Facultad de Ciencias Físicas y Matemáticas
Departamento de Ingeniería Eléctrica
EL3204 Análisis de Sistemas Dinámicos y Estimación
2025

Detección y Estimación
Version 2.5

Sebastián Espinosa[1] y Jorge F. Silva[1]

[1] *PhD. Electrical Engineering. Sebastián and Jorge are members of the Information and Decision Systems Group, Universidad de Chile.*

Resumen

Este apunte tiene como propósito introducir al lector en los fundamentos teóricos y principios esenciales de la teoría de detección y estimación, áreas clave en la toma de decisiones basada en información. A lo largo del documento, se desarrollará una formalización matemática rigurosa, complementada con resultados fundamentales que sustentan muchas aplicaciones modernas, desde las telecomunicaciones y el procesamiento de señales hasta la inteligencia artificial y la astrofísica.

Contenidos

Introducción

La teoría de detección y estimación en procesamiento de señales es un campo que reúne matemáticos e ingenieros eléctricos para concebir una teoría robusta respecto a las observaciones o señales recibidas en sistemas físicos. Tanto la teoría de detección como la de estimación usan una gran cantidad de herramientas estadísticas y, por lo tanto, es una aplicación directa de la teoría de probabilidades.

Un sistema, desde su concepción más general, es un ente que entrega información (variables observables) y, a partir de esta información, surge una pregunta natural que es cómo hacer buen uso de esta información. Así, por ejemplo, al monitorear procesos es posible identificar momentos en los que el proceso se aleja del punto de operación deseado, lo que permite *detectar* anomalías en el sistema. Por lo tanto, la detección consiste en poder reconocer situaciones específicas, dentro de un conjunto de opciones. Luego, lo ideal es poder calcular el valor de los parámetros y el vector de estados asociados al modelo del sistema. Debido a limitaciones del sistema, estos valores no podrán obtenerse de manera exacta y, por lo tanto, deben ser *estimados*. Por lo que la estimación corresponde a un proceso de decisión donde se entiende que los valores serán aproximaciones del valor real.

La estadística, si se resume en pocas palabras, corresponde a una disciplina cuya filosofía se basa en *la parte por el todo*, es decir, a partir de una muestra (conjunto

de observaciones)[1] el objetivo es poder inferir algún valor de interés aceptando la posibilidad de equivocarse. Es por esta misma razón que la estadística y la teoría de probabilidad presentan esta similitud debido que esta última provee un esquema robusto para caracterizar incertidumbre, la que proviene del hecho que tanto el modelo propuesto posee incertidumbre intrínseca (variables ocultas, perturbaciones, errores sistemáticos) así como también la cantidad de muestras pueden no ser lo suficientemente representativas. Pero se debe enfatizar que esta teoría va mucho más allá de la estadística como tal y, en particular, el enfoque de este apunte es desde un punto de vista más conceptual.

En este apunte explicaremos los procedimientos necesarios para que, a partir de un conjunto de observaciones, podamos tomar una decisión a partir de un criterio[2]. La teoría de probabilidad juega un rol importante en este caso, debido a que las observaciones o mediciones corresponden a fuentes de información que fueron adquiridas mediante sensores, lo que indica que estas observaciones estarán sujetas a incertidumbre o ruido. Esta naturaleza estocástica entonces permite caracterizar una fuente de información (observación) como una variable aleatoria X y gracias a este modelamiento nos permitirá utilizar la axiomática y herramientas vistas en probabilidades.

El siguiente ejemplo ilustra la idea de cómo la teoría de detección ayuda a formalizar la idea de toma de decisiones en un contexto aplicado.

Ejemplo 0.1. La estudiante PAT tiene una moneda de 100 pesos. Mientras lanzaba la moneda varias veces se le ocurrió la idea de modelar el experimento asociado al lanzamiento de la moneda y ver qué sale. Rápidamente se da cuenta que el modelo propuesto es simple de caracterizar mediante variables aleatorias (observaciones y sus resultados) y su espacio de probabilidad inducido:

- $R_X = \{0, 1\}$
- $\mathcal{F} = \mathcal{P}(R_X)$
- $P_X(X = 1) = P_X(X = 0) = \frac{1}{2}$

Este modelo representa fielmente algo tan simple como lanzar una moneda. El supuesto de que $P_X(X = 1) = P_X(X = 0) = \frac{1}{2}$ representa la equiprobabilidad en los resultados y resulta natural plantearlo de esta manera porque se asume que la moneda tiene un

[1] En estadística como tal se llama muestra poblacional
[2] Para ser aún más precisos, esta teoría cabe en una categoría más general que se llama Teoría de Juegos

comportamiento esperado. Ahora bien, PAT lanza la moneda 30 veces y se da cuenta que obtuvo 28 caras. Según la teoría, es completamente probable obtener este resultado, pero tal resultado es muy bajo, lo que abre las puertas para preguntarse si es que la moneda realmente es equilibrada.

PAT rápidamente entonces se plantea la posibilidad de que el modelo cambió, es decir, *decidir* si existió alguna perturbación en el modelo original o simplemente ver si en realidad el modelo original planteado estaba malo. El proceso anterior en donde hubo, probablemente, un cambio de modelo, se conoce como *detección*.

PAT decide finalmente que el modelo original no era el correcto y le gustaría proponer otro. Dado que el experimento sigue siendo lanzar una moneda y ver qué sale, se da cuenta que el cambio natural que se le debe hacer al modelo original es el valor de la probabilidad de obtener cara. Pero, ¿qué valor se le puede dar? para responder a esta interrogante, una forma sería recolectar una gran cantidad de datos (observaciones), es decir, muchos lanzamientos de la moneda y, mediante alguna técnica estadística, *estimar* el valor de la probabilidad p de obtener cara (formalmente sería estimar el valor de $p = P_X(X = 1)$).

Naturalmente, repitiendo lo dicho anteriormente, tanto la decisión de PAT de cambiar el modelo así como la estimación de p estarán sujetas a errores lo que abre la posibilidad de equivocarse en la decisión o estimación. Cabe preguntarse entonces si es posible no equivocarse o, al menos, equivocarse lo menos posible. Todo esto será abordado con mayor detalle en las unidades que serán vistas a continuación.

0.1. Juego Estadístico

En un nivel más formal, la teoría de detección y estimación puede entenderse como un caso particular dentro del marco general de los juegos estadísticos, donde la idea fundamental consiste en tomar una decisión que minimice el costo esperado asociado a las posibles consecuencias.

Más precisamente, un juego estadístico se define formalmente como la interacción entre dos "jugadores": la Naturaleza, que selecciona un parámetro desconocido, y el Estadístico, que elige una regla de decisión con el objetivo de minimizar la pérdida inducida por dicha elección. A continuación daremos su definición formal

---

Definición 0.1. (Juego Estadístico) Un juego estadístico es una tripleta (Θ, Δ, L) dotada de un espacio muestral $\mathbb{X}$ y una variable aleatoria X. Cada uno de estos elementos cumple un rol bien definido dentro del proceso de toma de decisiones bajo incertidumbre:

- Un espacio de parámetros $\Theta \subset \mathbb{R}^k$, $k \in \mathbb{N}$. Corresponde a un conjunto de posibles estados de la naturaleza.
- Un espacio de decisión Δ que corresponde a todas las posibles decisiones disponibles para el agente que toma la decisión.
- Un espacio de observación $\mathbb{X}$, que puede o no ser numérico. Para este apunte asumiremos que el espacio de observación es numérico y, además, lo dotaremos de una variable aleatoria $X : \Omega \to \mathbb{X}$, recordando que Ω es el espacio muestral o dominio abstracto del experimento. Luego, en términos simples $X \in \mathbb{X}$ corresponde a una variable de observación.[3] Además, dado un $\theta \in \Theta$ se inducirá una distribución de probabilidad $P_X(\cdot|\Theta = \theta)$ que caracterizará el modelo del sistema.
- Una regla de decisión $r : \mathbb{X} \to \Delta$, o alternativamente llamada estrategia o función de decisión que provee una unión entre las observaciones (y por ende el estado de la naturaleza por medio de $P_X(\cdot|\Theta = \theta)$) y las decisiones.
- Una función de costo $L : \Theta \times \Delta \to \mathbb{R}$ que favorece o penaliza la decisión tomada respecto a lo observado.

Una idea ilustrativa de la teoría de estimación y detección se presenta en la Figura 1

Observaciones 0.1.

- La tripleta (Θ, Δ, L) es una representación general de un juego estadístico. En términos esenciales, la idea es a partir de un conjunto de observaciones ($\mathbb{X}$) tomar una decisión (Δ) que me entregue la menor pérdida posible (L).
- En gran parte de este apunte, y sin tanta pérdida de generalidad, asumiremos que tanto el espacio de decisión como el de parámetros son idénticos ($\Theta = \Delta$),

[3] En teoría estadística se admite que el espacio de observaciones $\mathbb{X}$ no sea numérico, ya que en muchos casos las observaciones pueden no ser números reales, como el resultado del lanzamiento de una moneda, o bien representar vectores, señales, imágenes, trayectorias o funciones continuas. En tales situaciones no se habla estrictamente de variables aleatorias —que por definición son funciones medibles con valores numéricos—, sino de funciones medibles o variables aleatorias generalizadas, es decir, funciones aleatorias con valores en un espacio medible más general.

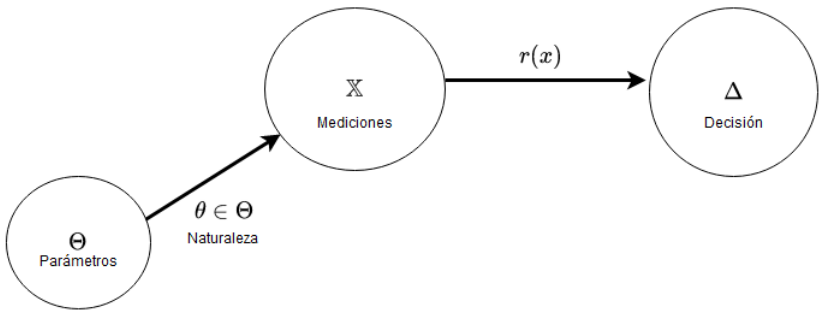


Figura 1: Idea general del problema de toma de decisiones bajo incertidumbre.

en cuyo caso los problemas de decisión y estimación pasan a llamarse detección paramétrica y estimación paramétrica respectivamente. Si el parámetro en cambio tiene una probabilidad asociada (es un objeto aleatorio) en este caso los problemas pasan a llamarse detección Bayesiana y estimación Bayesiana. Gráficamente la teoría de estimación paramétrica y detección paramétrica se presenta en la Figura 2

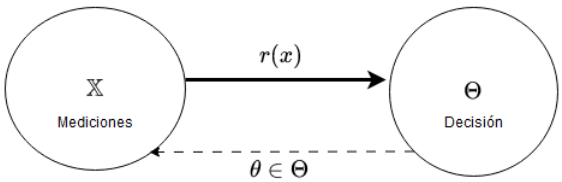


Figura 2: Idea general del problema de toma de decisiones.

Con estas hipótesis y dependiendo de la naturaleza de Θ podemos clasificar el juego estadístico como sigue:

- Si Θ es finito o numerable y además es determinístico: Detección Paramétrica.

- Si Θ es finito o numerable y además es aleatorio: Detección Bayesiana.
- Si Θ es no numerable y además es determinístico: Estimación Paramétrica.
- Si Θ es no numerable y además es aleatorio: Estimación Bayesiana.

Las unidades que serán vistas en adelante abordarán estas teorías, cada una con sus diferencias y/o similitudes con respecto al resto.

0.1.1. Análisis del Juego Estadístico

Una vez definido el juego estadístico, el siguiente paso consiste en **analizar el juego**, es decir, evaluar y comparar las posibles reglas de decisión. El objetivo de este análisis es determinar qué regla $r : \mathbb{X} \to \Delta$ resulta "mejor" frente a la incertidumbre introducida por la naturaleza (representada por el parámetro θ) y la aleatoriedad de la observación X.

Dado que las observaciones son aleatorias, no tiene sentido evaluar la pérdida punto a punto (es decir, solo para un valor específico de x), pues el estadístico no controla la realización del experimento. Lo que se busca es medir el desempeño **en promedio** bajo la distribución P_θ asociada al verdadero estado θ.

Por ello, se introduce la **función de riesgo**, definida como el valor esperado de la pérdida:

$$R(\theta, r) = \mathbb{E}(L(\theta, r(X))) \tag{1}$$

Esta expectativa representa el costo esperado de utilizar la regla r cuando el parámetro verdadero es θ. El uso del valor esperado no es un artificio técnico, sino una consecuencia directa de la aleatoriedad inherente al experimento: el riesgo refleja la calidad promedio de la regla de decisión antes de observar los datos.

Es importante notar que, aunque las reglas de decisión se aplican observación a observación, la evaluación de su calidad no puede realizarse "punto a punto". Si se intentara minimizar la pérdida $L(\theta, r(x))$ para cada x, ello implicaría conocer de antemano la observación (y en muchos casos incluso el propio θ), lo cual contradice el carácter incierto del experimento. En cambio, el riesgo $R(\theta, r)$ se define ex ante, es decir, antes de conocer el valor de X, promediando sobre su distribución bajo P_θ.

De este modo, el riesgo cuantifica el desempeño promedio del procedimiento estadístico y permite comparar reglas de decisión en términos de su eficacia general,

independientemente de la realización particular de los datos. En otras palabras, la pérdida mide la calidad de una decisión, mientras que el riesgo mide la calidad de una *regla de decisión*.

Una característica notable del marco de minimización del riesgo es su gran generalidad: al modificar la forma de la función de pérdida $L(\theta, r(x))$, se obtienen problemas de naturaleza completamente distinta. Cada elección de L determina la forma en que se cuantifica la discrepancia entre el verdadero estado θ y la acción $r(x)$ adoptada por el estadístico.

Por ejemplo, si se define la pérdida como

$$L(\theta, r(x)) = \mathbb{1}_{\{\theta \neq r(x)\}}, \tag{2}$$

el riesgo representa la probabilidad de clasificación errónea, y el problema resultante es el de **clasificación** o **reconocimiento de patrones**.

Si, en cambio, se adopta una pérdida cuadrática, con reglas parametrizadas por θ:

$$L(\theta, r(x)) = (y - r_\theta(x))^2, \tag{3}$$

el riesgo coincide con el error cuadrático medio, dando lugar al problema clásico de **regresión**. En problemas de regresión, la pérdida se define sobre el par (X, Y) generado por θ, y mide la discrepancia entre la respuesta Y y la predicción $r(X)$.

Finalmente, si la cantidad a decidir de interés es una densidad $f_X(x|\theta)$ y se define

$$L(\theta, r(x)) = -\log r_X(x|\theta), \tag{4}$$

la minimización del riesgo conduce al problema de **estimación de densidades**. Esta diversidad revela que los nombres de las distintas ramas —regresión, clasificación y estimación de densidades— no responden a diferencias esenciales en el fundamento teórico, sino a elecciones particulares de la función de pérdida utilizada. Desde la perspectiva de la teoría estadística, todos estos problemas son instancias de un mismo principio unificador: la **minimización del riesgo esperado**, donde la variación en la función L determina el tipo de tarea estadística que se aborda.

Antes de continuar con la materia formal, se presentará un breve repaso de probabilidades que servirá como base conceptual para el desarrollo posterior.

0.2. Repaso de Probabilidades

Este apunte se construye sobre conocimientos básicos de la teoría de probabilidades. En esta sección se dará un repaso por los elementos centrales que constituyen esta teoría, se dará énfasis en las definiciones, notación y aplicación práctica de cada concepto. Para una mayor comprensión pueden referirse al apunte correspondiente del curso de Probabilidades y Procesos Estocásticos.

0.2.1. Espacios de Probabilidad

Un espacio de probabilidad se define como una tripleta $(\Omega, \mathcal{F}, \mathbb{P})$, donde:

- Ω: Espacio muestral, conjunto de todos los resultados posibles.
- $\mathcal{F}$: σ-álgebra de subconjuntos de Ω, contiene los eventos medibles.
- $\mathbb{P}$: Medida de probabilidad, función que asigna a cada evento $A \in \mathcal{F}$ un número $\mathbb{P}(A)$ tal que:
  - $0 \leq \mathbb{P}(A) \leq 1$
  - $\mathbb{P}(\Omega) = 1$
  - Si $\{A_1, A_2, ...\} \subseteq \mathcal{F}$ es una colección numerable de conjuntos **disjuntos**[4] en $\mathcal{F}$ entonces:

$$\mathbb{P}\left(\bigcup_{i\in\mathbb{N}} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i) \text{ } (\sigma - \text{aditividad}).$$

    En particular, sea $n \in \mathbb{N}$, si $\{A_1, A_2, ..., A_n\} \subseteq \mathcal{F}$ es una colección finita de eventos disjuntos entonces:

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i).$$

    Esto último de obtiene al establecer $A_i = \emptyset$ para todo $i \in \{n+1, n+2, ...\}$.

Veremos ahora propiedades adicionales que cumplen las medidas de probabilidad.

Propiedades adicionales: Sean $A, B \in \mathcal{F}$, tenemos que:

1- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ (complemento).

[4] $\forall i, j \in \mathbb{N}, i \neq j, A_i \cap A_j = \emptyset$

2- $\mathbb{P}(\emptyset) = 0$.

3- Si $A \subseteq B \Rightarrow \mathbb{P}(B \backslash A) = \mathbb{P}(B) - \mathbb{P}(A)$ (diferencia).

4- Si $A \subseteq B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$ (monotonía).

5- $0 \leq \mathbb{P}(A) \leq 1$.

6- Sean A, B $\mathcal{F}$ no necesariamente disjuntos entonces: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

7- Si $\{A_1, A_2, ...\} \subseteq \mathcal{F}$ es una colección numerable de conjuntos no necesariamente disjuntos entonces:

$$\mathbb{P}\left(\bigcup_{i\in\mathbb{N}} A_i\right) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i) \text{ } (\sigma - \text{subaditividad}).$$

En particular, sea $n \in \mathbb{N}$, si $\{A_1, A_2, ..., A_n\} \subseteq \mathcal{F}$ es una colección finita de eventos entonces:

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathbb{P}(A_i).$$

8- Si $\{A_1, A_2, ...\} \subseteq \mathcal{F}$ es una colección numerable de conjuntos en $\mathcal{F}$ tales que $A_1 \subseteq A_2 \subseteq A_3 \subseteq ...$ entonces:

$$\mathbb{P}\left(\bigcup_{i\in\mathbb{N}} A_i\right) = \lim_{n\to\infty} \mathbb{P}(A_n) \text{ (continuidad por arriba).}$$

En particular, si $\{B_1, B_2, ...\} \subseteq \mathcal{F}$ es una colección cualquiera, entonces:

$$\mathbb{P}\left(\bigcup_{i\in\mathbb{N}} B_i\right) = \lim_{n\to\infty} \mathbb{P}\left(\bigcup_{i=1}^{n} B_i\right).$$

9- Si $\{A_1, A_2, ...\} \subseteq \mathcal{F}$ es una colección numerable de conjuntos en $\mathcal{F}$ tales que $A_1 \supseteq A_2 \supseteq A_3 \supseteq ...$ entonces:

$$\mathbb{P}\left(\bigcap_{i\in\mathbb{N}} A_i\right) = \lim_{n\to\infty} \mathbb{P}(A_n) \text{ (continuidad por abajo).}$$

En particular, si $\{B_1, B_2, ...\} \subseteq \mathcal{F}$ es una colección cualquiera, entonces:

$$\mathbb{P}\left(\bigcap_{i\in\mathbb{N}} B_i\right) = \lim_{n\to\infty} \mathbb{P}\left(\bigcap_{i=1}^{n} B_i\right).$$

Conceptos y propiedades importantes:

- Independencia: Sean $A, B \in \mathcal{F}$, se dice que A y B son independientes si: $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$. La ocurrencia de un evento no depende de la ocurrencia del otro.
- Probabilidad condicional: Sean $A, B \in \mathcal{F}$, se define la probabilidad condicional de A dado Bcomo $\mathbb{P}(A|B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$
- Regla de Bayes: $\mathbb{P}(A|B) = \dfrac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$
- Probabilidades Totales: Si $\{A_1, A_2, ..., A_n\} \subseteq \mathcal{F}$ es una partición de Ω, entonces:

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B|A_i)\mathbb{P}(A_i)$$

0.2.2. Variables Aleatorias

Una variable aleatoria es una función $X : \Omega \to \mathbb{R}$ que asigna un valor numérico a cada resultado del experimento aleatorio. Gracias a las variables aleatorias ahora es posible inducir una nueva medida de probabilidad, llamada medida de probabilidad inducida.

Sea X una v.a. y sea $A \in \mathcal{B}$, definimos la medida de probabilidad inducida $P_X : \mathcal{B}(\mathbb{R}) \to [0, 1]$ como la función:

$$P_X(A) \triangleq \mathbb{P}(X^{-1}(A)). \tag{5}$$

La ecuación (5) se conoce como **distribución** de la variable aleatoria. Además, gracias a esta definición hemos creado (inducido) un nuevo espacio de probabilidad denotado como $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$. A partir de ahora, si se conoce la distribución ya no es necesario devolverse al espacio de probabilidad original y más aún, se heredan las propiedades dadas por la axiomática de la medida de probabilidad.

Las siguientes son notaciones usuales para la probabilidad de un evento A en el espacio inducido, recordar que A puede ser un singleton, un intervalo, o una combinación de estos. Sea $a, b \in \mathbb{R}$ con $a < b$ y $A \in \mathcal{F}$:

- $P_X(\{a\}) = P_X(X = a) = \mathbb{P}(X(w) = a) = \mathbb{P}(\{w \in \Omega | X(w) = a\}) = \mathbb{P}(X^{-1}(\{a\}))$.
- $P_X([a, b]) = P_X(a \leq X \leq b) = \mathbb{P}(a \leq X(w) \leq b) = \mathbb{P}(\{w \in \Omega | a \leq X(w) \leq b\}) = \mathbb{P}(X^{-1}([a, b]))$.

- $P_X(]-\infty, b]) = P_X(X \leq b) = \mathbb{P}(X(w) \leq b) = \mathbb{P}(\{w \in \Omega | X(w) \leq b\}) = \mathbb{P}(X^{-1}(]-\infty, b]))$.
- $P_X([a, \infty[) = P_X(a \leq X) = \mathbb{P}(a \leq X(w)) = \mathbb{P}(\{w \in \Omega | a \leq X(w)\}) = \mathbb{P}(X^{-1}([a, \infty[))$.
- $P_X(A) = P_X(X \in A) = \mathbb{P}(X(w) \in A) = \mathbb{P}(\{w \in \Omega | X(w) \in A\}) = \mathbb{P}(X^{-1}(A))$

En particular, para el evento $A =]-\infty, x]$, se define la función de probabilidad acumulada $F_X(\cdot)$ como $F_X(x) = P_X(X \leq x) = P_X(]-\infty, x)$.

Las variables aleatorias se clasifican en dos grandes tipos dependiendo del recorrido de la variable aleatoria R_X:

- Una variable aleatoria X es discreta si toma un conjunto finito o numerable de valores posibles. Su comportamiento se describe mediante una **función de probabilidad de masa (pmf)** $p_X(x)$, luego el calcular la probabilidad en este contexto se puede hacer como:

$$P_X(A) = \sum_{x \in A} p_X(x) \tag{6}$$

- Una variable aleatoria X es continua si puede tomar infinitos valores en un intervalo (no numerables) y su comportamiento se describe mediante una **función de densidad de probabilidad (pdf)** $f_X(x)$, luego calcular la probabilidad de que X caiga en un intervalo se calcula mediante una integral:

$$P_X(A) = \int_a^b f_X(x)\, dx \tag{7}$$

  Recordar también que en el caso de variables continuas $P_X(X = a) = \int_a^a f_X(x)\, dx = 0$.

Por lo tanto, si se conoce la función de probabilidad de masa para el caso discreto o la función de densidad de probabibilidad para el caso continuo, es todo lo necesario para poder calcular probabilidades en este nuevo espacio ya que se limita a usar las expresiones en (6) y (7).

En muchos contextos, se requiere modelar más de una variable aleatoria a la vez. Por ejemplo, al analizar la relación entre la temperatura y la humedad, la señal y el ruido, o la entrada y salida de un sistema, se considera una pareja de variables aleatorias (X, Y).

La extensión es directa

Sea $(\Omega, \mathcal{F}, \mathbb{P})$ un espacio de probabilidad, $B \in \mathcal{B}(\mathbb{R}^2)$ y un vector aleatorio (X, Y), definimos la medida inducida $P_{X,Y}$ como:

$$P_{X,Y}(B) = \mathbb{P}(\{w \in \Omega | (X(w), Y(w)) \in B\}). \tag{8}$$

Análogo al caso unidimensional mencionaremos las notaciones más usadas. Sean $a, b, c, d \in \mathbb{R}$:

- $P_{X,Y}(\{a\} \times \{b\}) = P_{X,Y}(X = a, Y = b) = \mathbb{P}(\{w \in \Omega | X(w) = a \wedge Y(w) = b\})$.
- $P_{X,Y}([a, b] \times [c, d]) = P_{X,Y}(a \leq X \leq b, c \leq Y \leq d) = \mathbb{P}(\{w \in \Omega | a \leq X(w) \leq b \wedge c \leq Y(w) \leq d\})$.
- Si el evento se puede describir como un producto cartesiano $C \times D$, entonces:

  $$P_{X,Y}(C \times D) = P_{X,Y}(X \in C, Y \in D) = \mathbb{P}(X^{-1}(C) \cap Y^{-1}(D)).$$

- Si no se puede escribir como un producto cartesiano, se escribe de manera general:

  $$P_{X,Y}(B) = P_{X,Y}((X, Y) \in B) = \mathbb{P}(\{w \in \Omega | (X(w), Y(w)) \in B\})$$

- $P_{X,Y}(\cdot)$ es una medida de probabilidad, la demostración es directa del caso unidimensional.
- Una distribución de dos variables aleatorias se llama distribución **conjunta**.

También podemos reducir el cálculo de las probabilidades según la naturaleza del recorrido de X y de Y.

- Si X y Y son variables discretas, el comportamiento conjunto se describe mediante la función de masa de probabilidad conjunta $p_{X,Y}(x, y) = P_{X,Y}(X = x, Y = y)$. Luego el cálculo de probabilidad de un evento A es

  $$P_{X,Y}(A) = \sum_{(x,y) \in A} p_{X,Y}(x, y) \tag{9}$$

  La función de probabilidad de masa condicional se define como: $p_{X|Y}(x \mid y) = \dfrac{p_{X,Y}(x, y)}{p_Y(y)}$ si $p_Y(y) > 0$.

  Las función de probabilidad de masa marginal se obtiene sumando sobre la otra variable: $p_X(x) = \sum_{y \in R_Y} p_{X,Y}(x, y)$, $p_Y(y) = \sum_{x \in R_X} p_{X,Y}(x, y)$

- Si X y Y son variables continuas, se utiliza la función de densidad conjunta $f_{X,Y}(x,y)$Luego el cálculo de probabilidad de un evento A es

$$P_{X,Y}(A) = \iint_A f_{X,Y}(x,y)dxdy \tag{10}$$

  La función de densidad condicional se define por: $f_{X|Y}(x \mid y) = \dfrac{f_{X,Y}(x,y)}{f_Y(y)}$ si $f_Y(y) > 0$.

  Las densidades marginales se obtienen integrando: $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy$, $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx$

Cuando se poseen n variables aleatorias, es decir, X_1,..., X_n se conoce como vector aleatorio, el cálculo de las probabibilidades en este caso serían una extensión n-dimensional de las expresiones en (9) y (10).

Conceptos y propiedades importantes en variables aleatorias:

- Regla de Bayes:
  - Discreto: $p_{X|Y}(x \mid y) = \dfrac{p_{Y|X}(y \mid x)\,p_X(x)}{p_Y(y)}$
  - Continuo: $f_{X|Y}(x \mid y) = \dfrac{f_{Y|X}(y \mid x)\,f_X(x)}{f_Y(y)}$
- Probabilidades totales:
  - Discreto: $P_Y(Y \in A) = \sum\limits_{x \in R_X} P_{Y|X}(A \mid X = x)\,p_X(x)$
  - Continuo: $P_Y(Y \in A) = \int\limits_{-\infty}^{\infty} P_{Y|X}(A \mid X = x)\,f_X(x)\,dx$
- Independencia:
  - Discreto: $(\forall x, y \in R) p_{X,Y}(x,y) = p_X(x) p_Y(y)$
  - Continuo: $(\forall x, y \in R) f_{X,Y}(x,y) = f_X(x) f_Y(y)$

0.2.3. Esperanza, Varianza y Esperanza Condicional

En el análisis de fenómenos aleatorios, no basta con conocer la probabilidad de cada resultado posible. En la práctica, muchas veces se requiere resumir el comportamiento

de una variable aleatoria en términos de valores representativos que faciliten la toma de decisiones. En este contexto, la esperanza o valor esperado entrega una medida central de tendencia, mientras que la varianza indica qué tan dispersos están los valores posibles respecto a dicho centro. Estas dos nociones son esenciales para evaluar riesgo, confiabilidad y eficiencia en sistemas que operan bajo incertidumbre.

Por ejemplo, entre dos estimadores de una cantidad desconocida, es habitual preferir aquel con menor varianza, ya que sus resultados fluctúan menos. Del mismo modo, la esperanza permite anticipar el valor promedio de una magnitud aleatoria en múltiples repeticiones del experimento. Ahora recordaremos sus definiciones:

Sea X una variable aleatoria.

- Definimos la esperanza de X como:
  - Discreto: $\mathbb{E}(X) = \sum_{x \in R_X} x \cdot p_X(x)$
  - Continuo: $\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx$

  Además, el teorema de la esperanza de funciones de variables aleatorios es un resultado central para el cálculo de éstas. Supongamos que tenemos la variable aleatoria X y $h : R_X \to \mathbb{R}$ medible. Sea $Y = h(X)$, el valor esperado de Y se calcula como:

  $$\mathbb{E}(h(X)) = \begin{cases} \int_{-\infty}^{\infty} h(x) f_X(x) dx & \text{Continuo} \\ \sum_{x \in R_X} h(x) p_X(x) & \text{Discreto} \end{cases}$$

  El teorema nos permite calcular de manera más directa la esperanza sobre funciones de variables aleatorias. Lo único que se necesita es tener la distribución de P_X no de P_Y, por lo que no se necesita tener la distribución inducida. Este teorema también suele llamarse Teorema del Estadístico Inconsciente.
- Definimos la varianza de X como $Var(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$

Algunas propiedades útiles para la esperanza. Consideremos dos variables aleatorias X, Y y dos funciones medibles $h, g : R_X \to \mathbb{R}$, tenemos que:

1- Sean $\alpha, \beta \in \mathbb{R}$. $\mathbb{E}(\alpha X + \beta) = \alpha \mathbb{E}(X) + \beta$.
2- Si $X \geq 0 \Rightarrow \mathbb{E}(X) \geq 0$.
3- Si $X \geq 0 \Rightarrow (\mathbb{E}(X) = 0 \Leftrightarrow P_X(X = 0) = 1)$.
4- $|\mathbb{E}(X)| \leq \mathbb{E}(|X|)$.

5- Sea $M > 0$ y $|X| \leq M$, $|\mathbb{E}(X)| \leq M$.
6- Sea $X \leq Y \Rightarrow \mathbb{E}(X) \leq \mathbb{E}(Y)$.
7- Sea $g(X) \leq h(X) \Rightarrow \mathbb{E}(g(X)) \leq \mathbb{E}(h(X))$.

Algunas propiedades útiles para la varianza. Sean X, Y variables aleatorias, $\alpha, \beta \in \mathbb{R}$, $\mathbb{E}(X^2)$ y $\mathbb{E}(Y^2)$ son finitos, tenemos que:

1- $Var(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$.
2- $Var(\alpha) = 0$.
3- $Var(\alpha X) = \alpha^2 Var(X)$, en particular, $Var(-X) = Var(X)$.
4- $Var(\alpha X + \beta) = \alpha^2 Var(X)$.
5- $Var(X) \geq 0$.
6- Si X e Y son independientes entonces $Var(X + Y) = Var(X) + Var(Y)$.

También se pueden extender estas definiciones al caso vectorial. Sea $X_1^n = (X_1, ..., X_n)^T$ un vector aleatorio. Entonces:

- Esperanza vectorial: $\mathbb{E}(X_1^n) = (\mathbb{E}(X_1), \mathbb{E}(X_2), \ldots, \mathbb{E}(X_n))^T$
- Matriz de varianza-covarianza:

$$Cov(X_1^n) = \mathbb{E}((X_1^n - \mathbb{E}(X_1^n))(X_1^n - \mathbb{E}(X_1^n))^T).$$

que es una matriz simétrica de dimensión $n \times n$, donde la entrada (i, j) corresponde a:

$$Cov(X_i, X_j) = \mathbb{E}((X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j)))$$

Una de las distribuciones más usadas en ingeniería es la distribución normal. Las razones son variadas, pero principalmente su simpleza de modelamiento y sus propiedades *ideales* la convierten en una opción sólida frente a fenómenos observables en la naturaleza.

Primero repasaremos su distribución y posteriormente mostraremos sus propiedades más características. Sea $X \sim N(\mu, \sigma^2)$, se dice que entonces X sigue una distribución normal de media μ y varianza σ^2. El recorrido de la variable aleatoria es $\mathbb{R}$ y su función de densidad de probabilidad está caracterizada por:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad \forall x \in \mathbb{R}.$$

Propiedades: Se puede verificar que si $X \sim N(\mu, \sigma^2)$, con $\alpha, \beta \in \mathbb{R}$, $\alpha \neq 0$, entonces:

1- $\mathbb{E}(X) = \mu$.
2- $Var(X) = \sigma^2$.
3- $Y = \alpha X + \beta$, luego $Y \sim N(\alpha\mu + \beta, \alpha^2\sigma^2)$.
4- Si $X \sim N(\mu_1, \sigma_1^2)$ e $Y \sim N(\mu_2, \sigma_2^2)$ independientes entonces $Z = X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Para el caso multivariado consideremos el vector X_1^n a valores en $\mathbb{R}^n$. El vector X_1^n sigue una distribución normal multivariada $N(\mu_1^n, \Sigma)$ si es que su función de densidad conjunta está dada por:

$$f_{X_1,...,X_n}(x_1, ..., x_n) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x_1^n - \mu_1^n)^t \Sigma^{-1}(x_1^n - \mu_1^n)}.$$

De donde $\mathbb{E}(X_1^n) = \mu_1^n$ es el vector de media y $\Sigma = Cov(X_1^n)$ es la matriz de covarianza. $|\Sigma|$ corresponde al determinante de la matriz de covarianza $Cov(X_1^n)$.

Hasta ahora, hemos considerado la esperanza y la varianza como medidas globales del comportamiento de una variable aleatoria, sin incorporar información adicional. Sin embargo, en muchos contextos es razonable preguntar cuál sería el valor esperado de una variable si conociéramos parte del sistema o el resultado de otra variable relacionada. Este tipo de razonamiento da lugar al concepto de **esperanza condicional**, que permite refinar nuestras estimaciones al incorporar nueva información. Sean X, Y dos variables aleatorias tal que Y es integrable. Entonces la esperanza condicional de Y dado X es una función $g(X) : R_X \to \mathbb{R}$ medible tal que $(\forall x \in R_X)$:

- Si X e Y son continuas:

$$g(x) = \mathbb{E}(Y|X = x) = \begin{cases} \frac{\int_{\mathbb{R}} y f_{X,Y}(x,y)dy}{f_X(x)} & \text{si } f_X(x) > 0 \\ 0 & \text{otro caso.} \end{cases}$$

- Si X e Y son discretas:

$$g(x) = \mathbb{E}(Y|X = x) = \begin{cases} \frac{\sum_{y \in R_Y} y p_{X,Y}(x,y)}{p_X(x)} & \text{si } p_X(x) > 0 \\ 0 & \text{otro caso.} \end{cases}$$

0.2.4. Convergencia de Variables Aleatorias

En muchas situaciones de interés, no estamos frente a un solo experimento aislado, sino frente a secuencias o colecciones de variables aleatorias: la evolución de una inversión

financiera a lo largo del tiempo o la repetición de un experimento físico miles de veces. En todos estos contextos, surge de manera natural la necesidad de entender qué sucede cuando el número de observaciones o experimentos crece indefinidamente. Consideremos una secuencia $(X_n)_{n\in\mathbb{N}}$ de variables aleatorias, tenemos los siguientes tipos de convergencia:

- Convergencia casi segura: $X_n \xrightarrow{c.s.} X$ si $\mathbb{P}\left(\{w \in \Omega : \lim_{n\to\infty} X_n(w) = X(w)\}\right) = 1$
- Convergencia en probabilidad: $X_n \xrightarrow{P} X$ si $(\forall \epsilon > 0) \lim_{n\to\infty} \mathbb{P}\left(\{w \in \Omega : |X_n(w) - X(w)| > \epsilon\}\right) = 0$
- Convergencia en distribución: $X_n \xrightarrow{d} X$ si $(\forall x \in \mathbb{R}) \lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ en todos los puntos de continuidad de F_X.

Los resultados más importantes en convergencia son los siguientes:

- Ley de los grandes números: Consideremos una secuencia de variables aleatorias $(X_n)_{n\in\mathbb{N}}$ independientes e idénticamente distribuidas a valores en $\mathbb{R}$ tales que $(\forall i \in \mathbb{N})\mathbb{E}(X_i) = \mu < \infty$. Entonces:

$$\text{Versión Débil } \frac{\sum_{i=1}^{n} X_i}{n} \xrightarrow{P} \mu \qquad \text{Versión Fuerte } \frac{\sum_{i=1}^{n} X_i}{n} \xrightarrow{c.s.} \mu$$

- Teorema central del límite: Sea $(X_n)_{n\in\mathbb{N}}$ una secuencia de variables aleatorias independientes e idénticamente distribuidas con esperanza $(\forall i \in \mathbb{N})\mathbb{E}(X_i) = \mu < \infty$ y varianza $(\forall i \in \mathbb{N})Var(X_i) = \sigma^2 < \infty$, tenemos que:

$$\frac{\sum_{i=1}^{n} \frac{X_i}{n} - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} Z,$$

  donde $Z \sim N(0, 1)$.

Para finalizar se entrrega una tabla con las distribuciones más usadas, sus parámetros y elementos de interés.

Distribución	**Tipo**	**Parámetros**	**Espacio Muestral**	**P.M.F./P.D.F.**	$\mathbb{E}(X)$	$Var(X)$	**F.G.M**
Bernoulli	Discreta	$\rho \in (0,1)$	$\{0,1\}$	$p_X(x) = \rho^x(1-\rho)^{1-x}$	ρ	$\rho(1-\rho)$	$M_X(t) = 1-\rho+\rho e^t$
Binomial	Discreta	$n \in \mathbb{N}, \rho \in (0,1)$	$\{0,1,\dots,n\}$	$p_X(k) = \binom{n}{k}\rho^k(1-\rho)^{n-k}$	$n\rho$	$n\rho(1-\rho)$	$M_X(t) = (1-\rho+\rho e^t)^n$
Binomial negativa	Discreta	$r \in \mathbb{N}, \rho \in (0,1)$	$\{r, r+1, \dots\}$	$p_X(k) = \binom{k-1}{r-1}\rho^r(1-\rho)^{k-r}$	$\frac{r}{\rho}$	$\frac{r(1-\rho)}{\rho^2}$	$M_X(t) = \left(\frac{\rho e^t}{1-(1-\rho)e^t}\right)^r$
Geométrica	Discreta	$\rho \in (0,1)$	$\{1,2,3,\dots\}$	$p_X(k) = (1-\rho)^{k-1}\rho$	$\frac{1}{\rho}$	$\frac{1-\rho}{\rho^2}$	$M_X(t) = \frac{\rho e^t}{1-(1-\rho)e^t}$
Poisson	Discreta	$\lambda > 0$	$\{0,1,2,\dots\}$	$p_X(k) = \frac{\lambda^k e^{-\lambda}}{k!}$	λ	λ	$M_X(t) = e^{\lambda(e^t-1)}$
Uniforme continua	Continua	$a < b$	$[a,b]$	$f_X(x) = \frac{1}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$M_X(t) = \frac{e^{tb}-e^{ta}}{t(b-a)}$
Exponencial	Continua	$\lambda > 0$	$[0,\infty)$	$f_X(x) = \lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$M_X(t) = \frac{\lambda}{\lambda - t},\ t < \lambda$
Normal (Gaussiana)	Continua	$\mu \in \mathbb{R}, \sigma > 0$	$\mathbb{R}$	$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2	$M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$
Chi-cuadrado	Continua	$k > 0$	$[0,\infty)$	$f_X(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$	k	$2k$	$M_X(t) = (1-2t)^{-k/2},\ t < 1/2$
t de Student	Continua	$\nu > 0$	$\mathbb{R}$	–	0 (si $\nu > 1$)	$\frac{\nu}{\nu-2}$ (si $\nu > 2$)	No definida.
Cauchy	Continua	$x_0 \in \mathbb{R}, \gamma > 0$	$\mathbb{R}$	$f_X(x) = \frac{1}{\pi\gamma[1+((x-x_0)/\gamma)^2]}$	No definida	No definida	No definida
Weibull	Continua	$k > 0, \lambda > 0$	$[0,\infty)$	$f_X(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}$	$\lambda\Gamma(1+\frac{1}{k})$	$\lambda^2[\Gamma(1+\frac{2}{k}) - (\Gamma(1+\frac{1}{k}))^2]$	$M_X(t) = \sum_{n=0}^{\infty} \frac{t^n}{n!}\lambda^n\Gamma\left(1+\frac{n}{k}\right), k \geq 1$
Rayleigh	Continua	$\sigma > 0$	$[0,\infty)$	$f_X(x) = \frac{x}{\sigma^2} e^{-x^2/(2\sigma^2)}$	$\sigma\sqrt{\pi/2}$	$\frac{4-\pi}{2}\sigma^2$	$M_X(t) = \sum_{n=0}^{\infty} \frac{t^n}{n!}\sigma^n 2^{n/2}\Gamma\left(1+\frac{n}{2}\right)$

Cuadro 1: Distribuciones de probabilidad con sus parámetros, soporte, funciones y características principales.

1

Unidad I: Detección Paramétrica

El problema de detección se entiende como el problema de inferir una variable θ discreta (que toma una cantidad finita o numerable de posibles valores) a partir de una variable aleatoria (o vector aleatorio) de observación X. Por ejemplo, las observaciones pueden provenir de una distribución de probabilidad que se conoce en su totalidad salvo por su esperanza, luego, en este contexto, θ representa todos los posibles candidatos a esperanzas o medias desconocidas.

Un problema de detección lo que busca en esencia es entender cuando un modelo cambió a otro debido a distintos factores (perturbaciones externas o errores sistemáticos). La decisión de si un modelo cambió dependerá de las observaciones recibidas y, por lo tanto, de las distribuciones inducidas debido al supuesto cambio de los modelos.

Para que el problema de inferencia sea definido como de detección θ puede tomar una cantidad finita o a lo más numerable de opciones. En caso de que θ tome valores infinitos no numerables se entenderá como estimación.

Ejemplos emblemáticos del problema de detección son el problema de test de hipótesis, los problemas de reconocimiento de patrones y los problemas de inferencia presentes en los sistemas de detección en comunicaciones digitales. Los problemas de detección en

resumen abarcan cualquier desafío en el que se posee una fuente de información y se debe decidir alguna de las finitas o numerables opciones que se poseen bajo algún criterio de optimalidad.

1.1. Formalización del Problema de Detección Paramétrica

Consideremos el caso de detección binario, es decir θ la variable a inferir pertenece al conjunto $\Theta = \{0, 1\}$. El objetivo es decidir, a partir de una o varias observaciones si $\theta = 0$ o bien $\theta = 1$. En este problema, las observaciones se modelan como variables (vectores) aleatorias X_1^n con $n \in \mathbb{N}$, las dos posibles opciones se llaman tradicionalmente hipótesis. Luego, un test de hipótesis binario es una instancia del problema de detección paramétrico cuando $\Theta = \{0, 1\}$ posee las siguientes componentes:

- Un espacio de observación $\mathbb{X}$ y variables aleatorias (o variables de observaciones) X que toman valores en $\mathbb{X}$. El valor particular que pueda tomar X se simboliza con minúscula ($X = x$), x también se conoce como realización, observación o dato. $\mathbb{X}$ es un espacio numérico y puede ser multidimensional, por ejemplo, $\mathbb{X} = \mathbb{R}^n$ con $n \in \mathbb{N}$ en cuyo caso se posee un vector aleatorio $X_1^n = (X_1, ..., X_n) \in \mathbb{X}$ cuyas observaciones o realizaciones son $x_1^n = (x_1, ..., x_n)$.[1]
- Un espacio de parámetros Θ binario, típicamente $\Theta = \{0, 1\}$. También se conoce como el espacio de llegada o el espacio donde nos interesa inferir el parámetro.
- Un espacio de decisión Δ que es, como indica su nombre, lo que se decidirá a partir de las observaciones recibidas. Recordar que también, dadas las hipótesis simplificatorias mencionadas en la introducción, este espacio coincide con el espacio de parámetros, luego $\Delta = \Theta$.
- Dos distribuciones de probabilidad indexadas por $\theta \in \Theta$, es decir, $P_X(\cdot|\theta = 0)$ y $P_X(\cdot|\theta = 1)$ tradicionalmente conocidas como hipótesis.
- Una regla, detector o test $\pi : \mathbb{X} \to \Theta = \{0, 1\}$ que será la función que tomará una decisión en base a algún criterio.

[1] En este apunte en algunos casos se suele hablar del vector de observaciones X (con mayúscula), ya que se entiende como el modelo estocástico subyacente, es decir, de variables aleatorias. Siendo rigurosos, la observación o realización es el acto de observar, luego es un valor fijo (es decir, el x en minúscula), sin embargo, no se debe olvidar que el modelamiento de las observaciones se hace a través de variables aleatorias ya que son éstas las que modelan incertidumbre. Entonces al hablar de vector de observación X_1^n en realidad nos referiremos al vector aleatorio antes de observar los datos. En algunos libros X_1^n también se conoce como fuente de información o fuente de observación o vector variable de observación.

El objetivo es decidir, a partir de observación(es) si $\theta = 0$ o bien $\theta = 1$.

Matemáticamente el problema se suele describir como:

$$\begin{aligned} &H_0 : \theta = 0 \Rightarrow X_1^n \sim P_{X_1^n}(\cdot|\theta = 0) \quad \text{(Hipótesis Nula)} \\ &H_1 : \theta = 1 \Rightarrow X_1^n \sim P_{X_1^n}(\cdot|\theta = 1) \quad \text{(Hipótesis Alternativa)}, \end{aligned} \tag{1.1}$$

$P_{X_1^n}(\cdot|\theta = 0)$ (respectivamente $P_{X_1^n}(\cdot|\theta = 1)$) representa la distribución de probabilidad inducida por X_1^n en caso de que $\theta = 0$ (respectivamente $\theta = 1$) sea la correcta en las observaciones.

En adelante nos gustaría establecer un criterio para decidir una hipótesis o la otra. Supongamos que tenemos una función π (en adelante se llamará regla o test) que va desde el espacio de las observaciones $\mathbb{X}$ al espacio de las decisiones Θ. Como estamos en el escenario binario el espacio de decisión es $\Theta = \{0, 1\}$ donde dado π tenemos que:

$$\begin{aligned} &\pi(x_1^n) = \pi(x_1, ..., x_n) = 0 \quad \text{(Aceptar } H_0\text{)} \\ &\pi(x_1^n) = \pi(x_1, ..., x_n) = 1 \quad \text{(Rechazar } H_0\text{)}. \end{aligned} \tag{1.2}$$

Esto significa que π será nuestro detector[2]. Lo que nos interesa saber es si el detector tiene un buen comportamiento, para eso introduciremos algunas definiciones de desempeño. Dada la regla o detector $\pi : \mathbb{X} \to \Theta = \{0, 1\}$ podemos definir las siguientes medidas.

Definición 1.1. (Tamaño del Test) Dada una regla $\pi : \mathbb{X} \mapsto \{0, 1\}$, se define el **tamaño**

[2] En estadística, es más correcto decir no rechazar H_0, ya que en realidad no se obtuvo la evidencia suficiente para rechazarla, pero tampoco estamos en condiciones de asegurar que es cierta

de π como:

$$\begin{aligned}
\alpha_\pi &\triangleq \underbrace{P_{X_1^n}(\pi(X_1^n)=1|\theta=0)}_{\text{rechazar } H_0 \text{ dado } H_0} \\
&= P_{X_1^n}(X_1^n \in \pi^{-1}(\{1\})|\theta=0) \\
&= \int_{\{(x_1,...,x_n)\in\mathbb{R}^n \,:\, \pi(x_1^n)=1\}} f_{X_1^n}(x_1,...,x_n|\theta=0)dx_1...dx_n \\
&= \int_{\pi^{-1}(\{1\})} f_{X_1^n}(x_1,...,x_n|\theta=0)dx_1...dx_n \\
&= \int_{\pi^{-1}(\{1\})} 1\cdot f_{X_1^n}(x_1,...,x_n|\theta=0)dx_1...dx_n + \int_{\pi^{-1}(\{0\})} 0\cdot f_{X_1^n}(x_1,...,x_n|\theta=0)dx_1...dx_n \\
&= \int_{\pi^{-1}(\{1\})} \pi(x_1^n) f_{X_1^n}(x_1,...,x_n|\theta=0)dx_1...dx_n + \int_{\pi^{-1}(\{0\})} \pi(x_1^n) f_{X_1^n}(x_1,...,x_n|\theta=0)dx_1...dx_n \\
&= \int_{\mathbb{R}^n} \pi(x_1^n)\cdot f_{X_1^n}(x_1,...,x_n|\theta=0)dx_1...dx_n \\
&= \mathbb{E}_{X_1^n}(\pi(X_1^n)|\theta=0) \qquad (1.3)
\end{aligned}$$

Hemos asumido que el espacio de observación $\mathbb{X}=\mathbb{R}^n$, y que por tanto el vector aleatorio está dotado una densidad de probabilidad $f_{X_1^n}$ (la expresión es análoga si es un espacio discreto). Notar que en la quinta igualdad se introdujo un 0 conveniente y, aprovechándose del recorrido de π, permite escribir la integral como en la sexta igualdad. Se puede hacer el razonamiento inverso, es decir, la esperanza se puede reducir al calculo de la probabilidad en el espacio donde se decidió 1 (que es lo que se observa en la cuarta igualdad), esto hace que la integral ya no sea sobre todo $\mathbb{R}^n$ sino que solamente una zona de ella ($\pi^{-1}(\{1\}) = \{x_1^n \in \mathbb{X} : \pi(x_1^n)=1\}$). La ventaja de escribir el tamaño del test en términos de esperanza es que, gracias a la ley de los grandes números, es posible calcular un valor aproximado de este valor a partir una cantidad suficiente de muestras u observaciones.

Notar que α_π es la probabilidad "condicional"[3] de que la regla decida la hipótesis alternativa ($\theta=1$) cuando la correcta era la hipótesis nula ($\theta=0$). Dicho de otra manera, α_π corresponde a la probabilidad de rechazar H_0 cuando H_0 es correcto, la probabilidad de falsa alarma, el error de tipo I, o el tamaño del test, todos estos nombres representan la misma probabilidad de error.

---

[3] En rigor no es una probabilidad condicional debido a que θ no es una variable aleatoria.

Definición 1.2. (Poder del Test) Dada una regla $\pi : \mathbb{X} \mapsto \{0, 1\}$, se define el **poder de** π como:

$$\begin{aligned}
\beta_\pi &\triangleq \underbrace{P_{X_1^n}(\pi(X_1^n) = 1|\theta = 1)}_{\text{aceptar } H_1 \text{ dado } H_1} \\
&= P_{X_1^n}(X_1^n \in \pi^{-1}(\{1\})|\theta = 1) \\
&= \int_{\{(x_1,...,x_n)\in\mathbb{R}^n \,:\, \pi(x_1^n)=1\}} f_{X_1^n}(x_1, ..., x_n|\theta = 1)dx_1...dx_n \\
&= \int_{\pi^{-1}(\{1\})} f_{X_1^n}(x_1, ..., x_n|\theta = 1)dx_1...dx_n \\
&= \int_{\pi^{-1}(\{1\})} 1 \cdot f_{X_1^n}(x_1, ..., x_n|\theta = 1)dx_1...dx_n + \int_{\pi^{-1}(\{0\})} 0 \cdot f_{X_1^n}(x_1, ..., x_n|\theta = 1)dx_1...dx_n \\
&= \int_{\pi^{-1}(\{1\})} \pi(x_1^n) f_{X_1^n}(x_1, ..., x_n|\theta = 1)dx_1...dx_n + \int_{\pi^{-1}(\{0\})} \pi(x_1^n) f_{X_1^n}(x_1, ..., x_n|\theta = 1)dx_1...dx_n \\
&= \int_{\mathbb{R}^n} \pi(x_1^n) \cdot f_{X_1^n}(x_1, ..., x_n|\theta = 1)dx_1...dx_n \\
&= \mathbb{E}_{X_1^n}(\pi(X_1^n)|\theta = 1) \qquad (1.4)
\end{aligned}$$

Los argumentos para obtener las igualdades en (1.4) son análogos al caso del tamaño del test presentados anteriormente. Este valor indica la probabilidad de correcta detección de la hipótesis alternativa. Notar que $P_{X_1^n}(\pi(X_1^n) = 0|\theta = 1)$ es la probabilidad de no detección o el error de tipo II que corresponde precisamente a $1 - \beta_\pi$[4].

A partir de las dos definiciones anteriores de tamaño y poder del test, nos damos cuenta que corresponden a criterios bien establecidos para escoger un test. La pregunta que se viene ahora es: ¿cómo elegir un buen test? podemos notar inmediatamente que no basta con pedir que el tamaño del test sea 0, porque al hacer eso es muy probable que el error de tipo II sea 1. Por lo general, los problemas de detección que ofrecen un error de tipo I o II de valor 0 son casos extraordinarios que no abordan un problema más realista desde el punto de vista de la ingeniería. Luego se debe tener en consideración que no lograremos en la práctica que el error sea 0.

[4] En algunos libros, por notación, β_π corresponde al error de tipo II, en este apunte dicho error es $1 - \beta_\pi$.

Entonces, para hablar de un buen test lo que sí se puede pedir es que ofrezca el mejor *compromiso* entre los dos tipos errores, es decir, encontrar el mejor balance entre tamaño y test de forma tal que ningún otro test ofrezca un mejor desempeño. Lo anterior nos permite formalizar el concepto de test óptimo:

Definición 1.3. (Optimalidad de un Test) Consideremos un test π^* de tamaño α_{π^*}, i.e.,

$$\alpha_{\pi^*} = \mathbb{E}_{X_1^n}(\pi^*(X_1^n)|\theta = 0). \tag{1.5}$$

π^* se dirá óptimo para su tamaño si, $\forall \pi \in F(\mathbb{X}, \Theta)$[5] tal que

$$\alpha_\pi = \mathbb{E}_{X_1^n}(\pi(X_1^n)|\theta = 0) \leq \alpha_{\pi^*} \tag{1.6}$$

se tiene que:

$$\beta_\pi \leq \beta_{\pi^*} = \mathbb{E}_{X_1^n}(\pi^*(X_1^n)|\theta = 1). \tag{1.7}$$

Esto nos dice que si π^* es óptimo para su tamaño α, cualquier otro test de tamaño menor que α (i.e. con menor error de tipo I), tendrá necesariamente un menor poder de test que el test óptimo (en consecuencia tendrá un mayor error de tipo II). En otras palabras podemos decir que π^* es una de las soluciones al problema de decisión óptimo de tamaño α si:

$$\max_{\pi \in F(\mathbb{X}, \Theta)} \mathbb{E}_{X_1^n}(\pi(X_1^n)|\theta = 1) \quad \text{sujeto a} \quad \alpha_\pi \leq \alpha \tag{1.8}$$

Por lo tanto si π^* es solución al problema (1.8) entonces ofrece el máximo poder para su tamaño α_{π^*} y, en consecuencia, ofrece el mejor compromiso entre tamaño y poder.

La pregunta que se debe resolver ahora es de qué manera podemos diseñar un test, y si existe un test que sea óptimo en el sentido de mejor compromiso entre tamaño y poder del test. La respuesta es afirmativa y está dado por el Lema de Neyman-Pearson.

1.2. Lema de Neyman-Pearson

El resultado central de esta sección es el llamado **Lema de Neyman-Pearson** que permite caracterizar de forma cerrada una familia de test óptimos en el sentido de la Definición 1.3. Este resultado nos entrega una receta concreta para poder encontrar test óptimos. Antes de introducir el resultado necesitamos considerar una familia más general del test que permitan la toma de decisiones aleatorias en ciertas circunstancias que garanticen la optimalidad del test.

[5] $F(\mathbb{X}, \Theta)$ es el conjunto de reglas que van de $\mathbb{X}$ a Θ

1.2.1. Test Aleatorios

Recordar que $\mathbb{X}$ es un espacio arbitrario numérico y sus elementos serán denotados como $x \in \mathbb{X}$, luego x podría representar un vector o un escalar (son las observaciones). Definimos el concepto de test binario aleatorio de la siguiente forma:

Definición 1.4. (Test Aleatorio) Un test o regla $\tilde{\pi} : \Omega \times \mathbb{X} \longrightarrow \Theta$ se dice aleatorio si está conformado por dos condiciones:

- Una función de 3 estados $\phi : \mathbb{X} \longrightarrow \{0, 1, 2\}$
- Una variable aleatoria binaria (distribución Bernoulli) $\rho : \Omega \longrightarrow \{0, 1\}$[6] caracterizada por $p = \mathbb{P}(\rho(w) = 1)$.

Luego el test aleatorio se puede escribir, $\forall x \in \mathbb{X}$, como

$$\tilde{\pi}(w, x) = \mathbb{1}_{\phi^{-1}(\{1\})}(x) + \rho(w) \cdot \mathbb{1}_{\phi^{-1}(\{2\})}(x), \tag{1.9}$$

donde $\mathbb{1}_A(x)$ es la función indicatriz del conjunto $A \subset \mathbb{X}$. Por otro lado $\phi^{-1}(\{1\})$ y $\phi^{-1}(\{2\})$ corresponden al conjunto preimagen de 1 y 2, respectivamente.

Observaciones 1.1.

- Un test aleatorio se puede ver como un test de tres estados donde en dos de ellos tiene una salida determinista (0 o 1) y en uno de ellos aleatoria (dado por la variable aleatoria $\rho(w)$).
- La función ϕ particiona el espacio de observación $\mathbb{X}$ en tres componentes $\left\{\phi^{-1}(\{0\}), \phi^{-1}(\{1\}), \phi^{-1}(\{2\})\right\}$

De (1.9) podemos notar que cuando $x \in \phi^{-1}(\{0\}) \Rightarrow \tilde{\pi}(w, x) = 0$, cuando $x \in \phi^{-1}(\{1\}) \Rightarrow \tilde{\pi}(w, x) = 1$ y cuando $x \in \phi^{-1}(\{2\}) \Rightarrow \tilde{\pi}(w, x) = \rho(w)$.

Por tanto solo cuando $x \in \phi^{-1}(\{2\})$, el test ofrece un comportamiento aleatorio gobernado por la variable $\rho(w)$. En otras palabras, los elementos en $\phi^{-1}(\{2\})$ no se sabe con certeza si toman el valor 0 o 1 y es la variable $\rho(w)$ (bernoulli) la que asigna 1 con

[6] Recordar que Ω corresponde al espacio muestral original o espacio madre

probabilidad p o 0 con probabilidad $1-p$.

Redefiniendo $A_0 \triangleq \phi^{-1}(\{0\})$, $A_1 \triangleq \phi^{-1}(\{1\})$, $A_2 \triangleq \phi^{-1}(\{2\})$, tenemos que de la Definición 1.4 una regla aleatoria $\tilde{\pi}$ se caracteriza completamente por una partición del espacio $\{A_0, A_1, A_2\}$ y p que es la probabilidad de $\rho(w) = 1$ ($p = \mathbb{E}(\rho)$[7]), donde

$$\tilde{\pi}(w, x) \triangleq \begin{cases} 1 & \text{si } x \in A_1 \\ 0 & \text{si } x \in A_0 \\ \rho(w) & \text{si } x \in A_2 \end{cases} \tag{1.10}$$

Una manera equivalente de definir esta regla, en el contexto de funciones de variables aleatorias (y por tanto asumiendo que X es una variable aleatoria) es la siguiente $\tilde{\pi} : \{0,1\} \times \mathbb{X} \to \{0,1\}$:

$$\tilde{\pi}(\rho, X) \triangleq \begin{cases} 1 & \text{si } X \in A_1 \\ 0 & \text{si } X \in A_0 \\ \rho & \text{si } X \in A_2 \end{cases} \tag{1.11}$$

$$= \begin{cases} 1 & \text{si } X \in A_1 \vee (X \in A_2 \wedge \rho = 1) \\ 0 & \text{si } X \in A_0 \vee (X \in A_2 \wedge \rho = 0) \end{cases} \tag{1.12}$$

La expresión en (1.12) representa el test considerado como una variable aleatoria, es decir, en un sentido formal y teórico. En cambio, la expresión en (1.10) define el test como función de una realización específica de los datos, x, lo cual es más práctico para su aplicación, ya que se basa en observaciones concretas. En los desarrollos siguientes trabajaremos con ambas notaciones de reglas. El tamaño del test para esta regla está dado por:

$$\begin{aligned} \alpha_{\tilde{\pi}} &= \mathbb{P}(\tilde{\pi}(w, X(w)) = 1 | \theta = 0) \\ &= \mathbb{E}_{\rho,X}(\tilde{\pi}(\rho, X) | \theta = 0)) \\ &= \mathbb{E}_{\rho}(\mathbb{E}_X(\tilde{\pi}(\rho, X) | \theta = 0)) \\ &= \mathbb{E}_{\rho}(\mathbb{E}_X(\mathbb{1}_{A_1}(X) + \rho \cdot \mathbb{1}_{A_2}(X) | \theta = 0)) \\ &= \mathbb{E}_{\rho}(\mathbb{P}(X(w) \in A_1 | \theta = 0) + \rho \cdot \mathbb{P}(X(w) \in A_2 | \theta = 0)) \\ &= \mathbb{P}(X(w) \in A_1 | \theta = 0) + p \cdot \mathbb{P}(X(w) \in A_2 | \theta = 0) \\ &= P_X(X \in A_1 | \theta = 0) + p \cdot P_X(X \in A_2 | \theta = 0), \end{aligned} \tag{1.13}$$

donde se asume que ρ es independiente a X y por tanto el parámetro θ es el que incide exclusivamente en la determinación de las estadísticas de X. Análogamente, el poder del

[7] Recordar que en una distribución Bernoulli el parámetro p corresponde al valor esperado de la variable aleatoria $\rho(w)$.

