Notas de clase

ESTADÍSTICA

Esta versión: 16 de marzo de 2022

Última versión: github.com/GAMES-Uchile/Curso-Estadistica

Felipe Tobar
Magister Data Science
Universidad de Chile

ftobar@dim.uchile.cl
www.dim.uchile.cl/~ftobar

Prefacio

Estas notas de clase contienen el material en base al cual se ha dictado el curso de Estadística: Teoría y Aplicaciones (MDS7101) del *Master of Data Science* el año 2022 y Estadística (MA3402) en el Departamento de Ingeniería Matemática en 2019 y 2020, ambos programas en la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile. Esta es una versión preliminar de lo que espero que alguna vez sea un *Apunte de Estadística*, pero por ahora hay parte incompletas y en desarrollo.

En el proceso de realizar este curso he recibido ayuda invaluable de varias personas. Me gustaría agradecer a Joaquín Fontbona y Daniel Remenik por proponerme dictar este curso en el Departamento de Ingeniería Matemática y por su constante disposición a discutir contenidos de éste. Además, muchas gracias a Natacha Astromujoff por su infinita ayuda en todo lo referente a la administración del curso (no solo en este curso sino que en todos). También agradezco al Centro de Modelamiento Matemático, por ser mi segundo hogar las primeras dos veces que dicté el curso. Finalmente, he sido muy afortunado de contar con el profesionalismo y entrega de un grupo espectacular de ayudates: ¡Gracias Francisco Vásquez, Arie Wortsman, y Bruno Moreno por todo el apoyo en el desarollo de este curso curso y en la producción de este apunte!

Felipe Tobar,
Santiago,
marzo 2022.

Índice general

Capítulo 1

Introducción

En este capítulo se motivará el estudio de la estadística, dando a conocer su campo de acción, junto con su relación con otras disciplinas de la matemática teórica y aplicada.

1.1. Motivación

Consideremos el siguiente escenario. Una moneda es lanzada al aire 99 veces, y en todas ellas observamos una *cara* (y ningún *sello*). En esta inusual situación, le preguntamos a una colega matemática cuál es la probabilidad de que el siguiente lanzamiento resulte sello. Ella no duda en responder "$\frac{1}{2}$". Implícitamente, nuestra colega ha asumido que la moneda no está *cargada*, es decir, que la probabilidad de observar cara o sello es la misma y consecuentemente la probabilidad de obtener el resultado de las 99 caras tiene probabilidad $(1/2)^{99}$, al igual que cualquiera de los otros 2^{99} posibles resultados (secuencias) de este experimento. El supuesto de que la moneda no está cargada puede venir del hecho que ella no tiene evidencia sobre la forma y/o composición de la moneda que le permitan confiar que ésta está cargada, entonces, ante esta falta de información, nuestra colega asume igual probabilidad de obtener cara o sello.

En este curso, estudiaremos una vía alternativa para evaluar si la moneda está cargada no, prescindiendo del conocimiento los aspectos físicos de la moneda. De hecho, notemos que el obtener 99 caras seguidas sugiere fuertemente que la moneda sí está cargada: si asumimos que la moneda no está cargada, en 99 lanzamientos existe una probabilidad de

$$1 - (1/2)^{99} = 0{,}99999999999999999999999999999984222782\ldots$$

de ver al menos un sello. Con lo que nos gustaría decir que la moneda está cargada como *por contradicción*. En la misma línea, ante el resultado mencionado anteriormente, podemos decir que la probabilidad de que la moneda **esté cargada** es mayor que la probabilidad de que no lo esté. Este ejemplo del lanzamiento de una moneda desconocida es una ilustración para escenarios generales donde desconocemos las propiedades físicas pero tenemos *evidencia empírica*, *datos*, *realizaciones* (en caso que consideramos que estos fenómenos son expresiones de una variable aleatoria). En este escenario, aflora naturalmente la siguiente pregunta: ¿Será posible usar datos para obtener mejores predicciones o decisiones?

Pareciese entonces que la estadística tiene que ver con las probabilidades, pues ambas hablan de *realizaciones* y de *probabilidad de ocurrencia*. Sin embargo, es precisamente al considerar el *uso de datos* para dilucidar las propiedades intrínsecas de un objeto o fenómeno general, lo que nos lleva a entender la diferencia entre las probabilidades y la estadística. La primera se dedica al estudio del comportamiento de los fenómenos naturales asumiendo que conocemos sus propiedades, tal como el caso descrito en el Ejemplo 1.1.

Ejemplo 1.1 (enfoque de las probabilidades)**.**

Asuma que tiene un dado de 6 caras **no cargado**. ¿Cuál es la probabilidad de que, dentro de $2N$ lanzamientos, obtenga más de N resultados pares?

La estadística, por el contrario, se dedica a entender las propiedades inherentes de los objetos/fenómenos (que usualmente son asumidas en el estudio de Probabilidades) desde sus realizaciones como en el ejemplo 1.2.

Ejemplo 1.2 (enfoque de la estadística)**.**

Ante la observación de una secuencia de N lanzamientos de un dado, cuya media y desviación estándar (muestral) están dadas por $\bar{x}$ y $\bar{s}$, ¿cuál es la probabilidad de que el dado esté cargado?

La diferencia entre ambas disciplinas es muy sutil para el no experto, pero informalmente podemos postular que el objetivo de la inferencia estadística está en la *dirección opuesta* al de las probabilidades: mientras que la última asume parámetros para predecir resultados, la primera usa resultados para estimar parámetros. Otra forma coloquial de ilustrar esta diferencia es decir que en probabilidades estudiamos las consecuencias de un mundo ideal, mientras que en estadística verificamos hasta qué punto nuestro mundo es ideal. Un diagrama de la relación entre probabilidades y estadística se ilustra a continuación.

$\rightarrow\rightarrow\rightarrow$	Probabilidades	$\rightarrow\rightarrow\rightarrow$
general		particular
población		muestra
modelo		datos
$\leftarrow\leftarrow\leftarrow$	Estadística	$\leftarrow\leftarrow\leftarrow$

Por esta razón, hay quienes dicen que la inferencia estadística es una **probabilidad inversa**.

Si bien hay diferencias claras que pueden ser identificadas en ambos enfoques, los recursos de las probabilidades y de la estadística suelen usarse en conjunto para problemas como el enunciado en el Ejemplo 1.3. En este caso, lo natural es en primer lugar usar los recursos de la estadística para identificar las propiedades del dado. Luego, podemos usar probabilidades para predecir el comportamiento del dado en el futuro. En este curso nos dedicaremos también a preguntas de este tipo, en donde realizamos ambos pasos de forma simultánea.

Ejemplo 1.3 (probabilidades y estadística)**.**

Considere que de N lanzamientos de un dado, el cuál no sabemos si está cargado o no, se han obtenido cantidades $s_1, s_2, s_3, s_4, s_5, s_6$ de 1's, 2's, 3's, 4's, 5's y 6's respectivamente.

¿Cuál es la probabilidad de obtener un 4 en los siguientes dos lanzamientos?

Lo anterior entonces nos deja en posición para bosquejar lo que puede ser una definición de la Estadística.

1.2. ¿Qué es la estadística?

Si bien hay un sinnúmero de definiciones, en base a lo postulado por David Spiegelhalter, consideraremos que la estadística es:

Un conjunto de principios y procedimientos para obtener y procesar evidencia cuantitativa para apoyar la toma de decisiones, hacer juicios, entender fenómenos naturales y hacer predicciones

Con lo que la Estadística no es únicamente *análisis de datos*, sino que también considera: diseño de experimentos, exploración de datos de forma gráfica, interpretación informal de datos, análisis formal estadístico, comunicación de resultados de forma clara, modelación y presentación de incertidumbre.

Desde un punto de vista más conceptual, podemos entender la estadística como una forma de razonamiento inductivo. Recordemos que una desventaja del razonamiento/lógica deductivo/a es que de alguna forma todas las consecuencias están incluidas en las premisas, con lo que, *uno no aprende nada*. El razonamiento inductivo por el contrario, nos permite aprender de observaciones de nuestro entorno, de forma empírica, a costa de no tener seguridad de lo que aprendemos. En este sentido, podemos entender las probabilidades como un ejemplo de lógica inductiva y la estadística como deductiva, donde mediante la modelación de la incertidumbre, la estadística representa un entorno para estudiar el problema de inducción: generalizar en base a observaciones. Podemos adelantar que no es posible aprender solo de observaciones pero sí disminuir nuestra incertidumbre, en particular, las observaciones solo nos permiten descartar hechos con seguridad, mas nunca confirmarlos, como se ilustra en la siguiente cita.

> Los animales domésticos esperan su alimento cuando ven la persona que habitualmente se lo da. Sabemos que todas estas expectativas, más bien burdas, de uniformidad, están sujetas a error. El hombre que daba de comer todos los días al pollo, a la postre le tuerce el cuello, demostrando con ello que hubiesen sido útiles al pollo opiniones más afinadas sobre la uniformidad de la naturaleza. (Bertrand Russell Los problemas de la filosofía)

Finalmente, hemos mencionado varias veces el concepto *aprender* durante esta sección. Esto es porque la Estadística ha sido instrumental en el desarrollo del Aprendizaje de Máquinas (AM), una componente de la Inteligencia Artificial que permite construir sistemas inteligentes de forma autónoma (sin la necesidad de que que éstos sean explícitamente programados). En su objetivo, el AM permite que estas máquinas *aprendan* del mundo mendiante observaciones donde los modelos estadísticos y el uso de datos para su ajuste es fundamental, desde ahí el rol de la estadística en el AM y el uso del término *aprender* (modelos) como una alternativa

al término más clásico *ajustar*. En el mismo contexto, la estadística ha jugado un rol preponderante en disciplinas como minería de datos, Big Data, ciencia/análisis de datos y tantos otros.

1.3. Enfoques frecuentista y Bayesiano

La estadística moderna considera principalmente dos enfoques distintos para abordar el problema de inferencia, ambos enfoques son complementarios. Su diferencia fundamental reside en el significado que cada uno le da a la probabilidad.

El primero de ellos es el enfoque clásico conocido como **frecuentista**. En este enfoque, la probabilidad adquiere el significado al que estamos acostumbrados: Casos favorables divido en casos totales. Teniendo esto en cuenta, el enfoque frecuentista define la probabilidad como una *frecuencia límite*, es decir, la probabilidad de un evento es la razón entre las veces que ocurre y el total de las veces, cuando éste último tiene a infinito. Dos características directas de esta definición son que i) las probabilidad de ocurrencia de un hecho depende de la naturaleza de éste, y ii) no tiene sentido definir probabilidades de eventos que son irrepetibles.

Las herramientas frecuentistas fueron desarrolladas hasta inicios del siglo pasado, como respuesta al tratamiento informal de las probabilidades existente hasta ese entonces, y su introducción fue muy exitosa en el sentido de equipar a las probabilidades con tratamiento matemático riguroso. Sin embargo, el enfoque frecuentitsta tiene limitantes, además de los dos puntos mencionados arriba, un problema relacionado con este enfoque es que no brinda un tratamiento natural para el problema de inferencia que permita incluir incertidumbre o sesgos del observador, como por ejemplo el *conocimiento experto*.

El segundo enfoque es el **Bayesiano**, el que si bien data de antes de la introducción del tratamiento formal del frecuentismo, recientemente ha sido retomado y complementado con los avances teóricos frecuentistas. El paradigma bayesiano postula que la probabilidad es una medida de incertidumbre (y no de frecuencia límite) o grado de creencia en la ocurrencia de un evento. Consecuentemente, este enfoque es subjetivo, pues la incertidumbre está en los ojos del observador, y además es perfectamente correcto definir probabilidades sobre hechos que no son repetibles.

En resumen, el enfoque clásico o *frecuentista*, asume lo siguiente:

- El concepto de probabilidad está relacionado con frecuencias límites, es decir, la probabilidad de un evento es la razón de veces que este ocurre versus las veces que no ocurre (usualmente referido como *casos favorables dividido por casos totales*). En este sentido, la probabilidad es una propiedad del mundo real.
- Los parámetros son constantes (fijos) y desconocidos, es decir, no existe *aleatoriedad* relacionada a los parámetros, por ende no podemos construir enunciados probabilísticos con respecto a ellos
- El procedimiento estadístico debe comportarse bien en el largo plazo, un ejemplo de esto

es que un $(1-\alpha)$-intervalo de confianza debe capturar (asintóticamente) el parámetro una fracción $1-\alpha$ de las veces luego de infinitos experimentos.

Por otro lado, el **enfoque bayesiano** se caracteriza por lo siguiente:

- La probabilidad es subjetiva y denota un grado de *creencia*, es decir, la aleatoriedad de un evento no solo es intrínseca de éste sino también de nuestra observación
- Lo anterior permite considerar aleatoriedad en los parámetros, pues el hecho de que éstos sean fijos no quiere decir que los conozcamos.
- Podemos considerar los parámetros como VAs y, consecuentemente, calcular su distribución de probabilidad. Inferencias puntuales o la incidencia de este parámetro en otras VAs está completamente determinada por su distribución.

