forall i, j$

- X_i y X_j son independientes.
- X_i y X_j tienen la misma distribución.

Diferencia $> \epsilon$ y $\geq \epsilon$

Nos preguntamos de la diferencia entre $\mathbb{P}(X \geq \epsilon)$ y $\mathbb{P}(X > \epsilon)$.
Vimos que en el caso discreto en general $\mathbb{P}(X \geq \epsilon) \neq \mathbb{P}(X > \epsilon)$ (pensar en la función de distribución de la $\mathcal{B}(1/2)$).
¡Es el único caso!

Proposición

Sea X una variable aleatoria con densidad f_X. Entonces, $\forall \epsilon$, $\mathbb{P}(X \geq \epsilon) = \mathbb{P}(X > \epsilon)$

1.1.2. Calculo con medidas de probabilidad

Prueba.

Eso viene del hecho que el valor del integral $\int_{\epsilon}^{+\infty} f_X(x)dx$ no toma en cuenta si el intervalo $[\epsilon, \infty)$ o (ϵ, ∞) sea abierto o cerrado en ϵ. □

Comentario: Todas la otras probabilidades/esperanzas con desigualdades $\leq$ o $\geq$ son iguales a las mismas probabilidades/esperanzas reemplazando $\leq$ por $>$ y $\geq$ por $<$.

1.1.3. Herramientas de probabilidad para caracterización de distribuciones en $\mathbb{R}$

Por el momento, sabemos que se puede describir una variable aleatoria con

- Su definición $X : (\Omega, \Sigma, \mu) \to (\mathbb{R}, \mathcal{B}, P_X)$,
- Su función $F_X : \mathbb{R} \to [0, 1]$,
- Su ley P_X,
- Cuando existe, su densidad f_X.

Problema: Encontrar mas herramientas que pueden caracterizar la distribución de X.
Por ejemplo, conocer X es suficiente para poder calcular $\mathbb{E}[g(X)]$ para cualquier función g medible.
¿Que podemos decir de la recíproca?
¿Cuando el conocimiento de $\mathbb{E}[g(X)]$ para una colección de g es suficiente para recuperar la distribución de X?

1.1.3. Herramientas de probabilidad para caracterización de distribuciones en $\mathbb{R}$

Función F_X

En el caso de la clase de funciones $\mathcal{G} = \{g_t(x) = \mathbb{1}_{x \leq t} : \ t \in \mathbb{R}\}$, tenemos que $\forall g_t \in \mathcal{G}$, $\mathbb{E}[g_t(X)] = \mathbb{E}[\mathbb{1}_{X \leq t}] = \mathbb{P}(X \leq t) = F_X(t)$. Como ya sabemos que F_X caracteriza la variable X, sabemos que la clase de funciones $\mathcal{G}$ caracteriza X.

Momentos

Sea $\mathcal{G} = \{x \mapsto x^p : p \in \{1, 2, \dots\}\}$. Para una función $g(x) = x^p$, su valor esperado $\mathbb{E}[g(X)] = \mathbb{E}[X^p]$ se llama *momento de orden p* de X.

¡Cuidado! Para algunas variables aleatorias los momentos caracterizan la variable y para algunas otras no...

Entonces, hay teoremas!

1.1.3. Herramientas de probabilidad para caracterización de distribuciones en $\mathbb{R}$

Teorema

Sea P una medida de probabilidad en $\mathbb{R}$, tal que $\forall p \geq 1$, $m_p = \int_{\mathbb{R}} x^p dP(x)$ existe. Supongamos que

$$z \mapsto \sum_{p \geq 1} \frac{m_p}{p!} z^p$$

es de radio de convergencia $R > 0$. Entonces P es completamente definida por sus momentos.

Ejemplo: Los momentos de una variable gaussiana $\mathcal{N}(0, \sigma^2)$ cumplen $m_{n+1} = n\sigma^2 m_{n-1}$ lo que proba que $(m_{n+1}/(n+1)!)/(m_{n-1}/(n-1)!) = \sigma^2/(n+1) \to 0$ y entonces la serie tiene un radio de convergencia $R = +\infty$. **Una variable gaussiana es unicamente definida por sus momentos.**

1.1.3. Herramientas de probabilidad para caracterización de distribuciones en $\mathbb{R}$

Ejemplo:

- La distribución Γ de densidad sobre $\mathbb{R}_+$ dada por

$$x \mapsto \frac{x^{s-1}e^{-x}}{\Gamma(s)}$$

  satisface los supuestos del teorema $\rightarrow$ Una variable Γ es caracterizada por sus momentos.
- Si se sabe que $X \sim \mathcal{N}(\mu, \sigma^2)$ desconocidos, los momentos de orden 1 y 2 son suficientes para caracterizar X.
  Concretamente, $\mathbb{E}[X] = \mu$ y $\sigma^2 = \mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - \mu^2$.
- Pero por ejemplo si $N \sim \mathcal{N}(0,1)$, N^3 o $\exp(N)$ no se caracterizan por sus momentos.

Comentario: Vamos a ver una técnica genérica para crear estimadores basada sobre el calculo de momentos. $\rightarrow$ Método de los momentos.

1.1.3. Herramientas de probabilidad para caracterización de distribuciones en $\mathbb{R}$

Función generatriz de momentos

Sea $\mathcal{G} = \{x \mapsto e^{\lambda x} : \lambda \in \mathbb{R}_+\}$. Para $g(x) = e^{\lambda x}$,
$\mathbb{E}[g(X)] = \mathbb{E}\left[e^{\lambda X}\right]$.
¡Cuidado! De nuevo esa cantidad puede ser infinita en cual caso lo siguiente no tiene sentido.
Cuando $\lambda \in \mathbb{R}_+$ y $\mathbb{E}\left[e^{\lambda X}\right] < +\infty$, uno define la función

$$\begin{array}{rccc} M_X & : & \mathbb{R}_+ & \to & \mathbb{R}_+^* \\ & & \lambda & \mapsto & \mathbb{E}\left[e^{\lambda X}\right] \end{array}$$

Proposición

La función M_X caracteriza la distribución de X.

Proposición (mas general)

Supongamos que existe $\lambda_0 \in \mathbb{R}_+^*$ tq $\forall [0, \lambda_0)$, $\mathbb{E}\left[e^{\lambda X}\right] < +\infty$ entonce la función M_X definida sobre $[0, \lambda_0)$ caracteriza X.

1.1.3. Herramientas de probabilidad para caracterización de distribuciones en $\mathbb{R}$

Ejercicio: Mostrar que si $\mathbb{E}\left[e^{\lambda_0 X}\right] < +\infty$ para un $\lambda_0 \in \mathbb{R}_+^*$ implica que $\forall [0, \lambda_0)$, $\mathbb{E}\left[e^{\lambda X}\right] < +\infty$.

En general si M_X existe, el teorema de convergencia dominada implica que M_X es derivable y

$$M_X'(\lambda) = \mathbb{E}\left[Xe^{\lambda X}\right]$$

lo que permite ver que $M_X'(0) = \mathbb{E}[X]$. De la misma manera $M_X^{(p)}(\lambda) = \mathbb{E}\left[X^p e^{\lambda X}\right]$ y entonces $M_X^{(p)}(0) = \mathbb{E}[X^p]$.

Generatriz de momentos!

1.1.3. Herramientas de probabilidad para caracterización de distribuciones en $\mathbb{R}$

Función característica

Sea $\mathcal{G} = \{x \mapsto e^{itx} : t \in \mathbb{R}\}$. Para $g(x) = e^{itx}$,
$\mathbb{E}[g(X)] = \mathbb{E}\left[e^{itX}\right]$.

Una exponencial compleja tiene módulo 1: $|e^{itx}| = 1 \leftarrow$ integrable, entonces $\forall t\mathbb{R}$, $\mathbb{E}\left[e^{itX}\right]$ existe y está en $\mathbb{C}$.

La función

$$\begin{array}{rccc} \phi_X & : & \mathbb{R} & \to & \mathbb{C} \\ & & t & \mapsto & \mathbb{E}\left[e^{itX}\right] \end{array}$$

se llama *función característica* de X.

1.1.3. Herramientas de probabilidad para caracterización de distribuciones en $\mathbb{R}$

Enunciamos algunas propiedades directas de la función característica.

Proposición

Sea X una variable aleatoria real y ϕ_X su función característica, entonces

1. $\phi_X(0) = 1$,
2. $|\phi_X(t)| \leq 1$,
3. $t \mapsto \phi_X(t)$ *es continua sobre* $\mathbb{R}$,
4. $\phi_{aX+b}(t) = e^{ib}\phi_X(at)$,
5. *Si para* $p \in \mathbb{N}$, $\mathbb{E}\left[|X|^p\right] < \infty$,
   $\phi_X^{(p)}(t) = \mathbb{E}\left[(iX)^p e^{itX}\right]$ y $\phi_X^{(p)}(0) = i^p \mathbb{E}\left[X^p\right]$.

Prueba.

Se usan los teoremas clásicos de análisis. □

1.1.3. Herramientas de probabilidad para caracterización de distribuciones en $\mathbb{R}$

Proposición

Dos variables X y Y son de misma distribución si y solo si

$$\forall t \in \mathbb{R}, \quad \phi_X(t) = \phi_Y(t).$$

Prueba.

Es una consecuencia del teorema de inversión de Fourier. De hecho $\phi_X(t) = \hat{f}_X(-t)$. □

Proposición

Sean X y Y dos variables aleatorias independientes. Entonces, $\forall t \in \mathbb{R}$,

$$\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t).$$

1.1.3. Herramientas de probabilidad para caracterización de distribuciones en $\mathbb{R}$

Prueba.

Sea $t \in \mathbb{R}$,

$$\begin{aligned}\phi_{X+Y}(t) &= \mathbb{E}\left[e^{it(X+Y)}\right] = \mathbb{E}\left[e^{itX}e^{itY}\right] \\ &= \mathbb{E}\left[e^{itX}\right]\mathbb{E}\left[e^{itY}\right] = \phi_X(t)\phi_Y(t).\end{aligned}$$

□

Ejemplo: Sea $X \sim \mathcal{N}(\mu, \sigma^2)$, entonces $\phi_X(t) = e^{it\mu - \frac{\sigma^2 t^2}{2}}$. Para ver eso, empezamos viendo que X y $\sigma N + \mu$ donde $N \sim \mathcal{N}(0,1)$ tienen la misma distribución. Entonces, por la Proposición,

$$\phi_X(t) = e^{it\mu}\phi_N(\sigma t),$$

y entonces es suficiente calcular ϕ_N. La densidad f_N de N es simétrica, por lo tanto $\forall t \in \mathbb{R}$, $\phi_N(t) = \phi_N(-t)$.

1.1.3. Herramientas de probabilidad para caracterización de distribuciones en $\mathbb{R}$

Ademas,

$$\phi_N(t) = \frac{\phi_N(t) + \phi_N(-t)}{2} = \int_{\mathbb{R}} \frac{e^{itx} + e^{-itx}}{2} f_N(x)dx$$
$$= \int_{\mathbb{R}} \cos(tx) \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dz$$

lo que asegura que $\phi_N(t)$ es real. Derivando bajo la integral y usando la integral por partes,

$$\phi'_N(t) = \int_{\mathbb{R}} \sin(tz) \frac{-z}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$
$$= -\int_{\mathbb{R}} t\cos(tz) \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = -t\phi_N(t).$$

Esta ecuación diferencial sencilla tiene como familia de soluciones $\phi(t) = e^{-t^2/2} + C$, pero $\phi(0) = 1$ por lo tanto $C = 0$. Finalmente, la única solución es $\phi_N(t) = e^{-t^2/2}$.

1.1.3. Herramientas de probabilidad para caracterización de distribuciones en $\mathbb{R}$

Caracterización con funciones de prueba

El conjunto $\mathcal{G} = \{g : \mathbb{R} \to \mathbb{R} : g \text{ es lipschitz y acotada}\}$ es también un conjunto de caracterización de la distribución de X. Se usa en el Lema de Portmanteau de caracterización de la convergencia en distribución.

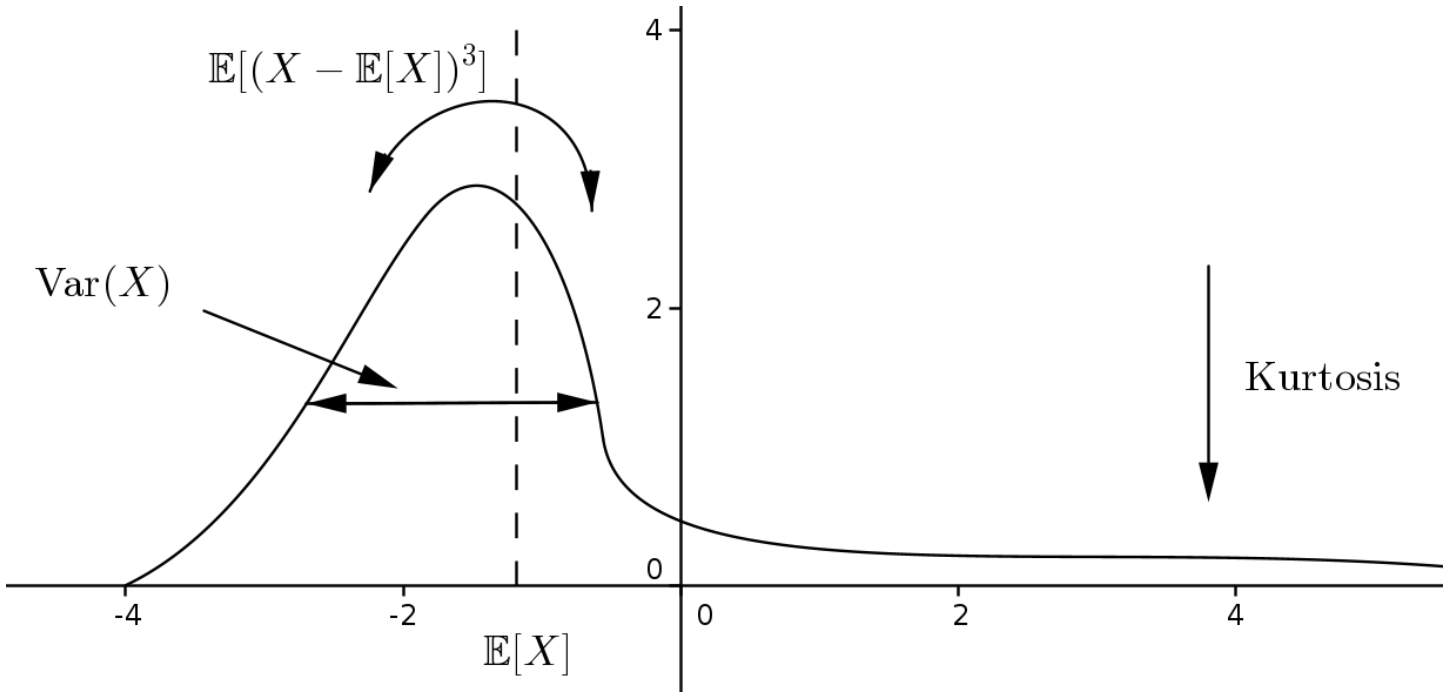
Localización, dispersión, simetría

Hay tres cantidades importantes para describir una distribución cualitativamente.

- $\mathbb{E}[X]$ mide la localización de la distribución.
- $\text{Var}(X)$ mide la dispersión
- $\mathbb{E}\left[(X - \mathbb{E}[X])^3\right]$ mide la simetría.

1.1.3. Herramientas de probabilidad para caracterización de distribuciones en $\mathbb{R}$

La imagen que sirve a interpretar esas cantidades es



A veces se habla de Kurtosis dado por

$$K = \frac{\mathbb{E}\left[(X-\mathbb{E}[X])^4\right]}{\mathrm{Var}(X)^2}$$

Es una medida de la concentración de X o del aplastamiento de las colas de la distribución de X.

1.Preliminares

1.2.Vectores aleatorios

1.2.1. Espacios productos

Basado en las definiciones de variables aleatorias reales vamos a definir vectores aleatorios. Consideramos los espacios vectoriales E de dimensión finita n. Por el teorema de existencia de base, todos los espacios vectoriales están en biyección con $\mathbb{R}^n$. Por esa razón, hacemos toda la teoría unicamente por $E = \mathbb{R}^n$.

σ-algebra producto

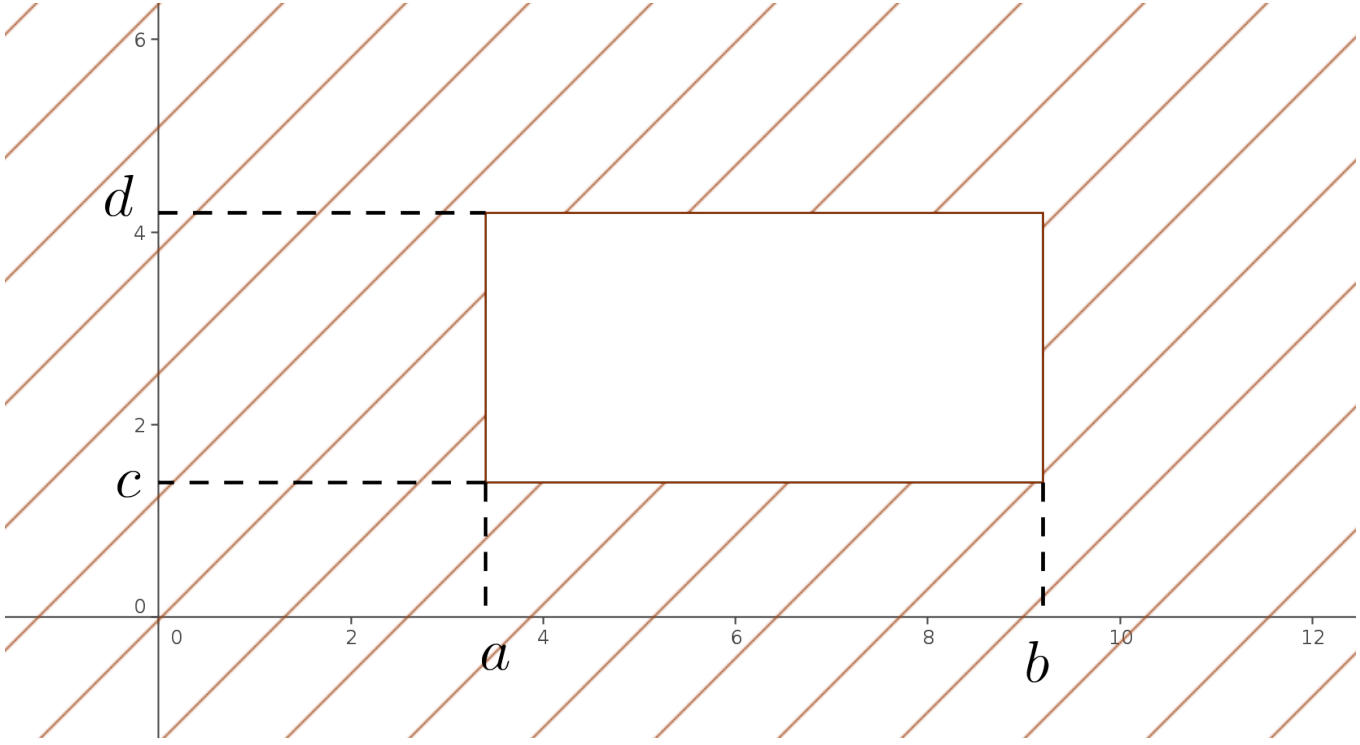
Sean $\mathcal{B}_1$ y $\mathcal{B}_2$ dos σ-algebras sobre $\mathbb{R}$. ¿Como definir a partir de $\mathcal{B}_1$ y $\mathcal{B}_2$ una σ-algebra sobre $\mathbb{R}^2$?

Idea: Considerar $\mathcal{C} = \{B_1 \times B_2 : B_1 \in \mathcal{B}_1 \text{ y } B_2 \in \mathcal{B}_2\}$.

$$\longrightarrow \mathcal{C} \text{ no es una } \sigma\text{-algebra!}$$

1.2.1. Espacios productos

En efecto, si $[a, b] \in \mathcal{B}_1$ y $[c, d] \in \mathcal{B}_2$ entonces el conjunto $([a, b] \times [c, d])^c$ representado:



no se representa de la forma $B_1 \times B_2$!

1.2.1. Espacios productos

Finalmente, uno denota $\mathcal{B} = \sigma(\mathcal{C})$ la σ-algebra **minima** que contiene $\mathcal{C}$. (i.e. es la intersección de todas las σ-algebras con contienen $\mathcal{C}$).

En $\mathbb{R}^2$, denotamos el ensemble de los borelianos $\mathcal{B}^2$.
Ejercicio: Para divertirse, demostrar que $\mathcal{B}^2 = \sigma(\mathcal{B} \times \mathcal{B})$.

Medidas productos

Ya vimos que las medidas P sobre $\mathbb{R}$ son unicamente definidas por los valores de $P([a, b])$.
De la misma manera, las medidas de $\mathbb{R}^2$ son unicamente definidas sobre los conjuntos de la forma $[a, b] \times [c, d]$.

Proposición

Una medida μ sobre un espacio medible (Ω, Σ) es completamente definida por los valores $\mu(C)$ donde $C \in \mathcal{C}$, un π-system que genera Σ.

Un π-system es tq $A, B \in \mathcal{C} \implies A \cap B \in \mathcal{C}$.

1.2.1. Espacios productos

Prueba.

Ver un curso de medida (o el Williams) □

Sean dos medidas μ_1 y μ_2 de $\mathbb{R}$. Para cada a, b, c, d, definimos

$$\mu([a,b] \times [c,d]) = \mu_1([a,b])\mu_2([c,d])$$

que se nota $\mu = \mu_1 \otimes \mu_2$ y se llama *medida producto* de μ_1 y μ_2.

Espacio producto: El espacio $(\mathbb{R}^2, \mathcal{B}^2, \mu)$ se llama *espacio producto* de $(\mathbb{R}, \mathcal{B}, \mu_1)$ y $(\mathbb{R}, \mathcal{B}, \mu_2)$.

Pregunta-Ejercicio: Pensar en la diferencia/relación entre espacio producto y variables independientes.

1.2.1. Espacios productos

- Las medidas que se escriben $\mu([a,b]\times[c,d]) = \nu_1([a,b])\nu_2([c,d])$ se llaman medidas productos.
- Las medidas productos son un conjunto de las medidas de $\mathbb{R}^2$. En efecto, es suficiente de verificar los tres axiomas de definición.
- La función $F_{X,Y}(x,y) = \mathbb{E}\left[\mathbb{1}_{(-\infty,x]\times(-\infty,y]}\right] = \mathbb{P}(X \leq x, Y \leq y)$ se llama *función de distribución multivariada*.
- Se puede recuperar las medidas originales μ_1 y μ_2 tomando, por ejemplo, los conjuntos $[a,b]\times\mathbb{R}$.

Se puede generalizar para cualquier medida $\rightarrow$ Marginales.

1.2.1. Espacios productos

Marginales

Es una descripción de la medida a través de la medida sobre cada componente del espacio global.

- Sea μ una medida finita sobre $(\mathbb{R}^2, \mathcal{B}^2)$. Para cada $B \in \mathcal{B}$ (borelianos de $\mathbb{R}$) la aplicación

$$B \mapsto \mu(B \times \mathbb{R})$$

  define una medida de $(\mathbb{R}, \mathcal{B})$ que llamamos *marginal* de μ en la primera componente/coordenada. La denotamos $\mu_1(B) = \mu(B \times \mathbb{R})$.
- Igualmente, definimos $\mu_2(B) = \mu(\mathbb{R} \times B)$.
- Cuando μ es una medida de probabilidad producto,

$$\forall a, b,\ \mu_1([a,b]) = \mu([a,b] \times \mathbb{R}) = \mu_1([a,b])\mu_2(\mathbb{R}) = \mu_1([a,b])$$

  y la notación es consistente.

1.2.1. Espacios productos

Es directo de tener

Proposición

Las marginales de una medida de probabilidad son probabilidades.

Prueba.

Se puede verificar los axiomas de medida sobre $\mu_1(B) = \mu(B \times \mathbb{R})$ y $\mu_1(\mathbb{R}) = \mu(\mathbb{R} \times \mathbb{R}) = 1$. □

Cálculos de marginales

- En el caso general, μ_1 es completamente caracterizada por los valores $\mu([a, b] \times \mathbb{R})$ para cada a, b.
- Si μ tiene una densidad f, $\forall B \in \mathcal{B}^2$,

$$\mu(B) = \int_B f(x, y)dxdy \qquad \text{(por definición)}$$

1.2.1. Espacios productos

- Entonces $\forall a, b$ tq $a \leq b$,

$$\begin{aligned}\mu_1([a,b]) &= \mu([a,b] \times \mathbb{R}) \\ &= \int_{[a,b]} f(x,y)dxdy \\ &\overset{Fubini}{=} \int_a^b \Big(\int_{\mathbb{R}} f(x,y)dy \Big) dx \\ &= \int_a^b g(x)dx\end{aligned}$$

  donde $g(x) = \int_R f(x,y)dy$. Entonces μ_1 tiene densidad dada por $g(x)$.

- De la misma manera, μ_2 tiene la densidad dada por

$$h(y) = \int_{\mathbb{R}} f(x,y)dx.$$

1.2.1. Espacios productos

Espacios $(\mathbb{R}^n, \mathcal{B}^n, P_X)$

Las definiciones que vimos en dimensión 2 se generalizan en dimensión n.

- $\mathbb{R}^n = \underbrace{\mathbb{R} \times \cdots \times \mathbb{R}}_{n\text{veces}}$.
- $\mathcal{B}^n = \sigma(\{B_1 \times \cdots \times B_n : \forall i,\ B_i \in \mathcal{B}^1\})$.
- P_X es la distribución de X como elemento de $\mathbb{R}^n$.

La variable aleatoria definida en el espacio $(\mathbb{R}^n, \mathcal{B}^n, P_X)$ se llama *vector aleatorio* de dimensión n.

Por ejemplo, la primera marginal sera dada por

$$f_1(x_1) = \int_{\mathbb{R}^{n-1}} f(x_1, x_2, \ldots, x_n) dx_2 \ldots dx_n.$$

1.2.2. Cálculos multivariados

En esta sección, vamos a usar todo el potencial de los cálculos de $\mathbb{E}[g(X)]$. Otra vez, empezamos con los conceptos en $\mathbb{R}^2$.

Probar que una probabilidad es producto

Si P es una producto sobre $\mathbb{R}^2$, $\exists P_1, P_2$ tal que $P = P_1 \otimes P_2$.
Sean $f, g : \mathbb{R} \longrightarrow \mathbb{R}$ integrables,

$$\begin{aligned}
\mathbb{E}[f(X)g(Y)] &= \int_{\mathbb{R}^2} f(x)g(y)dP(x,y) = \int\int_{\mathbb{R}\times\mathbb{R}} f(x)g(y)dP_1(x)dP_2(y) \\
&= \int f(x)\Big(\int_{\mathbb{R}} g(y)dP_2(y)\Big)dP_1(x) \\
&= \Big(\int_{\mathbb{R}} f(x)dP_1(x)\Big)\Big(\int_{\mathbb{R}} g(y)dP_2(y)\Big) \\
&= \mathbb{E}[f(X)]\,\mathbb{E}[g(Y)].
\end{aligned}$$

Lo que demuestra que X y Y son independientes $\Leftrightarrow$ el vector (X, Y) tiene una ley $P_{X,Y}$ producto.

1.2.2. Cálculos multivariados

Sean X and Y dos variables reales, lo siguientes hechos con equivalentes:

1. X y Y son independientes.
2. (X, Y) tiene una ley producto.
3. $\forall f, g$ funciones lipschitz acotadas $\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\,\mathbb{E}[g(Y)]$.
4. $F_{X,Y}(x, y) = F_X(x)F_Y(y)$ (que sería de definición de $F_{X,Y}(x, y)$?)
5. $\forall a, b, c, d$, $\mathbb{P}_{X,Y}([a, b] \times [c, d]) = \mathbb{P}_X([a, b])\,\mathbb{P}_Y([c, d])$
6. Si X y Y tienen densidad, $f_{X,Y}(x, y) = f_X(x)f_Y(y)$.
7. Si X y Y tienen densidad, $\forall a, b, c, d$

$$\int_{[a,b]\times[c,d]} f_{X,Y}(x, y)dxdy = \int_a^b f_X(x)dx \int_c^d f_Y(y)dy$$

1.2.2. Cálculos multivariados

Ejemplo de cálculo de marginales

Imaginemos que tenemos n cajas tal que la caja k contiene k bolas numeradas $1, \ldots, k$. Se selecciona una caja X al azar y luego una bola Y al azar en la caja X. El soporte del vector es $\Gamma = \{(k,j) \in \{1,\ldots,n\}^2 : j \leq k\}$.

$$\forall k,j \in \{1,\ldots,n\}^2, \quad \mathbb{P}(X=k, Y=j) = \frac{1}{n} \times \frac{1}{k} \mathbb{1}_{j \leq k}.$$

La marginal de Y es

$$\begin{aligned} \mathbb{P}(Y=j) &= \sum_{k=1}^{n} \mathbb{P}(X=k, Y=j) \\ &= \sum_{k=j}^{n} \frac{1}{nk} = \frac{1}{n} \sum_{k=j}^{n} \frac{1}{k} \end{aligned}$$

La marginal de X es $\mathbb{P}(X=k) = 1/n$ lo que permite ver que la ley de (X,Y) no es una probabilidad producto.