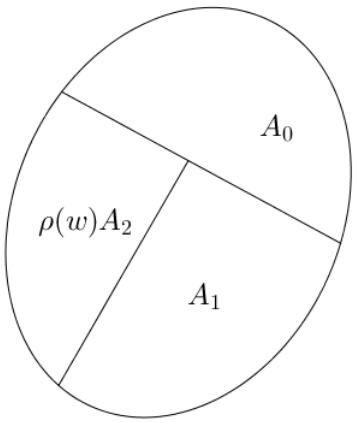


Figura 1.1: Partición de $\mathbb{X}$ inducida por una regla de decisión aleatoria.

test esta dado por:

$$\begin{aligned}
\beta_{\tilde{\pi}} &= \mathbb{P}(\tilde{\pi}(w, X(w)) = 1|\theta = 1) \\
&= \mathbb{E}_{\rho,X}(\tilde{\pi}(\rho, X)|\theta = 1)) \\
&= \mathbb{E}_{\rho}(\mathbb{E}_X(\tilde{\pi}(\rho, X)|\theta = 1)) \\
&= \mathbb{E}_{\rho}(\mathbb{E}_X(\mathbb{1}_{A_1}(X) + \rho \cdot \mathbb{1}_{A_2}(X)|\theta = 1)) \\
&= \mathbb{E}_{\rho}(\mathbb{P}(X(w) \in A_1|\theta = 0) + \rho \cdot \mathbb{P}(X(w) \in A_2|\theta = 1)) \\
&= \mathbb{P}(X(w) \in A_1|\theta = 1) + p \cdot \mathbb{P}(X(w) \in A_2|\theta = 1) \\
&= P_X(X \in A_1|\theta = 1) + p \cdot P_X(X \in A_2|\theta = 1), \qquad (1.14)
\end{aligned}$$

Es posible construir un test aleatorio por medio de la composición o mezcla (aleatoria) de test determinísticos.

Proposición 1.1. Sea $\tilde{\pi}$ un test aleatorio caracterizado por $\{A_0, A_1, A_2\}$ y $p \in (0, 1)$. Consideremos los test determinísticos[8]

$$\pi_1(x) = \begin{cases} 0 & \text{si } x \in A_0 \cup A_2 \\ 1 & \text{si } x \in A_1 \end{cases} \qquad (1.15)$$

[8] Notar que π_1 y π_2 son determinísticos en el sentido que no dependen de ρ.

$$\pi_2(x) = \begin{cases} 0 & \text{si } x \in A_0 \\ 1 & \text{si } x \in A_1 \cup A_2 \end{cases} \tag{1.16}$$

y una variable aleatoria binaria $\rho(w)$ con $\mathbb{P}(\rho(w) = 1) = p$, entonces que se tiene que el test aleatorio $\tilde{\pi}$ puede escribirse de la siguiente maneras:

$$\tilde{\pi}(w, x) = \pi_1(x)(1 - \rho(w)) + \pi_2(x)\rho(w). \tag{1.17}$$

Demostración: Propuesto. □

Por otro lado, la combinación lineal de test aleatorios es un test aleatorio, apoyado por siguiente resultado:

Proposición 1.2. Sea $\rho(w)$ una variable aleatoria binaria arbitraria y $\pi_1(\cdot)$, $\pi_2(\cdot)$ dos test aleatorios arbitrarios, entonces

$$\pi_{12}(w, x) = \pi_1(w, x) \cdot \rho(w) + \pi_2(w, x) \cdot (1 - \rho(w)) \tag{1.18}$$

es un test aleatorio.

Demostración: Propuesto. □

1.2.2. Resultado Principal

A continuación introduciremos el resultado principal conocido como el Lema de Neyman-Pearson. Este Lema suele presentarse levemente distinto a cómo se presentará acá, la razón principal es debido a que un test aleatorio definido en (1.10) se puede definir de otra manera, por lo que el Lema también cambia en presentación. Por completitud, en la sección 1.6 se presenta la definición alternativa de test aleatorio y la versión del Lema en este contexto

---

Teorema 1.1. (Lema de Neyman-Pearson) Sea $\Theta = \{0, 1\}$ y X la variable aleatoria con su correspondiente observación x en $\mathbb{X}$ y dos distribuciones factibles $\{P_X(\cdot|\theta) : \theta \in \{0, 1\}\})$ que definen el problema en (1.1) (es decir que para $\theta = 0$ existe una distribución $P_X(\cdot|\theta = 0)$ y para $\theta = 1$ existe una distribución $P_X(\cdot|\theta = 1)$).

Para un $\nu > 0$ arbitrario y una variable aleatoria binaria $\rho(w)$, se tiene que el test aleatorio de la forma:

$$\pi_\nu(w,x) = \begin{cases} 1 & \text{si } L(x|\theta=1) > \nu L(x|\theta=0) \\ 0 & \text{si } L(x|\theta=1) < \nu L(x|\theta=0) \\ \rho(w) & \text{si } L(x|\theta=1) = \nu L(x|\theta=0) \end{cases} \tag{1.19}$$

es óptimo para su tamaño entre $]0,1[$ en el sentido de la Definición 1.3. Dado que X es una variable aleatoria, el test se puede reescribir como una función de variable aleatoria notando que:

$$\pi_\nu(\rho,X) = \begin{cases} 1 & \text{si } L(X|\theta=1) > \nu L(X|\theta=0) \\ 0 & \text{si } L(X|\theta=1) < \nu L(X|\theta=0) \\ \rho & \text{si } L(X|\theta=1) = \nu L(X|\theta=0) \end{cases} \tag{1.20}$$

con ρ independiente de X. Además el test de la forma (haciendo $\nu \to \infty$):

$$\pi(x) = \begin{cases} 1 & \text{si } L(x|\theta=0) = 0 \\ 0 & \text{si } L(x|\theta=0) > 0 \end{cases} \tag{1.21}$$

es óptimo de tamaño 0 y el test de la forma (haciendo $\nu \to 0^+$):

$$(\forall x \in \mathbb{X})\pi(x) = 1. \tag{1.22}$$

es óptimo de tamaño 1.

Observaciones 1.2.

- $L(x|\theta)$ es la función de verosimilitud. $L(x|\theta)$ cambiará dependiendo si $\mathbb{X}$ es un espacio continuo o discreto. Así, en el caso continuo $L(x|\theta)$ corresponde a la densidad de X evaluado en $X = x$ y en el caso discreto $L(x|\theta)$ será la función de probabilidad de masa evaluado en $X = x$. Más precisamente, asumiendo que $\mathbb{X} = \mathbb{R}^n$, definimos la función de verosimilitud $L : \Theta \to \mathbb{R}^+ \cup \{0\}$ como:
  - Para cada θ, si X_1^n es un vector aleatorio y admite una función de probabilidad de masa conjunta $p_{X_1^n}(\cdot|\theta)$ la verosimilitud es, para $x_1^n \in \mathbb{R}^n$:

$$L(X_1 = x_1, X_2 = x_2, ..., X_n = x_n|\theta) = p_{X_1^n}(x_1, x_2, ..., x_n|\theta).$$

- Para cada θ, si X_1^n es un vector aleatorio que induce una distribución continua y admite una densidad conjunta $f_{X_1^n}(\cdot|\theta)$ la verosimilitud es, para $x_1^n \in \mathbb{R}^n$:

$$L(X_1 = x_1, X_2 = x_2, ..., X_n = x_n|\theta) = f_{X_1^n}(x_1, x_2, ..., x_n|\theta).$$

La función de verosimilitud es una función del parámetro θ dado un vector de observaciones, luego no es más que la probabilidad de masa conjunta o densidad de probabilidad conjunta evaluada en las observaciones por cada valor de θ. Sin embargo, al ser función de X, indexada por θ, la función de verosimilitud es también una variable aleatoria y por eso se puede escribir como $L(X|\theta)$. Normalmente la función de verosimilitud en estadística tiene un énfasis distinto ya que la variable de observación $x \in \mathbb{X}$ está fija ya que la observación es un dato, luego la verosimilitud es función de θ[9], pero podemos apreciar que, independiente de la notación, el énfasis es el mismo.

- $\forall \alpha \in [0, 1]$ existe un test aleatorio de la forma en (1.19), (1.21) y (1.22) tal que su tamaño de test es α (existencia). Esto siempre y cuando exista al menos un conjunto $A \subset \mathbb{X}$ tal que $(\forall x \in A)|L(x|\theta = 0) - L(x|\theta = 0)| \neq 0$, es decir, debe existir algún grado de acoplamiento entre las verosimilitudes.
- El test $\pi(\cdot)$ en (1.19), (1.21) y (1.22) es único para su tamaño salvo soluciones que difieren de π en un conjunto de probabilidad cero respecto a $L(X|\theta = 1)$ y $L(X|\theta = 0)$.
- El Teorema 1.1 nos dice que la razón $\frac{L(X|\theta=1)}{L(X|\theta=0)}$ de probabilidades (o likelihood ratio) es la información suficiente que permite construir una familia de test óptimo en el sentido de la Definición 1.3.
- La intepretación de este test es que existen 3 zonas de decisión, si se cumple la condición $L(X|\theta = 1) > \nu L(X|\theta = 0)$ se decide 1, si se cumple $L(X|\theta = 1) < \nu L(X|\theta = 0)$ se decide 0 y si se cumple $L(X|\theta = 1) = \nu L(X|\theta = 0)$ existe incertidumbre y, por lo tanto, se deja al azar. Este último conjunto se conoce como un evento de borde y es aquél que dada su naturaleza no entrega información suficiente para decidir, por lo que se deja al azar tomar la decisión, más adelante veremos que también sirve como mecanismo de ajuste para lograr el tamaño de test que deseemos.

---

[9] Por esta razón, suele escribirse a veces como $L(x_1^n, \theta)$ o $L(\theta)$, omitiendo la dependencia de x_1^n.

1.2.3. Demostración

Para la demostración de este resultado, verificaremos su optimalidad y existencia.

[Optimalidad]: El resultado se demostrará para el caso continuo, el caso discreto es análogo. Necesitamos demostrar que $\forall \nu \in \mathbb{R}^+$ y $\forall p \in [0,1]$, $\pi_\nu(\rho, X)$ de parámetros $\{A_0^\nu, A_1^\nu, A_2^\nu, p\}$, con

$$\begin{aligned} A_0^\nu &\triangleq \{x \in \mathbb{X} : f_X(x|\theta=1) < \nu f_X(x|\theta=0)\} \\ A_1^\nu &\triangleq \{x \in \mathbb{X} : f_X(x|\theta=1) > \nu f_X(x|\theta=0)\} \\ A_2^\nu &\triangleq \{x \in \mathbb{X} : f_X(x|\theta=1) = \nu f_X(x|\theta=0)\}, \end{aligned} \tag{1.23}$$

es óptimo dado su tamaño $\mathbb{E}(\pi_\nu(\rho, X)|\theta=0) = \alpha_{\pi_\nu}$. En otras palabras, si existe otro test aleatorio π tal que: $\alpha_\pi \leq \alpha_{\pi_\nu}$ entonces seria suficiente verificar que

$$\beta_\pi \leq \beta_{\pi_\nu}. \tag{1.24}$$

Consideremos para estos efectos el siguiente desarrollo descompuesto en tres integrales. Fijemos un valor ρ_1 para el test óptimo y un ρ para el test arbitrario, tenemos que:

$$\begin{aligned} &\mathbb{E}_X(\pi_\nu(\rho_1, X) - \pi(\rho, X)|\theta=1) - \nu(\mathbb{E}_X(\pi^v(\rho_1, X) - \pi(\rho, X)|\theta=0)) \\ &= \int_{\mathbb{X}=A_0^\nu \cup A_1^\nu \cup A_2^\nu} (\pi_\nu(\rho_1, x) - \pi(\rho, x))(f_X(x|\theta=1) - \nu f_X(x|\theta=0))dx \\ &= \underbrace{\int_{A_0^\nu} -\pi(\rho, x) \underbrace{(f_X(x|\theta=1) - \nu f_X(x|\theta=0))}_{<0 \text{ de (1.23)}} dx}_{\geq 0} \\ &+ \underbrace{\int_{A_1^\nu} (1 - \pi(\rho, x)) \underbrace{(f_X(x|\theta=1) - \nu f_X(x|\theta=0))}_{>0 \text{ de (1.23)}} dx}_{\geq 0} \\ &+ \underbrace{\int_{A_2^\nu} \underbrace{(\pi_\nu(\rho_1, x) - \pi(\rho, x))(f_X(x|\theta=1) - \nu f_X(x|\theta=0))}_{=0 \text{ de (1.23)}} dx}_{=0}. \end{aligned} \tag{1.25}$$

Esto lleva a que para todo ρ_1 y ρ arbitrarios:

$$\mathbb{E}_X(\pi_\nu(\rho_1, X)|\theta=1) - \mathbb{E}_X(\pi(\rho, X)|\theta=1) \geq \nu(\mathbb{E}_X(\pi_\nu(\rho_1, X)|\theta=0) - \mathbb{E}_X(\pi(\rho, X)|\theta=0)). \tag{1.26}$$

Tomando esperanza en ambos lados de (1.26) con respecto a ρ_1 y ρ (la parte aleatoria de π_ν y π, respectivamente) se tiene que:

$$\beta_{\pi_\nu} - \beta_\pi \geq \nu(\alpha_{\pi_\nu} - \alpha_\pi). \tag{1.27}$$

Finalmente como $\alpha_{\pi_\nu} \geq \alpha_\pi$, esto implica que $\beta_{\pi_\nu} \geq \beta_\pi$. Para el caso del test definido en (1.21) tenemos que su tamaño es:

$$\begin{aligned}
\alpha_{\pi^*} &= P_X(\pi^*(X) = 1|\theta = 0) \\
&= P_X(L(X|\theta = 0) = 0|\theta = 0) \\
&= \int_{\{x\in\mathbb{X}: f_X(x|\theta=0)=0\}} f_X(x|\theta = 0)dx \\
&= 0.
\end{aligned} \tag{1.28}$$

Lo que corrobora que es de tamaño 0. Supongamos ahora que existe otro test π tal que $\alpha_\pi = \mathbb{E}(\pi(X)|\theta = 0) = 0$, esto significa, por propiedad de la esperanza que $\pi(x) = 0$ en el conjunto $A = \{x \in \mathbb{X} : f_X(x|\theta = 0) > 0\}$. Notemos que $A^c = \{x \in \mathbb{X} : f_X(x|\theta = 0) = 0\}$. Analizemos ahora $\beta_{\pi^*} - \beta_\pi$:

$$\begin{aligned}
\beta_{\pi^*} - \beta_\pi &= \mathbb{E}(\pi^*(X) - \pi(X)|\theta = 1) \\
&= \underbrace{\int_A (\pi^*(x) - \pi(x))f_X(x|\theta = 1)dx}_{=0} + \int_{A^c} (\pi^*(x) - \pi(x))f_X(x|\theta = 1)dx \\
&= \int_{A^c} (1 - \pi(x))f_X(x|\theta = 1)dx \\
&\geq 0
\end{aligned} \tag{1.29}$$

Luego, $\beta_{\pi^*} \geq \beta_\pi$, lo que demuestra la optimalidad de tamaño 0 para este test. Finalmente el test definido en (1.22) posee tamaño 1 y poder 1, en efecto:

$$\begin{aligned}
\alpha_\pi &= P_X(\pi(X) = 1|\theta = 0) \\
&= P_X(\mathbb{X}|\theta = 0) \\
&= 1.
\end{aligned} \tag{1.30}$$

$$\begin{aligned}
\beta_\pi &= P_X(\pi(X) = 1|\theta = 1) \\
&= P_X(\mathbb{X}|\theta = 1) \\
&= 1.
\end{aligned} \tag{1.31}$$

Dado que tiene poder 1, necesariamente el poder de cualquier otro test será menor o igual, con lo que concluimos la demostración.

[Existencia]: Tenemos que mostrar que $\forall \alpha \in [0,1]$ existe un test aleatorio $\pi_\nu \to \{A_0, A_1, A_2, p\}$ donde ρ es su variable aleatoria binaria, tal que su tamaño del test es efectivamente α. Ya sabemos que los test de tamaño 0 y 1 existen por la construcciones en (1.21) y (1.22), veremos ahora el caso $\alpha \in]0,1[$

Para esto analizamos el tamaño del test de parámetros $\{A_0, A_1, A_2, p\}$:

$$\begin{aligned}\alpha_{\pi_\nu} &= \mathbb{E}_\rho(\mathbb{E}_X(\pi_\nu(\rho, X)|\theta = 0)) \\ &= \mathbb{E}_\rho(\mathbb{P}(X(w) \in A_1|\theta = 0) + \rho(w)\mathbb{P}(X(w) \in A_2|\theta = 0)) \\ &= \mathbb{P}(X(w) \in A_1|\theta = 0) + \mathbb{P}(X(w) \in A_2|\theta = 0) \cdot p \\ &= P_X(X \in A_1|\theta = 0) + p \cdot P_X(X \in A_2|\theta = 0). \end{aligned} \tag{1.32}$$

Asumamos ahora $\alpha \in]0,1[$, por definición, (y asumiendo que X admite densidad) el primer término en (1.32) corresponde a:

$$P_X\left(f_X(X|\theta = 1) > \nu f_X(X|\theta = 0)|\theta = 0\right) = P_X\left(\frac{f_X(X|\theta = 1)}{f_X(X|\theta = 0)} > \nu \middle| \theta = 0\right), \tag{1.33}$$

y el segundo término en (1.32) a:

$$P_X(f_X(X|\theta = 1) = \nu f_X(X|\theta = 0)|\theta = 0) = P_X\left(\frac{f_X(X|\theta = 1)}{f_X(X|\theta = 0)} = \nu \middle| \theta = 0\right). \tag{1.34}$$

Notar que resulta útil mirar la siguiente variable aleatoria $Y = H(X) = \frac{f_X(X|\theta=1)}{f_X(X|\theta=0)}$ (llamado razón de verosimilitud o *likelihood ratio*)[10] inducida por X, donde tenemos que:

$$\alpha_{\pi_\nu} = P_Y\left(Y > \nu|\theta = 0\right) + pP_Y\left(Y = \nu|\theta = 0\right). \tag{1.35}$$

En el caso que Y tenga una densidad bajo el modelo $\theta = 0$ entonces su función de distribución $F_Y(y|\theta = 0)$ es continua y por lo tanto $P_Y(Y = \nu|\theta = 0) = 0$[11].

Formalmente si Y tiene una densidad $f_Y(y)$ entonces se verifica que:

$$P_Y\left(Y > \nu \middle| \theta = 0\right) \quad \text{y} \quad P_Y\left(Y \geq \nu \middle| \theta = 0\right) \tag{1.36}$$

[10] Podemos asumir el caso donde $f_X(X|\theta = 0) \neq 0$ ya que el caso $f_X(X|\theta = 0) = 0$ tiene probabilidad 0.
[11] Recordar que esto es porque estamos pidiendo la integral sobre un único valor y no sobre un intervalo

son funciones continuas de ν y, por lo tanto, existe ν^* (como función de α) tal que

$$P_Y\left(Y > \nu^*(\alpha)\middle|\theta = 0\right) = \alpha. \tag{1.37}$$

Entonces, en el caso continuo, para todo $\alpha \in [0,1]$ existe un ν tal que $P_Y\left(Y > \nu|\theta = 0\right) = \alpha$ lo que resuelve el problema de existencia.

Supongamos ahora que $Y = H(X) = \frac{f_X(X|\theta=1)}{f_X(X|\theta=0)}$ no necesariamente admite densidad (y continuidad) y es tal que la función $F_Y(\nu) = P_Y(Y \leq \nu|\theta = 0)$ no toma el valor α, es decir, la función de distribución acumulada es discontinua en ν_0 y existe ν_0 tal que

$$1 - \alpha \leq F_Y(\nu_0) \Leftrightarrow P_Y\left(Y > \nu_0\middle|\theta = 0\right) \leq \alpha \text{ y} \tag{1.38}$$

$$(\forall \epsilon > 0) F_Y(\nu_0 - \epsilon) < 1 - \alpha \Leftrightarrow (\forall \epsilon > 0)\ \ P_Y\left(Y > \nu_0 - \epsilon\middle|\theta = 0\right) > \alpha. \tag{1.39}$$

Además vemos que $\{w \in \Omega : Y(w) > \nu_0\} \subseteq \{w \in \Omega : Y(w) \geq \nu_0\}$, por lo que la probabilidad $P_Y(Y > \nu|\theta = 0) = \mathbb{P}(Y(w) > \nu|\theta = 0)$ es decreciente con ν (otra forma de argumentar es que es sabido que $F_Y(\nu)$ es creciente y $P_Y(Y > \nu|\theta = 0)$ es su complemento). Las siguientes proposiciones nos ayudarán a concluir el resultado pedido

---

Proposición 1.3.

$$\lim_{\epsilon \to 0} P_Y\left(Y > \nu_0 - \epsilon\right) - P_Y\left(Y > \nu_0\right) = P_Y\left(Y = \nu_0\right) \Leftrightarrow \lim_{\epsilon \to 0} P_Y\left(Y > \nu_0 - \epsilon\right) = P_Y\left(Y \geq \nu_0\right) \tag{1.40}$$

Demostración: Es una aplicación directa de la continuidad de la medida en probabilidades. □

Proposición 1.4. La condición en (1.38) y (1.39) se observa si y solo si $P_Y(Y = \nu_0|\theta = 0) > 0$.[12]

Demostración: Utilizando la Proposición 1.3 y restando (1.39) con (1.38) se obtiene inmediatamente lo pedido. □

[12] En otras palabras cuando la función de distribución de Y es discontinua en ν_0, ver Figura 1.2.

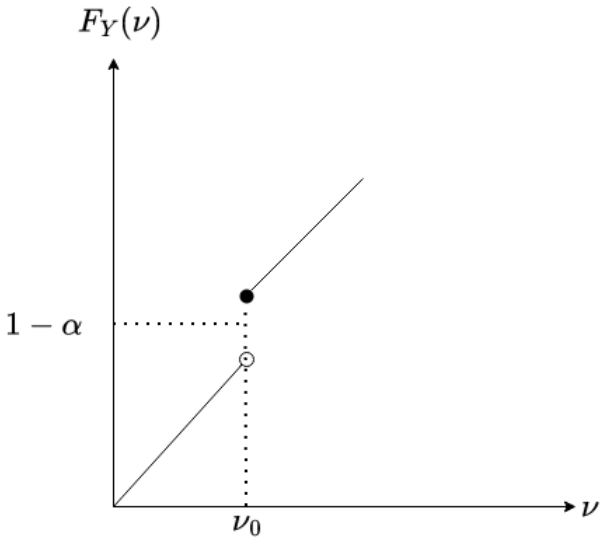


Figura 1.2: Gráfico de la función $F_Y(\nu) = P_Y(Y \leq \nu|\theta = 0)$ bajo la condición en (1.38) y (1.39), recordar que la función de distribución acumulada es continua por la derecha.

Entonces, de las condiciones en (1.38), (1.39) la discontinuidad de $F_Y(\nu_0)$ y el resultado en (1.40), tenemos que:

$$P_Y\left(Y > \nu_0 \middle| \theta = 0\right) \leq \alpha \text{ y } P_Y\left(Y \geq \nu_0|\theta = 0\right) > \alpha. \tag{1.41}$$

Con esto podemos considerar ν_0 como parámetro para definir $\{A_0^{\nu_0}, A_1^{\nu_0}, A_2^{\nu_0}\}$ y un $p \in [0, 1[$ como combinación convexa y solución del cálculo del tamaño del test:

$$P_Y\left(Y > \nu_0|\theta = 0\right) + p \cdot P_Y\left(Y = \nu_0|\theta = 0\right) = \alpha \tag{1.42}$$

es decir:

$$p = \frac{\alpha - P_Y\left(Y > \nu_0|\theta = 0\right)}{P_Y\left(Y = \nu_0|\theta = 0\right)} \in [0, 1[. \tag{1.43}$$

Lo anterior nos dice que si Y presenta un punto de discontinuidad, entonces mediante el ajuste del valor de p es posible de todas maneras lograr un tamaño de test α arbitrario,

esto último es muy usado cuando se poseen probabilidades de masa que no son continuas. Con esto concluimos la demostración de la existencia.

1.2.4. Discusión del Resultado

1- Si Y tiene función de densidad de probabilidad. y, en consecuencia, $\forall \nu \in \mathbb{R}^+$ $P_X(f_X(X|\theta=1) = \nu f_X(X|\theta=0)|\theta=0) = 0$, el test óptimo de Neyman Pearson puede expresarse de forma determinística como:

$$\pi_\nu(x) = \begin{cases} 1 & \text{si } f_X(x|\theta=1) > \nu f_X(x|\theta=0) \\ 0 & \text{si } f_X(x|\theta=1) \leq \nu f_X(x|\theta=0), \end{cases} \tag{1.44}$$

o en su defecto como:

$$\tilde{\pi}_\nu(x) = \begin{cases} 1 & \text{si } f_X(x|\theta=1) \geq \nu f_X(x|\theta=0) \\ 0 & \text{si } f_X(x|\theta=1) < \nu f_X(x|\theta=0). \end{cases} \tag{1.45}$$

En esta caso $\pi_\nu(x)$, $\tilde{\pi}_\nu(x)$ ofrecen el mismo desempeño en términos que:

$$\mathbb{E}_X(\pi_\nu(X)|\theta=0) = \mathbb{E}_X(\tilde{\pi}_\nu(X)|\theta=0) = \alpha_{\pi_\nu} \tag{1.46}$$

$$\mathbb{E}_X(\pi_\nu(X)|\theta=1) = \mathbb{E}_X(\tilde{\pi}_\nu(X)|\theta=1) = \beta_{\pi_\nu} \tag{1.47}$$

Los test expresados en (1.44) y (1.45) se conocen como **test de verosimilitud**. Que corresponde a una versión particular del test de Neyman-Pearson sujeto a distribuciones continuas.

2- Si $Y(X) = \frac{f_X(X|\theta=1)}{f_X(X|\theta=0)}$ admite densidad, entonces el test para el parámetro ν está dado por:

$$\pi_\nu(x) = \begin{cases} 1 & \text{si } Y(x) \geq \nu \\ 0 & \text{si } Y(x) < \nu \end{cases} \tag{1.48}$$

Por lo que se tiene que:

$$\alpha_{\pi_\nu} = \mathbb{E}_X(\pi_\nu(X)|\theta=0) = P_Y(Y \geq y|\theta=0) = \int_\nu^\infty f_Y(y|\theta=0)dy \tag{1.49}$$

$$\beta_{\pi_\nu} = \mathbb{E}_X(\pi_\nu(X)|\theta=1) = P_Y(Y \geq y|\theta=1) = \int_\nu^\infty f_Y(y|\theta=1)dy. \tag{1.50}$$

Sin embargo determinar expresiones cerradas para la distribución de Y puede ser un problema difícil.

1.3. Curva ROC (Receiver Operating Characteristic)

Dado un problema de decisión binario como en la ecuación (1.1), el Lema de Neyman Pearson nos entrega una familia[13] de test óptimos $\{\pi_\alpha(\cdot) : \forall \alpha \in [0,1]\}$ donde sabemos que:

$$\beta_{\pi_\alpha} = \max_{\pi \in \mathbb{F}(\mathbb{X},\Theta) \text{ con } \alpha_\pi \leq \alpha} \beta_\pi, \tag{1.51}$$

por tanto el conjunto de pares $\{(\alpha, \beta_{\pi_\alpha}) : \alpha \in [0,1]\}$ ofrece el compromiso óptimo para el problema en (1.1) entre los errores de tipo I y tipo II.

Definimos la curva ROC asociado al test de Neyman-Pearson como:

$$f_{ROC}(\alpha) = \beta_{\pi_\alpha}, \ \forall \alpha \in [0,1]. \tag{1.52}$$

Es decir la curva ROC es la función que asocia el mejor poder del test por cada error de tipo I en $[0,1]$.

Proposición 1.5. Se puede verificar que la curva ROC asociada al test de Neyman-Pearson:

1- $f_{ROC}(\alpha)$ es una función no decreciente.
2- $f_{ROC}(1) = 1$. Notar que no necesariamente $f_{ROC}(0) = 0$.
3- $f_{ROC}(\alpha)$ es una función cóncava.

Demostración:
1- Tomemos $\alpha_1 \leq \alpha_2$, tenemos lo siguiente:

$$\begin{aligned} f_{ROC}(\alpha_1) &= \beta_{\pi_{\alpha_1}} \\ &= \max_{\pi \in \mathbb{F}(\mathbb{X},\Theta) \text{ con } \alpha_\pi \leq \alpha_1} \beta_\pi, \\ &\leq \max_{\pi \in \mathbb{F}(\mathbb{X},\Theta) \text{ con } \alpha_\pi \leq \alpha_2} \beta_\pi, \\ &= \beta_{\pi_{\alpha_2}} \\ &= f_{ROC}(\alpha_2), \end{aligned} \tag{1.53}$$

donde la desigualdad proviene del hecho que el máximo se está tomando sobre un conjunto más grande, luego el espacio de búsqueda es mayor lo que permite que se encuentre un test con mejor desempeño.

[13] Se le dice familia ya que es una cantidad no numerable de reglas, indexadas por α o por ν

2- Basta elegir el test de la forma $(\forall x \in \mathbb{X})\pi(x) = 1$.

3- Supongamos que la curva es convexa. Tomemos dos puntos arbitrarios (α_1, β_1) y (α_2, β_2). Ambos puntos están asociados a dos test aleatorios π^1 y π^2 con sus respectivos diseños ν_1, p_1 y ν_2, p_2 respectivamente. Sabemos de la Proposición 1.2 que la combinación convexa entre dos test aleatorios es un test aleatorio y que, entonces, se puede elegir un test aleatorio tal que su tamaño sea $p\alpha_1 + (1-p)\alpha_2$ con $p \in [0,1]$ y, análogamente, su poder sea $p\beta_1 + (1-p)\beta_2$. Si asumimos que la curva ROC es convexa, entonces el valor del poder $p\beta_1+(1-p)\beta_2$ estará por encima del valor del poder dado por el Lema de Neyman-Pearson. Esto es una contradicción porque el test de Neyman-Pearson sabemos que es óptimo, es decir, ningún otro test ofrece mejor compromiso entre tamaño y poder. La Figura 1.3 muestra gráficamente el argumento utilizado.

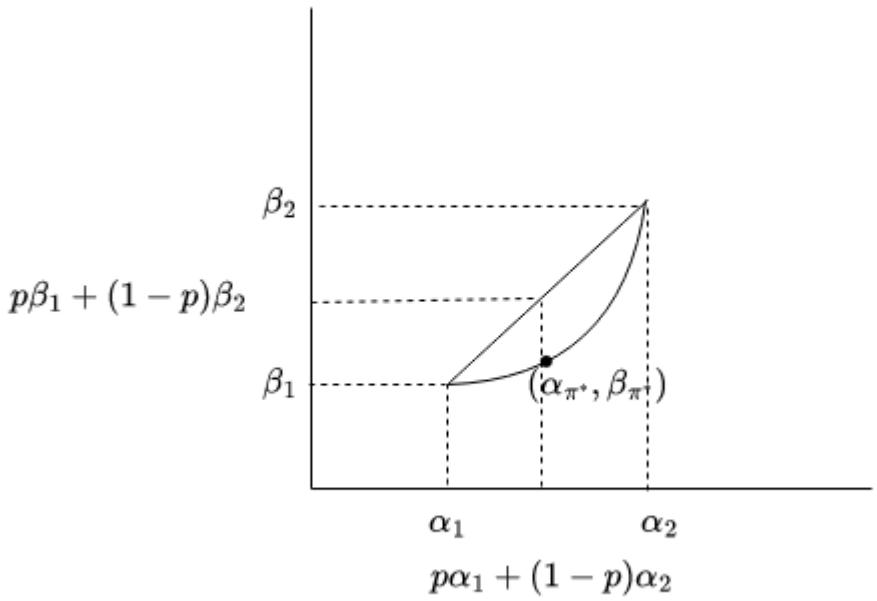


Figura 1.3: Propiedad de convexidad.

□

Una ilustración de una curva ROC típica es presentada en la Figura 1.4. La curva ROC expresa la complejidad del problema de inferencia en el sentido que evidencia el compromiso óptimo alcanzable entre los dos errores que definen este problema.

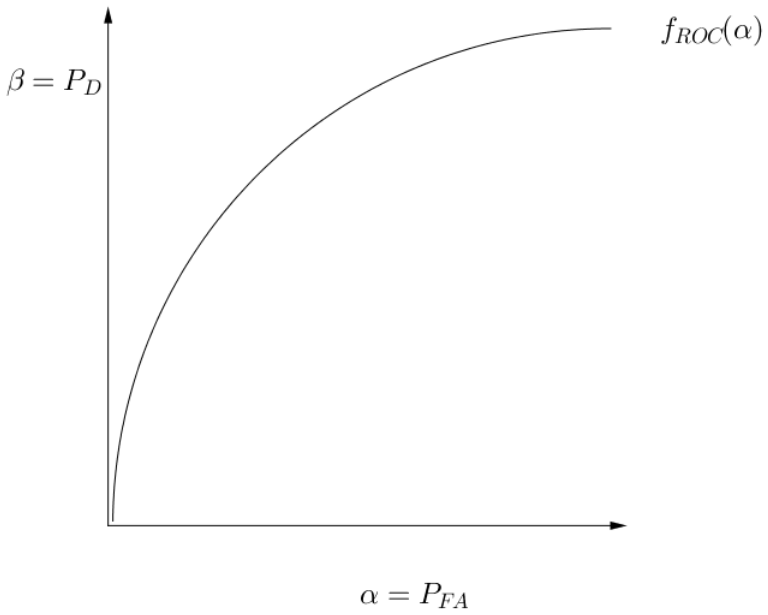


Figura 1.4: Ilustración de la curva ROC para un problema de detección binario.

Observaciones 1.3.

- La curva ROC es una curva teórica ya que depende del test de Neyman-Pearson que a su vez depende de las distribuciones, algo que en la práctica no suele tenerse.
- Es posible generar otras curvas ROC, usando otras reglas de decisión, estas curvas tendrán un mejor desempeño en la medida que se acerquen cada vez más a la curva entregada por el Lema de Neyman-Pearson. Por lo tanto, el desempeño

de una regla de decisión arbitraria será mejor en la medida que se acerque a la entregada por la curva ROC del Lema de Neyman-Pearson.

- El desempeño de la curva de Neyman-Pearson puede mejorar aún más en la medida que se tengan muchas observaciones que provengan de vectores aleatorios independientes e idénticamente distribuidos (i.i.d.), esto es, $\alpha_n \to 0$ y $\beta_n \to 1$ si es que se posee un vector aleatorio $X_1^n \in \mathbb{X}$ i.i.d..

La Figura 1.5 muestra los distintos comportamientos de los detectores, dado que en la práctica no se puede obtener el test óptimo que está dado por el Lema de Neyman-Pearson, se busca un test que se pueda acercar tanto al entregado por el Lema.

Un test tendrá mejor desempeño en la medida que para un valor dado de error de tipo I, el poder del test es lo más grande posible y, consecuentemente, el error de tipo II es más pequeño. En las secciones siguientes veremos aplicaciones directas del Lema de

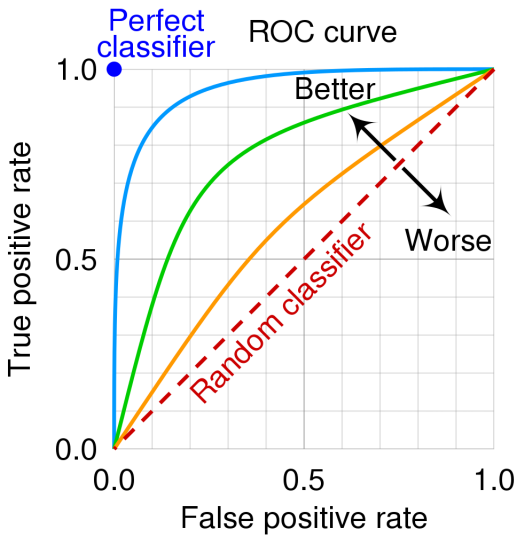


Figura 1.5: Fuente: https://en.wikipedia.org/wiki/Receiver_operating_characteristic

Neyman-Pearson, obtendremos soluciones cerradas e interpretaciones de este resultado para casos emblemáticos.

1.4. Caso de Estudio 1: Ruido Gaussiano

El caso de distribuciones Gaussianas es emblemático tanto por su simplicidad analítica, como por su amplio uso como modelo de observación, en particular en problemas de comunicaciones digitales y reconocimiento de patrones. Veremos una instancia básica de este problema en el siguiente ejemplo:

Ejemplo 1.1. Consideremos $\Theta = \{0, 1\}$ y

$$H_0 : \theta = 0 : X \sim \mathcal{N}(\mu_0, \sigma^2) \rightarrow L(x|\theta = 0) = f_X(x|\theta = 0) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_0)^2}{2\sigma^2}}$$
$$H_1 : \theta = 1 : X \sim \mathcal{N}(\mu_1, \sigma^2) \rightarrow L(x|\theta = 1) = f_X(x|\theta = 1) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}}, \tag{1.54}$$

donde se asume que $\mu_0 < \mu_1$. Estas probabilidades de observación se obtienen por ejemplo en el caso del modelo de ruido aditivo Gaussiano en comunicaciones, donde por uso de canal se transmite una señal de dos posibles estados (binaria) por medio de la regla:

$$H_0 : S = \mu_0$$
$$H_1 : S = \mu_1, \tag{1.55}$$

y las observaciones (en el receptor) están dadas por la variable

$$X = S + Z \tag{1.56}$$

donde $Z \sim \mathcal{N}(0, \sigma^2)$ modela el ruido agregado por el canal de comunicaciones.

En este caso dado un test π lo que debe hacer es decidir si la observación x recibida proviene de una Gaussiana de media μ_0 o de media μ_1.

Aplicaremos entonces el Lema de Neyman-Pearson para determinar la forma de los test óptimos en este caso. Es importante primero caracterizar la función de verosimilitud para cada hipótesis (recordando que la verosimilitud corresponde a la función de densidad de probabilidad o función de probabilidad de masa inducida por la variable aleatoria observada), así, tenemos que, para una observación $x \in \mathbb{X} = \mathbb{R}$

$$f_X(x|\theta = 0) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu_0)^2}{2\sigma^2}} \tag{1.57}$$

y

$$f_X(x|\theta = 1) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu_1)^2}{2\sigma^2}} \tag{1.58}$$

El test se plantea como, dado $\nu > 0$:

$$\pi_\nu(\rho, x) = \begin{cases} 1 & \text{si } \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu_1)^2}{2\sigma^2}} > \nu \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu_0)^2}{2\sigma^2}} \\ 0 & \text{si } \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu_1)^2}{2\sigma^2}} < \nu \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu_0)^2}{2\sigma^2}} \\ \rho & \text{si } \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu_1)^2}{2\sigma^2}} = \nu \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu_0)^2}{2\sigma^2}} \end{cases} \tag{1.59}$$

En general se requerirá expresar este test de forma más amigable, de modo de determinar de forma explícita la partición que genera este test sobre las observaciones. Dicho lo anterior, podemos trabajar una de las desigualdades del test y dejarla más clara.

$$\begin{aligned} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu_1)^2}{2\sigma^2}} &> \nu \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu_0)^2}{2\sigma^2}} \\ e^{\frac{-(x-\mu_1)^2}{2\sigma^2}} &> \nu e^{\frac{-(x-\mu_0)^2}{2\sigma^2}} \\ -\frac{(x-\mu_1)^2}{2\sigma^2} &> \log(\nu) - \frac{(x-\mu_0)^2}{2\sigma^2} \\ \frac{(x-\mu_0)^2}{2\sigma^2} - \frac{(x-\mu_1)^2}{2\sigma^2} &> \log(\nu) \\ (x-\mu_0)^2 - (x-\mu_1)^2 &> 2\log(\nu)\sigma^2 \end{aligned}$$

$$\begin{aligned} x^2 - 2x\mu_0 + \mu_0^2 - (x^2 - 2x\mu_1 + \mu_1^2) &> 2\log(\nu)\sigma^2 \\ x^2 - 2x\mu_0 + \mu_0^2 - x^2 + 2x\mu_1 - \mu_1^2 &> 2\log(\nu)\sigma^2 \\ x(2\mu_1 - 2\mu_0) + \mu_0^2 - \mu_1^2 &> 2\log(\nu)\sigma^2 \\ x &> \frac{2\log(\nu)\sigma^2 + \mu_1^2 - \mu_0^2}{2\mu_1 - 2\mu_0} \end{aligned} \tag{1.60}$$

Luego, el test de Neyman-Pearson se puede expresar de la siguiente forma:

$$\pi_\nu(\rho, x) = \begin{cases} 1 & \text{si } x > \frac{2\log(\nu)\sigma^2 + \mu_1^2 - \mu_0^2}{2\mu_1 - 2\mu_0} \\ 0 & \text{si } x < \frac{2\log(\nu)\sigma^2 + \mu_1^2 - \mu_0^2}{2\mu_1 - 2\mu_0} \\ \rho & \text{si } x = \frac{2\log(\nu)\sigma^2 + \mu_1^2 - \mu_0^2}{2\mu_1 - 2\mu_0} \end{cases} \tag{1.61}$$

Observaciones 1.4.

- Noten que si definimos $A_1^\nu = \left\{x \in \mathbb{R} : x > \frac{2\log(\nu)\sigma^2 + \mu_1^2 - \mu_0^2}{2\mu_1 - 2\mu_0}\right\}$, $A_0^\nu = \left\{x \in \mathbb{R} : x < \frac{2\log(\nu)\sigma^2 + \mu_1^2 - \mu_0^2}{2\mu_1 - 2\mu_0}\right\}$ y $A_2^\nu = \left\{x \in \mathbb{R} : x = \frac{2\log(\nu)\sigma^2 + \mu_1^2 - \mu_0^2}{2\mu_1 - 2\mu_0}\right\}$, la unión

de estos tres conjuntos forman $\mathbb{R}$ y además entre ellos son disjuntos, luego forman una partición del espacio.

- En palabras simples el test nos indica que para decidir, a partir de una observación, entre una hipótesis u otra, se debe verificar que este valor supere o no un umbral que depende de ν y las medias de las Gaussianas.
- Notemos que el evento (sea bajo $\theta = 0$ o $\theta = 1$) $\left\{x \in \mathbb{R} : x = \frac{2\log(\nu)\sigma^2+\mu_1^2-\mu_0^2}{2\mu_1-2\mu_0}\right\}$ es de probabilidad 0 debido a que la probabilidad de X se calcula sobre una distribución continua, por lo que no tiene masa. Dicho de otro modo, la probabilidad de un singleton siempre es 0 sobre cualquier distribución continua. Lo anterior es un argumento suficiente para transformar el test aleatorio en uno determinístico, entregando el conjunto A_2^ν a cualquiera de los otros dos conjuntos (A_0^ν o A_1^ν). El test pasa a ser entonces

$$\pi_\nu(x) = \begin{cases} 1 & \text{si } x \geq \frac{2\log(\nu)\sigma^2+\mu_1^2-\mu_0^2}{2\mu_1-2\mu_0} \\ 0 & \text{si } x < \frac{2\log(\nu)\sigma^2+\mu_1^2-\mu_0^2}{2\mu_1-2\mu_0} \end{cases} \tag{1.62}$$

o también puede ser

$$\pi_\nu(x) = \begin{cases} 1 & \text{si } x > \frac{2\log(\nu)\sigma^2+\mu_1^2-\mu_0^2}{2\mu_1-2\mu_0} \\ 0 & \text{si } x \leq \frac{2\log(\nu)\sigma^2+\mu_1^2-\mu_0^2}{2\mu_1-2\mu_0} \end{cases} \tag{1.63}$$

Cualquiera de los dos es correcto. Lo importante es que en estas situaciones un test aleatorio (que tiene una tercera variable $\rho(w)$) pasa a ser determinístico (que solo tiene dos opciones 0 o 1). Con lo anterior los conjuntos $A_0^\nu = \pi_\nu^{-1}(\{0\})$ y $A_1^\nu = \pi_\nu^{-1}(\{1\})$ ya quedan determinados de la siguiente manera:
Si se adopta la ecuación (1.62):

$$A_0^\nu = \pi_\nu^{-1}(\{0\}) = \left\{x \in \mathbb{R} : x < \frac{2\log(\nu)\sigma^2 + \mu_1^2 - \mu_0^2}{2\mu_1 - 2\mu_0}\right\} \tag{1.64}$$

$$A_1^\nu = \pi_\nu^{-1}(\{1\}) = \left\{x \in \mathbb{R} : x \geq \frac{2\log(\nu)\sigma^2 + \mu_1^2 - \mu_0^2}{2\mu_1 - 2\mu_0}\right\} \tag{1.65}$$

Si se adopta la ecuación (1.63):

$$A_0^\nu = \pi_\nu^{-1}(\{0\}) = \left\{x \in \mathbb{R} : x \leq \frac{2\log(\nu)\sigma^2 + \mu_1^2 - \mu_0^2}{2\mu_1 - 2\mu_0}\right\} \tag{1.66}$$

$$A_1^\nu = \pi_\nu^{-1}(\{1\}) = \left\{x \in \mathbb{R} : x > \frac{2\log(\nu)\sigma^2 + \mu_1^2 - \mu_0^2}{2\mu_1 - 2\mu_0}\right\} \tag{1.67}$$

(Basta elegir una opción para que el problema este resuelto).

Ahora calcularemos el tamaño y poder del test. Recordemos que:

$$\alpha_{\pi_\nu} = \mathbb{E}(\pi_\nu(\rho, X) = 1|\theta = 0) \tag{1.68}$$
$$= P_X(X \in A_1^\nu|\theta = 0) + pP_X(X \in A_2^\nu|\theta = 0) \tag{1.69}$$

Esto quiere decir que corresponde a la probabilidad de observar la variable aleatoria X en el conjunto A_1^ν más la probabilidad de observar la variable aleatoria X en el conjunto A_2^ν por p (la probabilidad de que $\rho = 1$), dado que en realidad la hipótesis correcta era $\theta = 0$, con lo anterior, es claro que se debe integrar lo siguiente

$$\begin{aligned}
\alpha_{\pi_\nu} &= \mathbb{E}(\pi_\nu(\rho, X)|\theta = 0) \\
&= P_X(X \in A_1^\nu|\theta = 0) + pP_X(X \in A_2^\nu|\theta = 0) \\
&= \int_{\frac{2\log(\nu)\sigma^2+\mu_1^2-\mu_0^2}{2\mu_1-2\mu_0}}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu_0)^2}{2\sigma^2}} dx + \int_{\frac{2\log(\nu)\sigma^2+\mu_1^2-\mu_0^2}{2\mu_1-2\mu_0}}^{\frac{2\log(\nu)\sigma^2+\mu_1^2-\mu_0^2}{2\mu_1-2\mu_0}} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu_0)^2}{2\sigma^2}} \\
&= \int_{\frac{2\log(\nu)\sigma^2+\mu_1^2-\mu_0^2}{2\mu_1-2\mu_0}}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu_0)^2}{2\sigma^2}} dx
\end{aligned} \tag{1.70}$$

Similarmente el poder del test se puede calcular como

$$\begin{aligned}
\beta_{\pi_\nu} &= \mathbb{E}(\pi_\nu(\rho, X)|\theta = 1) \\
&= \int_{\frac{2\log(\nu)\sigma^2+\mu_1^2-\mu_0^2}{2\mu_1-2\mu_0}}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu_1)^2}{2\sigma^2}} dx
\end{aligned} \tag{1.71}$$

Resultará útil considerar la función $Q(x) \triangleq P_Z(Z \geq x)$ donde $Z \sim \mathcal{N}(0, 1)$, es decir:

$$Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy. \tag{1.72}$$

Entonces si $X \sim N(\mu, \sigma^2)$ se tiene que $\left(\frac{X-\mu}{\sigma}\right) \sim N(0, 1)$, luego

$$\begin{aligned}
P_X(X \geq x) &= P_X\left(\left(\frac{X-\mu}{\sigma}\right) \geq \frac{x-\mu}{\sigma}\right) \\
&= Q\left(\frac{x-\mu}{\sigma}\right).
\end{aligned} \tag{1.73}$$

Por lo tanto se puede verificar que:

$$\alpha_{\pi_\nu} = Q\left(\frac{\tau(\nu) - \mu_0}{\sigma}\right). \tag{1.74}$$

$$\beta_{\pi_\nu} = Q\left(\frac{\tau(\nu) - \mu_1}{\sigma}\right) \tag{1.75}$$

donde $\tau(\nu) = \frac{2\log(\nu)\sigma^2 + \mu_1^2 - \mu_0^2}{2\mu_1 - 2\mu_0}$.

Observaciones 1.5.

- En general, para test de variables aleatorias continuas es normal que el evento $X \in A_2$ sea de probabilidad cero por lo que suele obviarse en el cálculo del error de tipo I o II, para este caso se decidió ser más explicitó solamente por completitud.
- El resultado anterior entrega un compromiso entre el error de tipo I y el valor de ν. Se observa que existe una relación entre ambos cuya formula explicita no es directa de determinar analíticamente. Sin embargo, la intuición detrás es que a mayor ν es de esperarse un menor error de tipo I (con el compromiso que aumenta el error de tipo II).
- Si bien fijar el ν me entrega un error de tipo I, en la práctica el procedimiento es inverso, es decir, se pide un error α_π (típicamente 0.05) con el que a partir de imponer eso, es posible despejar ν (numéricamente).
- La Figura 1.6 muestra una representación gráfica de la zona de decisión para H_0 y H_1, la ilustración muestra el evidente compromiso entre los dos tipos de errores y el rol del umbral $\tau(\nu) = \frac{2\log(\nu)\sigma^2 + \mu_1^2 - \mu_0^2}{2\mu_1 - 2\mu_0}$. Se aprecia que aumentar el umbral (color azul) significa disminuir el error de tipo 1 (color rojo), sin embargo, esto también hará disminuir el poder del test (color verde).

Consideremos ahora el caso de múltiples mediciones i.i.d.

$$\begin{aligned} H_0 &: \theta = 0 : (X_1, ..., X_n) \sim \mathcal{N}(\mu_0, \sigma^2) \to L(x_1, ..., x_n|\theta = 0) = f_{X_1^n}(x_1, ..., x_n|\theta = 0) \\ H_1 &: \theta = 1 : (X_1, ..., X_n) \sim \mathcal{N}(\mu_1, \sigma^2) \to L(x_1, ..., x_n|\theta = 1) = f_{X_1^n}(x_1, ..., x_n|\theta = 1), \end{aligned} \tag{1.76}$$

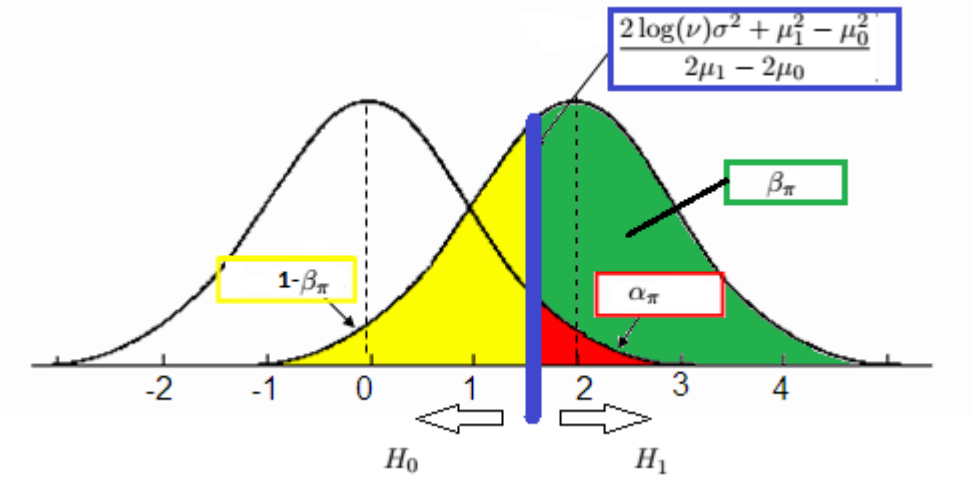


Figura 1.6: Región de decisión para un test de Neyman-Pearson, caso Gaussiano univariado.

Bajo las mismas hipótesis del planteamiento anterior, caracterizaremos la familia de test óptimos en el sentido de Neyman-Pearson para un $\nu \in \mathbb{R}^+$. Nuevamente es importante primero expresar la función de verosimilitud para cada hipótesis. En este caso poseemos n observaciones independientes e idénticamente distribuidas (i.i.d.) y recordando que la verosimilitud conjunta equivale al producto de las marginales cuando son i.i.d, se tiene que para $x_1^n \in \mathbb{X} = \mathbb{R}^n$.

$$f_{X_1^n}(x_1, ..., x_n|\theta = 0) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_i-\mu_0)^2}{2\sigma^2}} \tag{1.77}$$

y

$$f_{X_1^n}(x_1, ..., x_n|\theta = 1) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_i-\mu_1)^2}{2\sigma^2}} \tag{1.78}$$

El test se plantea como

$$\pi_\nu(\rho, x_1, ..., x_n) = \begin{cases} 1 & \text{si } \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_i-\mu_1)^2}{2\sigma^2}} > \nu \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_i-\mu_0)^2}{2\sigma^2}} \\ 0 & \text{si } \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_i-\mu_1)^2}{2\sigma^2}} < \nu \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_i-\mu_0)^2}{2\sigma^2}} \\ \rho & \text{si } \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_i-\mu_1)^2}{2\sigma^2}} = \nu \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_i-\mu_0)^2}{2\sigma^2}} \end{cases} \tag{1.79}$$

Como en el ejemplo anterior, es mejor expresar la partición de una forma más cómoda. De esta forma trabajamos una de las desigualdades

$$\begin{aligned}
\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_i-\mu_1)^2}{2\sigma^2}} &> \nu \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_i-\mu_0)^2}{2\sigma^2}} \\
\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{\sum_{i=1}^{n} \frac{-(x_i-\mu_1)^2}{2\sigma^2}} &> \nu \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{\sum_{i=1}^{n} \frac{-(x_i-\mu_0)^2}{2\sigma^2}} \\
e^{\sum_{i=1}^{n} \frac{-(x_i-\mu_1)^2}{2\sigma^2}} &> \nu e^{\sum_{i=1}^{n} \frac{-(x_i-\mu_0)^2}{2\sigma^2}} \\
-\sum_{i=1}^{n} \frac{(x_i-\mu_1)^2}{2\sigma^2} &> \log(\nu) - \sum_{i=1}^{n} \frac{(x_i-\mu_0)^2}{2\sigma^2} \\
-\sum_{i=1}^{n} (x_i-\mu_1)^2 &> 2\sigma^2 \log(\nu) - \sum_{i=1}^{n} (x_i-\mu_0)^2 \\
\sum_{i=1}^{n} (x_i-\mu_0)^2 - \sum_{i=1}^{n} (x_i-\mu_1)^2 &> 2\sigma^2 \log(\nu)
\end{aligned}$$

$$\begin{aligned}
\sum_{i=1}^{n} x_i^2 - 2x_i\mu_0 + \mu_0^2 - (x_i^2 - 2x_i\mu_1 + \mu_1^2) &> 2\sigma^2 \log(\nu) \\
\sum_{i=1}^{n} -2x_i\mu_0 + \mu_0^2 + 2x_i\mu_1 - \mu_1^2 &> 2\sigma^2 \log(\nu) \\
\sum_{i=1}^{n} x_i(2\mu_1 - 2\mu_0) + \mu_0^2 - \mu_1^2 &> 2\sigma^2 \log(\nu) \\
(2\mu_1 - 2\mu_0) \sum_{i=1}^{n} x_i + n\mu_0^2 - n\mu_1^2 &> 2\sigma^2 \log(\nu) \\
(2\mu_1 - 2\mu_0) n\overline{x} + n\mu_0^2 - n\mu_1^2 &> 2\sigma^2 \log(\nu) \quad \text{con } \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x} \\
\overline{x} &> \frac{2\sigma^2 \log(\nu) - n\mu_0^2 + n\mu_1^2}{(2\mu_1 - 2\mu_0)n} \qquad (1.80)
\end{aligned}$$

donde $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ Con lo que el test ahora queda

$$\pi_\nu(\rho, x_1, ..., x_n) = \begin{cases} 1 & \text{si } \overline{x} > \frac{2\sigma^2 \log(\nu) - n\mu_0^2 + n\mu_1^2}{(2\mu_1 - 2\mu_0)n} \\ 0 & \text{si } \overline{x} < \frac{2\sigma^2 \log(\nu) - n\mu_0^2 + n\mu_1^2}{(2\mu_1 - 2\mu_0)n} \\ \rho(w) & \text{si } \overline{x} = \frac{2\sigma^2 \log(\nu) - n\mu_0^2 + n\mu_1^2}{(2\mu_1 - 2\mu_0)n} \end{cases} \tag{1.81}$$

Nuevamente como estamos en un espacio continuo y el evento $\left\{x_1^n \in \mathbb{R}^n : \overline{x} = \frac{2\sigma^2 \log(\nu) - n\mu_0^2 + n\mu_1^2}{(2\mu_1 - 2\mu_0)n}\right\}$ tiene probabilidad 0 (sea para $\theta = 0$ o $\theta = 1$) ya que la variable aleatoria $\overline{X}$ sigue una distribución normal (la combinación lineal de Gaussianas es Gaussiana), luego, estamos pidiendo la probabilidad de un singleton sobre una variable continua. Podemos entonces reducir este test a uno determinístico, dejándolo como

$$\pi_\nu(x_1, ..., x_n) = \begin{cases} 1 & \text{si } \overline{x} \geq \frac{2\sigma^2 \log(\nu) - n\mu_0^2 + n\mu_1^2}{(2\mu_1 - 2\mu_0)n} \\ 0 & \text{si } \overline{x} < \frac{2\sigma^2 \log(\nu) - n\mu_0^2 + n\mu_1^2}{(2\mu_1 - 2\mu_0)n} \end{cases} \tag{1.82}$$

Para calcular el error de tipo I, trabajando en el espacio $\mathbb{R}^n$ puede no resultar una tarea fácil, sin embargo, se puede recordar que $\overline{X} \sim N(\mu, \sigma^2/n)$, con lo que se define la variable aleatoria $Y = \overline{X}$, luego

$$\alpha_{\pi_\nu} = P_{X_1^n}(\pi_\nu(X_1^n) = 1 | \theta = 0) \tag{1.83}$$

$$\alpha_{\pi_\nu} = P_{X_1^n}\left(\overline{X} \geq \frac{2\sigma^2 \log(\nu) - n\mu_0^2 + n\mu_1^2}{(2\mu_1 - 2\mu_0)n} \middle| \theta = 0\right) \tag{1.84}$$

$$\alpha_{\pi_\nu} = P_{X_1^n}\left(\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \geq \frac{\frac{2\sigma^2 \log(\nu) - n\mu_0^2 + n\mu_1^2}{(2\mu_1 - 2\mu_0)n} - \mu_0}{\sigma/\sqrt{n}} \middle| \theta = 0\right) \tag{1.85}$$

$$\alpha_{\pi_\nu} = P_Y\left(Y \geq \frac{\frac{2\sigma^2 \log(\nu) - n\mu_0^2 + n\mu_1^2}{(2\mu_1 - 2\mu_0)n} - \mu_0}{\sigma/\sqrt{n}}\right) \tag{1.86}$$

$$\alpha_{\pi_\nu} = Q\left(\frac{\frac{2\sigma^2 \log(\nu) - n\mu_0^2 + n\mu_1^2}{(2\mu_1 - 2\mu_0)n} - \mu_0}{\sigma/\sqrt{n}}\right) \tag{1.87}$$

donde $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-y^2/2} dy$. Se ocupó el clásico resultado de la distribución normalizada $N(0, 1)$. Como observarán, nuevamente existe una relación entre α y ν y que finalmente sintetiza el resultado visto en test de hipótesis "se acepta H_0 si el promedio es menor que cierto umbral".

Propuesto 1.1. Genere la curva ROC del test óptimo explorando un rango de valores $\mu \in \mathbb{R}$ y de σ^2 para el caso que $\mu_0 = -\mu_1 = -1$. Comente sus resultados.

Propuesto 1.2. Si definimos $d = |\mu_0 - \mu_1|$ y con ellos $SNR = \frac{d}{\sigma} = \frac{|\mu_0 - \mu_1|}{\sigma}$. Encontrar una expresión para α_{π_ν} y β_{π_ν} en (1.74) y (1.75) como función de ν, SNR y σ^2.

1.5. Caso de Estudio 2: Detección Binaria con Observaciones Discretas

El siguiente ejemplo es un modelo simplificado de un sistema de comunicaciones óptico. En este problema las observaciones son discretas por lo que no es posible reducir el problema a un test determinístico.