Existen ventajas y desventajas para ambos enfoques, lo cual hace que ambos sean considerados en distintas aplicaciones. Si bien el enfoque bayesiano es muy antiguo, la estadística clásica ha privilegiado un punto de vista frecuentista, mientras que disciplinas como minería de datos y aprendizaje de máquinas se inclinan por el enfoque bayesiano. De todas formas actualmente ambos métodos se consideran en base a sus propios méritos.

Capítulo 2

Repaso de Probabilidades

Este capítulo contiene un breve repaso sobre conceptos claves en la teoría de probabilidad, los cuales deben ser dominados para entender los contenidos del curso. Además, se hará una introducción a la teoría de la información, dando definiciones y teoremas claves de esta.

2.1. Notación y conceptos básicos

2.1.1. Espacio de probabilidad

Sea Ω un conjunto cualquiera, se considera a $\mathcal{P}(\Omega)$ el conjunto potencia de Ω dado por $\mathcal{P}(\Omega) = \{A : A \subseteq \Omega\}$, es decir, el conjunto de todos los subconjuntos de Ω. Notar que el conjunto vacío $\{\} = \phi$ siempre es un elemento en $\mathcal{P}(\Omega)$. También nos referimos a $\mathcal{P}(\Omega)$ como "las partes de Ω."

Definición 2.1 ($\sigma-$álgebra)**.**
Una $\sigma-$álgebra es un subconjunto $\Sigma \subseteq \mathcal{P}(\Omega)$ con las siguientes propiedades:

1. $\Omega \in \Sigma$, $\phi \in \Sigma$,
2. Si $A \in \Sigma$, entonces $A^c \in \Sigma$,
3. Si $\{A_i\}_{i\in\mathbb{N}} \subseteq \Sigma$, entonces $\bigcup_{i\in\mathbb{N}} A_i \in \Sigma$.

En adelante, se considera a Σ una $\sigma-$álgebra sobre Ω. Además, a la dupla (Ω, Σ) se le denomina *espacio medible*.

Definición 2.2 (Medida)**.**
Se dice que una función $\mu : \Sigma \to \mathbb{R}$ es una medida sobre Σ si es que μ cumple que:

1. Positividad: $\forall A \in \Sigma$, $\mu(A) \geq 0$,
2. $\mu(\phi) = 0$,
3. σ-aditividad: Si $\{A_i\}_{i\in\mathbb{N}}$ es un conjunto disjunto, es decir, que $A_i \cap A_j = \phi, \forall i \neq j$,

entonces:

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

En el caso $\mu(\Omega) = 1$, μ es una *medida de probabilidad*.

Antes de introducir lo que entenderemos como espacio de probabilidad, es necesario entender lo que es la σ-álgebra de Borel. Para esto, necesitamos la siguiente definición:

Definición 2.3 (σ-álgebra generada por un conjunto)**.**
Sea $\mathcal{L} \subseteq \mathcal{P}(\Omega)$ una familia de partes de Ω. Consideremos a $\mathcal{J}$ como $\mathcal{J}(\mathcal{L}) = \{\mathcal{B} : \mathcal{B}$ es σ-álgebra en $\Omega, \mathcal{L} \subseteq \mathcal{B}\}$, es decir, $\mathcal{J}(\mathcal{L})$ es el conjunto de todas las σ-álgebras sobre Ω que contienen al conjunto $\mathcal{J}$. La σ-álgebra *generada por un conjunto* $\mathcal{L}$, $\sigma(\mathcal{L})$ está dada por:

$$\sigma(\mathcal{L}) = \bigcap_{\mathcal{B}\in\mathcal{J}} \mathcal{B}.$$

Basicamente, $\sigma(\mathcal{L})$ se entiende como la *σ-álgebra más pequeña que contiene a $\mathcal{L}$*. Es decir, si σ_0 es cualquier $\sigma-$álgebra que contiene a $\mathcal{L}$, entonces necesariamente $\sigma(\mathcal{L}) \subseteq \sigma_0$.

En el caso que $\Omega = \mathbb{R}$, se dotará de este espacio de una σ-álgebra especial, la σ-álgebra de los *Borelianos*, la cual es clave en la teoría de probabilidades. Si se considera que $\mathcal{L} = \{(-\infty, x] : x \in \mathbb{R}\}$, la σ-álgebra de Borel $\sigma(\mathcal{L})$ se denotará por $\mathbb{B}(\mathbb{R})$ o simplemente por $\mathbb{B}$. A los elementos de $\mathbb{B}$ se le llaman los *borelianos*.

Consideremos a μ una medida, a la tripleta (Ω, Σ, μ) se le denomina espacio de medida, en el caso de que μ sea una medida de probabilidad, entonces (Ω, Σ, μ) es un espacio de probabilidad. En un espacio de probabilidad, a los elementos $\omega \in \Omega$ se le denominan eventos. Se denotará a las medidas de probabilidad como $\mathbb{P}$.

2.1.2. Propiedades de un espacio de probabilidad

Consideremos a $(\Omega, \Sigma, \mathbb{P})$ un espacio de probabilidad y a $A, B \in \Sigma$. Se verifican las siguientes propiedades:

1. Se cumple que $\mathbb{P}(B\backslash A) = \mathbb{P}(B) - \mathbb{P}(B \cap A)$.
2. $\mathbb{P}$ es creciente, esto es:
$$A \subseteq B \implies \mathbb{P}(A) \leq \mathbb{P}(B).$$
3. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
4. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

2.1.3. Probabilidad condicional

En adelante, se considera a $(\Omega, \Sigma, \mathbb{P})$ un espacio de probabilidad fijo.

Definición 2.4.
Sean $A, B \in \Sigma$ con $\mathbb{P}(B) > 0$. La probabilidad condicional de A dado B es:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Es directo verificar las siguientes propiedades:

1. Sean $A, B \in \Sigma$ con $\mathbb{P}(A) > 0$ y $\mathbb{P}(B) > 0$. Se tiene que:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)},$$

   lo cual es conocido como el Teorema de Bayes.

2. Si $\{A_n : n \in \mathbb{N}\} \subseteq \Sigma$ es una familia disjunta, entonces se tiene que:

$$\mathbb{P}\left(\bigcup_{n\in\mathbb{N}} A_n|B\right) = \sum_{n\in\mathbb{N}} \mathbb{P}(A_n|B).$$

Para más propiedades revisar: G. Grimmett & D. Stirzaker, *Probability and Random Processes*, Oxford Press, 2001.

2.2. Teoría de la Información

La teoría de la información es el estudio científico de las leyes matemáticas que rigen la transmisión y el procesamiento de la información, además de *medir* la información y la representación de la misma. Esta teoría, tiene intersección con muchos campos, desde el lenguaje y criptografía, hasta las probabilidades. En esta sección, se introducirán conceptos y definiciones básicas de la teoría de la información, como la entropía y la información mutua.

2.2.1. Entropía

Consideramos a X e Y dos variables aleatorias (VA) discretas con función de densidad p_X y q_Y y posibles valores $\mathcal{X} = x_1, x_2, ...$ e $\mathcal{Y} = y_1, y_2, \ldots$. La Entropía de la VA X se define de la siguiente forma:

$$\begin{aligned} H(X) &= \mathbb{E}[-\log p_X] \\ &= -\sum_{n\geq 1} p_X(x_n)\log p_X(x_n). \end{aligned}$$

La entropía, en analogía a la termodinámica, es una medida de *desorden* en los valores que puede tomar una VA, es decir, mientras menos predecible es el valor de la VA, mayor es su entropía. Se dice también que la Entropía es la cantidad de *bits* necesarios para describir X.

Se definen las entropías condicionales de X dado Y como:

$$\begin{aligned} H(X|Y=y) &= -\sum_n p_{x_n|y}\log(p_{X|Y}(x_n|y)) \\ H(X|Y) &= -\sum_i \sum_j p_{X,Y}(i,j)\log(p_{X|Y}(i|j)), \end{aligned}$$

se define análogamente a $H(X,Y)$, la entropía conjunta considerando la densidad conjunta $p_{X,Y}$.

Propiedad 2.1 (Propiedades de la entropía)**.**
Algunas propiedades de la entropía son las siguientes:

- $0 \leq H(X) \leq \log(|\mathcal{X}|)$,
- $0 \leq H(X|Y) \leq H(X)$.

2.2.2. Información mutua

La información mutua I entre X e Y se define como:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= \sum_{i,j} p_{X,Y}(i,j)\log\frac{p_{X,Y}(i,j)}{p_X(i)p_Y(j)}, \end{aligned}$$

la cual tiene las siguientes propiedades:

Propiedad 2.2 (Propiedades Información Mutua)**.**
Sean X e Y dos VAs, tenemos

1. No negatividad: $I(X;Y) \geq 0$
2. Simetría: $I(X;Y) = I(Y;X)$
3. Regla de la cadena: $I(X;Y,Z) = I(X;Z) + I(X;Y|Z)$

2.2.3. Relación entre Entropía e Información Mutua

Algunas de las relaciones que existen entre la información mutua y la entropía son las siguientes:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \\ &= H(X,Y) - H(Y|X) - H(X|Y) \end{aligned}$$

las cuales, visualmente se obtienen dadas la figura 2.1.

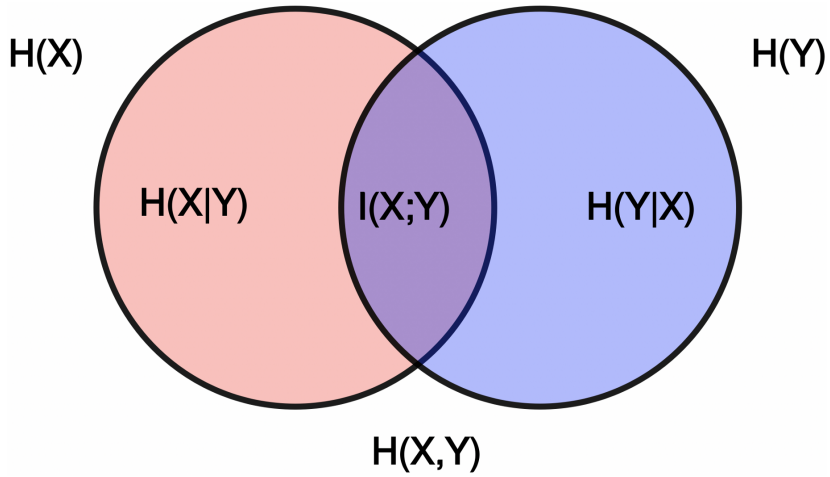

Figura 2.1: Relación entre H e I

2.2.4. Divergencia de Kullback-Leibler

Definición 2.5 (Divergencia de Kullback-Liebler)**.**
Para dos densidades de probabilidad f y g, definidas sobre un mismo conjunto de partida $\mathcal{X}$, la divergencia de Kullback-Leibler entre ellas está definida mediante

$$\mathrm{KL}\left(f \| g\right) = \int_{\mathcal{X}} f(x) \log\left(\frac{f(x)}{g(x)}\right) \mathrm{d}x. \tag{2.1}$$

Observación 2.1.

La divergencia KL es siempre positiva $\forall f, g$ (desigualdad de Gibbs):

$$\begin{aligned} -\mathrm{KL}\left(f\|g\right) &= \int_{\mathcal{X}} f(x) \log\left(\frac{g(x)}{f(x)}\right) \mathrm{d}x \\ &\leq \log\left(\int_{\mathcal{X}} f(x) \frac{g(x)}{f(x)}\right) \mathrm{d}x, \qquad \text{(Jensen's)} \\ &= \log\left(\int_{\mathcal{X}} g(x)\right) \mathrm{d}x \\ &= \log 1 = 0., \end{aligned}$$

Además, como $\log(\cdot)$ es estrictamente convexo, la igualdad $\mathrm{KL}\left(f\|g\right) = 0$ solo se cumple si el argumento $\frac{g(x)}{f(x)}$ es constante, lo cual se tiene solo para $g(x) = f(x)$.

Otra propiedad clave de la divergencia KL es que puede ser infinita y es asimétrica, por esta razón nos referimos a KL como divergencia y no *distancia*. La intuición detrás de la KL es que es una medida de *error* de estimar la densidad f mediante la densidad g. Algunas propiedades son las siguientes:

Propiedad 2.3. 1. Relación con la información mutua:

$$I(X;Y) = KL(p_{X,Y}||p_X p_Y)$$

2.2.5. Data Processing Inequality

Teorema 2.1 (Data Processing Inequality (DPI))**.**

Sea una cadena de markov dada por las variables aleatorias $X \to Y \to Z$, se tiene que:

$$I(X;Y) \geq I(X;Z)$$

la cual se llama Data Processing Inequality (DPI), con igualdad si y solo si $X \to Z \to Y$ es una cadena de Markov

Como corolario directo de la DPI, se tiene que para cualquier función determinista $g(\cdot)$ se cumple que:

$$I(X;Y) \geq I(X;g(Y))$$

y entonces no existe ninguna función de Y la cual pueda aumentar la información que Y contiene sobre X.