1.2.2. Cálculos multivariados

Ejemplo de cálculo de marginales

Sea (X, Y) un vector aleatorio de densidad $f_{X,Y}(x, y) = 8x(x - y)\mathbb{1}_{0 \leq y < x \leq 1}$, entonces

$$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x, y)dy$$
$$= \int_0^x 8x(x - y)dy = 8x^3 - 8x\left[\frac{1}{2}y^2\right]_0^x = 4x^3.$$

Y

$$f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x, y)dx = \int_y^1 8x(x - y)dx$$
$$= 8\left[\frac{1}{3}x^3\right]_y^1 - 8y\left[\frac{1}{2}x^2\right]_y^1 = 8\left(\frac{1}{3} - \frac{y^3}{3}\right) - 8y\left(\frac{1}{2} - \frac{y^2}{2}\right)$$
$$= \frac{8}{3} - 4y + \frac{4}{3}y^3.$$

Entonces $f_{X,Y}(x, y) \neq f_X(x)f_Y(y)$.

1.2.2. Cálculos multivariados

Ejemplo de cálculo de marginales

Sean X and Y dos variables independientes de distribución $\mathcal{E}(\lambda)$ y $\mathcal{E}(\nu)$. Calcular $P(X > Y)$.
El vector (X, Y) tiene la densidad

$$f_{X,Y}(x, y) = \lambda e^{-\lambda x} \mathbb{1}_{x \geq 0} \times \nu e^{-\nu y} \mathbb{1}_{y \geq 0}$$

Entonces

$$\begin{aligned}
\mathbb{P}(X > Y) &= \int_{\mathbb{R}^2} \mathbb{1}_{x>y} f_{X,Y}(x, y) dx dy \\
&= \int_{\mathbb{R}} \nu e^{-\nu y} \mathbb{1}_{y \geq 0} \Big(\int_y^{+\infty} \lambda e^{-\lambda x} \mathbb{1}_{x \geq 0} dx \Big) dy \\
&= \int_0^{+\infty} \nu e^{-\nu y} e^{-\lambda y} dy \\
&= \nu \Big[\frac{1}{\lambda + \nu} e^{-(\lambda+\nu)y} \Big]_0^{+\infty} = \frac{\nu}{\lambda + \nu}
\end{aligned}$$

1.2.2. Cálculos multivariados

Transformar un vector (X, Y)

Sabemos que $\mathbb{E}[g(X, Y)]$ para todo g lipschitz caracteriza la distribución de (X, Y).
Si uno transforma (X, Y) en $(U, V) = \phi(X, Y) \in \mathbb{R}^2$ tal que la transformada es $\mathcal{C}^1$ y biyectiva, se puede caracterizar la distribución de (U, V) con la técnica del cambio de variables siguiente. Sea g lipschitz y acotada.

$$\begin{aligned}\mathbb{E}[g(U, V)] &= \mathbb{E}[g(\phi(X, Y))] \\ &= \int_{\mathbb{R}^2} g(\phi(x, y)) f_{X,Y}(x, y) dxdy \\ &= \int_D g(u, v) f_{X,Y}(\phi^{-1}(u, v)) |J_{\phi^{-1}}| dudv\end{aligned}$$

Esto siendo cierto para cada función g lipschitz y acotada, tenemos que $f_{U,V}(u, v) = f_{X,Y}(\phi^{-1}(u, v)) |J_{\phi^{-1}}| \mathbb{1}_D(u, v)$.

1.2.2. Cálculos multivariados

Ejemplo :

Sean X y Y dos variables independientes de distribución $\mathcal{E}(\lambda)$ y $\mathcal{E}(\nu)$. Consideramos el vector aleatorio $(X+Y, 2X-Y)$. La función ϕ está dada por $\phi(x,y) = (x+y, 2x-y)$ y $\phi^{-1}(u,v) = (\frac{u+v}{3}, \frac{2u-v}{3})$. Tenemos

$$|J_{\phi^{-1}}| = \begin{vmatrix} \frac{1}{3} & \frac{1}{3} \\ \frac{2}{3} & -\frac{1}{3} \end{vmatrix} = \frac{1}{3},$$

y por las restricciones

$$x \geq 0 \Leftrightarrow u+v \geq 0 \Leftrightarrow v \geq -u$$
$$y \geq 0 \Leftrightarrow 2u-v \geq 0 \Leftrightarrow v \leq 2u$$

el dominio vale $D = \{(u,v) : -u \leq v \leq 2u \text{ y } u \geq 0\}$, lo que nos da

$$f_{U,V}(u,v) = \frac{\lambda\nu}{3} e^{-(\lambda+2\nu)u/3} e^{-(\lambda-\nu)v/3} \mathbb{1}_{-u\leq v\leq 2u} \mathbb{1}_{u\geq 0}.$$

1.2.2. Cálculos multivariados

Ahora se puede calcular la marginal f_U por ejemplo. Si $\lambda \neq \nu$, para $u \geq 0$,

$$\begin{aligned} f_U(u) &= \int_{\mathbb{R}} f_{U,V}(u,v)dv \\ &= \frac{\lambda\nu}{3} e^{-(\lambda+2\nu)u/3} \int_{-u}^{2u} e^{-(\lambda-\nu)v/3} dv \\ &= \frac{\lambda\nu}{3} e^{-(\lambda+2\nu)u/3} \left[\frac{-3}{\lambda-\nu} e^{-(\lambda-\nu)v/3} \right]_{-u}^{2u} \\ &= \frac{\lambda\nu}{\nu-\lambda} e^{-(\lambda+2\nu)u/3} \left[e^{-(\lambda-\nu)v/3} \right]_{-u}^{2u} \\ &= \frac{\lambda\nu}{\nu-\lambda} \left(e^{-\lambda u} - e^{-\nu u} \right) \end{aligned}$$

Si $\lambda = \nu$, $f_U(u) = \frac{\lambda^2}{3} e^{-\lambda u} \times 3u = \lambda^2 u e^{-\lambda u}$ sobre $\mathbb{R}_+$. es una variable aleatoria Gamma.

1.2.2. Cálculos multivariados

Si $\lambda = \nu$

$$
\begin{aligned}
f_V(v) &= \int_{\mathbb{R}} f_{U,V}(u,v)du = \frac{\lambda^2}{3}\int_{\max(v/2,-v)}^{+\infty} e^{-\lambda u}du \\
&= \frac{\lambda}{3}\left[-e^{-\lambda u}\right]_{\max(v/2,-v)}^{+\infty} \\
&= \frac{\lambda}{3}e^{-\lambda \max(v/2,-v)} \\
&= \frac{\lambda}{3}\left(e^{-\frac{\lambda}{2}v}\mathbb{1}_{v\geq 0} + e^{\lambda v}\mathbb{1}_{v<0}\right).
\end{aligned}
$$

Esa variable es distribuida sobre todo $\mathbb{R}$ y tiene dos colas asimétricas del lado izquierdo y derecho.

1.2.2. Cálculos multivariados

Notación $\mathbb{E}[g(X)]$ para $g : \mathbb{R}^n \to \mathbb{R}^m$

Para definir $\mathbb{E}[g(X)]$ usamos las coordenadas. Denotamos $g_1, \ldots, g_m$ las funciones componentes de g:

$$g(x) = \begin{pmatrix} g_1(x) \\ \vdots \\ g_m(x) \end{pmatrix}$$

y definimos usando la definición $\mathbb{E}[g_i(X)]$ para funciones reales

$$\mathbb{E}[g(X)] = \begin{pmatrix} \mathbb{E}[g_1(X)] \\ \vdots \\ \mathbb{E}[g_m(X)] \end{pmatrix}$$

Como consecuencia, el objecto $\mathbb{E}[g(X)]$ es un vector de $\mathbb{R}^m$. Por ejemplo, $\mathbb{E}[(X, Y)] = (\mathbb{E}[X], \mathbb{E}[Y])$.

1.2.2. Cálculos multivariados

Notación $\mathbb{E}[g(X)]$ para $g : \mathbb{R}^n \to \mathbb{C}$

Tratamos $\mathbb{C}$ como un $\mathbb{R}$-espacio vectorial absolutamente normal. Entonces $g(x) = \Re(g(x)) + i\Im(g(x))$ lo que implica que $\mathbb{E}[g(X)] = \mathbb{E}[\Re(g(X))] + i\mathbb{E}[\Im(g(X))]$.

Notación $\mathbb{E}[g(X)]$ para $g : \mathbb{R}^n \to \mathcal{M}_{m\times\ell}(\mathbb{R})$

También, $\mathcal{M}_{m\times\ell}(\mathbb{R})$ es un espacio lineal. Su base es $E_{i,j}$ las matrices con puros 0 salvo en la posición i, j donde vale 1. Entonces,

$$g(x) = \begin{pmatrix} g_{1,1}(x) & \cdots & g_{1,\ell}(x) \\ \vdots & & \vdots \\ g_{m,1}(x) & \cdots & g_{m,\ell}(x) \end{pmatrix}$$

y entonces

$$\mathbb{E}[g(X)] = \begin{pmatrix} \mathbb{E}[g_{1,1}(X)] & \cdots & \mathbb{E}[g_{1,\ell}(X)] \\ \vdots & & \vdots \\ \mathbb{E}[g_{m,1}(X)] & \cdots & \mathbb{E}[g_{m,\ell}(X)] \end{pmatrix}$$

1.2.2. Cálculos multivariados

Caso particular

Sea X un vector aleatorio de $\mathbb{R}^n$. La esperanza de la función $g : \mathbb{R} \to \mathcal{M}_{n\times n}(\mathbb{R})$ tal que $g(x) = (x - \mathbb{E}[X])(x - \mathbb{E}[X])^T$ se llama *matriz de (varianza-)covarianza* de X. Se escribe,

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}\left[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T\right] \\
&= \begin{pmatrix} & \vdots & \\ \cdots & \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])] & \cdots \\ & \vdots & \end{pmatrix} \\
&= \begin{pmatrix} \mathrm{Var}(X_1) & \cdots & \mathrm{Cov}(X_1, X_n) \\ & \ddots & \vdots \\ & & \mathrm{Var}(X_n) \end{pmatrix}
\end{aligned}$$

y la cantidad $\mathrm{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])]$ se llama *covarianza* de X_i y X_j.

1.2.2. Cálculos multivariados

Función característica multivariada

La función característica de dimensión n es la esperanza de la función

$$\begin{array}{rccc} g_t & : & \mathbb{R}^n & \to & \mathbb{C} \\ & & x & \mapsto & e^{i(t.x)} \end{array}$$

donde $t.x$ designa es producto interior de los dos vectores t y x de $\mathbb{R}^n$. De nuevo, denotamos

$$\phi_X(t) = \mathbb{E}\left[e^{i(t.X)}\right]$$

y la llamamos *función característica* de X.

Estamos usando la integración de una función a valores vectoriales.

Proposición

La función característica caracteriza la distribución del vector X.

1.2.3. Ley de vectores aleatorios

De manera análoga a las variables aleatorias reales, podemos hablar de la ley P_X del vector aleatorio X. Cada variable-componente del vector aleatorio tienen también una distribución.

Cuidado! La ley/distribución de un vector aleatorio no está completamente definido por sus marginales!!

Ejemplo: Sea $Z_1 = (X, Y)$ y $Z_2 = (W, 1 - W)$ donde las tres variables reales X, Y, W siguen una distribución $\mathcal{U}([0, 1])$ y son independientes. En particular, los dos vectores Z_1 y Z_2 tienen las mismas leyes marginales ($\mathcal{U}([0, 1])$), pero no tienen la misma distribución/ley. En efecto,

$$\mathbb{P}_{Z_1}((1/2, 1] \times (1/2, 1]) = \mathbb{P}_X((1/2, 1])\,\mathbb{P}_Y((1/2, 1]) = (1/2)^2$$
$$\mathbb{P}_{Z_2}((1/2, 1] \times (1/2, 1]) = \mathbb{P}_W(W \geq 1/2 \text{ y } W < 1/2) = 0.$$

En particular, cuando uno describe un vector aleatorio, no es suficiente de solo calcular sus marginales.

1.2.3. Ley de vectores aleatorios

Función generatriz

Supongamos que $\exists r > 0$ tal que $\forall t \in B_2(0, r)$, $\mathbb{E}\left[e^{t.X}\right] < \infty$.
Entonces, la función

$$\begin{array}{rccc} M_X & : & B_2(0,r) & \to & \mathbb{R} \\ & & t & \mapsto & \mathbb{E}\left[e^{(t.x)}\right] \end{array}$$

existe y se llama *función generatriz de los momentos mixtos* de X.

De nuevo, se puede derivar la función M_X y obtener

$$\partial_i M_X(t) = \mathbb{E}\left[X_i e^{t.X}\right]$$
$$\partial_i M_X(0) = \mathbb{E}\left[X_i\right]$$

Ademas, uno puede calcular $\partial_i \partial_j M_X(0) = \mathbb{E}\left[X_i X_j\right]$.

1.2.3. Ley de vectores aleatorios

Extendiendo el calculo anterior, se ve que se puede calcular todos los momentos mixtos del vector X:

$$m_{\alpha_1,\ldots,\alpha_n} = \mathbb{E}\left[X_1^{\alpha_1} \ldots X_n^{\alpha_n}\right].$$

Una pregunta natural es de saber si los momentos caracterizan la distribución de X.

- Los momentos simples $\rightarrow$ seguramente que no!
- Los momentos mixtos $\rightarrow$ Hay condiciones! Los teoremas son generalizaciones en dimensión n del teorema de caracterización por momentos en dimensión 1.
- Hay una grande literatura! Es el tópico de probabilidades libres en espacios de probabilidades no conmutativas. Los interesados puedes ir a ver Topics in Random Matrix Theory.

1.2.3. Ley de vectores aleatorios

Funciones de prueba

Tenemos de nuevo:

Proposición

Un vector aleatorio $X \in \mathbb{R}$ está completamente definido por los valores de $\mathbb{E}[h(X)]$ para cada $h : \mathbb{R}^n \to \mathbb{R}$ lipschitz y acotada.

Cuantiles

No hay una manera obvia de definir los cuantiles en $\mathbb{R}^n$! (**Dibujar para darse cuenta**)

Una definición posible de la mediana empírica (sobre un conjunto finito de datos) es:

Sean $X_1, \ldots, X_k$ una muestra de v.a. de $\mathbb{R}^n$. La cantidad

$$\overline{m}_X = \operatorname*{argmin}_{y \in \mathbb{R}^n} \sum_{i=1}^{k} \|y - X_i\|_2$$

se llama *mediana geométrica* de la muestra $X_1, \ldots, X_k$

1.2.3. Ley de vectores aleatorios

Una otra manera de definir una mediana puede, por ejemplo, ser la profundidad de Tukey.

Sea $X_1, \ldots, X_k \in \mathbb{R}^n$ y $p \in \mathbb{R}^n$. Definimos

$$Tukey(p) = \min_{H \in \mathcal{H}_p} |H \cap \{X_1, \ldots, X_k\}|$$

donde $\mathcal{H}_p$ son los hiperplanos que contienen p, como la *profundidad de Tukey*. Es una medida de la centralidad del punto p. Por eso, podemos definir

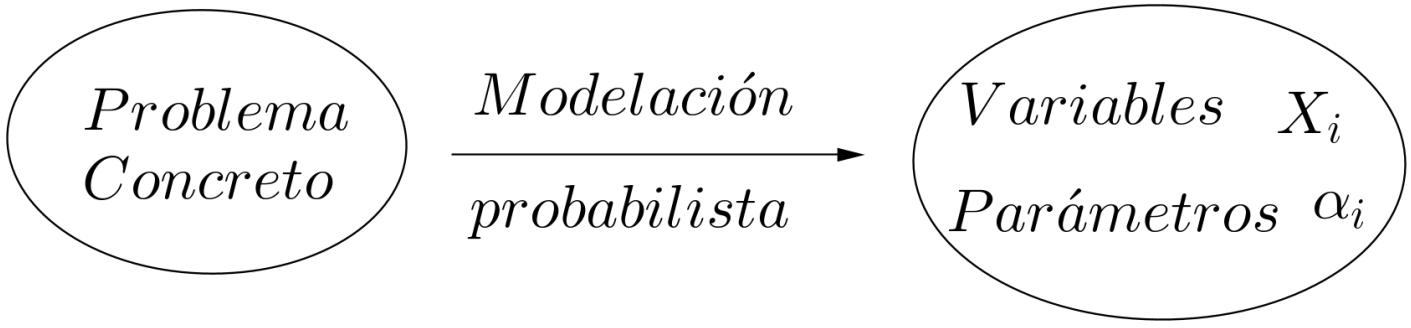
$$\text{med}(X_1, \ldots, X_k) = \operatorname*{argmax}_{p \in \mathbb{R}^n} Tukey(p)$$

1.Preliminares

1.3.Modelos de probabilidad, modelos de estadística

1.3.1. Diferencia entre probabilidad y estadística

Que es según ustedes la diferencia entre probabilidad y estadística?



Mundo probabilista

- Se suponen **conocidos** los parámetros.
- Se estudian las variables.

Mundo Estadístico

- Se suponen **conocidas** las variables.
- Se estudian los parámetros.

1.3.1. Diferencia entre probabilidad y estadística

En el estudio de modelos estadísticos se presentan dos contextos diferentes. Gracias a varias realizaciones de la misma distribución, queremos adivinar el parámetro desconocido de dos maneras:

Mundo Frecuentista

- Los parámetros se consideran como elementos deterministas de $\mathbb{R}^k$.
- Los resultados toman la forma de intervalos de confianza alrededor del punto de estimación.

Mundo Bayesiano

- Los parámetros se consideran como variables aleatorias.
- Los resultados toman la forma de distribuciones lo mas concentrado posible alrededor del verdadero parámetro.

1.3.2. Modelos estadísticos

Espacios $(\Omega, \mathcal{B}, P_\theta, \theta \in \Theta)$

- En la modelación probabilista, los parámetros α_i están en el espacio Θ y los denotamos θ. Cuando hay mas que uno, seguimos denotando $\theta = (\theta_1, \ldots, \theta_k) \in \mathbb{R}^k$.
- Para parámetros $\theta \in \Theta$, la distribución de las variables X_i (que van a tener la misma distribución) se escribe P_θ.
- Ejemplo: $\Omega = \mathbb{R}$, $\mathcal{B}$ los borelianos, $\Theta = \mathbb{R}$ y $\forall \theta \in \Theta$, $P_\theta \sim \mathcal{N}(\theta, 1)$.
- El ensemble $(\Omega, \mathcal{B}, P_\theta, \Theta)$ se llama *modelo de probabilidad*.

Modelos parámetricos y no-parámetricos

- Cuando el espacio Θ es un sub-espacio de $\mathbb{R}^k$ se dice que el modelo es *parámetrico*.
- Cuando no es el caso, se dice que el modelo es *no-parámetrico*.

1.3.2. Modelos estadísticos

Ejemplos:

- El modelo anterior $(\mathbb{R}, \mathcal{B}, \mathbb{R}, (\mathcal{N}(\theta, 1))_{\theta \in \Theta})$ es parámetrico.
- El modelo $(\mathbb{R}, \mathcal{B}, \Theta, (P_f)_{f \in \Theta})$ donde Θ es el espacio de las funciones $f : \mathbb{R} \to [0, 1]$ lipschitz y creciente tal que $\lim_{-\infty} f = 0$ y $\lim_{+\infty} f = 1$.

  Una probabilidad P_f está definida por su función de distribución $F = f$.

  El modelo es no-parámetrico.
- A veces la denominación de modelos semi-parámetricos está usada. Es un caso particular de modelos no-parámetricos. Se puede usar cuando:
  - Una parte de θ está en espacios $\mathbb{R}^k$.
  - El espacio Θ tiene una estructura diferencial particular que permite una cierta discretización.

1.3.2. Modelos estadísticos

Conocido vs Desconocido: En el contexto estadístico,

- θ es desconocido,
- Θ es conocido. (Por abuso de vocabulario, lo llamamos también modelo)

Observación

Una *observación* es una variable aleatoria X tal que $P_X = P_{\theta_0}$ donde θ_0 es el parámetro desconocido que buscamos.

Muestra

Para $n \in \mathbb{N}$, una *n-muestra* de ley ν es una colección $X_1, \ldots, X_n$ de variables i.i.d. de distribución común ν.

Cuando una observación tiene la forma de una n-muestra de ley ν_θ, entonces $P_\theta = (\nu_\theta)^{\otimes n}$. El espacio asociado $(\Omega^n, \mathcal{B}, P_\theta, \theta \in \Theta)$ se llama *modelo estadístico*.

1.3.2. Modelos estadísticos

Modelos identifiables

Para determinar un parámetro $\theta \in \Theta$ sin ambigüedad la función $\theta \mapsto P_\theta$ es inyectiva.

Un modelo tal que $\theta \mapsto P_\theta$ es inyectiva se llama *modelo identifiable*.

Error de especificación

Es posible que la verdadera ley P_{θ_0} que buscamos no esté en la clase de leyes candidatas $(P_\theta)_{\theta\in\Theta}$. Eso tiene que ver con la surjetividad de $\theta \mapsto P_\theta$. En este caso, no se podrá ajustar perfectamente el modelo estadístico a los datos. En el caso parámetrico, denotamos

$$E(\Theta) = \min_{\theta\in\Theta} \|\theta - \theta_0\|$$

que llamamos el *error de especificación* del modelo.

1.Preliminares

1.4.Cantidades empíricas

1.4.1. Distribución empírica

El primer paso para definir cantidades que permiten definir estimadores de nuestros parámetros desconocidos es a través de lo que llamamos cantidades empíricas.

Definición de P_n y P_nf

En todo lo que sigue, supongamos tener acceso a una muestra $X_1, \ldots, X_n$ de ley común P_{θ_0}.

- Denotamos $P_n = \sum_{i=1}^n \frac{1}{n}\delta_{X_i}$ la medida (aleatoria) que carga con masa $\frac{1}{n}$ cada punto X_i de la muestra. Esa medida se llama *medida empírica* asociada con la muestra $X_1, \ldots, X_n$.
- Para una medida P y una función f integrable, denotamos

$$Pf = \int f dP$$

  la integral al respecto de P.
- Para la medida $P = P_n$, el calculo anterior da

$$P_nf = \int f dP_n = \frac{1}{n}\sum_{i=1}^n f(X_i)$$

1.4.1. Distribución empírica

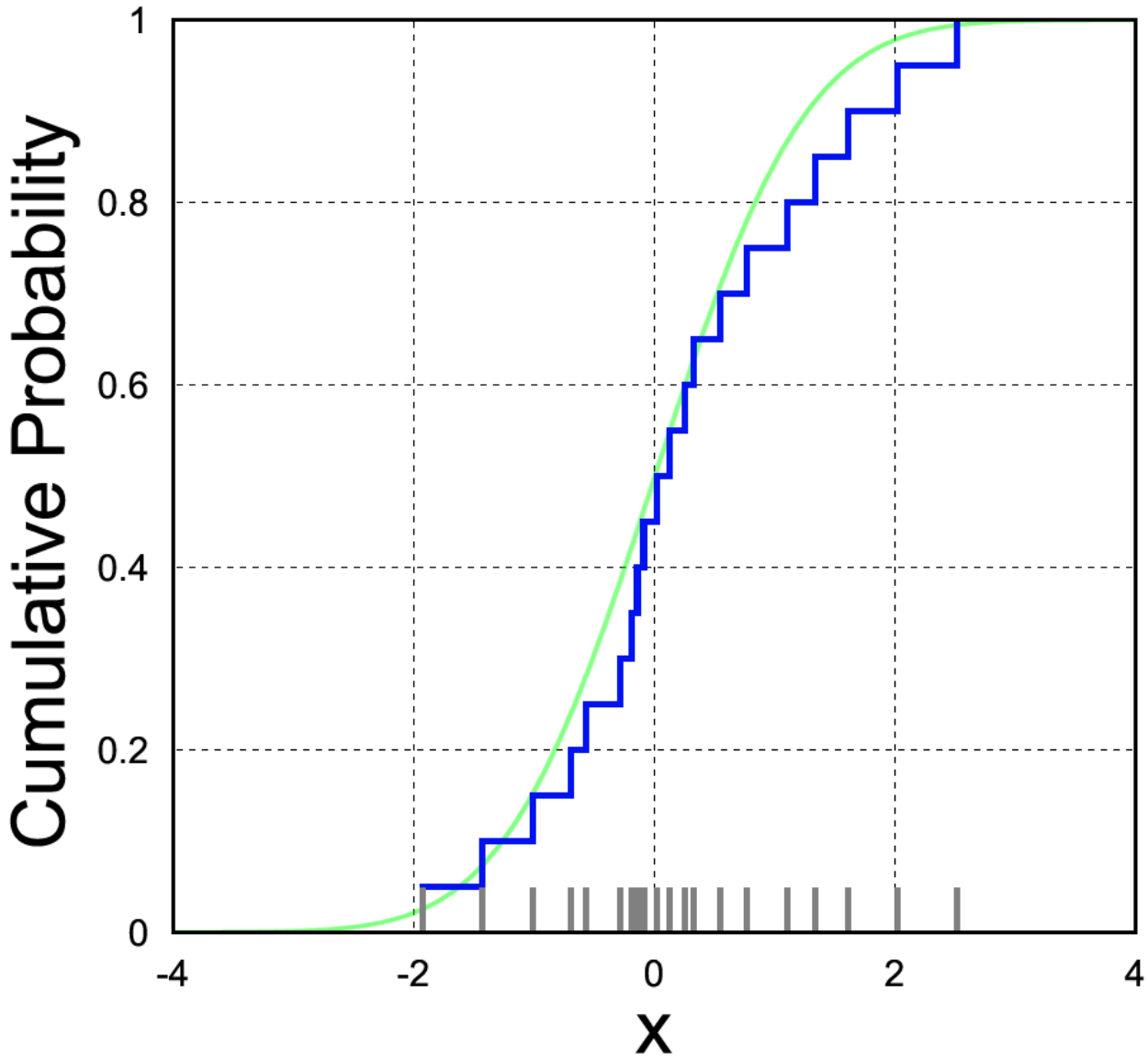
Esta definición es útil para definir varias cantidades que dependen de la muestra $X_1, \ldots, X_n$.

Función de distribución empírica, momentos empíricos

- Si tomamos $f = \mathrm{id}$, entonces $P_n f = \frac{1}{n} \sum_{i=1}^{n} X_i$. Es la *media empírica*.
- Si tomamos $f_t(x) = \mathbb{1}_{x \leq t}$, entonces $P_n f_t = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq t}$. La función $F_n : t \mapsto P_n f_t$ y se llama *función de distribución empírica*.
- Si tomamos $f(x) = x^k$, entonces $P_n f = \frac{1}{n} \sum_{i=1}^{n} X_i^k$ y se llama el momento empírico de orden k de la distribución P_{θ_0}.

1.4.1. Distribución empírica

El dibujo de la función de distribución empírica:



1.4.1. Distribución empírica

Veremos al final de este capitulo que por el teorema de la ley de grandes números:

Proposición

Sean $X_1, \ldots, X_n$ variables i.i.d. de distribución común X. Entonces, for all $t \in \mathbb{R}$,

$$F_n(t) \underset{n \to +\infty}{\longrightarrow} F_X(t)$$

De hecho el resultado es aún mas fuerte en caso de variables a densidad.

Proposición (Glivenko-Cantelli)

Sean $X_1, \ldots, X_n$ variables i.i.d. de distribución común X que tiene densidad. Entonces, for all $t \in \mathbb{R}$,

$$\sup_{t \in \mathbb{R}} |F_n(t) - F_X(t)| \underset{n \to +\infty}{\longrightarrow} 0$$

1.4.1. Distribución empírica

Las mismas convergencias se pueden probar sobre las otras cantidades empíricas que encontramos antes bajo condiciones de existencia de momentos.