Ejemplo 1.2. Se tiene $\theta \in \{0, 1\}$ parámetro fijo que representa el estado de una variable binaria que se transmite por un canal de comunicaciones digitales. La variable observada en el receptor es X con valores en $\mathbb{N}$ (la cantidad de fotones medidos por un detector óptico). El modelo de observación dice que $X \sim Poisson(\lambda)$ donde

$$\begin{aligned} \lambda &= \lambda_0 \quad \text{si } \theta = 0 \\ \lambda &= \lambda_1 \quad \text{si } \theta = 1 \end{aligned} \tag{1.88}$$

es decir

$$\begin{aligned} L(x|\theta = 0) &= P_X(X = x|\theta = 0) = e^{-\lambda_0} \frac{\lambda_0^x}{x!} \\ L(x|\theta = 1) &= P_X(X = x|\theta = 1) = e^{-\lambda_1} \frac{\lambda_1^x}{x!} \end{aligned} \tag{1.89}$$

Estudiemos la forma de la familia de test óptimo que nos ofrece el Lemma de Neyman-Pearson y, en particular, encontraremos los parámetros para el test óptimo de tamaño $\alpha \in (0, 1)$. Asumiremos el caso no trivial donde $\lambda_1 > \lambda_0$. La función de razón de verosimilitud está dado por:

$$\frac{f_X(x|\theta = 1)}{f_X(x|\theta = 0)} = e^{\lambda_0 - \lambda_1} \left(\frac{\lambda_1}{\lambda_0} \right)^x. \tag{1.90}$$

Por lo tanto decidir H_1 corresponde al siguiente conjunto:

$$\begin{aligned} A_1^\nu &= \left\{ x \in \mathbb{N} : e^{\lambda_0 - \lambda_1} \left(\frac{\lambda_1}{\lambda_0} \right)^x > \nu \right\} \\ &= \left\{ x \in \mathbb{N} : \lambda_0 - \lambda_1 + x \ln\left(\frac{\lambda_1}{\lambda_0} \right) > \ln(\nu) \right\} \\ &= \left\{ x \in \mathbb{N} : x > \frac{\ln(\nu) + (\lambda_1 - \lambda_0)}{\ln(\lambda_1) - \ln(\lambda_0)} \right\}. \end{aligned} \tag{1.91}$$

De forma mas general, la partición $\{A_0^\nu, A_1^\nu, A_2^\nu\}$ inducida por la familia de test óptimos es la siguiente (ver Teorema 1.1):

$$\begin{aligned} A_0^\nu &= \left\{ x \in \mathbb{N} : x < \frac{\ln(\nu) + (\lambda_1 - \lambda_0)}{\ln(\lambda_1) - \ln(\lambda_0)} \right\} \\ A_1^\nu &= \left\{ x \in \mathbb{N} : x > \frac{\ln(\nu) + (\lambda_1 - \lambda_0)}{\ln(\lambda_1) - \ln(\lambda_0)} \right\} \\ A_2^\nu &= \left\{ x \in \mathbb{N} : x = \frac{\ln(\nu) + (\lambda_1 - \lambda_0)}{\ln(\lambda_1) - \ln(\lambda_0)} \right\}, \end{aligned} \tag{1.92}$$

con $\lambda_0, \lambda_1 \in \mathbb{R}^+ \backslash \{0\}$ y $\nu > 0$. Notar que A_2 puede ser vacío. Si adicionalmente p es la probabilidad que la variable $\rho(w)$ tome el valor 1, entonces el test queda descrito por $\pi = \{A_1^\nu, A_0^\nu, A_2^\nu, p\}$ y en particular por los parámetros ν y p. El Lema de Neyman-Pearson se plantea como, dado $x \in \mathbb{N}$ y $\nu > 0$:

$$\pi_\nu(\rho, x) = \begin{cases} 1 & \text{si } x > \frac{\ln(\nu)+(\lambda_1-\lambda_0)}{\ln(\lambda_1)-\ln(\lambda_0)} \\ 0 & \text{si } x > \frac{\ln(\nu)+(\lambda_1-\lambda_0)}{\ln(\lambda_1)-\ln(\lambda_0)} \\ \rho & \text{si } x = \frac{\ln(\nu)+(\lambda_1-\lambda_0)}{\ln(\lambda_1)-\ln(\lambda_0)} \end{cases} \tag{1.93}$$

Por tanto la expresión para el tamaño del test esta dada por:

$$\begin{aligned} \alpha_\pi &= P_X(X \in A_1^\nu | \theta = 0) + p \cdot P_X(X \in A_2^\nu | \theta = 0) \\ &= \sum_{x > \frac{\ln(\nu)+(\lambda_1-\lambda_0)}{\ln(\lambda_1)-\ln(\lambda_0)}}^{\infty} e^{-\lambda_0} \frac{\lambda_0^x}{x!} + \mathbb{1}_{\mathbb{N}} \left(\underbrace{\frac{\ln(\nu) + (\lambda_1 - \lambda_0)}{\ln(\lambda_1) - \ln(\lambda_0)}}_{x_\nu \triangleq} \right) \cdot p e^{-\lambda_0} \frac{\lambda_0^{x_\nu}}{x_\nu!}, \end{aligned} \tag{1.94}$$

recordando que $\mathbb{1}_{\mathbb{N}}(x_\nu)$ vale uno si x_ν es natural y 0 si no. Dado un $\alpha \in [0, 1]$ arbitrario, nos pondremos en el caso que existe un test determinístico ($p = 0$) tal que sea de tamaño

α. A partir de (1.94), esto equivale a pedir que $\exists x(\alpha) \in \mathbb{N}$ tal que:

$$1 - \alpha = \sum_{x=1}^{x(\alpha)} e^{-\lambda_0} \frac{\lambda_0^x}{x!}. \tag{1.95}$$

Bajo la identidad en (1.95) se toma ν_α (el subíndice α indica que ν es función de α por ser este último un parámetro de diseño) como solución de:

$$\begin{aligned} \ln(\nu_\alpha) &= x(\alpha) \ln\left(\frac{\lambda_1}{\lambda_0}\right) - (\lambda_1 - \lambda_0) \\ \nu_\alpha &= e^{x(\alpha)\ln\left(\frac{\lambda_1}{\lambda_0}\right) - (\lambda_1 - \lambda_0)}. \end{aligned} \tag{1.96}$$

Si por el contrario para un α dado no es posible encontrar solución para (1.95) para un $x(\alpha)$ entero positivo, necesariamente se debe recurrir a un test aleatorio. En este caso podemos considerar:

$$x_0(\alpha)^* = \arg\max_{x_0 \in \mathbb{N}} \left\{ \sum_{x=x_0+1}^{\infty} e^{-\lambda_0} \frac{\lambda_0^x}{x!} \right\} \text{ tal que } \sum_{x=x_0(\alpha)^*+1}^{\infty} e^{-\lambda_0} \frac{\lambda_0^x}{x!} < \alpha \tag{1.97}$$

Es decir el natural $x_0(\alpha)^*$ que maximice $f(x_0) = \sum_{x=x_0+1}^{\infty} e^{-\lambda_0} \frac{\lambda_0^x}{x!}$ y que garantice que la suma a partir de $x_0(\alpha)^*+1$ sea menor que α. Por lo tanto la suma desde $x_0(\alpha)^*$ será mayor que α[14], y tenemos que

$$\sum_{x=x_0(\alpha)^*+1}^{\infty} e^{-\lambda_0} \frac{\lambda_0^x}{x!} + e^{-\lambda_0} \frac{\lambda_0^{x_0(\alpha)^*}}{x_0(\alpha)^*!} > \alpha \Rightarrow \exists p_\alpha \in [0,1] \tag{1.98}$$

tal que

$$\sum_{x=x_0(\alpha)^*+1}^{\infty} e^{-\lambda_0} \frac{\lambda_0^x}{x!} + e^{-\lambda_0} \frac{\lambda_0^{x_0(\alpha)^*}}{x_0(\alpha)^*!} p_\alpha = \alpha. \tag{1.99}$$

Lo anterior debido a la garantía de existencia del Lema de Neyman-Pearson[15]. Finalmente, el test optimo está dado por los parámetros $\nu_\alpha \to \{A_0^{\nu_\alpha}, A_1^{\nu_\alpha}, A_2^{\nu_\alpha}\}$ y $p_\alpha \in (0,1)$.

---

[14] La función $f(x_0) = \sum_{x=x_0+1}^{\infty} e^{-\lambda_0} \frac{\lambda_0^x}{x!}$ es decreciente

[15] Esto también se puede deducir por el teorema de los valores intermedios

1.6. Anexo: Test Aleatorios y Lema de Neyman-Pearson

Ya vimos una definición de un test aleatorio, en esta sección hablaremos de una segunda definición, que es consistente con lo hablado anteriormente pero a veces es más usada en la comunidad estadística. Sea $\gamma \in [0,1]$, un test aleatorio se puede definir como una regla $\pi_\gamma : \mathbb{X} \to [0,1]$ tal que admite la siguiente estructura para $\{A_0, A_1, A_2\}$ partición de $\mathbb{X}$.

$$\pi_\gamma(x) \triangleq \begin{cases} 1 & \text{si } x \in A_1 \\ 0 & \text{si } x \in A_0 \\ \gamma & \text{si } x \in A_2 \end{cases} \tag{1.100}$$

Notar que esta definición no es compatible con la de una regla de decisión impuesta en el caso de detección binaria, esto porque en este caso el recorrido de π es un continuo y no el caso discreto $\{0,1\}$, por lo que este test en realidad es tal que $\pi_\gamma : \mathbb{X} \to \Delta$ donde $\Delta \neq \Theta$ lo que contradice nuestras hipótesis simplificatorias del problema de detección. Ahora bien, si se adopta este test como regla de decisión y se cambia el espacio de decisión a $\Delta = [0,1]$ entonces el tamaño del test se debe redefinir como sigue:

$$\begin{aligned} \alpha_\pi &= \mathbb{E}_X(\pi_\gamma(X)|\theta = 0) \\ &= P_X(X \in A_1|\theta = 0) + \gamma \cdot P_X(X \in A_2|\theta = 0), \end{aligned} \tag{1.101}$$

y el poder se redefine como:

$$\begin{aligned} \beta_\pi &= \mathbb{E}_X(\pi_\gamma(X)|\theta = 1) \\ &= P_X(X \in A_1|\theta = 1) + \gamma \cdot P_X(X \in A_2|\theta = 1), \end{aligned} \tag{1.102}$$

En este caso las definiciones de tamaño y poder son levemente distintas porque involucra directamente el cálculo de la esperanza. En las Definiciones 1.1 y 1.2 vemos que se pueden interpretar como la probabilidad de error del test. Sin embargo al computar la esperanza de estas redefiniciones recuperamos el mismo resultado del error de tipo I y poder de un test aleatorio, luego es consistente. El Lema de Neyman-Pearson se puede escribir de la siguiente manera:

Teorema 1.2. (Lema de Neyman-Pearson, versión alternativa) Sea $\Theta = \{0,1\}$ y X la variable aleatoria con su correspondiente observación x en $\mathbb{X}$ y dos distribuciones factibles $\{P_X(\cdot|\theta) : \theta \in \{0,1\}\}$ que definen el problema en (1.1) (es decir que para $\theta = 0$ existe una distribución $P_X(\cdot|\theta = 0)$ y para $\theta = 1$ existe una distribución $P_X(\cdot|\theta = 1)$).

Para un $\nu > 0$ arbitrario y $\gamma \in [0, 1]$, se tiene que el test aleatorio de la forma:

$$\pi_\gamma^\nu(x) = \begin{cases} 1 & \text{si } L(x|\theta = 1) > \nu L(x|\theta = 0) \\ 0 & \text{si } L(x|\theta = 1) < \nu L(x|\theta = 0) \\ \gamma & \text{si } L(x|\theta = 1) = \nu L(x|\theta = 0) \end{cases} \tag{1.103}$$

es óptimo para su tamaño entre $]0, 1[$ en el sentido de la Definición 1.3.

Observamos que el enunciado es equivalente al propuesto en el Teorema 1.1, solamente posee una doble indexación en γ y ν, los resultados de optimalidad y existencia son exactamente iguales.

1.7. Problemas

Se presentan a continuación una sección de problemas relacionados con detección paramétrica.

Problema 1.1. (Detección de variables con distribución Poisson)
Considere una variable aleatoria X con distribución Poisson de parámetro λ.

$$P_X(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}, \tag{1.104}$$

a) Determine la función generadora de momentos de X, es decir:

$$M_X(t) = \sum_{k\geq 0} P_X(X=k)\cdot e^{tk}, \tag{1.105}$$

y verifique que es igual a $e^{\lambda\cdot(e^t-1)}$.

b) Considere X_1,..,X_n variables aleatorias independientes e idénticamente distribuidas (i.i.d.) con distribución Poisson de parámetro λ. Verifique que $X = \sum_{i=1}^n X_i$ es Poisson de parámetro $n\lambda$. *Indicación*: Considere los resultados de probabilidades respecto a suma de variables aleatorias y las propiedades de la función generadora de momentos.

c) Considere el problema de detección binario en el escenario paramétrico, donde $\Theta = \{0,1\}$ y se tiene que:

$$\theta = 0 \Rightarrow X \sim Poisson(\lambda_0), \tag{1.106}$$

$$\theta = 1 \Rightarrow X \sim Poisson(\lambda_1) \tag{1.107}$$

con $\lambda_1 > \lambda_0$. Determine la forma general de la familia de test óptimos dados por el Lema de Neyman-Pearson, y analice la forma de las zonas de decisión considerando que $\lambda_1 > \lambda_0$. Comente.

d) Encuentre el test óptimo para el tamaño $\alpha = 0{,}01$. Considere $\lambda_0 = 2$ y $\lambda_1 = 4$. *Indicación*: Notar que un test aleatorio podría ser necesario.

e) Encuentre los valores de tamaño α sobre los cuales los test deterministicos son óptimos o en su defecto la condición que se debe cumplir para ello.

Problema 1.2. (Detección de símbolos sobre ruido aditivo Gaussiano)
Considere el problema clásico de comunicaciones digitales, de la detección de símbolos binarios contaminadas por ruido aditivo Gaussiano. En este caso $\Theta = \{0, 1\}$ y la variable aleatoria de observación dado $\theta \in \Theta$ esta dada por:

$$X = S_\theta + N \tag{1.108}$$

con $S_0 = \mu$ amd $S_1 = -\mu$, $\mu > 0$ y $N \sim \mathcal{N}(0, \sigma^2)$. Del Lema de Neyman-Pearson, se sabe que la familia de test óptimos $\{\pi_\eta(\cdot) : \eta \in \mathbb{R}\}$, es determinística y ofrece la siguiente estructura:

$$\pi_\eta(x) = 1, \text{ si } \ln(l(x)) > \eta \tag{1.109}$$

$$\pi_\eta(x) = 0, \text{ si } \ln(l(x)) \leq \eta \tag{1.110}$$

donde $l(x) = \frac{f_X(x|\theta=1)}{f_X(x|\theta=0)}$ es la razón de verosimilitud.

a) Verifique que la regla de decisión en este caso reduce a: $\pi_\eta(x) = 1$ si $x < \tau_\eta$ y $\pi_\eta(x) = 0$ de lo contrario. Encuentre una expresión para τ_η.
b) Verifique que $Y = \ln(l(X))$ es una variable aleatoria Gaussiana y determine su media y varianza para los dos escenarios $\theta = 0$ y $\theta = 1$.
c) Encuentre expresiones para el poder y el tamaño de $\pi_\eta(\cdot)$ como función de los parámetros del problema (σ^2, μ,η) y la función $Q(z) = \int_z^\infty \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$.
d) Considere $\sigma^2 = 1$, $\mu = 1$,y con ello genere la curva ROC cubriendo un rango representativo de pares de valores de tamaño y poder (utilice Python o el lenguaje de programación que desee para crear la curva).
e) Repita el computo anterior, considerando los siguientes valores para la varianza del ruido $\sigma^2 = 10^{-3}, 10^{-2}, 10^{-1}, 1, 10^2, 10^3$. Analice los resultados obtenidos y comente sobre la complejidad del problema de decisión.

Problema 1.3. (Múltiples mediciones)
Considere el mismo escenario del Problema 1.2, pero asuma que se tienen múltiples mediciones (o en su defecto transmisiones sucesivas del mismo símbolo),

$$X_1, X_2, \ldots, X_n$$

y donde $X_i = S_\theta + N_i$ $(i = 1, .., n)$, para lo cual $N_1, .., N_n$ son variables aleatorias i.i.d. que siquen una $\mathcal{N}(0, \sigma^2)$. Ahora la regla de decisión enfrenta el vector aleatorio Gaussiano $X_1^n = (X_1, .., X_n)$ con valores en $\mathbb{R}^n$ y va al espacio de decisión $\Theta = \{0, 1\}$.

a) Condicionado a los valores de $\theta \in \Theta$, determine la distribución de X_1^n y sus parámetros.
b) Analice la familia de test óptimos y verifique que $\forall x_1^n \in \mathbb{R}^n$

$$\log l(x_1^n) = \frac{2}{\sigma^2}\bar{\mu}^t \cdot x_1^n$$

donde $\bar{\mu} = (\mu, \mu, ..., \mu) \in \mathbb{R}^n$. Específicamente para $n = 2$ y $\eta = 0$, determine gráficamente las zonas de decisión, es decir:

$$A_0 = \pi_\eta^{-1}(\{0\}) = \left\{x_1^2 \in \mathbb{R}^2 : \ln l(x_1^2) \leq \eta\right\},$$

$$A_1 = \pi_\eta^{-1}(\{1\}) = \left\{x_1^2 \in \mathbb{R}^2 : \ln l(x_1^2) > \eta\right\}.$$

c) Considere $\mu = 1$, $\sigma^2 = 10$ y $n = 1, 10, 10^2, 10^3$, respectivamente. Para estos distintos escenarios determine el test óptimo $\pi_\eta^n : \mathbb{R}^n \to \{0, 1\}$, es decir determine η, tal que:

$$\alpha_{\pi_\eta^n} = \mathbb{E}(\pi_\eta^n(X_1^n)|\theta = 0) = 0{,}01$$

y con ello grafique $\beta_{\pi_\eta^n} = \mathbb{E}(\pi_\eta^n(X_1^n)|\theta = 1)$ como función de n. Comente que observa en el poder del test y cual es la influencia en el número de mediciones.
d) Complemente el análisis anterior generando la curva ROC completa para los escenarios $n = 1, 10, 10^2, 10^3$. Comente si este resultado es consistente con lo observado en el punto anterior.

Problema 1.4. Considere un problema de detección binario $\Theta = \{0, 1\}$ donde la variable aleatoria fuente de observación X toma valores en la recta real $\mathbb{X} = \mathbb{R}$ y sigue las estadísticas como función del parámetro θ dadas por:

$$\theta = 0 : X \sim Unifome[0, 1]$$

$$\theta = 1 : X \sim Unifome[0, K]$$

con $K > 1$.

a) Determine la familia de test óptimos en el sentido del Lema de Neyman-Pearson. Considere solamente la familias de test óptimas de tamaño entre $]0,1[$

b) Fije un umbral $\tau \in \mathbb{R}$ y considere el siguiente test determinístico:

$$\pi_\tau(x) = 1 \text{ si } \log \frac{f_X(x|\theta=1)}{f_X(x|\theta=0)} \geq \tau \tag{1.111}$$

y $\pi_\tau(x) = 0$ si la condición en (1.111) no se cumple[16]. Determine las regiones de decisión de π_τ, es decir los conjuntos $A_0^\tau = \pi_\tau^{-1}(\{0\})$ y $A_1^\tau = \pi_\tau^{-1}(\{1\})$. Especifique como cambian dichas regiones como función de τ e identifique rangos concretos en el espacio de posibles valores de τ.

c) Del punto anterior, determine las expresiones para el poder y tamaño del test como función del valor de τ. Recordar que:

$$\alpha_{\pi_\tau} = P_X(\pi_\tau(X) = 1|\theta = 0)$$
$$\beta_{\pi_\tau} = P_X(\pi_\tau(X) = 1|\theta = 1)$$

d) Determine la curva ROC. Es posible obtener la curva ROC completa (para todos los tamaños) con test determinísticos? Justifique su respuesta.

e) Vuelva al punto b) y d) y discuta que pasa con las regiones de decisión y la curva ROC si $K \to \infty$.

Problema 1.5. Considere una secuencia binaria de largo n, un vector $s_1^n = (s_1, ..., s_n) \in \{0,1\}^n$ transmitida por un canal binario simétrico (BSC). La probabilidad condicional de observar $x_1^n = (x_1, ..., x_n) \in \{0,1\}^n$ a la salida del canal dado que se transmite la secuencia $s_1^n = (s_1, ..., s_n)$ esta dada por el siguiente modelo:

$$\begin{aligned} P_{X_1^n|S_1^n}(X_1^n = x_1^n|S_1^n = s_1^n) &= \prod_{i=1}^n P_{X_i|S_i}(X_i = x_i|S_i = s_i) \\ &= \prod_{i=1}^n (\epsilon \cdot \mathbb{1}_{\{x_i \neq s_i\}} + (1-\epsilon) \cdot \mathbb{1}_{\{x_i = s_i\}}) \end{aligned} \tag{1.112}$$

donde $\epsilon \in (0,1)$ es la probabilidad de error del canal.

[16] Considere para estos efectos que $\log \frac{0}{0} \triangleq \lim_{x\to 0} \log \frac{x}{x} = 0$.

a) Encuentre una expresión para $P_{X_1^n|S_1^n}(X_1^n = x_1^n|S_1^n = s_1^n)$ como función de

$$d_H((x_1, ..., x_n); (s_1, ..., s_n)) = d_H(x_1^n; s_1^n) = \sum_{i=1}^{n} \mathbb{1}_{\{x_i \neq s_i\}}, \tag{1.113}$$

conocida como la *distancia de Hamming* entre las palabras binarias.

b) Si definimos el conjunto $\bar{B}_k(s_1^n) = \{x_1^n \in \{0,1\}^n : d_H(x_1^n; s_1^n) \leq k\} \subset \{0,1\}^n$ para todo $k \in \{0, ..., n\}$, determine una expresión para.

$$\eta_k = P_{X_1^n|S_1^n}(\bar{B}_k(s_1^n)|s_1^n). \tag{1.114}$$

De una interpretación a esta probabilidad del punto de vista del problema de transmitir s_1^n y recibir x_1^n. *Indicación*: Notar que $d_H(x_1^n; s_1^n) = k$ equivale a decir que hay k-bits donde x_1^n difiere de s_1^n. Puede ser útil, en primera instancia, considerar el conjunto

$$\bar{A}_k(s_1^n) = \{x_1^n \in \{0,1\}^n : d_H(x_1^n; s_1^n) = k\} \tag{1.115}$$

$\subset \{0,1\}^n$ y determinar

$$\varsigma_k = P_{X_1^n|S_1^n}(\bar{A}_k(s_1^n)|s_1^n). \tag{1.116}$$

c) Considere que tenemos dos hipótesis, $\Theta = \{0,1\}$, y que dado $\theta = 0$ entonces se transmite $(0, 0, ..., 0) \in \{0,1\}^n$ y que dado $\theta = 1$ se transmite $(1, 1, ..., 1) \in \{0,1\}^n$. Utilice el Lema de Neyman-Pearson para determinar la forma de la familia de test óptimos en este problema. *Indicación*: Notar que en este caso la función de verosimilitud se construye como:

$$L(x_1, ..., x_n|\theta) = P_{X_1^n|S_1^n}(X_1^n = x_1^n|\theta, \theta, ..., \theta). \tag{1.117}$$

d) Restrinja el análisis al conjunto de decisión

$$A_1 = \left\{ x_1^n \in \{0,1\}^n : \frac{L(x_1, ..., x_n|\theta = 1)}{L(x_1, ..., x_n|\theta = 0)} > v \right\} \tag{1.118}$$

de parámetro v. Verifique que este conjunto esta dado por la regla de mínima distancia, es decir $(x_1, ..., x_n) \in A_1$ si, y solo si,

$$d_H((x_1, ..., x_n); (1, ..., 1)) < d_H((x_1, .., x_n); (0, 0, .., 0)) + \tau(v, \epsilon), \tag{1.119}$$

y determine la expresión de $\tau(v, \epsilon) \in \mathbb{R}$, función de v y ϵ, con $\epsilon < \frac{1}{2}$. Repita el mismo análisis y determine los conjuntos

$$A_0 = \left\{ x_1^n \in \{0,1\}^n : \frac{L(x_1, ..., x_n|\theta = 1)}{L(x_1, ..., x_n|\theta = 0)} < v \right\} \tag{1.120}$$

$$A_2 = \left\{ x_1^n \in \{0,1\}^n : \frac{L(x_1,...,x_n|\theta=1)}{L(x_1,...,x_n|\theta=0)} = v \right\} \tag{1.121}$$

como función de la regla de minina distancia sugerida en (1.119). *Indicación*: Utilice lo obtenido en el punto a).

e) Considere n par, $v = 1$ y $\mathbb{P}(\rho(w) = 1) = 0{,}5$. Muestre primero que $\tau(v = 1, \epsilon < 0{,}5) = 0$.

- Verifique que en el caso n par, $A_2 \neq \emptyset$, caracterice el conjunto y determine su cardinalidad.
- Encuentre expresiones para el tamaño y el poder del test.

Indicación: Será de gran utilidad obtener las expresiones obtenidas en (1.114) y (1.116). En particular, asocie los conjuntos A_0, A_1 y A_2 a los conjuntos $\bar{B}_k(s_1^n)$ y $\bar{A}_k(s_1^n)$ del punto b).

f) (PENDIENTE)[17] Consideremos el problema del punto c), pero en un contexto Bayesiano, done $p_\Theta(0) = p_\Theta(1) = 0{,}5$. Determine el test Bayesiano óptimo para la función de costo 0-1 (es decir $L_{0,0} = L_{1,1} = 0$ y $L_{1,0} = L_{0,1} = 1$) y verifique formalmente que la solución esta dada por la siguiente estructura:

$$\pi^*(x_1^n) = \begin{cases} 1 & \text{si } d_H((x_1,...,x_n);(1,...,1)) < d_H((x_1,...,x_n);(0,...,0)) \\ 0 & \text{si } d_H((x_1,...,x_n);(1,...,1)) > d_H((x_1,...,x_n);(0,...,0)) \\ I & \text{si } d_H((x_1,...,x_n);(1,...,1)) = d_H((x_1,...,x_n);(0,...,0)) \end{cases} \tag{1.122}$$

donde I denota indiferencia, es decir, 1 o 0.

Problema 1.6. El caso de distribuciones Gaussianas[18] es emblemático tanto por su simplicidad analítica, como por su amplio uso como modelo de observación, en particular en problemas de comunicaciones digitales y reconocimiento de patrones. Consideremos $\mathbb{X} = \mathbb{R}^n$, $\Theta = \{0,1\}$ y una secuencia $X_1, ..., X_n$ de variables aleatorias independientes e

[17] Este problema requiere conocimientos del contexto Bayesiano.

[18] Si X sigue una distribución normal de parámetros μ, σ^2 entonces su función de densidad de probabilidad es $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $x \in \mathbb{R}$

idénticamente distribuidas (i.i.d.) para un cierto $n \in \mathbb{N}$ bajo las siguientes hipótesis:

$$H_0 : \theta = 0 : X \sim N(0,1) \rightarrow f_{X_1^n}(x_1, ..., x_n | \theta = 0) = \prod_{i=1}^{n} f_{X_i}(x_i | \theta = 0)$$

$$H_1 : \theta = 1 : X \sim N(0,\sigma^2) \rightarrow f_{X_1^n}(x_1, ..., x_n | \theta = 1) = \prod_{i=1}^{n} f_{X_i}(x_i | \theta = 1),$$

donde se asume que $\sigma^2 > 1$.

a) Sea $n \in \mathbb{N}$, plantee la familia de test óptimos en el sentido de Neyman-Pearson $\{\pi_\nu(\cdot) : \nu \in \mathbb{R}^+\}$. Argumente brevemente que en este caso la familia de test óptimos es determinística. En particular, verifique que para un $\nu \in \mathbb{R}^+$ arbitrario el test tiene la siguiente forma cerrada:

$$\pi_\nu^{-1}(\{1\}) = \left\{ x \in \mathbb{X} : \sum_{i=1}^{n} x_i^2 \geq \gamma \right\},$$

con $\gamma = \gamma(\nu, \sigma^2, n)$ función de ν, σ^2, n que se debe determinar explícitamente.

b) Considere el caso $n = 2$ y $\nu \geq 1$. Para un $\nu \geq 1$ fijo calcule de forma cerrada el tamaño del test α_π y poder del test β_π. Puede usar el hecho que si $X_i \sim N(0, \sigma^2)$, $i \in \{1, 2\}$ independientes entonces la variable aleatoria $Y = \sum_{i=1}^{2} \frac{X_i^2}{\sigma^2}$, tiene función de densidad de probabilidad:

$$f_Y(y) = \begin{cases} \frac{1}{2} e^{-y/2} & \text{para } y > 0 \\ 0 & \text{para } y \leq 0 \end{cases}$$

c) Para el caso $n = 2$, demuestre que el desempeño de la familia de test óptimos esta dado por la siguiente curva ROC

$$\beta_\pi = f_{ROC}(\alpha_\pi) = \alpha_\pi^{1/\sigma^2}$$

Indicación: Compare los resultados de la parte b) y despeje ν.

Problema 1.7. El caso de distribuciones exponenciales[19] es emblemático tanto por su simplicidad analítica, así como por su uso como modelamiento del tiempo de falla en sistemas complejos. Consideremos $\mathbb{X} = \mathbb{R}^+$, $\Theta = \{0, 1\}$ y una observación $X = x$ a valores en $\mathbb{R}^+$ bajo las siguientes hipótesis:

$$\begin{aligned} H_0 :&\theta = 0 : X \sim exp(\lambda_0) \to f_X(x|\theta = 0) \\ H_1 :&\theta = 1 : X \sim exp(\lambda_1) \to f_X(x|\theta = 1), \end{aligned} \tag{1.123}$$

donde se asume que $\lambda_1 > \lambda_0 > 0$.

a) Plantee la familia de test óptimos en el sentido de Neyman-Pearson $\{\pi_\tau(\cdot) : \tau \in \mathbb{R}^+\}$ (no considere los test de tamaño 0 ni 1). Argumente que en este caso la familia de test óptimos es determinística. En particular, verifique que para un $\tau \in \mathbb{R}^+$ arbitrario el test tiene la siguiente forma cerrada:

$$\pi_\tau^{-1}(\{1\}) = \{x \in \mathbb{X} : x \leq \gamma\}, \tag{1.124}$$

con $\gamma = \gamma(\tau, \lambda_0, \lambda_1)$ función de $\tau, \lambda_0, \lambda_1$ que se debe determinar explícitamente.

b) Considere $\tau < \frac{\lambda_1}{\lambda_0}$ y calcule de forma cerrada el tamaño del test α_π y poder del test β_π.

c) Demuestre que el desempeño de la familia de test óptimos esta dado por la siguiente curva ROC

$$\beta_\pi = f_{ROC}(\alpha_\pi) = 1 - (1 - \alpha_\pi)^{\lambda_1/\lambda_0} \tag{1.125}$$

Problema 1.8. Considere $\mathbb{X} = \mathbb{R}$, una variable aleatoria X continua y las siguientes hipótesis

$$\begin{aligned} H_0 :&\theta = 0 : X \sim P_X(\cdot|\theta_0) \to f_X(x|\theta = 0) \\ H_1 :&\theta = 1 : X \sim P_X(\cdot|\theta_0) \to f_X(x|\theta = 1), \end{aligned} \tag{1.126}$$

Además, se sabe que existe un conjunto $A \subset \mathbb{R}$ tal que:

$$(\forall x \in A)(f_X(x|\theta = 1) > 0 \wedge f_X(x|\theta = 0) = 0) \ \wedge \ (\forall x \in A^c)(f_X(x|\theta = 1) = 0 \wedge f_X(x|\theta = 0) > 0) \tag{1.127}$$

[19] Si X sigue una distribución exponencial de parámetro λ entonces su f.d.p es $f_X(x) = \lambda e^{-\lambda x}$, $x > 0$

Demuestre que existe un test óptimo perfecto, es decir, que existe un test con tamaño $\alpha_\pi = 0$ y poder $\beta_\pi = 1$. Indique claramente la forma de este test, verifique que $\alpha_\pi = 0$ y $\beta_\pi = 1$ y comente por qué en este caso es posible obtener este resultado.

2

Unidad II: Detección Bayesiana

A diferencia del problema de detección paramétrica, la teoría Bayesiana ofrece un nueva visión en el sentido que requiere aceptar la premisa de que la naturaleza especificó una distribución de probabilidad sobre Θ conocida como distribución *a priori* o *prior*. Esto en un principio puede ser objeto de controversia ya que se requiere conocer tal distribución.

Desde un punto de vista subjetivista, no es necesario creer que la naturaleza realmente escogió un estado θ de acuerdo a una distribución a priori; más bien, la distribución a priori se puede ver simplemente como un reflejo de la creencia de la decisión hecha por el agente o detector respecto al verdadero estado. La adquisición de datos (evidencia), usualmente observaciones, actúa de forma tal que modifica la creencia del agente sobre dicho estado de la naturaleza. De hecho, demostraremos más adelante, que toda regla de decisión *buena* es esencialmente una regla Bayesiana con alguna distribución a priori.

En esta unidad, entonces, la variable a inferir Θ se modela como una variable aleatoria en un conjunto finito que depende (estadísticamente) de la observación x, por lo que deja de ser un parámetro a diferencia del caso de detección paramétrico. Esto entrega una flexibilidad mayor ya que es posible modelar el problema usando la Teoría de Bayes.

2.1. Formalización del Problema de Detección Bayesiano

En el contexto Bayesiano Θ se modela como una variable aleatoria con distribución P_Θ en $\mathcal{A} = \{1, ..., k\}$, $k \in \mathbb{N}$, (es decir a valores finitos), P_Θ se le llama distribución a priori y e induce una función de probabilidad de masa. En este contexto tenemos que, dado $\Theta = \theta$ se tiene una probabilidad condicional de la variable aleatoria X que está dada por:

$$\mathbb{P}(X(w) \in B | \Theta(w) = \theta). \tag{2.1}$$

Alternativamente esta probabilidad se puede caracterizar por la distribución inducida $P_{X|\Theta}(B|\theta)$ en $\mathbb{X}$. De esta forma tenemos que $\forall B \subseteq \mathbb{X}$:

$$\begin{aligned} \mathbb{P}(X(w) \in B, \Theta(w) = \theta) &= P_{X,\Theta}(B, \{\theta\}) \\ &= \underbrace{P_\Theta(\{\theta\}) \cdot P_{X|\Theta}(B|\{\theta\})}_{\text{Regla de Probabilidad Condicional}} . \end{aligned} \tag{2.2}$$

Típicamente X toma valores en $\mathbb{X} = \mathbb{R}^n$, $n \in \mathbb{N}$ y Θ toma valores en $\mathcal{A} = \{1, ..., k\}$, por lo tanto, $p_\Theta(\theta) \triangleq P_\Theta(\Theta = \theta) = P_\Theta(\{\theta\})$ denota la función de probabilidad de masa de Θ. Por otro lado se tiene:

$$P_{X|\Theta}(B|\Theta = \theta) = \int_B f_{X|\Theta}(x|\theta)dx \tag{2.3}$$

donde $f_{X|\Theta}(x|\theta)$ denota la función de densidad de probabilidad condicional de X dado $\Theta = \theta$. Similarmente:

$$P_{X|\Theta}(B|\Theta = \theta) = \sum_{x \in B} p_{X|\Theta}(x|\theta), \tag{2.4}$$

donde $p_{X|\Theta}(x|\theta)$ es la función de masa condicional. Finalmente la distribución del vector conjunto (X, Θ) queda determinada por:

$$\begin{aligned} \mathbb{P}(X(w) \in B, \Theta(w) = \theta) &= P_{X,\Theta}(B, \{\theta\}) \\ &= p_\Theta(\theta) \cdot \int_B f_{X|\Theta}(x|\theta)dx, \end{aligned} \tag{2.5}$$

o bien si es discreto,

$$\begin{aligned} \mathbb{P}(X(w) \in B, \Theta(w) = \theta) &= P_{X,\Theta}(B, \{\theta\}) \\ &= p_\Theta(\theta) \cdot \sum_{x \in B} p_{X|\Theta}(x|\theta), \end{aligned} \tag{2.6}$$

para todo $B \subset \mathbb{X}$ y $\theta \in \mathcal{A}$.

Con este breve repaso, ahora podemos introducir los elementos que componen un problema de detección Bayesiano.

- Un espacio de observación $\mathbb{X}$ y variables aleatorias que toman valores en $\mathbb{X}$. $X = x$ se conoce como observación o dato. $\mathbb{X}$ es un espacio numérico abstracto y también puede ser multidimensional, por ejemplo, $\mathbb{X} = \mathbb{R}^n$ con $n \in \mathbb{N}$ en cuyo caso las observaciones x_1^n provienen de un vector aleatorio $X_1^n \in \mathbb{X}$.
- Un espacio de decisión $\mathcal{A}$ finito o numerable y una variable aleatoria Θ con valores en $\mathcal{A}$.
- Distribuciones de probabilidad condicionales indexadas por $\theta \in \Theta$, es decir, $P_X(\cdot|\Theta = \theta)$, $\theta \in \mathcal{A}$. Además se posee una distribución de probabibilidad sobre Θ, $P_\Theta(\cdot)$ la cual se conocerá como distribución *a priori* o *prior*. La distribución a priori *debe*ser discreta, por lo que está dotado de una función de probabilidad de masa $p_\Theta(\cdot)$
- Una regla, detector o test $\pi : \mathbb{X} \mapsto \mathcal{A}$ que será la función que tomará una decisión en base a algún criterio.
- Una función de costo o riesgo $L : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ que penaliza la incorrecta decisión. En adelante asumiremos que $L : \mathcal{A} \times \mathcal{A} \to \mathbb{R}^+ \cup \{0\}$.

Hablaremos más en detalle del riesgo ya que es un elemento nuevo respecto al caso paramétrico.

2.2. Riesgo Promedio Bayesiano

Para continuar con la formalización del problema de detección Bayesiano, se debe buscar la regla óptima respecto a un criterio, este criterio no es más que el promedio del costo dado por la variable aleatoria del riesgo $L(\Theta, \pi(X))$. Las siguientes definiciones ayudarán a establecer el criterio de optimalidad en el sentido Bayesiano:

Definición 2.1. (Riesgo Promedio) Consideremos una función de riesgo: $L : \mathcal{A} \times \mathcal{A} \to \mathbb{R}^+ \cup \{0\}$ que penaliza los errores en la toma de decisión y una regla de decisión: $\pi : \mathbb{X} \to \mathcal{A}$. Dado un θ que determina las estadísticas de las fuentes de observaciones $X \sim P_{X|\Theta}(\cdot|\Theta = \theta)$, definimos el riesgo promedio $R : \mathcal{A} \times F(\mathbb{X}, \mathcal{A}) \to \mathbb{R}^+ \cup \{0\}$ condicionado a θ como:

$$R(\theta, \pi) \triangleq \mathbb{E}(L(\theta, \pi(X))|\Theta = \theta) = \begin{cases} \underbrace{\int_{\mathbb{X}} L(\theta, \pi(x)) f_{X|\Theta}(x|\theta) dx}_{\text{Caso espacio continuo con f.d.p condicional}} \\ \underbrace{\sum_{x \in \mathbb{X}} L(\theta, \pi(x)) p_{X|\theta}(x|\theta)}_{\text{Caso espacio discreto con f.p.m condicional}} \end{cases} \tag{2.7}$$

Para asegurar que este riesgo está bien definido, nos restringiremos al conjunto de reglas o test tal que el riesgo $R(\theta, \pi)$ existe y es finito para todo $\theta \in \mathcal{A}$. La expresión (2.7) está condicionada a una realización de Θ. Por lo tanto $R(\Theta, \pi)$ es una variable aleatoria (función de Θ y π). Para cada regla π podemos definir su promedio respecto a θ, llamado Riesgo Promedio Bayesiano:

Definición 2.2. (Riesgo Promedio Bayesiano) Sea, $\pi \in F(\mathbb{X}, \mathcal{A})$, una distribución $P_\Theta(\cdot)$ discreta y su riesgo promedio $R(\theta, \pi)$. Definimos el Riesgo Promedio Bayesiano como el promedio de $R(\Theta, \pi)$ con respecto a la variable Θ (asumiremos el caso continuo para X dado Θ, el caso discreto es análogo)[1]:

$$\begin{aligned} r(\pi) &\triangleq \mathbb{E}_\Theta(R(\Theta, \pi)) \\ &= \sum_{\theta \in \mathcal{A}} R(\theta, \pi) \cdot p_\Theta(\theta) \\ &= \sum_{\theta \in \mathcal{A}} \mathbb{E}\left(L(\theta, \pi(X)) | \Theta = \theta\right) \cdot p_\Theta(\theta) \\ &= \sum_{\theta \in \mathcal{A}} p_\Theta(\theta) \cdot \int_{\mathbb{X}} L(\theta, \pi(x)) f_{X|\Theta}(x|\theta) dx \\ &= \sum_{\theta \in \mathcal{A}} \int_{\mathbb{X}} L(\theta, \pi(x)) \cdot p_\Theta(\theta) \cdot f_{X|\Theta}(x|\theta) dx \\ &= \sum_{\theta \in \mathcal{A}} \int_{\mathbb{X}} L(\theta, \pi(x)) \cdot \underbrace{f_{X,\Theta}(x, \theta)}_{\text{densidad conjunta}} dx \\ &= \mathbb{E}_{X,\Theta}(L(\Theta, \pi(X))). \end{aligned} \tag{2.8}$$

Con las definiciones anteriores ya podemos establecer una regla concreta óptima en el sentido Bayesiano.

2.3. Decisión Óptima: Distribución a Posteriori

Recapitulando, la regla óptima Bayesiana dependerá de los siguientes elementos previamente introducidos:

[1] Se debe tener un poco de cuidado, formalmente $f_{X,\Theta}(x, \theta)$ no es una densidad como tal ya que Θ no es continua, esta densidad en realidad es mixta pero adoptaremos este leve abuso de notación.

i) P_Θ distribución a priori dotado de una función de probabilidad de masa.
ii) $f_{X|\Theta}(\cdot|\theta)$, función de densidad de probabilidad condicional (o de masa según sea el caso).
iii) $L : \mathcal{A} \times \mathcal{A} \to \mathbb{R}^+$, función de costo.

Luego, la solución del problema de detección Bayesiana es aquella función que minimiza el riesgo promedio Bayesiano, es decir, se plantea como:

$$\begin{aligned}\pi^* &= \arg \min_{\pi \in F(\mathbb{X},\mathcal{A})} r(\pi) \\ &= \arg \min_{\pi \in F(\mathbb{X},\mathcal{A})} \mathbb{E}_{X,\Theta}(L(\Theta, \pi(X))). \end{aligned} \tag{2.9}$$

Por lo tanto, π^* es la regla que minimiza el riesgo Bayesiano. Esta expresión en principio no garantiza unicidad y además no da una manera cerrada de poder encontrar tal regla. Luego, si analizamos de forma más detallada la función objetivo en (2.9) tenemos lo siguiente:

$$\begin{aligned}\mathbb{E}_{X,\Theta}(L(\Theta, \pi(X))) &= \sum_{\theta \in \mathcal{A}} \int_{\mathbb{X}} L(\theta, \pi(x)) f_{X,\Theta}(x, \theta) dx \\ &= \int_{\mathbb{X}} \left[\sum_{\theta \in \mathcal{A}} L(\theta, \pi(x)) P_{\Theta|X}(\Theta = \theta|x) \right] f_X(x) dx. \end{aligned} \tag{2.10}$$

Se puede notar que el término $\sum_{\theta \in \mathcal{A}} L(\theta, \pi(x)) P_{\Theta|X}(\Theta = \theta|x)$ es función exclusiva de la evaluación de $\pi(\cdot)$ en el punto x y no de los restantes valores $\pi(y)$ que adopta en $y \in \mathbb{X}\backslash\{x\}$. Por lo tanto, minimizar (2.9) equivale a minimizar el argumento de la función (2.10) punto a punto, es decir, para cualquier observación arbitraria o $\forall x \in \mathbb{X}$, $\pi^*(x)$ es solución de:

$$\begin{aligned}\pi^*(x) &= \arg \min_{y \in \mathcal{A}} \sum_{\theta \in \mathcal{A}} L(\theta, y) P_{\Theta|X}(\Theta = \theta|x), \quad \forall x \in \mathbb{X} \\ &= \arg \min_{y \in \mathcal{A}} \mathbb{E}\left(L(\Theta, y)|X = x \right), \quad \forall x \in \mathbb{X}. \end{aligned} \tag{2.11}$$

Observaciones 2.1. Interpretando la regla óptima Bayesiana en (2.11), dada una observación x, $\pi(x)$ es la decisión que minimiza el riesgo promedio, respecto a la distribución a posteriori de Θ dado el evento $X = x$.

Por Bayes sabemos que la distribución a posteriori se obtiene como:

$$P_{\Theta|X}(\Theta = \theta|x) = \frac{f_{\Theta,X}(\theta, x)}{f_X(x)} = \frac{f_{X|\Theta}(x|\theta) p_\Theta(\theta)}{\sum_{\tilde{\theta} \in \mathcal{A}} f_{X|\Theta}(x|\tilde{\theta}) p_\Theta(\tilde{\theta})} \tag{2.12}$$

donde

$$f_X(x) = \sum_{\tilde{\theta}\in\mathcal{A}} f_{X,\Theta}(x,\tilde{\theta}) = \sum_{\tilde{\theta}\in\mathcal{A}} f_{X|\Theta}(x|\tilde{\theta})p_\Theta(\tilde{\theta}). \tag{2.13}$$

se obtuvo mediante el uso de probabilidades totales. De esta manera la regla de decisión óptima es solución de

$$\pi^*(x) = \arg\min_{y\in\mathcal{A}} \sum_{\theta\in\mathcal{A}} L(\theta,y) \frac{f_{X|\Theta}(x|\theta)p_\Theta(\theta)}{\sum\limits_{\tilde{\theta}\in\mathcal{A}} f_{X|\Theta}(x|\tilde{\theta})p_\Theta(\tilde{\theta})}, \quad \forall x \in \mathbb{X}. \tag{2.14}$$

Además, podemos notar que $f_X(x)$ es constante para cualquier elección de $y \in \mathcal{A}$, entonces la regla en (2.14) se puede escribir como:

$$\pi^*(x) = \arg\min_{y\in\mathcal{A}} \sum_{\theta\in\mathcal{A}} L(\theta,y) f_{X|\Theta}(x|\theta)p_\Theta(\theta), \quad \forall x \in \mathbb{X}. \tag{2.15}$$

La expresión (2.15) tiene la ventaja de ser general, pero a su vez difícil de manejar, veremos una función de costo particular que reduce el problema significativamente.

2.3.1. Función de costo $L_{0,1}$

Consideraremos el caso especial de la función de costo 0-1 en (2.16). Ésta juega un rol central en problemas de reconocimiento de patrones y comunicaciones digitales pues su costo promedio equivale a la probabilidad de error de decisión. La función de costo 0-1 esta dada por:

$$L_{0,1}(x,y) = \begin{cases} 0 & \text{si } x = y \\ 1 & \text{si } x \neq y \end{cases} \quad \forall x,y \in \mathcal{A} \tag{2.16}$$

Notar que el costo es simétrico y penaliza con el mismo valor el evento de error. Dada una regla π y un valor $\theta \in \mathcal{A}$ tenemos que el riesgo promedio condicional de la función $L_{0,1}$ es:

$$\begin{aligned} R_{0,1}(\theta,\pi) &= \mathbb{E}_X(L_{0,1}(\theta,\pi(X))|\Theta = \theta) \\ &= \int_{\mathbb{X}} L_{0,1}(\theta,\pi(x)) f_{X|\Theta}(x|\theta)dx. \end{aligned} \tag{2.17}$$

Como se desarrolló anteriormente en el caso paramétrico, sabemos que la regla π particiona el espacio de observación. Podemos definir la partición inducida por la regla π como $\{A_1,...,A_k\}$ donde tenemos que:

$$A_\theta = \pi^{-1}(\{\theta\}) \subset \mathbb{X} \quad \forall \theta \in \{1,...,k\} = \mathcal{A}. \tag{2.18}$$

Por definición se puede verificar que:

$$\forall x \in A_\theta \;\; L(\theta, \pi(x)) = 0 \tag{2.19}$$

$$\forall x \notin A_\theta \;\; L(\theta, \pi(x)) = 1, \tag{2.20}$$

por lo tanto se puede escribir la función de costo mediante la siguiente indicatriz:

$$L(\theta, \pi(x)) = \mathbb{1}_{A_\theta^c}(x). \tag{2.21}$$

Con esta identidad y gracias a la propiedad de la esperanza, tenemos que:

$$\begin{aligned} R_{0,1}(\theta, \pi) &= \int_{\mathbb{X}} \mathbb{1}_{A_\theta^c}(x) \cdot f_{X|\Theta}(x|\theta) dx \\ &= \int_{A_\theta^c} f_{X|\Theta}(x|\theta) dx \\ &= P_{X|\Theta}(A_\theta^c | \Theta = \theta) \qquad (2.22) \\ &= P_{X|\Theta}(X \in A_\theta^c | \theta) \\ &= P_{X|\Theta}(\pi(X) \neq \theta | \theta). \qquad (2.23) \end{aligned}$$

Observaciones 2.2. De la expresión (2.23) $R_{0,1}(\theta, \pi)$ representa **la probabilidad de error** de la regla π bajo la hipótesis $\Theta = \theta$.

La función de costo promedio 0-1 de la regla π es: .

$$\begin{aligned} r_{0,1}(\pi) &= \mathbb{E}_{X,\Theta}\{L_{0,1}(\Theta, \pi(X))\} \\ &= \sum_{\theta=1}^{k} p_\Theta(\theta) \cdot R_{0,1}(\theta, \pi) \\ &= \sum_{\theta=1}^{k} p_\Theta(\theta) \cdot P_{X|\Theta}(A_\theta^c | \Theta = \theta) \quad \text{de (2.22)} \\ &= P_{X,\Theta}\left(\bigcup_{\theta \in \mathcal{A}} A_\theta^c \times \{\theta\} \right) \\ &\underbrace{=}_{\text{definición de } \pi} P_{X,\Theta}(\{(x, \theta) \in \mathbb{X} \times \mathcal{A} : \pi(x) \neq \theta\}). \qquad (2.24) \end{aligned}$$

Alternativamente:

$$\begin{aligned} r_{0,1}(\pi) &= \sum_{\theta=1}^{k} p_\Theta(\theta) \cdot P_{X|\Theta}(A_\theta^c|\theta) \\ &= \sum_{\theta=1}^{k} p_\Theta(\theta) \cdot P_{X|\Theta}(\pi(X) \neq \theta|\Theta = \theta) \quad \text{de (2.23)} \\ &= \sum_{\theta=1}^{k} P_{X,\Theta}(\pi(X) \neq \theta, \Theta = \theta) \\ &= P_{X,\Theta}(\pi(X) \neq \Theta). \end{aligned} \tag{2.25}$$

Observaciones 2.3. La función de costo promedio $r_{0,1}(\pi)$ es la probabilidad de error de π respecto a la distribución conjunta de (X, Θ), ver (2.25). Por lo tanto $r_{0,1}(\pi)$ se entiende como la probabilidad de incorrecta clasificación. Del punto de vista de cómputo este valor es el promedio de los valores $\{R_{0,1}(\theta, \pi) : \theta \in \mathcal{A}\}$ con respecto a la distribución a priori de Θ, es decir:

$$P_{\text{error}}(\pi) = r_{0,1}(\pi) = \sum_{\theta=1}^{k} p_\Theta(\theta) \cdot R_{0,1}(\theta, \pi). \tag{2.26}$$

Vemos entonces que considerando la función de costo $L_{0,1}$ el riesgo promedio equivale a minimizar la probabilidad de error, respecto a la regla óptima en (2.11) tenemos que se reduce a:

$$\begin{aligned} \pi_{0,1}^*(x) &= \arg\min_{y \in \mathcal{A}} \sum_{\theta \in \mathcal{A}} L_{0,1}(\theta, y) P_{\Theta|X}(\Theta = \theta|x) \\ &= \arg\min_{y \in \mathcal{A}} \sum_{\theta \in \mathcal{A}\ \theta \neq y} P_{\Theta|X}(\Theta = \theta|x) \\ &= \arg\min_{y \in \mathcal{A}} P_{\Theta|X}(\mathcal{A} \backslash \{y\}|x) \\ &= \arg\min_{y \in \mathcal{A}} 1 - P_{\Theta|X}(\Theta = y|x) \\ &= \arg\max_{y \in \mathcal{A}} P_{\Theta|X}(\Theta = y|x), \end{aligned} \tag{2.27}$$

es decir, cuando la función de costo es $L_{0,1}$ la regla Bayesiana óptima $\pi_{0,1}^*(x)$ corresponde al criterio de **maximizar la probabilidad a posteriori** o regla MAP (*maximum a*

posteriori). Es posible seguir trabajando la expresión gracias a la regla de Bayes, con lo que:

$$\begin{aligned}\pi^*_{0,1}(x) &= \arg\max_{\theta\in\mathcal{A}} P_{\Theta|X}(\Theta=\theta|x)\\ &= \arg\max_{\theta\in\mathcal{A}} \frac{f_{\Theta,X}(\theta,x)}{f_X(x)}\\ &= \arg\max_{\theta\in\mathcal{A}} f_{\Theta,X}(\theta,x)\\ &= \arg\max_{\theta\in\mathcal{A}} f_{X|\Theta}(x|\theta)\cdot p_\Theta(\theta). \end{aligned} \tag{2.28}$$

Un caso particular a considerar es cuando $p_\Theta(\theta) = \frac{1}{|\mathcal{A}|}$ (distribución a priori equiprobable), se tiene que:

$$\pi^*_{0,1}(x) = \arg\max_{\theta\in\mathcal{A}} f_{X|\Theta}(x|\theta) \tag{2.29}$$

que corresponde al criterio de **máxima verosimilitud** o ML (*maximum likelihood*).

2.3.2. Función de costo cuadrático

Otra función de costo muy utilizada es la función de costo cuadrática $L_{MSE} : \mathcal{A}\times\Delta \to \mathbb{R}^+ \cup \{0\}$ (con $\Delta \subset \mathbb{R}$) definida como:

$$L_{MSE}(x,y) = (x-y)^2 \tag{2.30}$$

Podemos notar que al usar esta función de costo, el espacio de parámetro es distinto al de decisión, lo que será un caso extraordinario que no respeta la hipótesis simplificatoria, esto porque como veremos más adelante, la regla óptima en este caso entregará un valor aproximado, lo que se puede reinterpretar como un valor estimado[2]. El riesgo Bayesiano es:

$$\mathbb{E}_{X,\Theta}(L_{MSE}(\Theta,\pi(X))) = \sum_{\theta\in\mathcal{A}}\int_{\mathbb{X}} (\theta-\pi(x))^2 f_{X,\Theta}(x,\theta)dx. \tag{2.31}$$

La ecuación en (2.31) se conoce como el error cuadrático medio o *Mean Square Error*. Vamos a utilizar la expresión en (2.11) para encontrar la regla óptima y consideremos el siguiente operador:

$$\mathbb{E}(\Theta|X=x) = \sum_{\theta\in\mathcal{A}} \theta P_{\Theta|X}(\Theta=\theta|x), \tag{2.32}$$

[2] Veremos esto con más detalle en la Unidad 4

que corresponde a la esperanza condicional de Θ dado $X = x$, entonces, el argumento en (2.11) lo podemos descomponer como:

$$\begin{aligned}
\sum_{\theta\in\mathcal{A}}(\theta-y)^2 P_{\Theta|X}(\Theta=\theta|x) &= \sum_{\theta\in\mathcal{A}}(\theta-y)^2 P_{\Theta|X}(\Theta=\theta|x) \\
&= \sum_{\theta\in\mathcal{A}}(\theta-\mathbb{E}(\Theta|X=x)+\mathbb{E}(\Theta|X=x)-y)^2 P_{\Theta|X}(\Theta=\theta|x) \\
&= \sum_{\theta\in\mathcal{A}}(\theta-\mathbb{E}(\Theta|X=x))^2 P_{\Theta|X}(\Theta=\theta|x) + \sum_{\theta\in\mathcal{A}}(\mathbb{E}(\Theta|X=x)-y)^2 P_{\Theta|X}(\Theta=\theta|x) \\
&\quad + 2(\mathbb{E}(\Theta|X=x)-y)\sum_{\theta\in\mathcal{A}}(\theta-\mathbb{E}(\Theta|X=x))P_{\Theta|X}(\Theta=\theta|x) \quad \to 0 \\
&= \sum_{\theta\in\mathcal{A}}(\theta-\mathbb{E}(\Theta|X=x))^2 P_{\Theta|X}(\Theta=\theta|x) + (\mathbb{E}(\Theta|X=x)-y)^2
\end{aligned} \tag{2.33}$$

Podemos notar que

$$\sum_{\theta\in\mathcal{A}}(\theta-\mathbb{E}(\Theta|X=x))^2 P_{\Theta|X}(\Theta=\theta|x) = Var(\Theta|X=x) \tag{2.34}$$

es la varianza condicional de Θ dado $X = x$. Por lo tanto:

$$\begin{aligned}
\pi^*(x) &= \arg\min_{y\in\Delta} Var(\Theta|X=x) + (\mathbb{E}(\Theta|X=x)-y)^2 \\
&= \arg\min_{y\in\Delta}(\mathbb{E}(\Theta|X=x)-y)^2 \\
&= \mathbb{E}(\Theta|X=x).
\end{aligned} \tag{2.35}$$

La última igualdad es evidente a partir del hecho que si tomamos $y = \mathbb{E}(\Theta|X = x)$ la función $(\mathbb{E}(\Theta|X = x) - y)^2$ es mínima tomando valor 0. Luego, el detector óptimo que minimiza el error cuadrático medio corresponde a:

$$\pi_{MMSE}(x) = \mathbb{E}(\Theta|X=x) = \sum_{\theta\in\mathcal{A}}\theta P_{\Theta|X}(\Theta=\theta|x), \tag{2.36}$$

que es la esperanza condicional o la esperanza de la distribución a posteriori de Θ dado $X = x$.

Finalmente el riesgo Bayesiano mínimo o error cuadrático medio mínimo (MMSE)

está dado por la siguiente expresión:

$$\begin{aligned} MMSE &= \min_{\pi:\mathbb{X}\to\mathcal{A}} \mathbb{E}_{\Theta,X}(L(\Theta,\phi(X))) \\ &= \min_{\phi:\mathbb{X}\to\mathcal{A}} \mathbb{E}_{\Theta,X}((\Theta-\phi(X))^2) \\ &= \int_{\mathbb{X}} \left[\sum_{\theta\in\mathcal{A}} (\theta - \mathbb{E}(\Theta|X=x))^2 P_{\Theta|X}(\Theta=\theta|x)\right] f_X(x)dx \\ &= \int_{\mathbb{X}} Var(\Theta|X=x) f_X(x)dx \\ &= \mathbb{E}(Var(\Theta|X)), \end{aligned} \tag{2.37}$$

que corresponde al promedio de la varianza condicional.

2.4. Relación con el Lema de Neyman-Pearson

En esta sección veremos la conexión del test Bayesiano óptimo con el Lema de Neyman-Pearson, para esto, consideremos el caso $\mathcal{A} = \{0, 1\}$ (caso binario), y una función de costo arbitraria de la forma:

$\mathcal{A}$ \ $\mathcal{A}$	0	1
0	l_{00}	l_{01}
1	l_{10}	l_{11}

Nos encontramos con que la regla óptima es, para un $x \in \mathbb{X}$ y aplicando (2.11):

$$\begin{aligned} \pi^*(x) &= \arg\min_{y\in\mathcal{A}} \sum_{\theta\in\mathcal{A}} L(\theta,y) P_{\Theta|X}(\Theta=\theta|x), \\ &= \begin{cases} 1 & \text{si } \sum_{\theta\in\mathcal{A}} L(\theta,0) P_{\Theta|X}(\Theta=\theta|x) > \sum_{\theta\in\mathcal{A}} L(\theta,1) P_{\Theta|X}(\Theta=\theta|x) \\ 0 & \text{si } \sum_{\theta\in\mathcal{A}} L(\theta,0) P_{\Theta|X}(\Theta=\theta|x) < \sum_{\theta\in\mathcal{A}} L(\theta,1) P_{\Theta|X}(\Theta=\theta|x) \\ I & \text{si } \sum_{\theta\in\mathcal{A}} L(\theta,0) P_{\Theta|X}(\Theta=\theta|x) = \sum_{\theta\in\mathcal{A}} L(\theta,1) P_{\Theta|X}(\Theta=\theta|x) \end{cases} \end{aligned} \tag{2.38}$$

donde I indica indiferencia, es decir, 0 o 1. Observamos que:

$$\begin{aligned}
\sum_{\theta\in\mathcal{A}} L(\theta,0)P_{\Theta|X}(\Theta=\theta|x) &> \sum_{\theta\in\mathcal{A}} L(\theta,1)P_{\Theta|X}(\Theta=\theta|x)\\
L(0,0)P_{\Theta|X}(\Theta=0|x)+L(1,0)P_{\Theta|X}(\Theta=1|x) &> L(0,1)P_{\Theta|X}(\Theta=0|x)+L(1,1)P_{\Theta|X}(\Theta=1|x)\\
l_{10}P_{\Theta|X}(\Theta=1|x)-l_{11}P_{\Theta|X}(\Theta=1|x) &> l_{01}P_{\Theta|X}(\Theta=0|x)-l_{00}P_{\Theta|X}(\Theta=0|x)\\
(l_{10}-l_{11})P_{\Theta|X}(\Theta=1|x) &> (l_{01}-l_{00})P_{\Theta|X}(\Theta=0|x)\\
P_{\Theta|X}(\Theta=1|x) &> \frac{l_{01}-l_{00}}{l_{10}-l_{11}}\cdot P_{\Theta|X}(\Theta=0|x)\\
\frac{f_{X|\Theta}(x|\Theta=1)\cdot p_\Theta(1)}{f_X(x)} &> \frac{l_{01}-l_{00}}{l_{10}-l_{11}}\cdot\frac{f_{X|\Theta}(x|\Theta=0)\cdot p_\Theta(0)}{f_X(x)}\\
f_{X|\Theta}(x|\Theta=1) &> \frac{(l_{01}-l_{00})p_\Theta(0)}{(l_{10}-l_{11})p_\Theta(1)}\cdot f_{X|\Theta}(x|\Theta=0) \quad (2.39)
\end{aligned}$$

Y que el test entonces queda:

$$\pi^*(x)=\begin{cases} 1 & \text{si } f_{X|\Theta}(x|\Theta=1) > \frac{(l_{01}-l_{00})p_\Theta(0)}{(l_{10}-l_{11})p_\Theta(1)}\cdot f_{X|\Theta}(x|\Theta=0)\\ 0 & \text{si } f_{X|\Theta}(x|\Theta=1) < \frac{(l_{01}-l_{00})p_\Theta(0)}{(l_{10}-l_{11})p_\Theta(1)}\cdot f_{X|\Theta}(x|\Theta=0)\\ I & \text{si } f_{X|\Theta}(x|\Theta=1) = \frac{(l_{01}-l_{00})p_\Theta(0)}{(l_{10}-l_{11})p_\Theta(1)}\cdot f_{X|\Theta}(x|\Theta=0)\end{cases} \quad (2.40)$$

Observamos que la regla óptima es una instancia del test de Neyman-Pearson y el umbral es $\nu=\frac{(l_{01}-l_{00})p_\Theta(0)}{(l_{10}-l_{11})p_\Theta(1)}$. Por lo que la distribución a priori y los costos juegan un rol de ajuste del umbral y, por lo tanto, nuevamente un compromiso entre tamaño y poder del test. Si además particularizamos esto para la función de costo $L_{0,1}$ tenemos lo siguiente:

$$\pi^*_{0,1}(x)=\begin{cases} 1 & \text{si } f_{X|\Theta}(x|\Theta=1) > \frac{p_\Theta(0)}{p_\Theta(1)}\cdot f_{X|\Theta}(x|\Theta=0)\\ 0 & \text{si } f_{X|\Theta}(x|\Theta=1) < \frac{p_\Theta(0)}{p_\Theta(1)}\cdot f_{X|\Theta}(x|\Theta=0)\\ I & \text{si } f_{X|\Theta}(x|\Theta=1) = \frac{p_\Theta(0)}{p_\Theta(1)}\cdot f_{X|\Theta}(x|\Theta=0)\end{cases} \quad (2.41)$$

Sabemos que la función de riesgo promedio Bayesiano en este caso equivale a la probabilidad de error, por lo que:

$$\begin{aligned} r_{0,1}(\pi) &= P_{X,\Theta}(\pi(X) \neq \Theta) \\ &= \sum_{\theta=0}^{1} P_{X|\Theta}(\pi(X) \neq \theta|\Theta = \theta) P_{\Theta}(\Theta = \theta) \\ &= P_{X|\Theta}(\pi(X) \neq 0|\Theta = 0) p_{\Theta}(0) + P_{X|\Theta}(\pi(X) \neq 1|\Theta = 1) p_{\Theta}(1) \\ &= \underbrace{P_{X|\Theta}(\pi(X) = 1|\Theta = 0)}_{\alpha_\pi} p_{\Theta}(0) + \underbrace{P_{X|\Theta}(\pi(X) = 0|\Theta = 1)}_{1-\beta_\pi} p_{\Theta}(1). \end{aligned} \tag{2.42}$$

Nos encontramos con que la probabilidad de error no es más que la combinación convexa del error de tipo I y el error de tipo II donde sus ponderadores están dados por la distribución a priori. La distribución a priori entonces indica cuanto peso o importancia se le dará a cada tipo de error, lo que nuevamente entrega un compromiso entre ambos tipos de errores, similar al test de Neyman-Pearson.