Capítulo 3

Primeros Conceptos

3.1. Modelos estadísticos

En este curso, en particular, nos enfocaremos en *estadística matemática*, lo cual provee inferencia estadística formal basada en herramientas de probabilidades, álgebra y teoría de la medida. Para esto asumiremos que tenemos datos generados desde un modelo estadístico (o probabilístico, o *generativo*) desconocido, donde nuestro objetivo es usar estos datos para determnar dichos modelos con el fin último de aprender sobre el mecanismo subyacente de la generación de datos y hacer predicciones (usando el modelo aprendido). El primer paso para lograr este objetivo es definir el *Modelo Estadístico.*

Definición 3.1 (Modelo estadístico)**.**

Un modelo estadístico es un conjunto de distribuciones de probabilidad, que pueden ser consideradas como *candidatas* para el mecanismo de generación de datos.

En algunos casos, las distribuciones de ese conjunto pueden ser expresadas mediante parámetros, por ejemplo, en el caso de la distribución normal, expresada mediante su media y su varianza. En dichos casos, el objetivo de descubrir el mecanismo de generación de datos (la distribución) es simplemente descubrir sus parámetros. El objetivo entonces de definir el modelo estadístico (paramétrico o no) es delinear los posibles representaciones para el mecanismo de generación de datos y, en base a los datos y algún criterio de eficiencia, encontrar el(los) modelo(s) apropiado(s). En este contexto, antes de encontrar ese modelo, consideramos que nuestro modelo tiene *parámetros desconocidos.*

En el trayecto del curso, asumiremos que disponemos de un conjunto de datos x, que pertenece a un espacio abstracto $\mathfrak{X}$, donde típicamente $\mathfrak{X} = \mathbb{R}^n$; aunque también podremos tener datos funcionales, como por ejemplo $\mathfrak{X} = \{f : [0,1] \to \mathbb{R}\}$. Asumiremos entonces que x es la realización de una variable aleatoria $X \in \mathfrak{X}$; con lo que implícitamente asumimos que $\mathfrak{X}$ es un espacio medible con su respectiva σ-álgebra. Podemos entender nuestro modelo estadístico como el espacio de posibles hipótesis que explican los datos observados. En este sentido, una de las preguntas que debemos poder responder es ¿Cuál es la ley de X?, es decir, ¿ Cómo calcular $\mathbb{P}(X \in A)$ donde $A \in \beta(\mathfrak{X})$?, con $\beta(\mathfrak{X})$ los borelianos de $\mathfrak{X}$

Nos enfocaremos en modelos paramétricos, con lo cual para es necesario definir formalmente

los parámetros y el espacio de éstos.

Definición 3.2 (Parámetro y Espacio de Parámetros)**.**
En un problema de inferencia estadística, la (o las) característica(s) que determinan la distribución de las variables aleatorias estudiadas son llamadas parámetros. El conjunto Ω de todos los posibles valores de los parámetros se llama espacio de parámetros.

Regresando a la pregunta, no habrá una, sino muchas posibles medidas de probabilidad como candidatas a ser la ley de X. A esto nos referíamos arriba cuando mencionamos la familia paramétrica de probabilidades donde cada una de las cuales puede ser la que actúa para generar x a través de X. Encontrar la (o las) distribuciones, dentro de este conjunto, que son mas representativas de haber generados los datos, es un objetivo de inferencia estadística.

Denotaremos a la familia paramétrica $\mathcal{P}$ de la siguiente forma:

$$\mathcal{P} = \{\mathcal{P}_\theta | \theta \in \Omega, \}$$

donde $\mathcal{P}_\theta$ es una medida de probabilidad bajo un parámetro $\theta \in \Omega$ en el espacio de parámetros. En nuestro estudio (pero en general no tiene que ser así) consideraremos que Ω es finito dimensional, es decir, $\Omega \subseteq \mathbb{R}^n$. Escribimos entonces que:

$$\theta = (\theta_1, ..., \theta_n).$$

Dado todo lo anterior, en la formulación de un modelo estadístico completo para representar un fenómeno se debiese tener lo siguiente plenamente identificado lo siguiente:

- θ como parámetro a estimar
- Ω espacio de parámetros con $\Omega \subseteq \mathbb{R}^n$
- $\mathcal{P}_\theta$ probabilidad sobre $\mathfrak{X}$ (como función de θ)
- X vector aleatorio con valores en $\mathfrak{X}$
- x elemento genérico de $\mathfrak{X}$ y realización de X (datos).

Ejemplo 3.1 (Fábrica de computadores)**.**
Una compañía de fabricación de computadores desea estimar el tiempo de vida de un componente particular en sus computadores. Para ello, en primer lugar se recolectan datos de los computadores que se han usado bajo condiciones normales. Luego de ser asesorados por expertos, deciden usar una distribución normal para modelar el tiempo que se demorará un componente en fallar. Se busca modelar todos los componentes con un tiempo de vida promedio θ y varianza σ^2, con θ y σ^2 parámetros desconocidos. Si se tienen N componentes, las variables aleatorias que modelan la vida útil de cada componente serán identificadas como $X_1, ..X_N$, con $X_i \sim \mathcal{N}(\theta, \sigma^2)$. ¿Qué opina de este modelo?

La inferencia estadística es una herramienta que nos permitirá resolver muchos tipos de problemas. Los más importantes serán los de *identificación*, donde nuestro objetivo es descubrir

el modelo que genero los datos, y *predicción* donde se intenta estimar una cantidad que no ha sido observada aún. Por supuesto, buscamos alcanzar ambos objetivos de forma estadística, es decir, modelando apropiadamente la incertidumbre asociada.

3.2. Construcción de variables aleatorias mediante transformaciones

3.2.1. Chi-Cuadrado

Sean $Z_1, ..., Z_n$ variables aleatorias con $Z_i \sim \mathcal{N}(0,1)$ independientes, se tiene que la suma de sus cuadrados

$$Q = \sum_{i=1}^{n} Z_i^2$$

distribuye según una distribución *Chi-Cuadrado* con n grados de libertad. Esto lo denotamos como:

$$Q \sim \mathcal{X}_n^2$$

Algunas propiedades de esta distribución:

a. $\mathbb{E}(\mathcal{X}_n^2) = n$

b. $\mathbb{V}(\mathcal{X}_n^2) = 2n$

c. $M_{\mathcal{X}_n^2}(t) = (1-2t)^{\frac{-n}{2}}$ con $t < \frac{1}{2}$

3.2.2. T-Student

Sea $X_1, ..., X_n$ una muestra i.i.d de una variable aleatoria $X \sim \mathcal{N}(\mu, \sigma^2)$. Consideremos el promedio muestral:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

y la varianza muestral insesgada:

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

Entonces:

$$\frac{\overline{X}_n - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

siendo t_{n-1} la distribución t-student con $n-1$ grados de libertad.

Capítulo 4

Estadísticos

Recordemos que en la aplicación de la estadística, además de nuestros supuestos, solo contamos con *datos*, consecuentemente, todo lo que hagamos partirá desde el uso de éstos. En este sentido, definimos un estadístico como una función de (las realizaciones de) una variable aleatoria, definida desde el espacio muestral. Es decir, cualquier función *medible* de los datos. Recordar que Θ es el espacio de parámetros.

Definición 4.1 (Estadístico)**.**
Sea $(\mathcal{T}, \mathcal{A}, \mu)$ un espacio de probabilidad y $X \in \mathcal{X}$ una variable aleatoria con distribución paramétrica $\mathcal{P} = \{P_\theta \text{ t.q. } \theta \in \Theta\}$. Un estadístico es una función medible de la realización $X = x$ independiente del parámetro θ (y de la distribución P_θ).

$$\begin{aligned} T : \mathcal{X} &\to \mathcal{T} \\ x &\mapsto T(x). \end{aligned}$$

Observación 4.1.
Es muy relevante diferenciar el valor del estadístico $T(x)$ como función de los datos (considerados por nosotros como la realización $X = x$ de la variable aleatoria), de la aplicación de la función $T(\cdot)$ a la variable aleatoria X, es decir, $T(X)$. El primero es un valor "fijo"mientras que el segundo es una VA con propia distribución de probabilidad inducida por P_θ y por la función T (llamada distribución *pushforward* $T_\sharp P_\theta$).

En base a los datos x, algunos estadísticos pueden ser:

$$T(x) = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad T'(x) = x, \qquad T''(x) = \text{mín}(x), \qquad T'''(x) = c \in \mathbb{R}.$$

4.1. Suficiencia

En términos generales, el objetivo de un estadístico es *encapsular* o *resumir* la información contenida en una muestra (de datos) $x = (x_1, x_2, \ldots, x_n)$ que es de utilidad para determinar

(o estimar) un/el/los parámetros de la distribución de X o alguna otra propiedad de ésta. Por esta razón, la función identidad o el promedio parecen cumplir, al menos intuitivamente, con esta misión. Esto es por que se intuitivamente queremos extraer la mayor información posible de la data, esto lo logran el estadístico T (que resume todos los datos) y el estadístico T' (que contiene todos los datos). Por el contrario, notemos que el estadístico T'' *pierde información*, dado que solo se extrae el mínimo valor de toda la data obtenida, así, perdiendo la representación de la, e.g., dispersión de la muestra. El mismo análisis se puede hacer para el estadístico constante, el que no contiene información alguna de los datos.

Coloquialmente, la idea de suficiencia de un estadístico (con respecto a un parámetro) puede ser expresada como

> *"... no other statistic that can be calculated from the same sample provides any additional information as to the value of the parameter." (Ronald Fisher, On the mathematical foundations of theoretical statistics)*

Formalmente, definimos un estadístico mediante.

Definición 4.2 (Estadístico Suficiente)**.**
Sea $(S, \mathcal{A}, \mu)$ un espacio de probabilidad y $X \in \mathcal{X}$ una variable aleatoria con distribución paramétrica $\mathcal{P} = \{P_\theta \text{ t.q. } \theta \in \Theta\}$. Diremos que la función $T : \mathcal{X} \to \mathcal{T}$ es un estadístico suficiente para θ (o para X o para $\mathcal{P}$) si la ley condicional $X|T(X)$ no depende del parámetro θ, es decir,

$$P_\theta(X \in A|T(X)),\ A \in \mathcal{B}(X), \text{no depende de } \theta.$$

Observemos entonces que si $T(X)$ es un estadístico suficiente, entonces, existe una función que

$$H(\cdot,\cdot) : \mathcal{B}(X) \times \mathcal{T} \to [0,1]$$

que es una distribución de probabilidad en el primer argumento y es medible en el segundo argumento.

Para poder entender mejor el concepto de un Estadístico Suficiente, se dan los siguientes ejemplos:

Ejemplo 4.1 (Estadístico suficiente trivial)**.**
Para cualquier familia paramétrica $\mathcal{P}$, el estadístico definido por

$$T(x) = x$$

es suficiente. En efecto, $P_\theta(X \in A|X = x) = \mathbb{1}_A(x)$ no depende del parámetro de la familia.

Ejemplo 4.2 (Estadístico suficiente Bernoulli)**.**

Sea $x = (x_1, \ldots, x_n) \sim Ber(\theta)$, $\theta \in \Theta = [0, 1]$, es decir

$$P_\theta(X = x) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}.$$

Veamos que $T(x) = \sum_{i=1}^{n} x_i$ es un estadístico suficiente (por definición). En efecto

$$\begin{aligned} P(X = x|T(X) = t) &= \frac{P(T(X) = t|X = x)P(X = x)}{P(T(X) = t)} && \text{(T. Bayes)} \\ &= \frac{\mathbb{1}_{T(x)=t}\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} && \text{(modelo y suma de Bernoulli es Binomial)} \\ &= \mathbb{1}_{T(x)=t}\binom{n}{t}^{-1} && \text{(pues } T(x) = t\text{)} \end{aligned}$$

Consecuentemente, $T(x) = \sum_{i=1}^{n} x_i$ es estadístico suficiente.

Intuitivamente, nos gustaría poder verificar directamente de la suficiencia de un estadístico desde la distribución (o densidad) de una VA, o al menos verificar una condición más simple que la definición. Esto es porque verificar la no-dependencia de la distribución condicional $P(X|T)$ puede ser no trivial, engorroso o tedioso. Para esto enunciaremos el Teorema de Fisher-Neyman, el cual primero requiere revisar la siguiente definición.

Definición 4.3 (Familia Dominada)**.**

Una familia de modelos paramétricos $\mathcal{P} = \{P_\theta \text{ t.q. } \theta \in \Theta\}$ es dominada si existe una medida μ, tal que $\forall\theta \in \Theta, P_\theta$ es absolutamente continua con respecto a μ (denotado $P_\theta \ll \mu$), es decir,

$$\forall\theta \in \Theta, A \in \mathcal{B}(X), \mu(A) = 0 \Rightarrow P_\theta(A) = 0.$$

La definición anterior puede interpretarse de la siguiente forma: si una familia de paramétrica $\mathcal{P} = \{P_\theta \text{ t.q. } \theta \in \Theta\}$ es dominada por una medida μ, entonces ninguno de los modelos $P_\theta \in \mathcal{P}$ puede asignar medida (probabilidad) no nula a conjuntos que tienen medida cero bajo μ (la medida *dominante*). Una consecuencia fundamental de que la distribución P_θ esté dominada por μ está dada por el Teorema de Radon–Nikodym, el cual establece que si $P_\theta \ll \mu$, entonces la distribución P_θ tiene una densidad con respecto a μ, es decir,

$$\forall A \in \mathcal{B}(X), P_\theta(X \in A) = \int_A p_\theta(x)\mu(\mathrm{d}x),$$

donde $p_\theta(x)$ es conocida como la densidad de P_θ con respecto a θ (o también como la derivada de Radon–Nikodym $\frac{dP_\theta}{d\mu}$).