Proposición

Sean $X_1, \ldots, X_n$ variables i.i.d. de distribución común X. Supongamos que $\mathbb{E}[X] < +\infty$ (resp. $P|X|^k < \infty$ para algún $k > 0$). Entonces,

$$\frac{1}{n}\sum_{i=1}^{n} X_i = P_n x \underset{n\to+\infty}{\longrightarrow} \mathbb{E}[X]$$

(resp.

$$P_n x^k \underset{n\to+\infty}{\longrightarrow} P x^k \quad \text{también denotado} \quad (P_n - P)x^k \underset{n\to+\infty}{\longrightarrow} 0.)$$

1.4.2. Estimadores

Estamos listos para definir lo que llamaremos un estimador a lo largo de este curso.

Definición de estimador

Dado una muestra $X_1, \ldots, X_n$ de variables aleatorias de ley común P, definimos un *estimador* como una función $\Omega^n \to \Theta$ medible al respecto de la medida producto $P^{\otimes n}$ del vector aleatorio $X_1, \ldots, X_n$.

Comentario: Esa definición dice en otras palabras que un estimador es una función que depende de las variables $X_1, \ldots, X_n$ como única fuente de azar. Cuando se espera que un estimador sea una estimación de un cierto parámetro θ, el estimador se denota

$$\hat{\theta}_n(X_1 \ldots, X_n) \quad \text{o} \quad \hat{\theta}_n \quad \text{o} \quad \hat{\theta}$$

1.4.2. Estimadores

Vocabulario de estimadores

- Un estimador es en particular una variables aleatoria. Se usa todo lo anterior para estudiar lo.
- Se dice que un estimador es *sin sesgo* si $\mathbb{E}[\hat{\theta}] = \theta$. En el caso contrario, se dice que el estimador tiene *sesgo* dado por $\mathbb{E}[\hat{\theta}] - \theta$.
- Se dice que $\hat{\theta}$ es *consistente o converge* si

$$\hat{\theta} \underset{n \to +\infty}{\xrightarrow{\mathbb{P}}} \theta$$

  donde la convergencia se hace en probabilidad.
- Se dice que $\hat{\theta}$ es *fuertamente consistente* si

$$\hat{\theta} \underset{n \to +\infty}{\xrightarrow{c.s.}} \theta$$

1.4.2. Estimadores

Las dos convergencias precedentes se refieren a :

1. Se dice que X_n converge en *probabilidad* hasta X si $\forall \epsilon > 0$,

$$\mathbb{P}\left(|X_n - X| > \epsilon\right) \underset{n \to +\infty}{\longrightarrow} 0$$

2. Se dice que X_n converge en *casi seguramente* hasta X si

$$\mathbb{P}\left(\{\omega : X_n(\omega) - X(\omega) \text{ no converge a } 0\}\right) = 0$$

Riesgo

En general uno quiere tener mas información sobre la cualidad de un estimador que la consistencia. La cantidad la mas usada es seguramente

$$R(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right].$$

1.4.2. Estimadores

El riesgo $R(\hat{\theta})$ se puede estudiar con la siguiente observación.

$$\begin{aligned} R(\hat{\theta}) &= \mathbb{E}\left[(\hat{\theta}-\theta)^2\right] = \mathbb{E}\left[((\hat{\theta}-\mathbb{E}\left[\hat{\theta}\right])+(\mathbb{E}\left[\hat{\theta}\right]-\theta))^2\right] \\ &= \mathbb{E}\left[(\hat{\theta}-\mathbb{E}\left[\hat{\theta}\right])^2\right] + \mathbb{E}\left[(\mathbb{E}\left[\hat{\theta}\right]-\theta)^2\right] \\ &+ 2\mathbb{E}\left[(\hat{\theta}-\mathbb{E}\left[\hat{\theta}\right])(\mathbb{E}\left[\hat{\theta}\right]-\theta)\right] \\ &= \text{Var}\left(\hat{\theta}\right) + \text{Sesgo}(\hat{\theta})^2 \end{aligned}$$

Esa descomposición se llama *descomposición en varianza y sesgo*. La interpretación es que la cualidad de un estimador $\hat{\theta}$ depende de un equilibrio sutil entre su sesgo y su varianza. En consecuencia, si solo nos interesan los estimadores sin sesgo, la tarea es de encontrar él que tiene la varianza mínima.

1.4.2. Estimadores

Información de Fisher

Nos fijamos en el contexto particular donde

1. El espacio Θ es un abierto de $\mathbb{R}$.
2. Para cada $\theta \in \Theta$, P_θ tiene densidad f_θ.
3. La función $\theta \mapsto f_\theta(x)$ es derivable c.s.
4. Para cada función h, hay inversión de la derivación y la integral:

$$\frac{\partial}{\partial\theta}\int h(x)f_\theta(x)dx = \int h(x)\frac{\partial}{\partial\theta}f_\theta(x)dx$$

Bajo esas hipótesis, la cantidad

$$\frac{\partial}{\partial\theta}(\log f_\theta(X)) = \frac{\frac{\partial}{\partial\theta}(f_\theta(X))}{f_\theta(X)}$$

existe c.s., donde la variable X es de distribución P_θ.
No hay problema de división por 0 porque bajo $X \sim P_\theta$, tenemos $f_\theta(X) > 0$ c.s.

1.4.2. Estimadores

La *información de Fisher* es la función $I : \Theta \to [0, +\infty]$ donde

$$I(\theta) = \mathbb{E}_{P_\theta}\left[\left(\frac{\partial}{\partial\theta}(\log f_\theta(X))\right)^2\right]$$

Es una función que nos permitirá medir la factibilidad de bien estimar el parámetro θ.

Proposición

Si el soporte de f_θ no depende de θ, la información de Fisher se puede escribir

$$I(\theta) = Var_{P_\theta}\left(\frac{\partial}{\partial\theta}(\log f_\theta(X))\right)$$

1.4.2. Estimadores

Prueba.

La única cosa que probar es que $\mathbb{E}_{P_\theta}\left[\frac{\partial}{\partial\theta}(\log f_\theta(X))\right] = 0$.

Denotamos $S = \text{Supp} f_\theta$ el soporte de f_θ.

$$\begin{aligned}
\mathbb{E}_{P_\theta}\left[\frac{\partial}{\partial\theta}(\log f_\theta(X))\right] &= \int \frac{\frac{\partial}{\partial\theta}(f_\theta(x))}{f_\theta(x)} f_\theta(x)dx \\
&= \int \frac{\partial}{\partial\theta}(f_\theta(x))\mathbb{1}_{f_\theta(x)>0}dx \\
&= \int_S \frac{\partial}{\partial\theta}(f_\theta(x))dx \\
&= \frac{\partial}{\partial\theta}\int_S f_\theta(x)dx = \frac{\partial}{\partial\theta}1 = 0
\end{aligned}$$

donde se usó la propiedad 4. en la ultima linea. □

1.4.2. Estimadores

Ejemplo: Sea X una variable $\mathcal{P}(\theta)$ con $\theta > 0$. Entonces,

$$\begin{aligned} I(\theta) &= \mathbb{E}_{P_\theta}\left[\left(\frac{\partial}{\partial\theta}(\log f_\theta(X))\right)^2\right] \\ &= \mathbb{E}_{P_\theta}\left[\left(\frac{\partial}{\partial\theta}(-\theta + X\log\theta - \log X!)\right)^2\right] \\ &= \mathbb{E}_{P_\theta}\left[\left(\frac{X}{\theta} - 1\right)^2\right] = \frac{1}{\theta^2}\mathrm{Var}\,(X) = \frac{1}{\theta}. \end{aligned}$$

donde usamos que $\mathrm{Var}\,(X) = \theta$.

1.4.2. Estimadores

Si uno tiene acceso a una n-muestra $X_1, \ldots, X_n$ tq $\forall i, X_i \sim P_\theta$ y donde el soporte de P_θ no depende de θ, podemos calcular la información de Fisher de la muestra $X = (X_1, \ldots, X_n)$. En efecto, $f_X = f_\theta^{\otimes n}$ y

$$\begin{aligned} I_X(\theta) &= \mathrm{Var}_{P_\theta}\left(\frac{\partial}{\partial\theta}\log f_\theta^{\otimes n}(X)\right) \\ &= \mathrm{Var}_{P_\theta}\left(\frac{\partial}{\partial\theta}\log\prod_{i=1}^{n} f_\theta(X_i)\right) \\ &= \mathrm{Var}_{P_\theta}\left(\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f_\theta(X_i)\right) \\ &= n\mathrm{Var}_{P_\theta}\left(\frac{\partial}{\partial\theta}\log f_\theta(X_1)\right) = nI_{X_1}(\theta). \end{aligned}$$

Así que la información de Fisher de una muestra (de soporte fijo) es la suma de la información de Fisher de cada variable de la muestra.

1.4.2. Estimadores

En nuestro ejemplo de las variables de Poisson, $I_X(\theta) = \frac{n}{\theta}$.
El uso principal de la información de Fisher es lo siguiente

Teorema (Cramer-Rao)

Sea $g : \mathbb{R} \to \mathbb{R}$ una función derivable. Sea $\hat{\theta}$ un estimador del parámetro θ. La cantidad $g(\hat{\theta})$ es un estimador de $g(\theta)$ que supongamos sin sesgo. Entonces, $\forall \theta \in \Theta$,

$$Var_{P_\theta}\left(g(\hat{\theta})\right) \geq \frac{g'(\theta)^2}{I_X(\theta)}$$

donde $I_X(\theta)$ es la información de Fisher sobre la muestra $X = (X_1, \ldots, X_n)$.

En el ejemplo anterior, tenemos por ejemplo que para cada estimador $\hat{\theta}$,

$$\mathrm{Var}_{P_\theta}\left(\hat{\theta}\right) \geq \frac{\theta}{n}.$$

1.4.2. Estimadores

Comentarios de la cota de Cramer-Rao

- La cota de Cramer-Rao es también una cota inferior del Riesgo del estimador.
- Los estimadores tal que $\text{Var}_{P_\theta}\left(g(\hat{\theta})\right) = \frac{g'(\theta)^2}{I_X(\theta)}$ para todo $\theta \in \Theta$ se llaman *UMVU (por uniform minimum variance unbiased)*

Prueba.

Supongamos que $\mathbb{E}\left[g(\hat{\theta})^2\right] < \infty$ porque en el caso contrario el resultado del teorema es trivial. En lo que sigue denotamos $g(\hat{\theta}) = h(x)$ para insistir sobre la dependencia del estimador al vector de muestra x. Dado que el estimador sea sin sesgo, tenemos $\mathbb{E}_{P_\theta}\left[g(\hat{\theta})\right] = g(\theta)$ y entonces

$$\begin{aligned} g'(\theta) &= \frac{\partial}{\partial\theta}\int_S h(x) f_\theta(x) dx \\ &= \int_S h(x)\left(\frac{\partial}{\partial\theta} f_\theta(x)\right) dx \end{aligned}$$

1.4.2. Estimadores

Prueba (fin).

Continuando el calculo,

$$
\begin{aligned}
g'(\theta) &= \int_S h(x) \left(\frac{\partial}{\partial \theta} \log f_\theta(x) \right) f_\theta(x) dx \\
&= \mathbb{E}_{P_\theta} \left[h(X) \frac{\partial}{\partial \theta} \log f_\theta(X) \right] \\
&= \mathbb{E}_{P_\theta} \left[h(X) \frac{\partial}{\partial \theta} \log f_\theta(X) \right] - \mathbb{E}_{P_\theta} \left[h(X) \right] \mathbb{E}_{P_\theta} \left[\frac{\partial}{\partial \theta} \log f_\theta(X) \right] \\
&= \mathrm{Cov} \left(h(X), \frac{\partial}{\partial \theta} \log f_\theta(X) \right)
\end{aligned}
$$

Lo que permite ver que

$|g'(\theta)| \leq \sqrt{\mathrm{Var}\left(h(X)\right) \mathrm{Var}\left(\frac{\partial}{\partial \theta} \log f_\theta(X)\right)} = \sqrt{\mathrm{Var}\left(g(\hat{\theta})\right) I(\theta)}$ lo que termina la prueba. □

1.4.2. Estimadores

Podemos directamente dar el corolario siguiente

Corolario

Bajo las mismas hipótesis y si $X = (X_1, \ldots, X_n)$ *es una muestra tal que la información de Fisher de* X_1 *es* $I_{X_1}(\theta)$*, tenemos*

$$Var_{P_\theta}\left(g(\hat{\theta})\right) \geq \frac{g'(\theta)^2}{nI_{X_1}(\theta)}$$

Comentario: La varianza de los estimadores óptimos decrece en $1/n$.

Ejemplo: Sean $X_1, \ldots, X_n \sim \mathcal{P}(\theta)$ y sea $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ un estimador de θ. Calculamos la varianza

$$\mathrm{Var}_{P_\theta}\left(\overline{X}_n\right) = \frac{\theta}{n} = \frac{1}{nI_{X_1}(\theta)}$$

$\Rightarrow \overline{X}_n$ es un estimador óptimo!

1.Preliminares

1.5.Convergencias estocásticas

1.5.1. Definiciones

Convergencia en Probabilidad

Sea $(X_n)_{n\in\mathbb{N}}$ una sucesión de variables aleatorias de $\mathbb{R}^k$. Sea X una variable aleatoria de $\mathbb{R}^k$. Digamos que X_n *converge en probabilidad* a X si y solo si

$$\forall \epsilon > 0, \qquad \mathbb{P}\left(\|X_n - X\| > \epsilon\right) \underset{n\to+\infty}{\longrightarrow} 0.$$

Se nota $X_n \xrightarrow{\mathbb{P}} X$.

Convergencia en Distribución

Sea $(X_n)_{n\in\mathbb{N}}$ una sucesión de variables aleatorias de $\mathbb{R}^k$. Sea X una variable aleatoria de $\mathbb{R}^k$. Digamos que X_n *converge en distribución* a X si y solo si para cada función $h : \mathbb{R}^k \to \mathbb{R}$ continua y acotada

$$\mathbb{E}\left[h(X_n)\right] \underset{n\to+\infty}{\longrightarrow} \mathbb{E}\left[h(X)\right].$$

Se nota $X_n \xrightarrow{(d)} X$.

1.5.1. Definiciones

Ya vimos que la distribución se puede caracterizar de muchas maneras diferentes. La convergencia en distribución sigue el mismo esquema.

Proposición

La convergencia en distribución se puede probar con cada herramienta de caracterización de distribución. Mas concretamente, los siguientes son equivalentes

1. $X_n \xrightarrow{(d)} X$
2. $\forall t,\ \phi_{X_n}(t) \longrightarrow \phi_X(t)$
3. $\forall t,\ F_{X_n}(t) \longrightarrow F_X(t)$
4. $\forall t,\ M_{X_n}(t) \longrightarrow M_X(t)$
5. $\forall h$ *lipschitz y acotada*, $\mathbb{E}\left[h(X_n)\right] \longrightarrow \mathbb{E}\left[h(X)\right]$.

1.5.1. Definiciones

Convergencia casi seguramente

Sea $(X_n)_{n\in\mathbb{N}}$ una sucesión de variables aleatorias de $\mathbb{R}^k$. Sea X una variable aleatoria de $\mathbb{R}^k$. Digamos que X_n *converge casi seguramente* a X si y solo si el conjunto $\mathcal{C} = \{\omega \in \Omega : X_n(\omega) \to X(\omega)\}$ es de probabilidad $P(\mathcal{C}) = 1$.

Se nota $X_n \xrightarrow{(c.s.)} X$.

- La convergencia casi segura es el pendiente de la convergencia punto por punto en analisis funcional.
- La convergencia en distribución es la mas usada en estadística.
- La convergencia en distribución es la mas usada en probabilidad.

1.5.2. Implicaciones entre convergencias

Tenemos las implicaciones:

Proposición

Se tiene que c.s. $\Rightarrow \mathbb{P} \Rightarrow (d)$

Prueba.

Por definición, $\mathbb{P}(\mathcal{C}) = 1 \Rightarrow \mathbb{P}(\mathcal{C}^c) = 0$, donde

$$\mathcal{C} = \{\omega : \exists \epsilon > 0, \exists n_k \text{ sucesión } \text{ t.q. } \forall k, \|X_{n_k} - X\| > \epsilon\}$$

En particular, para cualquier $\epsilon > 0$ fijo,

$$\begin{aligned} 0 &= \mathbb{P}(\exists n_k \text{ sucesión } \text{ t.q. } \forall k, \|X_{n_k} - X\| > \epsilon) \\ &= \mathbb{P}\left(\limsup_n \{\|X_n - X\| > \epsilon\}\right) \\ &\geq \limsup_n \mathbb{P}(\|X_n - X\| > \epsilon) \end{aligned}$$

donde la ultima linea se cumple por el Lema de Fatou. □

1.5.2. Implicaciones entre convergencias

Prueba.

Entonces,
$0 \geq \limsup_n \mathbb{P}(\|X_n - X\| > \epsilon) \geq \liminf_n \mathbb{P}(\|X_n - X\| > \epsilon) \geq 0$ lo que implica la primera parte del resultado. Sea f una función lipschitz y acotada $|f| \leq K$ y

$$\forall x, y \in \mathbb{R}^k, |f(x) - f(y)| \leq \lambda \|x - y\|,$$

entonces,

$$\begin{aligned} |\mathbb{E}[f(X_n)] - \mathbb{E}[f(X)]| &\leq \mathbb{E}\left[|f(X_n) - f(X)| \mathbb{1}_{\|X_n - X\| \geq \epsilon}\right] \\ &\quad + \mathbb{E}\left[|f(X_n) - f(X)| \mathbb{1}_{\|X_n - X\| < \epsilon}\right] \\ &\leq 2K\mathbb{P}(\|X_n - X\| \geq \epsilon) + \lambda\epsilon \end{aligned}$$

Tomando ϵ suficientemente pequeño y luego n suficientemente grande, tenemos la segunda parte del resultado. □

1.5.2. Implicaciones entre convergencias

Un resultado útil

El siguiente resultado sera utilizado en el Lema de Slutsky.

Proposición

Sean $(X_n)_n$ *y* $(Y_n)_n$ *dos sucesiones de vectores aleatorios de* $\mathbb{R}^k$ *y sea* $X \in \mathbb{R}^k$ *un vector aleatorio. Supongamos que* $X_n \xrightarrow{(d)} X$ *y* $\|X_n - Y_n\| \xrightarrow{\mathbb{P}} 0$, *entonces* $Y_n \xrightarrow{(d)} X$.

Prueba.

Sea f λ-lipschitz y acotada. Haciendo las mismas cuentas

$$|\mathbb{E}[f(X_n)] - \mathbb{E}[f(Y_n)]| \leq 2K\mathbb{P}(\|X_n - Y_n\| \geq \epsilon) + \lambda\epsilon$$

lo que implica que $\mathbb{E}[f(X_n)] - \mathbb{E}[f(Y_n)] \underset{n\to+\infty}{\longrightarrow} 0$, y como $\mathbb{E}[f(X_n)] \underset{n\to+\infty}{\longrightarrow} \mathbb{E}[f(X)]$, llegamos al resultado. □

1.5.2. Implicaciones entre convergencias

Estabilidad de la cv estocástica, Lema de Slutsky

Es un hecho conocido que para la convergencia de sucesiones la noción de limite es estable al respecto de la suma. Mas concretamente, si para dos sucesiones $(x_n)_n$ y $(y_n)_n$, tenemos que $x_n \to x$ y al mismo tiempo $y_n \to y$, entonces $x_n + y_n \to x + y$. Que decir de $X_n + Y_n \longrightarrow$ cuando las dos sucesiones convergen estocásticamente?

Proposición

- *Si* $X_n \xrightarrow{c.s.} X$ *y* $Y_n \xrightarrow{c.s.} Y$, *entonces* $X_n + Y_n \xrightarrow{c.s.} X + Y$.
- *Si* $X_n \xrightarrow{\mathbb{P}} X$ *y* $Y_n \xrightarrow{\mathbb{P}} Y$, *entonces* $X_n + Y_n \xrightarrow{\mathbb{P}} X + Y$.
- *En general, no es verdad que* $X_n \xrightarrow{(d)} X$ *y* $Y_n \xrightarrow{(d)} Y \Rightarrow$ $X_n + Y_n \xrightarrow{(d)} X + Y$

1.5.2. Implicaciones entre convergencias

Prueba.

- Para la convergencia c.s., solo se tiene que ver que la intersección de dos eventos de probabilidad 1 tiene probabilidad 1.
- Es cierto que

$$\mathbb{P}\left(\|X_n + Y_n - X - Y\| \geq \epsilon\right) \leq \mathbb{P}\left(\|X_n - X\| \geq \epsilon/2\right) + \mathbb{P}\left(\|Y_n - Y\| \geq \epsilon/2\right).$$

  lo que prueba la propiedad.
- En general, la convergencia en distribución no define sin ambigüedad la variable limite completamente (solo su distribución).

□

1.5.2. Implicaciones entre convergencias

El lema de Slutsky da un caso positivo cuando el limite de una de las dos sucesión es una variable constante.

Lema

Si $X_n \xrightarrow{(d)} X$ *y* $Y_n \xrightarrow{(d)} c$ *donde* $c \in \mathbb{R}^k$ *es una vector constante, entonces*

$$X_n + Y_n \xrightarrow{(d)} X + c$$

Prueba.

Obviamente tenemos $X_n + c \xrightarrow{(d)} X + c$, porque podemos ver las funciones lipschitz y acotadas de $X_n + c$ como funciones lipschitz y acotadas de X_n. Ademas, $\|X_n + Y_n - (X_n + c)\| \xrightarrow{\mathbb{P}} 0$, entonces por el "resultado útil", tenemos $X_n + Y_n \xrightarrow{(d)} X + c$. □

1.5.3. Caracterización de convergencias en distribución

Exactamente como en el caso de caracterización de la distribución de variables/vectores aleatorios/aleatorios, podemos usar las cantidades que caracterizan la distribución para probar convergencia en distribución.

Teorema (Levy)

Sea $(X_n)_n$ *una sucesión de vectores aleatorios en* $\mathbb{R}^k$ *y denotamos* ϕ_{X_n} *la función característica de cada* X_n*. Supongamos que*

- $\forall t \in \mathbb{R}^k$, $g(t) := \lim_n \phi_{X_n}(t)$ *existe.*
- *La función* g *es continua en 0.*

Entonces, $\exists$ *una variable aleatoria* $X \in \mathbb{R}^k$ *tq* $g(t) = \phi_X(t)$ *y* $X_n \xrightarrow{(d)} X$.

Prueba.

Eso sale del alcance de este curso. Se usa la noción de tensión de variables aleatorias. □

1.5.3. Caracterización de convergencias en distribución

Caracterización de Cramer-Wold

Una consecuencia cómoda es el

Corolario (Cramer-Wold)

Sea $(X_n)_n$ una sucesión de vectores aleatorios de $\mathbb{R}^k$. Entonces

$$X_n \xrightarrow{(d)} X \Leftrightarrow \forall t \in \mathbb{R}^k, t^T X_n \xrightarrow{(d)} t^T X$$

La convergencia a la derecha es entre variables aleatorias reales!

Prueba.

$$\begin{aligned} X_n \xrightarrow{(d)} X &\Leftrightarrow \forall u \in \mathbb{R}^k, \phi_{X_n}(u) \to \phi_X(u) \\ &\Leftrightarrow \forall u \in \mathbb{R}^k, \mathbb{E}\left[e^{u^T X_n}\right] \to \mathbb{E}\left[e^{u^T X}\right] \\ &\Leftrightarrow \forall \lambda \in \mathbb{R}, \forall t \in \mathbb{R}^k, \mathbb{E}\left[e^{\lambda t^T X_n}\right] \to \mathbb{E}\left[e^{\lambda t^T X}\right] \\ &\Leftrightarrow \forall t \in \mathbb{R}^k, t^T X_n \xrightarrow{(d)} t^T X \end{aligned}$$

1.5.3. Caracterización de convergencias en distribución

Ejemplo: Avanzando nos un poco sobre la noción de vector gaussiano, si $Z \sim \mathcal{N}(\nu, \Sigma)$ entonces $\phi_Z(t) = e^{it^T\mu - \frac{1}{2}t^T\Sigma t}$.
En efecto, $t^T Z$ es una variables gaussiana (en $\mathbb{R}$) para cada $t \in \mathbb{R}^k$ y

$$\mathbb{E}\left[t^T Z\right] = t^T \mathbb{E}\left[Z\right] = t^T \mu$$

De la misma manera, denotando $Y = Z - \mu$

$$\begin{aligned}
\text{Var}\left(t^T Z\right) &= \mathbb{E}\left[(t^T Y)^2\right] = \mathbb{E}\left[(t^T Y)(t^T Y)^T\right] \\
&= \mathbb{E}\left[t^T Y Y^T t\right] \\
&= t^T \mathbb{E}\left[YY^T\right] t = t^T \Sigma t.
\end{aligned}$$

Como sabemos que para una variable gaussiana $\mathcal{N}(\mu, \sigma^2)$ real, $\phi_N(t) = e^{it\mu - \frac{1}{2}t^2\sigma^2}$ y que $\phi_Z(t) = \phi_{t^T Z}(1)$, tenemos la respuesta.

1.Preliminares

1.6.Ley de grandes números, Teorema del limite central, método delta

1.6.1. Ley de grandes números

Muchas veces en este curso encontraremos cantidades de la forma $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ donde las variables aleatorias X_i serán independientes. Los teoremas siguientes tratan del comportamiento asintótico de $\overline{X}_n$.

Forma débil

Teorema

Sean $X_1, \ldots, X_n$ vectores aleatorios i.i.d. de $\mathbb{R}^k$. Supongamos que $\mu = \mathbb{E}[X_1]$ existe, entonces $\overline{X}_n \xrightarrow{\mathbb{P}} \mu$.

Prueba.

Por la existencia de $\mathbb{E}[X_1]$, $\phi'_{X_1}(0) = i\mu$. Considerando el desarrollo limite $\phi_{X_1}(t) = 1 + it\mu + o(t)$, tenemos

$$\phi_{\overline{X}_n}(t) = \phi_{X_1}(t/n)^n \to e^{it\mu}$$

que es la función característica de una variable constante igual a μ. Entonces $\overline{X}_n \xrightarrow{(d)} \mu$ lo que es equivalente a $\overline{X}_n \xrightarrow{\mathbb{P}} \mu$. □

1.6.1. Ley de grandes números

Forma fuerte

De hecho, se puede probar una forma fuerte de la ley de grandes números:

Teorema

Sean $X_1, \ldots, X_n$ vectores aleatorios i.i.d. de $\mathbb{R}^k$. Supongamos que $\mu = \mathbb{E}[X_1]$ existe, entonces $\overline{X}_n \xrightarrow{c.s.} \mu$.

Prueba.

Admitido en este curso. Para eso se necesitan unos conceptos mas como la ley del 0-1 de Kolmogorov por ejemplo. □

Comentario: Si un parámetro desconocido se escribe como el valor esperado de una variable aleatoria, el estimador $\overline{X}_n$ será un estimador consistente del parámetro.

1.6.2. Teorema del limite central

El siguiente teorema se enfoca a describir el comportamiento de $\bar{X}_n$ alrededor de $\mathbb{E}[X_1]$.

Teorema (Central Limite (en $\mathbb{R}$))

Sean $X_1, \ldots, X_n$ variables aleatorias reales i.i.d. con $\mathbb{E}[X_1] = 0$ y $Var(X_1) = \sigma^2$. Entonces,

$$\sqrt{n}\bar{X}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \xrightarrow{(d)} \mathcal{N}(0, \sigma^2)$$

Eso es el resultado fundamental de Probabilidad. Eso explica en esencia la importancia de la distribución gaussiana y de su estudio intenso en la literatura.

1.6.2. Teorema del limite central

Prueba.

Sea $\phi = \phi_{X_1}$. Tenemos $\phi'(0) = i\mathbb{E}[X_1] = 0$ y $\phi''(0) = i^2\mathbb{E}\left[X_1^2\right] = -\sigma^2$.

Entonces,

$$\mathbb{E}\left[e^{it\sqrt{n}\bar{X}_n}\right] = \phi\left(\frac{t}{\sqrt{n}}\right)^n = \left(1 - \frac{t^2\sigma^2}{2n} + o(\frac{1}{n})\right)^n \underset{n\to+\infty}{\longrightarrow} e^{-\frac{t^2\sigma^2}{2}}.$$

Pero $\phi_{\mathcal{N}(0,\sigma^2)}(t) = e^{-\frac{t^2\sigma^2}{2}}$ es continua en 0.