2.5. Medidas de Desempeño

En clasificación (detección) nos vemos enfrentados a evaluar el desempeño de las reglas propuestas respecto a otras reglas. Si bien tenemos los criterios óptimos (Máxima a Posteriori en detección Bayesiana o Lema de Neyman-Pearson en detección paramétrica), no siempre podemos acceder a las distribuciones de los datos o incluso el espacio de los parámetros puede ser suficientemente grande que hace inviable calcular la probabilidad de error de manera analítica. A continuación mostraremos distintas medidas de desempeño, comúnmente llamadas métricas[3], que son las más usadas en clasificación y hablaremos de sus usos y ventajas.

Estas medidas serán definidas en clasificación binaria pero son fácilmente extensibles al caso multiclase.

Definición 2.3. (Verdadero Positivo, Falso Positivo, Verdadero Negativo, Falso Negativo) Consideremos $\Theta = \{0, 1\}$ y un total de N observaciones, de las cuales N_0 corrresponde a

[3] Notar que no necesariamente son métricas en un sentido matemático ya que no representan distancias

la clase 0 y N_1 corresponden a la clase 1. Además consideremos una regla $r : \mathbb{X} \to \{0, 1\}$. Definimos:

- Verdadero Positivo (tp): $\sum_{i=1}^{N} \mathbb{1}_{\pi^{-1}(\{1\})\times\{1\}}(x_i, \theta) = \sum_{i=1}^{N_1} \mathbb{1}_{\pi^{-1}(\{1\})}(x_i) = \sum_{i=1}^{N_1} \mathbb{1}_{\{\pi(x_i)=1\}}(x_i)$
- Verdadero Negativo (tn): $\sum_{i=1}^{N} \mathbb{1}_{\pi^{-1}(\{0\})\times\{0\}}(x_i, \theta) = \sum_{i=1}^{N_0} \mathbb{1}_{\pi^{-1}(\{0\})}(x_i) = \sum_{i=1}^{N_0} \mathbb{1}_{\{\pi(x_i)=0\}}(x_i)$
- Falso Positivo (fp): $\sum_{i=1}^{N} \mathbb{1}_{\pi^{-1}(\{1\})\times\{0\}}(x_i, \theta) = \sum_{i=1}^{N_0} \mathbb{1}_{\pi^{-1}(\{1\})}(x_i) = \sum_{i=1}^{N_0} \mathbb{1}_{\{\pi(x_i)=1\}}(x_i)$
- Falso Negativo (fn): $\sum_{i=1}^{N} \mathbb{1}_{\pi^{-1}(\{0\})\times\{1\}}(x_i, \theta) = \sum_{i=1}^{N_1} \mathbb{1}_{\pi^{-1}(\{0\})}(x_i) = \sum_{i=1}^{N_1} \mathbb{1}_{\{\pi(x_i)=0\}}(x_i)$

Notar que $N_0 = tn + fp$ y $N_1 = tp + fn$. La Figura 2.1 ilustra los distintos tipos de clasificación.

Definición 2.4. (Tasas) Consideremos $\Theta = \{0, 1\}$ y un total de N observaciones, de las cuales N_0 corrresponde a la clase 0 y N_1 corresponden a la clase 1. Además consideremos una regla $r : \mathbb{X} \to \{0, 1\}$. Definimos:

- Tasa de Verdadero Positivo (tpr):

$$\begin{aligned} P_{X|\Theta}(\pi(X) = 1|\Theta = 1) &= \frac{P_{X,\Theta}(\pi(X) = 1, \Theta = 1)}{p_\Theta(1)} \\ &= \frac{P_{X,\Theta}(\pi(X) = 1, \Theta = 1)}{P_{X,\Theta}(\pi(X) = 0, \Theta = 1) + P_{X,\Theta}(\pi(X) = 1, \Theta = 1)} \\ &\approx \frac{\frac{tp}{N}}{\frac{fn}{N} + \frac{tp}{N}} \\ &= \frac{tp}{fn + tp} \end{aligned} \tag{2.43}$$

- Tasa de Verdadero Negativo (tnr):

$$\begin{aligned} P_{X|\Theta}(\pi(X)=0|\Theta=0) &= \frac{P_{X,\Theta}(\pi(X)=0,\Theta=0)}{p_\Theta(0)} \\ &= \frac{P_{X,\Theta}(\pi(X)=0,\Theta=0)}{P_{X,\Theta}(\pi(X)=0,\Theta=0)+P_{X,\Theta}(\pi(X)=1,\Theta=0)} \\ &\approx \frac{\frac{tn}{N}}{\frac{tn}{N}+\frac{fp}{N}} \\ &= \frac{tn}{tn+fp} \end{aligned} \tag{2.44}$$

- Tasa de Falso Positivo (fpr):

$$\begin{aligned} P_{X|\Theta}(\pi(X)=1|\Theta=0) &= \frac{P_{X,\Theta}(\pi(X)=1,\Theta=0)}{p_\Theta(0)} \\ &= \frac{P_{X,\Theta}(\pi(X)=1,\Theta=0)}{P_{X,\Theta}(\pi(X)=0,\Theta=0)+P_{X,\Theta}(\pi(X)=1,\Theta=0)} \\ &\approx \frac{\frac{fp}{N}}{\frac{tn}{N}+\frac{fp}{N}} \\ &= \frac{fp}{tn+fp} \end{aligned} \tag{2.45}$$

- Tasa de Falso Negativo (fnr):

$$\begin{aligned} P_{X|\Theta}(\pi(X)=0|\Theta=1) &= \frac{P_{X,\Theta}(\pi(X)=0,\Theta=1)}{p_\Theta(1)} \\ &= \frac{P_{X,\Theta}(\pi(X)=1,\Theta=1)}{P_{X,\Theta}(\pi(X)=0,\Theta=1)+P_{X,\Theta}(\pi(X)=1,\Theta=1)} \\ &\approx \frac{\frac{fn}{N}}{\frac{fn}{N}+\frac{tp}{N}} \\ &= \frac{fn}{fn+tp} \end{aligned} \tag{2.46}$$

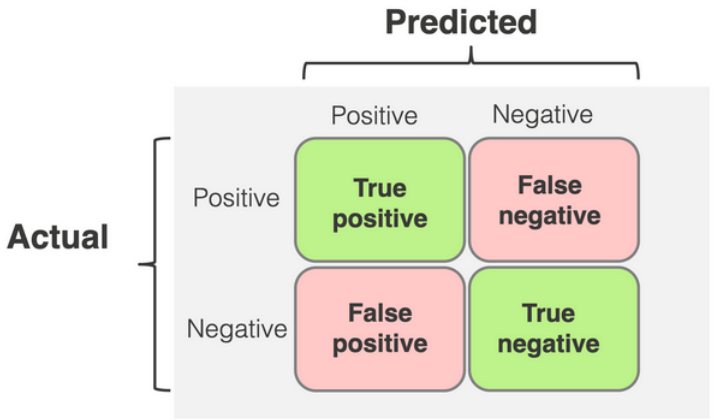


Figura 2.1: Ilustración de clasificación Binaria

2.5.1. Exactitud (Accuracy)

La exactitud se define como:

$$\begin{aligned} P_{X,\Theta}(\pi(X) = \Theta) &= P_{X,\Theta}(\pi(X) = 0, \Theta = 0) + P_{X,\Theta}(\pi(X) = 1, \Theta = 1) \\ &\approx \frac{tp}{N} + \frac{tn}{N} \\ &= \frac{tp + tn}{tp + tn + fp + fn} \end{aligned} \tag{2.47}$$

La exactitud es la medida por excelencia y mide la proporción de correctas predicciones, por definición es una aproximación de la probabilidad de correcta detección. La exactitud mide que tan cerca está la detección del valor correcto y se usan cuando los verdaderos negativos y los verdaderos positivos son los más relevantes, sin embargo, tiene limitaciones y es que puede dar interpretaciones erradas cuando las clases 0 y 1 son desbalanceadas. Por ejemplo, en un sistema de detección de spam, imaginemos por un momento que el spam es raro, basta tener un clasificador que diga siempre que no es spam, tendría una muy buena exactitud, pero no sería de mucha ayuda. Es por esto que las siguientes medidas complementan la anterior.

2.5.2. Exhaustividad (Recall)

La exhaustividad es lo mismo que que la Tasa de Verdadero Positivo (tnr), es decir, corresponde a:

$$P_{X|\Theta}(\pi(X) = 1|\Theta = 1) = \frac{tp}{fn + tp} \tag{2.48}$$

La exhaustividad es el poder del test, mide la proporción de predicciones positivas a lo largo de todas los datos positivos, es decir, se enfoca en qué tan bueno es el modelo en encontrar todos los positivos. Por ejemplo la exhaustividad mide, para todos los pacientes que tienen un ataque al corazón, cuantos de ellos fueron detectados con tener dicho problema. Una alta exhaustividad indica entonces que rara vez no detecta gente enferma. Dicho esto, se recomienda usar cuando el costo de los falsos negativos es muy alto. Ver Figura 2.2 para tener una idea ilustrativa.

2.5.3. Precisión (Precision)

La precisión se define como:

$$\begin{aligned} P_{\Theta|X}(\Theta = 1|\pi(X) = 1) &= \frac{P_{X,\Theta}(\pi(X) = 1, \Theta = 1)}{P_X(\pi(X) = 1)} \\ &= \frac{P_{X,\Theta}(\pi(X) = 1, \Theta = 1)}{P_{X,\Theta}(\pi(X) = 1, \Theta = 0) + P_{X,\Theta}(\pi(X) = 1, \Theta = 1)} \\ &\approx \frac{\frac{tp}{N}}{\frac{fp}{N} + \frac{tp}{N}} \\ &= \frac{tp}{fp + tp} \end{aligned} \tag{2.49}$$

La precisión mide la proporción de positivos correctos a lo largo de todas las predicciones hechas por el modelo. Por ejemplo la precisión mide cuantos pacientes realmente tienen una enfermedad al corazón a lo largo de cuantos de ellos fueron detectados con tener dicho problema. Una alta precisión entonces indica que uno es cuidadoso y rara vez se diagnostica erróneamente a un paciente. Dicho esto, se recomienda usar cuando el costo de los falsos positivos es muy alto. Ver Figura 2.2 para tener una idea ilustrativa.

2.5.4. F1-Score

Definimos la F1-Score como la media armónica entre la precisión y la exhaustividad, es decir:

$$F1 = \frac{2}{P^{-1}_{\Theta|X}(\Theta = 1|\pi(X) = 1) + P^{-1}_{X|\Theta}(\pi(X) = 1|\Theta = 1)} = \frac{2tp}{2tp + fn + fp} \tag{2.50}$$

Esta medida está limitada a clasificación binaria, ofrece un balance entre precisión y exhaustividad y, por lo tanto, mide que tan bien el clasificador hace este balance. La razón por la cual se ocupa media armónica y no aritmética es porque son probabilidades condicionales de distintas proyecciones, al ocupar media armónica, los recíprocos de ambas expresiones poseen denominador común (tp) y, por lo tanto, son comparables. F1-Score se utiliza cuando los falsos positivos y los falsos negativos son cruciales y cuando las clases 0 y 1 son desbalanceadas.

2.5.5. Matriz de Confusión

Cuando se tienen muchas clases, una manera visual de representar las correctas clasificaciones es mediante una matriz de confusión. Supongamos que tenemos $\Theta = K$, $K \in \mathbb{N}$ clases y $N = N_1 + ... + N_K$ datos. Además consideremos una regla $\pi : \mathbb{X} \to \{1, ..., K\}$. Definimos la matriz de confusión C_{KK} tal que

$$C_{ij} = \sum_{l=1}^{N} \mathbb{1}_{\pi^{-1}(\{j\})\times\{i\}}(x_l, \theta) = \sum_{l=1}^{N_i} \mathbb{1}_{\pi^{-1}(\{j\})}(x_l) \tag{2.51}$$

donde $i \in \{1, ..., K\}$ (filas) representa las instancias de las clases reales y j (columnas) representa las predicciones de cada clase. La Figura 2.3 muestra un típico caso de matriz de confusión, es de esperarse que los elementos en la diagonal concentren la mayor cantidad de elementos si se trata de un buen clasificador.

2.6. Caso de Estudio 1: Canal Binario Simétrico

El canal binario simétrico es un ejemplo básico en comunicaciones, la idea es que un bit de información (0 o 1) es transmitido por un canal hacia un receptor, quien debe decidir si el símbolo recibido corresponde al transmitido. Consideremos el siguiente canal de transmisión, modelado mediante probabilidades condicionales: Se tiene la siguiente relación:

$$P_{X|\Theta}(X = x|\Theta = 0) = \begin{cases} 1 - \epsilon & \text{si } x = 0 \\ \epsilon & \text{si } x = 1 \end{cases} \tag{2.52}$$

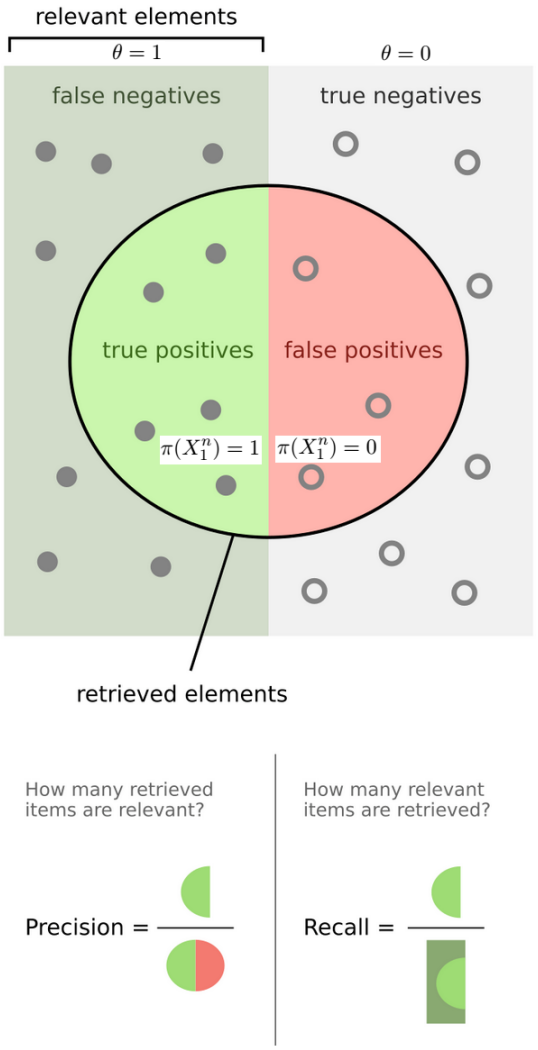


Figura 2.2: Ilustración de Exhaustividad y Precisión

$$P_{X|\Theta}(X = x|\Theta = 1) = \begin{cases} \epsilon & \text{si } x = 0 \\ 1 - \epsilon & \text{si } x = 1 \end{cases} \tag{2.53}$$

		Valor Predicho		
		Gato	**Perro**	**Conejo**
Valor real	**Gato**	5	3	0
	Perro	2	3	1
	Conejo	0	2	11

Figura 2.3: Matriz de Confusión para 3 clases

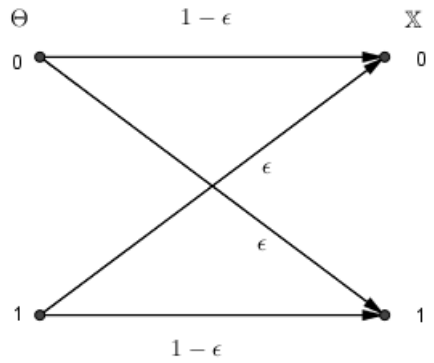


Figura 2.4: Canal Binario Simétrico

Es decir, la probabilidad de que el símbolo sea intercambiado al pasar por el canal es ϵ, por otra parte, la probabilidad de que el símbolo no cambie es $1-\epsilon$. Por otro lado, asumiremos que $p_\Theta(1) = p$ y $p_\Theta(0) = 1 - p$. En general consideremos una función de costo $L(v_1, v_2)$ $\forall v_1, v_2 \in \{0, 1\}$

$\mathcal{A}$ \ $\mathcal{A}$	0	1
0	$l_{00} = 0$	$l_{01} = 5$
1	$l_{10} = 7$	$l_{11} = 0$

Sabemos que la regla óptima dada una observación $x \in \{0,1\}$ está dada por (2.11), más precisamente,

$$\pi^*(x) = \arg\min_{y\in\mathcal{A}} \sum_{\theta\in\mathcal{A}} L(\theta,y) P_{\Theta|X}(\Theta=\theta|x), \quad \forall x \in \mathbb{X}. \tag{2.54}$$

Luego analizaremos la regla óptima según sea la observación recibida. Supongamos que $x = 1$, luego la regla óptima es:

$$\begin{aligned}
\pi^*(1) &= \arg\min_{\theta\in\{0,1\}} \left\{ L(0,\theta) P_{\Theta|X}(\Theta=0|X=1) + L(1,\theta) P_{\Theta|X}(\Theta=1|X=1) \right\} \\
&= \arg\min_{\theta} \left\{ \underbrace{L(0,0) P_{\Theta|X}(\Theta=0|X=1) + L(1,0) P_{\Theta|X}(\Theta=1|X=1)}_{\theta=0}, \underbrace{L(0,1) P_{\Theta|X}(\Theta=0|X=1) + L(1,1) P_{\Theta|X}(\Theta=1|X=1)}_{\theta=1} \right\} \\
&= \arg\min_{\theta} \left\{ \underbrace{L(1,0) P_{\Theta|X}(\Theta=1|X=1)}_{\theta=0}, \underbrace{L(0,1) P_{\Theta|X}(\Theta=0|X=1)}_{\theta=1} \right\} \\
&= \arg\min_{\theta} \left\{ \underbrace{L(1,0) \frac{f_{X,\Theta}(1,1)}{f_X(1)}}_{\theta=0}, \underbrace{L(0,1) \frac{f_{X,\Theta}(1,0)}{f_X(1)}}_{\theta=1} \right\} \\
&= \arg\min_{\theta} \left\{ \underbrace{L(1,0) \frac{f_{X,\Theta}(1,1)}{\sum_{\theta=0}^{1} f_{X,\Theta}(1,\theta)}}_{\theta=0}, \underbrace{L(0,1) \frac{f_{X,\Theta}(1,0)}{\sum_{\theta=0}^{1} f_{X,\Theta}(1,\theta)}}_{\theta=1} \right\} \\
&= \arg\min_{\theta} \left\{ \underbrace{l_{10} \frac{f_{X,\Theta}(1,1)}{f_{X,\Theta}(1,0) + f_{X,\Theta}(1,1)}}_{\theta=0}, \underbrace{l_{01} \frac{f_{X,\Theta}(1,0)}{f_{X,\Theta}(1,0) + f_{X,\Theta}(1,1)}}_{\theta=1} \right\} \\
&= \arg\min_{\theta} \left\{ \underbrace{l_{10} \frac{f_{X|\Theta}(1|1) p_\Theta(1)}{f_{X|\Theta}(1|0) p_\Theta(0) + f_{X|\Theta}(1|1) p_\Theta(1)}}_{\theta=0}, \underbrace{l_{01} \frac{f_{X|\Theta}(1|0) p_\Theta(0)}{f_{X|\Theta}(1|0) p_\Theta(0) + f_{X|\Theta}(1|1) p_\Theta(1)}}_{\theta=1} \right\} \\
&= \arg\min_{\theta} \left\{ \underbrace{l_{10} \frac{(1-\epsilon)p}{\epsilon(1-p) + (1-\epsilon)p}}_{\theta=0}, \underbrace{l_{01} \frac{\epsilon(1-p)}{\epsilon(1-p) + p(1-\epsilon)}}_{\theta=1} \right\}
\end{aligned} \tag{2.55}$$

Particularicemos el análisis cuando $p_\Theta(0) = p_\Theta(1) = 1/2$ y $\epsilon = 1/3$. La ecuación (2.55) reduce a:

$$\pi^*(1) = \arg\min_\theta \left\{ \underbrace{\frac{14}{3}}_{\theta=0}, \underbrace{\frac{5}{3}}_{\theta=1} \right\}$$
$$\pi^*(1) = 1 \quad \text{dado que } \frac{5}{3} < \frac{14}{3} \tag{2.56}$$

Análogamente, cuando $x = 0$, tenemos que:

$$\pi^*(0) = \arg\min_\theta \left\{ \underbrace{l_{10}\frac{p\epsilon}{p\epsilon + (1-\epsilon)(1-p)}}_{\theta=0}, \underbrace{l_{01}\frac{(1-\epsilon)(1-p)}{(1-\epsilon)(1-p)+p\epsilon}}_{\theta=1} \right\}$$
$$= \arg\min_\theta \left\{ \underbrace{\frac{7}{3}}_{\theta=0}, \underbrace{\frac{10}{3}}_{\theta=1} \right\}$$
$$\pi^*(0) = 0 \tag{2.57}$$

Por lo tanto la regla óptima en este caso es simplemente la función identidad. Lo anterior tiene sentido pues el canal no es lo suficientemente corrupto como para pensar que el símbolo recibido es distinto al que se transmite. Luego la decisión óptima corresponde a creerle al símbolo recibido. Notar también como afecta la decisión el costo de equivocarse (L_{10}, L_{01}) y la probabilidad p de enviar el símbolo p ya que distintas configuraciones de p, ϵ y los costos naturalmente afectarán la decisión final.

Propuesto 2.1. Analizar el caso $l_{01} = l_{10} = 1$, $l_{00} = l_{11} = 0$ (Regla MAP) como función de $p \in (0, 1)$.

Propuesto 2.2. Suponga $p = \frac{1}{2}$ y la función costo $L_{0,1}$, determine el régimen en ϵ donde $\pi^*(x) = x$ y por el contrario donde $\pi^*(x) = 1 - x$.

2.7. Caso de Estudio 2: Modelo Gaussiano

Para este ejemplo adoptaremos la notación $\overline{X}$ para denotar vectores en vez de lo usual X_1^n. Consideremos $\overline{m}_1 \in \mathbb{R}^n$ y que Θ toma valores en $\mathcal{A} = \{1, 2\}$ con probabilidad p_1 y p_2.

El modelo asume lo siguiente:

$$\overline{X} = \overline{m}_\Theta + \overline{N} \tag{2.58}$$

donde $\overline{X}$ es un vector de dimensión n y $\overline{N} \sim N(\overline{0}, \sigma^2 I_{n\times n})$ (donde $I_{n\times n}$ es la matriz identidad). Por lo tanto tenemos las siguiente probabilidades condicionales :

$$\begin{aligned}(\overline{X}|\Theta = 1) &\sim N(\overline{m}_1, \sigma^2 I_{n\times n})\\ (\overline{X}|\Theta = 2) &\sim N(\overline{m}_2, \sigma^2 I_{n\times n})\end{aligned} \tag{2.59}$$

El criterio óptimo bajo la regla $L_{0,1}$ dada una observación $\overline{x} \in \mathbb{R}^n$ es

$$\begin{aligned}\pi^*(\overline{x}) &= \arg \max_{\theta\in\{1,2\}} P_{\Theta|X}(\theta|\overline{x})\\ &= \arg \max_{\theta\in\{1,2\}} f_{X|\Theta}(\overline{x}|\theta)p_\Theta(\theta)\\ &= \arg \max_{\theta\in\{1,2\}} \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} e^{-\frac{1}{2}\left[(\overline{x}-\overline{m}_\theta)^t \frac{I}{\sigma^2}(\overline{x}-\overline{m}_\theta)\right]} p_\theta\end{aligned} \tag{2.60}$$

Consideremos la siguiente región de decisión, aquella donde se decide $\Theta = 1$:

$$S_{1,2} = \left\{\overline{x} \in \mathbb{R}^n : \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} e^{-\frac{1}{2}\left[(\overline{x}-\overline{m}_1)^t \frac{I}{\sigma^2}(\overline{x}-\overline{m}_1)\right]} p_1 > \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} e^{-\frac{1}{2}\left[(\overline{x}-\overline{m}_2)^t \frac{I}{\sigma^2}(\overline{x}-\overline{m}_2)\right]} p_2\right\}, \tag{2.61}$$

dicho de otra forma, esta zona corresponde a la zona donde $P_{\Theta|X}(\Theta = 1|X = \overline{x})$ es mayor que $P_{\Theta|X}(\Theta = 2|X = \overline{x})$. Analizamos la condición de pertenencia en $S_{1,2}$ con más detalle, tomando logaritmo:

$$\begin{aligned}-\frac{1}{2}\left[(\overline{x}-\overline{m}_1)^t \frac{I}{\sigma^2}(\overline{x}-\overline{m}_1)\right] + \frac{1}{2}\left[(\overline{x}-\overline{m}_2)^t \frac{I}{\sigma^2}(\overline{x}-\overline{m}_2)\right] &> \log\left(\frac{p_2}{p_1}\right)\\ \frac{1}{2\sigma^2}\left[||\overline{x}-\overline{m}_2||^2 - ||\overline{x}-\overline{m}_1||^2\right] &> \log\left(\frac{p_2}{p_1}\right)\end{aligned} \tag{2.62}$$

$$\begin{aligned}||\overline{x}||^2 - 2\langle\overline{x},\overline{m}_2\rangle + ||\overline{m}_2||^2 - ||\overline{x}||^2 + 2\langle\overline{x},\overline{m}_1\rangle - ||\overline{m}_1||^2 &> 2\sigma^2 \log\left(\frac{p_2}{p_1}\right)\\ \langle\overline{x},(\overline{m}_2-\overline{m}_1)\rangle &< \frac{||\overline{m}_2||^2 - ||\overline{m}_1||^2}{2} + \sigma^2\log\left(\frac{p_1}{p_2}\right)\end{aligned} \tag{2.63}$$

Es decir tenemos de (2.63) que:

$$S_{1,2} = \left\{\overline{x} \in \mathbb{R}^n : \langle\overline{x},(\overline{m}_2-\overline{m}_1)\rangle < \frac{||\overline{m}_2||^2 - ||\overline{m}_1||^2}{2} + \sigma^2\log\left(\frac{p_1}{p_2}\right)\right\}. \tag{2.64}$$

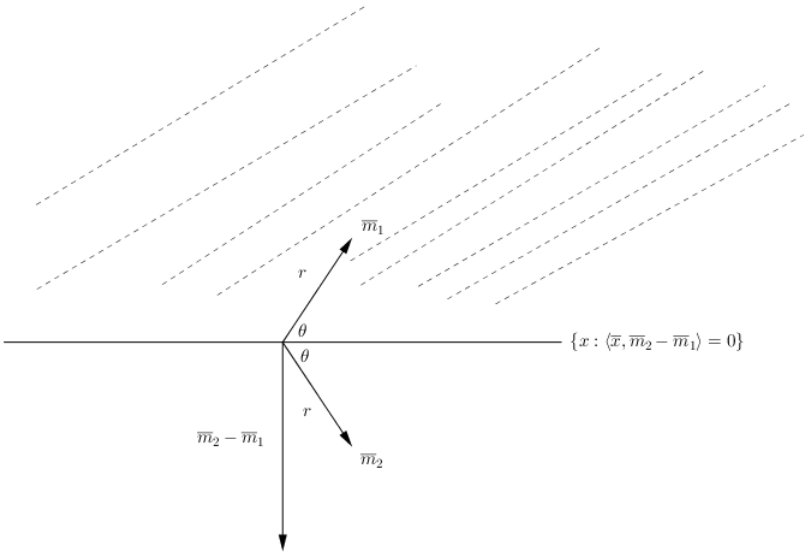


Figura 2.5: Diagrama región S_{12}

Si simplificamos al caso $||\overline{m}_1|| = ||\overline{m}_2|| = r$ y $p_2 = p_1$ la regla reduce a:

$$S_{1,2} = \{\overline{x} \in \mathbb{R}^n : \langle \overline{x}, (\overline{m}_2 - \overline{m}_1) \rangle \leq 0\}. \tag{2.65}$$

Supongamos ahora que estamos en el escenario equiprobable y además asumiendo la función de costo $L_{0,1}$, luego, $p_1 = p_2 = \frac{1}{2}$, vemos que el criterio de máxima verosimilitud implica la regla de mínima distancia (ver (2.62)):

$$\pi^*_{ML}(\overline{x}) = \arg \min_{\theta \in \{1,2\}} ||\overline{x} - \overline{m}_\theta|| \tag{2.66}$$

donde

$$S_{1,2} = \{\overline{x} \in \mathbb{R}^n : ||\overline{x} - \overline{m}_1|| < ||\overline{x} - \overline{m}_2||\}, \tag{2.67}$$

por lo tanto, $\pi^*_{ML}(\overline{x}) = 1$ si $||\overline{x} - \overline{m}_1|| < ||\overline{x} - \overline{m}_2||$.

Entonces cuando $p_1 = p_2 = \frac{1}{2}$ el criterio de máxima verosimilitud reduce a:

$$\pi_{ML}(\overline{x}) = \begin{cases} 1 & \text{si } ||\overline{x} - \overline{m}_1|| < ||\overline{x} - \overline{m}_2|| \\ 2 & \text{si } ||\overline{x} - \overline{m}_1|| \geq ||\overline{x} - \overline{m}_2|| \end{cases} \tag{2.68}$$

Por lo tanto

$$\begin{aligned} S_{12} = \pi(\{1\})^{-1} &= \{\overline{x} \in \mathbb{R}^n : \pi_{ML}(\overline{x}) = 1\} \\ &= \{\overline{x} \in \mathbb{R}^n : ||\overline{x} - \overline{m}_1|| < ||\overline{x} - \overline{m}_2||\} \\ &= \left\{ \overline{x} \in \mathbb{R}^n : \langle \overline{x}, (\overline{m}_2 - \overline{m}_1) \rangle < \frac{||\overline{m}_2||^2 - ||\overline{m}_1||^2}{2} \right\}. \end{aligned}$$

es la regla de mínima distancia. Finalmente evaluamos la probabilidad de error

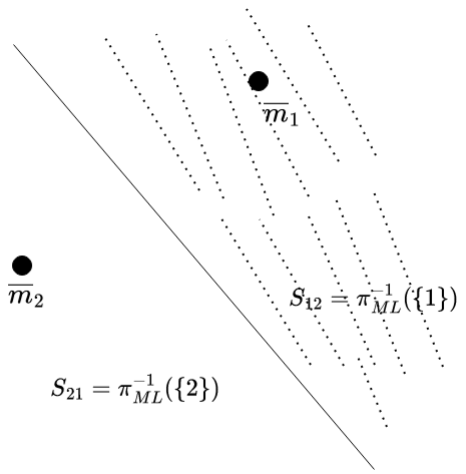


Figura 2.6: Diagrama región de mínima distancia cuando $n = 2$.

$$\begin{aligned} p_e &= \mathbb{E}_{X,\Theta}(L(\Theta, \pi(X))) \\ &= \sum_{\theta \in \{1,2\}} P_\Theta(\Theta = \theta) P_{X|\Theta}(\pi(X) \neq \theta | \Theta = \theta) \\ &= \frac{1}{2} P_{X|\Theta}(\pi(X) \neq 1 | \Theta = 1) + \frac{1}{2} P_{X|\Theta}(\pi(X) \neq 2 | \Theta = 2) \\ &= \frac{1}{2} P_{X|\Theta}(\pi(X) = 2 | \Theta = 1) + \frac{1}{2} P_{X|\Theta}(\pi(X) = 1 | \Theta = 2) \end{aligned}$$

Por simetría analizaremos solamente $P_{X|\Theta}(\pi(X) = 2|\Theta = 1)$, lo que nos lleva a:

$$\begin{aligned}
p_{\text{error},1} &= P_{\overline{X}}\left(\langle \overline{X}, (\overline{m}_2 - \overline{m}_1)\rangle \geq \frac{||\overline{m}_2||^2 - ||\overline{m}_1||^2}{2} \;\middle|\; \overline{X} = \overline{m}_1 + \overline{N}\right) \\
&= P_{\overline{N}}\left(\left\langle \overline{N}, (\overline{m}_2 - \overline{m}_1)\right\rangle + \overline{m}_1^t(\overline{m}_2 - \overline{m}_1) \geq \frac{||\overline{m}_2||^2 - ||\overline{m}_1||^2}{2}\right) \\
&= P_{\overline{N}}\left(\left\langle \overline{N}, (\overline{m}_2 - \overline{m}_1)\right\rangle + \langle \overline{m}_1, \overline{m}_2\rangle - ||\overline{m}_1||^2 \geq \frac{||\overline{m}_2||^2 - ||\overline{m}_1||^2}{2}\right) \\
&= P_{\overline{N}}\left(\left\langle \overline{N}, (\overline{m}_2 - \overline{m}_1)\right\rangle \geq \frac{||\overline{m}_2||^2 + ||\overline{m}_1||^2 - 2\langle \overline{m}_1, \overline{m}_2\rangle}{2}\right) \\
&= P_{\overline{N}}\left(\overline{N}^t(\overline{m}_2 - \overline{m}_1) \geq \frac{||\overline{m}_1 - \overline{m}_2||^2}{2}\right)
\end{aligned} \tag{2.69}$$

Notar que $\overline{N}$ es un vector Gaussiano multidimensional, lo que significa que $\overline{N}^t(\overline{m}_2 - \overline{m}_1)$ es una variable aleatoria Gaussiana de media $\mathbb{E}(\overline{N}^t(\overline{m}_2 - \overline{m}_1)) = 0$ y varianza:

$$\begin{aligned}
\mathbb{E}((\overline{N}^t(\overline{m}_2 - \overline{m}_1))^2\} &= \mathbb{E}\{(\overline{N}^t(\overline{m}_2 - \overline{m}_1))(\overline{N}^t(\overline{m}_2 - \overline{m}_1)))\} \\
&= \mathbb{E}((\overline{m}_2 - \overline{m}_1)^t\overline{N}\overline{N}^t(\overline{m}_2 - \overline{m}_1)) \\
&= (\overline{m}_2 - \overline{m}_1)^t\mathbb{E}(\overline{N}\overline{N}^t)(\overline{m}_2 - \overline{m}_1) \\
&= (\overline{m}_2 - \overline{m}_1)^t\sigma^2 I(\overline{m}_2 - \overline{m}_1) \\
&= \sigma^2||\overline{m}_2 - \overline{m}_1||^2
\end{aligned} \tag{2.70}$$

Luego, definiendo $Z = \overline{N}^t(\overline{m}_2 - \overline{m}_1)$, tenemos que

$$\begin{aligned}
P_{\overline{N}}\left(\overline{N}^t(\overline{m}_2 - \overline{m}_1) \geq \frac{||\overline{m}_1 - \overline{m}_2||^2}{2}\right) &= P_Z\left(Z \geq \frac{||\overline{m}_1 - \overline{m}_2||^2}{2}\right) \\
&= P_Z\left(\frac{Z}{\sigma||\overline{m}_2 - \overline{m}_1||} \geq \frac{||\overline{m}_1 - \overline{m}_2||^2}{2\sigma||\overline{m}_2 - \overline{m}_1||}\right) \\
&= Q\left(\frac{||\overline{m}_1 - \overline{m}_2||}{2\sigma}\right)
\end{aligned} \tag{2.71}$$

con $Q(z) = \int_z^\infty \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$. Finalmente,

$$p_{\text{error},1} = Q\left(\frac{||\overline{m}_1 - \overline{m}_2||}{2\sigma}\right) \tag{2.72}$$

La razón $SNR = \frac{||\overline{m}_1 - \overline{m}_2||}{\sigma}$ en (2.72) se conoce como la razón señal a ruido del problema de detección. Cuando se tiene una variable aleatoria Z positiva de esperanza finita, es posible utilizar la desigualdad de Markov.

$$P_Z(Z \geq z) \leq \frac{\mathbb{E}(Z)}{z}. \tag{2.73}$$

con $z \in \mathbb{R}^+$. Con esto podemos obtener una cota superior para la función $Q\left(\frac{||\overline{m}_1 - \overline{m}_2||}{2\sigma}\right)$, considerando $Z \sim N(0,1)$ y

$$\begin{aligned} P_Z\left(Z \geq \frac{||\overline{m}_1 - \overline{m}_2||}{2\sigma}\right) &\leq P_Z\left(Z^2 > \frac{||\overline{m}_1 - \overline{m}_2||^2}{4\sigma^2}\right) \\ &\leq \frac{\mathbb{E}(Z^2)4\sigma^2}{||\overline{m}_1 - \overline{m}_2||^2} \\ &= \frac{4\sigma^2}{||\overline{m}_1 - \overline{m}_2||^2} \\ &= \frac{4}{SNR^2}. \end{aligned} \tag{2.74}$$

Expresión que nos dice que aumentando la razón señal a ruido entonces es un criterio suficiente para minimizar la probabilidad de error.

2.8. Problemas

Se presentan a continuación una sección de problemas relacionados con detección Bayesiana.

Problema 2.1. (Detección y Criterio de Máxima Verosimilitud)
Considere el problema de diseñar un sistema de detección para un lector digital (por ejemplo un lector de CD). La idea es decodificar (detectar) símbolos binarios almacenados, por medio de mediciones secuenciales con ruido o errores de medición.

Formalmente consideremos Θ como la variable aleatoria en $\{0,1\}$ almacenada, y medimos una versión ruidosa de ella $X \in \{0,1\}$ (la variable de observación) donde se tiene que:

$$P_{X|\Theta}(X=0|\Theta=1) = P_{X|\Theta}(X=1|\Theta=0) = \epsilon \tag{2.75}$$

$$P_{X|\Theta}(X=0|\Theta=0) = P_{X|\Theta}(X=1|\Theta=1) = 1-\epsilon \tag{2.76}$$

con $0 < \epsilon < \frac{1}{2}$.

a) Para el problema de detectar Θ como función de X, determine la regla óptima de decisión $\pi^* : \{0,1\} \to \{0,1\}$, para la función de costo $L_{0,1}$ es decir:

$$\pi^* = \arg \min_{\pi:\{0,1\}\to\{0,1\}} \mathbb{E}_{X,\Theta}(L_{0,1}(\Theta, \pi(X))) \tag{2.77}$$

cuando $P(\Theta=1) = P(\Theta=0) = \frac{1}{2}$. Finalmente obtenga una expresión para la probabilidad de error de la regla óptima, es decir, determine

$$p_\epsilon = \mathbb{E}_{X,\Theta}(L_{0,1}(\Theta, \pi^*(X))). \tag{2.78}$$

b) La idea de esta parte es evaluar un esquema de codificación para mejorar el desempeño del detector de la parte a). Para ello consideremos un código $\mathcal{C} : \{0,1\} \to \{0,1\}^3$, donde las palabras binarias asociadas a los símbolos cero y uno las llamamos $\mathcal{C}(0) = (b_1, b_2, b_3)$ y $\mathcal{C}(1) = (c_1, c_2, c_3)$, respectivamente (luego $b_i, c_i \in \{0,1\}, i \in \{1,2,3\}$). Consideremos también Z como la nueva decisión y $\overline{\Theta}$ la señal codificada, es decir, $\overline{\Theta}$ queda dada por la siguiente regla o proceso de codificación:

$$\overline{\Theta} = (\Theta_1, \Theta_2, \Theta_3) = (b_1, b_2, b_3) \text{ si } Z = 0$$
$$\overline{\Theta} = (\Theta_1, \Theta_2, \Theta_3) = (c_1, c_2, c_3) \text{ si } Z = 1$$

Finalmente, lo que observamos es un vector aleatorio $\overline{X} = (X_1, X_2, X_3)$ (versión ruidosa de $\overline{\Theta}$), donde tenemos que, por independencia, lo siguiente:

$$P_{\overline{X}|\overline{\Theta}}((X_1, X_2, X_3) = (x_1, x_2, x_3)|(\Theta_1, \Theta_2, \Theta_3) = (\theta_1, \theta_2, \theta_3)) \tag{2.79}$$

$$= P_{X_1|\Theta_1}(X_1 = x_1|\Theta_1 = \theta_1)P_{X_2|\Theta_2}(X_2 = x_2|\Theta_2 = \theta_2)P_{X_3|\Theta_3}(X_3 = x_3|\Theta_3 = \theta_3) \tag{2.80}$$

y con la misma probabilidad de error

$$(\forall i \in \{1, 2, 3\})\ P_{X_i|\Theta_i}(X_i \neq b|\Theta_i = b) = \epsilon. \tag{2.81}$$

b.1) Determine las distribuciones condicionales, es decir, determine:

$$P_{\overline{X}|Z}(\overline{X} = (x_1, x_2, x_3)|Z = 0) \ \text{ y } \ P_{\overline{X}|Z}(\overline{X} = (x_1, x_2, x_3)|Z = 1) \tag{2.82}$$

como función de (b_1, b_2, b_3), (c_1, c_2, c_3) y ϵ. *Indicación*: Puede serle útil la función indicatriz $\mathbb{1}_{x_i \neq b_i}$ y $\mathbb{1}_{x_i \neq c_i}$.

b.2) Si $P_Z(Z = 1) = P_Z(Z = 0) = \frac{1}{2}$ determine la regla óptima

$$\pi^* : \{0, 1\}^3 \to \{0, 1\} \tag{2.83}$$

de detección de Z como función de $\overline{X}$ para la función costo $L_{0,1}$ y verifique que:

$$I_0 = \{(x_1, x_2, x_3) : \pi^*(x_1, x_2, x_3) = 0\} \tag{2.84}$$

$$= \{(x_1, x_2, x_3) : d_H(x_1, x_2, x_3; b_1, b_2, b_3) < d_H(x_1, x_2, x_3; c_1, c_2, c_3)\} \tag{2.85}$$

donde $d_H(x_1, x_2, x_3; y_1, y_2, y_3) = \mathbb{1}_{x_1 \neq y_1} + \mathbb{1}_{x_2 \neq y_2} + \mathbb{1}_{x_3 \neq y_3}$.

b.3) Determine una expresión para la nueva probabilidad de error

$$p_\epsilon = \mathbb{E}_{X,Z}(L_{0,1}(Z, \pi^*(\overline{X}))) \tag{2.86}$$

y demuestre que disminuye a medida que $d_H(b_1, b_2, b_3; c_1, c_2, c_3)$ aumenta. Con ello determine una condición sobre (b_1, b_2, b_3) y (c_1, c_2, c_3) (es decir sobre el código $\mathcal{C}$) para minimizar (2.86).

Problema 2.2. Considere un problema de detección binario $\Theta = \{0, 1\}$ en un contexto Bayesiano, donde $p = p_\Theta(1)$ y $1 - p = p_\Theta(0)$ y donde la probabilidad condicional de X dado $\Theta = \theta$ esta dada por la distribución $P_{X|\Theta}(\cdot|\theta)$ con densidad $f_{X|\Theta}(x|\theta)$. Considere una función de costo arbitraria con los siguientes valores: $L_{0,0}$, $L_{1,0}$, $L_{0,1}$ y $L_{1,1}$. Estos elementos definen la función de costo[4].

a) Dado $A \subset \mathbb{X}$ arbitrario, considere un test de la forma: $\pi_A(x) = 1_A(x)$, donde $1_A(x)$ es la función indicatriz de A. Determine expresiones para $P_{j,i} = P_{X|\Theta}(\pi_A(X) = i|\Theta = j)$ y con ello el riesgo del test dado por

$$r(\pi_A) = \mathbb{E}_{X,\Theta}(L(\Theta, \pi_A(X))).$$

b) Considere $L_{0,0} = L_{1,1} = 0$. Determine el test Bayesiano óptimo $\pi_{MAP}(x)$ y verifique que $\pi_{MAP}(x) = \pi_A(x)$ para un $A \subset \mathbb{X}$. Determine la forma del conjunto óptimo A, como función de $L_{0,1}$, $L_{1,0}$, p, $f_{X|\Theta}(x|0)$ y $f_{X|\Theta}(x|1)$.

c) Verifique que la solución Bayesiano óptima del punto anterior, es también óptima en el sentido de Neyman-Pearson, es decir en el sentido que ofrece un compromiso optimo entre poder y tamaño.
Para ello determine $\alpha_{\pi_{MAP}}$ y demuestre que no existe un test binario de tamaño menor que $\alpha_{\pi_{MAP}}$ tal que su poder sea mayor que $\beta_{\pi_{MAP}}$. *Indicación*: Encuentre una expresión para relacionar $r(\pi_{MAP})$ con $\alpha_{\pi_{MAP}}$ y $\beta_{\pi_{MAP}}$.

Problema 2.3. Se pide que implemente un sistema de decisión que detecte la presencia de una señal $s_t \triangleq s(t)$. Para eso suponga que se tiene un sistema que observa n muestras ruidosas de la señal $(s_k)_{k=1,\ldots,n}$.

En concreto se distinguen dos escenarios posibles de observación.

Presencia de señal $\Theta = 1$:

$$\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{pmatrix} + \begin{pmatrix} N_1 \\ N_2 \\ \vdots \\ N_n \end{pmatrix} \tag{2.87}$$

[4] $L_{i,j}$ es el costo de decidir j cuando el valor verdadero es que toma Θ es i.

Ausencia de señal $\Theta = 0$:

$$\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} N_1 \\ N_2 \\ \vdots \\ N_n \end{pmatrix} \tag{2.88}$$

donde $N_1, ..., N_n$ son variables aleatorias independientes que distribuyen $\mathcal{N}(0, \sigma^2)$.

a) Notar que dado el valor de $\Theta = \theta$, el vector $X_1^n = (X_1, ..., X_n)$ es un vector Gaussiano. Determine su vector de media y matriz de covarianza en ambos escenarios (presencia y ausencia de señal). *Indicación*: Notar que $X_1, ..., X_n$ son variables aleatorias independientes.

b) Del punto anterior determine la función de log-verosimilitud

$$L(x_1, ..., x_n|\theta) = \ln f_{X_1,...,X_n|\Theta}(x_1, ..., x_n|\theta)$$

y la solución del problema:

$$\tau_{ML}(x_1, ..., x_n) = \arg \max_{\theta \in \{0,1\}} L(x_1, ..., x_n|\theta). \tag{2.89}$$

Indicación: Se debe llegar a una expresión cerrada para $\tau_{ML}(x_1, ..., x_n)$, función de $x_1, ..., x_n$ y los parámetros conocidos del problema.

c) Determine la probabilidad de error del test del punto anterior cuando $p_\Theta(1) = p_\Theta(0) = \frac{1}{2}$.

d) Determine que pasa con la probabilidad de error del test óptimo en (2.89), si la potencia de la señal dada por $||s||^2 = \sum_{i=1}^n s(i)^2$ tiende a infinito, es decir, $\lim_{n\to\infty} ||s||^2 = \infty$

Problema 2.4. La estudiante PAT se encuentra en una sala de espías y debe detectar la presencia del símbolo $S_1 = 1$ o $S_2 = -1$ a partir de múltiples observaciones para su posterior decodificación. Por problemas de tiempo, solamente alcanza a recibir dos observaciones ruidosas de tal símbolo. Se sabe que la primera variable de observación X_1 es tal que $X_1 = N_1 + S_i$ $(i \in \{1, 2\})$ y la segunda variable de observación X_2 es tal que $X_2 = N_2 + S_i$ $(i \in \{1, 2\})$. Donde N_1 y N_2 son variables aleatorias independientes que

siguen una distribución $Laplace(0,1)$. Suponga que las señales S_1 y S_2 son transmitidas con igual probabilidad e independientes de N_1 y N_2..

Indicación: Una variable aleatoria X sigue una distribución $Laplace(\mu,\beta)$, $\mu \in \mathbb{R}$, $\beta > 0$ si su densidad está dada por

$$f_X(x) = \frac{1}{2\beta} e^{\frac{-|x-\mu|}{\beta}} \quad x \in \mathbb{R}.$$

Puede ocupar además el hecho que si $X \sim Laplace(\mu,\beta)$ entonces $X + \alpha \sim Laplace(\mu+\alpha,\beta)$, con $\alpha \in \mathbb{R}$.

El objetivo de la pregunta es ayudar a PAT a diseñar un esquema de decisión a partir de las observaciones recibidas. Para esto siga los siguientes pasos:

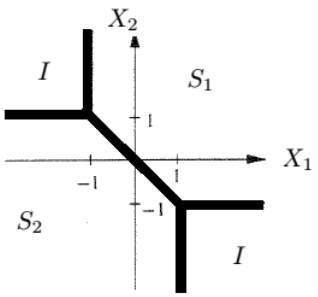
a) Considere el problema de detección Bayesiano, es decir, indique lo siguiente: El espacio de observación $\mathbb{X}$, el espacio de decisión $\mathcal{A}$, la distribución de la variable aleatoria Θ a inferir y la densidad condicional de las observaciones $f_{X|\Theta}(x|\theta)$, $x \in \mathbb{X}$.
b) Suponga que la función de costo del problema Bayesiano es la función $L_{0,1}$, es decir,

$$L_{0,1}(x,y) = \begin{cases} 1 & \text{si } x \neq y \\ 0 & \text{si } x = y. \end{cases}$$

Demuestre que en este caso la regla óptima $r^* : \mathbb{X} \to \mathcal{A}$ se puede expresar de la siguiente manera:

$$r^*(x) = \begin{cases} 1 & \text{si } \frac{f_{X|\Theta}(x|\theta_1)}{f_{X|\Theta}(x|\theta_2)} > \frac{P_\Theta(\Theta=\theta_2)}{P_\Theta(\Theta=\theta_1)} \\ -1 & \text{si } \frac{f_{X|\Theta}(x|\theta_1)}{f_{X|\Theta}(x|\theta_2)} < \frac{P_\Theta(\Theta=\theta_2)}{P_\Theta(\Theta=\theta_1)} \\ I & \text{si } \frac{f_{X|\Theta}(x|\theta_1)}{f_{X|\Theta}(x|\theta_2)} = \frac{P_\Theta(\Theta=\theta_2)}{P_\Theta(\Theta=\theta_1)} \end{cases}$$

Donde I indica *indiferencia*, es decir, -1 o 1.
c) Verifique que la región de decisión óptima está dada por la siguiente figura:



Para eso le ayudará ubicar en la figura las siguientes zonas:

- $X_1 = -X_2$,
- $X_1 \leq -1$, $X_2 \geq 1$
- $X_1 \geq 1$, $X_2 \leq -1$

Luego demuestre e interprete la razón por la que en estos casos la decisión óptima corresponde a la indiferencia.

Indicación: Simplifique lo que más pueda la región de decisión obtenida en **(b)** y analice lo obtenido con la región de la figura.

Problema 2.5. La estudiante PAT fue a una degustación a ciegas de una pizzería muy famosa. Lamentablemente PAT estaba muy congestionada por lo que le costaba identificar los sabores y además por ser una degustación a ciegas no sabía cuál pizza estaba comiendo. Se sabe que este local durante la degustación ofreció 3 tipos de pizzas: pizza con carne, pizza con piña y pizza GFX, además, la frecuencia con la cual salía cada pizza del horno era del 50 %, 20 % y 30 % respectivamente. Las pizzas pueden gustarle o no gustarle a PAT y, por experiencia, ella sabe que si la pizza es de carne le gusta el 40 % de las veces, si es de piña le gusta el 20 % de las veces y si es la GFX le gusta el 70 % de las veces. PAT probó la pizza recibida y le gustó. El objetivo de esta pregunta es que usted ayude a PAT a detectar cuál fue la pizza que comió. Para esto siga los siguientes pasos:

a) Plantee un espacio de observación y un espacio de decisión adecuado, indique la distribución de la variable aleatoria Θ asociado al espacio de decisión (distri-

bución a priori) y las densidades y/o probabilidades de masa condicionales de X dado $\Theta = \theta$.

b) Suponga que la función de costo del problema Bayesiano es la función $L_{0,1}$, es decir,

$$L_{0,1}(x, y) = \begin{cases} 1 & \text{si } x \neq y \\ 0 & \text{si } x = y. \end{cases} \tag{2.90}$$

Plantee la regla óptima de decisión y a partir de esto decida cuál fue la pizza que comió.

Problema 2.6. Considere un problema de detección Bayesiano con un espacio de observación $\mathbb{X}$ arbitrario tal que $\Theta = \{0, 1\}$, es decir, un problema de decisión binario. Además considere la función de costo $L_{0,1}$.

a) Demuestre que en este caso la regla óptima puede escribirse como:

$$r^*(x) = \begin{cases} 1 & \text{si } \eta(x) > 1/2 \\ 0 & \text{si } \eta(x) < 1/2 \\ I & \text{si } \eta(x) = 1/2 \end{cases} \tag{2.91}$$

donde $\eta(x) = P_{\Theta|X}(\Theta = 1|X = x)$ e I indica indiferencia, es decir, 0 o 1.

b) Demuestre que para cualquier otra regla de decisión $\pi : \mathbb{X} \to \{0, 1\}$ se tiene que:

$$P_{X,\Theta}(r^*(X) \neq \Theta) \leq P_{X,\Theta}(\pi(X) \neq \Theta).$$

Es decir, la regla r^* es aquella que minimiza la probabilidad de error. Para esto use el hecho que, dado $x \in \mathbb{X}$ y una regla $\pi : \mathbb{X} \to \{0, 1\}$ arbitraria entonces:

$$P_{\Theta|X}(\pi(X) \neq \Theta|X = x) = 1 - \left[\mathbb{1}_{\pi(x)=1}(x) \cdot \eta(x) + \mathbb{1}_{\pi(x)=0}(x) \cdot (1 - \eta(x))\right], \tag{2.92}$$

donde $\mathbb{1}_B(x)$ corresponde a la indicatriz, es decir, vale 1 si $x \in B$ y 0 en caso contrario.

Problema 2.1. Considere un cuerpo radiactivo que emite θ partículas, con $\theta \in \mathbb{N}$. Para detectar las partículas emitidas, se cuenta con un detector imperfecto, el cual detecta cada partícula emitida de forma independiente. Para modelar el proceso de detección, consideremos la variable aleatoria B_i que toma el valor 1 si la partícula i-ésima fue detectada y 0 si no, donde B_i distribuye Bernoulli de parámetro p ($P_{B_i}(B_i = 1) = p$).

Finalmente, la variable de observación X es el número de partículas totales detectadas dada por

$$X = \sum_{i=1}^{\theta} B_i \in \{0, \cdots, \theta\}$$

Notar que dados p y θ conocidos, X distribuye binomial de parámetros p y θ, es decir:

$$P_X(X = k|p, \theta) = \binom{\theta}{k} p^k (1-p)^{\theta-k}$$

Considere el problema de estimar la cantidad de partículas emitidas θ asumiendo conocido p, pero en un contexto Bayesiano, donde la cantidad de partículas emitidas distribuye Poisson de parámetro λ conocido, es decir:

$$p_\Theta(\theta) = \frac{\lambda^\theta}{\theta!} e^{-\lambda}, \quad \forall \theta \in \{0, 1, 2, \cdots\}$$

Luego, se busca el estimador que minimice el error cuadrático medio $\phi_{MMSE}(X)$, dada una observación de X. Para ello, siga los siguientes pasos:

a) Determine la probabilidad conjunta $P_{X,\Theta}(X = k, \Theta = \theta)$ y con ello muestre que la variable aleatoria X (número de partículas detectadas) distribuye Poisson de parámetro λp, es decir:

$$p_X(k) = \frac{(\lambda p)^k}{k!} e^{-\lambda p} \quad \forall k \in \{0, 1, 2, \cdots\}$$

b) Muestre que:

$$P_{\Theta|X}(\Theta = \theta|X = k) = \frac{(\lambda(1-p))^{\theta-k}}{(\theta-k)!} e^{-\lambda(1-p)}, \quad \text{si } \theta \geq k$$

y

$$P_{\Theta|X}(\theta|k) = 0 \quad \text{si } \theta < k$$

y con ello obtenga $\pi_{MMSE}(X)$. Comente sobre los regímenes $p \approx 1$ y $p \approx 0$

3

Unidad III: Estimación Paramétrica

Estimación es el proceso de toma de decisiones en un espacio de parámetros continuo. Hemos visto dos filosofías principales en detección: el Lema de Neyman-Pearson donde no hay distribuciones a priori en los parámetros; y el enfoque Bayesiano. donde se asume la existencia de una distribución a priori. Esta misma dicotomía existe en estimación, ya que podemos ver el parámetro (continuo) como una cantidad determinística (pero desconocida) o como una variable aleatoria. En el caso de estimación paramétrica el eje central de esta teoría está dominado por el principio de *máxima verosimilitud* que se verá en las siguientes secciones.

Formalmente el problema de estimación se entiende como la inferencia de una variable θ continua (que toma una cantidad no numerable de posibles valores) a partir de una variable aleatoria (o vector aleatorio) de observación X, donde naturalmente θ influyó en la observación recibida x.

Si hacemos el contraste, en detección uno elige una opción (decisión) dentro de un conjunto de posibles resultados mediante un detector o regla. En este caso dado que θ vive en un conjunto no numerable (continuo), entonces el objetivo es poder *acercarse* a dicho valor. El problema natural surge a que dicho valor o parámetro θ es desconocido, por lo que la regla a elegir, en este caso llamado estimador, debe cumplir un criterio de

optimalidad sin saber de antemano que el parámetro es el correcto.

En muchos ámbitos teóricos y prácticos, nos vemos enfrentados al problema de estimar un parámetro (o parámetros) de una distribución indexada por θ por medio de múltiples observaciones. Por lo que un supuesto central en estimación es que se tendrá acceso a un vector X_1^n independiente e idénticamente distribuido (i.i.d.). Esto último es de vital importancia porque vamos a pedir, como veremos más adelante, algún criterio de optimalidad cuando las observaciones tienden a infinito.

Antes de continuar la formalización del problema, daremos un ejemplo que ilustra la familia de distribuciones Bernoulli indexadas por $\theta \in [0, 1]$.

Ejemplo 3.1. Sea $\mathbb{X} = \{0, 1\}$, $\mathcal{F}_\Theta = \{P_X(\cdot|\theta) : \theta \in [0, 1]\}$, la familia de distribuciones asociadas a X donde

$$P_X(X = 1|\theta) = \theta \tag{3.1}$$

$$P_X(X = 0|\theta) = 1 - \theta \tag{3.2}$$

Supongamos que poseemos un vector aleatorio X_1^n independiente e idénticamente distribuido donde cada marginal tiene distribución $P_X(\cdot|\theta)$, la pregunta es estimar θ a partir de este vector de observaciones.

Vemos que este modelamiento se puede asociar a, por ejemplo, determinar si una moneda está cargada o no, y cuál es dicho valor de carga. Una solución plausible sería, a partir del vector de observaciones, contar cuántas veces se obtuvo el valor $X = 0$ y dividirlo respecto al total. Formalmente entonces un estimador $\tau_n : \mathbb{X}^n \to \Theta$ natural para este caso sería el siguiente:

$$\tau_n(x_1^n) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{1\}}(x_i), \tag{3.3}$$

donde $x_1^n \in \{0, 1\}^n$ y

$$\mathbb{1}_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases} \tag{3.4}$$

es la función indicatriz.

El estimador anterior corresponde al promedio empírico que no es más que contar la cantidad de veces que se repitió un resultado y se divide por el total de observaciones. Esto

corresponde a, gracias a la ley de los grandes números, a obtener un valor aproximado de la probabilidad de obtener 1 ya que:

$$\mathbb{E}\left(\tau_n(X_1^n)\right) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{1\}}(X_i)\right) = \frac{1}{n}\sum_{i=1}^{n}P_{X_i}(X_i = 1|\theta) = \frac{1}{n}\sum_{i=1}^{n}\theta = \theta. \tag{3.5}$$

Por lo que la solución propuesta de estimador tiene garantías de converger al parámetro que se busca inferir cuando las observaciones tienden a infinito. En lo que sigue formalizaremos el problema de estimación paramétrica y también encontraremos criterios para diseñar un estimador en algún sentido de optimalidad.

3.1. Formalización del Problema de Estimación Paramétrica

Un problema de estimación paramétrica se compone de 4 elementos centrales:

- Un espacio de observación $\mathbb{X}$ (típicamente $\mathbb{X} = \mathbb{R}$) y variables aleatorias que toman valores en $\mathbb{X}$. El valor particular $X = x$ se conoce como observación, realización o dato. Si se tienen $n \in \mathbb{N}$ observaciones o datos entonces $X_1^n = x_1^n \in \mathbb{X}^n$. X_1^n se conoce como vector aleatorio, vector variable de observación.
- Un espacio de parámetros Θ infinito no numerable. Es también el espacio de llegada o el espacio donde nos interesa inferir el parámetro.
- Una familia de distribuciones de probabilidad indexadas por $\theta \in \Theta$, es decir, considerando el vector aleatorio X_1^n, tenemos lo siguiente:

  $$\mathcal{F}_\Theta = \{P_X(\cdot|\theta) : \theta \in \Theta\}.$$

  $\mathcal{F}_\Theta$ se conoce como una familia de distribuciones de probabilidad parametrizadas mediante θ, es decir, por cada valor de θ se obtiene una distribución distinta (θ es el parámetro que indexa la familia) y Θ el universo de posibles parámetros factibles (por ejemplo $\theta \in \mathbb{R}^q$ con $q \in \mathbb{N}$). En este apunte nos concentraremos, salvo algunas excepciones. en la estimación de un sólo parámetro, luego $q = 1$.

  $\mathcal{F}_\Theta$ también se conoce como modelo paramétrico.
- Una función $\tau_n : \mathbb{X}^n \to \Theta$ donde $\tau_n(X_1^n) = \tau_n(X_1, ..., X_n)$. Las funciones de observaciones en el caso de detección se llaman reglas o test, en este caso se llaman estadísticos. Además, en este caso en particular cuando el recorrido del estadístico es Θ se le llama estimador (también se suele escribir como $\hat{\theta}_n(X_1^n)$).

Veamos un ejemplo concreto de familia de distribuciones:

Ejemplo 3.2. Consideremos la variable aleatoria X y una familia de distribuciones normales de media $\theta \in \mathbb{R}$ y varianza $\sigma^2 \in \mathbb{R}^+$, luego,

$$\mathcal{F}_\Theta = \{P_X(\cdot|\theta) : \theta \in \mathbb{R}\},$$

donde $P_X(\cdot|\theta)$ es una distribución que se caracteriza por su densidad de probabilidad dada por:

$$f_X(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\theta)^2}{2\sigma^2}}.$$

Luego esta familia de distribuciones es aquella que posee solamente distribuciones normales y cada una se diferencia de la otra por su valor de media $\theta \in \mathbb{R}$.