Con la noción de Familia Dominada y de densidad de probabilidad, podemos enunciar el siguiente teorema importante y fundamental que conecta la forma de la densidad de un

modelo paramétrico con la suficiencia de su estadístico.

Teorema 4.1 (Factorización, Neyman-Fisher)**.**
Sea $\mathcal{P} = \{P_\theta \text{ t.q. } \theta \in \Theta\}$ una familia dominada por μ, con p_θ la densidad de P_θ. Entonces, T es un estadístico suficiente si y solo si existen funciones apropiadas $g_\theta(\cdot)$ y $h(\cdot)$, i.e., medibles y no-negativas, tal que la densidad p_θ, $\theta \in \Omega$, admite la siguiente factorización:

$$p_\theta(x) = g_\theta(T(x))h(x). \tag{4.1}$$

Lo anterior se debe cumplir $\forall x \in \mathfrak{X}$ y $\mu - ctp$. También, se tiene que esto es condición necesaria y suficiente para decir que $T(X)$ es suficiente.

El Teorema de Neyman-Fisher es clave para evaluar, directamente de la densidad de un modelo, la suficiencia de un estadístico. Pues al identificar la expresión de la V.A. que interactúa con el parámetro (en la función g_θ) es posible determinar el estadístico suficiente. Antes de ver una demostración informal del Teorema 4.1, revisemos un par de ejemplos.

Ejemplo 4.3 (Factorización Bernoulli)**.**
Notemos que la densidad de Bernoulli (que es igual a su distribución por ser un modelo discreto) factoriza tal como se describe en el Teorema 4.1. En efecto, consideremos $x = (x_1, \ldots, x_n) \sim \text{Bernoulli}(\theta)$ y el estadístico $T(x) = \sum x_i$, entonces,

$$\mathbb{P}(X = x) = \underbrace{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}}_{g_\theta(T(x))} \cdot \underbrace{1}_{h(x)} \tag{4.2}$$

Con lo anterior, se tiene que $g_\theta(T(x))$ y $h(x)$ cumplen que son medibles no negativas con lo cual se cumplen las hipótesis del Teorema de Neyman-Fisher y entonces $T(X)$ es suficiente

Ejemplo 4.4 (Factorización Normal (varianza conocida))**.**
Consideremos ahora $x = (x_1, \ldots, x_n) \sim \mathcal{N}(\mu, \sigma^2)$, con σ^2 conocido y el estadístico

$T(x) = \frac{1}{n}\sum x_i$, entonces,

$$\begin{aligned}
p(X = x) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right) \\
&= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right) \\
&= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}((x_i - \bar{x}) + (\bar{x} - \mu))^2\right) \\
&= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + 2\cancel{(x_i - \bar{x})}(\bar{x} - \mu) + (\bar{x} - \mu)^2\right) \\
&= \underbrace{(2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)}_{h(x)} \underbrace{\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\bar{x} - \mu)^2\right)}_{g_\theta(T(x))}
\end{aligned}$$

Notamos que $h(x)$ no depende de los parámetros y solo depende de los datos, en cambio, $g_\theta(T(x))$ depende del parámetro y es función del estadístico $T(X)$. Nuevamente se cumplen que las funciones anteriores son medibles no-negativas y se cumplen las hipótesis del Teorema de Neyman-Fisher y entonces $T(X)$ es un estadístico suficiente.

A continuación, veremos la prueba del Teorema 4.1 para el caso discreto.

Demostración de Teorema Neyman-Fisher, caso discreto. Primero probamos la implicancia hacia la derecha ($\Rightarrow$), es decir, asumiendo que $T(X)$ es un estadístico suficiente, tenemos,

$$\begin{aligned}
p_\theta(X = x) &= P_\theta(X = x, T(X) = T(x)) \\
&= \underbrace{P_\theta(X = x | T(X) = T(x))}_{h(x),\text{ no depende de } \theta \text{ por hipótesis}} \underbrace{P_\theta(T(X) = T(x))}_{g_\theta(T(x))},
\end{aligned}$$

es decir, la factorización deseada.

Ahora probamos la implicancia hacia la izquierda ($\Leftarrow$), es decir, asumimos la factorización en la ecuación (4.1). En primer lugar, tenemos que el modelo se puede escribir como (Bayes)

$$p_\theta(X = x | T(X) = t) = \frac{p_\theta(T(X) = t | X = x) p_\theta(X = x)}{p_\theta(T(X) = t)}.$$

Donde $p_\theta(T(X) = t | X = x) = \mathbb{1}_{T(x)=t}$ y la hipótesis esto nos permite escribir

$$\begin{aligned}
p_\theta(X = x) &= g_\theta(T(x))h(x) \\
p_\theta(T(X) = t) &= \sum_{x';T(x')=t} p_\theta(X = x') = \sum_{x';T(x')=t} g_\theta(T(x'))h(x')
\end{aligned}$$

Incluyendo estas últimas dos expresiones en la ec. (4.1), tenemos

$$p_\theta(X = x|T(X) = t) = \frac{\mathbb{1}_{T(x)=t}\cancel{g_\theta(T(x))}h(x)}{\sum_{x';T(x')=t}\cancel{g_\theta(T(x'))}h(x')} = \frac{\mathbb{1}_{T(x)=t}h(x)}{\sum_{x';T(x')=t}h(x')} \tag{4.3}$$

donde los términos que se cancelan son todos iguales a $g_\theta(t)$.

Finalmente, como el lado derecho de la ecuación (4.3) no depende de θ, se concluye la demostración. ■

4.1.1. Particiones Suficientes

Un estadístico induce una partición en el conjunto de *outcomes* posibles. Es posible estudiar la suficiencia en términos de estas particiones, dadas por $\{x|T(x) = t\}$, para cada t.

Definición 4.4.
Una partición $\{B_1, ...B_k\}$ se dice suficiente si $f(x|x \in B_i)$ no depende de θ.

Teorema 4.2.
Un estadístico es suficiente si y sólo si la partición que induce es suficiente.

Demostración. Ejercicio. ■

Ejemplo 4.5.
Sean $X_1, X_2, X_3 \sim Bernoulli(\theta)$. Sea $T = \sum_i X_i$. En el Cuadro 4.1 se puede observar los *outcomes* y los estadísticos:

x^n	t	$p(x\|t)$
$(0,0,0)$	$t=0$	1
$(0,0,1)$	$t=1$	$1/3$
$(0,1,0)$	$t=1$	$1/3$
$(1,0,0)$	$t=1$	$1/3$
$(0,1,1)$	$t=2$	$1/3$
$(1,0,1)$	$t=2$	$1/3$
$(1,1,0)$	$t=2$	$1/3$
$(1,1,1)$	$t=3$	1

Cuadro 4.1: *Outcomes* y estadísticos (Bernoulli, $T = \sum_i X_i$)

Notemos que como $p(x|t)$ no depende de θ, T es un estadístico suficiente.

Observación 4.2.
Dos estadísticos pueden generar la misma partición. Por ejemplo T del ejemplo anterior, y $U = 3T$.

Observación 4.3.
Toda partición que refine a una partición suficiente será suficiente.

Veamos un ejemplo de un estadístico que no genera una partición suficiente (y por lo tanto, no es suficiente)

Ejemplo 4.6.
Sean X_1, X_2 y $X_3 \sim Bernoulli(\theta)$. Entonces $T = X_1$ no es suficiente. Veamos su partición en el Cuadro 4.2:

x^n	t	$p(x\|t)$
$(0,0,0)$	$t=0$	$(1-\theta)^2$
$(0,0,1)$	$t=0$	$\theta(1-\theta)$
$(0,1,0)$	$t=0$	$\theta(1-\theta)$
$(0,1,1)$	$t=0$	θ^2
$(1,0,0)$	$t=1$	$(1-\theta)^2$
$(1,0,1)$	$t=1$	$\theta(1-\theta)$
$(1,1,0)$	$t=1$	$\theta(1-\theta)$
$(1,1,1)$	$t=1$	θ^2

Cuadro 4.2: *Outcomes* y estadísticos (Bernoulli, $T = X_1$)

4.2. Suficiencia minimal

La idea de suficiencia del estadístico dice relación, coloquialmente, con la *información* contenida en el estadístico que permite *determinar* el parámetro θ. En ese sentido, se tiene la intuición que un estadístico es suficiente si no existe otro estadístico que pueda determinar de *mejor* formar el parámetro usando los mismos datos. En el extremo de esta intuición de

suficiencia, el estadístico puede ser simplemente todos los datos, i.e, $T(X) = X$ (estadístico trivial), en cuyo caso la suficiencia es directa como se vio en el Ejemplo 4.1. En esta sección, por el contrario, estamos interesados en estadísticos que son suficientes pero que contienen la mínima cantidad de información, pues considerar todos los datos puede ser redundante en cuanto a la determinación del parámetro.

Sin una definición formal de *información* aún, recordemos que los estadísticos representan un resumen o una compresión de los datos mediante la función $T(\cdot)$ medible. En este sentido, la aplicación de dicha función solo puede *quitar* o, a lo sumo, *mantener la información desde la preimagen a la imagen*. Esto nos permite definir el siguiente concepto:

Definición 4.5 (Estadístico Suficiente Minimal)**.**
Un estadístico $T : \mathcal{X} \to \mathcal{T}$ es suficiente minimal si

- $T(X)$ es suficiente, y
- $\forall T'(X)$ estadístico suficiente, existe una función f tal que $T(X) = f(T'(X))$.

Ejemplo 4.7.
Si $X_1, ..X_{2n}$ son observaciones i.i.d de una normal $\mathcal{N}(\theta, 1)$, $\theta \in \mathbb{R}$, entonces:

$$\bar{T} = \begin{pmatrix} \sum_{i=1}^{n} X_i \\ \sum_{i=n+1}^{2n} X_i \end{pmatrix}$$

es suficiente pero no minimal. Se puede demostrar que $T = \sum_{i=1}^{2n} X_i$ es suficente minimal.

Los estadísticos suficiente minimales están claramente definidos pero dicha definición no es útil para encontrar o construir estadístico suficiente minimales. El siguiente Teorema establece una condición que permite evaluar si un estadístico es suficiente minimal

Teorema 4.3 (Suficiencia minimal)**.**
Sea $\mathcal{P} = \{P_\theta \text{ t.q. } \theta \in \Theta\}$ una familia dominada con densidades $\{p_\theta \text{ t.q. } \theta \in \Theta\}$ y asuma que existe un estadístico $T(X)$ tal que para cada $x, y \in \mathcal{X}$:

$$\frac{p_\theta(x)}{p_\theta(y)} \text{ no depende de } \theta \Leftrightarrow T(x) = T(y) \tag{4.4}$$

entones, $T(X)$ es suficiente minimal.

Antes de probar este teorema, veamos un ejemplo aplicado a la distribución de Poisson.

Ejemplo 4.8.
Recordemos que la distribución de Poisson (de parámetro θ) modela la cantidad de eventos en un intervalo de tiempo de la forma y consideremos las observaciones $x =$

$(x_1, \ldots, x_n) \sim Poisson(\theta)$ con densidad

$$p_\theta(x) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \frac{e^{-n\theta}\theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}. \tag{4.5}$$

Notemos que la razón de estas densidades para dos observaciones $x, y \in \mathcal{X}$ toma la forma

$$\frac{p_\theta(x)}{p_\theta(y)} = \frac{\theta^{\sum_{i=1}^n x_i - \sum_{i=1}^n y_i}}{\prod_{i=1}^n x_i! / \prod_{i=1}^n y_i!}, \tag{4.6}$$

lo cual no depende de θ únicamente si $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$, consecuentemente, $T(x) = \sum_{i=1}^n x_i$ es un estadístico suficiente minimal de acuerdo al Teorema 4.3.

Demostración de Teorema 4.3. Recordemos que queremos demostrar que la ec. (4.4) implica que $T(\cdot)$ es estadístico suficiente minimal. Primero veremos que T es suficiente. Dada la partición inducida por el estadístico $T(X)$, para un valor $x \in \mathcal{X}$ consideremos $x_T \in \{x'; T(x') = T(x)\}$, entonces

$$p_\theta(x) = \underbrace{p_\theta(x)/p_\theta(x_T)}_{h(x) \text{ indep. } \theta} \underbrace{p_\theta(x_T)}_{q_\theta(T(x))} \tag{4.7}$$

donde la no dependencia de θ se tiene por el supuesto del Teorema.