Entonces, por el Teorema de Levy, $\sqrt{n}\bar{X}_n \xrightarrow{(d)} \mathcal{N}(0,\sigma^2)$. □

1.6.2. Teorema del limite central

Para pasar de la convergencia de variables a la convergencia de vectores usamos el Lema de Cramer-Wold.

Teorema

Sean $X_1, \ldots, X_n$ *vectores aleatorios de* $\mathbb{R}^k$ *i.i.d. tal que* $\mu = \mathbb{E}[X_1]$ *y* $\Sigma = \mathbb{E}\left[(X_1 - \mu)(X_1 - \mu)^T\right]$, *entonces*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \xrightarrow{(d)} \mathcal{N}_k(\mu, \Sigma)$$

donde $\mathcal{N}_k(\mu, \Sigma)$ *denota el vector gaussiano de valor esperado* μ *y de matriz de covarianza* Σ.

1.6.2. Teorema del limite central

Prueba.

Sea $t \in \mathbb{R}^k$ y sean $Y_i = t^T X_i - t^T \mu$. Las variables Y_i son reales y tal que $\mathbb{E}[Y_1] = 0$ y $\text{Var}(Y_1) = t^T \Sigma t$.

Entonces,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i \xrightarrow{(d)} \mathcal{N}(0, t^T \Sigma t).$$

Pero si $Z \sim \mathcal{N}_k(0, \Sigma)$, $t^T Z \sim \mathcal{N}(0, t^T \Sigma t)$. Lo que significa que

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} t^T (X_i - \mu) \xrightarrow{(d)} t^T Z.$$

Eso siendo cierto para cada $t \in \mathbb{R}^k$, el Lema de Cramer-Wold nos dice que $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \xrightarrow{(d)} Z$ en los vectores. □

1.6.2. Teorema del limite central

Teorema de Lindeberg-Feller

Existen muchas versiones del Teorema del Limite Central. Aquí introducimos una versión que permite introducir flexibilidad en el supuesto i.i.d.

Teorema (Lindeberg-Feller)

Para cada n, sean $X_{n,1}, \ldots, X_{n,k_n}$ vectores aleatorios independientes tal que

- *Para cada $\epsilon > 0$, $R_n = \sum_{i=1}^{k_n} \mathbb{E}\left[\|X_{n,i}\|^2 \mathbb{1}_{\|X_{n,i}\|>\epsilon}\right] \underset{n\to\infty}{\longrightarrow} 0$*
- *Se cumple $\sum_{i=1}^{k_n} Cov(X_{n,i}) \underset{n\to\infty}{\longrightarrow} \Sigma$*

Entonces,

$$\sum_{i=1}^{k_n} X_{n,i} - \mathbb{E}[X_{n,i}] \xrightarrow{(d)} \mathcal{N}(0, \Sigma)$$

Prueba.

Eso usa de nuevo el Teorema de Levy. □

1.6.3. Método Delta

Tenemos los bloques fundamentales

$$\bar{X}_n \longrightarrow c \quad \text{y} \quad \sqrt{n}(\bar{X}_n - c) \longrightarrow \mathcal{N}(0, \sigma^2)$$

Que decir de una transformación de $\bar{X}_n$ en $\ell(\bar{X}_n)$? Cuando converge a $\ell(c)$? Que decir de la renormalización $\sqrt{n}(\ell(\bar{X}_n) - \ell(c))$?

Teorema (Método Delta)

Sea una sucesión $(X_n)_{n\in\mathbb{N}}$ de vectores aleatorios de $\mathbb{R}^k$. Sea $(a_n)_{n\in\mathbb{N}}$ una sucesión determinista de reales y $\ell : \mathbb{R}^k \to \mathbb{R}^p$. Supongamos que

- $a_n \to +\infty$.
- $\exists c \in \mathbb{R}^k$ *un vector determinista y* Y *un vector aleatorio tal que* $a_n(X_n - c) \xrightarrow{(d)} Y$.
- ℓ *es derivable en* c *y* $D\ell(c) \in \mathcal{M}_{p,k}(\mathbb{R})$.

Entonces, $a_n(\ell(X_n) - \ell(c)) \xrightarrow{(d)} D\ell(c) \times Y$

1.6.3. Método Delta

Prueba.

Por el Lema de Slutsky, es suficiente de probar que

$$W_n := a_n(\ell(X_n) - \ell(c)) - D\ell(c) \times a_n(X_n - c) \xrightarrow{\mathbb{P}} 0.$$

En efecto, si denotamos $V_n = D\ell(c) \times a_n(X_n - c)$, sabemos que $V_n \xrightarrow{(d)} Y$ y que $W_n \xrightarrow{\mathbb{P}} 0$ lo que implica que $W_n + V_n \xrightarrow{(d)} D\ell(c) \times Y$.
Sea $\underline{\epsilon > 0}$ y $\delta > 0$.

$$\begin{aligned}\mathbb{P}(\|W_n\| > \epsilon) &\leq \mathbb{P}(\|W_n\| > \epsilon \text{ y } \|X_n - c\| \leq \delta) + \mathbb{P}(\|X_n - c\| > \delta) \\ &= \mathbb{P}(E) + \mathbb{P}(\|X_n - c\| > \delta)\end{aligned}$$

donde definimos $E = \{\|W_n\| > \epsilon \text{ y } \|X_n - c\| \leq \delta\}$.

□

1.6.3. Método Delta

Prueba.

Pero por definición de $D\ell(c)$, $\forall T > 0,\ \exists \delta_0 > 0$

$$\|X - c\| \leq \delta_0 \Longrightarrow \|\ell(X) - \ell(c) - D\ell(c)(X - c)\| \leq T\|X - c\|$$

Eso siendo una propiedad de la función ℓ, tenemos de nuevo $\forall n$ y $\forall \delta < \delta_0$,

$$\|X_n - c\| \leq \delta \Longrightarrow \|W_n\| \leq a_n T\|X_n - c\|.$$

Entonces bajo el evento E, $\epsilon < \|W_n\| \leq a_n T\|X_n - c\|$ lo que implica que

$$\mathbb{P}(E) \leq \mathbb{P}(a_n\|X_n - c\| > \epsilon/T)$$

y

$$\begin{aligned}\mathbb{P}(\|W_n\| > \epsilon) &\leq \mathbb{P}(a_n\|X_n - c\| > \epsilon/T) + \mathbb{P}(a_n\|X_n - c\| > a_n\delta) \\ &\leq 2\mathbb{P}(a_n\|X_n - c\| > \min(\epsilon/T, a_n\delta))\end{aligned}$$

□

1.6.3. Método Delta

Prueba.

Como $a_n \to +\infty$, existe $N_0 \in \mathbb{N}$ tal que $\forall n \geq N_0$, tenemos $\min(\epsilon/T, a_n\delta) = \epsilon/T$, lo que muestra

$$\begin{aligned}
\limsup_{n\to\infty} \mathbb{P}\left(\|W_n\| > \epsilon\right) &\leq 2\limsup_{n\to\infty} \mathbb{P}\left(a_n\|X_n - c\| > \min(\epsilon/T, a_n\delta)\right) \\
&= 2\limsup_{n\to\infty} \mathbb{P}\left(a_n\|X_n - c\| > \epsilon/T\right) \\
&= 2\mathbb{P}\left(\|Y\| > \epsilon/T\right)
\end{aligned}$$

La probabilidad de la izquierda siendo independiente de T, lo podemos elegir arbitrariamente pequeño para tener $\mathbb{P}\left(\|Y\| > \epsilon/T\right) \longrightarrow 0$. Finalmente, mostramos que $W_n \xrightarrow{\mathbb{P}} 0$ lo que prueba el resultado. □

1.6.3. Método Delta

La consecuencia mayor de este resultado es lo siguiente.

Corolario

Sea $X_1, \ldots, X_n$ una muestra de vectores aleatorios de $\mathbb{R}^k$ tq $\mathbb{E}[X_1] = \mu \in \mathbb{R}^k$ y $Var(X_1) = \Sigma \in \mathcal{M}_{k\times k}(\mathbb{R})$. Sea $\ell : \mathbb{R}^k \to \mathbb{R}$ una función diferenciable en μ de diferencia $D\ell(\mu)$, entonces,

$$\sqrt{n}(\ell(\overline{X}_n) - \ell(\mu)) \xrightarrow{(d)} \mathcal{N}(0, D\ell(\mu)^T \Sigma D\ell(\mu)).$$

donde $D\ell(\mu)^T \Sigma D\ell(\mu)$ es un real positivo.

Prueba.

Es una consecuencia directa del teorema del método delta. □

1.6.3. Método Delta

Ejemplo: Sea $X_i \sim \mathcal{P}(\theta)$. Hemos visto que $\overline{X}_n \xrightarrow{\mathbb{P}} \theta$. Pero sabemos que la varianza de una variable de Poisson es también igual a θ lo que permite pensar en un otro estimador:

$$\hat{s}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Tomando la notación $Y_i = X_i - \theta$, tenemos fácilmente

$$\mathbb{E}[Y_i] = 0, \quad \mathbb{E}\left[Y_i^2\right] = \theta, \quad \mathbb{E}\left[Y_i^3\right] = \theta \quad \text{y} \quad \mathbb{E}\left[Y_i^4\right] = \theta(1+3\theta),$$

y $\hat{s}_n^2(X_1, \ldots, X_n) = \hat{s}_n^2(Y_1, \ldots, Y_n)$.
Se ve que $\hat{s}_n^2 = \hat{\mu}_{2,n} - \hat{\mu}_{1,n}^2$ donde

$$\hat{\mu}_{2,n} = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 \qquad \text{y} \qquad \hat{\mu}_{1,n} = \overline{Y}_n$$

1.6.3. Método Delta

Por el teorema central límite,

$$\sqrt{n}\left(\begin{pmatrix}\hat{\mu}_{1,n}\\ \hat{\mu}_{2,n}\end{pmatrix}-\begin{pmatrix}0\\ \theta\end{pmatrix}\right)\xrightarrow{(d)}\mathcal{N}(0,\Sigma)$$

donde

$$\Sigma=\begin{pmatrix}\theta & \theta\\ \theta & \theta(1+3\theta)\end{pmatrix}$$

Vemos que $\hat{s}_n^2=\ell(\hat{\mu}_{1,n},\hat{\mu}_{2,n})$ con $\ell(x,y)=y-x^2$. Calculando la diferencial, $D\ell(0,\theta)=(0,1)^T$ y aplicando el corolario,

$$\sqrt{n}(\hat{s}_n^2-\theta)\xrightarrow{(d)}\mathcal{N}(0,\theta+3\theta^2)$$

Pero tenemos,

$$\sqrt{n}(\overline{X}_n-\theta)\xrightarrow{(d)}\mathcal{N}(0,\theta)$$

y $\overline{X}_n$ es un mejor estimador de θ que $\hat{s}_n^2$.

2.Fundamentos de estimación

2.1.Principios de estimación

2.1.1. Método de momentos

Para estimar una cantidad que se escribe como $\mathbb{E}[h(X)]$, usamos la convergencia

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i) \longrightarrow \mathbb{E}[h(X)]$$

si uno tiene una muestra $X_1, \ldots, X_n$.
En muchos contextos, las cantidades se escriben como funciones de momentos. Es lo que se llama **método de momentos**. Si denotamos

$$\mu_1 = \mathbb{E}_\theta[X] \quad \hat{\mu}_{1,n} \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\mu_2 = \mathbb{E}_\theta\left[X^2\right] \quad \hat{\mu}_{2,n} \frac{1}{n}\sum_{i=1}^{n} X_i^2,$$

la varianza no es un momento pero se escribe gracias a los momentos de orden 1 y 2.

2.1.1. Método de momentos

Método de momentos

Un estimador que viene del *método de momentos* tiene la forma

$$\hat{g}(\theta) = f(\hat{\mu}_{1,n}, \hat{\mu}_{2,n}, \ldots, \hat{\mu}_{k,n})$$

Ejemplo: Sea $X_1, \ldots, X_n$ una muestra de distribución exponencial de parámetro λ desconocido : $f_{X_1} : x \mapsto \mathbb{1}_{x>0}\lambda e^{-\lambda x}$. Sabemos que $\mathbb{E}[X_1] = \frac{1}{\lambda}$, entonces un estimado de λ por el método de momentos está dado por

$$\hat{g}(\lambda) = \frac{1}{\hat{\mu}_{1,n}}.$$

Convergencia

Esos estimadores se estudian gracias a los teoremas limites del fin del capítulo anterior. Bajo las buenas condiciones los estimadores $\hat{\mu}_{i,n}$ convergen y tienen normalidad asintótica y el método delta permite concluir por la función f.

2.1.2. Estimador de cuantiles

Es muy importante de tener una manera de estimar los cuantiles de algunas distribuciones.

- Para las pruebas de hipótesis.
- Para los intervalos de confianza.

Recordamos que la función de distribución está dada por $F_X(t) = \mathbb{P}(X \leq t)$ y que el cuantile de orden $\beta \in [0, 1]$ está dado por $q_\beta = F_X^{-1}(\beta)$ donde

$$F_X^{-1}(\beta) = \inf\{x : F_X(x) \geq \beta\}.$$

Cuantile empírico

El *cuantile empírico* de orden β de una muestra $X_1, \ldots, X_n$ está dado por $q_{n,\beta} = F_n^{-1}(\beta)$ donde

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq t}.$$

2.1.2. Estimador de cuantiles

Estadísticas de orden

La definición de cuantiles empíricos tiene que ver con la noción de estadísticas de orden. Uno puede reordenar la muestra $X_1, \ldots, X_n$ en $X_{(1)}, \ldots, X_{(n)}$ de tal manera que $\forall i, j$,

$$i \leq j, \Longrightarrow X_{(i)} \leq X_{(j)}$$

Con esas notaciones

$$q_{n,\beta} = X_{(\lceil n\beta \rceil)}$$

donde $\lceil \rceil$ designa la parte entera superior.

Convergencia

Teorema

Sea $\beta \in (0,1)$ tq F_X es creciente en la vecindad de q_β. Entonces,

$$q_{n,\beta} \xrightarrow{\mathbb{P}} q_\beta$$

2.1.2. Estimador de cuantiles

Prueba.

Por la crecimiento de F_X, $\forall \epsilon > 0$,

$$F_X(q_\beta + \epsilon) - \beta > 0 \qquad \text{y} \qquad \beta - F_X(q_\beta - \epsilon) > 0.$$

Entonces, para cada $\epsilon > 0$,

$$\begin{aligned}
\mathbb{P}\left(q_{n,\beta} > q_\beta + \epsilon\right) &= \mathbb{P}\left(\beta > F_n(q_\beta + \epsilon)\right) \\
&= \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \leq q_\beta + \epsilon} < \beta\right) \\
&\leq \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \leq q_\beta + \epsilon} - \mathbb{P}\left(X_1 \leq\right) q_\beta + \epsilon\right) < 0
\end{aligned}$$

Pero la LGN afirma que $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \leq q_\beta + \epsilon} \xrightarrow{\mathbb{P}} \mathbb{P}\left(X_1 \leq q_\beta + \epsilon\right)$ lo que implica $\mathbb{P}\left(q_{n,\beta} > q_\beta + \epsilon\right) \to 0$ y por simetría $q_{n,\beta} \xrightarrow{\mathbb{P}} q_\beta$. □

2.1.3. Máximo de verosimilitud

Vamos a hacer algunas hipótesis de trabajo:

- Tenemos acceso a una muestra $X_1, \ldots, X_n$ de una ley de probabilidad desconocida.
- La distribución de X tiene una densidad denotada f.
- Supongamos que f pertenece a una clase

$$\mathcal{F} = \{f(\cdot|\theta) : \theta \in \Theta\}$$

  Entonces $\exists \theta_0 \in \Theta$ tq $f(\cdot) = f(\cdot|\theta_0)$.
- Cuando el contexto lo exige, supondremos que $\forall \theta, \forall x, f(x|\theta) > 0$. *Eso siempre se reduce a una consideración de soporte.*

Ejemplos:

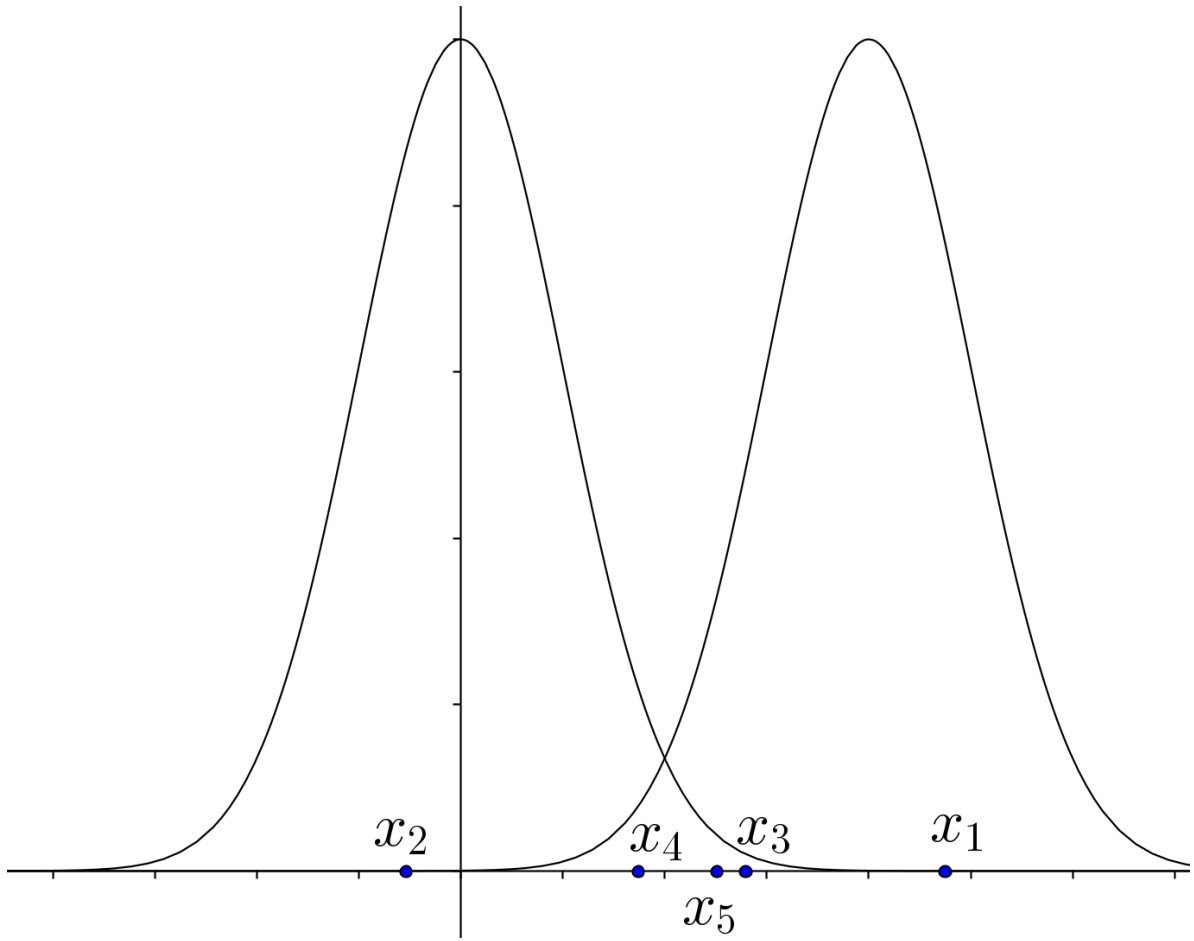
$$\mathcal{F} = \{x \mapsto \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma}} : (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+\}$$

$$\mathcal{F} = \{x \mapsto p^x(1-p)^{1-x} : p \in [0,1]\}$$

2.1.3. Máximo de verosimilitud

Problemática

Cómo se puede construir un estimador $\hat{\theta}$ del valor θ_0?



Donde las dos curvas corresponden a $\theta = 0$ y $\theta = 1$.

2.1.3. Máximo de verosimilitud

- Si solo tuvieramos x_1, que queremos contestar? $\hat{\theta} = 1$ porque

$$\mathbb{P}_0(X_1 = x_1) < \mathbb{P}_1(X_1 = x_1) \text{ o igualmente } f(x_1|0) < f(x_1|1)$$

- Si consideramos x_1 y x_2, sabemos que

$$\mathbb{P}_0((X_1, X_2) = (x_1, x_2)) = \mathbb{P}_0(X_1 = x_1)\,\mathbb{P}_0(X_2 = x_2)$$

y entonces comparamos los valores

$$f(x_1|0)f(x_2|0) \text{ y } f(x_1|1)f(x_2|1).$$

- Para $x_1, \ldots, x_n$, comparamos entonces

$$f(x_1|0) \times \cdots \times f(x_n|0) \text{ y } f(x_1|1) \times \cdots \times f(x_n|1)$$

y guardamos lo mas grande $\Rightarrow \hat{\theta} = 1$ aqui.

2.1.3. Máximo de verosimilitud

Verosimilitud

Para cada $\theta \in \Theta$, la cantidad

$$\prod_{i=1}^{n} f(X_i|\theta)$$

se llama *verosimilitud* del parámetro θ. El logaritmo de ese valor se llama *log-verosimilitud* de θ.

Máximo de verosimilitud

El valor definido por

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}} \prod_{i=1}^{n} f(X_i|\theta)$$

define un estimador de θ que se llama *estimador de máximo de verosimilitud*.

2.1.3. Máximo de verosimilitud

En general se prefiere la formula equivalente siguiente

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log f(X_i|\theta)$$

cuando el soporte de cada $f(\cdot|\theta)$ es $\mathbb{R}$ entero.
Ejemplo: Si tomamos la clase de funciones

$$\mathcal{F} := \left\{ x \mapsto \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}} \ : \ \mu \in \mathbb{R} \right\},$$

el estimador de máximo de verosimilitud está dado por

$$\hat{\mu} = \underset{\mu \in \mathbb{R}}{\operatorname{argmax}} \sum_{i=1}^{n} \log f(X_i|\mu)$$

2.1.3. Máximo de verosimilitud

Pero $\sum_{i=1}^{n} \log f(X_i|\mu) = \sum_{i=1}^{n} -\frac{(X_i-\mu)^2}{2} + n \log \frac{1}{\sqrt{2\pi}}$. Entonces es suficiente de minimizar

$$g(\mu) := \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{2}.$$

Pero

$$g'(\mu) = 2 \sum_{i=1}^{n} (X_i - \mu)$$

y $g'(\hat{\mu}) = 0 \Rightarrow \hat{\mu} = \overline{X}_n$. En este caso, el estimador de máximo de verosimilitud coincide con la media empírica.

No será siempre el caso. En general $\hat{\theta} \neq \overline{X}_n$

2.1.3. Máximo de verosimilitud

Por la LGN, sabemos que

$$\frac{1}{n}\sum_{i=1}^{n}\log f(X_i|\theta) \xrightarrow{\mathbb{P}} \mathbb{E}_{\theta_0}\left[f(X|\theta)\right]$$

para cada $\theta \in \Theta$. En lo que sigue, denotaremos

- $L_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\log f(X_i|\theta)$
- $L(\theta) = \mathbb{E}_{\theta_0}\left[\log f(X|\theta)\right]$

Lema

Para cada $\theta \in \Theta$,

$$L(\theta) \leq L(\theta_0)$$

Ademas, la desigualdad es estricta salvo para los θ tal que

$$\mathbb{P}_{\theta_0}\left(f(X|\theta) = f(X|\theta_0)\right) = 1.$$

2.1.3. Máximo de verosimilitud

Prueba.

$$L(\theta) - L(\theta_0) = \mathbb{E}_{\theta_0}\left[\log f(X|\theta) - \log f(X|\theta_0)\right]$$
$$= \mathbb{E}_{\theta_0}\left[\frac{\log f(X|\theta)}{\log f(X|\theta_0)}\right]$$

Usando que $\log t \leq t - 1$,

$$\mathbb{E}_{\theta_0}\left[\frac{\log f(X|\theta)}{\log f(X|\theta_0)}\right] \leq \mathbb{E}_{\theta_0}\left[\frac{f(X|\theta)}{f(X|\theta_0)} - 1\right]$$
$$= \int \left(\frac{f(x|\theta)}{f(x|\theta_0)} - 1\right) f(x|\theta_0)dx$$
$$= \int f(x|\theta)dx - \int f(x|\theta_0)dx = 1 - 1 = 0.$$

La desigualdad es estricta si y solo si la desigualdad $\log t \leq t - 1$ es estricta si y solo si $t \neq 1$ lo que ocurre si $\exists x$ tq $f(x|\theta) \neq f(x|\theta_0)$. □

2.1.3. Máximo de verosimilitud

Comentario: Como consecuencia directa,

$$\theta_0 = \underset{\theta \in \Theta}{\operatorname{argmax}}\, L(\theta).$$

En resumen:

- El máx de L_n es en $\hat{\theta}$.
- El máx de L es en θ_0.
- El máx de L puede no ser único pero eso implica que el modelo no es identifiable.

Teorema (informal)

Bajo condiciones de regularidad sobre la familia $\mathcal{F}$, $\hat{\theta}$ es un estimador consistente. Es decir

$$\hat{\theta} \xrightarrow{\mathbb{P}} \theta_0.$$

2.1.3. Máximo de verosimilitud

Teorema (formal)

Supongamos que

- $\sup_{\theta \in \Theta} |L_n(\theta) - L(\theta)| \xrightarrow{\mathbb{P}} 0$ *(regularidad)*
- $\forall \epsilon > 0$, $\sup_{\theta : |\theta - \theta_0| \geq \epsilon} L(\theta) < L(\theta_0)$. *(idenfiabilidad)*

Entonces,

$$\hat{\theta} \xrightarrow{\mathbb{P}} \theta_0.$$

Prueba.

$$\begin{aligned} 0 \leq L(\theta_0) - L(\hat{\theta}) &= L(\theta_0) - L_n(\theta_0) + L_n(\theta_0) - L(\hat{\theta}) \\ &\leq L(\theta_0) - L_n(\theta_0) + L_n(\hat{\theta}) - L(\hat{\theta}) \\ &\leq \underset{\downarrow \mathbb{P}}{L(\theta_0) - L_n(\theta_0)} \quad + \quad \underset{\downarrow \mathbb{P}}{\sup_{\theta \in \Theta} L_n(\theta) - L(\theta)} \\ &\qquad\qquad 0 \qquad\qquad\qquad\qquad 0 \end{aligned}$$

□

2.1.3. Máximo de verosimilitud

Por el segundo supuesto, $\forall \epsilon > 0$, $\exists \eta > 0$ tq $\forall \theta$ tq $|\theta - \theta_0| \geq \epsilon$,

$$L(\theta) < L(\theta_0) - \eta.$$

Lo que implica que

$$\left\{ |\hat{\theta} - \theta_0| \geq \epsilon \right\} \subset \left\{ L(\hat{\theta}) < L(\theta_0) - \eta \right\},$$

y entonces,

$$\mathbb{P}_{\theta_0} \left(|\hat{\theta} - \theta_0| \geq \epsilon \right) \leq \mathbb{P}_{\theta_0} \left(L(\hat{\theta}) < L(\theta_0) - \eta \right) \underset{n \to \infty}{\longrightarrow} 0.$$

Finalmente, eso demuestra que $\hat{\theta} \xrightarrow{\mathbb{P}} \theta_0$.

2.1.3. Máximo de verosimilitud

En contextos reales, muy seguido, se puede calcular $\hat{\theta}$ explícitamente y uno puede probar la convergencia directamente (ad hoc).

Normalidad asintótica

Queremos probar algo del tipo TCL,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{(d)} \mathcal{N}(0, \sigma^2_{MV})$$

Denotamos $\ell(X|\theta) = \log f(X|\theta)$. Recordamos que la información de Fisher está dada por

$$I(\theta) = \mathbb{E}_{\theta_0}\left[\ell'(X|\theta)^2\right].$$

Lema

Tenemos $I(\theta_0) = -\mathbb{E}_{\theta_0}\left[\ell''(X|\theta_0)\right]$.

2.1.3. Máximo de verosimilitud

Prueba.