Observaciones 3.1. Así como en detección, las distribuciones indexadas por θ pueden ser discretas o continuas, pero el parámetro θ debe estar en un conjunto no numerable para ser considerado un problema de estimación.

Entonces, formalmente vamos a considerar un vector aleatorio $X_1^n = (X_1, ..., X_n)$ con distribución conjunta $P_{X_1^n}(\cdot|\theta) \in \mathcal{F}_\Theta$. El problema de estimación paramétrica consiste en encontrar un estimador $\tau_n : \mathbb{X}^n \to \Theta$ donde $\tau_n(X_1^n) = \tau_n(X_1, ..., X_n)$[1] es el valor estimado.

Salvo casos específicos, en adelante vamos a considerar el problema de muchas observaciones independientes e idénticamente distribuidas (i.i.d.), es decir, tenemos:

$$\theta \in \Theta \to (X_1, ..., X_n) \sim P_{X_1^n}(\cdot|\theta)^n \tag{3.6}$$

lo que lleva a:

$$P_{X_1^n}(X_1 \in A_1, ..., X_n \in A_n|\theta) = P_X(A_1|\theta) \cdot P_X(A_2|\theta) \cdot ... \cdot P_X(A_n|\theta), \quad \forall A_1, ..., A_n \subseteq \mathbb{X}, \tag{3.7}$$

en otras palabras $X_1^n = (X_1, ..., X_n)$ son muestras i.i.d. con marginal $P_X(\cdot|\theta) \in \mathcal{F}_\Theta$.

[1] También se suele escribir como $\hat{\theta}_n(X_1^n)$.

Observaciones 3.2.

1- Un estimador $\tau_n(X_1^n) = \tau_n(X_1, ..., X_n)$ no es más que una función que va desde el espacio de observaciones a una decisión en el espacio de parámetros.
2- $\tau_n(X_1^n)$ es una variable aleatoria en Θ, dado que X_1^n es un vector aleatorio en $\mathbb{X}^n$.
3- En estadística se suele escribir como $\tau_n(x_1^n)$ y se refiere a una realización del estadístico, es decir, cuando ya tienes una muestra concreta (vector de observaciones) $X_1^n = x_1^n$. Entonces $\tau_n(x_1^n)$ es simplemente un número calculado a partir de esos datos observados. En esta unidad trabajaremos en general con su versión con variable aleatoria, salvo en los casos que se tengan ejemplos concretos.
4- Si $X_1^n \sim P_{X_1^n}(\cdot|\theta) \Rightarrow \tau_n(X_1^n) \sim P_{\theta\in\Theta}$ en Θ. En otras palabras al fijar θ la $P_{X^n}(\cdot|\theta)$ induce una distribución $P_{\theta\in\Theta}$ en Θ por medio de τ_n.

En resumen, cualquier función de $\mathbb{X}^n \to \Theta$ induce un estimador, estos estimadores pueden acercarse al valor real como no. Para saber si un estimador es una buena elección necesitamos proponer criterios para seleccionar uno, con alguna noción de optimalidad.

3.1.1. Supuestos Adicionales

Un principio básico que debe satisfacer la familia es la noción de identificabilidad o discriminabilidad. Además, pediremos ciertas condiciones de regularidad para garantizar intercambio de derivada con integral en caso de ser necesario. Estas condiciones

Definición 3.1. (Familias Distinguibles) Decimos que la familia paramétrica

$$\mathcal{F}_\Theta = \{P_X(\cdot|\theta) : \theta \in \Theta\},$$

es identificable o discriminable si existe un $A \subset \mathbb{X}$ tal que $\forall\theta, \theta' \in \Theta$ tal que $\theta \neq \theta'$ entonces $P_X(A|\theta) \neq P_X(A|\theta')$. Matemáticamente la distancia entre dos distribuciones $P_X(\cdot|\theta) \neq P_X(\cdot|\theta')$ se puede expresar como:

$$V(P_X(\cdot|\theta), P_X(\cdot|\theta')) = \sup_{A\subseteq\mathbb{X}} |P_X(A|\theta) - P_X(A|\theta')| > 0,$$

donde $V : \mathcal{F} \times \mathcal{F} \mapsto \mathbb{R}^+$ se llama distancia en variaciones totales.

Lo que se pide esencialmente es que exista al menos un evento donde las medidas de probabilidad difieran, de esta manera son distinguibles y por lo tanto sea posible plantear un problema de estimación. En adelante asumiremos que las familias paramétricas con las cuales trabajaremos son distinguibles.

Definición 3.2. (Condiciones de Regularidad) Las siguientes condiciones se asumirán para los cálculos que se realizarán más adelante. Estos supuestos evitarán problemas de indefiniciones matemáticas.

1- Para un conjunto $\Theta \subset \mathbb{R}$ se tiene que:

$$(\forall x \in \mathbb{X})(\forall \theta, \theta' \in \Theta) L(x|\theta) > 0 \Leftrightarrow L(x|\theta') > 0, \tag{3.8}$$

donde $L(x|\theta)$ es la función de verosimilitud. Esta definición implica que existe un conjunto $A \subset \mathbb{X}$:

$$(\forall x \in A)(\forall \theta \in \Theta) L(x|\theta) > 0 \tag{3.9}$$

$$(\forall x \in A^c)(\forall \theta \in \Theta) L(x|\theta) = 0. \tag{3.10}$$

2- Para cualquier $x \in A$ la función de verosimilitud $L : \Theta \to \mathbb{R}^+ \cup \{0\}$ es de clase C^2 (segundas derivadas continua respecto a θ).

3- $\int\limits_A \left| \frac{\partial L(x|\theta)}{\partial \theta} \right| dx < \infty$ (caso continuo) o $\sum\limits_{x \in A} \left| \frac{\partial L(x|\theta)}{\partial \theta} \right| < \infty$ (caso discreto)

4- $\frac{\partial}{\partial \theta} \int\limits_A L(x|\theta) dx = \int\limits_A \frac{\partial L(x|\theta)}{\partial \theta} dx$ (caso continuo) o $\frac{\partial}{\partial \theta} \sum\limits_{x \in A} L(x|\theta) = \sum\limits_{x \in A} \frac{\partial L(x|\theta)}{\partial \theta}$ (caso discreto).

5- $\frac{\partial}{\partial \theta} \int\limits_A \frac{\partial L(x|\theta)}{\partial \theta} dx = \int\limits_A \frac{\partial^2 L(x|\theta)}{\partial \theta^2} dx$ (caso continuo) o $\frac{\partial}{\partial \theta} \sum\limits_{x \in A} \frac{\partial L(x|\theta)}{\partial \theta} = \sum\limits_{x \in A} \frac{\partial^2 L(x|\theta)}{\partial \theta^2}$ (caso discreto).

Las últimas condiciones son garantías para poder intercambiar integral con derivada, esto siempre se puede hacer si es que la condición 3 se cumple.

3.2. Nociones de Optimalidad

Las siguientes definiciones son criterios básicos y esenciales para definir que un estimador tiene un comportamiento *deseado*. Estos criterios están basados en condiciones finitas o asintóticas respecto al número de observaciones.

Definición 3.3. (Consistencia) Sea $\theta \in \Theta$ arbitrario tal que $(X_1, ..., X_n) \sim P_{X_1^n}(\cdot|\theta)$, una secuencia de estimadores $(\tau_n)_{n\in\mathbb{N}}$ se dice consistente, si $\forall \epsilon > 0$ se cumple que:

$$\lim_{n\to\infty} P_{X_1^n}(|\tau_n(X_1^n) - \theta| > \epsilon) = 0. \tag{3.11}$$

o, alternativamente,

$$\tau_n(X_1^n) \xrightarrow{P} \theta. \tag{3.12}$$

Si, por otra parte,

$$\lim_{n\to\infty} \tau_n(X_1^n) \xrightarrow{\text{c.s.}} \theta.$$

Es decir, el estimador converge casi seguramente a θ, se dice que $(\tau_n)_{n\in\mathbb{N}}$ es fuertemente consistente.

Notar que (3.12) es equivalente a decir que $\forall \epsilon > 0$, $\forall \nu > 0$, $\exists n_0 \in \mathbb{N}$ $\forall n \geq n_0$

$$P_{X_1^n}\left(\{x_1^n \in \mathbb{X}^n : |\tau_n(x_1^n) - \theta| > \epsilon\}\right) < \nu. \tag{3.13}$$

En lenguaje de convergencia de variables aleatorias (3.12) y (3.13) equivale a decir que la secuencia $\tau_1(X_1), \tau_2(X_1^2), \tau_3(X_1^3), ..., \tau_n(X_1^n) \to \theta$ en probabilidad, por lo que el estimador, a medida que aumenta la cantidad de muestras, se acerca al parámetro desconocido con una alta probabilidad.

La definición de consistencia es una propiedad asintótica, es decir cuando $n \to \infty$ se cumple lo pedido. También es importante tener condiciones deseables en el régimen de muestras finitas.

Definición 3.4. (Estimador Insesgado) Sea $\theta \in \Theta$ arbitrario tal que $(X_1, ..., X_n) \sim P_{X_1^n}(\cdot|\theta)$, una secuencia de estimadores $(\tau_n)_{n\in\mathbb{N}}$ se dice insesgado si

$$\mathbb{E}_{X_1^n}(\tau_n(X_1^n)) = \theta. \tag{3.14}$$

Es decir que en promedio el estimador se acerca al parámetro desconocido. El valor $|\mathbb{E}_{X_1^n}(\tau_n(X_1^n)) - \theta|$ se llama sesgo.

Una propiedad más débil sobre una familia de estimadores $(\tau_n)_{n\in\mathbb{N}}$ es el concepto de asintóticamente insesgado:

Definición 3.5. (Estimador Asintóticamente Insesgado) Sea $\theta \in \Theta$ arbitrario tal que $(X_1, ..., X_n) \sim P_{X_1^n}(\cdot|\theta)$, una secuencia de estimadores $(\tau_n)_{n\in\mathbb{N}}$ se dice asintóticamente insesgado si

$$\lim_{n\to\infty} \mathbb{E}_{X_1^n}(\tau_n(X_1^n)) = \theta, \tag{3.15}$$

Con estas definiciones ahora vamos a ver un ejemplo, donde veremos un caso concreto y además propondremos un estimador para ver su desempeño.

Ejemplo 3.3. Consideremos el caso de una distribución normal donde se poseen n variables de observación independientes e idénticamente distribuidas $X_i \sim N(\mu, \sigma^2)$ $\forall i \in \{1, ..., n\}$. Consideremos el estimador denominado *media empírica* definido como:

$$\tau_n(X_1^n) = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{3.16}$$

Podemos ver que este estimador es insesgado, ya que:

$$\begin{aligned}
\mathbb{E}_{X_1^n}(\tau_n(X_1^n)) &= \mathbb{E}_{X_1^n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{X_i}(X_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu \\
&= \mu
\end{aligned} \tag{3.17}$$

Ahora veremos que el estimador es consistente, para esto, recordemos las desigualdades de Markov y Chebyshev.

Teorema 3.1. (Desigualdad de Markov) Sea una variable aleatoria X a valores en $\mathbb{R}^+$ o 0 con esperanza finita $\mathbb{E}(X)$, tenemos la siguiente desigualdad conocida como desigualdad de Markov:

$$(\forall \epsilon > 0)\ P_X(X > \epsilon) \leq \frac{\mathbb{E}(X)}{\epsilon}. \tag{3.18}$$

Teorema 3.2. (Desigualdad de Chebyshev) Sea una variable aleatoria X con esperanza finita $\mathbb{E}(X)$ y $\mathbb{E}(X^2)$ finito, tenemos que:

$$(\forall \epsilon > 0)\ P_X(|X - \mathbb{E}(X)| > \epsilon) \leq \frac{Var(X)}{\epsilon^2}. \tag{3.19}$$

Ahora consideramos $\tau_n(X_1^n) = \frac{1}{n}\sum_{i=1}^{n} X_i$, entonces $\mathbb{E}(\tau_n(X_1^n)) = \mu$. Adicionalmente:

$$\begin{aligned} Var(\tau_n(X_1^n)) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var\left(X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 \\ &= \frac{\sigma^2}{n}. \end{aligned} \tag{3.20}$$

Por lo tanto, para $\epsilon > 0$, aplicamos la desigualdad de Chebyshev:

$$0 \leq P_{X_1^n}\left(|\tau_n(X_1^n) - \mu| > \epsilon\right) \leq \frac{Var(\tau_n(X_1^n))}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \xrightarrow{n\to\infty} 0. \tag{3.21}$$

Con lo anterior vemos que $\tau_n(X_1^n) \xrightarrow{P} \mu.$, por lo que $\tau_n(X_1^n)$ es un estimador consistente de μ.

Observaciones 3.3. En general no existe una equivalencia entre consistencia y sesgo, es decir, pueden existir estimadores insesgados pero no consistentes, o bien estimadores consistentes e insesgados. Sin embargo, existe un criterio suficiente para demostrar cuando un estimador es consistente a partir de uno asintóticamente insesgado.

De este ejemplo se puede demostrar un resultado que dice que si $(\tau_n)_{n\in\mathbb{N}}$ es insesgado y su varianza converge a cero cuando $n \to \infty$, entonces $(\tau_n)_{n\in\mathbb{N}}$ es consistente. A continuación vamos a demostrar una variante más general:

Teorema 3.3. Consideremos la familia de estimadores de θ, $(\tau_n)_{n\in\mathbb{N}}$ tal que cumple las siguientes condiciones:

- $\lim_{n\to\infty} \mathbb{E}_{X_1^n}(\tau_n(X_1^n)) = \theta$ (asintóticamente insesgado).
- $\lim_{n\to\infty} Var(\tau_n(X_1^n)) = 0$ (varianza desvaneciente).

entonces $(\tau_n)_{n\in\mathbb{N}}$ es consistente.

Demostración: Sea $\epsilon > 0$, por la desigualdad de Markov tenemos que

$$P_{X_1^n}\left(|\tau_n(X_1^n) - \theta| > \epsilon\right) \leq \frac{\mathbb{E}\left(|\tau_n(X_1^n) - \theta|^2\right)}{\epsilon^2}. \tag{3.22}$$

Analicemos más en detalle la siguiente expresión:

$$\begin{aligned}
&\mathbb{E}\{(\tau_n(X_1^n) - \theta)^2\} \\
&= \mathbb{E}\{(\tau_n(X_1^n) - \mathbb{E}\{\tau_n(X_1^n)\} + \mathbb{E}\{\tau_n(X_1^n)\} - \theta)^2\} \\
&= \mathbb{E}\left[(\tau_n(X_1^n) - \mathbb{E}\{\tau_n(X_1^n)\})^2 - 2(\tau_n(X_1^n) - \mathbb{E}\{\tau_n(X_1^n)\})(\mathbb{E}\{\tau_n(X_1^n)\} - \theta) + (\mathbb{E}\{\tau_n(X_1^n)\} - \theta)^2\right] \\
&= Var(\tau_n(X_1^n)) - 2(\mathbb{E}\{\tau_n(X_1^n)\} - \mathbb{E}\{\tau_n(X_1^n)\})(\mathbb{E}\{\tau_n(X_1^n)\} - \theta) + (\mathbb{E}\{\tau_n(X_1^n)\} - \theta)^2 \\
&= Var(\tau_n(X_1^n)) - 2 \cdot 0 \cdot (\mathbb{E}\{\tau_n(X_1^n)\} - \theta) + (\mathbb{E}\{\tau_n(X_1^n)\} - \theta)^2 \\
&= Var(\tau_n(X_1^n)) + (\mathbb{E}\{\tau_n(X_1^n)\} - \theta)^2
\end{aligned} \tag{3.23}$$

Entonces:

$$\frac{\mathbb{E}(|\tau_n(X_1^n) - \theta|^2)}{\epsilon^2} \leq \frac{Var(\tau_n(X_1^n)) + (\mathbb{E}(\tau_n(X_1^n)) - \theta)^2}{\epsilon^2} \xrightarrow{n\to\infty} 0 \tag{3.24}$$

Por lo que para $\epsilon > 0$ tenemos que:

$$\lim_{n\to\infty} P_{X_1^n}\left(|\tau_n(X_1^n) - \theta| > \epsilon\right) = 0, \tag{3.25}$$

lo que equivale a la definición de consistencia

$$\tau_n(X_1^n) \xrightarrow{P} \theta. \tag{3.26}$$

□

Se pueden establecer otras variantes del resultado anterior, el siguiente teorema presenta una idea similar:

Teorema 3.4. Sea $(\tau_n)_{n\in\mathbb{N}}$ asintóticamente insesgado que sigue la siguiente estructura, es decir,

$$\mathbb{E}_{X_1^n}(\tau_n(X_1^n)) = \theta + k_n \quad \text{donde } k_n \xrightarrow{n\to\infty} 0.$$

Si adicionalmente se tiene que:

$$\lim_{n\to\infty} Var(\tau_n(X_1^n)) = 0,$$

entonces $(\tau_n)_{n\in\mathbb{N}}$ es consistente.

Demostración: Sea $\epsilon > 0$, por la desigualdad de Chebyshev tenemos que

$$P_{X_1^n}\left(|\tau_n(X_1^n) - \theta| > \epsilon\right) \leq \frac{\mathbb{E}\left(|\tau_n(X_1^n) - \theta|^2\right)}{\epsilon^2}. \tag{3.27}$$

Analicemos más en detalle la siguiente expresión:

$$\begin{aligned}(\tau_n(X_1^n) - \theta)^2 &= (\tau_n(X_1^n) - \mathbb{E}(\tau_n(X_1^n)) + k_n)^2 \\ &= k_n^2 - 2k_n(\tau_n(X_1^n) - \mathbb{E}(\tau_n(X_1^n))) + (\tau_n(X_1^n) - \mathbb{E}(\tau_n(X_1^n)))^2 \end{aligned} \tag{3.28}$$

Tomando esperanza en (3.28) y aplicando esto en (3.27)

$$\begin{aligned}\frac{\mathbb{E}(|\tau_n(X_1^n) - \theta|^2)}{\epsilon^2} &\leq \frac{k_n^2 + 2k_n\mathbb{E}_{X_1^n}(\tau_n(X_1^n) - \mathbb{E}(\tau_n(X_1^n))) + Var(\tau_n(X_1^n))}{\epsilon^2} \\ &= \frac{k_n^2 + Var(\tau_n(X_1^n))}{\epsilon^2} \xrightarrow{n\to\infty} 0 \end{aligned} \tag{3.29}$$

Por lo que nuevamente para $\epsilon > 0$ tenemos que:

$$\lim_{n\to\infty} P_{X_1^n}\left(|\tau_n(X_1^n) - \theta| > \epsilon\right) = 0, \tag{3.30}$$

lo que equivale a la definición de consistencia

$$\tau_n(X_1^n) \xrightarrow{P} \theta. \tag{3.31}$$

□

Para concluir esta sección veremos un ejemplo adicional sobre la distribución normal.

Ejemplo 3.4. Nuevamente consideremos el caso de una distribución normal donde se poseen n variables de observaciones independientes e idénticamente distribuidas $X_i \sim N(\mu, \sigma^2)\ \forall i \in \{1, ..., n\}$. Propondremos un estimador de la varianza conocido como la varianza empírica:

$$\tau_n^{\sigma^2}(X_1^n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

donde $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ es la media empírica y sabemos que $\mathbb{E}(\overline{X}_n)$ y $Var(\overline{X}_n) = \frac{\sigma^2}{n}$.

Mostraremos que τ_N^{σ} es sesgado pero que $(\tau_n)_{n\in\mathbb{N}}$ es asintoticamente insesgado.

$$\begin{aligned}
\mathbb{E}\left(\tau_n^{\sigma^2}(X_1^n)\right) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\right) \\
&= \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n}(X_i^2 - 2X_i\overline{X}_n + \overline{X}_n^2)\right) \\
&= \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n}X_i^2 - \frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}X_iX_j + \overline{X}_n^2\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}\mathbb{E}(X_i^2) - \frac{2}{n}\mathbb{E}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}X_iX_j\right) + n\mathbb{E}(\overline{X}_n^2)\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}\mathbb{E}(X_i^2) - \frac{2}{n}\mathbb{E}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}X_iX_j\right) + nVar(\overline{X}_n) + n(\mathbb{E}\left(\overline{X}_n\right))^2\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}\mathbb{E}(X_i^2) - \frac{2}{n}\left(\mathbb{E}\left(\sum_{i=1}^{n}X_i^2\right) + \mathbb{E}\left(\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}X_iX_j\right)\right) + \sigma^2 + n\mu^2\right) \\
&= \frac{1}{n}\left(\frac{(n-2)}{n}\sum_{i=1}^{n}(Var(X_i) + (\mathbb{E}(X_i))^2) - \frac{2}{n}\left(\mathbb{E}\left(\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}X_iX_j\right)\right) + \sigma^2 + n\mu^2\right)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n}\left(\frac{(n-2)}{n}\sum_{i=1}^{n}(\sigma^2+\mu^2) - \frac{2(n^2-n)\mu}{n} + \sigma^2 + n\mu^2\right) \\
&= \frac{1}{n}\left((n-2)(\sigma^2+\mu^2) - 2(n-1)\mu + \sigma^2 + n\mu^2\right) \qquad (3.32) \\
&= \frac{1}{n}\left((n-2)\sigma^2 + \sigma^2 + ((n-2) - 2(n-1) + n)\mu\right) \\
&= \frac{n-1}{n}\sigma^2 \qquad (3.33)
\end{aligned}
$$

Notar que se ocupó al propiedad que $Var(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$. Este estimador es sesgado, pero es reparable, si proponemos el siguiente estimador

$$\tau_n^I(X_1^n) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2, \qquad (3.34)$$

esta función corresponde a un estimador de σ^2 insesgado.

El desarrollo antes hecho ocupó desarrollo algebraicos sobre las sumatorias, existe otro camino un tanto más corto, por completitud será desarrollado nuevamente ocupando este método:

$$
\begin{aligned}
\mathbb{E}\left(\tau_n^{\sigma^2}(X_1^n)\right) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left((X_i - \overline{X}_n)^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} Var\left(X_i - \overline{X}_n\right) + \left(\mathbb{E}\left(X_i - \overline{X}_n\right)\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} Var\left(X_i - \overline{X}_n\right) + (\mu - \mu)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} Var\left(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} Var\left(\frac{n-1}{n}X_i - \frac{1}{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} X_j\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n}\left(\left(\frac{n-1}{n}\right)^2 Var\left(X_i\right) + Var\left(\frac{1}{n}\sum_{\substack{j=1\\ j\neq i}}^{n} X_j\right)\right)\\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\left(\frac{n-1}{n}\right)^2 Var\left(X_i\right) + \frac{(n-1)\sigma^2}{n^2}\right)\\
&= \frac{(n-1)^2}{n^2}\sigma^2 + \frac{(n-1)\sigma^2}{n^2}\\
&= \frac{n-1}{n}\sigma^2
\end{aligned}
\tag{3.35}
$$

Propuesto 3.1. Verifique que $\tau_n^{\sigma^2}(X_1^n)$ y $\tau_n^I(X_1^n)$ son estimadores consistentes de σ^2.

Los ejemplos anteriores nos muestran distintos estimadores en casos bien particulares de distribuciones. Si bien los criterios de optimalidad antes descritos son muy usados en estadística, existe un criterio adicional respecto a la varianza del estimador. En la sección siguiente queremos establecer un nuevo criterio para seleccionar un estimador en el sentido de mínima varianza.

Veremos que existe un límite fundamental (una cota) para la varianza de la familia de estimadores insesgados.

3.3. El Criterio de Mínima Varianza

Sabemos que para entender el comportamiento de una variable aleatoria podemos analizar dos medidas importantes; esperanza y varianza. Si la esperanza existe nos gustaría además ver si es que la variable aleatoria está concentrada entorno a ese valor o no. De manera análoga, si un estimador es insesgado, quisiéramos saber si la varianza sel estimador es lo suficientemente pequeña, así para poder dar garantías que el estimador tiene alta precisión y está concentrado en el parámetro desconocido. Cuando un estimador insesgado es de mínima varianza se le dice eficiente.

En adelante, si nos concentramos en la familia de estimadores insesgados, una pregunta fundamental es caracterizar el estimador de minima varianza, es decir, si es

posible, dada una familia de distribuciones paramétricas, encontrar aquél estimador insesgado que entregue la menor varianza posible. En esta linea uno de los resultados centrales de la teoría de estimación es la celebrada cota de Cramer-Rao que ofrece una expresión analítica para acotar la mínima varianza en un contexto de estimación paramétrica.

Consideremos nuevamente una familia de distribuciones indexadas por $\theta \in \Theta$:

$$\mathcal{F}_\Theta = \{P_X(\cdot|\theta) : \theta \in \Theta\}.$$

Nuevamente utilizaremos la función de verosimilitud que, dependiendo de si la distribución es discreta o continua, tendremos dos casos:

- Si X_1^n siguen una distribución discreta con función de probabilidad de masa conjunta $P_{X_1^n}(\cdot)$ la verosimilitud se define como, dado $x_1^n \in \mathbb{X}^n$:
$$L(X_1 = x_1, X_2 = x_2, ..., X_n = x_n|\theta) = p_{X_1^n}(x_1, x_2, ..., x_n|\theta).$$
- Si X_1^n siguen una distribución continua con función de densidad conjunta $f_{X_1^n}(\cdot)$ la verosimilitud se define como, dado $x_1^n \in \mathbb{X}^n$:
$$L(X_1 = x_1, X_2 = x_2, ..., X_n = x_n|\theta) = f_{X_1^n}(x_1, x_2, ..., x_n|\theta).$$
Recordar que la función de verosimilitud es una función de θ ya que las obsrvaciones están fijas, sin embargo, es una variable aleatoria al tener dependencia del vector X_1^n.

Recordar que la verosimilutd es función de θ, para $x_1^n \in \mathbb{X}^n$ y también se puede ver como un objeto aleatorio. Sin pérdida de generalidad, en los desarrollos siguientes asumiremos que tenemos una familia de distribuciones continuas en $\mathcal{F}_\Theta$ (el resultado es análogo para el caso discreto), entonces:

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1^n}(x_1, ..., x_n|\theta) dx_1 ... dx_n = 1. \tag{3.36}$$

Asumiendo que estamos en las condiciones de regularidad dada la Definición 3.2, se tiene la siguiente identidad:

$$\begin{aligned}
&\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\partial f_{X_1^n}(x_1, ..., x_n|\theta)}{\partial \theta} dx_1 ... dx_n = 0 \\
\Leftrightarrow &\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left[\frac{1}{f_{X_1^n}(x_1, ..., x_n|\theta)} \frac{\partial f_{X_1^n}(x_1, ..., x_n|\theta)}{\partial \theta} \right] f_{X_1^n}(x_1, ..., x_n|\theta) dx_1 ... dx_n = 0 \\
\Leftrightarrow &\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} \left(\ln f_{X_1^n}(x_1, ..., x_n|\theta) \right) f_{X_1^n}(x_1, ..., x_n|\theta) dx_1 ... dx_n = 0,
\end{aligned} \tag{3.37}$$

notar que la expresión en (3.37) es equivalente a:

$$\mathbb{E}_{X_1^n}\left(\frac{\partial \ln f_{X_1^n}(X_1,...,X_n|\theta)}{\partial\theta}\right)=0. \tag{3.38}$$

El resultado en (3.38) corresponde a una condición de regularidad que se debe verificar para los desarrollos posteriores. La función

$$\frac{\partial \ln f_{X_1^n}(X_1,...,X_n|\theta)}{\partial\theta}, \tag{3.39}$$

también suele llamarse score y sirve para medir la sensibilidad del logaritmo de la verosimilitud respecto a cambios infinitesimales de θ.

Ahora consideremos un estimador del parámetro θ arbitrario, dado por $\tau_n(\cdot):\mathbb{X}^n\to\Theta$ y, sin perdida de generalidad, que:

$$\mathbb{E}_{X_1^n}\left(\tau_n(X_1^n)\right)=f(\theta)\quad\forall\theta\in\Theta. \tag{3.40}$$

Es decir, que el sesgo es una función de θ. Asumiendo que $f(\theta)$ es diferenciable, y derivando (3.40), tenemos que:

$$\forall\theta\in\Theta\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\tau_n(x_1,...,x_n)\frac{\partial \ln f_{X_1^n}(x_1,...,x_n|\theta)}{\partial\theta}\cdot f_{X_1^n}(x_1,...,x_n|\theta)dx_1...dx_n=f'(\theta). \tag{3.41}$$

Por otro lado, de (3.37), multiplicando por $f(\theta)$ tenemos que:

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}f(\theta)\frac{\partial \ln f_{X_1^n}(x_1,...,x_n|\theta)}{\partial\theta}f_{X_1^n}(x_1,...,x_n|\theta)dx_1...dx_n=0, \tag{3.42}$$

restando (3.41) y (3.42):

$$\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}(\tau_n(x_1,...,x_n)-f(\theta))\frac{\partial \ln f_{X_1^n}(x_1,...,x_n|\theta)}{\partial\theta}f_{X_1^n}(x_1,...,x_n|\theta)dx_1...dx_n=f'(\theta). \tag{3.43}$$

La esperanza induce un producto interno, con tal contexto podremos hacer uso de la desigualdad de Cauchy-Schwarz que dice que para dos variables aleatorias X e Y:

$$|\mathbb{E}(XY)|^2\leq\mathbb{E}(X^2)\mathbb{E}(Y^2). \tag{3.44}$$

Por lo tanto aplicando (3.44) en (3.43)

$$\begin{aligned}(f'(\theta))^2 &= \left|\mathbb{E}_{X_1^n}\left((\tau_n(X_1^n) - f(\theta)) \cdot \frac{\partial \ln f_{X_1^n}(x_1, ..., x_n|\theta)}{\partial\theta}\right)\right|^2 \\ &\leq \mathbb{E}_{X_1^n}\left((\tau_n(X_1^n) - f(\theta))^2\right) \cdot \mathbb{E}_{X_1^n}\left(\left(\frac{\partial \ln f_{X_1^n}(X_1, ..., X_n|\theta)}{\partial\theta}\right)^2\right) \\ &= Var(\tau_n(X_1^n)) \cdot \mathbb{E}_{X_1^n}\left(\left(\frac{\partial \ln f_{X_1^n}(X_1, ..., X_n|\theta)}{\partial\theta}\right)^2\right), \end{aligned} \tag{3.45}$$

que es equivalente a decir que:

$$Var(\tau_n(X_1^n)) \geq \frac{f'(\theta)^2}{\mathbb{E}_{X_1^n}\left(\left(\frac{\partial \ln L(X_1,...,X_n|\theta)}{\partial\theta}\right)^2\right)}. \tag{3.46}$$

La expresión en (3.46) corresponde a una cota inferior para la varianza de la familia de estimadores dada la familia de distribuciones $\mathcal{F}_\Theta$. Notar que esta cota es universal en el sentido que es válida para cualquier estimador sobre la familia de modelos paramétricos indexados por $\theta \in \Theta$. Cuando se obtienen resultados del tipo universales se entienden también como *Límites Fundamentales*. En la sección que viene aplicaremos este resultado sobre los estimadores insesgados, lo que nos dará una cota fundamental ampliamente usada en estadística.

3.4. La Información de Fisher y La Cota de Cramér-Rao

En la desigualdad en (3.46) aparece un término en el denominador que le daremos especial atención, se conoce como la Información de Fisher y será el elemento central como cota para los estimadores insesgados.

Definición 3.6. (Información de Fisher) Bajo las condiciones de regularidad de la Definición 3.2, se define la Información de Fisher de $\mathcal{F}_\Theta$ asociada a n observaciones como:

$$\mathcal{I}_n(\theta) \triangleq \mathbb{E}_{X_1^n}\left(\left(\frac{\partial \ln L(X_1, ..., X_n|\theta)}{\partial\theta}\right)^2\right) \tag{3.47}$$

Observaciones 3.4.

- La Información de Fisher depende exclusivamente de la familia $\mathcal{F}_\Theta$ y de el número de observaciones, es decir, $\mathcal{I}_n(\theta)$ es independiente del estimador τ_n
- La desigualdad en (3.46) ofrece una cota inferior para la varianza del estimador $\tau_n(\cdot)$ sujeto a la condición en (3.40).

Además, gracias a la propiedad de la derivada del logaritmo y las condiciones de regularidad de la Definición 3.2, tenemos que la Información de Fisher se puede expresar de otra manera:

Proposición 3.1. Bajo las condiciones de regularidad de la Definición 3.2, tenemos que la Información de Fisher se puede expresar como:

$$\mathbb{E}_{X_1^n}\left(\left(\frac{\partial \ln L(X_1,...,X_n|\theta)}{\partial\theta}\right)^2\right) = -\mathbb{E}_{X_1^n}\left(\frac{\partial^2 \ln L(X_1,...,X_n|\theta)}{\partial\theta^2}\right). \tag{3.48}$$

Demostración: Nuevamente asumiremos el caso continuo para la verosimilitud. Desarro-

llaremos la expresión del lado derecho:

$$
\begin{aligned}
&- \mathbb{E}_{X_1^n}\left(\frac{\partial^2 \ln L(X_1,...,X_n|\theta)}{\partial\theta^2}\right)\\
&= -\mathbb{E}_{X_1^n}\left(\frac{\partial}{\partial\theta}\left[\frac{1}{f_{X_1^n}(X_1,...,X_n|\theta)}\frac{\partial f_{X_1^n}(X_1,...,X_n|\theta)}{\partial\theta}\right]\right)\\
&= -\mathbb{E}_{X_1^n}\left(-\frac{1}{(f_{X_1^n}(X_1,...,X_n|\theta))^2}\left(\frac{\partial f_{X_1^n}(X_1,...,X_n|\theta)}{\partial\theta}\right)^2+\frac{1}{f_{X_1^n}(X_1,...,X_n|\theta)}\frac{\partial^2 f_{X_1^n}(X_1,...,X_n|\theta)}{\partial\theta^2}\right)\\
&= -\mathbb{E}_{X_1^n}\left(-\left(\frac{\partial \ln L(X_1,...,X_n|\theta)}{\partial\theta}\right)^2+\frac{1}{f_{X_1^n}(X_1,...,X_n|\theta)}\frac{\partial^2 f_{X_1^n}(X_1,...,X_n|\theta)}{\partial\theta^2}\right)\\
&= \mathbb{E}_{X_1^n}\left(\left(\frac{\partial \ln L(X_1,...,X_n|\theta)}{\partial\theta}\right)^2\right)+\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{\partial^2 f_{X_1^n}(x_1,...,x_n|\theta)}{\partial\theta^2}dx_1...dx_n\\
&= \mathbb{E}_{X_1^n}\left(\left(\frac{\partial \ln L(X_1,...,X_n|\theta)}{\partial\theta}\right)^2\right)+\frac{\partial^2}{\partial\theta^2}\left[\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}f_{X_1^n}(x_1,...,x_n|\theta)dx_1...dx_n\right]\\
&= \mathbb{E}_{X_1^n}\left(\left(\frac{\partial \ln L(X_1,...,X_n|\theta)}{\partial\theta}\right)^2\right)+\frac{\partial^2}{\partial\theta^2}(1)\\
&= \mathbb{E}_{X_1^n}\left(\left(\frac{\partial \ln L(X_1,...,X_n|\theta)}{\partial\theta}\right)^2\right) \qquad (3.49)
\end{aligned}
$$

□

Una segunda propiedad es la aditividad de la Información de Fisher bajo distribuciones i.i.d.

Proposición 3.2. Bajo las condiciones de regularidad de la Definición 3.2, y considerando un vector aleatorio X_1^n i.i.d. tenemos que la Información de Fisher se puede expresar como:

$$\mathcal{I}_n(\theta) = \mathbb{E}_{X_1^n}\left(\left(\frac{\partial \ln L(X_1,X_2,...,X_n|\theta)}{\partial\theta}\right)^2\right) = n\mathcal{I}_1(\theta). \qquad (3.50)$$

donde

$$\mathcal{I}_1(\theta) = \mathbb{E}_{X_1}\left(\left(\frac{\partial \ln L(X_1|\theta_0)}{\partial\theta}\right)^2\right). \qquad (3.51)$$

Demostración:

$$\begin{aligned}
\mathcal{I}_n(\theta) = \mathbb{E}_{X_1^n}\left(\left(\frac{\partial \ln L(X_1, X_2, ..., X_n|\theta)}{\partial\theta}\right)^2\right) &= -\mathbb{E}_{X_1^n}\left(\frac{\partial^2 \ln L(X_1, ..., X_n|\theta)}{\partial\theta^2}\right) \\
&= -\mathbb{E}_{X_1^n}\left(\frac{\partial^2 \ln \prod_{i=1}^{n} L(X_i|\theta)}{\partial\theta^2}\right) \\
&= -\sum_{i=1}^{n}\mathbb{E}_{X_1^n}\left(\frac{\partial^2 \ln L(X_i|\theta)}{\partial\theta^2}\right) \\
&= -n\mathbb{E}_{X_1}\left(\frac{\partial^2 \ln L(X_1|\theta)}{\partial\theta^2}\right) \\
&= n\mathbb{E}_{X_1}\left(\left(\frac{\partial \ln L(X_1|\theta_0)}{\partial\theta_0}\right)^2\right). \qquad (3.52)
\end{aligned}$$

□

El resultado en (3.46) puede particularizarse al caso de los estimadores insesgados lo que nos llevará a la celebrada cota de Cramér-Rao. Formalmente consideremos la familia de estimadores insesgados, es decir, se tiene la siguiente familia:

$$\mathcal{T}_n \triangleq \{\tau_n : \mathbb{X}^n \to \Theta : \quad \mathbb{E}_{X_1^n}\left(\tau_n(X_1^n)\right) = \theta, \forall\theta \in \Theta\} \qquad (3.53)$$

entonces, observando que $f(\theta) = \theta \Rightarrow f'(\theta) = 1$, tenemos que $\forall\tau_n \in \mathcal{T}_n$:

$$Var(\tau_n(X_1^n)) \geq \frac{1}{\mathcal{I}_n(\theta)}. \qquad (3.54)$$

Dado que la cota es independiente de τ_n, entonces en particular se cumple para el estimador que minimice la varianza, es decir:

$$(\forall\theta \in \Theta) \qquad \min_{\tau_n \in \mathcal{T}_n} Var(\tau_n(X_1^n)) \geq \frac{1}{\mathbb{E}_{X_1^n}\left(\left(\frac{\partial \ln L(X_1,...,X_n|\theta)}{\partial\theta}\right)^2\right)}. \qquad (3.55)$$

Este resultado es conocido como la desigualdad de Cramér-Rao. Ahora enunciaremos este teorema emblemático:

Teorema 3.5. (Desigualdad de Crámer-Rao) Sea $\mathcal{T}_n$ la familia de estimadores insesgados de n observaciones sobre una familia de distribuciones $P_{X_1^n}(\cdot|\theta)$, $\theta \in \Theta$, entonces, asumiendo las condiciones de regularidad de la Definición 3.2, tenemos que:

$$(\forall \theta \in \Theta) \qquad \min_{\tau_n \in \mathcal{T}_n} \mathbb{E}_{X_1^n}\left((\tau_n(X_1^n) - \theta)^2\right) \geq \frac{1}{\mathcal{I}_n(\theta)}, \tag{3.56}$$

donde $\mathcal{I}_n(\theta)$ esta dada por:

$$\mathcal{I}_n(\theta) = \mathbb{E}_{X_1^n}\left(\left(\frac{\partial \ln L(X_1, ..., X_n|\theta)}{\partial \theta}\right)^2\right). \tag{3.57}$$

Observaciones 3.5.

- La información de Fisher se interpreta como la cantidad de información promedio que ofrecen las observaciones $(X_1, ..., X_n)$ para estimar el parámetro θ en un sentido de varianza.
- La desigualdad de Cramér-Rao ofrece una cota inferior para la mínima varianza de estimadores insesgados.
- Equivalentemente es una cota para el mínimo error cuadrático medio de estimadores insesgados (ya que el sesgo es 0).
- Si existe familia de estimadores insesgados por medio de la siguiente condición:

  $$\lim_{n\to\infty} Var(\tau_n(X_1^n)) = 0, \tag{3.58}$$

  entonces se tiene de (3.56) que necesariamente

  $$\lim_{n\to\infty} \mathcal{I}_n(\theta) = \infty. \tag{3.59}$$

- La cota de Cramér-Rao por lo general no se alcanza, esto significa entonces que, para una familia de distribuciones paramétricas, todos los estimadores insesgados serán estrictamente mayores que la cota. No obstante, la cota de Cramér-Rao sigue siendo un indicador de mínima precisión y lo que se busca en la práctica es que el estimador pueda estar lo más cercano posible a tal valor. En la siguiente subsección veremos condiciones necesarias y suficientes para alcanzar la cota.

3.4.1. Condiciones de Alcanzabilidad de la Cota Cramér-Rao

Este resultado da condiciones necesarias y suficientes para poder alcanzar la cota de Cramér-Rao. Se tiene el siguiente Teorema:

Teorema 3.6. (Condiciones de Alcanzabilidad de la Cota de Cramér-Rao) La cota de Cramér-Rao es alcanzable por un estimador insesgado, si y solo si, existe una función $\tau_n : \mathbb{X}^n \to \Theta$ (exclusiva de las observaciones y que no dependa del parámetro) tal que para todo $\theta \in \Theta$:

$$\frac{\partial \ln L(X_1, ..., X_n|\theta)}{\partial \theta} = \mathcal{I}_n(\theta) \cdot (\tau_n(X_1^n) - \theta). \tag{3.60}$$

En este caso el estimador de mínima varianza es $\tau_n(X_1^n)$ y su minima varianza $Var(\tau_n(X_1^n))$ es $\frac{1}{\mathcal{I}_n(\theta)}$.

Demostración: Del uso de la desigualdad de *Cauchy-Schwarz* una condición necesaria y suficiente para alcanzar la igualdad en (3.46), y en consecuencia la existencia de un estimador que alcance la cota de Cramér-Rao, es que $\frac{\partial \ln L(X_1,...,X_n|\theta)}{\partial \theta}$ sea colineal a $(\tau_n(X_1^n) - \theta)$ en el sentido que exista una función $A(\theta) \in \mathbb{R}$, donde:

$$\frac{\partial \ln L(X_1, ..., X_n|\theta)}{\partial \theta} = A(\theta) \cdot (\tau_n(X_1^n) - \theta), \tag{3.61}$$

con $A(\theta)$ es una constante que puede o no depender de θ. Reemplazando esta condición de co-linealidad en (3.43) se obtiene que:

$$Var(\tau_n(X_1^n)) = \frac{1}{A(\theta)}, \tag{3.62}$$

si además (3.61) se eleva al cuadrado y se toma esperanza tendremos que:

$$\mathbb{E}_{X_1^n}\left(\left(\frac{\partial \ln L(X_1, .., X_n|\theta)}{\partial \theta}\right)^2\right) = A(\theta)^2 \cdot Var(\tau_n(X_1^n)), \tag{3.63}$$

y reemplazando lo obtenido en (3.62) concluimos que:

$$A(\theta) = \mathcal{I}_n(\theta). \tag{3.64}$$

□

Observaciones 3.6. En general la familia de funciones que ofrecen la descomposición en (3.60) son de tipo exponenciales, por lo tanto, este resultado está bien restringido a dichas familias que permiten obtener soluciones analíticas y cerradas.

A continuación veremos dos ejemplos que nos ayudarán a ilustrar y demostrar la alcanzabilidad de la cota de Cramér-Rao en estos escenarios.

Ejemplo 3.5. Consideremos el caso de n variables de observaciones i.i.d. $(X_1, ..., X_n) \sim P_X(\cdot|\theta)^n$. tal que $X_i \sim N(\theta, \sigma^2)$, es decir, la densidad marginal asociada a cada observación x está dada por:

$$P_X(\cdot|\theta) \mapsto f_X(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\theta)^2}{2\sigma^2}} \quad \forall x \in \mathbb{R}. \tag{3.65}$$

donde σ es conocido y queremos estimar θ. Vemos que el logaritmo de la verosimilitud está dado por:

$$\begin{aligned} \ln L(X_1, ..., X_n|\theta) &= \ln\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X_i-\theta)^2}{2\sigma^2}}\right) \\ &= n \ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \sum_{i=1}^{n} \frac{(X_i-\theta)^2}{2\sigma^2}. \end{aligned} \tag{3.66}$$

y vemos que:

$$\frac{\partial \ln L(X_1, ..., X_n|\theta)}{\partial\theta} = \underbrace{\left(\frac{n}{\sigma^2}\right)}_{\mathcal{I}_n(\theta)} \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \theta\right)}_{(\tau_n^*(X_1^n)-\theta)} \tag{3.67}$$

y dado que $\mathbb{E}(\tau_n^*(X_1^n)) = \theta$ (insesgado), entonces se cumple la descomposición de (3.60). Por lo que el estimador de mínima varianza es $\tau_n^*(X_1^n) = \frac{1}{n}\sum_{i=1}^{n} X_i$ y alcanza la cota de Cramér-Rao cuyo valor es:

$$Var(\tau_n^*(X_1^n)) = \frac{1}{\mathcal{I}_n(\theta)} = \frac{\sigma^2}{n}. \tag{3.68}$$

El estimador $\tau_n^*(X_1^n) = \frac{1}{n}\sum_{i=1}^{n} X_i$ es de mínima varianza.

Ejemplo 3.6. Consideremos el caso de n variables de observación i.i.d. $(X_1, ..., X_n) \sim P_X(\cdot|\theta)^n$. tal que $X_i \sim Poisson(\theta)$, es decir, la función de probabilidad de masa asociada a cada observación x está dada por:

$$P_X(\cdot|\theta) \mapsto P_X(X = x|\theta) = \frac{e^{-\theta}\theta^x}{x!} \ \forall x \in \mathbb{N}. \tag{3.69}$$

Queremos estimar θ a partir de $(X_1, ..., X_n) \sim P_X(\cdot|\theta)^n$ (i.i.d.). Vemos que el logaritmo de la verosimilitud está dado por:

$$\begin{aligned} \ln L(X_1, ..., X_n|\theta) &= \ln\left(\prod_{i=1}^{n} \frac{e^{-\theta}\theta^{X_i}}{X_i!}\right) \\ &= \sum_{i=1}^{n} X_i \ln(\theta) - \ln(X_i!) - \theta. \end{aligned} \tag{3.70}$$

y vemos que:

$$\frac{\partial \ln L(X_1, ..., X_n|\theta)}{\partial\theta} = \underbrace{\left(\frac{n}{\theta}\right)}_{\mathcal{I}_n(\theta)} \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \theta\right)}_{(\tau_n^*(X_1^n)-\theta)}. \tag{3.71}$$

Dado que $\mathbb{E}(\tau_n^*(X_1^n)) = \theta$ (insesgado), entonces nuevamente se cumple la descomposición de (3.60). Por lo que el estimador de mínima varianza es $\tau_n^*(X_1^n) = \frac{1}{n}\sum_{i=1}^{n} X_i$ y alcanza la cota de Cramér-Rao cuyo valor es:

$$Var(\tau_n^*(X_1^n)) = \frac{1}{\mathcal{I}_n(\theta)} = \frac{\theta}{n}. \tag{3.72}$$

En un número mayoritario de escenarios de inferencia paramétrica, la familia de distribuciones $\mathcal{F}_\Theta$ no ofrecen la descomposición en (3.60) y por lo tanto

$$\min_{\tau_n \in \mathcal{T}_n} \mathbb{E}_{X_1^n}\left((\tau_n(X_1^n) - \theta)^2\right) > \frac{1}{\mathcal{I}_n(\theta)}. \tag{3.73}$$

De todas formas $\mathcal{I}_n(\theta)^{-1}$ se utiliza como una figura de mérito o indicador para, por un lado, evaluar que tan lejos es el desempeño de un estimador insesgado de la cota de Cramér-Rao de mínima varianza y, por otro lado, como un indicador de la complejidad del problema de inferencia y como este límite escala como función del numero de observaciones.

Podemos ver entonces que la cota de Cramér-Rao actúa como un límite fundamental del problema de estimación paramétrica. Es importante recalcar que esta desigualdad es válida solamente para la familia de estimadores insesgados por lo que de debe verificar dicha condición previamente. Para una cota más general para estimadores sesgados se puede utilizar la expresión en (3.46), la cual requiere poder calcular la derivada de $f(\theta)$.

3.4.2. Unicidad del Estimador de Mínima Varianza

Dado que no siempre se puede alcanzar la cota de Cramér-Rao, se puede buscar un desempeño entre distintos estimadores. El siguiente resultado es importante para el análisis numérico a la hora de buscar un estimador de mínima varianza y dice lo siguiente:

Teorema 3.7. Dada una familia $\mathcal{F}_\Theta$, si existe un estimador insesgado de minimiza varianza entonces es único (casi seguramente).

Demostración: Sea $\theta \in \Theta$, y un vector variable de observaciones $(X_1, ..., X_n) \sim P_{X_1^n}(\cdot|\theta)$ supongamos que existen dos estimadores $\tau_1(X_1^n)$ y $\tau_2(X_1^n)$ con $\tau_1 \neq \tau_2$ tales que son de mínima varianza, es decir solución del problema:

$$\min_{\tau_n \in \mathcal{T}_n} Var(\tau_n(X_1^n)) = V_0, \tag{3.74}$$

Sobre estos dos estimadores podemos proponer un tercer estimador definido como:

$$\tau_3(X_1^n) = \frac{1}{2}\tau_1(X_1^n) + \frac{1}{2}\tau_2(X_1^n) \tag{3.75}$$

Claramente $\tau_3(X_1^n)$ es insesgado ya que $\tau_1(X_1^n)$ y $\tau_2(X_1^n)$ lo son y, por lo tanto, $\tau_3 \in \mathcal{T}_n$. Al calcular su varianza tenemos que:

$$\begin{aligned} Var(\tau_3(X_1^n)) &= \mathbb{E}((\tau_3(X_1^n) - \theta)^2) \\ &= \frac{1}{4}(Var(\tau_1(X_1^n)) + Var(\tau_2(X_1^n)) + 2Cov(\tau_1(X_1^n), \tau_2(X_1^n))). \end{aligned}$$

Notando que, además,

$$\begin{aligned} (Cov(\tau_1(X_1^n), \tau_2(X_1^n)))^2 &= \left|\mathbb{E}\left[(\tau_1(X_1^n) - \theta)\,(\tau_2(X_1^n) - \theta)\right]\right|^2 \\ &\leq Var(\tau_1(X_1^n))Var(\tau_2(X_1^n)) \qquad \backslash\text{Cauchy-Schwarz} \\ &= V_0^2 \end{aligned} \tag{3.76}$$

Finalmente se tiene que:

$$Var(\tau_3(X_1^n)) \leq \frac{1}{4}(V_0 + V_0 + 2V_0) = V_0 \tag{3.77}$$

La desigualdad estricta no es factible pues contradice el hecho que $\tau_1(X_1^n)$ y $\tau_2(X_1^n)$ son estimadores de mínima varianza. Por lo tanto, $Var(\tau_3(X_1^n)) = V_0$, en ese sentido, la desigualdad de Cauchy-Schwarz se cumple con igualdad y necesariamente son linealmente dependientes, es decir,

$$\tau_1(X_1^n) - \theta = k_0(\tau_2(X_1^n) - \theta) \tag{3.78}$$

para cierto $k_0 \in \mathbb{R}$, reemplazando (3.78) en (3.76) obtenemos

$$k_0^2 V_0^2 = V_0^2 \Rightarrow k_0^2 = 1 \tag{3.79}$$

por lo tanto $\tau_1(X_1^n) = \tau_2(X_1^n)$, lo que contradice la hipótesis. Concluimos que el estimador de mínima varianza es único casi seguramente. □

3.5. Estimador de Máxima Verosimilitud

En la sección anterior encontramos un límite fundamental para un estimador insesgado de mínima varianza. Sin embargo, poco se ha dicho para obtener un estimador a partir de observaciones.

En esta sección veremos un criterio concreto de selección de parámetros (y por lo tanto obtener un estimador). Uno de los principios clásicos es el criterio de **máxima verosimilitud**.

El estimador de máxima verosimilitud o *maximum likelihood* esencialmente pide elegir, como valor estimado del parámetro, aquél parámetro donde la probabilidad de haber obtenido la observación recibida sea la mayor posible. Esto es, teniendo observaciones $x_1^n \in \mathbb{X}^n$, uno *mira hacia atrás* y calcula la probabilidad -desde el punto de vista quien realiza el experimento- que la muestra obtenida es observada.

Con lo dicho anteriormente podemos dar la definición formal de este estimador:

Definición 3.7. (Estimador de Máxima Verosimilitud) Consideremos nuevamente la familia paramétrica $\mathcal{F}_\Theta$, sea $\theta \in \Theta$ y un vector variable de observaciones $(X_1, ..., X_n) \sim P_{X_1^n}(\cdot|\theta)$, el estimador de máxima verosimilitud $\tau_{ML} : \mathbb{X}^n \to \Theta$ se define como:

$$\tau_{ML}(X_1^n) = \arg\max_{\theta \in \Theta} L(X_1, ..., X_n|\theta). \tag{3.80}$$

Lo que corresponde a una variable aleatoria. Notar que, en términos prácticos, se posee un conjunto de observaciones que ya están dados, y luego se denota por $X_1^n = x_1^n$, por lo que el estimador se suele escribir como:

$$\tau_{ML}(x_1^n) = \arg\max_{\theta\in\Theta} L(X_1 = x_1, ..., X_n = x_n|\theta). \tag{3.81}$$

Donde $\arg\max_{\theta\in\Theta} f(\theta)$ corresponde a el argumento $\theta \in \Theta$ que maximiza la función f. Normalmente las familias a optimizar son exponenciales, luego es conveniente aplicarle logaritmo[2] y trabajar sobre la log-verosimilitud, con esto el estimador de máxima verosimilitud también puede definirse como:

$$\tau_{ML}(X_1^n) = \arg\max_{\theta\in\Theta} \ln(L(X_1, ..., X_n|\theta)), \tag{3.82}$$

Podemos observar que al aplicar logaritmo, al ser una función creciente estricta, el resultado del estimador no cambia ya que estamos buscando el argumento que maximiza la función y no el valor máximo de la función.

Notamos entonces que el estimador de máxima verosimilitud equivale a encontrar el parámetro en Θ que mejor describa los datos en un sentido de probabilidad. En otras palabras el objetivo es encontrar el parámetro que hace las observaciones más probables dentro de la familia $\mathcal{F}_\Theta$.

Analizando más en detalle, el vector X_1^n suelen ser independiente, por lo que el estimador de máxima verosimilitud se puede escribir como:

$$\tau_{ML}(X_1^n) = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \ln(L(X_i|\theta)). \tag{3.83}$$

Para resolver la ecuación (3.82), y asumiendo las condiciones de regularidad de la Definición 3.2, la derivada nos dice que:

$$\frac{\partial \ln L(X_1, X_2, ..., X_n|\theta)}{\partial\theta} = \frac{1}{L(X_1, X_2, ..., X_n|\theta)} \cdot \frac{\partial L(X_1, X_2, ..., X_n|\theta)}{\partial\theta} = 0, \tag{3.84}$$

es decir, debemos encontrar $\theta^* = \tau_{ML}(X_1^n)$ como función de las observaciones que resuelva:

$$\frac{1}{L(X_1, ..., X_n|\theta^*)} \cdot \frac{\partial L(X_1, ..., X_n|\theta^*)}{\partial\theta} = 0 \tag{3.85}$$

[2] Por lo general se aplica logaritmo de base e, sin embargo, la definición no cambia en resultado si se aplica otra base mayor que 1.

Naturalmente si $\ln(L(X_1, ..., X_n|\theta))$ es cóncava, la solución de la ecuación anterior nos da el óptimo global del problema. En la práctica la condición de primer orden nos define el espacio de soluciones factibles, sobre las cuales podremos encontrar la solución optima. Sin embargo, para muchos problemas la solución del estimador de máxima verosimilitud no ofrece expresiones cerradas y solo es posible aproximar por medio de métodos numéricos tipo gradiente descendente.

Veamos el siguiente ejemplo para ver cómo aplicar el estimador en una situación concreta:

Ejemplo 3.7. Sea X_1^n un vector variable i.i.d. de observaciones tal que $(\forall i \in \{1, ..., n\})X_i \sim N(\theta, \sigma^2)$. Asumiremos σ^2 conocido y el problema de estimación se reduce a estimar θ (la media de la distribución normal). En este contexto la función de verosimilitud es:

$$L(X_1, ..., X_n|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(X_i-\theta)^2}{2\sigma^2}}. \tag{3.86}$$

Necesitamos encontrar el estimador de máxima verosimilitud, es decir, el valor de θ que maximiza (3.86). Como mencionamos anteriormente, conviene tomar logaritmo debido a que esta distribución es de tipo exponencial.

$$\log(L(X_1, ..., X_n|\theta)) = n \ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \sum_{i=1}^{n} \frac{(X_i - \theta)^2}{2\sigma^2}, \tag{3.87}$$

luego, el problema de estimación de máxima verosimilitud equivale a encontrar:

$$\tau_{ML}(X_1^n) = \arg\max_{\theta\in\mathbb{R}} \ln(L(X_1, ..., X_n|\theta)) \tag{3.88}$$

$$= \arg\min_{\theta\in\mathbb{R}} \sum_{i=1}^{n} \frac{(X_i - \theta)^2}{2\sigma^2}. \tag{3.89}$$

La última expresión corresponde a minimizar (en promedio) el error cuadrático entre la variable de observación X_i y la media $\mu = \mathbb{E}(X_i)$. Luego, al tomar la expresión:

$$\arg\min_{\theta\in\mathbb{R}} \sum_{i=1}^{n} \left(\frac{X_i - \theta}{\sqrt{2}\sigma}\right)^2 = \frac{1}{2\sigma^2} \cdot \arg\min_{\theta\in\mathbb{R}} \sum_{i=1}^{n} (X_i - \theta)^2, \tag{3.90}$$

y aplicando la condición de primer orden (derivar e igualar a 0) nos dice que:

$$\begin{aligned}\frac{\partial \log L(X_1,...,X_n|\theta)}{\partial\theta} = 0 \Rightarrow \sum_{i=1}^{n} \frac{(X_i-\theta^*)}{\sigma^2} &= 0 \\ \Leftrightarrow \sum_{i=1}^{n} \frac{(X_i-\theta^*)}{\sigma^2} &= 0 \\ \Leftrightarrow \sum_{i=1}^{n} X_i - n\theta^* &= 0 \\ \Leftrightarrow \theta^* = \tau_{ML}(X_1^n) &= \frac{1}{n}\sum_{i=1}^{n} X_i. \end{aligned} \tag{3.91}$$

Por otro lado,

$$\left.\frac{\partial^2 \log L(X_1,...,X_n|\theta)}{\partial\theta^2}\right|_{\mu=\hat{\theta}} = \frac{-n}{\sigma^2} < 0, \tag{3.92}$$

por lo que el valor encontrado es un máximo, con lo que $\tau_{ML}(X_1^n) = \frac{1}{n}\sum_{i=1}^{n} X_i$ es el estimador de máxima verosimilitud. Notemos que este estimador es el mismo propuesto en el Ejemplo 3.3, conocido como media empírica, ya sabemos que dicho estimador para el caso de distribuciones normales es insesgado y consistente. Sólo por completitud vamos a analizar su sesgo y consistencia:

$$\begin{aligned}\mathbb{E}(\tau_{ML}(X_1^n)) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{\sum_{i=1}^{n} \mathbb{E}\,(X_i)}{n} \\ &= \frac{\sum_{i=1}^{n} \theta}{n} \\ &= \mu, \end{aligned} \tag{3.93}$$

con lo que es insesgado. Para ver la consistencia, notemos que aplicando la desigualdad de

Chebyshev tenemos que, para $\epsilon > 0$:

$$\begin{aligned} P_{X_1^n}(|\tau_{ML}(X_1^n) - \theta| > \epsilon) &\leq \frac{Var(\tau_{ML}(X_1^n))}{\epsilon^2} \\ &= \frac{\sum_{i=1}^{n} Var(X_i)}{n^2\epsilon^2} \\ &= \frac{\sum_{i=1}^{n} \sigma^2}{n^2\epsilon^2} \\ &= \frac{\sigma^2}{n\epsilon^2}. \end{aligned} \tag{3.94}$$

Tomando $n \to \infty$ vemos que $\tau_{ML}(X_1^n) \xrightarrow{P} \theta$, luego el estimador es consistente.

Observaciones 3.7.

- Particularmente para el caso anterior, el estimador $\tau_{ML}(X_1^n) = \frac{1}{n}\sum_{i=1}^{n} X_i$ es consistente y se pudo haber demostrado como consecuencia directa de la ley débil de los grandes números.
- Del Ejemplo 3.5 notamos que este estimador alcanza la cota de Cramér-Rao, por ende también es de mínima varianza.

Propuesto 3.2. Sea X_1^n un vector variable i.i.d. de observaciones tal que $(\forall i \in \{1, ..., n\})X_i \sim N(\mu, \theta)$. Asumiremos μ conocido y el problema de estimación se reduce a estimar θ (la varianza de la distribución normal). Verifique que:

$$\tau_{ML}(X_1^n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2. \tag{3.95}$$

En las subsecciones que vienen estudiaremos propiedades importantes que posee este estimador, lo que lo convierte en la solución por excelencia para una gran cantidad de problemas de estimación paramétrica. En lo específico veremos que el estimador de máxima verosimilitud tiene una conexión con la Información de Fisher, es consistente y asintóticamente de mínima varianza.

3.5.1. Maxima Verosimilitud y Mínima Varianza

Existe una conexión importante entre la cota de Cramér-Rao y el estimador de máxima varosimilitud. Asumamos que la cota de Cramér-Rao se alcanza (ver Sección 3.4.1), entonces sabemos que existe un estimador $\tau_n^* : \mathbb{X}^n \to \Theta$ tal que:

$$\frac{\partial \ln L(X_1^n|\theta)}{\partial \theta} = \mathcal{I}_n(\theta)(\tau_n^*(X_1^n) - \theta), \tag{3.96}$$

La solución a las condiciones de la derivada en (3.84) nos dice entonces que:

$$\mathcal{I}_n(\theta)(\tau_n^*(X_1^n) - \theta) = 0 \Rightarrow \tau_{ML}(X_1^n) = \tau_n^*(X_1^n). \tag{3.97}$$

Para analizar si este óptimo local es al mismo tiempo global, calculamos la segunda derivada de la función objetivo evaluado en $\tau_{ML}(X_1^n)$:

$$\begin{aligned}
\left.\frac{\partial^2 \ln L(X_1, ..., X_n|\theta)}{\partial \theta^2}\right|_{\theta=\tau_{ML}} &= \left.\frac{\partial \mathcal{I}_n(\theta)}{\partial \theta}\right|_{\theta=\tau_{ML}} (\tau_n^*(X_1^n) - \tau_{ML}(X_1^n)) - \mathcal{I}_n(\tau_{ML}(X_1^n)) \\
&= -\mathcal{I}_n(\tau_{ML}(X_1^n)) \\
&< 0
\end{aligned} \tag{3.98}$$

De este análisis se desprende que la solución al problema $\frac{\partial \ln L(X_1, X_2, ..., X_n|\theta)}{\partial \theta}$ es única dado la forma en (3.96) y es equivalente a la solución que alcanza la mínima varianza.