Para probar que el estadístico es suficiente minimal, asumamos que existe otro estadístico suficiente $T'(X)$, y consideremos dos valores en el mismo subconjunto de la partición inducida por $T'(X)$, i.e., $x, y \in \mathcal{X}$, t.q. $T'(x) = T'(y)$, y veamos que (mediante la factorización de Neyman-Fisher) podemos escribir la razón de verosimilitudes de la forma

$$\frac{p_\theta(x)}{p_\theta(y)} = \frac{g'_\theta(T'(x))h'(x)}{g'_\theta(T'(y))h'(y)} = \frac{h'(x)}{h'(y)}, \quad \text{pues } T'(x) = T'(y) \tag{4.8}$$

consecuentemente, el enunciado nos permite aseverar que como $\frac{p_\theta(x)}{p_\theta(y)}$ no depende de θ, entonces $T(x) = T(y)$. Es decir, hemos mostrado que $T'(x) = T'(y)$ implica $T(x) = T(y)$, por lo que T es función de T'.

■

4.3. La familia exponencial

Hasta este punto, hemos considerado algunas distribuciones paramétricas, tales como Bernoulli, Gaussiana o Poisson, para ilustrar distintas propiedades y definiciones de los estadísticos. En esta sección, veremos que realmente todas estas distribuciones (y otras más) pueden escribirse de forma unificada. Para esto, consideremos la siguiente expresión llamada

log-normalizador (la razón de este nombre será clarificada en breve).

$$A(\eta) = \log \int_{\mathcal{X}} \exp\left(\sum_{i=1}^{s} \eta_i T_i(x)\right) h(x) \mathrm{d}x, \tag{4.9}$$

donde definimos lo siguiente:

- $\eta = [\eta_1, \ldots, \eta_s]^\top$ es el parámetro natural
- $T = [T_1, \ldots, T_s]^\top$ es un estadístico
- $h(x)$ es una función no-negativa

Definición 4.6 (La Familia Exponencial)**.**
Definamos la siguiente función de densidad de probabilidad parametrizada por $\eta \in \{\eta | A(\eta) < \infty\}$ con $\theta \in \Theta$

$$p_\theta(x) = \exp\left(\sum_{i=1}^{s} \eta_i(\theta) T_i(x) - A(\theta)\right) h(x) \tag{4.10}$$

donde el hecho que $p_\eta(x)$ integra uno puede claramente verificarse reemplazando la ecuación (4.9) en (4.10), con lo cual se puede ver que A definido en (4.9) es precisamente el logaritmo de la constante de normalización de la densidad definida en la ec. (4.10). Equivalemente, y comunmente, se utiliza que la familia exponencial viene dada por funciones de densidad de la forma:

$$p_\theta(x) = \exp\left(\sum_{i=1}^{s} \eta_i(\theta) T_i(x)\right) h(x) g(\theta)$$

donde g es una función no-negativa.

Muchas de las distribuciones que usalmente consideramos pertenecen a la familia exponencial, por ejemplo, la distribución normal, exponencial, gamma, chi-cuadrado, beta, Dirichlet, Bernoulli, categórica, Poisson, Wishart (inversa) y geométrica. Otras distribuciones solo pertenecen a la familia exponencial para una determinada elección de sus parámetros, como lo ilustra el siguiente ejemplo.

Ejemplo 4.9 (El modelo binomial pertenece a la familia exponencial)**.**
Recordemos la distribución binomial está dada por

$$\begin{aligned}\text{Bin}\,(x|\theta,n) &= \binom{n}{x}\theta^x(1-\theta)^{n-x}, \quad x\in\{0,1,2\ldots,n\}\\ &= \underbrace{\binom{n}{x}}_{h(x)}\exp\left(x\ \underbrace{\log\left(\frac{\theta}{1-\theta}\right)}_{\text{parámetro natural}}+\underbrace{n\log\left(1-\theta\right)}_{-A(\theta)}\right)\end{aligned}$$

consecuentemente, para que $h(x)$ sea únicamente una función de la variable aleatoria, entonces el número de intentos n tiene que se una cantidad conocida, **no un parámetro**.

Ejemplo 4.10 (El modelo normal pertenece a la familia exponencial)**.**
La distribución normal $\mathcal{N}(\mu,\sigma^2)$ tiene densidad:

$$p_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-1}{2\sigma^2}(x-\mu)^2} \tag{4.11}$$

$$= \frac{1}{\sqrt{2\pi}}\exp\left(\frac{\mu}{\sigma^2}x-\frac{-1}{2\sigma^2}x^2-\left(\frac{\mu^2}{2\sigma^2}+\log\sigma\right)\right) \tag{4.12}$$

donde $\theta=(\mu,\sigma^2)$. Esta es una familia exponencial de dos parámetros con

- Estadístico: $T_1(x)=x$, $T_2(x)=x^2$,
- Parámetro natural: $\nu_1(\theta)=\mu/\sigma^2$, $\nu_2(\theta)=-1/(2\sigma^2)$,
- $A(\theta)=\mu^2/(2\sigma^2)+\log\sigma$,
- $h(x)=1/\sqrt{2\pi}$.

Ejemplo 4.11.
Ejercicio Demuestre que las distribuciones *Poisson* y *Bernoulli* pertenecen a la familia exponencial.

Observación 4.4.
El estadístico T es un estadístico suficiente para ν en la familia exponencial. En efecto, notemos que

$$p_\theta(x) = \underbrace{\exp\left(\sum_{i=1}^{s}\eta_i(\theta)T_i(x)-A(\theta)\right)}_{g_\theta(T(x))}\underbrace{h(x)}_{h(x)} \tag{4.13}$$

consecuentemente, por el Teorema 4.1 (Neyman-Fisher), tenemos que T es un estadístico suficiente para ν.

La familia exponencial va a ser ampliamente usada durante el curso, lo cual se debe a sus propiedades favorables para el análisis estadístico. Por ejemplo, el producto de dos distribuciones de la familia exponencial también pertenece a la familia exponencial. En efecto, consideremos dos VA X_1, X_2, con distribuciones en la familia exponencial respectivamente dadas por

$$p_1(x_1) = h_1(x_1) \exp\left(\theta_1 T_1(x_1) - A_1(\theta_1)\right) \tag{4.14}$$
$$p_2(x_2) = h_2(x_2) \exp\left(\theta_2 T_2(x_2) - A_2(\theta_2)\right) \tag{4.15}$$

si asumimos que estas VA son independientes, entonces densidad conjunta de $X = (X_1, X_2) \sim p$ está dada por

$$\begin{aligned} p(x) &= p_1(x_1)p_2(x_2) \\ &= \underbrace{h_1(x_1)h_2(x_2)}_{h(x)} \exp\left(\underbrace{[\theta_1, \theta_2]}_{\theta} \underbrace{\begin{bmatrix} T_1(x_1) \\ T_2(x_2) \end{bmatrix}}_{T(x)} - \underbrace{(A_1(\theta_1) + A_2(\theta_2))}_{A(\theta)} \right), \end{aligned} \tag{4.16}$$

con lo que eligiendo $\theta = [\theta_1, \theta_2]$ y $T = [T_1, T_2]$, vemos que X está dado por una distribución de la familia exponencial.

Otra propiedad de las familia exponencial es la relación entre los momentos de la distribucion y el lognormalizador A. Denotando

$$Q(\theta) = \exp\left(A(\theta)\right) = \int_{\mathcal{X}} \exp\left(\theta T(x)\right) h(x) \mathrm{d}x, \tag{4.17}$$

podemos observar que la derivada de $A(\theta)$ está dada por

$$\begin{aligned} \frac{dA(\theta)}{d\theta} &= Q^{-1}(\theta) \frac{dQ(\theta)}{d\theta} \\ &= \frac{\int_{\mathcal{X}} T(x) \exp\left(\theta T(x)\right) h(x) \mathrm{d}x}{\int_{\mathcal{X}} \exp\left(\theta T(x)\right) h(x) \mathrm{d}x} \\ &= \frac{\int_{\mathcal{X}} T(x) \exp\left(\theta T(x) - A(\theta)\right) h(x) \mathrm{d}x}{\int_{\mathcal{X}} \exp\left(\theta T(x) - A(\theta)\right) h(x) \mathrm{d}x} \qquad \cdot A(\theta)/A(\theta) \\ &= \mathbb{E}\left(T(x)\right) \end{aligned} \tag{4.18}$$

Ejemplo 4.12.
Verificar para derivadas de orden superior (ejercicio)

Observación 4.5.
Consideremos el mapa

$$\theta \mapsto \mu = \frac{dA(\theta)}{d\theta} = \int_{\mathcal{X}} T(x) \exp\left(\theta T(x) - A(\theta)\right) h(x) \mathrm{d}x = \mathbb{E}\left(T(x)\right) \tag{4.19}$$

Para ciertas familias, este mapa es biyectivo (familia minimal $\rightarrow A(\theta)$ estrictamente convexo), es decir, podemos expresar el modelo mediante los parámetros μ en vez de θ, esto se llama *mean parametrisation* (MP), manteniendo la relación 1-1 entre modelos a distintas parametrizaciones. La MP es fundamental en el problema de estimación: ¿por qué?

4.4. Ejercicios

1. Estudiemos el problema de estimación del área de un rectángulo. Consideremos un rectángulo de lados a,b desconocidos. Sea $X = (X_1, ..., X_n)$ una MAS que representa el lado a) del rectángulo con $X_i \sim \mathcal{N}(a, \sigma^2)$ e $Y = (Y_1, ..., Y_n)$ otra MAS que representa el lado b) del rectángulo con $Y_i \sim \mathcal{N}(b, \sigma^2)$, σ conocido y X e Y independientes. Plantee el modelo paramétrico asociado y encuentre un estadístico suficiente.

2. Se desea estudiar el tiempo que demora en realizarse cierto proceso. Digamos que el proceso posee 2 etapas. Se consta de una MAS dada por $X = (X_1, ..., X_n)$ donde cada $X_i \sim \Gamma(2, \theta)$ la distribución Gamma con X_i el tiempo total que demora en realizarse un proceso completo (ambas etapas), θ es desconocido. Plantee el modelo parametrico y demuestre que $T(X) = \sum_{i=1}^{n} X_i$ es un estadístico suficiente.

3. Sea $X = (X_1, ..., X_n)$ una MAS con $X_i \sim U(\alpha, \beta)$ donde α, β son desconocidos. Demuestre que $T(X) = \left(\min_{i=1,...,n} X_i, \max_{i=1,...,n} X_i \right)$ es un estadístico suficiente para (α, β).

4. Sea $X = (X_1, ..., X_n)$ una MAS con $X_i \sim \Gamma(\alpha, \beta)$ la distribución Gamma dada por:

$$\Gamma(\alpha, \beta) \sim \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-\frac{1}{\beta}x}$$

   Demuestre que $T(X) = \left(\prod_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i \right)$ es un estadistico suficiente para (α, β)

5. Sea $X = (X_1, ..., X_n)$ una MAS con función de densidad dada por:

$$f(x \mid \theta) = \frac{\theta}{(1+x)^{1+\theta}}, 0 < x < \infty, \theta > 0$$

   Con θ desconocido. Encuentre un estadístico suficiente.

6. Sea $X = (X_1, ..., X_n)$ una MAS con una función de densidad dada por:

$$f(x|\theta) = \theta x^{\theta-1}, \theta > 0, x \in (0, 1)$$

   (i) ¿Es $\sum_{i=1}^{n} X_i$ un estadístico suficiente para θ?

   (ii) Encuentre un estadístico minimal suficiente para θ

(iii) ¿Es $T'(X) = \sum_{i=1}^{n} ln(X_i)$ un estadístico minimal para θ?

(iv) Encuentre un estadístico suficiente y completo para θ.

(v) Sea S(X) un estadístico ancilario para θ. Encuentre la correlación entre S(X) y $T'(X)$.

Capítulo 5

Estimadores

Recordemos que, dada una familia de modelos estadísticos y datos que asumimos vienen de un miembro de dicha familia, nuestro objetivo es obtener (estimar) el modelo particular que generó los datos, es decir, cuáles son los parámetros del modelo. En este capítulo se introducirá la noción de estimador, es decir, una función que busca estimar el parámetro mencionado anteriormente en base a los datos disponibles.

Definición 5.1.

][Estimador] Sea $g : \Omega \to \mathbb{R}^n$ tal que $g(\theta) = (g_1(\theta), ..., g_n(\theta))$ a valores en $\mathbb{R}$. Nos interesa estimar $g(\theta)$. Para estimar $g(\theta)$ usamos un **estimador** que es una función $\hat{g} : \mathfrak{X} \to g(\Omega)$ medible. Diremos que $\hat{g}(\theta)$ es la estimación de $g(\theta)$.

Observación 5.1.

Los estimadores son casos particulares de los estadísticos, pues son funciones de los datos que tienen por conjunto de llegada la imagen de Ω a través de $g(\cdot)$.

Observación 5.2.