Derivando dos veces tenemos,

$$\ell'(X|\theta) = \frac{f'(X|\theta)}{f(X|\theta)} \qquad \text{y} \qquad \ell''(X|\theta) = \frac{f''(X|\theta)}{f(X|\theta)} - \frac{f'(X|\theta)^2}{f(X|\theta)^2}$$

donde las derivadas se toman al respecto de θ. Como $\int f(x|\theta)dx = 1$ tenemos, derivando en θ,

$$\int f'(x|\theta)dx = 0 \qquad \text{y} \qquad \int f''(x|\theta)dx = 0$$

Entonces,

$$\begin{aligned}\mathbb{E}_{\theta_0}\left[\ell''(X|\theta_0)\right] &= \int \left(\frac{f''(x|\theta_0)}{f(x|\theta_0)} - \frac{f'(x|\theta_0)^2}{f(x|\theta_0)^2}\right) f(x|\theta_0)dx \\ &= \int f''(x|\theta_0)dx - \mathbb{E}_{\theta_0}\left[\ell'(X|\theta_0)^2\right] = -I(\theta_0).\end{aligned}$$

□

2.1.3. Máximo de verosimilitud

Teorema

Supongamos que

- $\theta \mapsto \ell(X|\theta) \in \mathcal{C}^3(\Theta)$ *c.s.*
- $\hat{\theta} \xrightarrow{\mathbb{P}} \theta_0$

Entonces,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{(d)} \mathcal{N}\left(0, \frac{1}{I(\theta_0)}\right).$$

Prueba.

El punto $\hat{\theta}$ siendo un extremo (en θ) de L_n, tenemos $L_n'(\hat{\theta}) = 0$.
Por el teorema del valor intermedio $\exists \hat{\theta}_1 \in (\hat{\theta}, \theta_0)$ tal que

$$0 = L_n'(\hat{\theta}) = L_n'(\theta_0) + L_n''(\hat{\theta}_1)(\hat{\theta} - \theta_0)$$

Entonces,

$$\sqrt{n}(\hat{\theta} - \theta_0) = -\sqrt{n}\frac{L_n'(\theta_0)}{L_n''(\hat{\theta}_1)}$$

2.1.3. Máximo de verosimilitud

Prueba.

Como $L'(\theta_0) = 0$ (porque θ_0 es un extremo de L),

$$\begin{aligned}\sqrt{n}L'_n(\theta_0) &= \sqrt{n}(L'_n(\theta_0) - L'(\theta_0)) \\ &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\ell'(X_i|\theta_0) - \mathbb{E}_{\theta_0}\left[\ell'(X|\theta_0)\right]\right) \\ &\xrightarrow[n\to\infty]{TCL} \mathcal{N}(0, \mathrm{Var}_{\theta_0}\left(\ell'(X|\theta_0)\right)) = \mathcal{N}(0, I(\theta_0))\end{aligned}$$

Por otro lado, $L''_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\ell''(X_i|\theta) \underset{n\to\infty}{\longrightarrow} \mathbb{E}_{\theta_0}\left[\ell''(X|\theta)\right]$ y $\hat{\theta} \xrightarrow{\mathbb{P}} \theta_0 \Rightarrow \hat{\theta}_1 \xrightarrow{\mathbb{P}} \theta_0$.

Entonces, $L''_n(\hat{\theta}_1) \longrightarrow \mathbb{E}_{\theta_0}\left[\ell''(X|\theta_0)\right] = -I(\theta_0)$. (Aqui se uso que ℓ'' es $\mathcal{C}^3$) Usando el Lema de Slutsky, tenemos

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{(d)} \mathcal{N}(0, \frac{I(\theta_0)}{I(\theta_0)^2})$$

□

2.1.3. Máximo de verosimilitud

Comentario: Necesitamos que $L_n'' \longrightarrow L''$ uniformemente en un compacto que contiene θ_0. Pero, tenemos que $L_n \in \mathcal{C}^3(\Theta)$ lo que implica que $L_n'' \in \mathcal{C}^1(\Theta)$. En particular, L_n'' es Lipschitz y eso da que L_n'' converge uniformemente hacia L''.

Ejemplo: Tomamos de nuevo

$$\mathcal{F} = \{x \mapsto p^x(1-p)^{1-x} : p \in [0,1]\}.$$

Entonces, $\log f(x|p) = x \log p + (1-x)\log(1-p)$ y

$$\ell'(x,p) = \frac{x}{p} - \frac{1-x}{1-p} \quad \text{y} \quad \ell'(x,p) = -\frac{x}{p^2} - \frac{1-x}{(1-p)^2}$$

Y entonces

$$I(p) = -\mathbb{E}_p\left[\ell''(x,p)\right] = \frac{\mathbb{E}_p[X]}{p^2} + \frac{1-\mathbb{E}_p[X]}{(1-p)^2} = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}.$$

2.1.3. Máximo de verosimilitud

Buscamos la solución $L_n'(p) = 0$.

$$L_n'(p) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i}{p} - \frac{1-X_i}{1-p}\right) = \frac{\overline{X}_n}{p} - \frac{1-\overline{X}_n}{1-p}$$

Y entonces,

$$\begin{aligned} L_n'(\hat{p}) = 0 &\Leftrightarrow \frac{\overline{X}_n}{p} = \frac{1-\overline{X}_n}{1-p} \\ &\Leftrightarrow (1-\hat{p})\overline{X}_n = \hat{p}(1-\overline{X}_n) \\ &\Leftrightarrow \hat{p} = \overline{X}_n \end{aligned}$$

El resultado anterior dice

$$\sqrt{n}(\hat{p} - p_0) \xrightarrow{(d)} \mathcal{N}(0, p_0(1-p_0)).$$

Es el resultado del TCL!

2.1.3. Máximo de verosimilitud

Ejemplo: Tomamos

$$\mathcal{F} := \{\alpha e^{-\alpha x} : \alpha \in \mathbb{R}_+^*\}$$

Entonces, $\log f(x|\alpha) = \log \alpha - \alpha x \Rightarrow (\log f(x|\alpha))'' = -\alpha^{-2}$. Lo que permite ver $I(\alpha) = \alpha^{-2}$. Calculando el máximo de verosimilitud, tenemos

$$L_n(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\alpha} - X_i\right) = \frac{1}{\alpha} - \overline{X}_n.$$

El estimador de máximo de verosimilitud está dado por $\hat{\alpha} = \overline{X}_n^{-1}$. El resultado del teorema nos dice

$$\sqrt{n}(\hat{\alpha} - \alpha_0) \xrightarrow{(d)} \mathcal{N}(0, \alpha_0^2).$$

2.Fundamentos de estimación

2.2.Regiones de confianza

2.2.1. Definiciones

Igual que antes, supongamos dado una muestra $X_1, \ldots, X_n$ de una variable X de distribución desconocida.

Region de confianza

Para todo $\alpha \in [0, 1]$, una *region de confianza de* $g(\theta)$ al nivel $1 - \alpha$ es un conjunto medible (entonces aleatorio) $\hat{C}$ (al respecto de $X_1, \ldots, X_n$) tal que para cada $\theta \in \Theta$,

$$\mathbb{P}_\theta \left(g(\theta) \in \hat{C} \right) \geq 1 - \alpha.$$

Cuando la desigualdad se vuelve una igualdad digamos que la region de confianza es exactamente de nivel α.

Cuidado: El supuesto de uniformidad es importante! No conocemos cual es el parámetro desconocido *a priori*.

2.2.1. Definiciones

Region de confianza asintótica

Para todo $\alpha \in [0,1]$, una *region de confianza asintótica de* $g(\theta)$ al nivel $1-\alpha$ es una **sucesión** de conjuntos medibles (entonces aleatorio) $\hat{C}_n$ (al respecto de $X_1, \ldots, X_n$) tal que para cada $\theta \in \Theta$,

$$\liminf_{n\to\infty} \mathbb{P}_\theta \left(g(\theta) \in \hat{C}_n \right) \geq 1-\alpha.$$

- Se ve que la noción es asintótica por la presencia del liminf que permite que la propiedad se verifique solamente para n suficientemente grande.
- Los valores usuales de α son 1%, 5% o 10%.
- Se interpreta con frases de la forma: La probabilidad que el verdadero parámetro se ubique en $\hat{C}$ en la experiencia considerada es de al menos $1-\alpha$.

2.2.2. Intervalos de confianza

En esta parte consideramos el caso real $\theta \in \mathbb{R}$. Entonces la region de confianza es un subconjunto (aleatorio) de $\mathbb{R}$. No tiene sentido de buscar conjuntos que no sean conexos.
$\Rightarrow$ Las regiones de confianza son **intervalos de confianza** $\hat{I}_\alpha$.
Vamos a usar los diferentes métodos vistos para construirlos.

- Los estimadores son la técnica preferida para construir intervalos de confianza.
- En general, si uno conoce muy bien el comportamiento de un estimador $\hat{\theta}$, uno puede dar intervalos de confianza precisos.
- Si uno usa la normalidad asintótica para crear un intervalo de confianza, el intervalo de confianza será **asintótico** y de *nivel exactamente* α.
- Si uno usa cotas uniformes, el intervalo de confianza será **no-asintótico** pero muy probablemente de *nivel exacto mal alto* que α.

2.2.2. Intervalos de confianza

Método de cuantiles

Eso corresponde al caso ideal si uno conoce los cuantiles de la variable aleatoria $\hat{\theta} - \theta_0$ completamente. Si denotamos q_β el cuantile de orden β de esa variable, entonces

$$\begin{aligned}\mathbb{P}\left(q_{\alpha/2} < \hat{\theta} - \theta_0 \leq q_{1-\alpha/2}\right) &= F_{\hat{\theta}-\theta_0}(q_{1-\alpha/2}) - F_{\hat{\theta}-\theta_0}(q_{\alpha/2}) \\ &\geq 1 - \alpha/2 - \alpha/2 \\ &= 1 - \alpha\end{aligned}$$

Eso se puede escribir de nuevo en

$$\mathbb{P}\left(\theta_0 \in [\hat{\theta} - q_{1-\alpha/2}, \hat{\theta} - q_{\alpha/2})\right) \geq 1 - \alpha$$

lo que permite dar el intervalo $\hat{I}_\alpha = [\hat{\theta} - q_{1-\alpha/2}, \hat{\theta} - q_{\alpha/2})$.

Pero en general no tenemos la información de q_α!

2.2.2. Intervalos de confianza

Ejemplo: Imaginamos que tengamos una muestra $X_1, \ldots, X_n$ de ley $\mathcal{N}(\theta_0, 1)$ con $\theta_0 \in \mathbb{R}$ desconocido. Sabemos que $\overline{X}_n \sim \mathcal{N}(\theta, \frac{1}{n})$. Entonces,

$$\sqrt{n}(\overline{X}_n - \theta_0) \sim \mathcal{N}(0, 1).$$

Los cuantiles de $\mathcal{N}(0, 1)$ son **completamente conocidos**. Por ejemplo si tomamos $\alpha = 5\%$, $q_{0.05} \simeq -1.645$ y $q_{0.025} \simeq -1.96$, se tendrá

$$\mathbb{P}\left(\sqrt{n}(\overline{X}_n - \theta_0) \in (-\infty, 1.645]\right) \simeq 0.95$$
$$\mathbb{P}\left(\sqrt{n}(\overline{X}_n - \theta_0) \in [-1.96, 1.96]\right) \simeq 0.95.$$

Entonces,

$$\hat{I}_1 = \left[\overline{X}_n - \frac{1.96}{\sqrt{n}}, \overline{X}_n + \frac{1.96}{\sqrt{n}}\right]$$
$$\hat{I}_2 = \left[\overline{X}_n - \frac{1.645}{\sqrt{n}}, +\infty\right]$$

son dos intervalos de confianza de nivel $1 - \alpha$.

2.2.2. Intervalos de confianza

Es bueno de guardar en mente una estrategia general para construir intervalos de confianza de θ_0.

Método general

1. Elegimos un estimador $\hat{\theta}$ de la cantidad desconocida θ_0.
2. Se calcula su distribución en función de θ.
3. Se transforma este estimador para obtener una variable aleatoria tal que su ley no depende mas de θ.
4. Se determinan los cuantiles de esa ley necesarios para construir un intervalo de confianza del valor desconocido.

La etapa 4. cambia cuando la información es incompleta. Las siguientes técnicas sirven justamente a reemplazar esa ultima etapa.

2.2.2. Intervalos de confianza

Intervalos de confianza por desigualdades de probabilidad

Cuando la varianza de nuestro estimador está acotada uniformemente en θ, podemos usar la desigualdad

Proposición (B-T)

Sean una variable aleatoria Y que tiene una varianza finita $Var(Y) < +\infty$. Entonces,

$$\mathbb{P}(|Y - \mathbb{E}[Y]| \geq \epsilon) \leq \frac{Var(Y)}{\epsilon^2}$$

Supongamos que el estimador $\hat{\theta}$ es sin sesgo. Entonces en nuestro caso $\text{Var}\left(\hat{\theta}\right) = \text{Var}\left(\hat{\theta} - \theta_0\right) \leq M$ tal que M no depende de θ, por lo tanto, $\forall\theta$,

$$\mathbb{P}\left(|\hat{\theta} - \theta_0| \geq \epsilon\right) \leq \frac{M}{\epsilon^2}.$$

Finalmente, tomar $\epsilon_0 = \sqrt{M/\alpha}$ nos da un intervalo de confianza $\hat{I} = [\hat{\theta} - \epsilon_0, \hat{\theta} + \epsilon_0]$ de nivel $1 - \alpha$.

2.2.2. Intervalos de confianza

Ejemplo: Sean $X_1, \ldots, X_n$ variable de Bernoulli $\mathcal{B}(\theta_0)$. Sabemos que $\mathrm{Var}(X_1) = \theta_0(1-\theta_0) \leq \frac{1}{4}$. Entonces para cada θ,

$$\mathbb{P}_\theta\left(|\overline{X}_n - \theta| \geq \epsilon\right) \leq \frac{1}{4n\epsilon^2}$$

Eligimos $\epsilon_0 = \frac{1}{2\sqrt{n\alpha}}$, lo que nos da un intervalo $\hat{I} = [\overline{X}_n - \frac{1}{2\sqrt{n\alpha}}, \overline{X}_n + \frac{1}{2\sqrt{n\alpha}}]$ de nivel $1-\alpha$.

- En la practica, este intervalo de confianza no es muy bueno.
- Si θ_0 es lejos de $1/2$ la varianza es lejos de la cota superior...
- La desigualdad de B-T no es muy precisa.

2.2.2. Intervalos de confianza

Casos de variables acotadas

Si las variables X_i viven en un espacio **acotado** tenemos un resultado mas fuerte:

Proposición (Hoeffding)

Sean $Y_1, \ldots, Y_n$ variables aleatorias independientes tq $\forall i,\ \mathbb{E}[Y_i] = 0$ y $a_i \leq Y_i \leq b_i$ casi seguramente (donde los a_i y b_i son deterministas). Entonces,

$$\mathbb{P}\left(\sum_{i=1}^{n} Y_i \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

En particular, si las Y_i son i.i.d. con $a \leq Y_i \leq b$,

$$\mathbb{P}\left(|\overline{Y}_n| \geq \epsilon\right) \leq 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right)$$

2.2.2. Intervalos de confianza

En el ejemplo anterior, tomamos $Y_i = X_i - \theta$, $a = -\theta$ y $b = 1 - \theta$. Vemos que el valor de ϵ tiene que ser tal que

$$2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right) = \alpha$$

lo que da

$$\epsilon = \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}$$

y finalmente el intervalo de confianza que podemos escribir es

$$\hat{I} = [\bar{X}_n - \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}, \bar{X}_n + \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}]$$

El tamaño de un intervalo así es $2\sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}$. El intervalo dado por el método de B-T es $\frac{1}{2\sqrt{n\alpha}}$. **Esa segunda técnica es mucho mejor.**

2.2.2. Intervalos de confianza

Vimos que tener información sobre la varianza de nuestro estimador es fundamental para poder crear intervalos de confianza validos. Siendo un parámetro también, la varianza es casi siempre *desconocida*.

Estimación consistente de σ^2

Supongamos que el parámetro desconocido θ es la esperanza de una variable aleatoria $Y_i = g(X_i)$. En el caso de tener una estimación consistente de σ^2, podemos usar el Lema de Slutsky y la normalidad asintótica del estimador $\bar{Y}_n$.

Supongamos que

$$\sqrt{n}(\bar{Y}_n - \mathbb{E}[Y]) \xrightarrow{(d)} \mathcal{N}(0, \sigma^2) \qquad \hat{\sigma}_n^2 \xrightarrow{\mathbb{P}} \sigma^2$$

entonces,

$$\frac{\sqrt{n}}{\hat{\sigma}_n}(\bar{Y}_n - \mathbb{E}[Y]) \xrightarrow{(d)} \mathcal{N}(0, 1)$$

2.2.2. Intervalos de confianza

Entonces, podemos crear el intervalo de confianza asintótico $\hat{I}_\alpha$ dado por

$$\hat{I}_\alpha = \left[\bar{Y}_n - q_{1-\alpha/2} \frac{\hat{\sigma}_n}{\sqrt{n}}, \bar{Y}_n - q_{\alpha/2} \frac{\hat{\sigma}_n}{\sqrt{n}} \right]$$

donde q_β son los cuantiles de la distribución normal estándar. Por ejemplo en el caso de $\alpha = 5\%$, el intervalo sera dado por

$$\hat{I}_\alpha = \left[\bar{Y}_n - 1.96 \frac{\hat{\sigma}_n}{\sqrt{n}}, \bar{Y}_n + 1.96 \frac{\hat{\sigma}_n}{\sqrt{n}} \right].$$

- De nuevo esa técnica tiene la ventaja de ser muy general.
- El intervalo de confianza es asintótico.

2.2.2. Intervalos de confianza

Estabilización de varianza

La idea de la técnica de estabilización de varianza es de usar el método delta para reemplazar la estimación de θ por la estimación de $\phi(\theta)$ por una función ϕ suficientemente suave. Supongamos

- $\eta = \phi(\theta)$ donde ϕ es derivable en θ y biyectiva.
- Tenemos un estimador $\hat{\theta}$ de θ pero de varianza desconocida $\sigma(\theta)$.
- $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{(d)} \mathcal{N}(0, \sigma(\theta)^2)$

Por el método delta,

$$\sqrt{n}(\phi(\hat{\theta}) - \eta) \xrightarrow{(d)} \mathcal{N}(0, \phi'(\theta)^2 \sigma(\theta)^2)$$

Eso permite ver que una buena elección de ϕ es tal que

$$\boxed{\phi'(\theta)\sigma(\theta) = 1}.$$

2.2.2. Intervalos de confianza

Entonces un intervalo de confianza de nivel α de η sera

$$\hat{I}_\alpha = \left[\phi(\hat{\theta}) - \frac{q_{1-\alpha/2}}{\sqrt{n}}, \phi(\hat{\theta}) - \frac{q_{\alpha/2}}{\sqrt{n}}\right].$$

donde los q_β designan los cuantiles de orden β de la distribución normal estándar. En consecuencia, el intervalo de confianza para $\theta = \phi^{-1}(\eta)$ sera dado por

$$\hat{I}_\alpha = \left[\phi^{-1}\left(\phi(\hat{\theta}) - \frac{q_{1-\alpha/2}}{\sqrt{n}}\right), \phi^{-1}\left(\phi(\hat{\theta}) - \frac{q_{\alpha/2}}{\sqrt{n}}\right)\right].$$

Reducimos el problema de la estimación de varianza a un problema de ecuaciones diferenciales:

$$\phi'(\theta) = \frac{1}{\sigma(\theta)}.$$

2.2.2. Intervalos de confianza

Ejemplo: Sean $X_1, \ldots, X_n$ una muestra de variables de Poisson de parámetro λ desconocido. Sabemos que $\bar{X}_n$ es un estimador consistente de λ y

$$\sqrt{n}(\bar{X}_n - \lambda) \xrightarrow{(d)} \mathcal{N}(0, \lambda).$$

Entonces, $\sigma(\lambda) = \sqrt{\lambda}$ y buscamos una función ϕ tal que

$$\phi'(\lambda) = \frac{1}{\sqrt{\lambda}} \text{ lo que permite fijar } \phi(\lambda) = 2\sqrt{\lambda}.$$

Entonces, por el método delta,

$$\sqrt{n}(2\sqrt{\bar{X}_n} - 2\sqrt{\lambda}) \xrightarrow{(d)} \mathcal{N}(0, 1)$$

y un intervalo de confianza para el parámetro λ de nivel 95% está dado por

$$\hat{I}_{5\%} = \left[\left(\sqrt{\bar{X}_n} - \frac{1.96}{2\sqrt{n}} \right)^2, \left(\sqrt{\bar{X}_n} + \frac{1.96}{2\sqrt{n}} \right)^2 \right]$$

2.2.2. Intervalos de confianza

Intervalos simultáneos

Cuando los parámetros que estimar son dos o mas, podemos combinar los intervalos de confianza gracias a la técnica de la cota unión.

Sea $\hat{I}_1$ un intervalo de confianza de nivel $1-\alpha$ para un parámetro θ_1 y $\hat{I}_2$ un intervalo de confianza de nivel $1-\beta$ para un parámetro θ_2. Entonces,

$$\mathbb{P}\left(\theta_1 \in \hat{I}_1 \text{ y } \theta_2 \in \hat{I}_2\right) \geq 1-\alpha-\beta.$$

Consecuencia: los dos parámetros pertenecen a sus intervalos de confianza respectivos **simultáneamente** con probabilidad al menos $1-\alpha-\beta$. Es equivalente de decir que $\hat{I}_1 \times \hat{I}_2$ es una region de confianza de nivel $1-\alpha-\beta$.

2.Fundamentos de estimación

2.3.Pruebas de hipótesis

2.3.1. Formalismo

Las ideas atrás de las pruebas de hipótesis son

- Formalizar un pregunta real por la cual la muestra $X_1, \ldots, X_n$ contiene información.
- Descartar hipótesis que parecen incompatibles con los datos $X_1, \ldots, X_n$.
- Cuidado: En general, la idea de las pruebas de hipótesis no es de confirmar un hecho!!

Prueba binaria

Llamamos *prueba de hipótesis* una función $T(X_1, \ldots, X_n)$ que toma sus valores en $\{0, 1\}$.

Una *distinción de caso* sobre la distribución desconocida P_θ es una partición

$$\Theta = \Theta_0 \cup \Theta_1 \text{ con } \Theta_0 \cap \Theta_1 = \emptyset.$$

Interpretación : Si la prueba de hipótesis T es perfecta, hemos de tener

$$T = 1 \Leftrightarrow \theta \in \Theta_1 \quad \text{y} \quad T = 0 \Leftrightarrow \theta \in \Theta_0$$

2.3.1. Formalismo

Hipótesis H_0 y H_1

Denotamos $H_0 : \theta \in \Theta_0$ y $H_1 : \theta \in \Theta_1$. Veremos que en la literatura estadística, las dos hipótesis H_0 y H_1 no tienen papeles intercambiables.

Error de primer especie

La función $\underline{\alpha} : \Theta_0 \to [0, 1]$ dada por

$$\underline{\alpha}(\theta) = \mathbb{P}_\theta \left(T(X_1, \ldots, X_n) = 1 \right)$$

se llama *error de primer especie*. Se dice que la prueba T *es de nivel* $\alpha \in [0, 1]$ cuando el real

$$\alpha^* = \sup_{\theta \in \Theta_0} \underline{\alpha}(\theta)$$

es menor o igual a α.

2.3.1. Formalismo

Interpretación: α mide la probabilidad máxima de hacer una equivocación en pensar que $\theta \in \Theta_1$ (porque $T = 1$) aunque $\theta \in \Theta_0$.

Error de segundo especie
La función $\underline{\beta} : \Theta_1 \to [0, 1]$ dada por

$$\underline{\beta}(\theta) = \mathbb{P}_\theta \left(T = 0 \right)$$

se llama *error de segundo especie*.

Potencia de T
La función $\pi : \Theta \to [0, 1]$ dada por

$$\pi(\theta) = \mathbb{P}_\theta \left(T = 1 \right) = \mathbb{E}_\theta \left[T = 1 \right]$$

se llama función de potencia del la prueba T.

2.3.1. Formalismo

La función π resume las funciones $\underline{\alpha}$ y $\underline{\beta}$. En efecto, sobre Θ_0, $\pi(\theta) = \underline{\alpha}(\theta)$ y sobre Θ_1, $\pi(\theta) = 1 - \underline{\beta}(\theta)$.

Uno espera que

- π sea cerca de 0 sobre el conjunto Θ_0,
- π sea cerca de 1 sobre el conjunto Θ_1.
- Pero en muchos casos hay una frontera entre Θ_0 y Θ_1 y la función π es continua.

Resulta que encontrar una buena prueba de hipótesis en una vez mas un juego de balance/optimización de dos cantidades dependiendo de la distribución desconocida.

Relación UMP(α)

Sean dos tests T_1 y T_2 de nivel α. Digamos que T_1 es *mas potente* que T_2 si $\forall \theta \in \Theta_1$, $1 - \underline{\beta}_1(\theta) \geq 1 - \underline{\beta}_2(\theta)$. Se dice que T es *UMP*(α) si es mas popente que cada prueba de nivel α.

2.3.1. Formalismo

Ejemplo: Si consideramos una muestra de variables $X_1, \ldots, X_n \sim \mathcal{N}(\theta, 1)$, podemos probar si

- $H_0 : \theta \geq 0$ o,
- $H_1 : \theta < 0$.

Con nuestro formalismo, eso sera $\Theta_0 = [0, +\infty)$ y $\Theta_1 = (-\infty, 0)$. Una manera natural de proceder es de basarse sobre la cantidad $\bar{X}_n$ y de considerar

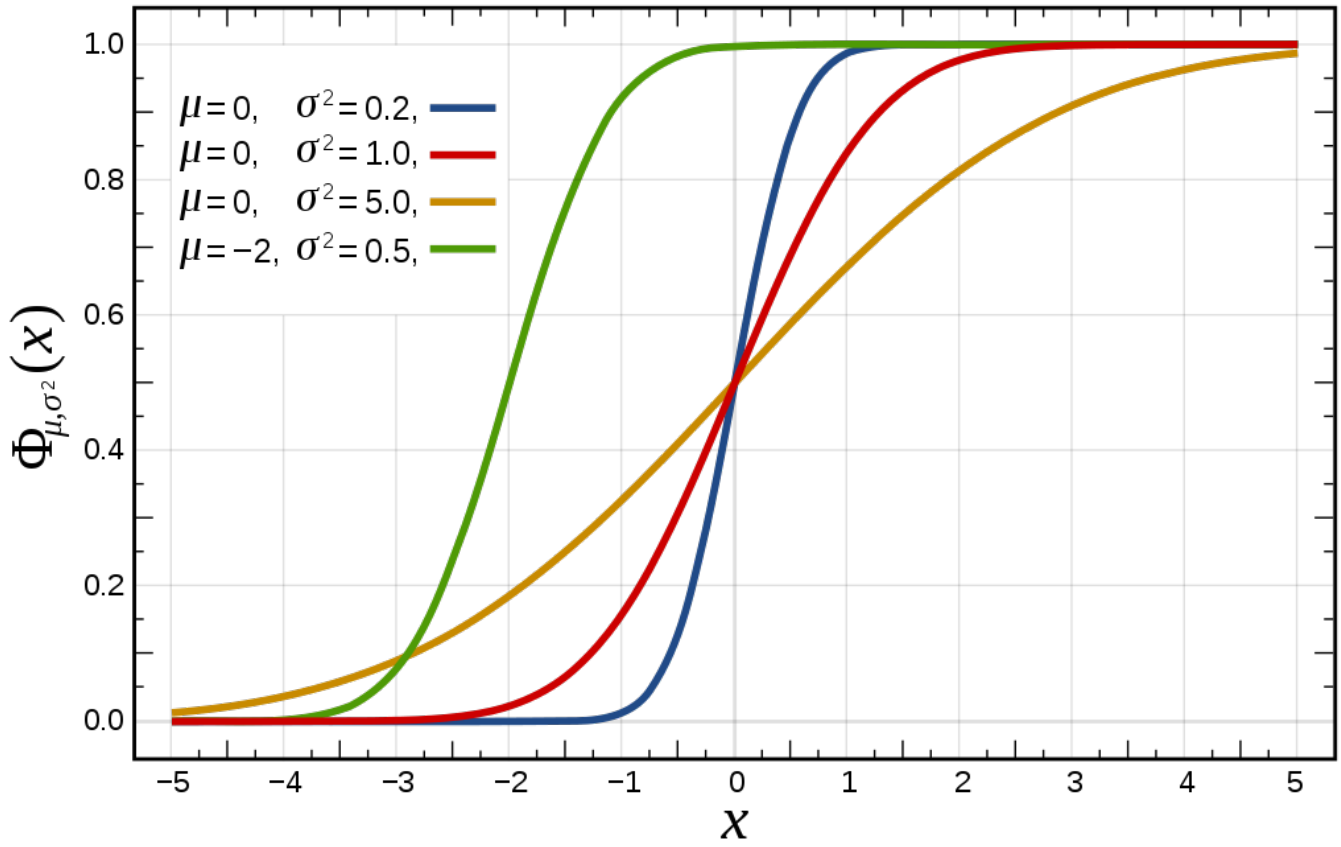
$$T = \mathbb{1}_{\bar{X}_n < 0}.$$

Basado a esa formula de definición, podemos calcular los errores α y β. Calculamos directamente la función de potencia. Como sabemos que $\bar{X}_n \sim \mathcal{N}(\theta, \frac{1}{n})$,

$$\pi(\theta) = \mathbb{P}_\theta \left(\bar{X}_n < 0 \right) = \Phi(-\theta\sqrt{n})$$

donde Φ es la función de distribución de una normal estándar.