La interesante conclusión de este análsis es que si existiese un estimador insesgado de mínima varianza, éste coincidirá con el entregado por el estimador de máxima verosimilitud.

La Información de Fisher tiene un vínculo con el estimador de máxima verosimilitud en el sentido que actúa como criterio para determinar la concavidad de la función de verosimilitud.

3.5.2. Consistencia del Estimador de Máxima Verosimilitud

Uno de los grandes resultados del estimador de máxima verosimilitud es que bajo el caso de distribuciones independientes e idénticamente distribuidas el estimador es consistente, es decir, converge en probabilidad a θ_0 donde $X \sim P_X(\cdot|\theta_0)$. Enunciaremos y demostraremos tal resultado

Teorema 3.8. Sea una familia $\mathcal{F}_\Theta$, $\theta_0 \in \Theta$, un vector $X_1^n \sim P_X(\cdot|\theta_0)^n$ (i.i.d) y el estimador $\tau_{ML}(X_1^n)$ como variable aleatoria en Θ. El estimador de máxima verosimilitud converge a θ_0 en probabilidad, es decir:

$$\tau_{ML}(X_1^n) \xrightarrow{P} \theta_0 \Leftrightarrow (\forall \epsilon > 0) \lim_{n\to\infty} P_{X_1^n}\left(|\tau_{ML}(X_1^n) - \theta_0| > \epsilon\right) = 0 \tag{3.99}$$

Demostración: Trabajaremos en el espacio continuo para las variables de observación ya que el caso discreto es análogo. Antes de continuar con la demostración es necesario pedir condiciones adicionales de regularidad. Estas condiciones introducen ciertos elementos que escapan de los contenidos del curso, sin embargo, serán explicados brevemente.

Definición 3.8. (Divergencia de Kullback-Leibler) Sean $\theta_0, \theta_1 \in \Theta$ y $X \sim P_X(\cdot|\theta_0)$. Definimos la divergencia de Kullback-Leibler[3] entre las densidades $f_X(x|\theta_0)$ y $f_X(x|\theta_1)$ como:

$$D(f_X(X|\theta_0)||f_X(X|\theta_1)) = \mathbb{E}\left(\log\left(\frac{f_X(X|\theta_0)}{f_X(X|\theta_1)}\right)\right) = \int_{-\infty}^{\infty} f_X(x|\theta_0)\log\left(\frac{f_X(x|\theta_0)}{f_X(x|\theta_1)}\right)dx. \tag{3.100}$$

este operador corresponde a la divergencia entre dos distribuciones, es muy usado en Teoría de la Información y sirve para medir similitudes entre dos distribuciones de probabilidad.

Más aún, se puede demostrar que la divergencia es positiva o 0, donde si se cumple esto último entonces las densidades inducidas por θ_0 y θ_1 son iguales en todos los puntos salvos aquellos de probabilidad 0. Vamos a demostrar la positividad, para esto utilizamos la desigualdad de Jensen,[4] vemos que:

$$\begin{aligned}
\mathbb{E}_X\left(\log\left(\frac{f_X(X|\theta_1)}{f_X(X|\theta_0)}\right)\right) &\leq \log\left(\mathbb{E}_X\left(\frac{f_X(X|\theta_1)}{f_X(X|\theta_0)}\right)\right) \\
&= \log\left(\int_{-\infty}^{\infty}\left(\frac{f_X(x|\theta_1)}{f_X(x|\theta_0)}\right)f_X(x|\theta_0)dx\right) \\
&= \log\underbrace{\left(\int_{-\infty}^{\infty} f_X(x|\theta_1)dx\right)}_{1} \\
&= 0.
\end{aligned} \tag{3.101}$$

[3] En este contexto el logaritmo puede ser base e, 2 o 10, no se pierde generalidad en ocupar uno u otro, para efectos de la demostración se usará base e.

[4] Si $f : \mathbb{R} \longrightarrow \mathbb{R}$ es una función cóncava, entonces para cualquier variable aleatoria X: $\mathbb{E}_X(f(X)) \leq f(\mathbb{E}_X(X))$.

Esto es equivalente a decir que

$$\mathbb{E}_X\left(\log\left(\frac{f_X(X|\theta_0)}{f_X(X|\theta_1)}\right)\right) \geq 0, \tag{3.102}$$

donde la igualdad se obtiene en (3.102), si y sólo si,

$$f_X(x|\theta_0) = f_X(x|\theta_1) \quad \forall x \in \mathbb{R}. \tag{3.103}$$

Con esta definición y esta propiedad, ahora pediremos lo siguiente:

Definición 3.9. (Condiciones de Regularidad Adicionales) Sean $\theta_0, \theta_1 \in \Theta$ y $X \sim P_X(\cdot|\theta_0)$. Sea $\epsilon > 0$ arbitrario, entonces:

$$\inf_{\theta_1 \in \Theta : |\theta_1 - \theta_0| \geq \epsilon} D(f_X(X|\theta_0)||f_X(X|\theta_1)) > 0, \tag{3.104}$$

$$\sup_{\theta_1 \in \Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{f_X(X_i|\theta_0)}{f_X(X_i|\theta_1)}\right) - D(f_X(X|\theta_0)||f_X(X|\theta_1))\right| \xrightarrow{c.s.} 0. \tag{3.105}$$

La condición en (3.104) indica un grado de identificabilidad análogo al caso de las familias distinguibles en la Definición 3.1, en caso que el espacio Θ sea compacto ambas definiciones son equivalentes. La condición (3.105) no es más que una versión de convergencia uniforme de la ley fuerte de los grandes números[5], es decir, convergencia en cualquier valor de $\theta \in \Theta$.

Asumiendo estos dos supuestos podemos proseguir en la demostración. Para esto, tomemos un $\theta_1 \in \Theta$ arbitrario tal que sea distinto a θ_0. La idea por tanto reduce a verificar que el siguiente evento:

$$\{x_1^n \in \mathbb{X}^n : L(x_1, ..., x_n|\theta_0) - L(x_1, ..., x_n|\theta_1) > 0\} \tag{3.106}$$

ocurre con alta probabilidad para n suficientemente grande (y en el limite con probabilidad 1 cuando n tiende a infinito). Recordando que el criterio de máxima verosimilitud es el argumento θ que maximiza $L(X_1, ..., X_n|\theta)$, por lo tanto, nos interesa elegir θ_0 y gracias a la condición en (3.104) podemos garantizar que necesariamente $\lim_{n\to\infty} \tau_{ML}(X_1^n) \xrightarrow{P} \theta_0$ ya que nos entrega un óptimo global. Veamos las siguientes igualdades:

$$\begin{aligned}\frac{1}{n}\log(L(X_1, ..., X_n|\theta_1)) &= \frac{1}{n}\sum_{i=1}^{n}\log(L(X_i|\theta_1)) \\ &= \frac{1}{n}\sum_{i=1}^{n}\log(f_X(X_i|\theta_1))\end{aligned} \tag{3.107}$$

[5] Se puede pedir convergencia en probabilidad para hacerlo menos fuerte

Análogamente se tiene que:

$$\frac{1}{n}\log(L(X_1,...,X_n|\theta_0)) = \frac{1}{n}\sum_{i=1}^{n}\log(f_X(X_i|\theta_0)). \tag{3.108}$$

Entonces tenemos lo siguiente:

$$\begin{aligned} L(X_1,...,X_n|\theta_0) - L(X_1,...,X_n|\theta_1) &= \frac{1}{n}\sum_{i=1}^{n}\log(f_X(X_i|\theta_0)) - \frac{1}{n}\sum_{i=1}^{n}\log(f_X(X_i|\theta_1)) \\ &= \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{f_X(X_i|\theta_0)}{f_X(X_i|\theta_1)}\right) \\ &\xrightarrow{c.s.} \mathbb{E}_X\left(\log\left(\frac{f_X(X|\theta_0)}{f_X(X|\theta_1)}\right)\right) \\ &> 0 \end{aligned} \tag{3.109}$$

y, por ende,

$$\mathbb{P}\left(\{w\in\Omega : \lim_{n\to\infty} L(X_1(w),...,X_n(w)|\theta_0) - L(X_1(w),...,X_n(w)|\theta_1) > 0\}\right) = 1. \tag{3.110}$$

Finalmente dado que (3.110) se cumple $\forall\theta_1 \neq \theta_0$ entonces la probabilidad en (3.110) más la condición en (3.104) se puede expresar como:

$$\mathbb{P}\left(\left\{w\in\Omega : \lim_{n\to\infty} \tau_{ML}(X_1^n(w)) = \theta_0\right\}\right) = 1 \tag{3.111}$$

Por lo tanto $\tau_{ML}(X_1^n)$ converge a θ_0 casi seguramente (y en consecuencia en probabilidad[6]), luego τ_{ML} es un estimador consistente. □

3.5.3. Condición de Normalidad Asintótica del Estimador de Máxima Verosimilitud

El siguiente resultado indica que el estimador de máxima verosimilitud es asintóticamente eficiente en el sentido que su varianza converge a la cota de Cramér-Rao cuando el número de observaciones se va a infinito.

[6] La convergencia casi segura de una secuencia aleatoria es más fuerte que la convergencia en probabilidad. Detalles en [7].

Teorema 3.9. Sea $\mathcal{F}_\Theta$ una familia de distribuciones y consideremos $\tau_{ML}(X_1^n)$ el estimador de máxima verosimilitud. Asumiendo las condiciones de regularidad de la Definición 3.2 y θ_0 es el valor tal que $X_1^n \sim P(\cdot|\theta_0)^n$ (i.i.d), se tiene que:

$$\sqrt{n}(\tau_{ML}(X_1^n) - \theta_0) \xrightarrow{d} Y, \tag{3.112}$$

con $Y \sim N\left(0, \frac{1}{\mathcal{I}_1(\theta_0)}\right)$ donde

$$\mathcal{I}_1(\theta_0) = \mathbb{E}_{X_1}\left(\left(\frac{\partial \log L(X_1|\theta_0)}{\partial \theta_0}\right)^2\right). \tag{3.113}$$

Es decir el estimador es asintóticamente eficiente y, además, se desprende que es consistente. Notar que no necesariamente es asintóticamente insesgado, por lo que la definición de eficiencia asintótica es distinta a la eficiencia finita.

Demostración: Antes de comenzar la demostración, utilizaremos un resultado de extensión de convergencia conocido como el Teorema de Slutsky.

Teorema 3.10. (Slustky) Sean $(X_n)_{n\in\mathbb{N}}$ e $(Y_n)_{n\in\mathbb{N}}$ dos secuencias de variables aleatorias tales que $X_n \xrightarrow{d} X$ e $Y_n \xrightarrow{d} c$, con $c \in \mathbb{R}$, entonces

$$X_n Y_n \xrightarrow{d} Xc \tag{3.114}$$

Dado que $\tau_{ML}(X_1^n) \to \theta_0$ casi seguramente (o con probabilidad 1), la idea es utilizar la hipotesis que la función $\ln L(X_1, ..., X_n|\theta)$ es dos veces diferenciable con respecto a θ. Con esto tomaremos la siguiente función $\frac{1}{\sqrt{n}}\frac{\partial \ln L(X_1,...,X_n|\theta)}{\partial \theta}$ y realizaremos un desarrollo en serie de Taylor entorno a θ_0 de orden 0 evaluado en $\theta = \tau_{ML}(X_1^n)$, es decir:

$$\frac{1}{\sqrt{n}}\frac{\partial \ln L(X_1, ..., X_n|\theta)}{\partial \theta}\bigg|_{\theta=\tau_{ML}} \tag{3.115}$$

$$= \frac{1}{\sqrt{n}}\frac{\partial \ln L(X_1, ..., X_n|\theta)}{\partial \theta}\bigg|_{\theta=\theta_0} + \frac{1}{\sqrt{n}}\frac{\partial^2 \ln L(X_1, ..., X_n|\theta)}{\partial \theta^2}\bigg|_{\theta=\tilde{\theta}}(\tau_{ML} - \theta_0), \tag{3.116}$$

con $\tilde{\theta} \in (\tau_{ML}, \theta_0)$. Lo primero que notamos es que por la consistencia del estimador de máxima verosimilitud $\tilde{\theta} \to \theta_0$ casi seguramente.

Por otro lado, dado que el estimador de máxima verosimilitud cumple la condición de primer orden (por definición maximiza la función $\ln L(X_1, ..., X_n|\theta)$), entonces:

$$\frac{1}{\sqrt{n}} \frac{\partial \ln L(X_1, ..., X_n|\theta)}{\partial \theta}\bigg|_{\theta=\tau_{ML}} = 0 \tag{3.117}$$

de (3.116) tenemos entonces que cuando n tiende a infinito:

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}} \frac{\partial \ln L(X_1, X_2, ..., X_n|\theta)}{\partial \theta}\bigg|_{\theta=\theta_0} = \lim_{n\to\infty} \frac{-\sqrt{n}}{n} \frac{\partial^2 \ln L(X_1, X_2, ..., X_n|\theta)}{\partial \theta^2}\bigg|_{\theta=\tilde{\theta}} (\tau_{ML} - \theta_0). \tag{3.118}$$

Vamos a analizar las expresiones de ambos lados de la identidad en (3.118).

Respecto al termino del lado derecho de (3.118), debido a la ley fuerte de los grandes números (notando que $\log(L(X_1, ..., X_n|\theta)) = \sum_{i=1}^{n} \log(L(X_i|\theta)))$ se tiene que:

$$\begin{aligned} \frac{-1}{n} \frac{\partial^2 \ln L(X_1, X_2, ..., X_n|\theta)}{\partial \theta^2}\bigg|_{\theta=\tilde{\theta}} &= -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \ln L(X_i|\theta)}{\partial \theta^2}\bigg|_{\theta=\tilde{\theta}} \\ &\xrightarrow{c.s.} -\mathbb{E}_{X_1}\left(\frac{\partial^2 \ln(f_{X_1}(X_1|\theta))}{\partial \theta^2}\bigg|_{\theta=\theta_0} \right) \\ &= \mathcal{I}_1(\theta_0) \end{aligned} \tag{3.119}$$

Respecto al término del lado izquierdo de (3.118), podemos notar que $\frac{\partial \ln(f_{X_1}(X_1|\theta))}{\partial \theta}\Big|_{\theta=\theta_0}$ es una variable aleatoria de media 0 y varianza $\mathcal{I}_1(\theta_0)$. Luego deducimos que de la aplicación del Teorema Central del Límite que:

$$\begin{aligned} \frac{1}{\sqrt{n}} \cdot \frac{1}{\sqrt{\mathcal{I}_1(\theta_0)}} \frac{\partial \ln L(X_1, X_2, ..., X_n|\theta)}{\partial \theta}\bigg|_{\theta=\theta_0} &= \frac{1}{\sqrt{n}} \frac{1}{\sqrt{\mathcal{I}_1(\theta_0)}} \sum_{i=1}^{n} \frac{\partial \ln L(X_i|\theta)}{\partial \theta}\bigg|_{\theta=\theta_0} \\ &= \frac{1}{\sqrt{n}} \cdot \frac{n}{\sqrt{\mathcal{I}_1(\theta_0)}} \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \ln L(X_i|\theta)}{\partial \theta}\bigg|_{\theta=\theta_0} \\ &= \frac{\frac{1}{n} \sum_{i=1}^{n} \frac{\partial \ln L(X_i|\theta)}{\partial \theta}\Big|_{\theta=\theta_0}}{\frac{\sqrt{\mathcal{I}_1(\theta_0)}}{\sqrt{n}}} \\ &\xrightarrow{d} N(0, 1) \end{aligned} \tag{3.120}$$

que equivale a decir que:

$$\frac{1}{\sqrt{n}}\frac{\partial \ln L(X_1, X_2, ..., X_n|\theta)}{\partial \theta}\bigg|_{\theta=\theta_0} \xrightarrow{d} N(0, \mathcal{I}_1(\theta_0)). \tag{3.121}$$

Finalmente regresando a (3.118), y tomando $n \to \infty$ tenemos que:

$$\underbrace{\frac{1}{\sqrt{n}}\frac{\partial \ln L(X_1, X_2, ..., X_n|\theta)}{\partial \theta}\bigg|_{\theta=\theta_0}}_{\xrightarrow{d} Y \sim N(0, \mathcal{I}_1(\theta_0))} = \sqrt{n}\underbrace{\frac{-1}{n}\frac{\partial^2 \ln L(X_1, X_2, ..., X_n|\theta)}{\partial \theta^2}\bigg|_{\theta=\theta_0}}_{\xrightarrow{c.s.} \mathcal{I}_1(\theta_0)}(\tau_{ML} - \theta_0) \tag{3.122}$$

Por lo tanto $\sqrt{n}(\tau_{ML}(X_1, .., X_n) - \theta_0)$ converge en distribución a una variable aleatoria $Z \sim \frac{1}{\mathcal{I}_1(\theta_0)}N(0, \mathcal{I}_1(\theta_0)) = N\left(0, \frac{1}{\mathcal{I}_1(\theta_0)}\right)$. Notar que el pasar dividiendo $\frac{-1}{n}\frac{\partial^2 \ln L(X_1, X_2, ..., X_n|\theta)}{\partial \theta^2}\Big|_{\theta=\theta_0}$ es posible de hacer gracias al teorema de Slutsky. □

Observaciones 3.8.

- El estimador de máxima verosimilitud puede no ser insesgado, pero al ser consistente y además le pedimos integrabilidad uniforme entonces se puede garantizar que sea asintóticamente insesgado.
- Los resultados de consistencia y normalidad asintótica son válidos cuando los modelos son independientes e idénticamente distribuidos , por lo tanto, si este supuesto no se cumple no se puede garantizar tales propiedades.
- No siempre el estimador de máxima verosimilitud ofrecerá una solución cerrada, por lo que muchas veces se requerirá utilizar algún optimizador y calcular el estimador de manera numérica.

El estimador de máxima verosimilitud es consistente en probabilidad (en consecuencia asintoticamente insesgado[7]) y adicionalmente su varianza converge (con n) a la mínima varianza dada por la cota de Cramér-Rao.

[7] En general la convergencia casi segura o en probabilidad no implican convergencia en media, salvo que el estimador sea uniformemente integrable

Por lo tanto para el caso de observaciones independientes e idénticamente distribuidas, no existe un mejor estimador con mejores propiedades de optimalidad que el de máxima verosimilitud, lo que lo convierte en el estimador por excelencia frente a observaciones i.i.d., cumpliendo las condiciones de regularidad de la Definición 3.2.

El siguiente ejemplo es un caso de estudio donde aplicaremos el estimador de máxima verosimilitud en un contexto multiparamétrico, es decir, Θ es un conjunto de más de una dimensión. En particular el objetivo será proponer un estimador para el vector de media y para la matriz de covarianza.

3.5.4. Caso de Estudio: Distribución Normal Multivariada

Consideremos un vector aleatorio $(X_1, ..., X_d)$ con valores en $\mathbb{R}^d$ tal que:

$$X_1^d \sim N(\overline{m}, K)$$

con $\overline{m} \in \mathbb{R}^d$ es el vector de media y $K = \mathbb{E}\{(X - \overline{m})(X - \overline{m})^t\} \in \mathbb{R}^{d\times d}$ la matriz de covarianza. El problema consiste en estimar $\overline{m}, K$ como función de n observaciones vectoriales i.i.d. $((X_1^d)_1, ..., (X_1^d)_n)$. Notar que este caso cada observación corresponde a un vector de dimensión d. Para reducir la notación diremos que $Y_i = (X_1^d)_i$, es decir, Y_i representa el vector i-ésimo. La función de verosimilitud conjunta en este caso es:

$$\begin{aligned} L(Y_1, ..., Y_n|\overline{m}, K) &= \prod_{i=1}^{n} \frac{1}{(\sqrt{2\pi})^m |K|^{1/2}} e^{\frac{-1}{2}(Y_i-\overline{m})^t K^{-1}(Y_i-\overline{m})} \\ &= (2\pi)^{-nm/2}|K|^{-n/2} e^{\frac{-1}{2}\sum_{i=1}^{n}(Y_i-\overline{m})^t K^{-1}(Y_i-\overline{m})} \end{aligned} \tag{3.123}$$

Luego

$$\ln(L(Y_1, ..., Y_n|\overline{m}, K)) = -\frac{nm}{2}\log(2\pi) - \frac{n}{2}\log|K| - \frac{1}{2}\sum_{i=1}^{n}(Y_i - \overline{m})^t K^{-1}(Y_i - \overline{m}) \tag{3.124}$$

Imponiendo las condiciones de primer orden, vamos a buscar el máximo de (3.124). Debido a la gran cantidad de operaciones matriciales que se utilizarán, vamos a introducir las

siguientes definiciones:

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad \text{(Media Empírica)} \tag{3.125}$$

$$\overline{S} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})(Y_i - \overline{Y})^t \quad \text{(Covarianza Empírica Muestral).} \tag{3.126}$$

El término cuadrático de (3.124) se puede re-escribir como:

$$\begin{aligned}(Y_i - \overline{m})^t K^{-1}(Y_i - \overline{m}) &= (Y_i - \overline{Y} + \overline{Y} - \overline{m})^t K^{-1}(Y_i - \overline{Y} + \overline{Y} - \overline{m}) \\ &= (Y_i - \overline{Y})^t K^{-1}(Y_i - \overline{Y}) + (\overline{Y} - \overline{m})^t K^{-1}(\overline{x} - \overline{m}) + 2(\overline{Y} - \overline{m})^t K^{-1}(Y_i - \overline{Y})\end{aligned} \tag{3.127}$$

donde al tomar sumatoria tenemos que

$$\begin{aligned}&\sum_{i=1}^{n}(Y_i - \overline{m})^t K^{-1}(Y_i - \overline{m}) \\ &= \sum_{i=1}^{n}(Y_i - \overline{Y})^t K^{-1}(Y_i - \overline{Y}) + \sum_{i=1}^{n}(\overline{Y} - \overline{m})^t K^{-1}(\overline{Y} - \overline{m}) + 2\sum_{i=1}^{n}(\overline{Y} - \overline{m})^t K^{-1}(Y_i - \overline{Y}) \quad \to 0 \\ &= \sum_{i=1}^{n}(Y_i - \overline{Y})^t K^{-1}(Y_i - \overline{Y}) + n \cdot (\overline{Y} - \overline{m})^t K^{-1}(\overline{Y} - \overline{m})\end{aligned} \tag{3.128}$$

El término $\sum_{i=1}^{n}(Y_i - \overline{Y})^t K^{-1}(Y_i - \overline{Y})$ se conoce como dispersión y $(\overline{Y} - \overline{m})^t K^{-1}(\overline{Y} - \overline{m})$ es el sesgo. Notemos que:

$$(Y_i - \overline{Y})^t K^{-1}(Y_i - \overline{Y}) = tr((Y_i - \overline{Y})(Y_i - \overline{Y})^t K^{-1}) \tag{3.129}$$

$$= tr(K^{-1}(Y_i - \overline{Y})(Y_i - \overline{Y})^t), \tag{3.130}$$

donde tr corresponde a la traza de una matriz, cumple que $tr(A) = tr(A^t)$ y es ciclica, i.e., $tr(ABC) = tr(CAB) = tr(BCA)$. Entonces volviendo a (3.128)

$$\begin{aligned}\sum_{i=1}^{n}(Y_i - \overline{m})^t K^{-1}(Y_i - \overline{m}) &= n \cdot (\overline{Y} - \overline{m})^t K^{-1}(\overline{Y} - \overline{m}) + tr\left(K^{-1}\sum_{i=1}^{n}(Y_i - \overline{Y})(Y_i - \overline{Y})^t\right) \\ &= n \cdot (\overline{Y} - \overline{m})^t K^{-1}(\overline{Y} - \overline{m}) + tr\left(K^{-1} n\overline{S}\right)\end{aligned} \tag{3.131}$$

Incorporando estos resultados tenemos que:

$$\log(L(Y_1, ..., Y_n|\overline{m}, K)) = -\frac{nm}{2}\log(2\pi) - \frac{n}{2}\log|K| - \frac{n}{2}tr\left(K^{-1}\overline{S}\right) - \frac{n}{2}(\overline{Y} - \overline{m})^t K^{-1}(\overline{Y} - \overline{m}). \tag{3.132}$$

Ahora estamos en condiciones de derivar, consideramos la verosimilitud en (3.132) función de $V = K^{-1}$ y $\overline{m}$ y la denotaremos como $\log(L(Y_1, ..., Y_n|\overline{m}, V)$. El objetivo es entonces estimar $\overline{m}$ y V.

Durante el desarrollo vamos a las siguientes identidades válidas para matrices V de dimensión $d \times d$, matriz simétrica A de dimensión $d \times d$ y vectores $\overline{m}$ de dimensión d:

1- $\frac{\partial \overline{m} A \overline{m}^t}{\partial \overline{m}} = 2A\overline{m}$
2- $\frac{\partial \log(|V|)}{\partial V} = (V^{-1})^t$
3- $\frac{\partial tr(V \cdot \overline{S})}{\partial V} = \overline{S}^t$.

Con este resultado, imponemos las condiciones de primer orden y pedimos que:

$$\frac{\partial \log(L(Y_1, ..., Y_n|\overline{m}, V)}{\partial \overline{m}} = 0 \quad \wedge \quad \frac{\partial \log(L(Y_1, ..., Y_n|\overline{m}, V)}{\partial V} = 0. \tag{3.133}$$

Esto implica que:

$$\begin{aligned} \frac{\partial \log(L(Y_1, ..., Y_n|\overline{m}, V)}{\partial \overline{m}} &= \frac{\partial}{\partial \overline{m}} \left(-\frac{n}{2} (\overline{Y} - \overline{m})^t K^{-1} (\overline{Y} - \overline{m}) \right) \\ &= -nK^{-1}(\overline{Y} - \overline{m}) = 0 \\ &\Rightarrow \widehat{\overline{m}}_{ML}(Y_1^n) = \overline{Y}. \end{aligned} \tag{3.134}$$

tenemos que:

$$\begin{aligned} \frac{\partial \log(L(Y_1, ..., Y_n|\overline{m}, V)}{\partial V} &= \frac{\partial}{\partial V} \left(\frac{n}{2} \log |V| - \frac{n}{2} tr\left(V\overline{S}\right) - \frac{n}{2} (\overline{Y} - \overline{m})^t V (\overline{Y} - \overline{m}) \right) \\ &= \frac{n}{2} \left((V^{-1})^t - \overline{S}^t - \frac{\partial}{\partial V} tr\left(V(\overline{Y} - \overline{m})(\overline{Y} - \overline{m})^t\right) \right) \\ &= \frac{n}{2} \left((V^{-1})^t - \overline{S}^t - ((\overline{Y} - \overline{m})(\overline{Y} - \overline{m})^t)^{-1} \right) = 0 \end{aligned} \tag{3.135}$$

Finalmente, tomando traspuesto, podemos notar que dado que $\widehat{\overline{m}}_{ML}(Y_1^n) = \overline{Y}$:

$$0 = V^{-1} - \overline{S} - (\overline{Y} - \overline{m})(\overline{Y} - \overline{m})^t = V^{-1} - \overline{S} - \cancelto{0}{(\overline{Y} - \overline{m})(\overline{Y} - \overline{m})^t}. \tag{3.136}$$

Por lo tanto $\widehat{K}_{ML}(Y_1^n) = \hat{V}_{ML}^{-1}(Y_1^n) = \overline{S}$, por lo que el estimador de máxima verosimilitud es:

$$\tau_{ML}(Y_1^n) = \begin{pmatrix} \widehat{\overline{m}}_{ML}(Y_1^n) \\ \widehat{K}_{ML}(Y_1^n) \end{pmatrix} = \begin{pmatrix} \overline{Y} \\ \overline{S} \end{pmatrix} \tag{3.137}$$

3.6. Estimador de Mínimo Error Cuadrático Medio

El estimador de mínimos cuadrados o Least Squares es un estimador muy usado debido a su simpleza de implementación. Uno de los supuestos centraes de este método es que no se conocen a las distribuciones de probabilidad del modelo asociado. Luego es un método muy prático. La idea central del uso de este estimador proviene del hecho de buscar minimizar la distancia entre el valor observado y el estimado.

Consideremos entonces el siguiente problema: dado un vector aleatorio Y_1^n buscamos estimar un vector $\bar{\theta} \in \mathbb{R}^m$ de tal forma de minimizar su error cuadrático, es decir, debemos minimizar:

$$||\bar{\theta} - \hat{\theta}(Y_1^n)||^2 = \sum_{i=1}^{m} (\theta_i - \hat{\theta}_i(Y_1^n))^2. \tag{3.138}$$

donde θ_i es la componente i-ésima del vector $\bar{\theta}$. Este caso es bien general debido a que el vector a estimar es multiparamétrico, pero, más aún, observamos que (3.138) es intratable dado que no conocemos $\bar{\theta}$ y, por lo tanto, encontrar un estimador $\hat{\theta}(Y_1^n)$ sin una hipótesis adicional es inviable. Adicional a lo anterior, en el criterio de mínimos cuadrados -por lo general- no poseemos la distribución de probabilidad parametrizada de Y_1^n dado $\bar{\theta}$, es decir, solamente tenemos acceso a datos, por lo que tampoco podemos dar en principios garantías de sesgo o consistencia.

Esta última observación motiva a replantearse y modificar el problema. Dado que lo que sí tenemos es un conjunto de observaciones, éstas representan una versión ruidosa de $\bar{\theta}$, por lo que el problema se puede replantear buscando minimizar la distancia entre el vector variable de observaciones Y_1^n con respecto al vector de parámetros $\bar{\theta}$ proyectado en el espacio de observaciones $\mathbb{R}^n$, es decir,

$$||Y_1^n - f(\mathbf{X}, \bar{\theta})||^2 = \sum_{i=1}^{n} (Y_i - f_i(\mathbf{X}, \bar{\theta}))^2. \tag{3.139}$$

donde $f : M_{n\times m} \times \mathbb{R}^m \to \mathbb{R}^n$ es una función parametrizada por $\bar{\theta}$ tal que entrega un ajuste entre $\mathbf{X}$ e Y_1^n. Luego f no es más que una función que proyecta la información de $\bar{\theta}$ al espacio que está Y_1^n. La matriz $\mathbf{X}$ es una matriz de soporte, suelen ser conjuntos de datos adicionales para facilitar encontrar la relación entre $\bar{\theta}$ e Y_1^n. La expresión $R_i \triangleq \theta_i - f_i(\mathbf{X}, \bar{\theta})$ se conoce como residuo y lo que se busca es que sea despreciable, es decir, $\mathbb{E}(R_i) = 0$. Esto significa que, entonces, los modelos de observación se pueden reescribir como:

$$Y = f_i(X_1^m, \bar{\theta}) + R \tag{3.140}$$

En consecuencia el problema de mínimos cuadrados se redujo a encontrar el vector de parámetros $\bar{\theta}$ tales que la función f mejor describa la relación entre $\mathbf{X}$ e Y_1^n, donde se acepta que exista un ruido de tipo aditivo despreciable. Esto último se conoce como el problema de **regresión**. Se desprende que la técnica de regresión es un mecanismo de ajustar datos y es aplicable en los modelos donde existe una dependencia aditiva del ruido.

La idea de la regresión es un razonamiento levemente inverso al que hemos usado en las unidades anteriores. En vez de encontrar la función que estime $\bar{\theta}$ usando el modelo dado por las distribuciones de las observaciones, se hará lo contrario, se llevarán los parámetros al espacio de observaciones y sobre eso estimar. Esta lógica se sustenta en el hecho que, al no conocer las distribuciones condicionales de las observaciones dado los parámetros, debemos utilizar las observaciones como criterio de optimalidad.

Para este obtener este estimador, llamado de mínimos cuadrados, se debe resolver el criterio de primer orden, es decir, encontrar la solución al siguiente sistema:

$$(\forall i \in \{1, ..., m\}) \quad \sum_{i=1}^{n} (Y_i - f_i(\mathbf{X}, \bar{\theta})) \frac{\partial f_i(\mathbf{X}, \bar{\theta})}{\theta_i}\bigg|_{\bar{\theta}=\theta_{LS}} = 0 \tag{3.141}$$

lo que entrega un sistema de ecuaciones que se debe resolver, por lo general aplicando métodos numéricos. Dado que f aún es implícito, nos limitaremos al caso del estimador lineal de mínimos cuadrados, es decir, $f(\mathbf{X}, \bar{\theta}) = \mathbf{X}\bar{\theta}$. Hay muchos problemas inversos en ingeniería que reducen al modelo lineal de observación, con lo que el modelo de observación en este caso se reduce a:

$$Y_1^n = \mathbf{X}\bar{\theta} + V_1^n, \tag{3.142}$$

donde:

1- $Y_1^n \in \mathbb{R}^n$ es el vector variable de observación (variable independiente)[8],
2- $\mathbf{X} \in M_{n\times m}$ es la matriz de proyección o el operador lineal que mapea el parámetro al espacio de observaciones (variable dependiente),
3- $\bar{\theta} \in \mathbb{R}^m$ es el parámetro a inferir, en este caso es un vector de parámetros.
4- $V_1^n \in \mathbb{R}^n$ representa un ruido aditivo con valores en $\mathbb{R}^n$.

[8] En la práctica, no se debe olvidar que se poseen datos fijos, luego esos datos no son más que las realizaciones de este vector (y_1^n)

a continuación vamos a despreciar el efecto del ruido, es decir, buscaremos minimizar:

$$\begin{aligned}\hat{\theta}_{LS}(\mathbf{X}, Y_1, ..., Y_n) &= \arg\min_{\bar{\theta}\in\mathbb{R}^m} \sum_{i=1}^{n} (Y_i - (\mathbf{X}\bar{\theta})_i)^2 \\ &= \arg\min_{\bar{\theta}\in\mathbb{R}^m} ||Y_1^n - \mathbf{X}\bar{\theta}||^2 \\ &= \arg\min_{\bar{\theta}\in\mathbb{R}^m} (Y_1^n - \mathbf{X}\bar{\theta})^t (Y_1^n - \mathbf{X}\bar{\theta}) \end{aligned} \tag{3.143}$$

Notar que $(\mathbf{X}\bar{\theta})_i$ es la fila i-ésima del problema. En este escenario hay varios casos:

1- $n \geq m$: Caso sobre-estimado. Más mediciones que grados de libertad.
2- $n = m$: Caso crítico.
3- $n < m$: Caso sub-estimado. Menos mediciones que grados de libertad.

En lo que sigue veremos una solución genérica para este problema. Vamos a considerar una matriz $W \in M_{n\times n}$ (llamada matriz de pesos) definida positiva y simétrica lo que transformará el problema de mínimos cuadrados a una versión ponderada (Weighted Least Squares).

El problema de estimación cuadrática ponderada se define como

$$\begin{aligned}\hat{\theta}_{WLS}(\mathbf{X}, Y_1, ..., Y_n) &= \arg\min_{\bar{\theta}\in\mathbb{R}^m} (Y_1^n - \mathbf{X}\bar{\theta})^t W (Y_1^n - \mathbf{X}\bar{\theta}) \\ &= \arg\min_{\bar{\theta}\in\mathbb{R}^m} \sum_{i=1}^{n} w_i (Y_i - (\mathbf{X}\bar{\theta})_i)^2, \end{aligned} \tag{3.144}$$

donde la última expresión se puede desprender solamente si W es diagonal cuya componente i-ésima es w_i. En adelante analizaremos la función objetivo:

$$\begin{aligned} J(\bar{\theta}) &= (Y_1^n - \mathbf{X}\bar{\theta})^t W (Y_1^n - \mathbf{X}\bar{\theta}) \\ &= (Y_1^n)^t W Y_1^n + \bar{\theta}^t \mathbf{X}^t W \mathbf{X}\bar{\theta} - (Y_1^n)^t W \mathbf{X}\bar{\theta} - \bar{\theta}^t \mathbf{X}^t W (Y_1^n) \\ &= (Y_1^n)^t W Y_1^n + \bar{\theta}^t \mathbf{X}^t W \mathbf{X}\bar{\theta} - 2(Y_1^n)^t W \mathbf{X}\bar{\theta}, \end{aligned} \tag{3.145}$$

donde se usó el hecho que $(Y_1^n)^t W \mathbf{X}\bar{\theta}$ es un escalar, luego es igual a su transpuesto. Aplicamos la condición de primer orden y las identidades válidas para toda matriz $A \in M_{m\times m}$ y $b \in \mathbb{R}^m$

1- $\nabla_{\bar{\theta}}(b^t\bar{\theta}) = \frac{\partial}{\partial\bar{\theta}}(b^t\bar{\theta}) = b$.

2- $\nabla_{\bar{\theta}}(\bar{\theta}^t A\bar{\theta}) = \frac{\partial}{\partial\bar{\theta}}(\bar{\theta}^t A\bar{\theta}) = 2A\bar{\theta}$.

Tenemos que:

$$\begin{aligned}\frac{\partial J(\bar{\theta})}{\partial\bar{\theta}} = 0 \Rightarrow \frac{\partial\left(\bar{\theta}^t\mathbf{X}^t W\mathbf{X}\bar{\theta}\right)}{\partial\bar{\theta}} - \frac{\partial\left(2(Y_1^n)^t W\mathbf{X}\bar{\theta}\right)}{\partial\bar{\theta}} = 0 \\ = 2\mathbf{X}^t W\mathbf{X}\bar{\theta} - 2((Y_1^n)^t W\mathbf{X})^t = 0 \\ \Rightarrow \hat{\theta}_{WLS} = (\mathbf{X}^t W\mathbf{X})^{-1}(\mathbf{X}^t W)Y_1^n \end{aligned} \tag{3.146}$$

En el caso particular que no se utilice una matriz de pesos se reduce al estimador lineal de mínimos cuadrados, dado por:

$$\hat{\theta}_{LS}(\mathbf{X}, Y_1, ..., Y_n) = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t Y_1^n \tag{3.147}$$

Observaciones 3.9. La condición de invertibilidad se da cuando W es definida no negativa, es invertible y el rango de $\mathbf{X}$ es completo, es decir, cuando $n \geq m$ y las columnas de X son linealmente independientes.

La ecuación (3.147) es una ecuación bien general que se puede aplicar siempre y cuando se asuman las siguientes hipótesis:

- Un modelo de observación tal que el ruido sea aditivo
- Los parámetros deben ser lineales, es decir, seguir la estructura funcional $f(\mathbf{X}, \bar{\theta}) = \mathbf{X}\bar{\theta}$. Veremos más adelante que la linealidad es sobre los parámetros, no necesariamente sobre las observaciones.

Es común en regresión ocupar pares de conjuntos de observaciones $(x_1, y_1), ..., (x_n, y_n)$ que provienen de distribuciones independientes e idénticamente distribuidos, la matriz de soporte $\mathbf{X}$ tendrá un conjunto extra de observaciones que motivarán encontrar la relación entre X e Y. En adelante veremos algunos ejemplos clásicos donde las hipótesis se cumplen y se puede ocupar estos resultados.

Ejemplo 3.8. (Modelo Agnóstico) Supongamos que tenemos un instrumento que mide una variable escalar por medio de la siguiente ecuación

$$Y_i = \theta + \underbrace{V_i}_{\text{ruido}} \qquad \forall i \in \{1, ..., n\} \tag{3.148}$$

Buscamos estimar θ por medio de variables de observación ruidosas Y_1^n. Se determinará el estimador lineal de mínimos cuadrados óptimo, para esto notemos que:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \cdot \theta + \begin{pmatrix} V_1 \\ \vdots \\ V_n \end{pmatrix} \tag{3.149}$$

$n \geq 1$, es un caso sobre-determinado. Aplicamos el estimador lineal de mínimos cuadrados y tenemos que:

$$\begin{aligned} \hat{\theta}_{LS}(Y_1, ..., Y_n) &= \left(\begin{pmatrix} 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \\ &= \frac{1}{n} \sum_{i=1}^{n} Y_i \quad | \text{ Media empírica} \end{aligned} \tag{3.150}$$

Nuevamente vemos que la media empírica aparece como resultado de optimalidad, esta vez en un contexto de regresión. La media empírica puede interpretarse como la solución más básica que corresponde a cuando se desea estimar θ sin tener apoyo adicional, lo que se conoce como modelo agnóstico. Básicamente el promedio empírico lo que busca es limpiar el ruido.

El siguiente ejemplo ilustra el caso más simple que corresponde al ajuste lineal, es decir, encontrar la función lineal que mejor describa la relación entre X_1^n e Y_1^n.

Ejemplo 3.9. (Regresión Lineal) Supongamos que tenemos un conjunto de puntos $\{(x_i, y_i)\}_{i=1}^n$. Consideremos el siguiente modelo lineal

$$Y_i = \alpha + \beta X_i + \underbrace{V_i}_{\text{ruido}} \qquad \forall i \in \{1, ..., n\} \tag{3.151}$$

Nos gustaría encontrar los parámetros α y β que mejor se ajusten al modelo lineal en el sentido de mínimo error cuadrático. Para esto entonces se determinará el estimador de mínimos cuadrados óptimo. Notemos que:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} \cdot \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} V_1 \\ \vdots \\ V_n \end{pmatrix} \tag{3.152}$$

$n \geq 2$, es un caso sobre-determinado. Aplicamos el estimador lineal de mínimos cuadrados y tenemos que:

$$\begin{aligned}
\hat{\theta}_{LS}((X_1,Y_1),...,(X_n,Y_n)) &= \left(\begin{pmatrix} 1 & \cdots & 1 \\ X_1 & \cdots & X_n \end{pmatrix} \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 & \cdots & 1 \\ X_1 & \cdots & X_n \end{pmatrix} \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \\
&= \begin{pmatrix} n & \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} X_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_i Y_i \end{pmatrix} \\
&= \frac{1}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} \begin{pmatrix} \sum_{i=1}^{n} X_i^2 & -\sum_{i=1}^{n} X_i \\ -\sum_{i=1}^{n} X_i & n \end{pmatrix} \begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_i Y_i \end{pmatrix} \\
&= \frac{1}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} \begin{pmatrix} \sum_{i=1}^{n} X_i^2 \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} X_i Y_i \\ n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i \end{pmatrix} \\
&= \begin{pmatrix} \hat{\alpha}((X_1,Y_1),...,(X_n,Y_n)) \\ \hat{\beta}((X_1,Y_1),...,(X_n,Y_n)) \end{pmatrix} \qquad (3.153)
\end{aligned}$$

Si nos concentramos en el parámetro $\hat{\beta}((X_1,Y_1),...,(X_n,Y_n))$, tenemos que:

$$\begin{aligned}
\hat{\beta}((X_1,Y_1),...,(X_n,Y_n)) &= \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \frac{1}{n}\sum_{i=1}^{n} X_i \frac{1}{n}\sum_{i=1}^{n} Y_i}{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2}. \qquad (3.154)
\end{aligned}$$

El numerador corresponde a la covarianza empírica entre X e Y. El denominador corresponde a la varianza empírica de X. Recordemos que la covarianza entre dos variables se interpreta como el grado de dependencia lineal entre dos variables, lo que tiene directa relación con que forme parte del coeficiente de $\hat{\beta}((X_1,Y_1),...,(X_n,Y_n))$, ya que

$\hat{\beta}((X_1, Y_1), ..., (X_n, Y_n))$ es la pendiente de la recta que asocia X con Y.

Observaciones 3.10.

- La regla anterior se puede extender a polinomios de mayor grado, por lo que el estimador lineal de mínimos cuadrados puede ser generalizado a escenarios más complejos siempre y cuando exista una relación lineal de parámetros (y no necesariamente lineal en las observaciones). Consideremos el siguiente modelo lineal:

$$Y_i = \alpha + \beta X_i + \gamma X_i^2 + \underbrace{V_i}_{\text{ruido}} \qquad \forall i \in \{1, ..., n\} \tag{3.155}$$

  Inmediatamente podemos realizar el ajuste al modelo lineal en el sentido de mínimo error cuadrático. Notemos que:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_1 & X_1^2 \\ \vdots & \vdots & \vdots \\ 1 & X_n & X_n^2 \end{pmatrix} \cdot \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} + \begin{pmatrix} V_1 \\ \vdots \\ V_n \end{pmatrix} \tag{3.156}$$

  Lo cual nos permite utilizar la expresión en (3.147).
- Similarmente puede ser extendido a funciones no necesariamente lineales pero que de todas formas exista un ajuste lineal de los parámetros mediante alguna transformación. Consideremos el siguiente modelo:

$$Y_i = Ke^{\beta X_i} + \underbrace{V_i}_{\text{ruido}} \qquad \forall i \in \{1, ..., n\} \tag{3.157}$$

  Tomando logaritmo y despreciando el ruido notemos que:

$$\ln(Y_i) = \ln(K) + \beta X_i \qquad \forall i \in \{1, ..., n\} \tag{3.158}$$

  Lo que nos entrega una relación lineal de la forma:

$$\begin{pmatrix} \ln(Y_1) \\ \vdots \\ \ln(Y_n) \end{pmatrix} = \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} \cdot \begin{pmatrix} \ln(K) \\ \beta \end{pmatrix}, \tag{3.159}$$

  y nuevamente nos permite utilizar la expresión en (3.147), lo que nos da un ajuste tipo exponencial.

- Dado que ya tenemos $\hat{\theta}_{LS}(\mathbf{X}, Y_1^n)$ podemos entonces calcular la mínima distancia a optimizar dada en (3.139). Al hacer esto nos encontramos con la siguiente expresión:

$$||Y_1^n - f(\mathbf{X}, \hat{\theta}_{LS}(\mathbf{X}, Y_1^n))||^2 = \sum_{i=1}^{n} (Y_i - f_i(\mathbf{X}, \hat{\theta}_{LS}(\mathbf{X}, Y_1^n)))^2 \tag{3.160}$$

El valor $(Y_i - f_i(\mathbf{X}, \hat{\theta}_{LS}(\mathbf{X}, Y_1^n)))$ corresponde al residuo ya que es la diferencia entre el valor observado y el valor estimado. Esto significa entonces que $\sigma_R^2 \triangleq \sum_{i=1}^{n} (Y_i - f_i(\mathbf{X}, \hat{\theta}_{LS}(\mathbf{X}, Y_1^n)))^2$ se interpreta como la varianza residual.

Por otro lado, conocemos el estimador más simple que corresponde a la media empírica $\frac{1}{n}\sum_{i=1}^{n} Y_i$, lo que también posee su propio error residual dado por $\sigma_Y^2 \triangleq \sum_{i=1}^{n} \left(Y_i - \frac{1}{n}\sum_{i=1}^{n} Y_i \right)^2$. Notemos que este valor corresponde a la varianza empírica de Y (salvo por un factor de n que falta)[9].

Entonces, con el objetivo de establecer un buen criterio de ajuste, definimos el coeficiente de determinación R^2 como:

$$R^2 = 1 - \frac{\sigma_R^2}{\sigma_Y^2}. \tag{3.161}$$

Un valor R^2 cercano a 1 significa una varianza residual 0 y el modelo explica con mucha precisión la variable Y. Por el contrario, un R^2 cercano a 0 significa un modelo de base, donde la predicción coincide con la media $\frac{1}{n}\sum_{i=1}^{n} Y_i$ y por lo tanto tiende a ser bastante pobre en desempeño ya que recordemos que la media empírica corresponde a la solución de un modelo agnóstico. Es posible tener coeficientes negativos lo que indicaría un ajuste con peor desempeño que haber utilizado la media empírica $\frac{1}{n}\sum_{i=1}^{n} Y_i$.

---

[9] Más precisamente la varianza empírica es $\frac{1}{n}\sum_{i=1}^{n} \left(Y_i - \frac{1}{n}\sum_{i=1}^{n} Y_i \right)^2$

3.7. Estadísticos Suficientes

En las secciones anteriores hemos visto dos maneras concretas de diseñar un estimador, mostramos garantías de optimalidad para el caso del estimador de máxima verosimiltud así como el estudiamos e interpretamos el estimador de mínimos cuadrados en un contexto de regresión.

El diseño de estimadores es un tema recurrente en estadística debido a que, formalmente, cualquier función de las observaciones puede ser considerado un estadístico, incluso la función $\tau_n(X_1^n) = 8$ es un estadístico. Luego, cuando buscamos a un buen estimador, resulta que no es necesario buscar todos los estimadores dentro de una familia arbitraria, sino que en una familia mucho más acotada.

La observación anterior entonces motiva a preguntarse si existen estimadores *clave* a partir de las observaciones que contengan toda la información sobre estos datos. De esto último podemos desprender la definición de estadístico suficiente.

Informalmente, un estadístico τ_n es suficiente si el estadista que conoce el valor de τ_n puede dar un desempeño igual de bueno para estimar θ que el estadista que conoce todo el vector de observaciones X_1^n. La definición formal es la siguiente:

Definición 3.10. (Estadístico Suficiente) Sea $\theta \in \Theta$ y una familia de distribuciones $\mathcal{F}_\Theta$, un estadístico[10] $\tau_n : \mathbb{X}^n \to \mathbb{R}$ se dice suficiente si por cada valor de $\tau_n = t$, la distribución condicional de X_1^n dado $\tau_n = t$ no depende de θ, es decir, es una función exclusiva de X_1^n

---

La definición anterior puede interpretarse de la siguiente manera: si la distribución de algo que se observa no depende de un parámetro, no puede proporcionarte información sobre él. Ahora bien, si la distribución de X_1^n depende del parámetro θ, y la distribución de X dado el estadístico suficiente τ_n no depende de θ, debe ser el caso de que toda la información sobre θ está contenida en τ_n; una vez que se conoce el valor de τ_n, el valor de X_1^n se vuelve irrelevante, porque la distribución condicional de X_1^n ya no depende de θ.

A partir de la definición no es evidente el poder encontrar un estadístico suficiente, sin embargo, el siguiente Teorema entrega una condicioń necesaria y suficiente para determinar uno.

[10] Notar que un estadístico es más general que un estimador porque el espacio de llegada puede ser distinto a Θ

Teorema 3.11. (Teorema de Factorización) Sea $\theta \in \Theta$ y una familia de distribuciones $\mathcal{F}_\Theta$, consideremos un vector aleatorio $X_1^n \sim P_{X_1^n}(\cdot|\theta)$. Un estadístico $\tau_n : \mathbb{X}^n \to \mathbb{R}$ es suficiente, si y sólo si, la verosimilitud puede descomponerse de la siguiente manera:

$$(\forall x_1^n \in \mathbb{X}^n) L(X_1 = x_1, ..., X_n = x_n|\theta) = u(x_1^n)v(\tau_n(x_1^n)|\theta) \tag{3.162}$$

Donde u y v son funciones no negativas. La función u puede depender de X_1^n, pero no del parámetro desconocido θ. La función v puede depender de θ, pero puede depender de x_1^n solo a través del valor de $\tau_n(X_1^n)$.

Demostración: Lo demostraremos para el caso discreto, el caso continuo se puede demostrar usando el teorema de cambio de variable multidimensional. Veremos ambas implicancias:

$\Leftarrow$ Sean $\theta \in \Theta$, $X_1^n = x_1^n$ y $\tau_n = t$ y asumimos que $P_{X_1^n}(X_1^n = x_1^n|\theta) = u(x_1^n)v(\tau_n(x_1^n)|\theta)$, vemos que:

$$\begin{aligned}
P_{X_1^n|\tau_n}(X_1^n = x_1^n|\tau_n = t, \theta) &= \frac{P_{X_1^n,\tau_n}(X_1^n = x_1^n, \tau_n = t|\theta)}{P_{\tau_n}(\tau_n = t|\theta)} \\
&= \frac{\mathbb{1}_{\tau_n^{-1}(\{t\})}(x_1^n)P_{X_1^n}(X_1^n = x_1^n|\theta)}{P_{\tau_n}(\tau_n = t|\theta)} \\
&= \frac{\mathbb{1}_{\tau_n^{-1}(\{t\})}(x_1^n)u(x_1^n)v(\tau_n(x_1^n) = t|\theta)}{P_{\tau_n}(\tau_n = t|\theta)} \\
&= \frac{\mathbb{1}_{\tau_n^{-1}(\{t\})}(x_1^n)u(x_1^n)v(\tau_n(x_1^n) = t|\theta)}{P_{X_1^n}(\tau_n^{-1}(\{t\})|\theta)} \\
&= \frac{\mathbb{1}_{\tau_n^{-1}(\{t\})}(x_1^n)u(x_1^n)v(\tau_n(x_1^n) = t|\theta)}{\sum\limits_{x_1^n \in \tau_n^{-1}(\{t\})} P_{X_1^n}(X_1^n = x_1^n|\theta)} \\
&= \frac{\mathbb{1}_{\tau_n^{-1}(\{t\})}(x_1^n)u(x_1^n)v(\tau_n(x_1^n) = t|\theta)}{\sum\limits_{x_1^n \in \tau_n^{-1}(\{t\})} u(x_1^n)v(\tau_n(x_1^n) = t|\theta)} \\
&= \frac{\mathbb{1}_{\tau_n^{-1}(\{t\})}(x_1^n)u(x_1^n)}{\sum\limits_{x_1^n \in \tau_n^{-1}(\{t\})} u(x_1^n)}.
\end{aligned} \tag{3.163}$$

Donde observamos que el resultado no depende de θ.

$\Rightarrow$ Si τ_n es suficiente, entonces, definamos:

$$v(\tau_n(X_1^n) = t|\theta) \triangleq P_{\tau_n}(\tau_n(X_1^n) = t|\theta) = \sum_{x_1^n \in \tau_n^{-1}(\{t\})} P_{X_1^n}(X_1^n = x_1^n|\theta) \quad (3.164)$$

y

$$u(x_1^n) \triangleq P_{X|\tau}(X_1^n = x_1^n|\tau_n(X_1^n) = \tau_n(x_1^n)), \quad (3.165)$$

donde (3.165) no depende de θ por suficiencia. Finalmente vemos que:

$$\begin{aligned} P_{X_1^n}(X_1^n = x_1^n|\theta) &= P_{X_1^n}(X_1^n = x_1^n, \tau_n(X_1^n) = \tau_n(x_1^n)|\theta) \\ &= P_{X_1^n}(X_1^n = x_1^n|\tau_n(X_1^n) = \tau_n(x_1^n))P_{\tau_n}(\tau_n(X_1^n) = \tau_n(x_1^n)|\theta) \\ &= u(x_1^n)v(\tau_n(X_1^n) = \tau_n(x_1^n)|\theta). \end{aligned} \quad (3.166)$$

□

Para finalizar esta sección veremos el ejemplo clásico de observaciones i.i.d. en distribuciones Gaussianas:

Ejemplo 3.10. Consideremos el caso de n variablea de observación i.i.d. $(X_1, ..., X_n) \sim P_X(\cdot|\theta)^n$. tal que $X_i \sim N(\theta, \sigma^2)$, es decir, la densidad marginal de cada observación está dada por:

$$P_X(\cdot|\theta) \mapsto f_X(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\theta)^2}{2\sigma^2}} \quad \forall x \in \mathbb{R}. \quad (3.167)$$

Para un $\theta \in \Theta$ obtenemos la verosimilitud

$$\begin{aligned} f_{X_1^n}(X_1, ..., X_n|\theta) &= \prod_{i=1}^{n} f_{X_i}(X_i|\theta) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(X_i-\theta)^2}{2\sigma^2}} \\ &= \frac{1}{(\sqrt{2\pi})^n \sigma^n} e^{\frac{-\sum_{i=1}^{n}(X_i-\theta)^2}{2\sigma^2}} \\ &= \frac{1}{(\sqrt{2\pi})^n \sigma^n} e^{\frac{-\left(\sum_{i=1}^{n} X_i^2 - 2X_i\theta + \theta^2\right)}{2\sigma^2}} \\ &= \frac{1}{(\sqrt{2\pi})^n \sigma^n} e^{\frac{-\sum_{i=1}^{n} X_i^2 + 2\sum_{i=1}^{n} X_i\theta - n\theta^2}{2\sigma^2}} \qquad (3.168) \\ &= \frac{1}{(\sqrt{2\pi})^n \sigma^n} e^{\frac{-\sum_{i=1}^{n} X_i^2}{2\sigma^2}} e^{\frac{2n\theta \frac{\sum_{i=1}^{n} X_i}{n} - n\theta^2}{2\sigma^2}} \\ &= \frac{1}{(\sqrt{2\pi})^n \sigma^n} e^{\frac{-\sum_{i=1}^{n} X_i^2}{2\sigma^2}} e^{\frac{2n\theta\tau_n(X_1^n) - n\theta^2}{2\sigma^2}}. \qquad (3.169) \end{aligned}$$

Identificamos:

$$u(X_1^n) = \frac{1}{(\sqrt{2\pi})^n \sigma^n} e^{\frac{-\sum_{i=1}^{n} X_i^2}{2\sigma^2}}, \qquad (3.170)$$

$$v(\tau_n(X_1^n)|\theta) = e^{\frac{2n\theta\tau_n(X_1^n) - n\theta^2}{2\sigma^2}}. \qquad (3.171)$$

Con lo que mostramos que el estadístico (estimador) $\tau_n(X_1^n) = \frac{\sum_{i=1}^{n} X_i}{n}$ es suficiente para estimar la media en el caso Gaussiano con varianza conocida.

3.8. Caso de Estudio: Astrometría y Fotometría

Contextualización y Modelamiento del Problema

Dos parámetros importantes para el estudio de la astronomía son la posición de los objetos luminosos en el cielo nocturno y la cantidad de luz (o más precisamente flujo) que llega desde el lugar donde se observa. La estimación de estos parámetros se conocen como

astrometría y fotometría, respectivamente. A través de las cámaras digitales CCD (*Charge Coupled Devices*) se puede abordar el problema de estimación (posición y flujo), contando la cantidad de fotones de la estrella que inciden en segmentos discretos del CCD llamados *pixeles*.

El astro o fuente puntual está en una posición u del cielo, emite un perfil de intensidad $F(x,u)$ de la forma (ver Fig. 3.1)

$$F(x,u) = F \cdot \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left(\frac{-(x-G(u))^2}{2\sigma_x^2}\right),\ \forall x\ \in \mathbb{R}. \tag{3.172}$$

Esta es la forma Gaussiana standard usada para modelar la propagación de luz y la difusión desde el astro al instrumento CCD. F corresponde a la intensidad total de luz emitida por el astro, σ_x corresponde al coeficiente de difusión, y por último $G(u)$ y x corresponden a puntos dentro del eje de medición del CCD. $G(u)$ es el mapeo entre la ubicación del astro en el espacio y su punto correspondiente en el eje de medición del CCD, de aquí en adelante dicho parámetro será renombrado como x_c, con lo que se tiene

$$F(x,x_c) = F \cdot \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left(\frac{-(x-x_c)^2}{2\sigma_x^2}\right),\ \forall x\ \in \mathbb{R}. \tag{3.173}$$

$F(x,x_c)$ en (3.173) e ilustrado en el manto Gaussiano de la Fig. 3.1 no es observado directamente en el CCD, sino que a través de tres fuente de perturbación, es decir, la medición en el arreglo de pixeles está sujeto a ruido. Estas fuentes de perturbación son:

a) Un perfil aditivo que captura la emisión de fotones, tanto de astros aledaños como de otros elementos del cielo nocturno (fotones que provienen de la luz de la luna por ejemplo) llamado *Background* (B) o Ruido de Fondo.
b) La cuantización espacial del perfil de la estrella al ser medida a través del arreglo de pixeles. Considerando el perfil dispersión gaussiano, la cuantización será (ver Fig. 3.2):

$$g_i(x_c) = \frac{1}{\sqrt{2\pi}\sigma_x} \int_{x_i-\Delta x/2}^{x_i+\Delta x/2} \exp\left(\frac{-(x-x_c)^2}{2\sigma_x^2}\right) dx \tag{3.174}$$

donde x_i corresponderá a la posición del pixel i-ésimo y Δx el tamaño de este (resolución del instrumento), el cual será constante a lo largo del arreglo.

c- El ruido de medición, el cual sigue una distribución de Poisson en cada pixel.

Integrando estos tres efectos, el modelo de observación del problema consiste en una colección de variables aleatorias independientes (observaciones) $\{I_i : i \in \mathbb{N}\}$ tales que

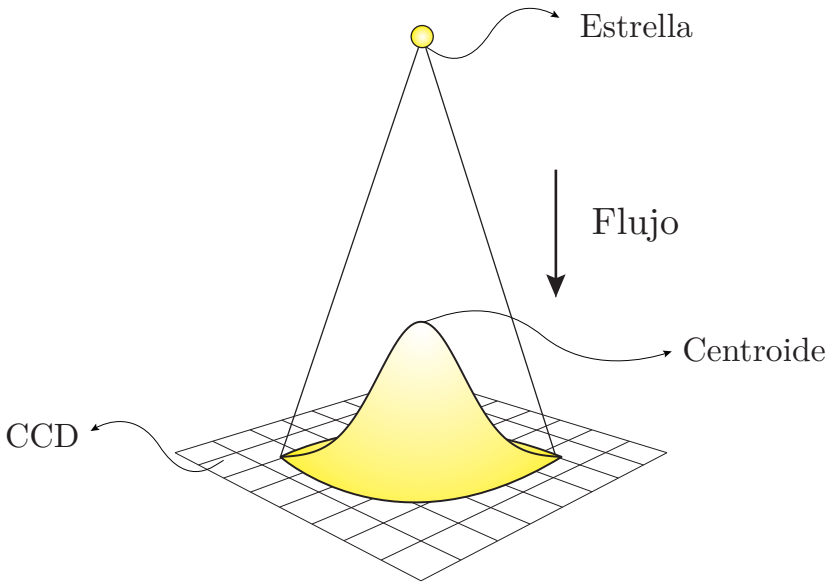


Figura 3.1: Dispersión de la luz en un arreglo de pixeles en una cámara CCD.

$$I_i \sim \text{Poisson}(\lambda_i(x_c, F)) \tag{3.175}$$

donde el parámetro $\lambda_i(x_c, F)$ (la media de la distribución Poisson)[11] recoge la observación en el pixel i-ésimo que se hubiera visto, es decir,

$$\lambda_i(x_c, F) = F \cdot g_i(x_c) + B. \tag{3.176}$$

Existen tres escenarios clásicos de estimación:

a- Astrometría: Se supondrán conocidos todos los parámetros salvo x_c y el problema es estimar x_c de $\{I_i, i \in \mathbb{N}\}$.
b- Fotometría, estimación de flujo: Se supondrán conocidos todos los parámetros salvo F y el problema es estimar F de $\{I_i, i \in \mathbb{N}\}$.
c- Fotometría, estimación de ruido de fondo: Se supondrán conocidos todos los parámetros salvo B y el problema es estimar B de $\{I_i, i \in \mathbb{N}\}$.