Los estimadores pueden ser usados para estimar el parámetro propiamente tal, en cuyo caso $g(\theta) = \theta$, o bien otras cantidades del modelo que son expresables a través de los parámetros. Por ejemplo, en el caso de un modelo Gaussiano, si bien el parámetro puede ser expresado como $\theta = [\mu, \sigma^2]$, podemos estar interesados en estimar el intervalo de confianza del 95 %, el cual está dado (aproximadamente) por

$$g(\theta) = [\mu - 2\sigma, \mu + 2\sigma]. \tag{5.1}$$

Ejemplo 5.1 (Estimador de la media Gaussiana)**.**

Consideremos $X = (X_1, \ldots, X_n) \sim \mathcal{N}(\mu, \sigma^2)$. Un estimador de $g(\theta) = g(\mu, \sigma) = \mu$ es el estadístico

$$\hat{g}(X) = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

5.1. Estimadores insesgados

Recordemos que nuestros estimadores, como función de la variable aleatoria X, son a su vez variables aleatorias. Consecuentemente, su estudio debe considerar sus propiedades aleatorias también. El primer paso para esto es la siguiente definición que dice relación con el valor esperado del estimador y el valor de la función $g(\theta)$ que éste estima.

Definición 5.2 (Estimador insesgado)**.**
Sea $\hat{g}(X)$ un estimador de $g(\theta)$. Este estimador es insesgado si

$$\mathbb{E}\left(\hat{g}(X)\right) = g(\theta),$$

donde el *sesgo* de $\hat{g}$ se define como

$$b_{\hat{g}}(\theta) = \mathbb{E}\left(\hat{g}(X)\right) - g(\theta).$$

Se dice también que un estimador es **asintoticamente insesgado** si es que:

$$\lim_n \mathbb{E}\left(\hat{g}(X_1, ..., X_n)\right) = g(\theta),$$

es decir, si el estimador solo se convierte en insesgado al usar *infinitos datos*.

Los estimadores insesgados juegan un rol relevante en el estudio y aplicación de la estadística, pues nos dicen que el estimador recupera efectivamente el parámetro *en promedio*. Sin embargo, uno no siempre debe poner exclusiva atención a ellos, pues el hecho que funcione en promedio no garantiza nada en cuanto a su dispersión (varianza) o cuántas muestras necesitamos para que el estimador sea confiable.

Los siguientes ejemplos ilustran el rol del estimador insesgado en dos familias paramétricas distintas.

Ejemplo 5.2 (Estimador insesgado de la media Gaussiana)**.**
El estimador de $g(\theta) = \mu$ descrito en el Ejemplo 5.1 es insesgado, en efecto:

$$\mathbb{E}\left(\hat{g}(X)\right) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu.$$

Veamos ahora un ejemplo de un estimador **sesgado** de la varianza y cómo se puede construir un estimador insesgado en base a éste.

Ejemplo 5.3 (Pythagoras)**.**
Consideremos una familia paramétrica $\mathcal{P} = \{P_\theta \text{ t.q. } \theta \in \Theta\}$ y denotemos por μ y σ^2 su media y su varianza respectivamente. Usando las observaciones $x_1, x_2, \ldots, x_n$,

calculemos la varianza del estimador de la media, dado por $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ mediante

$$\mathbb{V}_\theta(\bar{x}) = \mathbb{V}_\theta\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \underbrace{=}_{\text{i.i.d.}} \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}_\theta(x_i) = \frac{\sigma^2}{n} \tag{5.2}$$

es decir, el estimador de la media usando n muestras, tiene una varianza σ^2/n.

Consideremos ahora el siguiente estimador para la varianza:

$$S_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{5.3}$$

y notemos que la esperanza de dicho estimador es

$$\begin{aligned}
\mathbb{E}_\theta(S_2) &= \mathbb{E}_\theta\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu + \mu - \bar{x})^2\right) \\
&= \mathbb{E}_\theta\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 + 2\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(\mu-\bar{x}) + \frac{1}{n}\sum_{i=1}^{n}(\mu-\bar{x})^2\right) \\
&= \mathbb{E}_\theta\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 - 2(\mu-\bar{x})^2 + (\mu-\bar{x})^2\right) \\
&= \mathbb{E}_\theta\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 - (\mu-\bar{x})^2\right) \\
&= \mathbb{V}_\theta(x_i) - \mathbb{V}_\theta(\bar{x}) \quad \text{ver ecuación (5.2)} \\
&= \sigma^2 + \sigma^2/n = \left(\frac{n+1}{n}\right)\sigma^2
\end{aligned} \tag{5.4}$$

Esto quiere decir que el sesgo del estimador en la ecuación (5.3) es asintóticamente insesgado, es decir, que su sesgo tiende a cero cuando el número de muestas n tiende a infinito. Sin embargo, notemos que podemos corregir el estimador de la varianza multiplicando el estimador original, S_2 en la ecuación (5.3) por $n/(n+1)$, con lo que el estimador corregido denotado por

$$S_2' = \frac{n}{n+1}S_2 = \frac{1}{n+1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{5.5}$$

cumple con

$$\mathbb{E}_\theta(S_2') = \left(\frac{n}{n+1}\right)\mathbb{E}_\theta(S_2) \underbrace{=}_{\text{ec.(5.4)}} \left(\frac{n}{n+1}\right)\left(\frac{n+1}{n}\right)\sigma^2 = \sigma^2 \tag{5.6}$$

es decir, el estimador S_2' en la ecuación (5.5) es insesgado.

5.2. Funciones de pérdida

Una función de pérdida, también llamada función de costo, es una función a valores reales de dos argumentos que, intuitivamente, determina el costo de estimar uno de los argumentos mediante el otro. Como nuestro objetivo es estimar parámetros definimos entonces una función de costo de la siguiente forma. **Desde ahora consideraremos estimadores de $g(\theta) = \theta$ y todas las esperanzas serán con respecto a θ por simplicidad de notación.**

Definición 5.3 (Función de costo)**.**
Sea $\theta \in \Omega$ un parámetro y $a \in \Omega$ un estimador, entonces el costo de estimar θ mediante a está dado por la función de costo definida mediante:

$$L : (\Omega \times \Omega) \to \mathbb{R} \tag{5.7}$$

$$(\theta \times a) \mapsto L(\theta, a). \tag{5.8}$$

Ejemplo 5.4 (Función de costo cuadrática)**.**
Una función de costo ampliamente usada para comparar estimadores es el **error cuadrático**, el cual está dado por

$$L_2(\theta, a) = ||\theta - a||^2.$$

Pregunta: ¿por qué usamos el exponente igual a 2 y no otro?

Ejemplo 5.5 (Función de costo $0-1$)**.**
Cuando estimamos parámetros que no tiene relación de orden, podemos usar la función de costo $0-1$ dada por

$$L_{01}(\theta, a) = \mathbb{1}_{\theta \neq a}.$$

Ejemplo 5.6 (Divergencia de Kullback-Liebler)**.**
Cuando los parámetros a estimar son distribuciones de probabilidad, podemos usar la siguiente función de costo

$$L_{\mathrm{KL}}(\theta, a) = \sum_{i=1}^{D} \theta_i \log\left(\frac{\theta_i}{a_i}\right).$$

Como el estimador (que es el argumento de la función de pérdida) es una VA, también lo es la función de pérdida. Consecuentemente, podemos calcular la esperanza de la función de pérdida, lo cual conocemos como *riesgo*.

En particular, el riesgo asociado a la pérdida cuadrática en el Ejemplo 5.5 para un estimador

ϕ del parámetro θ, está dado por:

$$\begin{aligned} R(\theta,\phi) &= \mathbb{E}\left((\theta-\phi)^2\right) \\ &= \mathbb{E}\left((\theta-\bar{\phi}+\bar{\phi}-\phi)^2\right); \quad \text{denotando } \bar{\phi}=\mathbb{E}(\phi) \\ &= \mathbb{E}\left((\theta-\bar{\phi})^2+2(\theta-\bar{\phi})\cancel{(\bar{\phi}-\phi)}+(\bar{\phi}-\phi)^2\right) \\ &= \underbrace{(\theta-\bar{\phi})^2}_{=b_\phi^2 \text{ (sesgo}^2\text{)}} + \underbrace{\mathbb{E}\left((\bar{\phi}-\phi)^2\right)}_{=V_\phi \text{ (varianza)}}. \end{aligned} \tag{5.9}$$

Donde podemos ver unas de las razones de la consideración del costo cuadrático: su riesgo se divide intuitivamente en dos términos que expresan la exactitud (cuán sesgado es) y la precisión (cuán disperso es) del estimador.

5.3. Teorema de Rao-Blackwell

Comentario: Para una notación más clara, nos referimos a los estimadores estimadores $\phi = \hat{g}$ de θ en general para evitar la expresión más engorrosa estimador $\hat{g}(X)$ de $g(\theta)$.

Siguiendo el racional de la sección anterior, evaluaremos la bondad de distintos estimadores (sesgados o insesgados) mediante una función de *pérdida* o *costo* que compara el valor reportado por el estimador y el valor real del parámetro. Esto permite usar la función de pérdida como una métrica para comparar (la bondad de) dos o más estimadores.

El siguiente teorema establece que la información reportada por un estadístico suficiente (Definición 4.2), puede solo mejorar un estimador.

Teorema 5.1 (Teorema de Rao-Blackwell)**.**
Sea $\phi = \phi(X)$ un estimador de θ tal que $\mathbb{E}_\theta(\phi) < \infty, \forall\theta$. Asumamos que existe $T = T(X)$ estadístico suficiente para θ y sea $\phi^\star = \mathbb{E}_\theta(\phi|T)$. Entonces,

$$\mathbb{E}_\theta\left((\phi^\star-\theta)^2\right) \leq \mathbb{E}_\theta\left((\phi-\theta)^2\right), \forall\theta, \tag{5.10}$$

donde la desigualdad es estricta salvo en el caso donde ϕ es función de T.

En otras palabras, el Teo. de Rao-Blackwell establece que un estimador puede ser *mejorado* si es reemplazado por su esperanza condicional dado un estadístico suficiente. El proceso de mejorar un estimador poco eficiente de esta forma es conocido como *Rao-Blackwellización* y veremos un ejemplo a continuación.

Ejemplo 5.7.
Consideremos $X = (X_1, \ldots, X_n) \sim \text{Poisson}(\theta)$ y estimemos el parámetro θ. Para esto, consideremos el estimador básico $\phi = X_1$ y *Rao-Blackwellicémoslo* usando el estimador

suficiente $T = \sum_{i=1}^{n} X_i$, es decir,

$$\phi^* = \mathbb{E}_\theta \left(X_1 \middle| \sum_i X_i = t \right). \tag{5.11}$$

Para calcular esta esperanza condicional, observemos primero que

$$\sum_{j=1}^{n} \mathbb{E}_\theta \left(X_j \middle| \sum_{i=1}^{n} X_i = t \right) = \mathbb{E}_\theta \left(\sum_{j=1}^{n} X_j \middle| \sum_{i=1}^{n} X_i = t \right) = t, \tag{5.12}$$

y que como $X_1, \ldots, X_n$ son iid, entonces todos los términos dentro de la suma del lado izquierdo de la ecuación anterior son iguales. Consecuentemente, recuperamos el estimador

$$\phi^* = \frac{t}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i. \tag{5.13}$$

Antes de demostrar el Teorema 5.1 consideremos dos variable aleatorias $X \in \mathcal{X}$, $Y \in \mathcal{Y}$, y recordemos dos propiedades básicas. En primer lugar la ley de esperanzas totales, la cual establece que

$$\begin{aligned}
\mathbb{E}_Y \mathbb{E}_{X|Y}(X|Y) &= \int_{\mathcal{Y}} \int_{\mathcal{X}} x \mathrm{d}P(x|y) \mathrm{d}P(y) && \text{def. esperanza} \\
&= \int_{\mathcal{X}} x \int_{\mathcal{Y}} \mathrm{d}P(x|y) \mathrm{d}P(y) && \text{linealidad} \\
&= \int_{\mathcal{X}} x \int_{\mathcal{Y}} \mathrm{d}P(x, y) && \text{def. esperanza condicional} \\
&= \int_{\mathcal{X}} x \mathrm{d}P(x) = \mathbb{E}_X(X). && \text{def. esperanza}
\end{aligned} \tag{5.14}$$

En segundo lugar, recordemos (?) la desigualdad de Jensen, la cual para el caso particular del costo cuadrático, puede verificarse mediante

$$0 \leq \mathbb{V}(X) = \mathbb{E}\left(X^2\right) - \mathbb{E}(X)^2 \Rightarrow \mathbb{E}\left(X^2\right) \geq \mathbb{E}(X)^2. \tag{5.15}$$

La desigualdad de Jensen es geométricamente intuitiva, como se observa en la Figura 5.1. Al calcular la imagen de $\mathbb{E}(x)$ bajo una función convexa, podemos encontrar una recta tangente a ese punto $L(X) = aX + b$. Tendremos que $\mathbb{E}(\varphi(X')) \geq \mathbb{E}(L(X')) = \mathbb{E}[aX' + b] = a\mathbb{E}[X`] + b = L(\mathbb{E}(X'))$ para otro punto X'. Tomando $X' = X$, $\mathbb{E}(\varphi(X)) \geq \varphi(\mathbb{E}(X))$.

Volviendo a lo anterior, utilizando las expresiones en (5.14) y (5.15), podemos demostrar el teorema anterior.