2.3.1. Formalismo



Se ve que cuando $n \to \infty$, la función de potencia π se acerca de la función de potencia ideal. Pero,

$$\alpha^* = \sup_{\theta \geq 0} \pi(\theta) = \Phi(0) = 0.5$$

lo que no es genial...

2.3.1. Formalismo

En conclusión, se introduce una disimétrica entre las dos hipótesis. (Ver etapa 3.)

Metodología

Basándonos sobre el ejemplo, podemos pensar en una metodología para crear pruebas de hipótesis.

1. Identificar un estimador $\hat{\theta}$ del parámetro en juego en las hipótesis H_0 y H_1.
2. Se elige un nivel de prueba α.
3. Se calcula una *zona R de rechazo* tal que $\alpha^* \leq \alpha$. Es la zona donde queremos que $\theta \in R \Leftrightarrow T = 1$.
4. Calculamos la potencia de la prueba para confirmar la utilidad de la prueba de hipótesis.
5. (opcional) Calculamos el *p*-valor.

2.3.1. Formalismo

Una otra manera de cuantificar la cualidad de una prueba de hipótesis T es a través de la noción de p-valor.

p-valor

Ahora supongamos que tenemos a nuestra disposición una colección de tests T_α tal que para cada $\alpha \in [0,1]$, T_α es de nivel α. La cantidad

$$p(X_1, \ldots, X_n) = \sup\{\alpha : T_\alpha(X_1, \ldots, X_n) = 0\}$$

se llama *p-valor* de la (colección de) prueba de hipótesis T_α.

Cuando uno solo tiene una única prueba, o que solo hay un solo α disponible, tendremos $p = \alpha$. Por eso, ciertos autores confunden las dos nociones.

2.3.1. Formalismo

Comentarios

- El p-valor es una variable aleatoria en $[0, 1]$.
- En términos informales, el p-valor es el mayor nivel que autoriza a aceptar H_0. Entonces, es un indice de la credibilidad de H_0.
- En la practica, rechazamos H_0 cuando el p-valor es bajo un cierto nivel típico (por ejemplo 5%).
- Si p es grande, no significa que podemos aceptar H_0! La colección de pruebas de hipótesis puede no ser potente...
- Casi seguramente tenemos $\alpha \mapsto T_\alpha(X_1, \ldots, X_n)$ es creciente, lo que justifica la definición y permite ver que

$$p(X_1, \ldots, X_n) = \inf\{\alpha : T_\alpha(X_1, \ldots, X_n) = 1\}$$

- La sucesión de la regiones de rechazo es creciente (en el sentido de la inclusión).

2.3.1. Formalismo

Ejemplo: Imaginamos que nuestra prueba de hipótesis sea de la forma

$$T_\alpha = \mathbb{1}_{\bar{X}_n < k_\alpha}$$

donde $\alpha \mapsto k_\alpha$ es una función creciente de α para asegurar que T_α sea de nivel α tq $k_0 = -\infty$ y $k_1 = +\infty$. Condicionalmente a las variables $X_1, \dots, X_n$, $\bar{X}_n$ es fijo y podemos ver que el p-valor está caracterizado por

$$k_p = \bar{X}_n$$

Entonces, si retomamos el ejemplo anterior (con esa clase de T_α), tenemos $\alpha = \sup_{\theta \geq 0} \Phi((k_\alpha - \theta)\sqrt{n}) = \Phi(k_\alpha \sqrt{n})$ lo que permite deducir que

$$p = \Phi(\bar{X}_n \sqrt{n})$$

Comentario: En general cuando la prueba de hipótesis es de la forma $T_\alpha = \mathbb{1}_{h(X_1,\dots,X_n) \leq k_\alpha}$ donde $\alpha^* = \alpha$, el p-valor esta dado por

$$p(x_1, \dots, x_n) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta \left(h(X_1, \dots, X_n) \leq h(x_1, \dots, x_n) \right)$$

2.3.2. Ejemplos de pruebas de hipótesis

Vamos a estudiar varios contextos donde intentamos de definir buenos tests. En algunos contextos, podemos probar que esos tests son UMP(α).

Ajuste a un parámetro

Las dos alternativas son

$$H_0 : \theta = \theta_{ref}$$
$$H_1 : \theta \neq \theta_{ref} \quad \text{o} \quad \theta < \theta_{ref} \quad \text{o} \quad \theta > \theta_{ref}$$

Ejemplo: Supongamos tener $X_1, \ldots, X_n$ de ley de momento de orden 2 finito. Por el teorema central limite,

$$h_n(X) = \sqrt{n}\left(\frac{\bar{X}_n - \mu_{ref}}{\sqrt{\sigma_n^2}}\right)$$

converge a una variable normal estándar si $\mathbb{E}[X_i] = \mu_{ref}$.

2.3.2. Ejemplos de pruebas de hipótesis

Entonces podemos definir

$$T_\alpha = \mathbb{1}_{h_n(X) \in R_\alpha}$$

donde $R_\alpha = (-\infty, q_{\alpha/2}) \cup (q_{1-\alpha/2}, +\infty)$ donde q_β son los cuantiles de la variable normal estándar. Por definición T_α es de nivel α.

Si $X = (X_1, \ldots, X_n)$ es de media $\mu \neq \mu_{ref}$, $h_n(X) \xrightarrow{\mathbb{P}} \pm\infty$. Entonces, a partir de un cierto n suficiente grande, $h_n(X) \in R_\alpha$. Y finalmente $\underline{\beta}(\theta) \underset{n\to\infty}{\longrightarrow} 0$ lo que permite decir que esta prueba es asintóticamente optima.

2.3.2. Ejemplos de pruebas de hipótesis

Prueba de ajuste a una distribución

Sea una distribución definida por su función de distribución G. Sea F la función de distribución de la muestra $X_1, \ldots, X_n$.
Consideramos las alternativas

$$H_0 : F = G$$
$$H_1 : F \neq G$$

Consideramos la función de distribución empírica $F_n(x) = n^{-1} \sum \mathbb{1}_{X_i \leq x}$. La estadística

$$h_n(X_1, \ldots, X_n) = \sup_{x \in \mathbb{R}} |F_n(x) - G(x)|$$

permite de distinguir entre las dos alternativas. En particular, se puede probar que

$$\mathbb{P}_F \left(\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| > \frac{c}{\sqrt{n}} \right) \longrightarrow \alpha(c) = 2 \sum_{r=1}^{\infty} (-1)^{r-1} \exp(-2c^2 r^2)$$

2.3.2. Ejemplos de pruebas de hipótesis

Si tomamos c tal que $\alpha(c) = \alpha$, la prueba $T_\alpha = \mathbb{1}_{h_n(X) \in R_\alpha}$ con $R_\alpha = (-\infty, -c/\sqrt{n}) \cup (c/\sqrt{n}, +\infty)$ es de nivel asintótico α. Esa metodología se llama *prueba de Kolmogorov-Smirnov*.

Prueba de Shapiro-Wilk

Esa prueba sirve en el caso donde G sea una distribución normal. Se basa sobre la estadística

$$h_n(X) = \frac{\left(\sum_{i=1}^n a_i x_{(i)}\right)^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}$$

donde

- $x_{(i)}$ son las estadísticas de orden.
- $(a_1, \ldots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}}$,
- donde $m = (m_1, \ldots, m_n)^T$ son las esperanzas de $(N_{(1)}, \ldots, N_{(n)})$ y V la matriz de varianza-covarianza de este vector.

2.3.2. Ejemplos de pruebas de hipótesis

Prueba de Neyman-Pearson

El contexto de esta prueba es de distinguir entre las dos alternativas

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta_1.$$

Teorema (Neyman-Pearson)

Sea $\alpha \in (0, 1)$ *y denotamos*

$$h_n(X) = \frac{V_X(\theta_1)}{V_X(\theta_0)} = \frac{f_{\theta_1}(X)}{f_{\theta_0}(X)}$$

la razón de verosimilitud. Sea $T_\alpha = \mathbb{1}_{h_n(X) > k_\alpha}$ *tal que* $\mathbb{P}_{\theta_0}(T_\alpha = 1) = \alpha$. *Entonces, esta prueba de hipótesis es UMP*(α).

2.3.2. Ejemplos de pruebas de hipótesis

Prueba.

Sea ϕ una otra prueba de hipótesis de nivel α lo que se escribe

$$\mathbb{E}_{\theta_0}[\phi(X)] = \mathbb{P}_{\theta_0}(\phi = 1) \leq \alpha \quad \text{y} \quad \mathbb{E}_{\theta_0}[T(X) - \phi(X)] \geq 0.$$

La meta es de mostrar que $\mathbb{E}_{\theta_1}[T] \geq \mathbb{E}_{\theta_1}[\phi]$. Pero

$$\begin{aligned}
&\mathbb{E}_{\theta_1}[T(X) - \phi(X)] - k_\alpha \mathbb{E}_{\theta_0}[T(X) - \phi(X)] \\
&= \mathbb{E}_{\theta_0}\left[(T(X) - \phi(X))\frac{f_{\theta_1}(X)}{f_{\theta_0}(X)}\right] + \mathbb{E}_{\theta_1}\left[(T(X) - \phi(X))\mathbb{1}_{f_{\theta_0}(X)=0}\right] \\
&- k_\alpha \mathbb{E}_{\theta_0}[T(X) - \phi(X)] \\
&= \mathbb{E}_{\theta_0}\left[(T(X) - \phi(X))\left(\frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} - k_\alpha\right)\right] + \mathbb{E}_{\theta_1}\left[(1 - \phi(X))\mathbb{1}_{f_{\theta_0}(X)=0}\right] \\
&\geq 0
\end{aligned}$$

Entonces, $\mathbb{E}_{\theta_1}[T(X) - \phi(X)] \geq k_\alpha \mathbb{E}_{\theta_0}[T(X) - \phi(X)] \geq 0$ □

2.3.2. Ejemplos de pruebas de hipótesis

Ejemplo: Si tomamos una vez mas nuestro ejemplo $X_1, \dots, X_n \sim \mathcal{N}(\theta, 1)$ y que consideramos las dos alternativas

$$H_0 : \theta = 0$$
$$H_1 : \theta = 1.$$

La razón de verosimilitud es $h(X) = \frac{V_X(1)}{V_X(0)}$ donde

$$V_X(\theta) = \left(\frac{1}{\sqrt{2\pi}}\right) \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(X_i - \theta)^2\right).$$

Entonces, eso se escribe

$$h_n(X) = \exp\left(\frac{n}{2}(2\bar{X}_n - 1)\right)$$

Finalmente, el test toma la forma

$$T_\alpha = \mathbb{1}_{\bar{x}_n > k_\alpha}$$

3.Modelo gaussiano lineal

3.1.Vectores gaussianos

3.1.1. Repasos

Recordamos rápidamente los hechos fundamentales de los vectores gaussianos.

Definición

Sea $X = (X_1, \ldots, X_k)^T$ un vector aleatorio. Se dice que X es un *vector gaussiano* si y solo si $\forall t \in \mathbb{R}^k$, $t^T X$ es una variable gaussiana.

- Su esperanza está dada por $\mu = \mathbb{E}[X] = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_k])^T$.
- Su matriz de varianza-coravianza $\Sigma = \mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T\right]$.
- Es de función de característica, $\forall t \in \mathbb{R}^k$,

$$\phi_X(t) = \exp\left(it^T\mu - \frac{t^T\Sigma t}{2}\right).$$

3.1.1. Repasos

Linealidad y Independencia

Repasamos algunas propiedades:

Proposición

Sea $X \in \mathbb{R}^k$ un vector gaussiano $\mathcal{N}(\mu, \Sigma)$. Entonces

- $\forall A \in \mathcal{M}_{k' \times k}(\mathbb{R})$ *y* $b \in \mathbb{R}^{k'}$, *tenemos*

$$AX + b \sim \mathcal{N}(A\mu + b, A\Sigma A^T)$$

- *Las componentes X_i son tal que*

$$\forall i \neq j,\ X_i \text{ y } X_j \text{ son independientes} \Leftrightarrow Cov(X_i, X_j) = 0.$$

Comentario: El sentido $\Rightarrow$ es verdad para cada tipo de distribución.
El sentido $\Leftarrow$ es únicamente verdad para vectores gaussianos.

3.1.1. Repasos

En los capítulos anteriores, trabajamos con vectores gaussianos sin dar una definición limpia (ver sección TCL 1.6.2) Aquí tratamos el caso general.

Densidad de un vector gaussiano

Sea X un vector gaussiano de $\mathbb{R}^k$ y sean $\mu = \mathbb{E}[X]$ y $\Sigma = \mathrm{Var}(X)$. Tenemos

1. La distribución de X admite una densidad en $\mathbb{R}^k$ si y solo si Σ es invertible.
2. En este caso, $\forall x \in \mathbb{R}^k$,

$$f_{\mu,\Sigma}(x) = \left(\frac{1}{\sqrt{2\pi}}\right)^k \frac{1}{\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

3. Si Σ no es invertible, $X - \mu$ pertenece c.s. al sub-espacio de $\mathbb{R}^k$ generado por los vectores propios asociados a los valores propios distinto de zero de Σ.

3.1.1. Repasos

Prueba.

La matrix Σ es simétrica y positiva entonces es diagonalisable en una base orthonormal $(u_1, \ldots, u_k)$ asociada a los valores propios $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$.

Sea U la matriz ortogonal creada por los vectores u_j y denotamos r el rango de esa matriz. Entonces $\lambda_r > 0$ y $\lambda_{r+1} = \cdots = \lambda_k = 0$.

$$\Sigma = U\Gamma U^T = (U\sqrt{\Gamma})(U\sqrt{\Gamma})^T$$

donde

$$\sqrt{\Gamma} = \begin{pmatrix} \sqrt{\lambda_1} & & & & & \\ & \ddots & & & & \\ & & \sqrt{\lambda_r} & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix}$$

3.1.1. Repasos

Prueba.

Por linealidad, los dos vectores X y $\mu + U\sqrt{\Gamma}Y$ donde $Y \sim \mathcal{N}(0, \mathrm{Id})$ tienen la misma distribución. Por definición, $\mathrm{Im}(U\sqrt{\Gamma})$ es el espacio generado por los vectores propios de Σ lo que justifica 3.

Supongamos que Σ es invertible. Sea g una función continua y acotada.

$$\begin{aligned}
&\mathbb{E}[g(\mu + U\sqrt{\Gamma}Y)] \\
&\quad = \int_{\mathbb{R}^k} g(\mu + U\sqrt{\Gamma}Y) \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y_i^2}{2}\right) dy_i \\
&\quad = \int_{\mathbb{R}^k} \frac{g(X)}{(2\pi)^{k/2}\det(U\sqrt{\Gamma})} \exp\left(-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right) dX
\end{aligned}$$

por el cambio de variable $X = \mu + U\sqrt{\Gamma}Y$. □

3.1.1. Repasos

Prueba.
En efecto,

$$\begin{aligned}\|Y\|^2 &= \|(X-\mu)\sqrt{\Gamma}^{-1}U^{-1}\|^2 \\ &= \|\sqrt{\Gamma}^{-1}U^{-1}(X-\mu)\|^2 \\ &= (X-\mu)^T U\sqrt{\Gamma}^{-1}\sqrt{\Gamma}^{-1}U^{-1}(X-\mu) \\ &= (X-\mu)^T\Sigma^{-1}(X-\mu)\end{aligned}$$

y el determinante de la matriz Jacobiana está dado por

$$|\det(J)| = |\det(U\sqrt{\Gamma})| = |\det(\sqrt{\Gamma})| = \sqrt{\det(\Gamma)}$$

□

Comentario: El limite del TCL vectorial es un vector gaussiano!

3.1.2. Distribuciones de χ^2 y de Student

Distribución de χ^2

La distribución del χ^2 se define gracias a la estructura euclidiana del espacio. Consideramos el espacio vectorial $(\mathbb{R}^k, \langle \cdot, \cdot \rangle)$ y su norma euclidiana $\| \cdot \|_2$ asociada.

Def-Teorema (χ^2)

Sea $X = (X_1, \ldots, X_k)$ un vector gaussiano de $\mathbb{R}^k$ tal que $\mathbb{E}[X] = \mu$ y $Var(X) = Id$. Entonces, la distribución de la variable $\|X\|_2^2 = X_1^2 + \cdots + X_k^2$ depende unicamente de los parámetros k y $\|\mu\|_2$. Digamos que $\|X\|_2^2$ sigue una ley del χ^2 con k grados de libertad y centro $\|\mu\|_2^2$. Se nota

$$\|X\|_2^2 \sim \chi^2(k, \|\mu\|_2^2).$$

Cuando $\|\mu\| = 0$, denotamos $\|X\|_2^2 \sim \chi^2(k)$.

3.1.2. Distribuciones de χ^2 y de Student

Que hay que probar?

Prueba.

La distribución depende a priori de k y de μ.

Supongamos que $\|\mu\| > 0$. $X \sim (\mu, \mathrm{Id})$ y sea $Y \sim (\mu', \mathrm{Id})$ tq $\|\mu\|_2 = \|\mu'\|_2$. Entonces, existe una matriz ortogonal tq $\mu = U\mu'$.

Tenemos, $UY \sim \mathcal{N}(U\mu', U\mathrm{Id}U^T) = \mathcal{N}(\mu, \mathrm{Id})$ lo que prueba que X y UY tienen la misma distribución.

Pero, $\|Y\|_2^2 = \|UY\|_2^2 \sim \|X\|_2^2$, y entonces $\|X\|_2$ solo depende de $\|\mu\|_2$. □

3.1.2. Distribuciones de χ^2 y de Student

Se puede calcular $\mathbb{E}[X]$ y $\mathrm{Var}(X)$ por $X \sim \chi^2(k)$.

- Nos basamos sobre el hecho: si $N \sim \mathcal{N}(0,1)$ entonces $\mathbb{E}\left[N^2\right] = 1$ y $\mathbb{E}\left[N^4\right] = 3$. Entonces,

$$\mathbb{E}[X] = \sum_{i=1}^{k} \mathbb{E}\left[X_i^2\right] = k$$

$$\mathbb{E}\left[X^2\right] = \mathbb{E}\left[\left(\sum_{i=1}^{k} X_i^2\right)^2\right] = \sum_{i=1}^{k} \mathbb{E}\left[X_i^4\right] + \sum_{i \neq j} \mathbb{E}\left[X_i^2\right] \mathbb{E}\left[X_j^2\right]$$

$$= 3k + k^2 - k$$

Entonces, $\mathrm{Var}(X) = 2k + k^2 - k^2 = 2k$.

3.1.2. Distribuciones de χ^2 y de Student

- Se puede calcular la densidad de $\chi^2(k)$.

$$f_{\chi^2(k)}(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$$

donde $\Gamma(a) = \int_0^{+\infty} t^{a-1} e^{-t} dt$, con la técnica del cambio de variable ...

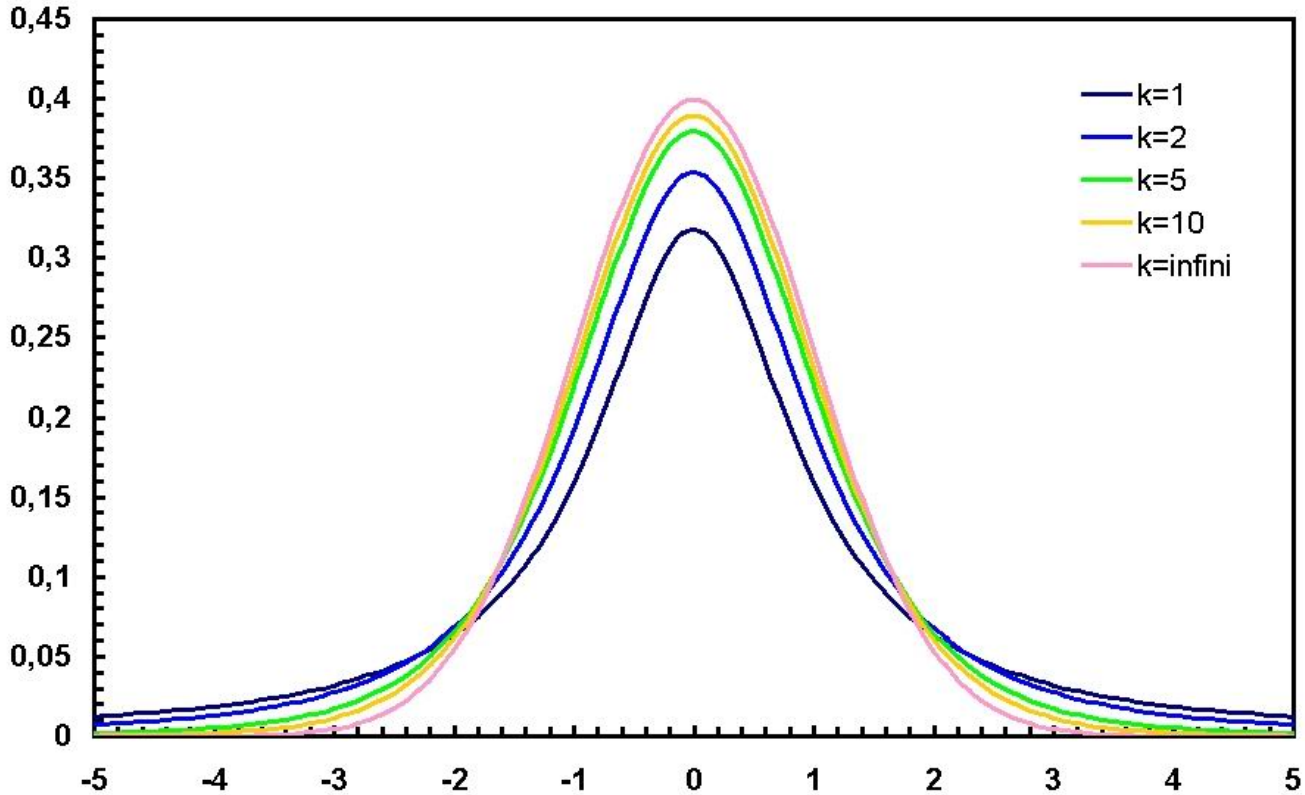
Ley de Student

Sea $X \sim \mathcal{N}(\mu, 1)$ y $Y \sim \chi^2(k)$. La distribución de

$$Z = \frac{X}{\sqrt{\frac{Y}{k}}}$$

se llama *distribución de Student* con k grados de libertad. Se denota $Z \sim \tau(k, \mu)$.
Cuando $\mu = 0$, se nota $Z \sim \tau(k)$.

3.1.2. Distribuciones de χ^2 y de Student



- Hay la convergencia $\tau(n) \xrightarrow{(d)} \mathcal{N}(0,1)$ (Slutsky y TCL)
- Las colas de $\tau(k)$ son mas pesadas que las de $\mathcal{N}(0,1)$.
- Como $\mathcal{N}(0,1)$, $\tau(k)$ es simétrica.
- Su densidad esta dada por

$$f_{\tau(k)}(x) = \frac{\Gamma((k+1)/2)}{\sqrt{k\pi}\Gamma(k/2)}(1+\frac{x^2}{k})^{-\frac{k+1}{2}}$$

3.1.3. Teorema de Cochran

Una herramienta fundamental es

Teorema (Cochran)

Sea $X \sim \mathcal{N}(\mu, Id)$ *y* $E_1 \oplus \cdots \oplus E_r = \mathbb{R}^k$ *una descomposición del espacio en sub-espacios ortogonales de dimensiones respectivas* $d_1, \ldots, d_r$.
Sean $\Pi_{E_1}, \ldots, \Pi_{E_r}$ *las proyecciones ortogonales sobre* $E_1, \ldots, E_r$.
Entonces,

- *Los vectores* $\Pi_{E_1} X, \ldots, \Pi_{E_r} X$ *son gaussianos y independientes.*
- *For all* $j = 1, \ldots, r$

$$\|\Pi_{E_j} X\|_2^2 \sim \chi^2(d_j, \|\Pi_{E_j} \mu\|_2^2)$$

3.1.3. Teorema de Cochran

Prueba.

Para todo $j \in \{1, \dots, n\}$, sea $(e_{j,1}, \dots, e_{j,d_j})$ una base ortonormal de E_j. Entonces,

$$\Pi_{E_j} X = \sum_{k=1}^{d_j} (e_{j,k}^T X) e_{j,k}.$$

Sea $U = \left((e_{1,1}, \dots, e_{1,d_1}), \dots, (e_{r,1}, \dots, e_{r,d_r})\right)^T$ la matriz ortogonal de cambio de variables. Es tal que $UU^T = \mathrm{Id}$. Tenemos

$$UX \sim \mathcal{N}(U\mu, \mathrm{Id}).$$

En particular, las coordenadas de UX son independientes (porque $\mathrm{Var}\,(UX) = \mathrm{Id}$). Esas variables son las

$$e_{j,k}^T X$$

Entonces, los grupos formados de esas variables forman grupos de variables independientes. □

3.1.3. Teorema de Cochran

Prueba.

Y $\forall j, k$,

$$e_{j,k}^T X \sim \mathcal{N}(e_{j,k}^T \mu, 1)$$

Entonces,

$$\|\Pi_{E_j} X\|_2^2 = \sum_{k=1}^{d_j} (e_{j,k}^T X)^2$$

es una suma de variables gaussianas independientes formando una vector gaussiano de esperanza $\Pi_{E_j}\mu$. Entonces $\|\Pi_{E_j} X\|_2^2 \sim \chi^2(d_j, \|\Pi_{E_j}\mu\|_2^2)$. □

Independencia de los estimadores μ y σ^2

Nos ponemos en el contexto de una muestra de variables gaussianas de distribución $\mathcal{N}(\mu, \sigma^2)$ con μ y σ^2 desconocidos. La log-verosimilitud se maximiza por

$$\hat{\mu}_n = \bar{X}_n \qquad \text{y} \qquad \hat{s}_n^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

3.1.3. Teorema de Cochran

Proposición

Sean $X_1, \ldots, X_n$ variables i.i.d. de distribución $\mathcal{N}(\mu, \sigma^2)$.
Entonces,

- *$\hat{\mu}_n$ y $\hat{s}_n^2$ son independientes.*
- $\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$.
- $\frac{n\hat{s}_n^2}{\sigma^2} \sim \chi^2(n-1)$.
- $\frac{\bar{X}_n - \mu}{\sqrt{\hat{s}_n^2/(n-1)}} \sim \tau(n-1)$.

Prueba.

Sean $Y_i = \frac{X_i - \mu}{\sigma}$ de tal manera que $X_i \sim \mathcal{N}(0,1)$. Sea $E = \text{Vect}(e)$ donde $e = (1, \ldots, 1)^T$ con $\mathbb{R}^n = E \oplus E^{\perp}$.
Entonces

$$\Pi_E Y = \langle \frac{e}{\sqrt{n}}, Y \rangle \frac{e}{\sqrt{n}} = \frac{1}{n}(e^T Y)e = \frac{1}{n}\sum_{i=1}^{n} Y_i e = \bar{Y}_n e.$$

□

3.1.3. Teorema de Cochran

Por el otro lado,

$$\Pi_{E^\perp} Y = Y - \Pi_E Y = \begin{pmatrix} Y_1 - \bar{Y}_n \\ \vdots \\ Y_n - \bar{Y}_n \end{pmatrix}.$$

Por el teorema de Cochran, lo vectores gaussianos

$$\bar{Y}_n e \qquad \text{y} \qquad Y - \Pi_E Y$$

son independientes así que lo mismo ocurre para $\bar{X}_n e$ y $X - \Pi_E X$. Como $\hat{\mu}_n$ es una función de $\bar{X}_n e$ y $\hat{s}_n^2$ es una función de $X - \Pi_E X$, tenemos que $\hat{\mu}_n$ y $\hat{s}_n^2$ son independientes. Finalmente,

$$\frac{n\hat{s}_n^2}{\sigma^2} = \|\Pi_{E^\perp} Y\|_2^2 \sim \chi^2(n-1)$$

y $\frac{\bar{X}_n - \mu}{\sqrt{\hat{s}_n^2/(n-1)}} \sim \tau(n-1)$.