[11] Recuerde que si $X \sim Poisson(\lambda)$ entonces $P_X(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$

Se asumirá una cantidad finita de observaciones $\{I_i, i = 1, ..., n\}$, donde se asume una buena cobertura del objeto que se mide, en el sentido que:

$$\sum_{i=1}^{n} g_i(x_c) \approx \sum_{i\in\mathbb{Z}} g_i(x_c) \approx \frac{1}{\sqrt{2\pi}\sigma_x} \int_{-\infty}^{\infty} \exp\left(\frac{-(x-x_c)^2}{2\sigma_x^2}\right) dx = 1 \tag{3.177}$$

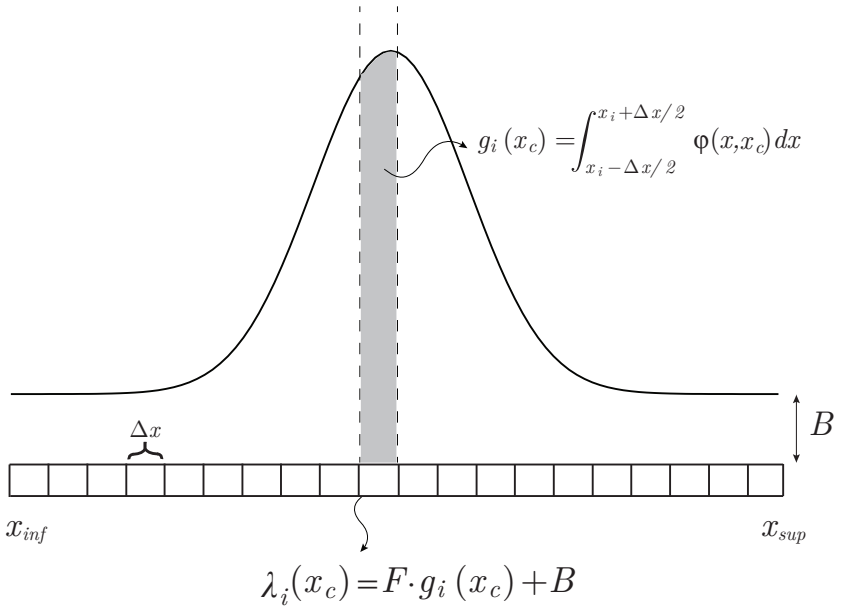


Figura 3.2: Descripción del modelo de adquisición digital (discreta) de datos en un arreglo unidimensional de pixeles

---

Problema 3.1. (Resultados Preliminares). Verifique que se cumple lo siguiente:

3.1. $\dfrac{\partial g_i(x_c)}{\partial x_c} = \dfrac{1}{\sqrt{2\pi}\sigma_x}\left[\exp\left(-\left(\dfrac{x_i - \frac{\Delta x}{2} - x_c}{\sqrt{2}\sigma_x}\right)^2\right) - \exp\left(-\left(\dfrac{x_i + \frac{\Delta x}{2} - x_c}{\sqrt{2}\sigma_x}\right)^2\right)\right]$

3.2. $\displaystyle\sum_{i=1}^{N} \frac{\partial g_i(x_c)}{\partial x_c} \approx 0$

3.3. $\mathrm{E}\left(I^2\right) = \lambda^2 + \lambda$, para ello utilice propiedades de la varianza de la distribución de Poisson.

3.8.1. Astrometría, Estimación de Posición

Suponiendo conocidos los parámetros F, σ_x y B se tiene que el vector variable de observaciones $I_1^n = (I_1, ..., I_n) \ \in \mathbb{X}^n$ sigue una distribución de probabilidad según (3.175) la cual depende del parámetro $x_c \in \Theta$, siendo $\Theta = \mathbb{R}$ el espacio de parámetros posibles. Considere la familia de estimadores insesgados:

$$\mathcal{T}^n = \{\tau_n : \mathbb{X}^n \to \Theta; \mathbb{E}_(\tau_n(I_1^n)) = \theta \ \text{ para todo } \theta \in \Theta\} \tag{3.178}$$

el siguiente problema tiene como objetivo encontrar un estimador $\hat{x}_c : \mathbb{X}^n \to \Theta \in \mathcal{T}^n$, tal que:

$$\begin{aligned} \hat{x}_c(I_1^n) &\triangleq \arg \min_{\tau_n \in \mathcal{T}^n} Var(\tau_n(I_1^n)) \\ &= \arg \min_{\tau_n \in \mathcal{T}^n} \mathbb{E}\left((\tau_n(I_1^n) - x_c)^2\right). \end{aligned} \tag{3.179}$$

Problema 3.2. (Límites Fundamentales)

a) Demuestre que la cota de Cramér-Rao para el parámetro x_c está dada por:

$$Var(\hat{x}_c(I_1^n)) \geq \frac{1}{\displaystyle\sum_{i=1}^{n} \frac{\left(F \frac{\partial g_i(x_c)}{\partial x_c}\right)^2}{F \cdot g_i(x_c) + B}} \tag{3.180}$$

b) Analice si existe algún estimador insesgado de x_c que alcance la cota de Cramér-Rao.

3.8.2. Fotometría, Estimación del Flujo

Suponiendo conocidos los parámetros x_c, σ_x y B se tiene que el vector variable de observaciones $I_1^n = (I_1, ..., I_n) \ \in \mathbb{X}^n$ sigue una distribución de probabilidad según (3.175) la cual depende del parámetro $F \ \in \Theta$, siendo $\Theta = \mathbb{R}^+$ el espacio de parámetros posibles.

Problema 3.3. (Límites Fundamentales)

a) Determine una expresión cerrada para:

$$\ln L(I_1, .., I_n|F). \tag{3.181}$$

b) Verifique la siguiente identidad:

$$\frac{d}{dF} \ln L(I_1, .., I_n|F) = \sum_{k=1}^{n} \left[\frac{g_k(x_c) \cdot I_k}{\lambda_k(x_c, F)} - g_k(x_c) \right]. \tag{3.182}$$

c) Si definimos la variable aleatoria $Y_k = \frac{g_k(x_c) \cdot I_k}{\lambda_k(x_c, F)} - g_k(x_c)$, verificar que es una variable aleatoria de media cero. Con ello demuestre que la Información de Fisher del problema está dada por:

$$\mathcal{I}_n(F) \triangleq \mathbb{E}\left\{ \left(\frac{d}{dF} \ln L(I_1, .., I_n|F) \right)^2 \right\} = \sum_{k=1}^{n} \frac{g_k^2(x_c)}{F \cdot g_k(x_c) + B}. \tag{3.183}$$

Indicación: Recordar la propiedad de la varianza sobre la suma de variables aleatorias independientes.

d) Considere el régimen de alta relación señal a ruido cuando se cumple que: $F g_k(x_c) \gg B$. Demuestre en este caso que:

$$min_{\tau_n \in T_n} Var(\tau_n(I_1, .., I_n)) \geq F \tag{3.184}$$

donde T_n denota la familia de estimadores insesgados.

e) Verifique si en este problema existe un estimador insesgado que alcance la cota de Cramer-Rao.

Problema 3.4. (Estimador Mínimos Cuadrados) Para el problema de fotometría presentado anteriormente, analizaremos el estimador Least Square, solución del siguiente problema de optimización:

$$F_{LS}^*(I_1, .., I_n) = \arg \min_{F \geq 0} \sum_{k=1}^{n} (I_k - \lambda_k(x_c, F))^2. \tag{3.185}$$

a) Determine una expresión cerrada para $F^*_{LS}(I_1, .., I_n)$ como función de los datos medidos.
b) Verifique si F^*_{LS} es un estimador insesgado de F y determine la varianza del estimador.
c) Compare la varianza del estimador LS con la cota de Cramér-Rao de la pregunta anterior. Comente.

3.9. Problemas

Se presentan a continuación una sección de problemas relacionados con estimación paramétrica.

Problema 3.5. Bajo las condiciones de regularidad de la Definición 3.2, verifique si el vector aleatorio $X_1, ..., X_n$ es i.i.d. con distribución $P_{X_1^n}(\cdot|\theta)$ entonces

$$\mathbb{E}_{X_1^n}\left(\left(\frac{\partial \ln L(X_1, ..., X_n|\theta)}{\partial \theta}\right)^2\right) = -\mathbb{E}_{X_1^n}\left(\frac{\partial^2 \ln L(X_1, ..., X_n|\theta)}{\partial \theta^2}\right). \tag{3.186}$$

Problema 3.6. Compruebe que si el vector aleatorio $X_1, ..., X_n$ es i.i.d. con distribución $P_{X_1^n}(\cdot|\theta)$ entonces la información de Fisher es aditiva, es decir:

$$\mathcal{I}_n(\theta) \triangleq \mathbb{E}_{X_1^n}\left(\left(\frac{\partial \ln L(X_1, X_2, ..., X_n|\theta)}{\partial \theta}\right)^2\right) = n \cdot \mathcal{I}_1(\theta). \tag{3.187}$$

Problema 3.7. Muestre que para cualquier estimador τ_n de θ su error de estimación se puede descomponer como varianza mas sesgo, es decir:

$$\mathbb{E}_{X_1^n}\left((\tau_n(X_1^n) - \theta)^2\right) = Var(\tau_n(X_1^n)) + \left(\mathbb{E}_{X_1^n}(\tau_n(X_1^n)) - \theta\right)^2. \tag{3.188}$$

Problema 3.8. Considere el problema de estimación paramétrico sobre la familia $\mathcal{F}_\Theta$ visto en clase.

a) Considere que $(X_1, ..., X_n)$ un vector i.i.d. con valores en $\{0, 1\}$ que sigue una distribución Bernoulli de parámetro $\theta \in [0, 1]$. Es decir $P_{X_i}(X_i = 1|\theta) = \theta$. Determine una expresión para $L(X_1, .., X_n|\theta)$ y verifique que ofrece la siguiente descomposición:

$$P_{X_1^n}(X_1^n = x_1^n|\theta) = P_{X_1^n}(\tau_n(X_1^n) = \tau_n(x_1^n)|\theta) \cdot P_{X_1^n}(X_1^n = x_1^n|\tau_n(X_1^n) = \tau_n(x_1^n), \theta)$$

donde

$$\tau_n(X_1^n) = X_1 + ... + X_n \tag{3.189}$$

y

$$P_{X_1^n}(X_1^n = x_1^n | \tau_n(X_1^n) = \tau_n(x_1^n), \theta) = \frac{1}{\binom{n}{\tau_n(x_1^n)}}. \tag{3.190}$$

En otras palabras se tiene que τ_n en Eq.(3.189) es un estadístico suficiente para inferir θ.

b) Del punto anterior verifique que el estimador de máxima verosimilitud esta dado por:

$$\tau_{ML}(X_1^n) = \tau_n(X_1^n)/n.$$

c) Finalmente demuestre que $\tau_{ML}(X_1^n)$ es insesgado, calcule su varianza y demuestre que es mínima.

Problema 3.9. Sea $X_1, ..., X_n$ una secuencia i.i.d. que sigue una distribución exponencial, es decir, su densidad está dada por $f_X(x|\theta) = \theta \cdot e^{-\theta x}$ con $x \in \mathbb{R}^+ \cup \{0\}$,

a) Determine el estimador de máxima verosimilitud
b) Determine $\mathbb{E}_{X_1^n}(\tau_{ML}(X_1^n))$
c) ¿Es este estimador de mínima varianza?

Problema 3.10. Sea $X_1, ..., X_n$ un vector aleatorio i.i.d. uniformemente distribuida en $[0, \theta]$ con θ.

a) Determine el estimador de máxima verosimilitud de θ y verifique que esta dado por

$$\tau_{ML}(X_1^n) = \text{máx}\{X_i : i = 1, ..., n\}.$$

b) Demuestre que $\tau_{ML}(X_1, ..., X_n)$ es sesgado.

Problema 3.11. Considere un sistema de modulación AM que genera la señal discreta

$$s_k = A \cdot \cos\left(\frac{2\pi}{T} \cdot k\right) \quad k \in \{1, .., n\} \tag{3.191}$$

que depende del parámetro A y donde $T > 0$ es un número entero conocido. El vector s_1^n no es observable directamente, si no que por medio de un ruido aditivo:

$$X_k = s_k + N_k \tag{3.192}$$

donde $N_1, N_2, .., N_n$ son variables aleatorias Gaussianas independientes e idénticamente distribuidas con media cero y varianza σ^2.

a) Notar que $X_1, ..., X_n$ es un vector Gaussiano. Con ello determine su vector de media y matriz de covarianza.
b) Del punto anterior determine la función de verosimilitud $L(X_1, ..., X_n|A)$ y con ello el estimador de máxima verosimilitud de A dadas las variables de observación $X_1, ..., X_n$. Es decir la solución de:

$$\hat{A}_{ML}(X_1^n) = \arg \max_{A \in \mathbb{R}^+} \ln L(X_1, ..., X_n|A). \tag{3.193}$$

Indicación: Se debe llegar a una expresión cerrada función de $X_1, ..., X_n$ y parámetros conocidos del problema.
c) Verifique que $\hat{A}_{ML}(X_1^n)$ es insesgado y determine su varianza.
d) Demuestre que $\hat{A}_{ML}(X_1^n)$ es el estimador insesgado de A de mínima varianza.
e) Demuestre que $\hat{A}_{ML}(X_1^n)$ es un estimador consistente de A cuando $n \to \infty$.

Problema 3.12. Considere un cuerpo radiactivo que emite θ partículas, con $\theta \in \mathbb{N}$. Para detectar las partículas emitidas, se cuenta con un detector imperfecto, el cual detecta cada partícula emitida de forma independiente. Para modelar el proceso de detección, consideremos la variable aleatoria B_i que toma el valor 1 si la partícula i-ésima fue detectada y 0 si no, donde B_i distribuye Bernoulli de parámetro p ($P_{B_i}(B_i = 1) = p$). Finalmente, la variable de observación X es el número de partículas totales detectadas dada por

$$X = \sum_{i=1}^{\theta} B_i.$$

Notar que dados p y θ conocidos, X distribuye binomial de parámetros p y θ, es decir:

$$P_X(X=k) = \binom{\theta}{k} p^k (1-p)^{\theta-k}$$

a) Asuma que conoce la cantidad de partículas emitidas θ. Determine el estimador de máxima verosimilitud del parámetro p dada una variable de observación de $X \in \{0, \cdots, \theta\}$.
b) Ahora considere que se cuenta con n realizaciones i.i.d. de la variable $X \in \{0, \cdots, \theta\}$. Determine la información de Fisher asociada al parámetro p, la cota de Cramér-Rao y verifique si existe un estimador que la alcance.
c) Ahora considere conocido el parámetro p. Determine el estimador de máxima verosimilitud del parámetro θ (cantidad de partículas emitidas) τ_{ML} dada una observación de X. Analice si el estimador es insesgado y determine su varianza. Utilice la aproximación $\log \theta! \approx \theta \log \theta - \theta$.

Problema 3.13. En muchas aplicaciones de laboratorio, es posible obtener valores con alta precisión llamado *Media de Población*. Este valor puede ser beneficioso ya que permite obtener estimadores (sesgados) con un error de estimación menor que la cota de Cramér-Rao. En particular, considere el caso Gaussiano, i.e, la densidad está dada por:

$$f_X(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\theta)^2}{2\sigma^2}} \quad \forall x \in \mathbb{R} \tag{3.194}$$

donde σ es conocido y queremos estimar θ a partir de $(X_1, ..., X_n)$, variables de observaciones independientes e idénticamente distribuidas. Se pide:

a) Verifique que la información de Fisher está dada por

$$\mathcal{I}_n(\theta) = \mathbb{E}_{X_1^n}\left(\left(\frac{\partial \ln L(X_1, ..., X_n|\theta)}{\partial \theta}\right)^2\right) = \frac{n}{\sigma^2}. \tag{3.195}$$

b) Demuestre la existencia de un estimador insesgado que alcance la cota de Cramér-Rao, es decir, encontrar y explicitar τ_n^* tal que

$$\mathbb{E}_{X_1^n}((\tau_n^*(X_1^n) - \theta)^2) = \frac{1}{\mathcal{I}_n(\theta)}. \tag{3.196}$$

Indicación: Utilizar la condición de alcanzabilidad de la función de verosimilitud.

c) Si definimos la media de población como

$$\gamma = \frac{\sigma}{\theta}$$

y estudiamos un estimador de θ de la siguiente forma:

$$\tau_n^C(X_1^n) = C \cdot \sum_{i=1}^{n} X_i \tag{3.197}$$

con $C \in \mathbb{R}^+$ el parámetro a definir.

Muestre que el valor de C óptimo (que minimiza el error cuadrático medio de estimación), es decir el C^* solución del problema:

$$\min_{C \in \mathbb{R}^+} \mathbb{E}_{X_1^n} \left(\left(\tau_n^C(X_1^n) - \theta \right)^2 \right), \tag{3.198}$$

está dado por:

$$C^* = (N + \gamma^2)^{-1}. \tag{3.199}$$

d) Verifique que :

$$\tau_n^{C^*}(X_1^n) = C^* \sum_{i=1}^{n} X_i \tag{3.200}$$

es sesgado y contraste este estimador con el obtenido en el punto (b). En particular verifique que $\tau_n^{C^*}(\cdot)$ tiene una varianza menor que el estimador $\tau_n^*(\cdot)$ encontrado en (b).

e) Se define la eficiencia de $\tau_n^{C^*}$ relativa a τ_n^* como el cuociente de sus errores cuadráticos medios, es decir:

$$eficiencia(\tau_n^{C^*}) = \frac{\mathbb{E}_{X_1^n} \left((\tau_n^*(X_1^n) - \theta)^2 \right)}{\mathbb{E}_{X_1^n} \left(\left(\tau_n^{C^*}(X_1^n) - \theta \right)^2 \right)}. \tag{3.201}$$

Verifique que $\tau_n^{C^*}$ es asintóticamente insesgado y que $\tau_n^{C^*}$ es eficiente en el sentido que:

$$\lim_{n \to \infty} eficiencia(\tau_n^{C^*}) = 1. \tag{3.202}$$

Problema 3.14.
El caso de distribuciones exponenciales es emblemático tanto por su simplicidad analítica, así como por su uso como modelamiento del tiempo de falla en sistemas complejos. Consideremos $\mathbb{X} = \mathbb{R}^n$, $\Theta = \mathbb{R}^+$ y una secuencia $X_1^n = X_1, ..., X_n$ de variables aleatorias independientes e identicamente distribuidas (i.i.d.) tales que $X_i \sim Exponencial\left(\frac{1}{\lambda}\right)$, $i \in \{1, ..., n\}$ con $\lambda > 0$ un parámetro.

Indicación: Si $X \sim Exponencial\,(\alpha)$ entonces su función de densidad de probabilidad es

$$f_X(x) = \begin{cases} \alpha e^{-\alpha x} & \text{para } x > 0 \\ 0 & \text{para } x \leq 0 \end{cases}$$

Donde además $\mathbb{E}(X) = \frac{1}{\alpha}$ y $Var(X) = \frac{1}{\alpha^2}$.

a) Obtenga el estimador de máxima verosimilitud $\hat{\lambda}_{ML}(\cdot)$ del parámetro λ.
b) Calcule sesgo y varianza del estimador obtenido en la parte a).
c) Demuestre que la Información de Fisher asociado al parámetro λ está dado por:

$$\mathcal{I}_n(\lambda) = \frac{n}{\lambda^2}.$$

¿Es el estimador encontrado en la parte **(a)** de mínima varianza?
d) La estudiante PAT desconoce la distribución asociada a las observaciones recibidas, sin embargo, es muy astuta por lo que propone el siguiente modelo de observación:

$$Y_i = \frac{n+1}{n}\lambda + V_i \quad i \in \{1, ..., n\}$$

Donde V_i corresponde al ruido y puede asumir que es despreciable. Calcule el estimador lineal de mínimos cuadrados. Además obtenga su varianza y compárela con la cota de Cramér-Rao. Comente por qué se tiene este resultado.
e) (PENDIENTE)[12] Suponga ahora que λ es una variable aleatoria que sigue una distribución $U(a, b)$, con $a, b \in \mathbb{R}^+$. Demuestre que el estimador Máxima a Posteriori $\phi_{MAP}(\cdot)$ dadas las variables de observación i.i.d. $X_i \sim Exponencial\left(\frac{1}{\lambda}\right)$,

[12] Se necesitan herramientas de Estimación Bayesiana

$i \in \{1, ..., n\}$ está dado por:

$$\phi_{MAP}(x_1, ..., x_n) = \begin{cases} a & \text{si } a > \hat{\lambda}_{ML}(x_1, ..., x_n) \\ \hat{\lambda}_{ML}(x_1, ..., x_n) & \text{si } a \leq \hat{\lambda}_{ML}(x_1, ..., x_n) \leq b \\ b & \text{si } b < \hat{\lambda}_{ML}(x_1, ..., x_n) \end{cases}$$

donde $\hat{\lambda}_{ML}(\cdot)$ es el estimador encontrado en la parte **(a)**.

Indicación: Si $X \sim U(a, b)$ entonces su función de densidad de probabilidad es

$$f_X(x) = \begin{cases} \dfrac{1}{b-a} & \text{si } a \leq x \leq b \\ 0 & \sim \end{cases}$$

4

Unidad IV: Estimación Bayesiana

En la unidad anterior ya hablamos del principio de estimación paramétrica. Vimos que formalmente el problema de estimación se entiende como la inferencia de una variable θ continua (que toma una cantidad no numerable de posibles valores) a partir de una variable aleatoria (o vector aleatorio) de observación X, donde naturalmente θ influyó en la observación x.

En este contexto Bayesiano el problema de inferencia la idea nuevamente es plantear un problema de decisión sobre θ, pero donde el parámetro θ es ahora un objeto aleatorio con valores en $\mathcal{A} = \mathbb{R}^d$. De manera análoga a lo que ocurrió en el caso de detección Bayesiana, en este caso tenemos acceso a un conocimiento *a priori* y además el proceso de inferencia, como veremos más adelante, requerirá el uso del Teorema de Bayes de manera frecuente.

Los métodos Bayesianos tienen en común la asignación de una probabilidad como medida de credibilidad de las hipótesis. En este contexto, la inferencia se entiende como un proceso de actualización de las medidas de credibilidad al conocerse nuevas evidencias u observaciones. Mediante la aplicación del Teorema de Bayes se busca obtener las probabilidades de las hipótesis condicionadas a las evidencias que se conocen.

Si recordamos la regla de Bayes tenemos lo siguiente:

$$P_{\Theta|X}(\Theta = \theta|X = x) = \frac{P_{X|\Theta}(X = x|\Theta = \theta)p_{\Theta}(\theta)}{P_X(X = x)} \tag{4.1}$$

En inferencia Bayesiana cada uno de los términos tiene una interpretación, los cuales se mencionarán a continuación:

- $P_{\Theta}(\Theta = \theta)$ se conoce como la información *a priori*, corresponde a la información inicial que se posee. Es una suposición que se tiene para luego ir actualizándola con las observaciones.
- $P_{X|\Theta}(X = x|\Theta = \theta)$ se conoce como la verosimilitud. Este concepto ya ha aparecido con anterioridad y normalmente este valor se asocia a los datos X que surgieron dada la naturaleza de Θ.
- $P_{\Theta|X}(\Theta = \theta|X = x)$ se conoce como la información *a posteriori*. Es el resultado obtenido después de haber actualizado el conocimiento *a priori* con las observaciones recibidas. Esto permite enriquecer la suposición inicial dada por la información *a priori*.
- $P_X(X = x)$ es un normalizador, es un valor que permite que la regla de bayes siga actuando como medida de probabilidad.

Esencialmente la regla de Bayes establece que

$$\text{Nuevo Conocimiento} \propto \text{Evidencia} \cdot \text{Conocimiento Inicial}$$

Veremos en adelante que la derivación formal es análoga al caso paramétrico, pero en este contexto podremos usar una mayor cantidad de herramientas del cálculo. Para comenzar, dado que ahora poseemos una variable o vector aleatorio Θ significa que tenemos una distribución de probabilidad:

$$P_{\Theta}(B) = \mathbb{P}(\Theta(w) \in B) \tag{4.2}$$

Esta distribución se conoce normalmente como distribución **a priori**. En particular, en estimación Θ es no numerable por lo que vamos a suponer que está dotado de una función de densidad de probabilidad dada por:

$$f_{\Theta}(\theta) \quad \forall \theta \in \mathcal{A}, \tag{4.3}$$

y, por lo tanto,

$$(\forall B \subset \mathbb{R}^d) \quad P_{\Theta}(B) = \int_B f_{\Theta}(\theta_1, ..., \theta_d) d\theta_1 ... d\theta_d. \tag{4.4}$$

Antes de formalizar el problema, se dará un pequeño contraste entre estimación paramétrica y Bayesiana.

4.0.1. Estimación Paramétrica versus Estimación Bayesiana

Como veremos más adelante, la formalización cambia levemente entre un problema u otro. Estimación paramétrica también se conoce como estimación frecuentista. Esto porque está basado en análisis empíricos inspirados en resultados asintóticos de la Ley de los Grande Números.

Para los frecuentistas, los parámetros se consideran "verdaderos fijos y desconocidos" mientras que los datos son aleatorios. Sin embargo, en la realidad, nunca se conocerá el valor del parámetro, pero esto en estimación paramétrica no es un problema en la medida que se pueda desarrollar teoría y métodos que entreguen algún grado de universalidad (sesgo o consistencia).

Para los Bayesianos, en cambio, los datos son fijos (pese a que provengan de una fuente aleatoria), mientras que el parámetro es desconocido y no se puede conocer. Las nociones de sesgo son irrelevantes en este caso porque solo se puede imaginar que el sesgo es relevante en un mundo donde ya conoces el parámetro. Pero si ya se conoce el parámetro, no tiene sentido entonces querer obtener un estimador.

Sin embargo, si se observan los métodos Bayesianos desde una perspectiva frecuentista, es importante notar que a veces inducir sesgo puede ser útil. Esto puede deberse a que se desea minimizar la varianza, o alternativamente, la disminución puede ser útil en muestras finitas. Por otra parte, es totalmente posible que la elección de un prior sea inapropiada y se termine desviando las estimaciones en la dirección incorrecta.

Como último comentario, la noción de sesgo no tiene realmente el mismo significado en estadística Bayesiana. El sesgo es una propiedad de un estimador, no de una estimación. El concepto de sesgo está condicionado a un valor verdadero del parámetro, esto último es un término que no existe en estimación Bayesiana

4.1. Formalización del Problema de Estimación Bayesiana

Un problema de estimación Bayesiano se compone de 5 elementos centrales:

- Un espacio de observación $\mathbb{X}$ y variables aleatorias de observación que toman valores en $\mathbb{X}$. El valor particular que toma la variable X, es decir, x se conoce como observación, realización o dato. $\mathbb{X}$ es un espacio numérico abstracto y también puede ser multidimensional, por ejemplo, $\mathbb{X} = \mathbb{R}^n$ con $n \in \mathbb{N}$ en cuyo caso las observaciones corresponden a un vector aleatorio $X_1^n \in \mathbb{X}$.
- Un espacio de decisión $\mathcal{A}$ infinito no numerable y una variable aleatoria Θ con valores en $\mathcal{A}$. Además se posee una distribución de probabibilidad sobre Θ, P_Θ la cual se conocerá como distribución *a priori* o *prior*. La distribución a priori *debe* ser continua, por lo que está dotado de una función de densidad de probabilidad $f_\Theta(\cdot)$
- Distribuciones de probabilidad condicionales indexadas por $\theta \in \Theta$, es decir, $\mathcal{F}_\mathcal{A} = \{P_X(\cdot|\Theta = \theta), \theta \in \mathcal{A}\}$.
- Un estimador $\phi : \mathbb{X} \mapsto \mathcal{A}$ que será la función que tomará una decisión en base a algún criterio.
- Una función de costo o riesgo $L : \mathcal{A} \times \mathcal{A} \to \mathbb{R}^+$ que penaliza la incorrecta decisión.

Análogo al caso de detección Bayesiano hablaremos más en detalle del riesgo ya que es un elemento nuevo respecto al caso paramétrico.

4.2. Riesgo Promedio Bayesiano

Para continuar con la formalización del problema de detección Bayesiano, se debe buscar la regla óptima respecto a un criterio, este criterio no es más que el promedio del costo dado por la variable aleatoria del riesgo $L(\Theta, \phi(X))$.

En este caso el problema de estimar Θ a partir de X, se basa en minimizar una función de riesgo $L : \mathcal{A} \times \mathcal{A} \to \mathbb{R}^+$ o también llamada función de costo o error.

Para cada $\theta_1, \theta_2 \in \mathcal{A}$, $L(\theta_1, \theta_2)$ cuantifica el error de estimar θ_2 cuando el parámetro real es θ_1. Las siguientes definiciones ayudarán a establecer el criterio de optimalidad en el sentido Bayesiano:

Definición 4.1. (Riesgo Promedio) Consideremos una función de riesgo: $L : \mathcal{A} \times \mathcal{A} \to \mathbb{R}^+ \cup \{0\}$ que penaliza los errores en la toma de decisión y un estimador: $\phi : \mathbb{X} \to \mathcal{A}$. Dado un θ que determina las estadísticas de las observaciones $X \sim P_{X|\Theta}(\cdot|\Theta = \theta)$, definimos el

riesgo promedio $R : \mathcal{A} \times F(\mathbb{X}, \mathcal{A}) \to \mathbb{R}^+ \cup \{0\}$ condicionado a θ como:

$$R(\theta, \phi) \triangleq \mathbb{E}(L(\theta, \phi(X))|\Theta = \theta) = \begin{cases} \underbrace{\int_{\mathbb{X}} L(\theta, \phi(x)) f_{X|\Theta}(x|\theta) dx}_{\text{Caso espacio continuo con f.d.p condicional}} \\ \underbrace{\sum_{x \in \mathbb{X}} L(\theta, \phi(x)) P_{X=x|\Theta=\theta}(x|\theta)}_{\text{Caso espacio condicinal discreto}} \end{cases} \tag{4.5}$$

Para asegurar que este riesgo está bien definido, nos restringiremos al conjunto de reglas o test tal que el riesgo $R(\theta, \phi)$ existe y es finito para todo $\theta \in \mathcal{A}$. La expresión (4.5) está condicionada a una realización de Θ. Por lo tanto $R(\Theta, \phi)$ es una variable aleatoria (función de Θ y ϕ). Para cada regla ϕ podemos definir su promedio respecto a θ, llamado Riesgo Promedio Bayesiano:

Definición 4.2. (Riesgo Promedio Bayesiano) Sea, $\phi \in F(\mathbb{X}, \mathcal{A})$, una distribución P_Θ que admite densidad y su riesgo promedio $R(\theta, \phi)$. Definimos el Riesgo Promedio Bayesiano como el promedio de $R(\Theta, \phi)$ con respecto a la variable Θ (asumiremos el caso continuo para X dado Θ, el caso discreto es análogo):

$$\begin{aligned} r(\phi) &\triangleq \mathbb{E}_\Theta(R(\Theta, \phi)) \\ &= \int_{\mathcal{A}} R(\theta, \phi) \cdot f_\Theta(\theta) d\theta \\ &= \int_{\mathcal{A}} \mathbb{E}\left(L(\theta, \phi(X))|\Theta = \theta\right) \cdot f_\Theta(\theta) d\theta \\ &= \int_{\mathcal{A}} \int_{\mathbb{X}} L(\theta, \phi(x)) \cdot f_{X|\Theta}(x|\theta) \cdot f_\Theta(\theta) dx d\theta \\ &= \int_{\mathcal{A}} \int_{\mathbb{X}} L(\theta, \phi(x)) f_{X,\Theta}(x, \theta) dx d\theta \\ &= \mathbb{E}_{\Theta,X}(L(\Theta, \phi(X))). \end{aligned} \tag{4.6}$$

Análogo al problema de detección Bayesiana, se buscará establecer un criterio óptimo general, lo que posteriormente se convertirán en estimadores concretos dependiendo al función de costo utilizada.

4.3. Decisión Óptima: Distribución a Posteriori

Recapitulando, la regla óptima Bayesiana dependerá de los siguientes elementos previamente introducidos:

i) P_Θ distribución a priori dotado de una función de densidad de probabilidad $f_\Theta(\theta)$.
ii) $f_{X|\Theta}(\cdot|\theta)$, función de densidad de probabilidad condicional (o de masa según sea el caso).
iii) $L : \mathcal{A} \times \mathcal{A} \to \mathbb{R}^+$, función de costo.

Luego, la solución del problema de detección Bayesiana es aquella función que minimiza el riesgo promedio Bayesiano, es decir, se plantea como:

$$\begin{aligned}\phi^* &= \arg \min_{\phi:\mathbb{X}\to\mathcal{A}} r(\phi) \\ &= \arg \min_{\phi:\mathbb{X}\to\mathcal{A}} \mathbb{E}_{\Theta,X}(L(\Theta, \phi(X))).\end{aligned} \tag{4.7}$$

Por lo tanto, ϕ^* es la regla que minimiza el riesgo Bayesiano. Esta expresión en principio no garantiza unicidad y además no da una manera cerrada de poder encontrar tal regla. Luego, si analizamos de forma más detallada la función objetivo en (4.7) tenemos lo siguiente:

$$\begin{aligned}\mathbb{E}_{\Theta,X}(L(\Theta, \phi(X))) &= \int_{\mathcal{A}} \int_{\mathbb{X}} L(\theta, \phi(x)) f_{X,\Theta}(x, \theta) dx d\theta \\ &= \int_{\mathbb{X}} \int_{\mathcal{A}} L(\theta, \phi(x)) f_{\Theta|X}(\theta|x) d\theta f_X(x) dx\end{aligned} \tag{4.8}$$

Se puede notar que el término

$$\int_{\mathcal{A}} L(\theta, \phi(x)) f_{\Theta|X}(\theta|x) d\theta \tag{4.9}$$

es función exclusiva de la evaluación de $\phi(\cdot)$ en el punto x y no de los restantes valores $\phi(y)$ que adopta en $y \in \mathbb{X}\backslash\{x\}$. Por lo tanto, minimizar (4.7) equivale a minimizar el argumento de la función (4.8) punto a punto, es decir, para cualquier observación o $\forall x \in \mathbb{X}$, $\phi^*(x)$ es solución de:

$$\begin{aligned}\phi^*(x) &= \arg \min_{y\in\mathcal{A}} \int_{\mathcal{A}} L(\theta, y) f_{\Theta|X}(\theta|x) d\theta, \quad \forall x \in \mathbb{X} \\ &= \arg \min_{y\in\mathcal{A}} \mathbb{E}\left(L(\Theta, y)|X = x\right), \quad \forall x \in \mathbb{X}\end{aligned} \tag{4.10}$$

Observaciones 4.1. Interpretando la regla óptima Bayesiana en (4.10), dada una observación x, $\phi(x)$ es la decisión que minimiza el riesgo promedio, respecto a la distribución a posteriori de Θ dado el evento (observación) $X = x$.

Por Bayes sabemos que la distribución a posteriori se obtiene como:

$$f_{\Theta|X}(\theta|x) = \frac{f_{\Theta,X}(\theta, x)}{f_X(x)} = \frac{f_{X|\Theta}(x|\theta) f_\Theta(\theta)}{\int\limits_{\mathcal{A}} f_{X|\Theta}(x|\tilde{\theta}) f_\Theta(\tilde{\theta}) d\tilde{\theta}} \tag{4.11}$$

donde

$$f_X(x) = \int_{\mathcal{A}} f_{X,\Theta}(x, \tilde{\theta}) d\tilde{\theta} = \int_{\mathcal{A}} f_{X|\Theta}(x|\tilde{\theta}) f_\Theta(\tilde{\theta}) d\tilde{\theta}. \tag{4.12}$$

se obtuvo mediante el uso de probabilidades totales. De esta manera la regla de decisión óptima es solución de

$$\phi^*(x) = \arg\min_{y \in \mathcal{A}} \int_{\mathcal{A}} L(\theta, y) \frac{f_{X|\Theta}(x|\theta) f_\Theta(\theta) d\theta}{\int\limits_{\mathcal{A}} f_{X|\Theta}(x|\tilde{\theta}) f_\Theta(\tilde{\theta}) d\tilde{\theta}}, \quad \forall x \in \mathbb{X} \tag{4.13}$$

Además, podemos notar que $f_X(x)$ es constante para cualquier elección de $y \in \mathcal{A}$, entonces la regla en (4.13) se puede escribir como:

$$\phi^*(x) = \arg\min_{y \in \mathcal{A}} \int_{\mathcal{A}} L(\theta, y) f_{X|\Theta}(x|\theta) f_\Theta(\theta) d\theta, \quad \forall x \in \mathbb{X} \tag{4.14}$$

La expresión (4.14) tiene la ventaja de ser general, pero a su vez difícil de manejar, veremos una función de costo particular que reduce el problema significativamente.

Observaciones 4.2. Notar que el problema de caracterizar $\phi^*(x)$ en (2.15) equivale a encontrar la constante $y \in \mathcal{A}$, que mejor estima Θ, cuando ésta sigue la distribución:

$$\Theta \sim P_{\Theta|X}(\cdot|x). \tag{4.15}$$

En otras palabras, $\phi^*(x) = y^*$ es el centroide óptimo o la constante que minimiza el riesgo de estimar Θ condicionado a $X = x$.

En lo que sigue consideraremos distintas funciones de costos muy usadas en estimación Bayesiana las cuales nos entregarán estimadores con expresiones cerradas.

4.3.1. Costo de Tipo Cuadrático

En este escenario consideramos $\mathcal{A} = \mathbb{R}$ y estamos interesados en el error cuadrático medio, es decir, la función de costo es $L(\theta_0, \theta_1) = (\theta_0 - \theta_1)^2$. En este caso ϕ^* se conoce como el estimador de mínimo error cuadrático medio o Minimum Mean Square Error (MMSE) estimator. La ecuación (4.10) se reduce a, tomando $x \in \mathbb{X}$:

$$\phi^*(x) = \arg\min_{y \in \mathcal{A}} \int_{\mathcal{A}} (\theta - y)^2 f_{\Theta|X}(\theta|x) d\theta \tag{4.16}$$

Consideremos el siguiente operador:

$$\mathbb{E}(\Theta|X = x) = \int_{\mathcal{A}} \theta f_{\Theta|X}(\theta|x) d\theta \tag{4.17}$$

que corresponde a la esperanza condicional de Θ dado $X = x$, entonces, el argumento en (4.16) lo podemos descomponer como:

$$\begin{aligned}
\int_{\mathcal{A}} (\theta - y)^2 f_{\Theta|X}(\theta|x) d\theta &= \int_{\mathcal{A}} (\theta - \mathbb{E}(\Theta|X = x) + \mathbb{E}(\Theta|X = x) - y)^2 f_{\Theta|X}(\theta|x) d\theta \\
&= \int_{\mathcal{A}} (\theta - \mathbb{E}(\Theta|X = x))^2 f_{\Theta|X}(\theta|x) d\theta + \int_{\mathcal{A}} (\mathbb{E}(\Theta|X = x) - y)^2 f_{\Theta|X}(\theta|x) d\theta \\
&\quad + 2(\mathbb{E}(\Theta|X = x) - y) \int_{\mathcal{A}} (\theta - \mathbb{E}(\Theta|X = x)) f_{\Theta|X}(\theta|x) d\theta \quad \rightarrow 0 \\
&= \int_{\mathcal{A}} (\theta - \mathbb{E}(\Theta|X = x))^2 f_{\Theta|X}(\theta|x) d\theta + (\mathbb{E}(\Theta|X = x) - y)^2
\end{aligned} \tag{4.18}$$

Podemos notar que

$$\int_{\mathcal{A}} (\theta - \mathbb{E}(\Theta|X = x))^2 f_{\Theta|X}(\theta|x) d\theta = Var(\Theta|X = x) \tag{4.19}$$

es la varianza condicional de Θ dado $X = x$. Por lo tanto:

$$\begin{aligned}
\phi^*(x) &= \arg\min_{y \in \mathcal{A}} Var(\Theta|X = x) + (\mathbb{E}(\Theta|X = x) - y)^2 \\
&= \arg\min_{y \in \mathcal{A}} (\mathbb{E}(\Theta|X = x) - y)^2 \\
&= \mathbb{E}(\Theta|X = x).
\end{aligned} \tag{4.20}$$

La última igualdad es evidente a partir del hecho que si tomamos $y = \mathbb{E}(\Theta|X = x)$ la función $(\mathbb{E}(\Theta|X = x) - y)^2$ es mínima tomando valor 0.

Observaciones 4.3. El estimador óptimo que minimiza el error cuadrático medio corresponde a:

$$\phi_{MMSE}(x) = \mathbb{E}(\Theta|X = x) = \int_{\mathcal{A}} \theta f_{\Theta|X}(\theta|x)d\theta, \tag{4.21}$$

que es la esperanza condicional o la esperanza de la distribución a posteriori de Θ dado $X = x$.

Finalmente el riesgo Bayesiano mínimo o error cuadrático medio mínimo (MMSE) está dado por la siguiente expresión

$$\begin{aligned} MMSE &= \min_{\phi:\mathbb{X}\to\mathcal{A}} \mathbb{E}_{\Theta,X}(L(\Theta, \phi(X))) \\ &= \min_{\phi:\mathbb{X}\to\mathcal{A}} \mathbb{E}_{\Theta,X}((\Theta - \phi(X))^2) \\ &= \int_{\mathbb{X}} \left[\int_{\mathcal{A}} (\theta - \mathbb{E}(\Theta|X = x))^2 f_{\Theta|X}(\theta|x)d\theta \right] f_X(x)dx \\ &= \int_{\mathbb{X}} Var(\Theta|X = x) f_X(x)dx \\ &= \mathbb{E}(Var(\Theta|X)), \end{aligned} \tag{4.22}$$

que corresponde al promedio de la varianza condicional.

4.3.2. Costo Tipo Uniforme

En este escenario consideremos $\mathcal{A} = \mathbb{R}$, estamos interesados en el error uniforme definido como:

$$L(\theta, y) = \begin{cases} 1 & \text{si } |\theta - y| > \frac{\Delta}{2} \\ 0 & \text{si } |\theta - y| \leq \frac{\Delta}{2} \end{cases} \tag{4.23}$$

$\forall \Delta > 0$. En este caso la ecuación (4.10) se reduce a, tomando $x \in \mathbb{X}$:

$$\begin{aligned} \phi^*_{MAP}(x) \arg\min_{y\in\mathcal{A}} & \int_{\mathcal{A}} L(\theta, y) f_{\Theta|X}(\theta|x)d\theta \\ \arg\min_{y\in\mathcal{A}} & \int_{|\theta-y|>\frac{\Delta}{2}} f_{\Theta|X}(\theta|x)d\theta \\ = \arg\min_{y\in\mathcal{A}} & 1 - \int_{|\theta-y|\leq\frac{\Delta}{2}} f_{\Theta|X}(\theta|x)d\theta \end{aligned}$$

$$
\begin{aligned}
&= \arg\min_{y\in\mathcal{A}} 1 - \int_{y-\Delta/2}^{y+\Delta/2} f_{\Theta|X}(\theta|x)d\theta \\
&= \arg\max_{y\in\mathcal{A}} \int_{y-\Delta/2}^{y+\Delta/2} f_{\Theta|X}(\theta|x)d\theta \qquad (4.24)
\end{aligned}
$$

Notar que:

$$
\int_{y-\Delta/2}^{y+\Delta/2} f_{\Theta|X}(\theta|x)d\theta \leq \sup_{\theta\in\mathcal{A}} f_{\Theta|X}(\theta|x)\frac{\Delta}{2}. \qquad (4.25)
$$

Por lo que el máximo se encuentra eligiendo

$$
\phi^*_{MAP}(x) = \arg\max_{y\in\mathcal{A}} f_{\Theta|X}(y|x). \qquad (4.26)
$$

Podemos ver entonces que esta expresión corresponde a elegir el estimador que máximiza la distribución a posteriori (también conocida como regla MAP o maximum a posteriori), la misma regla que el escenario de detección Bayesiano asociada a la función de costo $0-1$.

Nuevamente nos encontramos con que las verosmilitudes son de tipo exponenciales, entonces aprovechándonos del crecimiento del logaritmo y de la regla de Bayes podemos escribir (4.26) como:

$$
\begin{aligned}
\phi^*_{MAP}(x) &= \arg\max_{y\in\mathcal{A}} \ln(f_{\theta|X}(y|x)) \\
&= \arg\max_{y\in\mathcal{A}} \ln\left(\frac{f_{X|\Theta}(x|y)f_\Theta(y)}{f_X(x)}\right) \\
&= \arg\max_{y\in\mathcal{A}} \ln\left(f_{X|\Theta}(x|y)f_\Theta(y)\right) \\
&= \arg\max_{y\in\mathcal{A}} \ln\left(f_{X|\Theta}(x|y)\right) + \ln\left(f_\Theta(y)\right) \qquad (4.27)
\end{aligned}
$$

El término $\ln\left(f_{X|\Theta}(x|y)\right)$ corresponde a la función de log-verosimilitud visto en estimación paramétrica. Por lo tanto el estimador MAP es una extensión del estimador de máxima verosimilitud en el caso Bayesiano donde ahora se considera la información adicional dada por la densidad de Θ. Notar entonces que (4.27) puede resolverse aplicando el criterio de primer orden siempre y cuando las densidades condicionales y marginales lo permitan.

Veremos un ejemplo concreto donde aplicaremos este estimador.

Ejemplo 4.1. Sea $\mathcal{A} = \mathbb{R}^+ \cup \{0\}$, supongamos que tenemos la secuencia de variables de observaciones $X_i = \Theta + W_i$, $i = 1, ..., n$. El parámetro aleatorio Θ es desconocido y sigue la siguiente función de densidad de probabilidad a priori

$$f_\Theta(\theta) = \lambda e^{-\lambda\theta}$$

donde $\lambda > 0$, es decir, $\Theta \sim exponencial(\lambda)$ y $W_i \sim N(0, \sigma^2)$ e independiente de Θ. Vamos a encontrar el estimador MAP, para esto entonces debemos resolver, dado $X_1^n = x_1^n$

$$\phi^*_{MAP}(x_1^n) = \arg \max_{y \in \mathbb{R}^+ \cup \{0\}} \ln\left(f_{X_1^n|\Theta}(x_1, ..., x_n|y)\right) + \ln\left(f_\Theta(y)\right). \tag{4.28}$$

Podemos ver que, dado $\Theta = y$, entonces $X_i \sim N(y, \sigma^2)$ ya que estamos sumando un valor constante con una variable normal de media 0, luego:

$$f_{X_1^n|\Theta}(x_1, ..., x_n|y) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_i-y)^2}{2\sigma^2}} \tag{4.29}$$

y con esto,

$$\begin{aligned}\phi^*_{MAP}(x_1^n) &= \arg \max_{y \in \mathbb{R}^+ \cup \{0\}} \ln\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_i-y)^2}{2\sigma^2}}\right) + \ln\left(\lambda e^{-\lambda y}\right) \\ &= \arg \max_{y \in \mathbb{R}^+ \cup \{0\}} -n\ln\left(\sqrt{2\pi}\sigma\right) - \sum_{i=1}^{n} \frac{(x_i - y)^2}{2\sigma^2} - \lambda y + \ln(\lambda). \end{aligned} \tag{4.30}$$

Aplicando el criterio de primer orden respecto a y tenemos que:

$$\frac{\partial}{\partial y}\left(-n\ln\left(\sqrt{2\pi}\sigma\right) - \sum_{i=1}^{n} \frac{(x_i - y)^2}{2\sigma^2} - \lambda y + \ln(\lambda)\right) = \sum_{i=1}^{n} \frac{(x_i - y)}{\sigma^2} - \lambda \tag{4.31}$$

Despejando y, obtenemos:

$$\begin{aligned} \sum_{i=1}^{n} \frac{(x_i - y)}{\sigma^2} - \lambda &= 0 \\ \sum_{i=1}^{n} x_i - yn - \lambda\sigma^2 &= 0 \\ y &= \frac{\sum_{i=1}^{n} x_i - \lambda\sigma^2}{n}, \end{aligned} \tag{4.32}$$

Por lo tanto,

$$\phi^*_{MAP}(x_1^n) = \begin{cases} \frac{\sum_{i=1}^{n} x_i - \lambda\sigma^2}{n}. & \text{si } \sum_{i=1}^{n} x_i - \lambda\sigma^2 \geq 0 \\ 0 & \sim \end{cases} \tag{4.33}$$

Notar que la condición $\sum_{i=1}^{n} x_i - \lambda\sigma^2 \geq 0$ es para pedir que no se salga del rango de Θ, en caso contrario se tendría que la derivada $\sum_{i=1}^{n} x_i - yn - \lambda\sigma^2 < 0$, es decir, negativa en todo punto $y \in \mathbb{R}^+ \cup \{0\}$, luego, la función es decreciente estricta y el máximo se obtiene tomando el menor valor posible para y, concluyendo que $y = 0$.

4.3.3. Costo Tipo Absoluto

En este escenario consideremos nuevamente $\mathcal{A} = \mathbb{R}$, estamos interesados en el error absoluto de la forma, dado $y, \theta \in \mathcal{A}$, $L(\theta, y) = |\theta - y|$. En este caso la ecuación (4.10) se reduce a, tomando $x \in \mathbb{X}$:

$$\phi^*_{abs}(x) = \arg\min_{y \in \mathcal{A}} \int_{-\infty}^{\infty} |\theta - y| f_{\Theta|X}(\theta|x) d\theta \tag{4.34}$$

Notar que el argumento en (4.34) se puede descomponer de la siguiente manera:

$$\int_{-\infty}^{\infty} |\theta - y| f_{\Theta|X}(\theta|x) d\theta = \int_{-\infty}^{\phi^*_{abs}(x)} (y - \theta) f_{\Theta|X}(\theta|x) d\theta + \int_{\phi^*_{abs}(x)}^{\infty} (\theta - y) f_{\Theta|X}(\theta|x) d\theta \tag{4.35}$$

Por lo que al establecer las condiciones de primer orden se obtiene que

$$\int_{-\infty}^{\phi^*_{abs}(x)} f_{\Theta|X}(\theta|x) d\theta = \int_{\phi^*_{abs}(x)}^{\infty} f_{\Theta|X}(\theta|x) d\theta \tag{4.36}$$

Este resultado nos indica que el estimador óptimo corresponde a la mediana de la densidad de probabilidad a posteriori.

4.4. Ortogonalidad y Estimación de Mínimos Cuadrados

En esta sección se verá la esperanza condicional desde una óptica de distancia. Ya vimos que, bajo un costo cuadrático corresponde al estimador óptimo Bayesiano. Sin

embargo, este estimador sigue siendo óptimo incluso en un escenario multiparamétrico donde $\mathcal{A} \subset \mathbb{R}^k$, $k \in \mathbb{N}$. En lo que sigue hablaremos brevemente de los espacios de Hilbert lo que nos dará un fondo conceptual para definir el estimador óptimo cuando la función de costo es la cuadrática en mayores dimensiones

Supongamos que las variables aleatorias X, Θ poseen segundo momento finito ($\mathbb{E}(X^2) < \infty$). Definamos la siguiente expresión:

$$\langle X, \Theta \rangle = \mathbb{E}(X\Theta).$$

Vemos que este operador define un producto interno y se verifican las siguientes propiedades para X, Θ, Y y $\alpha, \beta \in \mathbb{R}$:

- $\langle X, X \rangle \geq 0$.
- $\langle \alpha X + \beta Z, \Theta \rangle = \alpha \langle X, \Theta \rangle + \beta \langle Z, \Theta \rangle$.

Como lo anterior define un producto interno podemos hablar de una norma (un distancia en términos simples) dada por:

$$||X||^2 = \langle X, X \rangle.$$

Al ser norma, cumple con todos los criterios para ser tal (positividad, homogeneidad y desigualdad triangular). Un espacio vectorial $\mathcal{H}$ con producto interno y cerrado se llama espacio de Hilbert. Similarmente podemos definir el producto interno vectorial (es decir, para pares de vectores aleatorios X_1^n, Θ_1^n) como:

$$\langle X_1^n, \Theta_1^n \rangle = \sum_{i=1}^{n} \mathbb{E}(X_i \Theta_i) = \mathbb{E}\left((\Theta_1^n)^T X_1^n\right)$$

así como su norma:

$$||X_1^n||^2 = \langle X_1^n, X_1^n \rangle.$$

Lo que también corresponde a un espacio de Hilbert.

En lo que sigue hablaremos de un espacio de Hilbert que sigue la siguiente estructura:

$$\mathcal{H} = \{Z_1^m : \Omega \to \mathbb{R}^m | \mathbb{E}\left((Z_1^m)^T Z_1^m\right) < \infty\}.$$

Es decir, este espacio no es más que vectores aleatorios con producto interno finito. Dado $X_1^n \in \mathbb{R}^n$, consideremos el siguiente subespacio de Hilbert:

$$\mathcal{G} = \{g(X_1^n) : \mathbb{R}^n \to \mathbb{R}^m | \mathbb{E}\left((g(X_1^n))^T g(X_1^n)\right) < \infty\}$$

El espacio $\mathcal{G}$ es un espacio más pequeño contenido en $\mathcal{H}$ ya que está limitado a funciones de un vector X_1^n previamente fijado.

Dado $\Theta_1^m \in \mathcal{H}$ y $X_1^n \in \mathbb{R}^n$ nos gustaría encontrar la función $g(X_1^n) : \mathbb{R}^n \to \mathbb{R}^m$ solución de

$$g^*(X_1^n) = \min_{g:\mathbb{R}^n \to \mathbb{R}^m} ||g(X_1^n) - \Theta_1^m||^2 = \min_{g:\mathbb{R}^n \to \mathbb{R}^m} \sum_{i=1}^{m} \mathbb{E}((g_i(X_1^n) - \Theta_i)^2) \tag{4.37}$$

donde $g_i(X_1^n)$ corresponde a la componente i-ésima de la función g. La ecuación (4.37) corresponde al error cuadrático medio (MSE en inglés). Una propiedad interesante y central en esta sección es que la solución al problema de la ecuación (4.37) es la esperanza condicional de Θ_1^m dado X_1^n, es decir,

$$g^*(X_1^n) = \mathbb{E}(\Theta_1^m | X_1^n).$$

Por lo tanto, el problema recién planteado se conoce como un problema de proyección: la función óptima $g^*(X_1^n)$ corresponde a la **proyección ortogonal** de Y_1^m sobre el subespacio de funciones medibles con respecto a X_1^n, denotado por $\mathcal{G}$. Esta proyección se define en el espacio de variables aleatorias de varianza finita, dotado del producto interno inducido por la esperanza. Bajo esta estructura, la esperanza condicional $\mathbb{E}(\Theta_1^m \mid X_1^n)$ no solo minimiza el error cuadrático medio, sino que además es aquella que cumple la condición de ortogonalidad. Esta interpretación geométrica refuerza el papel de la esperanza condicional como la mejor aproximación de Y_1^m a partir de la información contenida en X_1^n. Para demostrar tal resultado, primero enunciaremos un resultado equivalente al del problema de minimización de error cuadrático medio planteado anteriormente usando herramientas de álgebra lineal. Más precisamente, sea $y \in \mathcal{H}$, un espacio vectorial $\mathcal{H}$, $\mathcal{G}$ un subespacio y sea f la proyección de y en $\mathcal{G}$ entonces $\forall z \in \mathcal{G}$.

$$\langle y - f, z \rangle = 0.$$

Dicho resultado se puede ver ilustrativamente en la Figura 4.1. Luego para nuestro problema, bastaría demostrar que $\forall g(X_1^n) \in \mathcal{G}$.

$$\langle \Theta_1^m - \mathbb{E}(\Theta_1^m | X_1^n), g(X_1^n) \rangle = 0.$$

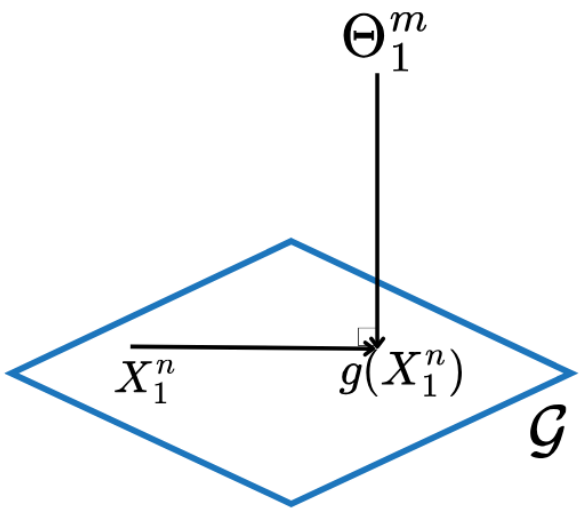


Figura 4.1: Proyección ortogonal de Θ_1^m sobre un subespacio $\mathcal{G}$. El residuo $\Theta_1^m - \mathbb{E}(\Theta_1^m|X_1^n)$ es ortogonal a todo elemento en $\mathcal{G}$.

Realizando el cálculo tenemos que:

$$\begin{aligned}
\langle \Theta_1^m - \mathbb{E}(\Theta_1^m|X_1^n), g(X_1^n)\rangle &= \mathbb{E}(g(X_1^n)^T(\Theta_1^m - \mathbb{E}(\Theta_1^m|X_1^n))) \\
&= \mathbb{E}(\mathbb{E}(g(X_1^n)^T(\Theta_1^m - \mathbb{E}(\Theta_1^m|X_1^n))|X_1^n)) \\
&= \mathbb{E}(g(X_1^n)^T\mathbb{E}((\Theta_1^m - \mathbb{E}(\Theta_1^m|X_1^n))|X_1^n)) \\
&= \mathbb{E}(g(X_1^n)^T(\mathbb{E}(\Theta_1^m|X_1^n) - \mathbb{E}(\Theta_1^m|X_1^n))) \\
&= 0.
\end{aligned}$$

Observaciones 4.4.

- La esperanza condicional se interpreta como la solución al problema de mínima distancia (proyección ortogonal).
- Suponiendo que las funciones de X son una superficie e Θ un punto fuera de esta superficie, la esperanza condicional es el punto en la superficie más cercano a Θ. Gracias a esta interpretación geométrica es más intuitivo entender el por qué $\mathbb{E}(g(X)|X) = g(X)$ debido a que no es necesario proyectar.

Una de las grandes desventajas de la esperanza condicional es la complejidad para poder calcular dicha expresión la que por lo general involucran integrales de espacios m-dimensionales. Debido a lo anterior, se suelen proponer soluciones alternativas, restringiendo el espacio de búsqueda de funciones permitidas. Esto se conoce como **regresión**.

Veremos el caso de regresión lineal y entregaremos el resultado de la función óptima para este problema.

Consideremos el problema de estimar Θ_1^m dado X_1^n pero sobre transformaciones lineales de X_1^n, es decir, tenemos el siguiente conjunto:

$$\mathcal{M} = \{g(X_1^n) = HX_1^n + \beta_1^m | \mathbb{E}\left((g(X_1^n))^T g(X_1^n)\right) < \infty, H \in M_{m\times n}, \beta_1^m \in \mathbb{R}^m\}$$

Nos gustaría encontrar la función óptima que minimice el erro cuadrático medio pero condicionado a que las funciones solamente pueden ser transformaciones lineales de X_1^n. Esto se reduce a encontrar la matriz óptima H^* y el vector óptimo $(\beta_1^m)^*$ que minimiza el error cuadrático medio. La condición de ortogonalidad nos pide que para cualquier función $g(X_1^n) \in \mathcal{M}$ se cumpla que:

$$\langle \Theta_1^m - (H^* X_1^n + (\beta_1^m)^*), g(X_1^n)\rangle = 0 \tag{4.38}$$

Dado que se debe cumplir para cualquier g, eligiendo convenientemente $g(X_1^n) = e_i$ (vector canónico), tenemos que:

$$\langle \Theta_1^m - H^* X_1^n - (\beta_1^m)^*, e_i\rangle = 0 \Rightarrow (\beta_i)^* = \mathbb{E}(\Theta_i) - H^*\mathbb{E}(X_i) \tag{4.39}$$

Similarmente eligiendo $g(X_1^n) = GX_1^n$, donde G es una matriz de $m \times n$ tenemos que:

$$\begin{aligned}
\langle \Theta_1^m - H^* X_1^n - (\beta_1^m)^*, GX_1^n\rangle &= \mathbb{E}(tr((\Theta_1^m - H^* X_1^n - (\beta_1^m)^*) \cdot (GX_1^n)^T)) \\
&= tr(\mathbb{E}((\Theta_1^m - H^* X_1^n - (\beta_1^m)^*) \cdot (GX_1^n)^T)) \\
&= tr(\mathbb{E}((\Theta_1^m - H^* X_1^n - (\beta_1^m)^*) \cdot (X_1^n)^T G^T)) \\
&= 0.
\end{aligned}$$

Donde tr corresponde a la traza de una matriz. Como la condición anterior se debe cumplir para cualquier G, luego necesariamente $\mathbb{E}((\Theta_1^m - H^* X_1^n - (\beta_1^m)^*) \cdot (X_1^n)^T) = 0$. Pero como

$\mathbb{E}((\Theta_1^m - H^* X_1^n - (\beta_1^m)^*)) = 0$, se tiene que:

$$\begin{aligned} 0 = \mathbb{E}((\Theta_1^m - H^* X_1^n - (\beta_1^m)^*) \cdot (X_1^n)^T) &= Cov((\Theta_1^m - H^* X_1^n - (\beta_1^m)^*), X_1^n) \\ &= Cov(\Theta_1^m, X_1^n) - H^* Cov(X_1^n, X_1^n) \\ H^* &= Cov(\Theta_1^m, X_1^n) Cov(X_1^n)^{-1}. \end{aligned}$$

Con lo que finalmente el estimador lineal óptimo es:

$$g^*(X_1^n) = \mathbb{E}(\Theta_1^n) + Cov(\Theta_1^m, X_1^n) Cov(X_1^n)^{-1}(X_1^n - \mathbb{E}(X_1^n))$$

La solución anterior se conoce como la fórmula de regresión lineal la cual conecta directamente con la sección 3.6, más precisamente (3.153). La diferencia sutil es que en el estimador de mínimos cuadrados estamos interesados en los parámetros que permiten relacionar X_1^n con Y_1^n mediante alguna función lineal $Y_i = \alpha + \beta X_i$. Es decir, buscamos los parámetros que inducen la función lineal que relacionan X_1^n con Y_1^n. En este caso, en cambio, queremos la regla $\phi(X_1^n)$ que se acerque a Θ_1^m minimizando su error cuadrático. Vemos entonces que el estimador de mínimos cuadrados es una instancia del estimador Bayesiano ya que este último contiene los parámetros buscados en el caso paramétrico.

4.5. Caso de Estudio: Distribución Conjunta Normal Multivariada

En este ejemplo analizaremos la extensión de la distribución normal multivariada a densidades condicionales, el gran resultado es que si se posee un vector aleatorio de distribución normal, entonces la distribución condicional también será normal.