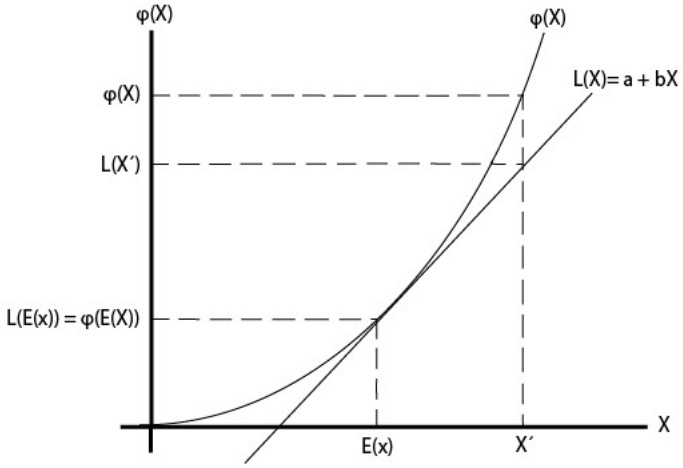

Figura 5.1: Intuición geométrica de la desigualdad de Jensen

Demostración de Teorema 5.1. La varianza del estimador $\phi^\star$ está dada por

$$\begin{aligned}
\mathbb{E}_\theta\left((\phi^\star-\theta)^2\right) &= \mathbb{E}_\theta\left(\left(\mathbb{E}_\theta\left(\phi|T\right)-\theta\right)^2\right) && \text{def.}\\
&= \mathbb{E}_\theta\left(\left(\mathbb{E}_\theta\left(\phi-\theta|T\right)\right)^2\right) && \text{linealidad}\\
&\leq \mathbb{E}_\theta\left(\mathbb{E}_\theta\left((\phi-\theta)^2|T\right)\right) && \text{Jensen}\\
&= \mathbb{E}_\theta\left((\phi-\theta)^2\right) && \text{ley esperanzas totales}
\end{aligned}$$

Donde las esperanzas exteriores son con respecto a T y las interiores con respecto a X (o equivalentemente a ϕ). Observemos además que la desigualdad anterior viene de la expresión en la ecuación (5.15), por lo que la igualdad es obtenida si $\mathbb{V}\left(\phi-\theta|T\right)=0$, es decir, la VA $\phi-\theta$ tiene que ser constante para cada valor de T, es decir, ϕ es función de T. Intuitivamente podemos entender esto como que si el estadístico ya fue considerado en el estimador, entonces conocer el valor del estadístico no reporta información adicional. ■

Observación 5.3.

Notemos que si el estimador ϕ es insesgado, su *Rao-Blackwellización* ϕ^* también lo es, en efecto

$$\mathbb{E}_\theta\left(\phi^*\right)=\mathbb{E}_\theta\left(\mathbb{E}_\theta\left(\phi|T\right)\right)=\mathbb{E}_\theta\left(\phi\right)=\theta, \tag{5.16}$$

donde la segunda igualdad está dada por la ley de esperanzas totales y la tercera por el supuesto de que ϕ es insesgado.

5.4. Varianza uniformemente mínima

Observemos que, en base al riesgo cuadrático definido en la ecuación (5.9), si un estimador es insesgado (Definición 5.2) entonces su riesgo cuadrático es únicamente su varianza. Esto motiva la siguiente definición de optimalidad para estimadores insesgados.

Definición 5.4 (Estimador insesgado de varianza uniformemente mínima)**.**
El estimador $\phi = \phi(X)$ de θ es un estimador insesgado de varianza uniformemente mínima (EIVUM) si es insesgado y además si $\forall \phi' : \mathcal{X} \to \Theta$ estimador insesgado se tiene

$$\mathbb{V}_\theta(\phi) \leq \mathbb{V}_\theta(\phi'), \forall \theta \in \Theta. \tag{5.17}$$

Es decir, el EIVUM es el estimador insesgado que tiene menor varianza de todos los estimadores insesgados (y puede no ser único).

Ejemplo 5.8.
Consideremos $X = (X_1, \ldots, X_n) \sim \text{Ber}(\theta)$ y los siguientes estimadores de θ

- $\phi_1(X) = X_1$
- $\phi_2(X) = \frac{1}{2}(X_1 + X_2)$
- $\phi_3(X) = \frac{1}{n}\sum_{i=1}^n X_i$

Observemos que todos estos estimadores son insesgados, pues como $\forall i, \mathbb{E}_\theta(X_i) = \theta$, entonces

$$\mathbb{E}_\theta(\phi_1(X)) = \mathbb{E}_\theta(\phi_2(X)) = \mathbb{E}_\theta(\phi_3(X)) = \theta. \tag{5.18}$$

Veamos ahora que la varianza de $\phi_3(X)$ está dada por

$$\mathbb{V}_\theta(\phi_3(X)) = \mathbb{V}_\theta\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\sum_{i=1}^n \mathbb{V}_\theta(X_i) = \frac{\theta(1-\theta)}{n} \tag{5.19}$$

pues $\mathbb{V}_\theta(X_i) = \mathbb{E}_\theta((\theta - X_i)^2) = \mathbb{E}_\theta(X_i^2) - \theta^2 = (0^2 \cdot (1-\theta) + 1^2 \cdot \theta) - \theta^2 = \theta(1-\theta)$. Consecuentemente, la varianza de los estimadores considerados decae como la inversa del número de muestras $1/n$.

Con las definiciones anteriores, podemos mencionar el siguiente teorema, el cual conecta la noción de estadístico completo con la de EIVUM.

Teorema 5.2 (Teorema de Lehmann-Scheffé)**.**
Sea X una VA con distribución paramétrica $\mathcal{P} = \{P_\theta \text{ t.q. } \theta \in \Theta\}$ y T un estadístico suficiente y completo para θ. Si el estimador $\phi = \phi(T)$ de θ es insesgado, entonces ϕ es el único EIVUM.

Es decir, el Teorema de Lehmann-Scheffé nos permite verificar que un estimador es el (único) EIVUM, si éste es insesgado y es función de un estadístico suficiente y completo.

Demostración. Veamos en primer lugar que es posible construir un estimador en función del estadístico suficiente $\phi(T)$ que tiene menor o igual varianza que un estimador arbitrario $\phi'(X)$. En efecto, el Teorema de Rao-Blackwell establece que el estimador

$$\phi(T) = \mathbb{E}_\theta(\phi'(X)|T), \tag{5.20}$$

tiene efectivamente menor (o igual) varianza que $\phi'(X)$.

Ahora veamos que solo existe un único estimador insesgado que es función del estadístico completo T. Asumiendo que existiesen dos estimadores insesgados de θ que son funciones de T, denotados $\phi_1(T), \phi_2(T)$, entonces, $\mathbb{E}_\theta\left(\phi_1(T) - \phi_2(T)\right) = 0$, es decir, $\phi(T) = \phi_1(T) - \phi_2(T)$ es un estimado insesgado de 0. Luego, como T es completo, entonces, $\phi(T) = 0$ es identicamente nulo, lo cual implica que $\phi_1(T) = \phi_2(T)$ c.s.-P_θ.

Hemos probado que (i) para un estimador arbitrario, se puede construir un estimador que es función de T el cual tiene menor o igual varianza que el estimador original y, (ii) el estimador insesgado $\phi(T)$ es único. Consecuentemente, $\phi(T)$ es el único EIVUM. ■

El Teorema de Lehmann-Scheffé da una receta para encontrar el EIVUM: simplemente es necesario encontrar un estadístico completo y construir un estimador insesgado en base a éste, esto garantiza que el estimador construido es el **único** EIVUM.

Ejemplo 5.9 (EIVUM para Bernoulli)**.**

Recordemos que en el Ejemplo 5.10 vimos que el estadístico $T = \sum_{i=1}^{n} X_i$ es completo para $X \sim \text{Ber}(\theta)$. Como el estimador de θ dado por $\phi(T) = T/n$ es insesgado,

$$\mathbb{E}_\theta\left(\phi(T)\right) = \mathbb{E}_\theta\left(T/n\right) = \sum_{i=1}^{n} \mathbb{E}_\theta\left(X_i\right)/n = \theta, \tag{5.21}$$

entonces $\phi(T) = T/n$ es el EIVUM para θ en $\text{Ber}(\theta)$ y es único.

Observación 5.4.

Lectura personal: Estadístico auxiliar (ancilliary) y teoremas de Basur y de Bahadur.

Definición 5.5.

Un estadístico T=T(X) se dice **ancilario** si su distribución no depende del parámetro a estimar.

Teorema 5.3 (Teorema de Basu)**.**

Si T es completo y suficiente para θ y V es ancilario para θ, entonces T y V son independientes

5.4.1. Información de Fisher

Para entender esta propiedad, primero definamos la función de puntaje o *score function* como la función aleatoria definida por la derivada de la log-verosimilitud, es decir,

$$S_\theta(X) = \frac{\partial \log p_\theta(X)}{\partial \theta}. \tag{5.22}$$

Observación 5.5.

La esperanza de la función de puntaje cero. En efecto, derivando la igualdad fundamental $1 = \int_{\mathcal{X}} p_\theta(x)\mathrm{d}x$ con respecto a θ, obtenemos

$$0 = \int_{\mathcal{X}} \frac{\partial p_\theta(X)}{\partial \theta}\mathrm{d}x = \int_{\mathcal{X}} \frac{1}{p_\theta(X)}\frac{\partial p_\theta(X)}{\partial \theta} p_\theta(X)\mathrm{d}x = \int_{\mathcal{X}} \frac{\partial \log p_\theta(X)}{\partial \theta} p_\theta(X)\mathrm{d}x = \mathbb{E}_\theta\left(S_\theta(X)\right) \tag{5.23}$$

Sorprendente.

Además, veamos que al derivar por segunda vez la función de puntaje, obtenemos:

$$\begin{aligned} 0 &= \int_{\mathcal{X}} \frac{\partial}{\partial \theta}\left(\frac{\partial \log p_\theta(X)}{\partial \theta} p_\theta(X)\right)\mathrm{d}x \\ &= \int_{\mathcal{X}} \left(\frac{\partial^2 \log p_\theta(X)}{\partial \theta^2} p_\theta(X) + \frac{\partial \log p_\theta(X)}{\partial \theta}\frac{\partial p_\theta(X)}{\partial \theta}\right)\mathrm{d}x \\ &= \mathbb{E}_\theta\left(\frac{\partial^2 \log p_\theta(X)}{\partial \theta^2}\right) + \mathbb{E}_\theta\left(\left(\frac{\partial \log p_\theta(X)}{\partial \theta}\right)^2\right). \end{aligned}$$

Cada uno de los dos términos de la ecuación anterior tiene la misma magnitud (uno es negativo y el otro es positivo), lo cual motiva la siguiente definición.

Definición 5.6 (Información de Fisher)**.**

La cantidad denotada mediante

$$I(\theta) = \mathbb{E}_\theta\left(\left(\frac{\partial \log p_\theta(X)}{\partial \theta}\right)^2\right) = -\mathbb{E}_\theta\left(\frac{\partial^2 \log p_\theta(X)}{\partial \theta^2}\right), \tag{5.24}$$

es conocida como información de Fisher. Además, como la esperanza de la función de puntaje es cero, la varianza de $I(\theta)$ puede ser expresada como

$$\mathbb{V}_\theta\left(S_\theta(X)\right) = \mathbb{E}_\theta\left(S_\theta(X)^2\right) - \cancel{\mathbb{E}_\theta\left(S_\theta(X)\right)^2} = \mathbb{E}_\theta\left(\left(\frac{\partial \log p_\theta(X)}{\partial \theta}\right)^2\right). \tag{5.25}$$

Consecuentemente, la información de Fisher también es la varianza de la función de pérdida, con lo que contamos con tres expresiones para poder calcular $I(\theta)$.

Ejercicio 5.1 (Cálculo de la información de Fisher para Bernoulli)**.**

Consideremos $X \sim \text{Ber}\,(\theta)$, entonces,

$$\begin{aligned} I(\theta) &= -\mathbb{E}_\theta\left(\frac{\partial^2}{\partial\theta^2}\log\left(\theta^X(1-\theta)^{1-X}\right)\right) \\ &= -\mathbb{E}_\theta\left(\frac{\partial^2}{\partial\theta^2}X\log\theta + \frac{\partial^2}{\partial\theta^2}(1-X)\log\left(1-\theta\right)\right) \\ &= \mathbb{E}_\theta\left(X\theta^{-2} + (1-X)(1-\theta)^{-2}\right) \\ &= \theta^{-1} + (1-\theta)^{-1} \\ &= \frac{1}{\theta(1-\theta)}. \end{aligned} \tag{5.26}$$

Ejercicio 5.2 (Cálculo de la información de Fisher para Poisson)**.**
Consideremos $X \sim \text{Poisson}\,(\theta)$, entonces,

$$\begin{aligned} I(\theta) &= \mathbb{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log\left(\frac{\theta^X e^{-\theta}}{X!}\right)\right)^2\right) \\ &= \mathbb{E}_\theta\left(\left(\frac{\partial}{\partial\theta}X\log\theta - \frac{\partial}{\partial\theta}\theta - \frac{\partial}{\partial\theta}\log(X!)\right)^2\right) \\ &= \mathbb{E}_\theta\left(\left(X\theta^{-1} - 1\right)^2\right) \\ &= \mathbb{E}_\theta\left(X^2\theta^{-2} - 2X\theta^{-1} + 1\right) \\ &= (\theta+\theta^2)\theta^{-2} - 2\theta\theta^{-1} + 1 \\ &= \theta^{-1}. \end{aligned}$$

Hasta ahora hemos calculado la función de puntaje en base a la verosimilitud de solo una una variable aleatoria. Si considerásemos la verosimilitud evaluada calculada para un conjunto de observaciones (IID), tenemos que

$$S_\theta(X_1,\ldots,X_n) = \frac{\partial\log\prod_{i=1}^n p_\theta(X_i)}{\partial\theta} = \sum_{i=1}^n \frac{\partial\log p_\theta(X_i)}{\partial\theta} = \sum_{i=1}^n S_\theta(X_i). \tag{5.27}$$

De igual forma, para la información de Fisher, tenemos,

$$I_n(\theta) = \mathbb{V}_\theta\left(\sum_{i=1}^n S_\theta(X_i)\right) = nI(\theta). \tag{5.28}$$

Observación 5.6.
La expresión anterior confirma la intuición sobre la información de Fisher en cuanto a *cuán informativa* es una muestra X para estimar el parámetro θ: Si una muestra tiene tiene una información de Fisher $I(\theta)$, entonces n muestras independientes del mismo

modelo tendrán n veces dicha información.