3.1.3. Teorema de Cochran

Ejemplo:

- Si denotamos $t = t_{n-1,1-\alpha/2}$ el cuantile de orden $1 - \frac{\alpha}{2}$ de $\tau(n-1)$, definimos

$$I_{n,\alpha} = \left[\bar{X}_n - t\sqrt{\hat{s}_n^2/(n-1)}, \bar{X}_n + t\sqrt{\hat{s}_n^2/(n-1)}\right]$$

de tal manera que $\mathbb{P}(\mu \in I_{n,\alpha}) = 1 - \alpha$.

- Si denotamos $C = C_{n-1,1-\alpha}$ el cuantile de orden $1 - \alpha$ de $\chi^2(n-1)$ entonces

$$J_{n,\alpha} = \left[0, \frac{n\hat{s}_n^2}{C}\right]$$

es un intervalo de confianza de orden $1 - \alpha$ de σ^2.

- Por independencia,

$$\mathbb{P}\left((\mu, \sigma^2) \in I_{n,\alpha} \times J_{n,\alpha}\right) = (1-\alpha)^2 \geq 1 - 2\alpha.$$

3.Modelo gaussiano lineal

3.2.Modelo gaussiano lineal

3.2.1. Dos modelos

Uno tiene acceso a un vector de observaciones $Y = (Y_1, \ldots, Y_n)^T \in \mathbb{R}^n$. Denotamos m su vector de esperanzas.

Modelo gaussiano lineal (F1)

- Digamos que Y sigue un *modelo gaussiano lineal* si

$$Y = m + \epsilon \tag{F1}$$

donde $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. Los parámetros desconocidos son m y σ^2.

- Uno hace la hipótesis que $m \in V$ donde V es un sub-espacio conocido de $\mathbb{R}^n$.

Comentario: Aquí es común de tener que $\dim(V) = p$ sea mucho mas pequeño que n. Uno piensa de la manera siguiente:

"Tratamos de encontrar p variables para explicar un vector de dimensión n."

3.2.1. Dos modelos

Vamos a ver una segunda formulación basada en el formalismo del modelo estadístico.

- Conjunto de parámetros : $\Theta = V \times \mathbb{R}_+$.
- Para cada $\theta = (m, \sigma^2)$, $P_\theta = \mathcal{N}(m, \sigma^2 I_n)$
- El modelo es $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), P_\theta, \theta \in V \times \mathbb{R}_+)$.

Podemos escribir:

Modelo gaussiano lineal (F2)

- Digamos que Y sigue un *modelo gaussiano lineal* si

$$Y = X\beta + \epsilon \tag{F2}$$

  donde $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$.
- El parámetro $\beta = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ es desconocido y la matrix $X \in \mathcal{M}_{n\times p}(\mathbb{R})$ es conocida. Se llama *matriz de variables explicativas*.

Dar la información de X es equivalente a dar la información de V.

3.2.1. Dos modelos

Comentario: Si $t \in \{1, \ldots, n\}$,

$$Y_t = \sum_{j=1}^{p} \beta_j X_{t,j} + \epsilon_t.$$

- Y_t es el resultado.
- Los β_j dan la dependencia lineal.
- Los $X_{t,j}$ son los condiciones de la experiencia (valores de los parámetros de experiencia).
- Los ϵ_t es el impacto de otros factores exteriores no tomados en cuenta en la modelación.

3.2.2. Ejemplos

Ejemplo 1: Una empresa de automobiles tiene buenas razones de pensar que el precio Y_t que una persona t está dispuesta a pagar para un coche se relaciona linealmente a su ingreso $X_{t,1}$ y a la distancia $X_{t,2}$ recorrida cada día por esa persona t. La modelación propuesta es entonces

$$Y_t = \mu + \beta_1 X_{t,1} + \beta_2 X_{t,2} + \epsilon_t$$

con $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$. Es la formulación (F2).

Si uno pone

$$m_t = \mu + \beta_1 X_{t,1} + \beta_2 X_{t,2}$$

y $m = (m_1, \ldots, m_n)^T$ pasamos a la formulación (F1) con

$$V = \text{Vect}(X_0, X_1, X_2)$$

con $X_0 = (1, \ldots, 1)^T$, $X_1 = (X_{1,1}, \ldots, X_{n,1})^T$ y $X_2 = (X_{1,2}, \ldots, X_{n,2})^T$.

3.2.2. Ejemplos

Ejemplo 2: Una empresa agro-química quiere probar un nuevo fertilizante orgánico. Lo compara con su mejor fertilizante actual. Quiere probar la diferencia de efecto en dos tipos de cultivos : Cereales o plantas verdes. Dispone de $n = 4q$ terrenos donde aplica q veces cada combinación posible. Denotamos c_1, c_2, c_3, c_4 el rendimiento medio del cultivo en cada de los cuatro casos referido a un rendimiento de referencia μ. La modelación se escribe $\forall t \in \{1, \ldots, q\}$ y $\forall k \in \{1, 2, 3, 4\}$,

$$Y_{t+(k-1)q} = \mu + c_k + \epsilon_{t+(k-1)q}$$

con $\epsilon_{t+(k-1)q} \sim \mathcal{N}(0, \sigma^2)$.

Ejercicio : Escribir la matriz X que corresponde en este caso.

3.2.2. Ejemplos

Comentarios:

- Si $\text{Ker}X \neq \{0\}$ el modelo **no** es identifiable.
  Pedir que $\text{Ker}X = \{0\}$ es equivalente a $\text{rg}(X) = p$ o a $\text{Im}(X) = V$.
- En el ejemplo 2, el modelo no es identifiable.
- Cuando el modelo no es identifiable, reducimos los parámetros a un conjunto ma pequeño. Por ejemplo $\mathbb{R}^p \to V'$ tq $\forall m \in \text{Im}(X), \exists! \beta \in V'$ con $m = X\beta$.

  Por ejemplo, podemos tomar $V' = (\text{Ker}X)^{\perp}$.

En el ejemplo 2, podemos tomar

$$\beta \in V' = \{(\mu, c_1, c_2, c_3, c_4)^T : c_1 + c_2 + c_3 + c_4 = 0\}$$

o

$$\beta \in V' = \{(\mu, c_1, c_2, c_3, c_4)^T : \mu = 0\}$$

3.2.3. Estimadores de m y σ^2

Vamos a dar dos resultados equivalentes basados en las dos formulaciones del modelo gaussiano lineal.

Teorema (F1)

Tomamos la formulación (F1) : $Y = m + \epsilon$.

- *El estimador MLE (de máximo de verosimilitud) es* $\hat{\theta}_n = (\hat{m}, \hat{s}_n^2)$ *donde*

$$\hat{m} = \Pi_V Y$$
$$\hat{s}_n^2 = \frac{1}{n} \| Y - \Pi_V Y \|_2^2$$

- *Los estimadores* $\hat{m}$ *y* $\hat{s}_n^2$ *son independientes y*

$$\hat{m} \sim \mathcal{N}(m, \sigma^2 \Pi_V) \qquad y \qquad \frac{n}{\sigma^2} \hat{s}_n^2 \sim \chi^2(n-p)$$

Prueba.

Es una otra aplicación de Cochran. (con el hecho $\Pi_V I_n \Pi_V^T = \Pi_V$)

□

3.2.3. Estimadores de m y σ^2

Comentarios

- $\hat{m}$ no tiene sesgo pero $\mathbb{E}\left[\hat{s}_n^2\right] = \sigma^2 \frac{n-p}{n}$ y entonces $\hat{s}_n^2$ tiene sesgo (que depende de p y n).
- Un otro estimador sin sesgo de la varianza es $\hat{\sigma}_n^2 = \frac{1}{n-p}\|Y - \Pi_V Y\|_2^2$.
- $\hat{m}$ es también el estimador de mínimos cuadrados:

$$\hat{m} \in \arg\min_{m \in V} \|Y - m\|_2$$

- Recordamos que p es la dimensión del espacio V y que por definición $V = \text{Im } X$.

3.2.3. Estimadores de m y σ^2

Propongamos una otra formulación del teorema en la formulación (F2).

Teorema (F2)

Supongamos que Ker $X = \{0\}$ (i.e. $X^T X$ es invertible y el modelo es identifiable). Tomamos la formulación (F2): $Y = X\beta + \epsilon$. Entonces,

- *El estimador MLE es $\hat{\theta}_n = (\hat{\beta}_n, \hat{s}_n^2)$ donde*

$$\hat{\beta}_n = (X^T X)^{-1} X^T Y \qquad y \qquad \hat{s}_n^2 = \frac{1}{n} \| Y - X\hat{\beta}_n \|_2^2.$$

- *Los dos estimadores $\hat{\beta}_n$ y $\hat{s}_n^2$ son independientes y*

$$\hat{\beta}_n \sim \mathcal{N}(\beta, \sigma^2 (X^T X)^{-1}) \qquad \frac{n}{\sigma^2} \hat{s}_n^2 \sim \chi^2(n-p).$$

3.2.3. Estimadores de m y σ^2

Prueba.

Es una nueva formulación. Vamos a ver que $\Pi_V = X(X^TX)^{-1}X^T$. Por ahora ponemos $A := X(X^TX)^{-1}X^T$. Tenemos

- Calculando el cuadrado

$$A^2 = X(X^TX)^{-1}(X^TX)(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T = A.$$

  Entonces A es un proyector.

- Tenemos también

$$A^T = \left(X(X^TX)^{-1}X^T\right)^T = (X^T)^T((X^TX)^{-1})^TX^T = A,$$

  lo que muestra que el proyector A es ortogonal.

- Vemos directamente que Im $A \subset V$. Pero vemos que por el supuesto que Ker $X = \{0\}$, rg$A = p$ lo que implica que Im $A = V$.

□

3.2.3. Estimadores de m y σ^2

Prueba.

Esos tres hechos prueban que $A = \Pi_V$. Entonces,

$$\Pi_V Y = \hat{m} = X\hat{\beta}_n$$

lo que implica que $\hat{\beta}_n = (X^T X)^{-1} X^T Y$. Como $Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$, tenemos

$$\begin{aligned}\hat{\beta}_n &\sim \mathcal{N}\big((X^T X)^{-1} X^T X\beta \ , \ \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1}\big) \\ &\sim \mathcal{N}\big(\beta, \sigma^2 (X^T X)^{-1}\big).\end{aligned}$$

Las formulas sobre $\hat{s}_n^2$ siendo idénticas al teorema anterior tenemos también el resultado para la varianza. La independencia sucede directamente del teorema (F1). □

3.2.4. Optimalidad

Podemos dar un hecho adicional sobre el estimador $\hat{\beta}_n$.

Proposición

El estimador $\hat{\beta}_n$ tiene la matriz de covarianza minimal adentro de los estimadores lineales sin sesgo de β. Mas precisamente, el orden es : $A \leq B$ si $\forall u$,

$$u^T A u \leq u^T B u.$$

Prueba.

Sea $\tilde{\beta}_n = CY$ un otro estimador sin sesgo y lineal de β. Entonces

$$\mathbb{E}\left[\tilde{\beta}_n\right] = \mathbb{E}\left[CY\right] = CX\beta = \beta.$$

Eso implica que $CX = I_p$. Calculando la varianza

$$\mathrm{Var}\left(\tilde{\beta}_n\right) = \mathrm{Var}\left(CY\right) = C\mathrm{Var}\left(Y\right)C^T = \sigma^2 CC^T.$$

3.2.4. Optimalidad

Prueba.

Entonces,

$$\begin{aligned}
&\text{Var}\left(\tilde{\beta}_n\right) = \sigma^2 CC^T \\
&= \sigma^2(C-(X^TX)^{-1}X^T+(X^TX)^{-1}X^T)(C-(X^TX)^{-1}X^T+(X^TX)^{-1}X^T)^T \\
&= \sigma^2(C-(X^TX)^{-1}X^T)(C-(X^TX)^{-1}X^T)^T + \sigma^2(X^TX)^{-1} \\
&\geq \sigma^2(X^TX)^{-1} \\
&= \text{Var}\left(\hat{\beta}_n\right).
\end{aligned}$$

El producto cruzado vale 0 porque

$$\begin{aligned}
(C-(X^TX)^{-1}X^T)((X^TX)^{-1}X^T)^T & \\
&= CX(X^TX)^{-1} - (X^TX)^{-1}X^TX(X^TX)^{-1} \\
&= (X^TX)^{-1} - (X^TX)^{-1} = 0.
\end{aligned}$$

4.Regresión lineal

4.1.Modelo y estimación

4.1.1. Generalización del MGL

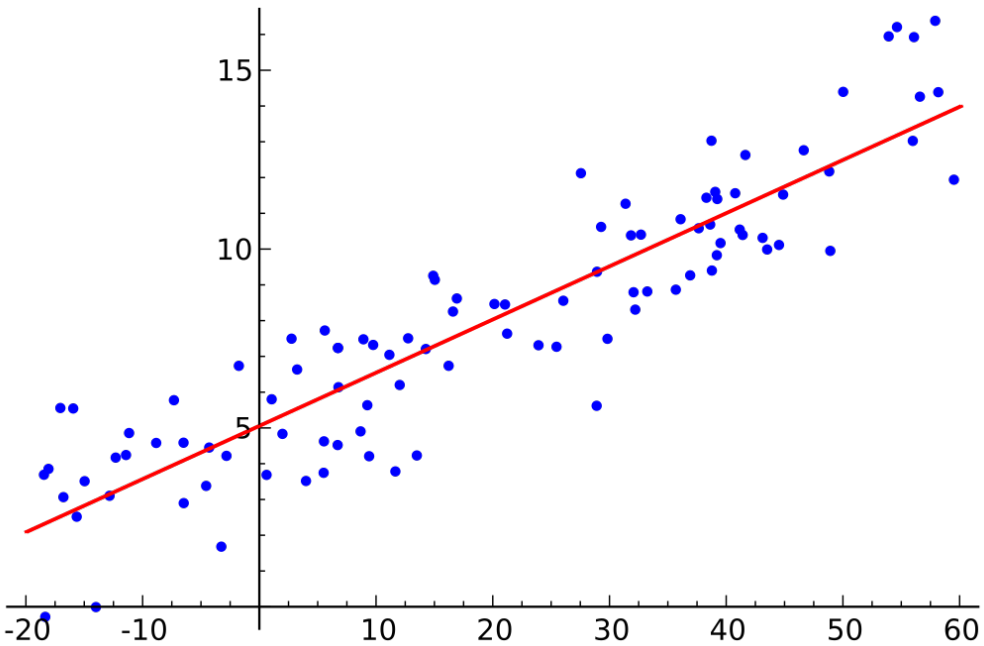
Retomamos la formulación (F2) del modelo gaussiano lineal. Nos interesamos a datos de la forma $(X_1, Y_1), \ldots, (X_n, Y_n)$ donde $X_t \in \mathbb{R}^q$ y $Y_t \in \mathbb{R}$. El modelo de interés es

$$Y_t = \mu + \beta_1 X_{t,1} + \cdots + \beta_q X_{t,q} + \epsilon_t, \qquad \forall t \in \{1, \ldots, n\}$$

Los errores ϵ_t no son gaussianos en general.

Regresión sencilla $q = 1$

En este caso $\forall t, Y_t = \mu + \beta X_t$ y es el caso donde buscamos la recta que pega lo mejor a un conjunto de datos del plano $\mathbb{R}^2$.

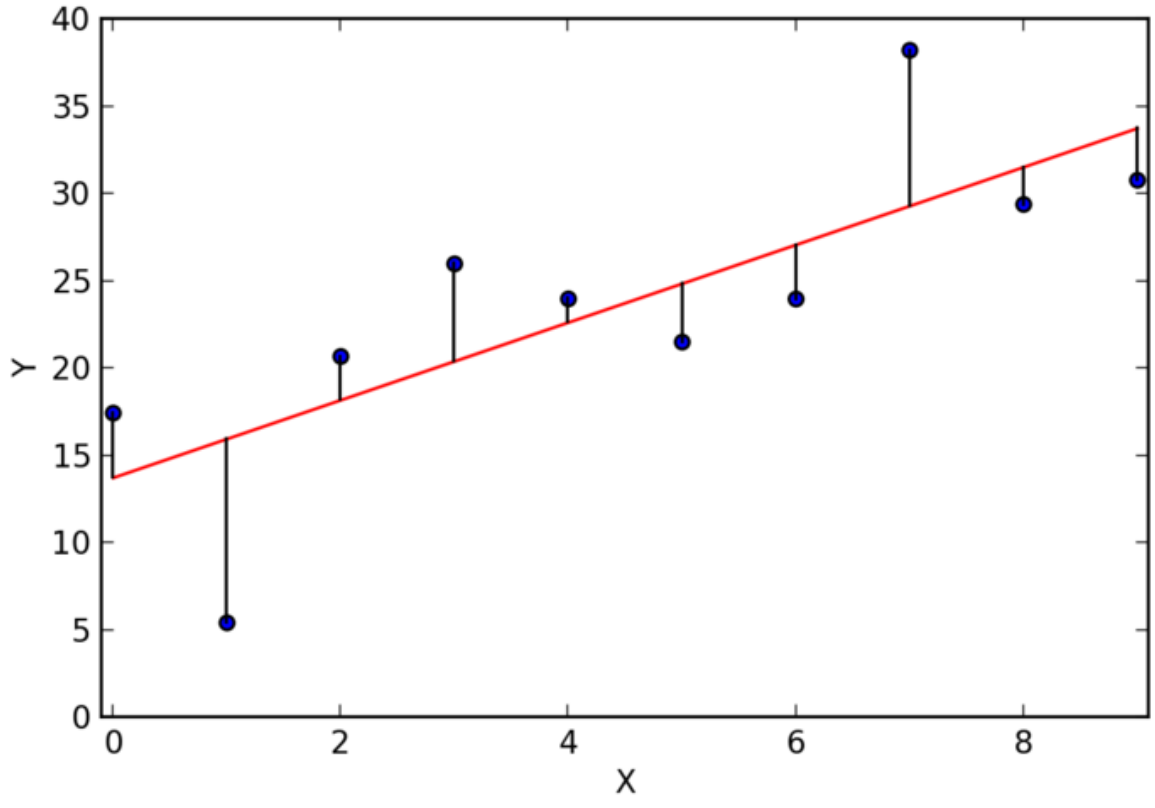


4.1.2. El estimador de mínimos cuadrados

Sean $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ y $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$.

Los estimador de mínimos cuadrados son $\hat{\theta}_n = (\hat{\mu}_n, \hat{\beta}_n)$ con

$$\hat{\beta}_n = \frac{\frac{1}{n}\sum_{i=1}^n (X_t - \overline{X}_n)(Y_t - \overline{Y}_n)}{\frac{1}{n}\sum_{i=1}^n (X_t - \overline{X}_n)^2} \qquad \text{y} \qquad \hat{\mu}_n = \overline{Y}_n - \hat{\beta}_n \overline{X}_n$$



4.1.2. El estimador de mínimos cuadrados

En el caso del modelo gaussiano lineal, podemos dar la distribución a n finito del vector de estimadores $\hat{\theta}_n = (\hat{\mu}_n, \hat{\beta}_n)$:

$$\hat{\theta}_n \sim \mathcal{N}\left(\begin{pmatrix} \mu \\ \beta \end{pmatrix}, \frac{\sigma^2}{\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i^2 & -\overline{X}_n \\ -\overline{X}_n & 1 \end{pmatrix} \right)$$

Hipótesis sobre el modelo

Hay varias maneras clásicas de poner condiciones sobre los términos de error ϵ_t.

- Las $X_{t,i}$ son variables aleatorias tal que las $X_{t,i}$ y ϵ_t son **independientes**. Es util cuando el usuario no tiene control sobre las condiciones del experimento.
- Las $X_{t,i}$ son variables aleatorias tal que $\forall t$,

$$\mathbb{E}\left[\epsilon_t | (X_{t,i})_i\right] = 0 \quad \text{y} \quad \text{Var}\left(\epsilon_t | (X_{t,i})_i\right) = \sigma^2$$

El segundo caso es mas general que el caso 1.

4.1.2. El estimador de mínimos cuadrados

- Las variables ϵ_t dado X es una distribución de vector gaussiano tal que

$$\mathbb{E}[\epsilon|X] = 0 \qquad \text{y} \qquad \text{Var}(\epsilon|X) = \begin{pmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_n^2 \end{pmatrix}$$

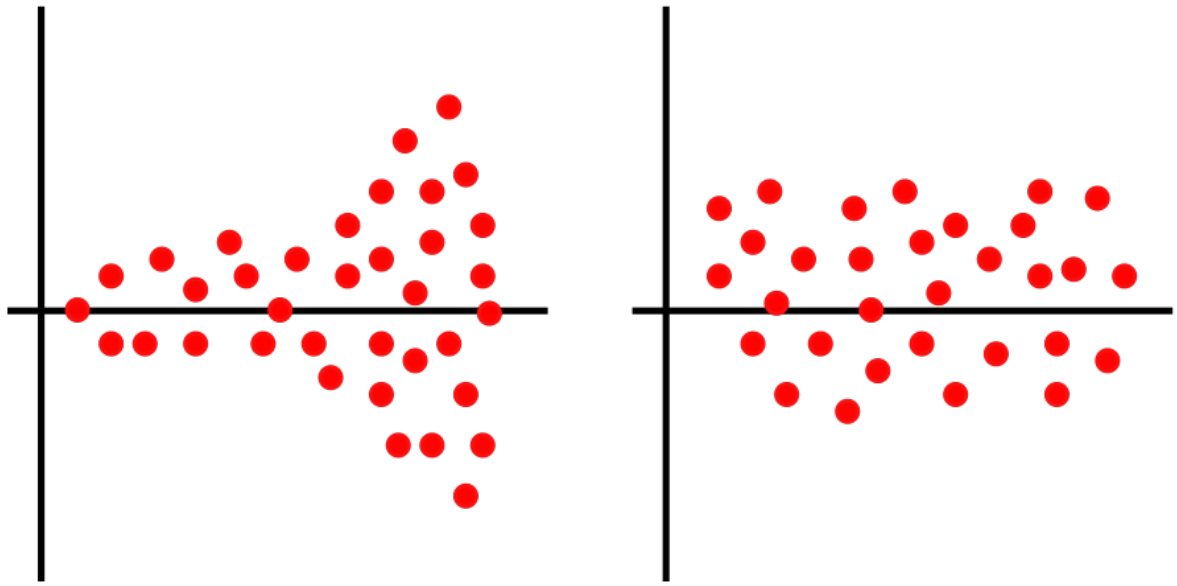
  Este ultimo caso se llama **Heterocedastico**.
  El caso con $\text{Var}(\epsilon|X) = \sigma^2 I_n$ se llama **Homocedastico**.

Para hacer la distinción entre esos dos casos, uno hace lo que se llama el análisis de los residuos.

1. Escribimos la estimación $\hat{\beta}$ de β.
2. Definimos $\forall t \in \{1, \ldots, n\}$, $\hat{\epsilon}_t = Y_t - (X\hat{\beta})_t$.
3. Uno guarda solo los t (en T_0) tal que $|\hat{\epsilon}_t| \geq 2\hat{\sigma}_n$ (por ejemplo)
4. Uno dibuja $((X\hat{\beta})_t, \hat{\epsilon}_t), \forall t \in T_0$.

4.1.2. El estimador de mínimos cuadrados

Observamos una diferencia cualitativa en los dos dibujos.



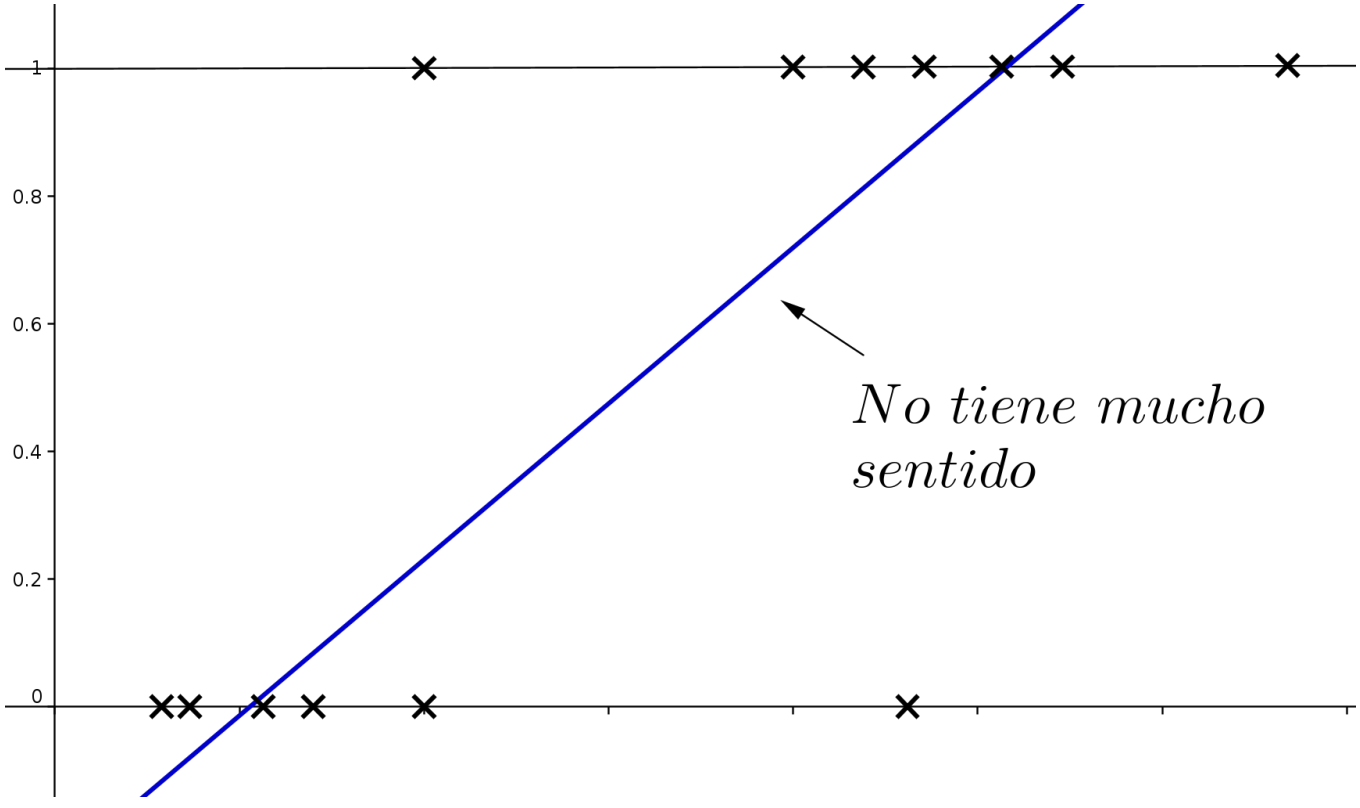
El caso Homocedastico da un dibujo de una nube equilibrada alrededor del eje horizontal. Al contrario, en el caso Heterocedastico, la nube es desformada.

4.Regresión lineal

4.2.Regresión logística

4.2.1. Modelo de la regresión logística

El contexto se presenta cuando uno quiere estudiar el efecto binario de las variables $(X_{t,i})_i$ sobre el resultado Y_t.
$\Rightarrow Y_t \in \{0, 1\}$.



4.2.1. Modelo de la regresión logística

Hipótesis

- Sea $\text{logit}(x) = \log\left(\frac{x}{1-x}\right)$ para $x \in (0,1)$.
- Sea $\text{expit}(x) = \frac{e^x}{1+e^x}$ para $x \in \mathbb{R}$.
- Esas dos funciones son inversas una de la otra.