Sea $\overline{X}$ e Y vectores aleatorios[1] con valores en $\mathbb{R}^n$ y $\mathbb{R}^m$ respectivamente, con distribución normal multivariada y parámetros

$$\begin{aligned} N(\mu_{\overline{X}}, \Sigma_{\overline{X}}) \Rightarrow \mu_{\overline{X}} = \mathbb{E}(\overline{X}) \qquad & \Sigma_{\overline{X}} = \mathbb{E}((\overline{X} - \mu_{\overline{X}})(\overline{X} - \mu_{\overline{X}})^t) \\ N(\mu_{\overline{Y}}, \Sigma_{\overline{Y}}) \Rightarrow \mu_{\overline{Y}} = \mathbb{E}(\overline{Y}) \qquad & \Sigma_{\overline{Y}} = \mathbb{E}((\overline{Y} - \mu_{\overline{Y}})(\overline{Y} - \mu_{\overline{Y}})^t) \end{aligned} \tag{4.40}$$

Adicionalmente consideremos la concatenación $Z = \begin{pmatrix} \overline{X} \\ \overline{Y} \end{pmatrix}$ con valores en $\mathbb{R}^{n+m}$ cuya distribución conjunta es Normal y además:

$$\mu_{\overline{Z}} = \mathbb{E}(\overline{Z}) = \begin{pmatrix} \mu_{\overline{X}} \\ \mu_{\overline{Y}} \end{pmatrix} \tag{4.41}$$

[1] Para evitar sobercargar la notación, utilizaremos por esta vez $\overline{X}$ e $\overline{Y}$ en vez de X_1^n e Y_1^m, respectivamente. Por lo tanto se utilizarán operaciones matriciales.

y

$$\begin{aligned}\Sigma_{\overline{Z}} &= \mathbb{E}((\overline{Z}-\mu_{\overline{Z}})(\overline{Z}-\mu_{\overline{Z}})^t) \\ &= \begin{pmatrix}\Sigma_{\overline{X}} & \Sigma_{\overline{XY}} \\ \Sigma_{\overline{YX}} & \Sigma_{\overline{Y}}\end{pmatrix}\end{aligned} \tag{4.42}$$

Donde

$$\Sigma_{\overline{YX}} = \mathbb{E}((\overline{Y}-\mu_{\overline{Y}})(\overline{X}-\mu_{\overline{X}})^t) \in M_{m\times n} \tag{4.43}$$

$$\Sigma_{\overline{XY}} = \mathbb{E}((\overline{X}-\mu_{\overline{X}})(\overline{Y}-\mu_{\overline{Y}})^t) \in M_{n\times m} \tag{4.44}$$

Entonces se tiene que la distribución de $\overline{Y}$ dado $\overline{X}=\overline{x}$ es Gaussiana de parámetros:

$$\mu_{\overline{Y}|\overline{X}}(\overline{x}) = \mathbb{E}(\overline{Y}|\overline{X}=\overline{x}) = \mu_{\overline{Y}} + \Sigma_{\overline{YX}}\Sigma_{\overline{X}}^{-1}(\overline{x}-\mu_{\overline{X}}) \tag{4.45}$$

y la covarianza es:

$$\Sigma_{\overline{Y}|\overline{X}} = \mathbb{E}((\overline{Y}-\mu_{\overline{Y}|\overline{X}}(\overline{X}))(\overline{Y}-\mu_{\overline{Y}|\overline{X}}(\overline{X}))^t) = \Sigma_{\overline{Y}} - \Sigma_{\overline{YX}}\cdot\Sigma_{\overline{X}}^{-1}\cdot\Sigma_{\overline{XY}} \tag{4.46}$$

Observaciones 4.5.

1- Notar que un corolario de este resultado es que el estimador MMSE en (4.45) de $\overline{Y}$ dado $\overline{X}$ coincide con el estimador lineal de mínimos cuadrados en (4.4) un resultado sorprendente y muy útil en la comunidad de señales.
2- El error cuadrático medio o rieso Bayesiano tomando el costo cuadrático está dado por la expresión (4.46)

Demostración: Podemos utilizar la definición de distribución normal multivariada y probabilidad condicional para caracterizar la densidad condicional $\overline{Y}$ dado X, así tenemos que para $\overline{x},\overline{y}\in\mathbb{R}^n$:

$$\begin{aligned}f_{\overline{Y}|\overline{X}}(\overline{y}|\overline{x}) &= \frac{f_{\overline{Z}}(\overline{x},\overline{y})}{f_X(\overline{x})} \\ &= \frac{1}{\left[(2\pi)^m\left|\Sigma_{\overline{Z}}\right|\right]^{1/2}\left|\Sigma_{\overline{X}}\right|^{-1/2}} \\ &\cdot\exp\left(-\frac{1}{2}\left[\begin{pmatrix}\overline{x}-\mu_{\overline{X}}\\ \overline{y}-\mu_{\overline{Y}}\end{pmatrix}^t\Sigma_{\overline{Z}}^{-1}\begin{pmatrix}\overline{x}-\mu_{\overline{X}}\\ \overline{y}-\mu_{\overline{Y}}\end{pmatrix} - (\overline{x}-\mu_{\overline{X}})^t\Sigma_{\overline{X}}^{-1}(\overline{x}-\mu_{\overline{X}})\right]\right)\end{aligned} \tag{4.47}$$

Vamos a utilizar el siguiente resultado válido para inversas de bloques de matrices:

$$\Sigma_{\overline{Z}}^{-1} = \begin{pmatrix} \Sigma_{\overline{X}} & \Sigma_{\overline{XY}} \\ \Sigma_{\overline{YX}} & \Sigma_{\overline{Y}} \end{pmatrix}^{-1} = \begin{pmatrix} A & B \\ B^t & C \end{pmatrix} \tag{4.48}$$

donde:

- $C \triangleq \Sigma_{\overline{Y}|\overline{X}}^{-1} = (\Sigma_{\overline{Y}} - \Sigma_{\overline{YX}}\Sigma_{\overline{X}}^{-1}\Sigma_{\overline{XY}})^{-1}$.
- $A \triangleq \Sigma_{\overline{X}}^{-1} + \Sigma_{\overline{X}}^{-1}\Sigma_{\overline{XY}}\Sigma_{\overline{Y}|\overline{X}}^{-1}\Sigma_{\overline{YX}}\Sigma_{\overline{X}}^{-1}$.
- $B \triangleq -\Sigma_{\overline{X}}^{-1}\Sigma_{\overline{XY}}\Sigma_{\overline{Y}|\overline{X}}^{-1}$

Si aplicamos este resultado en (4.47) se tiene que

$$\begin{aligned}
f_{\overline{Y}|\overline{X}}(\overline{y}|\overline{x}) &= \frac{1}{\left[(2\pi)^m \left|\Sigma_{\overline{Z}}\right|\right]^{1/2} \left|\Sigma_{\overline{X}}\right|^{-1/2}} \\
&\cdot \exp\left(-\frac{1}{2}\left[(\overline{x}-\mu_{\overline{X}})^t A(\overline{x}-\mu_{\overline{X}}) + 2(\overline{x}-\mu_{\overline{X}})^t B(\overline{y}-\mu_{\overline{Y}})\right.\right. \\
&\left.\left.+(\overline{y}-\mu_{\overline{Y}})^t C(\overline{y}-\mu_{\overline{Y}}) - (\overline{x}-\mu_{\overline{X}})^t\Sigma_{\overline{X}}^{-1}(\overline{x}-\mu_{\overline{X}})\right]\right) \\
&= \frac{1}{\left[(2\pi)^m \left|\Sigma_{\overline{Z}}\right|\right]^{1/2} \left|\Sigma_{\overline{X}}\right|^{-1/2}} \\
&\cdot \exp\left(-\frac{1}{2}\left[(\overline{x}-\mu_{\overline{X}})^t(\Sigma_{\overline{X}}^{-1} + \Sigma_{\overline{X}}^{-1}\Sigma_{\overline{XY}}\Sigma_{\overline{Y}|\overline{X}}^{-1}\Sigma_{\overline{YX}}\Sigma_{\overline{X}}^{-1})(\overline{x}-\mu_{\overline{X}})\right.\right. \\
&- 2(\overline{x}-\mu_{\overline{X}})^t\Sigma_{\overline{X}}^{-1}\Sigma_{\overline{XY}}\Sigma_{\overline{Y}|\overline{X}}^{-1}(\overline{y}-\mu_{\overline{Y}}) + (\overline{y}-\mu_{\overline{Y}})^t\Sigma_{\overline{Y}|\overline{X}}^{-1}(\overline{y}-\mu_{\overline{Y}}) \\
&\left.\left.-(\overline{x}-\mu_{\overline{X}})^t\Sigma_{\overline{X}}^{-1}(\overline{x}-\mu_{\overline{X}})\right]\right) \\
&= \frac{1}{\left[(2\pi)^m \left|\Sigma_{\overline{Z}}\right|\right]^{1/2} \left|\Sigma_{\overline{X}}\right|^{-1/2}} \\
&\cdot \exp\left(-\frac{1}{2}\left[(\overline{x}-\mu_{\overline{X}})^t\Sigma_{\overline{X}}^{-1}\Sigma_{\overline{XY}}\Sigma_{\overline{Y}|\overline{X}}^{-1}\Sigma_{\overline{YX}}\Sigma_{\overline{X}}^{-1}(\overline{x}-\mu_{\overline{X}})\right.\right. \\
&\left.\left.-2(\overline{x}-\mu_{\overline{X}})^t\Sigma_{\overline{X}}^{-1}\Sigma_{\overline{XY}}\Sigma_{\overline{Y}|\overline{X}}^{-1}(\overline{y}-\mu_{\overline{Y}}) + (\overline{y}-\mu_{\overline{Y}})^t\Sigma_{\overline{Y}|\overline{X}}^{-1}(\overline{y}-\mu_{\overline{Y}})\right]\right)
\end{aligned} \tag{4.49}$$

Al ver la estructura dentro de la exponencial identificamos una forma cuadrática. Llamemos

provisoriamente $P = \Sigma_{\overline{YX}}\Sigma_{\overline{X}}^{-1}(\overline{x} - \mu_{\overline{X}})$ y $T = \overline{y} - \mu_{\overline{Y}}$, tenemos que:

$$\begin{aligned} f_{\overline{Y}|\overline{X}}(\overline{y}|\overline{x}) &= \frac{1}{\left[(2\pi)^m \left|\Sigma_{\overline{Z}}\right|\right]^{1/2} \left|\Sigma_{\overline{X}}\right|^{-1/2}} \exp\left(-\frac{1}{2}\left[P^t\Sigma_{\overline{Y}|\overline{X}}^{-1}P - 2P^t\Sigma_{\overline{Y}|\overline{X}}^{-1}T + T^t\Sigma_{\overline{Y}|\overline{X}}^{-1}T\right]\right) \\ &= \frac{1}{\left[(2\pi)^m \left|\Sigma_{\overline{Z}}\right|\right]^{1/2} \left|\Sigma_{\overline{X}}\right|^{-1/2}} \exp\left(-\frac{1}{2}(P-T)^t\Sigma_{\overline{Y}|\overline{X}}^{-1}(P-T)\right) \end{aligned} \tag{4.50}$$

Identificamos:

$$\begin{aligned} P - T &= \Sigma_{\overline{YX}}\Sigma_{\overline{X}}^{-1}(\overline{x} - \mu_{\overline{X}}) - \overline{y} + \mu_{\overline{Y}} \\ &= -(\overline{y} - (\mu_{\overline{Y}} + \Sigma_{\overline{YX}}\Sigma_{\overline{X}}^{-1}(\overline{x} - \mu_{\overline{X}}))) \\ &= -(\overline{y} - \mu_{\overline{Y}|\overline{X}}(\overline{x})), \end{aligned} \tag{4.51}$$

donde se definió $\mu_{\overline{Y}|\overline{X}}(\overline{x}) \triangleq \mu_{\overline{Y}} + \Sigma_{\overline{YX}}\Sigma_{\overline{X}}^{-1}(\overline{x} - \mu_{\overline{X}})$ con lo que:

$$f_{\overline{Y}|\overline{X}}(\overline{y}|\overline{x}) = \frac{1}{\left[(2\pi)^m \left|\Sigma_{\overline{Z}}\right|\right]^{1/2} \left|\Sigma_{\overline{X}}\right|^{-1/2}} \exp\left(-\frac{1}{2}(\overline{y} - \mu_{\overline{Y}|\overline{X}}(\overline{x}))^t\Sigma_{\overline{Y}|\overline{X}}^{-1}(\overline{y} - \mu_{\overline{Y}|\overline{X}}(\overline{x}))\right) \tag{4.52}$$

Ahora basta ver que, por propiedades de determinantes en bloques:

$$\left|\Sigma_{\overline{Z}}\right| = \left|\Sigma_{\overline{X}}\right| \left|\Sigma_{\overline{Y}} - \Sigma_{\overline{YX}}\Sigma_{\overline{X}}^{-1}\Sigma_{\overline{XY}}\right| = \left|\Sigma_{\overline{X}}\right| \left|\Sigma_{\overline{Y}|\overline{X}}\right|. \tag{4.53}$$

Deducimos entonces que:

$$\begin{aligned} f_{\overline{Y}|\overline{X}}(\overline{y}|\overline{x}) &= \frac{1}{\left[(2\pi)^m \left|\Sigma_{\overline{Z}}\right|\right]^{1/2} \left|\Sigma_{\overline{X}}\right|^{-1/2}} \exp\left(-\frac{1}{2}(\overline{y} - \mu_{\overline{Y}|\overline{X}}(\overline{x}))^t\Sigma_{\overline{Y}|\overline{X}}^{-1}(\overline{y} - \mu_{\overline{Y}|\overline{X}}(\overline{x}))\right) \\ &= \frac{1}{\left[(2\pi)^m \left|\Sigma_{\overline{Y}|\overline{X}}\right|\right]^{1/2}} \exp\left(-\frac{1}{2}(\overline{y} - \mu_{\overline{Y}|\overline{X}}(\overline{x}))^t\Sigma_{\overline{Y}|\overline{X}}^{-1}(\overline{y} - \mu_{\overline{Y}|\overline{X}}(\overline{x}))\right), \end{aligned} \tag{4.54}$$

por lo que encontramos una estructura de una distribución normal multivariada cuyos parámetros son:

$$\overline{Y}|\overline{X} \sim N(\mu_{\overline{Y}|\overline{X}}(\overline{X}), \Sigma_{\overline{Y}|\overline{X}}), \tag{4.55}$$

donde

$$\Sigma_{\overline{Y}|\overline{X}} \triangleq \Sigma_{\overline{Y}} - \Sigma_{\overline{YX}} \cdot \Sigma_{\overline{X}}^{-1}\Sigma_{\overline{XY}} \tag{4.56}$$

y

$$\mathbb{E}(\overline{Y}|\overline{X}) \triangleq \mu_{\overline{Y}|\overline{X}}(\overline{X}) = \mu_{\overline{Y}} + \Sigma_{\overline{XY}}\Sigma_{\overline{X}}^{-1}(\overline{X} - \mu_{\overline{X}}) \tag{4.57}$$

□

4.6. Problemas

Se presentan a continuación una sección de problemas relacionados con estimación Bayesiana.

Problema 4.1. Considere el problema de estimar $\theta \in \mathbb{R}$ dada una observación $X \in \mathbb{R}$, se sabe que la distribución condicional de Θ dado X está dotada de la siguiente función de densidad condicional definida como:

$$f_{\Theta|X}(\theta|x) = \begin{cases} e^{-(\theta-x)}, \text{ si } \theta > x \\ 0, \text{ si } x > \theta \end{cases}$$

Encuentre el estimador de mínimo error cuadrático medio y MAP.

Problema 4.2. Considere que $X_1^m \sim \mathcal{N}(\overline{\theta}, K_X)$ con valores en $\mathbb{R}^m$, donde $\overline{\theta}$ es el vector de media y K_X una matriz de covarianza (invertible). Además considere que conocemos K_X y queremos estimar $\overline{\theta}$ por medio de mediciones lineales indirectas dadas por el siguiente modelo:

$$Z_1^n = HX_1^m + N_1^n \tag{4.58}$$

donde H es una matriz de $n \times m$ y $N_1^n \sim \mathcal{N}(0, \sigma^2 \cdot I_{n\times n})$, y N_1^n es independiente de X_1^n (este modelo se conoce como Canal Lineal más Ruido Aditivo Gaussiano).

a) Verifique que $Z_1^n \sim \mathcal{N}(\mu_Z, K_Z)$ y determine específicamente su vector de media y matriz de covarianza como función de σ^2,H, K_X y $\overline{\theta}$.

b) Asuma que $n > m$ y que H es de rango completo. Adicionalmente considere K_X, H y σ^2 conocidos. Con esto determine una expresión para el estimador de máxima verosimilitud dada una observación de Z_1^n.
Indicación: Reduzca el problema a un problema tipo mínimos cuadrados y con ello utilice la expresión cerrada que da solución a ese criterio.

c) En el mismo escenario del modelo aditivo presentado en (4.58), considere en cambio el problema de estimar X_1^m (una variable de estado) como función de una observación de Z_1^n dados todos los parámetros $\overline{\theta}$, K_X, H y σ^2 conocidos. Para ello primero determine una expresión para la matriz de covarianza cruzada:

$$K_{XZ} = \mathbb{E}\left((X_1^m - \overline{\theta}) \cdot (Z_1^n - \mu_Z)^t\right).$$

d) Finalmente obtenga una expresión para el estimador de mínimo error cuadrático medio (MMSE) de X_1^m dada una observación Z_1^n.

Problema 4.3. La secuencia X_i, $i = 1, ..., n$ es observada y sigue la siguiente función de densidad de probabilidad condicionada a $\Theta = \mu$

$$f_{X|\Theta}(x|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Las observaciones son independientes cuando son condicionadas por μ. La media μ sigue una distribución a priori

$$\mu \sim N(\mu_0, \sigma_0^2).$$

Encuentre el estimador de mínimo error cuadrático medio y MAP para μ.

5

Unidad V: Tópicos en Procesamiento de Información

Vimos que en las unidades anteriores los problemas de detección y estimación provienen de un contexto más amplio derivado de la Teoría de Juegos, donde analizamos una instancia particular de esta teoría.

En esta unidad veremos algunos tópicos que no encajan directamente con la teoría vista anteriormente, sin embargo, es una instancia de procesamiento de información que facilita la toma de decisiones bajo contexto de incertidumbre. Los temas que se cubrirán en estas notas son herramientas muy usadas en procesamiento de señales.

5.1. Test de Hipótesis Compuesto

En la Unidad I vimos el problema de detección paramétrica la decisión coincide con el parámetro a decidir. Los test de hipótesis pueden ser extendidos a más de dos parámetros en cuyo caso el test pasa a llamarse M-ario si es que $|\Theta| = M$, $M \in \mathbb{N}$. Cuando la decisión (ya sea en test de hipótesis binario o M-ario) es de cardinal 1, se conoce como test de hipótesis simple. Supongamos ahora el siguiente espacio de parámetros $\Theta = \mathbb{R} = \{\theta_0\} \cup \mathbb{R}\backslash\{\theta_0\}$ es decir, el espacio de parámetros se puede dividir en

dos conjuntos $\{\theta_0\}$ y la otra decisión sería su complemento. Si el espacio de parámetros $\Theta = \Theta_0 \cup \Theta_1 \cup ... \cup \Theta_{M-1}$ con $\{\Theta_0, \Theta_1, ..., \Theta_{M-1}\}$ partición de Θ es tal que al menos uno de estos conjuntos Θ_0 es de cardinal mayor que 1, el test de hipótesis pasará a llamarse compuesto.

Los siguientes ejemplos son formulaciones de test de hipótesis compuestos:

$$\begin{aligned} \Theta =& \mathbb{R} \\ H_0 :& \theta = \theta_0 \\ H_1 :& \theta \neq \theta_0, \end{aligned} \tag{5.1}$$

$$\begin{aligned} \Theta =& \mathbb{R}^+ \\ H_0 :& \theta = 0 \\ H_1 :& \theta > 0, \end{aligned} \tag{5.2}$$

De manera más general definimos un test de hipótesis binario compuesto como:

$$\begin{aligned} \Theta =& \Theta_0 \cup \Theta_1 \\ H_0 :& \theta \in \Theta_0 \\ H_1 :& \theta \in \Theta_1, \end{aligned} \tag{5.3}$$

Donde $\Theta_0 \cap \Theta_1 = \emptyset$. Notamos que en este caso además $\Delta \neq \Theta$ ya que $\Delta = \{H_0, H_1\}$, o bien $\Delta = \{0, 1\}$ donde la decisión será si quedarse con θ_0 (elegir 0) o si cambió (elegir 1). La ventaja de los test compuestos es que son más generales que los simples y más versátiles, resulta natural plantear la hipótesis de que un parámetro cambió ($\theta \neq \theta_0$) respecto a su condición actual ($\theta = \theta_0$) sin saber a cuál valor cambió realmente θ, lo importante es solamente detectar el cambio[1].

Lamentablemente en el caso de test de hipótesis compuestos no existe un equivalente al test de Neyman-Pearson, es decir, no existe un criterio que garantice optimalidad en el sentido de mejor compromiso entre los dos tipos de errores. Sin embargo, existen extensiones que, bajo ciertas hipótesis adicionales, permiten garantizar la optimalidad en casos compuestos. Introduciremos ciertas definiciones que son extensiones directas de lo visto en la Unidad I.

[1] Este tipo de análisis también se conoce como concept drift o model drift.

Definición 5.1. (Tamaño de Test) Para un test de hipótesis compuesto, un test o regla $\pi : \mathbb{X} \to \{0, 1\}$ se dice de tamaño $\alpha \in [0, 1]$ si:

$$\sup_{\theta \in \Theta_0} \mathbb{E}(\pi(X)|\theta \in \Theta_0) = \alpha \tag{5.4}$$

Definición 5.2. (Test Uniformemente más Poderoso) Para un test de hipótesis compuesto, un test o regla $\pi : \mathbb{X} \to \{0, 1\}$ se dice uniformemente poderoso (UMP) de tamaño $\alpha \in [0, 1]$ si, para cualquier otro test $\tilde{\pi} : \mathbb{X} \to \{0, 1\}$ también de tamaño α entonces:

$$(\forall \theta \in \Theta_1) \;\; \mathbb{E}(\pi(X)|\theta \in \Theta_1) \geq \mathbb{E}(\tilde{\pi}(X)|\theta \in \Theta_1). \tag{5.5}$$

Estas definiciones son extensiones naturales de las vistas en el caso de test de hipótesis binario simple, sin embargo, (5.5) es una condición muy exigente. Si bien el test de Neyman-Pearson es óptimo para el caso $\theta_1 \in \Theta_1$ para un tamaño α dado, no existen garantías de que también lo sea para el caso $\theta_2 \in \Theta_1$ con $\theta_1 \neq \theta_2$. Pese a ello, en adelante veremos condiciones para poder garantizar la existencia test UMP de tamaño α. El siguiente ejemplo servirá como motivación para encontrar condiciones de suficiencia para un test UMP.

Ejemplo 5.1. Consideremos $X \sim N(\theta, \sigma^2)$, $\Theta_0 =]-\infty, \mu_0]$ y $\Theta_1 =]\mu_0, \infty[$ quisiéramos detectar θ bajo las hipótesis

$$\begin{aligned} \Theta =& \Theta_0 \cup \Theta_1 \\ H_0 :& \theta \leq \mu_0 \\ H_1 :& \theta > \mu_0 \end{aligned} \tag{5.6}$$

Nos gustaría encontrar el test π UMP bajo la condición

$$(\forall \theta \leq \mu_0)\mathbb{E}(\pi(X)|\theta \in]-\infty, \mu_0]) \leq \alpha \tag{5.7}$$

Para resolver esto, primero resolvamos un caso más simple, es decir, $H_0' : \theta = \mu_0$, $H_1' = \mu_1$ con $\mu_0 < \mu_1$. Notamos que este caso es el mismo que el ejemplo 1.1, en cuyo caso sabemos que si aplicamos el Lema de Neyman-Pearson con un umbral ν tenemos que el test puede reescribirse como (ver (1.62)):

$$\pi(x) = \begin{cases} 1 & \text{si } x \geq \tau \\ 0 & \text{si } x < \tau \end{cases} \tag{5.8}$$

con $\tau = \frac{2\log(\nu)\sigma^2+\mu_1^2-\mu_0^2}{2\mu_1-2\mu_0}$. Podemos observar que los errores de tipo I y II son (ver 1.75 y 1.74):

$$\beta = Q\left(\frac{\tau-\mu_1}{\sigma}\right) \tag{5.9}$$

$$\alpha = Q\left(\frac{\tau-\mu_0}{\sigma}\right). \tag{5.10}$$

Esto significa que:

$$\tau = \sigma^2 Q^{-1}(\alpha) + \mu_0. \tag{5.11}$$

y, por lo tanto, el test se puede escribir como:

$$\pi(x) = \begin{cases} 1 & \text{si } x \geq \sigma^2 Q^{-1}(\alpha) + \mu_0 \\ 0 & \text{si } x < \sigma^2 Q^{-1}(\alpha) + \mu_0 \end{cases} \tag{5.12}$$

para lograr un tamaño α, la elección de τ no depende de μ_1, sea cual sea su valor entre $]\mu_0, \infty[$. Esto significa que en este caso el Lemma de Neyman-Pearson es UMP para el caso $H_0 : \theta = \mu_0$, $H_1 : \theta > \mu_0$. Para completar el análisis, falta extender el análisis al caso $\theta \leq \mu_0$. Para esto basta ver que se cumpla (5.7), en efecto, sea $\theta' \leq \mu_0$, tenemos que:

$$\begin{aligned} \mathbb{E}(\pi(X)|\theta = \mu_0) &= Q\left(\frac{\tau-\mu_0}{\sigma}\right) \\ &\geq Q\left(\frac{\tau-\theta'}{\sigma}\right) \\ &= \mathbb{E}(\pi(X)|\theta = \theta'). \end{aligned} \tag{5.13}$$

Lo que cumple la condición de tamaño α según la Definición 5.4, lo que significa que el test en (5.12) es UMP según las hipótesis establecidas en (5.6).

En resumen en este caso encontramos un test UMP donde se pide detectar la media de una distribución normal sujeto a las hipótesis $H_0 : \theta \leq \mu_0$ contra $H_1 : \theta > \mu_0$. Este tipo de test se llaman *one-sided* y la decisión es bastante simple: rechazar H_0 si x supera un umbral τ y aceptarlo en caso contrario. El umbral se elije de forma tal que tenga tamaño α.

Con todo el análisis visto anteriormente, estamos en condiciones de dar un criterio para lograr test UMP. Para esto introduciremos el concepto de razón de verosimilitud monótono:

Definición 5.3. (Razón de Verosimilitud Monótono) Una familia de distribuciones $\{P_X(\cdot|\theta \in \Theta)\}$ tiene una razón de verosimilitud monótona si para $\theta_1 < \theta_2$, $\theta_1, \theta_2 \in \Theta$ la función:

$$ML(X) = \frac{L(X|\theta = \theta_2)}{L(X|\theta = \theta_1)} \tag{5.14}$$

es creciente en los puntos donde al menos una de las dos verosimilitudes es positiva. Si $L(X|\theta = \theta_2) > 0$ y $L(X|\theta = \theta_1) = 0$ se definirá como ∞.

Luego si una familia de distribuciones tiene razón de verosimilitud monótono, mientras mayor sea el valor de la observación, es mas probable que provenga de H_1.

El siguiente teorema es una extensión del Lema de Neyman-Pearson para test compuestos y razón de verosimilitud monótono.

Teorema 5.1. (Extensión de Karlin y Rubin) Sea $x_0 \in \mathbb{R}$, un test aleatorio de la forma

$$\pi(w, x) = \begin{cases} 1 & \text{si } x > x_0 \\ 0 & \text{si } x < x_0 \\ \rho(w) & \text{si } x = x_0 \end{cases} \tag{5.15}$$

con $\rho \in Bernoulli(p)$. Cualquier test de la forma en (5.15) es UMP para las hipótesis $H_0 : \theta \leq \theta_0$ contra $H_1 : \theta > \theta_0$ siempre que le tamaño no sea 0. Para cualquier $\alpha \in]0, 1]$ existe un $x_0 \in \mathbb{R}$ y un $p \in [0, 1]$ tal que el test sea de tamaño $\alpha \in]0, 1]$ para las hipótesis $H_0 : \theta \leq \theta_0$ contra $H_1 : \theta > \theta_0$.

Demostración: Consideremos una familia de distribuciones con razón de verosimilitud monótona, si $x > x_0$ entonces $ML(x) \geq ML(x_0)$ luego, cualquier test de la forma (5.15) se puede escribir como un test de Neyman-Pearson con $\nu > 0$ para $H_0 : \theta = \theta_0$ contra $H_1 : \theta = \theta_1$ con $\theta_0 < \theta_1$ tal que:

$$\pi(w, x) = \begin{cases} 1 & \text{si } L(x|\theta_1) > \nu L(x|\theta_0) \\ 0 & \text{si } L(x|\theta_1) < \nu L(x|\theta_0) \\ \rho(w) & \text{si } L(x|\theta_1) = \nu L(x|\theta_0) \end{cases} \tag{5.16}$$

Luego, gracias al Lema de Neyman-Pearson el test en (5.15) es óptimo en poder ya que es independiente de θ_1 siempre y cuando $\theta_1 > \theta_0$ (así se preserva la hipótesis de monotonía). Similarmente, es óptimo de tamaño α ya que al ser un test de Neyman-Pearson es alcanzable para un θ fijo, por lo tanto directamente se puede elegir un $x_0 \in \mathbb{R}$ tal que $(\forall \theta \leq \theta_0)\mathbb{E}(\pi(\rho, X)|\theta \in]-\infty, \theta_0]) \leq \alpha$. □

5.2. Test de Verosimilitud Generalizado

En la sección anterior encontramos condiciones suficientes para encontrar un test óptimo para test compuestos one-sided. Si bien la familia de distribuciones con razón de verosimilitud monótona se cumple para distribuciones de tipo exponencial, no siempre es el caso, lo que dificulta ocupar el criterio de Karlin y Rubin.

En esta sección abordaremos un test muy usado en estadística y veremos cómo a partir de este test es posible derivar funciones (estadísticos) muy usados en problemas de test de hipótesis compuestos, este test se llama razón de veromisilitud generalizado (GLRT):[2]

Definición 5.4. (Test de Razón de Verosimilitud Generalizado) Consideremos un espacio de observación $\mathbb{X}$ y un espacio de parámetros $\Theta = \Theta_0 \cup \Theta_1$ con $\Theta_0 \cap \Theta_1 = \emptyset$, una familia de distribuciones $\{P_X(\cdot|\theta \in \Theta)\}$ y un test de hipótesis $H_0 : \theta \in \Theta_0$ contra $H_0 : \theta \in \Theta_1$ con $\Delta = \{0, 1\}$. Definimos el test de verosimilitud generalizado, dado $\nu > 0$, como:

$$\pi_{GRLT}(X) = \begin{cases} 1 & \text{si } \frac{\sup\limits_{\theta\in\Theta} L(X|\theta\in\Theta)}{\sup\limits_{\theta\in\Theta_0} L(X|\theta\in\Theta_0)} \geq \nu \\ 0 & \sim \end{cases} \tag{5.17}$$

La razón

$$\frac{\sup\limits_{\theta\in\Theta} L(X|\theta \in \Theta)}{\sup\limits_{\theta\in\Theta_0} L(X|\theta \in \Theta_0)} \tag{5.18}$$

se conoce como razón de verosimilitud generalizado.

El GLRT es una extensión del test de verosimilitud, además, directamente se observa que cuando $\Theta_0 = \{0\}$ y $\Theta_1 = \{1\}$ se recupera el test de verosimilitud. Este test es general, por lo que en principio no es posible dar garantías que sea UMP, sin embargo, permite entregar expresiones cerradas para una gran familia de test de hipótesis conocidos. Veamos los siguientes ejemplos para familiarizarse con distintos test estadísticos:

Ejemplo 5.2. Consideremos $\mathbb{X} = \mathbb{R}$ y un vector variable X_1^n i.i.d. tal que $(\forall i \in \{1, ..., n\})$ $X_i \sim N(\theta, \sigma^2)$, $\Theta_0 = \{\mu_0\}$ y $\Theta_1 = \mathbb{R}\backslash\{\mu_0\}$ quisiéramos detectar μ_0 bajo las

[2] Generalised Likelihood Test Ratio

hipótesis:

$$\begin{aligned}\Theta =&\Theta_0 \cup \Theta_1\\ H_0 :&\theta = \mu_0\\ H_1 :&\theta \neq \mu_0\end{aligned} \tag{5.19}$$

Para esto propondremos el GLRT y calcularemos la razón de verosimilitud generalizada:

$$\begin{aligned}\frac{\sup\limits_{\theta\in\Theta} L(X_1^n|\theta\in\Theta)}{\sup\limits_{\theta\in\Theta_0} L(X_1^n|\theta\in\Theta_0)} &= \frac{\sup\limits_{\theta\in\Theta} L(X_1^n|\theta\in\Theta)}{L(X_1^n|\theta=\mu_0)}\\ &= \frac{L\left(X_1^n|\theta=\overline{X}\right)}{L(X_1^n|\theta=\mu_0)}\\ &= \frac{e^{-\sum_{i=1}^n \frac{(X_i-\overline{X})^2}{2\sigma^2}}}{e^{-\sum_{i=1}^n \frac{(X_i-\mu_0)^2}{2\sigma^2}}}\\ &= e^{\sum\limits_{i=1}^n \frac{(X_i-\mu_0)^2-(X_i-\overline{X})^2}{2\sigma^2}}\\ &= e^{\sum\limits_{i=1}^n \frac{X_i^2-2X_i\mu_0+\mu_0^2-X_i^2+2X_i\overline{X}-\overline{X}^2}{2\sigma^2}}\\ &= e^{\sum\limits_{i=1}^n \frac{-2X_i\mu_0+\mu_0^2+2X_i\overline{X}-\overline{X}^2}{2\sigma^2}}\\ &= e^{\frac{-2n\overline{X}\mu_0+n\mu_0^2+2n\overline{X}^2-n\overline{X}^2}{2\sigma^2}}\\ &= e^{\frac{-2n\overline{X}\mu_0+n\mu_0^2+n\overline{X}^2}{2\sigma^2}}\\ &= e^{\frac{n(\overline{X}-\mu_0)^2}{2\sigma^2}}\end{aligned} \tag{5.20}$$

donde $\overline{X} = \frac{1}{n}\sum\limits_{i=1}^n X_i$ corresponde a la media empírica solución de $\arg\max\limits_{\theta\in\Theta} L(X_1^n|\theta\in\Theta)$ que equivale al criterio de máxima verosimilitud visto en estimación paramétrica (ver (3.91)). Esto nos dice que el GLRT se puede escribir como:

$$\pi_{GLRT}(X_1,...,X_n) = \begin{cases} 1 & \text{si } \frac{n\left(\overline{X}-\mu_0\right)^2}{\sigma^2} \geq \log(2\nu) \\ 0 & \sim \end{cases} \tag{5.21}$$

Podemos notar que el test anterior puede resolverse de dos maneras. La primera forma es

notar que:

$$\begin{aligned}\frac{n\left(\overline{X}-\mu_0\right)^2}{\sigma^2} &= \frac{n\left(\frac{1}{n}\sum_{i=1}^n X_i-\mu_0\right)^2}{\sigma^2}\\ &= n\left(\frac{1}{n}\sum_{i=1}^n\frac{(X_i-\mu_0)}{\sigma}\right)^2\\ &= \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n\frac{(X_i-\mu_0)}{\sigma}\right)^2\\ &= Z^2\end{aligned} \tag{5.22}$$

donde $Z\sim N(0,1)$, luego $Z^2\sim\chi_1^2$, lo que corresponde a una distribución chi-cuadrado con 1 grado de libertad. Por completitud la distribución chi-cuadrado con n grados de libertad es una variable aleatoria $Y\sim\chi_n^2$ cuya densidad es:

$$f_Y(y)=\begin{cases}\frac{y^{n/2-1}e^{-y/2}}{2^{n/2}\Gamma(y/2)},\ \text{si } y\geq 0\\ 0, \qquad \sim\end{cases} \tag{5.23}$$

Es interesante además notar que si $X_i\sim N(0,1)$ independientes para todo $i\in\{1,...,n\}$ entonces $Z=\sum_{i=1}^n X_i\sim\chi_n^2$.

La segunda forma de evaluar el test es reescribiéndolo en términos del error de tipo I:

$$\begin{aligned}P_X(\pi(X)=1|\theta=\mu_0) &= P_X\left(\frac{n\left(\overline{X}-\mu_0\right)^2}{\sigma^2}\geq\log(2\nu)\middle|\theta=\mu_0\right)\\ &= P_X\left(|Z|\geq\sqrt{\log(2\nu)}\middle|\theta=\mu_0\right)\quad Z\sim N(0,1)\\ &= 2P_X\left(Z\geq\sqrt{\log(2\nu)}\middle|\theta=\mu_0\right)\\ &= 2Q(\sqrt{\log(2\nu)})\end{aligned} \tag{5.24}$$

Luego, si se desea un error de tipo I α, basta elegir:

$$\sqrt{\log(2\nu)}=Q^{-1}\left(\frac{\alpha}{2}\right) \tag{5.25}$$

Lo que permite escribir el test

$$\pi_{GLRT}(X_1,...,X_n) = \begin{cases} 1 & \text{si } \left|\overline{X}-\mu_0\right| \geq \frac{\sigma Q^{-1}\left(\frac{\alpha}{2}\right)}{\sqrt{n}} \\ 0 & \text{si } \mu_0 - \frac{\sigma Q^{-1}\left(\frac{\alpha}{2}\right)}{\sqrt{n}} < \overline{X} < \mu_0 + \frac{\sigma Q^{-1}\left(\frac{\alpha}{2}\right)}{\sqrt{n}} \end{cases} \tag{5.26}$$

El intervalo $\left[\overline{x} - \frac{\sigma Q^{-1}\left(\frac{\alpha}{2}\right)}{\sqrt{n}}, \overline{x} + \frac{\sigma Q^{-1}\left(\frac{\alpha}{2}\right)}{\sqrt{n}}\right]$ se conoce como un **intervalo de confianza**, indica un intervalo donde, con probabilidad $1-\alpha$, el parámetro desconocido estará en dicho intervalo. Tipicamente $\alpha = 0{,}05$ lo que se conoce como nivel de significancia del $95\,\%$.

Ejemplo 5.3. Consideremos $\mathbb{X} = \mathbb{R}^n$ y un vector variable X_1^n i.i.d. tal que $(\forall i \in \{1,...,n\})\ X_i \sim N(\mu,\theta)$, $\Theta_0 = \{\mu_0\} \times \mathbb{R}^+$ y $\Theta_1 = \mathbb{R} \times \mathbb{R}^+ \backslash (\{\mu_0\} \times \mathbb{R}^+)$, con σ_0^2 **desconocido**, quisiéramos detectar μ_0 bajo las hipótesis:

$$\begin{aligned} \Theta =& \Theta_0 \cup \Theta_1 \\ H_0 :& \theta = \mu_0 \\ H_1 :& \theta \neq \mu_0 \end{aligned} \tag{5.27}$$

Nuevamente propondremos el GLRT y calcularemos la razón de verosimilitud generalizada, sin embargo, al ser σ_0^2 desconocido, también debe ser estimado. Cuando μ_0 es conocido el estimador de máxima verosimilitud para σ_0^2 es $\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$, en cambio, cuando μ_0 es

conocido el estimador de máxima verosimilitud para σ_0^2 es $\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$:

$$
\begin{aligned}
\frac{\sup_{\theta\in\Theta} L(X_1^n|\theta\in\Theta)}{\sup_{\theta\in\Theta_0} L(X_1^n|\theta\in\Theta_0)} &= \frac{\sup_{\theta\in\Theta} L(X_1^n|\theta\in\Theta)}{L(X_1^n|\theta=\mu_0)} \\
&= \frac{L\left(X_1^n|\theta=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\right)}{L(X_1^n|\theta=\mu_0)} \\
&= \frac{\frac{1}{(2\pi)^{n/2}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\right)^{n/2}} e^{-\sum_{i=1}^{n}\frac{(X_i-\overline{X})^2}{\frac{2}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2}}}{\frac{1}{(2\pi)^{n/2}\left(\frac{1}{n}\sum\limits_{i=1}^{n}(X_i-\mu)^2\right)^{n/2}} e^{-\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\frac{2}{n}\sum\limits_{i=1}^{n}(X_i-\mu)^2}}} \\
&= \left(\frac{\sum_{i=1}^{n}(X_i-\mu_0)^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\right)^{n/2} \\
&= \left(\frac{\sum_{i=1}^{n}(X_i-\overline{X}+\overline{X}-\mu_0)^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\right)^{n/2} \\
&= \left(\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2-2(X_i-\overline{X})(\overline{X}-\mu_0)+(\overline{X}-\mu_0)^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\right)^{n/2} \\
&= \left(\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2+n(\overline{X}-\mu_0)^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\right)^{n/2} \\
&= \left(1+\frac{n(\overline{X}-\mu_0)^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\right)^{n/2} \qquad (5.28)
\end{aligned}
$$

Gracias a la monotonía de (5.28) y que el umbral es arbitrario podemos proponer entonces

el siguiente estadístico

$$\begin{aligned}\frac{n(n-1)(\overline{X}-\mu_0)^2}{\sum_{i=1}^n (X_i-\overline{X})^2} &= \left(\frac{\sqrt{n(n-1)}(\overline{X}-\mu_0)}{\sqrt{\sum_{i=1}^n (X_i-\overline{X})^2}}\right)^2 \\ &= \left(\frac{\overline{X}-\mu_0}{\frac{\sqrt{\sum_{i=1}^n (X_i-\overline{X})^2}}{\sqrt{n(n-1)}}}\right)^2 \\ &= \left(\frac{\overline{X}-\mu_0}{\sqrt{\frac{S}{n}}}\right)^2 \end{aligned} \tag{5.29}$$

La variable $S = \frac{1}{n-1}\sum_{i=1}^n (X_i-\overline{X})^2$ se conoce como la varianza muestral, corresponde a un estimador insesgado de la varianza σ_0^2. La variable

$$t_{n-1} = \frac{\overline{X}-\mu_0}{\sqrt{\frac{S}{n}}} \tag{5.30}$$

es una variable aleatoria de distribución t-student con $n-1$ grados de libertad. Más precisamente una variable aleatoria t_n deriva de la siguiente expresión:

$$t_n = \frac{Z}{\sqrt{Y/n}} \tag{5.31}$$

donde $Z \sim N(0,1)$ e $Y \sim \chi_k$ con Z e Y independientes. En particular si $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ e $Y = \frac{\sum_{i=1}^n (X_i-\overline{X})^2}{\sigma^2} \sim {\chi_{n-1}}^3$ tenemos que

$$\begin{aligned}\frac{Z}{\sqrt{Y/(n-1)}} &= \frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{\sum_{i=1}^n (X_i-\overline{X})^2}{\sigma^2(n-1)}}} \\ &= \frac{\overline{X}-\mu}{\sqrt{\frac{S}{n}}}, \end{aligned} \tag{5.32}$$

lo que corrobora que posea distribución t_{n-1}.La distribución de t-student es muy usada en estadística y es una generalización de la distribución normal standard. Su función de

[3] Se puede demostrar que Z e Y son independientes.

densidad de probabilidad es:

$$f_{t_\tau}(t) = \frac{\Gamma\left(\frac{\tau+1}{2}\right)}{\sqrt{\pi\tau}\Gamma\left(\frac{\tau}{2}\right)}\left(1+\frac{t^2}{\tau}\right)^{-\frac{\tau+1}{2}} \tag{5.33}$$

donde τ es el grado de libertad, es un real positivo, pero en general suele ser un número natural. Entonces el test se puede escribir como, para $K \in \mathbb{R}$:

$$\pi_{GLRT}(X_1,...,X_n) = \begin{cases} 1 & \text{si } \left|\frac{\overline{X}-\mu_0}{\sqrt{\frac{S}{n}}}\right| \geq |K| \\ 0 & \sim \end{cases} \tag{5.34}$$

Notar que la región de decisión en (5.34) posee un valor absoluto, análogo al caso (5.26), lo que se conoce como un test *two-sided* debido a que existen dos zonas de rechazo de la hipótesis nula.

5.3. Transformada de Karhunen-Loéve

Consideremos un vector aleatorio X_1^m de media 0[4] y sea $\Sigma = cov(X_1^m) = \mathbb{E}((X_1^m)(X_1^m)^t)$ su matriz de varianza-covarianza. Sabemos que esta matriz Σ es simétrica, lo que significa que es diagonalizable y además admite una base ortonormal de vectores propios por lo que $\Sigma = \mathbf{U}\Lambda\mathbf{U}^t$ con $\mathbf{U}$ matriz de $m \times m$ de vectores propios ortonormales tal que $\mathbf{U}\mathbf{U}^t = \mathbf{U}^t\mathbf{U} = I$ y Λ una matriz diagonal.

Sea $Y_1^m = \mathbf{U}^t X_1^m$. Vemos que Y_1^m es un vector aleatorio cuyas componentes no están correlacionadas ya que:

$$\begin{aligned} \mathbb{E}((Y_1^m)(Y_1^m)^t) &= \mathbb{E}((\mathbf{U}^t X_1^m)(\mathbf{U}^t X_1^m)^t) \\ &= \mathbb{E}(\mathbf{U}^t(X_1^m(X_1^m)^t)\mathbf{U}) \\ &= \mathbf{U}^t\mathbb{E}(X_1^m(X_1^m)^t)\mathbf{U} \\ &= \mathbf{U}^t\mathbf{U}\Lambda\mathbf{U}^t\mathbf{U} \\ &= \Lambda. \end{aligned} \tag{5.35}$$

Naturalmente es posible recuperar el vector X_1^m mediante

$$X_1^m = \mathbf{U}Y_1^m = \sum_{i=1}^{m} {}^iU_1^m Y_i \tag{5.36}$$

[4] Si bien en un principio parece un supuesto fuerte, si un vector X_1^m no es de media cero, puede redefinirse uno nuevo como $Y_1^m = X_1^m - \mathbb{E}(X_1^m)$, por lo que no se pierde generalidad.

donde ${}^iU_1^m$ es el vector propio i-ésimo asociado a la matriz $\mathbf{U}$. La expresión en (5.36) nos dice que podemos construir un vector aleatorio como una combinación lineal de vectores ortogonales cuyas coeficientes son variables aleatorias no correlacionadas. Esta representación se conoce como la transformada de Karhunen-Loéve.

La transformada de Karhunen-Loéve podría servir para transmitir el vector X_1^m si, tanto el emisor como el receptor, conocen su matriz de varianza-covarianza y su descomposición. Entonces, enviando los Y_i son suficientes para representar X_1^m y se necesitan m números distintos para lograr tal representación.

Supongamos ahora que X_1^m es un vector con m de muy alta dimensión tal que lograr su representación total sea muy costosa. Nos gustaría aproximar X_1^m usando una menor cantidad de componentes. ¿Cuál es la mejor representación con la limitación de utilizar menores componentes?

Consideremos $\hat{X}_1^m = \mathbf{K}X_1^m$ una aproximación del vector X_1^m donde $\mathbf{K}$ es una matriz de $m \times m$ de rango $r < m$. Esta representación se llama representación de rango r de X_1^m ya que se necesitan r piezas de información para aproximar X_1^m.

El objetivo ahora es determinar de manera explícita $\mathbf{K}$, para esto, buscaremos la matriz $\mathbf{K}$ que minimice el error cuadrático medio entre X_1^m y su representación $\hat{X}_1^m$. Pediremos además que la matriz $\mathbf{K}$ sea simétrica[5] de esta manera podemos representar $\mathbf{K}$ como:

$$\mathbf{K} = \mathbf{U}D_r\mathbf{U}^t \tag{5.37}$$

[5] Este supuesto se inspira del hecho que la transformada de Karhunen-Loéve tanto $\mathbf{U}$ como $\mathbf{U}^t$ sirven como proyección de X_1^m (ver (5.36))

Donde $\mathbf{U} = ({}^1U_1^m|{}^2U_1^m|\ldots|{}^mU_1^m)$ es la matriz de vectores propios ortonormales y D_r es una matriz diagonal tal que

$$\begin{pmatrix} \mu_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \mu_2 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \mu_r & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \tag{5.38}$$

esto debido a que K es de rango $r < m$, luego posee $m - r$ filas linealmente dependientes y por tanto $m - r$ valores propios 0[6]. Con estas condiciones calculamos el error cuadrático medio o MSE en función de $\mathbf{K}$:

$$\begin{aligned} MSE(\mathbf{K}) &= ||X_1^m - \hat{X}_1^m||^2 \\ &= \mathbb{E}((X_1^m - \hat{X}_1^m)^t(X_1^m - \hat{X}_1^m)) \\ &= tr\left[\mathbb{E}((X_1^m - \hat{X}_1^m)(X_1^m - \hat{X}_1^m)^t)\right] \\ &= tr\left[\mathbb{E}((X_1^m - \mathbf{K}X_1^m)(X_1^m - \mathbf{K}X_1^m)^t)\right] \\ &= tr\left[\mathbb{E}((I - \mathbf{K})X_1^m(X_1^m)^t(I - \mathbf{K})^t)\right] \\ &= tr\left[(I - \mathbf{K})\Sigma(I - \mathbf{K})^t\right] \\ &= tr\left[(I - \mathbf{U}D_r\mathbf{U}^t)\Sigma(I - \mathbf{U}D_r\mathbf{U}^t)^t\right] \\ &= tr\left[(\mathbf{U}\mathbf{U}^t - \mathbf{U}D_r\mathbf{U}^t)\Sigma(\mathbf{U}\mathbf{U}^t - \mathbf{U}D_r\mathbf{U}^t)^t\right] \\ &= tr\left[\mathbf{U}(I - D_r)\mathbf{U}^t\Sigma\mathbf{U}(I - D_r)\mathbf{U}^t\right] \\ &= tr\left[(I - D_r)\mathbf{U}^t\mathbf{U}(I - D_r)\mathbf{U}^t\Sigma\mathbf{U}\right] \\ &= tr\left[(I - D_r)^2\mathbf{U}^t\Sigma\mathbf{U}\right] \end{aligned} \tag{5.39}$$

donde tr corresponde a la traza de una matriz, cumple que $tr(A) = tr(A^t)$ y es ciclica, i.e., $tr(ABC) = tr(CAB) = tr(BCA)$. Desarrollamos la traza, notando además que $\mathbf{U}_{ij} = {}^jU_i$

[6] $det(A - \lambda I) = det(A) = 0$ entonces A no es invertible

(la componente i-ésima del j-ésimo vector) y obtenemos lo siguiente:

$$\begin{aligned}
tr\left[(I-D_r)^2\mathbf{U}^t\Sigma\mathbf{U}\right] &= \sum_{i=1}^{m}\left[(I-D_r)^2\mathbf{U}^t\Sigma\mathbf{U}\right]_{ii} \\
&= \sum_{i=1}^{m}\sum_{k=1}^{m}((I-D_r)^2)_{ik}(\mathbf{U}^t\Sigma\mathbf{U})_{ki} \\
&= \sum_{i=1}^{m}((I-D_r)^2)_{ii}(\mathbf{U}^t\Sigma\mathbf{U})_{ii} \\
&= \sum_{i=1}^{m}((I-D_r)^2)_{ii}\sum_{j=1}^{m}(\mathbf{U}^t)_{ij}(\Sigma\mathbf{U})_{ji} \\
&= \sum_{i=1}^{m}((I-D_r)^2)_{ii}\sum_{j=1}^{m}(\mathbf{U})_{ji}(\Sigma\mathbf{U})_{ji} \\
&= \sum_{i=1}^{m}((I-D_r)^2)_{ii}\sum_{j=1}^{m}(\mathbf{U})_{ji}\sum_{l=1}^{m}\Sigma_{jl}\mathbf{U}_{li} \\
&= \sum_{i=1}^{m}((I-D_r)^2)_{ii}\sum_{j=1}^{m}(\mathbf{U})_{ji}\sum_{l=1}^{m}\Sigma_{jl}{}^iU_l \\
&= \sum_{i=1}^{m}((I-D_r)^2)_{ii}\sum_{j=1}^{m}(\mathbf{U})_{ji}(\Sigma^iU_1^m)_{j1} \\
&= \sum_{i=1}^{m}((I-D_r)^2)_{ii}\sum_{j=1}^{m}{}^iU_j(\Sigma U_i)_{j1} \\
&= \sum_{i=1}^{m}((I-D_r)^2)_{ii}({}^iU_1^m)^t\Sigma({}^iU_1^m) \\
&= \sum_{i=1}^{r}(1-\mu_i)^2({}^iU_1^m)^t\Sigma({}^iU_1^m) + \sum_{i=r+1}^{m}({}^iU_1^m)^t\Sigma({}^iU_1^m) \qquad (5.40)
\end{aligned}$$

Después de este humilde desarrollo nos damos cuenta que para minimizar el $MSE(K)$ se debe considerar $\mu_i = 1$ parar todo $i \in \{1, ..., r\}$. Entonces lo que se debe hacer es minimizar

$$\sum_{i=r+1}^{m}({}^iU_1^m)^t\Sigma({}^iU_1^m), \qquad (5.41)$$

reconocemos que $({}^{i}U_1^m)^t\Sigma({}^{i}U_1^m)$ corresponde a una forma cuadrática cuya cota inferior es $\lambda_{min}({}^{min}U_1^m)^t({}^{min}U_1^m)$ donde λ_{min} corresponde al valor propio de menor valor de Σ y ${}^{min}U_1^m$ su vector propio asociado. Con esto deducimos que entonces $\sum_{i=r+1}^{m}({}^{i}U_1^m)^t\Sigma({}^{i}U_1^m)$ debe contener los $m-r$ valores propios más pequeños y, por lo tanto,

$$\mathbf{K} = \mathbf{U}I_r\mathbf{U}^t, \tag{5.42}$$

es la matriz diagonalizada donde I_r es una matriz diagonal tal que vale 1 en la diagonal para los r valores propios más grandes de Σ y $\mathbf{U}$ es la matriz de vectores propios tal que los primeros r vectores propios son asociados a los r mayores valores propios de Σ.

La interpretación de este resultado es como sigue: Para obtener la mejor aproximación de X_1^m usando solamente r piezas de información, se debe enviar los valores Y_i asociados a los r valores propios más grandes de Σ.

5.4. Análisis de Componentes Principales

Una de las desventajas de la transformada de Karhunen-Loéve es que no es práctica. En general es complicado determinar Σ y los vectores propios asociados a la señal. Sin embargo, es posible aplicar este método indirectamente sobre fuentes de información notando que la transformada de Karhunen-Loéve es una proyección ya sea sobre un vector aleatorio o muestras de éste.

Antes de continuar, modificaremos levemente la notación para que sea menos pesada. Consideremos un vector aleatorio $\overline{X} = X_1^m$ de media 0 y consideremos ahora $(\overline{x}_1, ..., \overline{x}_n)$, es decir, n muestras del vector $\overline{X}$, no olvidar que este vector tiene dimensión m. Con estas n muestras podemos calcular la matriz de varianza-covarianza S empírica insesgada como:

$$S = \frac{1}{n-1}\sum_{i=1}^{n}\overline{x}_i(\overline{x}_i)^t \tag{5.43}$$

Nos gustaría nuevamente una aproximación del vector $\overline{X}$ manteniendo la mayor información posible sobre éste, para esto, queremos un nuevo vector Y_1^r tal que la $j-$ésima componente corresponda a una combinación lineal (proyección) de la data que represente la $j-$ésima mayor porción de la varianza de las observaciones. Esta selección de las componentes (es decir Y_j) de la data recibida se llama j-ésima componente principal.

Formalmente, consideremos Y_1^r las r componentes principales de la data proveniente de $Y_j = (\overline{v}_j)^t X_1^m$ donde $\overline{v}_j$ es el vector que proyecta la información dada por $\overline{X}$. Así, por ejemplo, la primera componente principal es elegida tal que $Y_1 = (\overline{v}_1)^t \overline{X}$ y $\overline{v}_1$ es elegido de forma tal que la varianza empírica de Y_1 sea máxima, sujeto a la condición que $||\overline{v}_1|| = 1$. Dado que solamente tenemos acceso a n muestras y no la distribución real tendremos que ocupar la representación empírica dada por $y_i = (\overline{v}_1)^t(\overline{x}_i)$ para cada $i \in \{1, ..., n\}$. La varianza empírica de Y_1 será entonces

$$\begin{aligned}
\hat{\sigma}^2_{Y_1} &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i)^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}((\overline{v}_1)^t(\overline{x}_i))^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(\overline{v}_1)^t(\overline{x}_i)(\overline{v}_1)^t(\overline{x}_i) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(\overline{v}_1)^t(\overline{x}_i)(\overline{x}_i)^t(\overline{v}_1) \\
&= (\overline{v}_1)^t S(\overline{v}_1). \qquad (5.44)
\end{aligned}$$

Observamos que maximizar (5.44) sujeto a $||\overline{v}_1|| = 1$ es el problema visto en (5.41): $\overline{v}_1$ es el vector propio normalizado que corresponde al mayor valor propio asociado a S, λ_1, luego:

$$\hat{\sigma}^2_{Y_1} = (\overline{v}_1)^t S(\overline{v}_1) = \lambda_1(\overline{v}_1)^t(\overline{v}_1) = \lambda_1. \qquad (5.45)$$

La siguiente componente principal, Y_2 se elije de forma tal que Y_2 no esté correlacionada con Y_1 lo que implica la condición $(\overline{v}_1)^t(\overline{v}_2) = 0$ lo que significa que $(\overline{v}_2)$ será el vector propio asociado al segundo mayor valor propio de S y así. Los vectores propios calculados para obtener las componentes principales se llaman direcciones de componentes principales. Si gran parte de la varianza de la señal está contenida en las componentes principales, estas componentes pueden usarse en vez de las observaciones para muchas aplicaciones estadísticas.

Observaciones 5.1.

- El análisis de componentes principales entonces corresponde a una reducción de la dimensionalidad del vector de observaciones. Los usos son variados, pero

esta técnica se aplica principalmente en los siguientes contextos: para reducir el espacio asignado en los datos, para acelerar el proceso de aprendizaje del algoritmo usado y cuando se desea visualizar la información (para esto se usan 2 o 3 componentes principales).

- Esta técnica no elimina las características originales, sino que se reinterprentan, es decir, al aplicar análisis de componentes principales se crearán nuevas características a partir del vector de características inicial. El cuidado de esto es que las nuevas características obtenidas son menos interpretables que las originales ya que buscan capturar la mayor varianza. Dado lo anterior esta técnica "resume"las características de lo observado.
- Existe una equivalencia entre el análisis de componentes principales y mínimos cuadrados. Como se verá en el Ejemplo 5.4, encontrar la dirección principal es equivalente a encontrar la proyección del vector que minimiza la distancia entre la observación y su proyección. Es decir, maximizar varianza equivale a minimizar error cuadrático.

Ejemplo 5.4. La Figura 5.1 muestra 200 muestras de un vector bidimensional $\overline{X} = X_1^2 = (X_1, X_2)$, dados por $(\overline{x}_1, ..., \overline{x}_{200})$. Se calcula la matriz de varianza-covarianza empírica y se obtiene que:

$$S = \frac{1}{200-1} \sum_{i=1}^{200} \overline{x}_i (\overline{x}_i)^t = \begin{pmatrix} 24{,}1893 & 10{,}6075 \\ 10{,}6075 & 6{,}38059 \end{pmatrix} \tag{5.46}$$

Los vectores propios y valores propios asociados a esta matriz S son los siguientes:

$$\overline{v}_1 = \begin{pmatrix} 0{,}9064 \\ 0{,}4225 \end{pmatrix} \text{ y } \lambda_1 = 29{,}1343. \tag{5.47}$$

$$\overline{v}_2 = \begin{pmatrix} -0{,}4225 \\ 0{,}9064 \end{pmatrix} \text{ y } \lambda_2 = 1{,}4355. \tag{5.48}$$

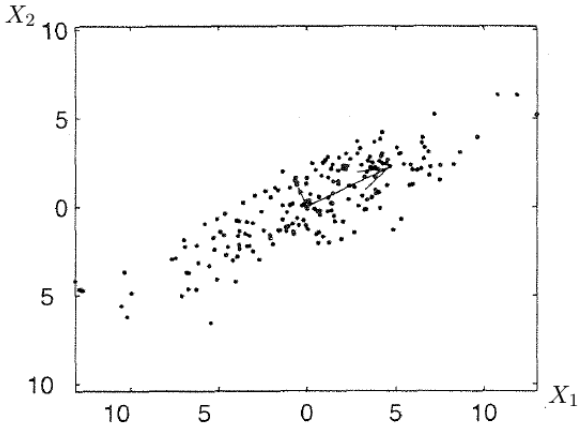


Figura 5.1: Scatter Plot del análisis de componentes principales.

La Figura 5.1 muestra además el gráfico de los vectores propios, las direcciones de componentes principales de la data, aumentado por la raíz de su correspondiente valor propio. La primera dirección principal sería:

$$Y_1 = (\overline{v}_1)^t \overline{X} \tag{5.49}$$

$$= 0{,}9064X_1 + 0{,}4225X_2 \tag{5.50}$$

que representa el $100 \cdot \frac{29{,}1343}{29{,}1343+1{,}4355} \approx 95\,\%$ de la varianza total del vector aleatorio X_1^2 y, por lo tanto, una buena aproximación del vector X_1^2.

Referencias

[1] Breiman, L. (1992). "Probability".
[2] Meyer, P. (1992). "Probabilidad y Aplicaciones Estadisticas".
[3] Ross, S. (1997). "A First Course in Probability".
[4] Todd K. Moon & W. C. Stirling, "Mathematical methods and algorithms for signal processing", NJ: Prentice hall, USA, 2000.
[5] Ravi R. Mazumdar, (2002) "Notes on Probability and Stochastic Processes,"Purdue University.
[6] Jaime San Martín, "Teoría de la medida", Editorial Universitaria, Chile, 2018.
[7] Espinosa, S. (2023). "Probabilidad y Procesos Estocásticos", versión 1.0.

Agradecimientos especiales para:

- Matías Carvajal por dar correcciones e ideas para mejorar este apunte.
- José Pablo Araya y Cristóbal Allendes por ser excelentes auxiliares. Siempre con la mejor disposición y apoyo.
- Jo porque ella al saber que estaba haciendo un apunte, exigió uno de manera imperiosa. Además citó mi apunte en su tesis de magíster, un ejemplo a seguir.
- VALERIA porque nuevamente me entrega los consejos precisos, quien desde el minuto 0 creyó en mi asegurando que sería el mejor profesor, sin ella el apunte nunca hubiese existido, la mejor hermana del alma ♡.
- PAT, obvio, sin ella los problemas diseñados pierden total sentido ya que es fuente de inspiración inagotable de preguntas creativas (su identidad será secreta hasta el fin de los tiempos). TKM PAT.