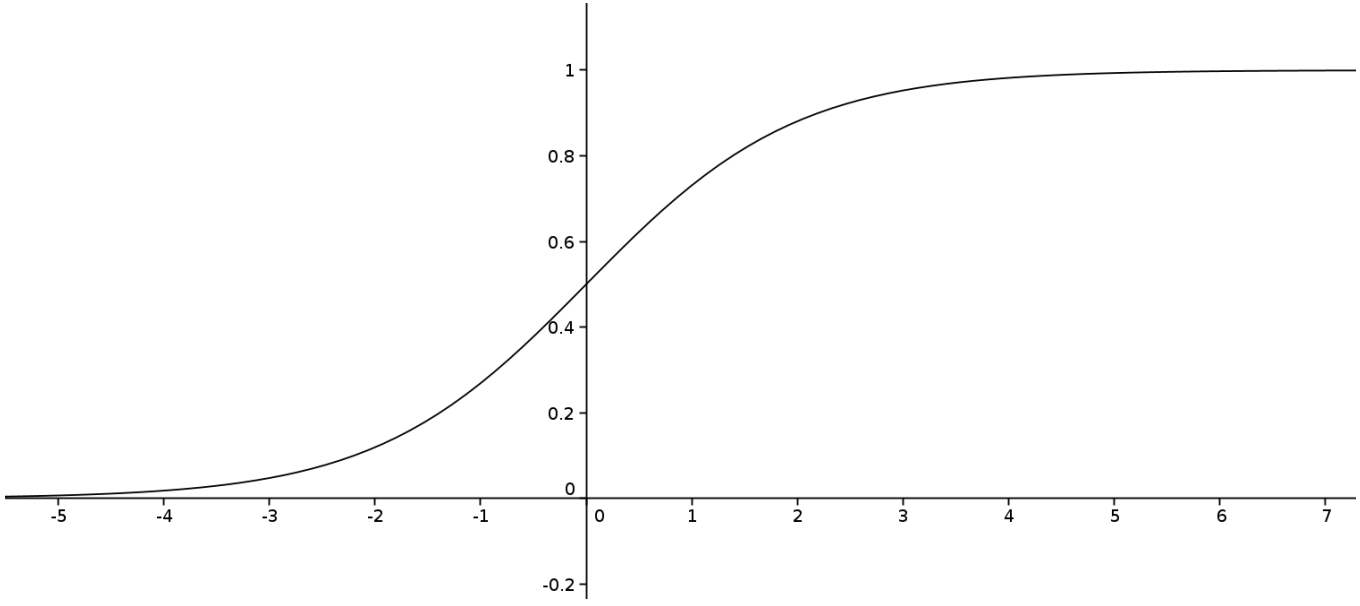


Figure: La función expit

4.2.1. Modelo de la regresión logística

Regresión logística

- Para X fijo, definimos $p = \text{expit}(\beta X + \mu)$ un parámetro en $(0, 1)$. Supongamos que $Y \sim \mathcal{B}(p)$.
  Ese modelo estadístico se llama *modelo de regresión logística* de Y sobre X.
- Cuando uno tiene una muestra $(X_1, Y_1), \ldots, (X_n, Y_n)$, lo anterior se escribe $\forall t$, $Y \sim \mathcal{B}(p_t)$ con $p_t = \text{expit}(\beta X_t + \mu)$.
- Si los X_t son vectoriales, βX_t es un producto interior (como en el caso de la regresión lineal).

Otra formulación:

Una formulación equivalente es de ver que $\forall t$, $Y_t \sim \mathcal{B}(p_t)$ y que

$$\text{logit}(p_t) = \mu + \beta_1 X_{t,1} + \ldots, \beta_q X_{t,q}.$$

La regresión logística es la regresión lineal (teórica) de $\text{logit}(p_t)$ sobre X.

4.2.2. Cálculos del estimador

Construcción

En lo que sigue denotamos $p(1|X) = \mathbb{P}(Y = 1|X)$ y $p(0|X) = \mathbb{P}(Y = 0|X)$. Vimos que el modelo se escribe también

$$\log \frac{p(1|X)}{p(0|X)} = \mu + \beta_1 X_1 + \ldots, \beta_q X_q$$

A la izquierda tenemos una función de X que regresamos linealmente sobre X. En la segunda formulación eso se escribe

$$p(X) = \mathbb{P}(Y|X) = \frac{e^{\mu+\beta_1 X_1+\ldots,\beta_q X_q}}{1 + e^{\mu+\beta_1 X_1+\ldots,\beta_q X_q}}.$$

Estimación

En practica, no tenemos en la muestra los valores de las funciones $p(1|X)$ y $p(0|X)$.

$\Rightarrow$ no podemos usar directamente la técnica de los mínimos cuadrados.

4.2.2. Cálculos del estimador

Verosimilitud de $(Y_t, (X_{t,i})_i)$

- Si $Y_t = 1$, la verosimilitud vale $\mathbb{P}(Y_t = 1 | (X_{t,i})_i) = p_t$.
- Si $Y_t = 0$, la verosimilitud vale $\mathbb{P}(Y_t = 0 | (X_{t,i})_i) = 1 - p_t$.

Entonces la verosimilitud asociada a este modelo esta dada por

$$L(\beta) = \prod_{t=1}^{n} p_t^{Y_t} (1 - p_t)^{1 - Y_t}$$

donde

$$p_t = \frac{e^{\mu + \beta_1 X_{t,1} + \ldots, \beta_q X_{t,q}}}{1 + e^{\mu + \beta_1 X_{t,1} + \ldots, \beta_q X_{t,q}}} \qquad (*)$$

Vemos, sin dificultad, que L es dos veces diferenciable. Entonces, para encontrar el máximo de L, podemos usar las técnicas numéricas de optimización.

$\rightarrow$ Por ejemplo, el algoritmo de Newton.

4.2.2. Cálculos del estimador

Presentamos dos algoritmos (entre muchos!) que dan aproximaciones del estimador de máximo de verosimilitud $\hat{\beta}$.

Algoritmo 1

1. Sea β^0 una inicialización (ej. $\beta^0 = (0, \ldots, 0)$).
2. Iterativamente calculamos

$$\beta^{i+1} = \beta^i - \left(\frac{\partial^2 L}{\partial\beta_i \partial\beta_j}\right)^{-1}_{i,j} \times \left(\frac{\partial L}{\partial\beta_i}\right)_i .$$

   donde $\left(\frac{\partial^2 L}{\partial\beta_i \partial\beta_j}\right)_{i,j}$ es la matriz hessiana de L y $\left(\frac{\partial L}{\partial\beta_i}\right)_i$ el gradiente de L.
3. Stop cuando $|\beta^{i+1} - \beta^i|$ es mas pequeño que un cierto valor umbral.

La etapa 2. sigue las ideas del famoso algoritmo de Newton para encontrar el mínimo de una función.

4.2.2. Cálculos del estimador

Algoritmo 2

1. Sea β^0 una inicialización.
2. Calcular p_t^0 usando (*)
3. Iterativamente :
   - 3.1 Sea $Z_t^i = \text{logit}(p_t^i) + \frac{Y_t - p_t^i}{p_t^i(1-p_t^i)}$, para cada $t \in \{1, \ldots, n\}$.
   - 3.2 Sea
   $$W^i = \begin{pmatrix} p_1^i(1-p_1^i) & & \\ & \ddots & \\ & & p_n^i(1-p_n^i) \end{pmatrix}$$
   - 3.3 Calculamos $\beta^{i+1} = (X^T W^i X)^{-1} X^T W Z^i$.
   - 3.4 Calculamos p_t^{i+1} usando (*).
4. Stop cuando $|\beta^{i+1} - \beta^i|$ es mas pequeño que un cierto valor umbral.

La etapa 3.3 es una regresión lineal (caso Heterocedastico con pesos dados por W^i) de Z^i sobre X.

4.Regresión lineal

4.3.Validación del modelo

4.3.1. Prueba de Student de una relación afín

Validación del modelo

Retomamos la relación (F2): $Y = X\beta + \epsilon$ y $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$.
En la validación del modelo lineal, preguntamos si hay una relación afín $c^T\beta = a$ sobre el vector β. Consideramos c y a conocidos en lo que sigue.

$$H_0 : c^T\beta = a$$
$$H_1 : c^T\beta \neq a$$

Para distinguir entre esas dos hipótesis usamos la estadística

$$h_n = \frac{c^T\beta - a}{\hat{\sigma}_n \sqrt{c^T(X^TX)^{-1}c}}$$

donde $\hat{\beta}$ y $\hat{\sigma}_n$ son los estimadores de β y σ. Sabemos que

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X^TX)^{-1})$$
$$\frac{\hat{\sigma}_n^2(n-p)}{\sigma^2} \sim \chi^2(n-p).$$

4.3.1. Prueba de Student de una relación afín

Eso muestra que h_n tiene una distribución de Student $\tau(n-p)$. Sea $t_{n-p,1-\frac{\alpha}{2}}$ el cuantile de orden $1-\frac{\alpha}{2}$ de $\tau(n-p)$. Sea

$$T(X,Y) = \mathbb{1}_{|h_n| > t_{n-p,1-\frac{\alpha}{2}}}$$

Cuando h_n es "grande" es menos probable de tener $c^T\beta = a$ y entonces rechazamos H_0.

Proposición

La prueba de hipótesis T es de nivel α.

Prueba.

Si $c^T\beta = a$,

$$\mathbb{P}_\beta\left(T(X,Y)=1\right) = 2\mathbb{P}_\beta\left(h_n > t_{n-p,1-\frac{\alpha}{2}}\right) = 2\frac{\alpha}{2} = \alpha. \qquad \square$$

Comentario: Un caso particular consiste en probar la influencia de un factor X_i sobre el valor de Y. Por ejemplo,

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0$$

4.3.2. Prueba de Fisher

En la formulación (F1): $Y = m + \epsilon$ y $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$.
La prueba de Fisher se interesa a la pregunta de si m pertenece a un sub-espacio lineal particular $W \subset V$ tal que $\dim(W) = q$.

$$H_0 : m \in W$$
$$H_1 : m \in V \backslash W$$

Distribución de Fisher

Sean $U \sim \chi^2(k)$ y $V \sim \chi^2(\ell)$ independientes. Digamos que $Z = \frac{U/k}{V/\ell}$ tiene la *distribución de Fisher* $\mathcal{F}(k, \ell)$ con grados de libertad k y ℓ.

Podemos retomar la formulación vectorial y notar que

$$\mathbb{R}^n = (W \oplus W^{\perp V}) \oplus V^{\perp}$$

donde $W^{\perp V}$ es el ortogonal de W adentro de V. En termino de proyectores

$$I_n = \Pi_W + (\Pi_V - \Pi_W) + (I_n - \Pi_V)$$

4.3.2. Prueba de Fisher

Entonces tenemos el resultado

Teorema

Si $m \in W$ entonces

$$F = \frac{(\|\Pi_W Y - \Pi_V Y\|^2)/(p-q)}{(\|Y - \Pi_V Y\|^2)/(n-p)} \sim \mathcal{F}(p-q, n-p)$$

y F es independiente de $\Pi_W Y$.

Prueba.

Es un nuevo uso de Cochran! (Ver que $m \in W$ es necesario para tener $\Pi_W m = \Pi_V m$) □

Cuando $m \notin W$, F tendrá valores mas altas. Podemos definir

$$T(X, Y) = \mathbb{1}_{F > f_{p-q,n-p,1-\alpha}}$$

donde $f_{p-q,n-p,1-\alpha}$ es el cuantile de orden $1-\alpha$ de $\mathcal{F}(p-q, n-p)$.

4.3.2. Prueba de Fisher

Comentario: Cuando $q = p - 1$ el sub-espacio W es un hiperplano de V.
Entonces, existe $d \in V$ tal que $m \in W \Leftrightarrow d^T m = 0 \Leftrightarrow c^T \beta = 0$ con $c = X^T d$.
La prueba de hipótesis de Fisher es entonces equivalente a la prueba de Student bilateral del parágrafo anterior. En efecto, si una variable $Z \sim \tau(n - p)$ entonces $Z^2 \sim \mathcal{F}(1, n - p)$.

Probar una sola relación afín se puede hacer igualmente por la prueba de Student o de Fisher.

Ejemplo: Si uno quiere comprobar simultáneamente $\beta_2 = 2\beta_1$ y $\beta_3 = 0$, definimos

$$C = \begin{pmatrix} -2 & 1 & 0 & 0 & \dots \\ 0 & 0 & 1 & 0 & \dots \end{pmatrix}$$

y probar las relaciones lineales es exactamente probar $C\beta = 0$ o $\beta \in \text{Ker}(C) = W$.

4.3.3. Prueba de Wald de varias relaciones afín

Este prueba de hipótesis se usa cuando uno quiere comprobar varias relaciones afín a la vez. Vimos que se puede resumir en

$$H_0 : C\beta = a$$
$$H_1 : C\beta \neq a$$

donde se supone que C es de rango máximo ($\Leftrightarrow$ las relaciones afín son independientes). Denotamos $k \leq p$ el número de relaciones afín que se consideran. $\mathrm{rg}(C) = k = \dim(W)$. Entonces, la matriz $C(X^TX)^{-1}C^T$ es simétrica y definida positiva. En particular es invertible y podemos considerar la estadística

$$W = \frac{\left((C\hat{\beta}_n - a)^T(C(X^TX)^{-1}C^T)^{-1}(C\hat{\beta}_n - a)\right) \Big/ k}{\|Y - X\hat{\beta}_n\|_2^2/(n-p)}$$

Proposición

Tenemos $W \sim \mathcal{F}(k, n-p)$.

4.3.3. Prueba de Wald de varias relaciones afín

Prueba.

Denotamos Δ la raíz cuadrada de la matriz $C(X^TX)^{-1}C^T$.
Para todo β, sabemos que la transformada lineal $C\hat{\beta}_n - a \sim \mathcal{N}(C\beta - a, \sigma^2 C(X^TX)^{-1}C^T)$. Y en particular, bajo la hipótesis H_0,

$$\Delta^{-1}(C\hat{\beta}_n - a) \sim \mathcal{N}(0, \sigma^2 I_k).$$

Así, bajo H_0,

$$\begin{aligned}\frac{1}{\sigma^2}(C\hat{\beta}_n - a)^T(C(X^TX)^{-1}C^T)^{-1}(C\hat{\beta}_n - a) \\ &= \frac{1}{\sigma^2}(C\hat{\beta}_n - a)^T\Delta^{-2}(C\hat{\beta}_n - a) \\ &= \frac{1}{\sigma^2}\|\Delta^{-1}(C\hat{\beta}_n - a)\|_2^2 \sim \chi^2(k)\end{aligned}$$

Se termina la prueba notando que $\|Y - X\hat{\beta}_n\|_2^2 \sim \chi^2(n-p)$ y que por el teorema de Cochran, $\hat{\beta}_n$ y $Y - X\hat{\beta}_n$ son independientes. □

4.3.3. Prueba de Wald de varias relaciones afín

Prueba de Wald

Caracterizamos la distribución de W en el caso de la hipótesis H_0. Bajo H_1, W tiene la tendencia de tomar valores mas grandes. Entonces, podemos definir

$$T(X, Y) = \mathbb{1}_{W > f_{k,n-p,1-\alpha}}$$

donde $f_{k,n-p,1-\alpha}$ es el cuantile de orden $1-\alpha$ de la distribución de Fisher de k y $n-p$ grados de libertad.

Entonces tenemos el

Proposición

La prueba de hipótesis T es de nivel α.

Un resultado anexo que obtenemos es que el conjunto

$$\varepsilon = \left\{ a \in \mathbb{R}^k : W \leq f_{k,n-p,1-\alpha} \right\}$$

es una elipsoide de confianza de nivel (exacto) igual a $1-\alpha$ de $C\beta$.

4.Regresión lineal

4.4.Predicción

4.4.1. Error de predicción

Paradigma

En el contexto de la regresión sencilla,

$$Y_t = \beta_0 + \beta_1 X_t + \epsilon_t \qquad \text{con} \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2).$$

Después de la regresión es muy común de querer dar un valor de Y para una nueva variable explicativa X^*.

El valor real que buscamos es

$$Y^* = \beta_0 + \beta_1 X^*.$$

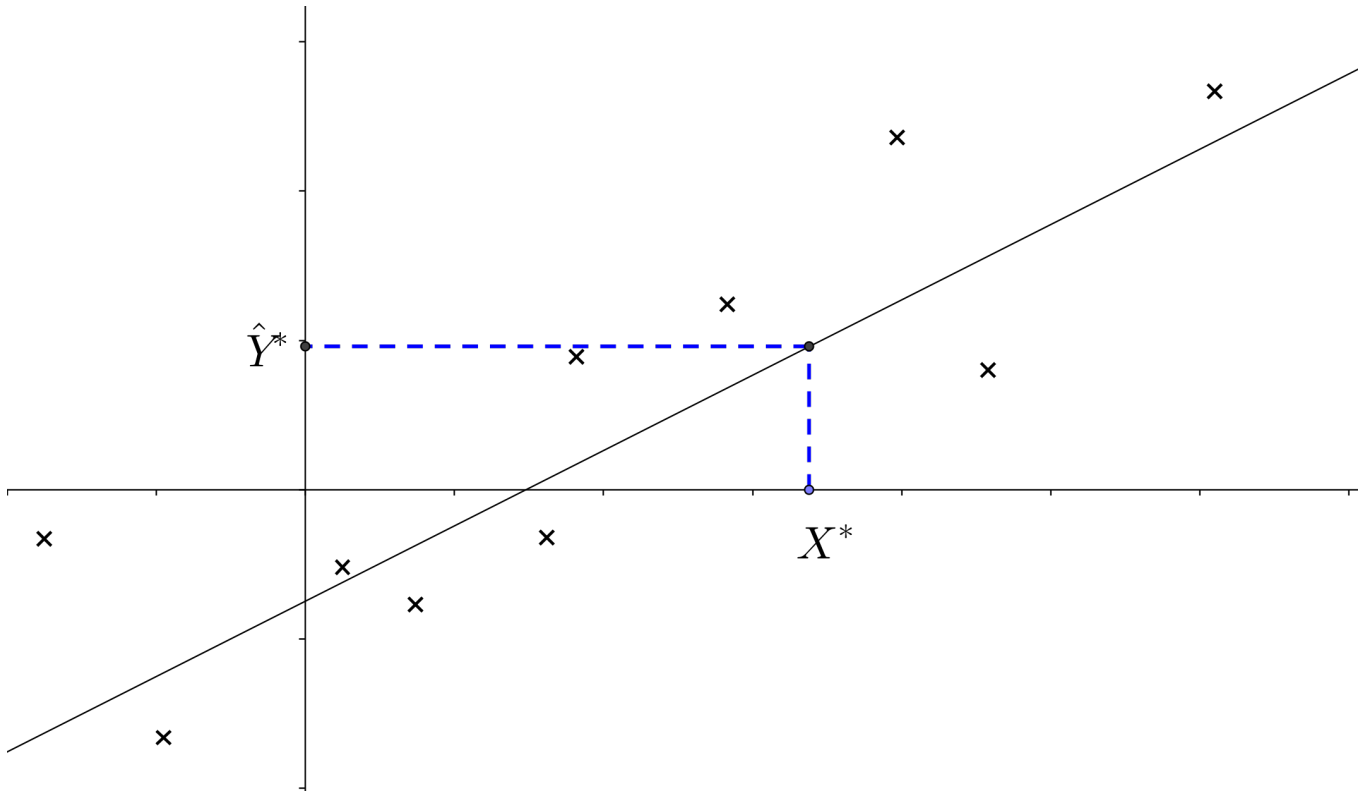
Denotamos $\hat{Y}^*$ un estimador de este valor. La notación toma en cuenta que $\hat{Y}^*$ es una variable aleatoria que depende de toda la clase $(X_1, Y_1), \ldots, (X_n, Y_n)$.

La variable aleatoria $(X^*, \hat{Y}^*)$ se llama *predicción* de la variable (X^*, Y^*).

4.4.1. Error de predicción

Denotamos

$$\theta = \beta_0 + \beta_1 X^*$$
$$\hat{\theta} = \hat{\beta}_0 + \hat{\beta}_1 X^*$$

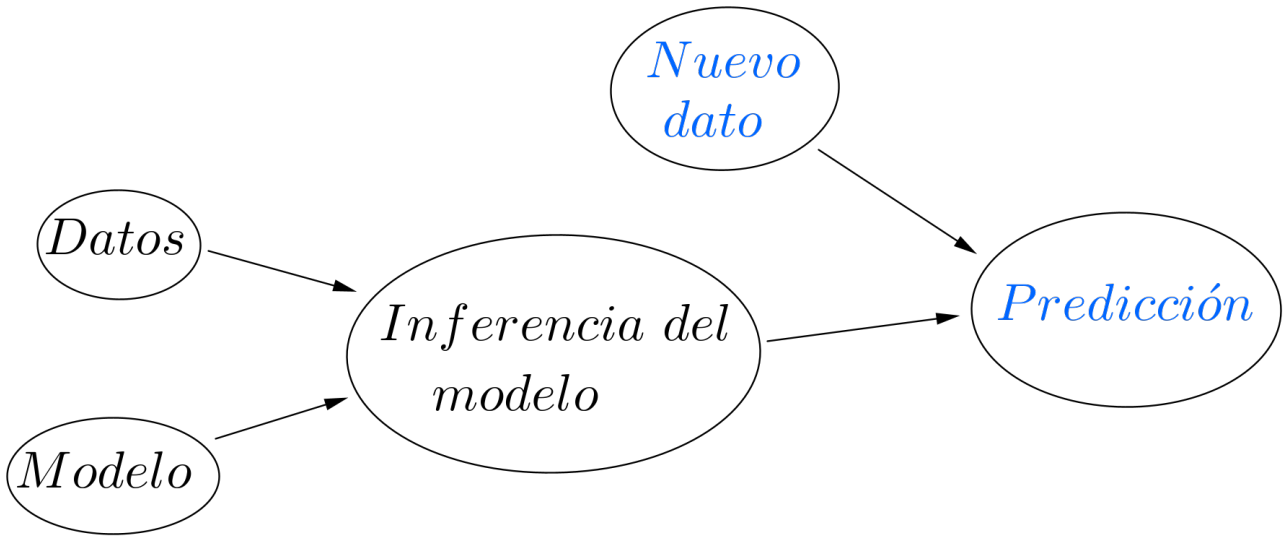


Definimos $\hat{Y}^* = \hat{\theta}$ para predecir $Y^* = \theta + \epsilon^*$.

4.4.1. Error de predicción

Se sabe que

- $\hat{Y}^* = \hat{\theta} \sim \mathcal{N}\left(\theta, \text{Var}\left(\hat{\theta}\right)\right)$
- $Y^* \sim \mathcal{N}(\theta, \sigma^2)$.
- Estamos estimando una variable aleatoria Y^* gracias a una otra variable aleatoria $\hat{Y}^*$.



4.4.1. Error de predicción

Error de predicción

Para dar un intervalo de confianza de Y^*, uno estudia

$$\hat{Y}^* - Y^* = \underbrace{\hat{\theta} - \theta}_{\mathcal{N}(0,\mathrm{Var}(\hat{\theta}))} - \underbrace{\epsilon^*}_{\mathcal{N}(0,\sigma^2)}.$$

Entonces, $\hat{Y}^* - Y^* \sim \mathcal{N}(0, \mathrm{Var}(\hat{\theta}) + \sigma^2)$. Esta cantidad se llama *error de predicción* de Y^*.

Calculo de Var$(\hat{Y}^* - Y^*)$

Por lo que probamos antes

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \frac{\sigma^2}{\sum_{t=1}^n (X_t - \bar{X}_n)^2} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_t^2 & -\bar{X}_n \\ -\bar{X}_n & 1 \end{pmatrix} \right)$$

4.4.2. Intervalos de predicción

Entonces

$$\begin{aligned}\text{Var}\left(\hat{\theta}\right) &= \text{Var}\left(\hat{\beta}_0\right) + \text{Var}\left(\hat{\beta}_1\right) X^{*2} + 2\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) X^* \\ &= \frac{\sigma^2}{\sum_{t=1}^{n}(X_t - \bar{X}_n)^2}\left[\frac{1}{n}\sum_{i=1}^{n} X_t^2 + X^{*2} + 2(-\bar{X}_n)X^*\right] \\ &= \frac{\sigma^2 \sum_{t=1}^{n}(X_t - X^*)^2}{n\sum_{t=1}^{n}(X_t - \bar{X}_n)^2}.\end{aligned}$$

Por lo tanto,

$$\text{Var}\left(\hat{Y}^* - Y^*\right) = \sigma^2\left(\frac{\sum_{t=1}^{n}(X_t - X^*)^2}{n\sum_{t=1}^{n}(X_t - \bar{X}_n)^2} + 1\right) =: \sigma^2 a_n(X^*)$$

4.4.2. Intervalos de predicción

Distribución de $\hat{Y}^* - Y^*$

Los teoremas anteriores aseguran que

$$\frac{\hat{s}_n^2}{\sigma^2}(n-2) \sim \chi^2(n-2).$$

Entonces, por definición de la distribución de Student,

$$\frac{\hat{Y}^* - Y^*}{\hat{s}_n \sqrt{a_n(X^*)}} \sim \tau(n-2).$$

Intervalo de predicción

Un intervalo de confianza para el valor de Y^* es

$$I_\alpha = [\hat{Y}^* - t_{n-2,1-\frac{\alpha}{2}} \hat{s}_n \sqrt{a_n(X^*)}, \hat{Y}^* + t_{n-2,1-\frac{\alpha}{2}} \hat{s}_n \sqrt{a_n(X^*)}]$$

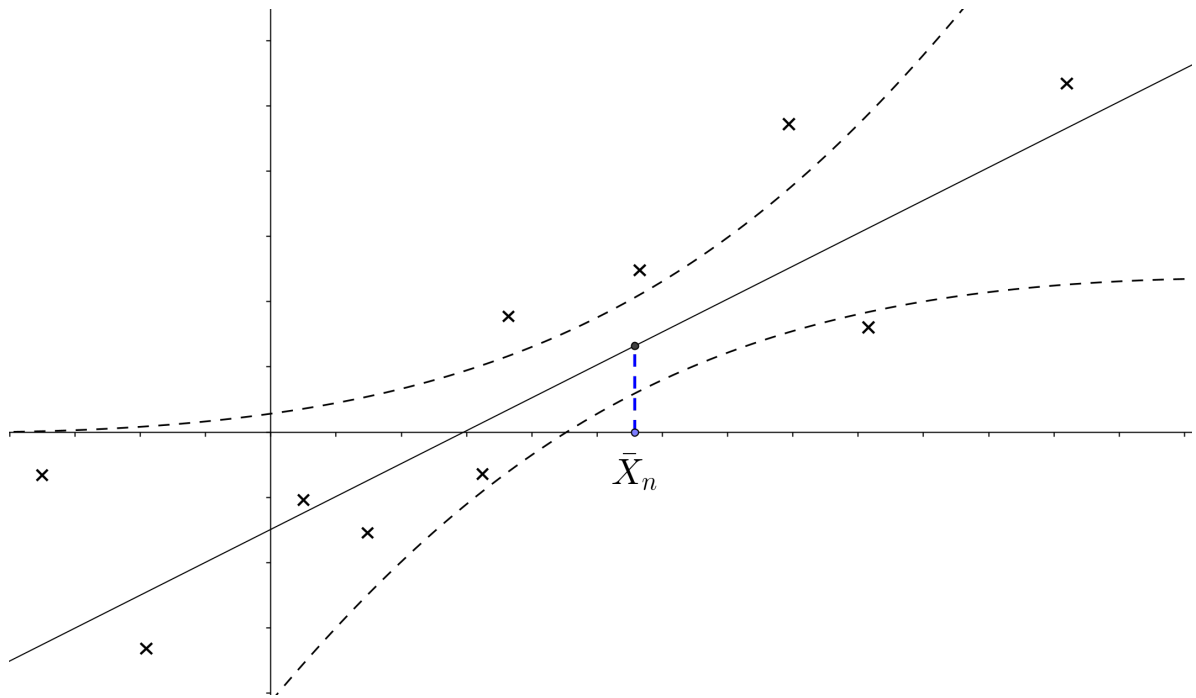
donde $t_{n-2,1-\frac{\alpha}{2}}$ es el cuantile de orden $1-\alpha/2$ de $\tau(n-2)$.

4.4.2. Intervalos de predicción

Interpretación de $a_n(X^*)$

Podemos descomponer el coeficiente $a_n(X^*)$ en

$$a_n(X^*) = \underbrace{1}_{\text{Aleatorio de } Y^*} + \underbrace{\frac{\sum_{t=1}^{n}(X_t - X^*)^2}{n\sum_{t=1}^{n}(X_t - \bar{X}_n)^2}}_{\text{Incertidumbre de la predicción}}$$



El punto de predicción con la mejor precisión es para $X^* = \bar{X}_n$.

4.4.2. Intervalos de predicción

Comentarios:

- Este estudio se puede generalizar en dimensiones mas alta ($\dim(\beta) = 2 \to \dim(\beta) = p$)
- De nuevo el punto de mejor predicción se ubica en $\bar{X}_n$, el centro de masa.
- Todos los cálculos son explícitos, por ejemplo

$$\text{Var}\left(\hat{\theta}\right) = \text{Var}\left(X^*\beta\right) = \sigma^2 X^*(X^T X)^{-1} X^{*T}$$

- La incertidumbre de la predicción dependerá de los valores propios de $(X^T X)^{-1}$. Por esa razón, ese matriz se llama *matriz de precisión* del problema de regresión.
- Para evitar problemas de intervalos de confianza de tamaño $\to \infty$, podemos usar una estimación local.

$$\to \text{Estimadores kernel}$$

Gracias por su atención

Suerte para el examen final.