5.4.2. Cota de Cramer Rao

Veamos ahora una desigualdad interesante para la información de Fisher y su relación con estimadores. Consideremos un estimador insesgado, es decir,

$$\mathbb{E}_\theta\left(\hat{\theta}(X)-\theta\right)=\int_{\mathcal{X}}(\hat{\theta}(X)-\theta)p_\theta(X)\mathrm{d}x=0. \tag{5.29}$$

Derivando esta expresión con respecto a θ, obtenemos

$$\begin{aligned}0&=\frac{\partial}{\partial\theta}\int_{\mathcal{X}}(\hat{\theta}(X)-\theta)p_\theta(X)\mathrm{d}x\\&=-\int_{\mathcal{X}}p_\theta(X)\mathrm{d}x+\int_{\mathcal{X}}(\hat{\theta}(X)-\theta)\frac{\partial p_\theta(X)}{\partial\theta}\mathrm{d}x\\&=-1+\int_{\mathcal{X}}(\hat{\theta}(X)-\theta)\frac{\partial\log p_\theta(X)}{\partial\theta}p_\theta(X)\mathrm{d}x.\end{aligned}$$

Lo que implica que

$$\begin{aligned}1&=\left(\int_{\mathcal{X}}(\hat{\theta}(X)-\theta)\frac{\partial\log p_\theta(X)}{\partial\theta}p_\theta(X)\mathrm{d}x\right)^2\\&=\left(\int_{\mathcal{X}}(\hat{\theta}(X)-\theta)\sqrt{p_\theta(X)}\sqrt{p_\theta(X)}\frac{\partial\log p_\theta(X)}{\partial\theta}\mathrm{d}x\right)^2\\&\leq\int_{\mathcal{X}}(\hat{\theta}(X)-\theta)^2p_\theta(X)\mathrm{d}x\int\left(\frac{\partial\log p_\theta(X)}{\partial\theta}\right)^2p_\theta(X)\mathrm{d}x.\end{aligned}$$

Notemos que la primera integral es la varianza del estimador insesgado $\hat{\theta}$ y la segunda es la esperanza del cuadrado de la función de puntaje (o la información de Fisher). Con esto, podemos enunciar el siguiente resultado

Definición 5.7 (Cota de Cramer-Rao)**.**
Sea $X_1,\ldots,X_n\sim p_\theta$ y $nI(\theta)$ su información de Fisher. Entonces para todo estimador insesgado θ' tenemos

$$\mathbb{V}_\theta\left(\theta'\right)\geq\left(nI(\theta)\right)^{-1},\quad\forall\theta\in\Theta \tag{5.30}$$

La cota de Cramer-Rao es un elemento fundamental en el estudio estadístico, pues establece que cualquier estimador insesgado tiene necesariamente una varianza que está por sobre el recíproco de la información de Fisher. Es decir, la varianza de un EIVUM se encuentra acotada inferiormente.

5.5. Completitud

Otra propiedad de los estimadores que permite estudiar su capacidad de estimar es la de *completitud*. A continuación definimos esta propiedad para el caso general de un estadístico, no necesariamente un estimador.

Definición 5.8 (Estadístico completo)**.**
Un estadístico $T(X)$ es completo si para toda función f, se tiene que

$$\mathbb{E}_\theta\left(f(T)|\theta\right) = 0, \forall\theta \in \Theta \Rightarrow \mathbb{P}_\theta\left(f(T) = 0\right) = 1, \forall\theta \in \Theta. \tag{5.31}$$

Intuitivamente entonces, podemos entender la noción de completitud como lo siguiente: un estadístico es completo si la única forma de construir un estimador insesgado de cero a partir de él es aplicándole la función idénticamente nula. Veamos un ejemplo de la distribución Bernoulli, donde el estadístico $T(x) = \sum x_i$ es efectivamente completo.

Ejemplo 5.10.
Sea $x = (x_1, \ldots, x_n)$ observaciones de $X \sim \text{Ber}(\theta)$, recordemos que $T(x) = \sum x_i \sim \text{Bin}(n, \theta)$, por lo que la esperanza de $f(T)$ está dada por

$$\mathbb{E}_\theta\left(f(T)\right) = \sum_{t=0}^{n} f(t)\binom{n}{t}\theta^t(1-\theta)^{n-t} = (1-\theta)^n \sum_{t=0}^{n} f(t)\binom{n}{t}\left(\frac{\theta}{1-\theta}\right)^t, \tag{5.32}$$

es decir un polinomio de grado n en $r = \theta/(1-\theta) \in \mathbb{R}_+$. Entonces, $\mathbb{E}_\theta\left(f(T)\right) = 0, \forall\theta$, implica que necesariamente los pesos de este polinomio son todos idénticamente nulos, es decir, $f(t) = 0, \forall t$, lo que a su vez implica $\mathbb{P}_\theta\left(f(T) = 0\right) = 1$. Consecuentemente, $T(x) = \sum x_i \sim \text{Bin}(n, \theta)$ es un estadístico completo.

El concepto de completitud dice relación con la construcción de estimadores usando estadísticos, lo cual puede ser ilustrado mediante el siguiente ejemplo

Ejemplo 5.11.
Consideremos dos estimadores, ϕ_1, ϕ_2 insesgados de θ distintos, es decir,

$$\mathbb{E}\left(\phi_1\right) = \mathbb{E}\left(\phi_2\right) = \theta,\ \mathbb{P}\left(\phi_1 \neq \phi_2\right) > 0. \tag{5.33}$$

Definamos ahora $\phi = \phi_1 - \phi_2$, donde verificamos que $\mathbb{E}\left(\phi\right) = 0, \forall\theta$, es decir, ϕ es un estimador insesgado de cero. Como nuestra hipótesis en la ecuación anterior dice que $\mathbb{P}\left(\phi_1 - \phi_2 = 0\right) > 0$, de acuerdo a la definición de estadístico completo, ϕ no es completo.

Ejemplo 5.12 (Estimador de la taza de la distribución exponencial)**.**
Consideremos $X \sim Exp(\theta)$, donde $Exp(x|\theta) = \theta\exp(-\theta x), \theta > 0$. Veamos en primer lugar que el estadístico trivial $T(X) = X$ es completo. En efecto, para una función

cualquiera $f(\cdot)$, como $\theta > 0$ tenemos

$$\mathbb{E}_\theta\left(f(X)\right) = \int_0^\infty f(x)\theta\exp(-\theta x)\mathrm{d}x = 0 \Rightarrow \int_0^\infty f(x)\exp(-\theta x)\mathrm{d}x = 0 \tag{5.34}$$

con lo cual si el lado derecho de la expresión anterior se cumple $\forall\theta$, entonces necesariamente $f(x) = 0$ (¿por qué?).

En segundo lugar, asumamos que existe un estimador insesgado $\hat{g}(X)$ de $g(\theta) = \theta$. Es decir,

$$\mathbb{E}_\theta\left(\hat{g}(X)\right) = \int_0^\infty \hat{g}(x)\theta\exp(-\theta x)\mathrm{d}x = \theta, \forall\theta,$$

lo cual es equivalente a $\int_0^\infty \hat{g}(x)\exp(-\theta x)\mathrm{d}x = 1, \forall\theta$, y también a (al derivar ambos lados de esta expresión c.r.a. θ)

$$\int_0^\infty x\hat{g}(x)\exp(-\theta x)\mathrm{d}x = 0, \forall\theta. \tag{5.35}$$

Esta última expresión es equivalente a que $\mathbb{E}\left(X\hat{g}(X)\right) = 0$, con lo que podemos utilizar el hecho de que X es un estadístico completo para decir que la función $X\hat{g}(X) = 0$ c.s. $\forall\theta$, y consecuentemente $\hat{g}(X) = 0$ c.s. $\forall\theta$.

Hemos mostrado que el supuesto de la existencia de un estimador (denotado $\hat{g}(X)$) insesgado para el parámetro del modelo exponencial $\theta > 0$, resulta en la contradicción $\hat{g}(X) = 0$ c.s. Consecuentemente, no es posible construir estimadores insesgados para θ en la distribución exponencial.

5.6. Ejercicios

1. Considere una Muestra Aleatoria Simple (MAS) $X = (X_1, ..., X_n)$ donde $X_i \sim \mathcal{N}(\mu, \sigma^2)$, $\forall i = 1, ..., n$ con μ y σ son parámetros desconocidos. Se consideran al estimador de la varianza y la media como:

$$S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(\overline{X}_n - X_i)^2, \quad \hat{\mu} := \overline{X}_n$$

donde $\overline{X}_n$ denota al promedio de $X_1, ..., X_n$, es decir $\overline{X}_n = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$.

   a) Demuestre que S^2 y $\hat{\mu}$ son independientes.

2. Sea $X = (X_1, ..., X_n)$ una MAS de un modelo Poisson de parámetro λ. Se busca estimar insesgadamente la función generadora de momentos $M_X(t) = \mathbb{E}[e^{tX}]$. Para esto, siga los siguientes pasos:

   a) Muestre que $M_{X_i}(t) = e^{-\lambda(1-e^t)}$, con $t \in \mathbb{R}$

b) Compruebe que $\widehat{M_X}(t) = e^{tX_1}$ es un estimador insesgado de $M_X(t)$.

c) Muestre que el estadístico $T(X) = \sum_{i=1}^{n} X_i$ tiene ley Poisson de parámetro $n\lambda$.

d) Calcule $\mathbb{P}(X_1 = k|S = s)$. ¿Qué ley sigue?

e) Muestre que $T(X)$ es un estadístico suficiente y completo para λ

f) Encuentre el estimador EIVUM para $M_X(t)$

3. Sea una MAS $X = (X_1, .., X_n)$ con n observaciones independientes del modelo gaussiano $\mathcal{N}(\mu, \sigma^2)$ y otra MAS $Y = (Y_1, ...Y_n)$ con n observaaciones independientes del modelo gaussiano $\mathcal{N}(\nu, \sigma^2)$. Se supone que X e Y son vectores independientes y que los parámetros μ, ν, σ son desconocidos y no están sujetos a ninguna restricción.

   a) Plantee el modelo paramétrico relacionado a la situación planteada. Compruebe $\mathcal{P}$ pertenece a la clase exponencial y que es de rango completo en la parametrización natural.

   b) Analice los siguientes estadísticos e indique si son suficientes y/o minimales:

   - $T_1 = \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} Y_i, \sum_{i=1}^{n} X_i^2, \sum_{i=1}^{n} Y_i^2\right)$
   - $T_2 = \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} Y_i, \sum_{i,j=1}^{n} Y_i X_j\right)$
   - $T_3 = \left(\sum_{i=1}^{n} (X_i + Y_i), \sum_{i=1}^{n} X_i^2, \sum_{i,j=1}^{n} Y_i^2\right)$

   c) Muestre que $S = \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} Y_i, \sum_{i,j=1}^{n} Y_i^2 + X_i^2\right)$ es un estadístico suficiente completo para $\mathcal{P}$

   d) Encuentre el EIVUM para las siguientes funciones del parámetro θ:

   - $g_1(\theta) = \mu + \nu$
   - $g_2(\theta) = \mu\nu$
   - $g_3(\theta) = \sigma^2$

   Se suele presentar cuando un vector bidimensional (por ejemplo, el que representa la velocidad del viento) tiene sus dos componentes, ortogonales, independientes y siguen una distribución normal. Su valor absoluto seguirá entonces una distribución de Rayleigh

4. Sea $X = (X_1, ..., X_n)$ una MAS de $n \in \mathbb{N}$ observaciones con $X_n \sim Unif(\theta, 2\theta)$. Considere que para la MAS anterior, sus estadísticos de orden $X_{(1)}, ..., X_{(n)}$ poseen una densidad dada por:

$$f_{X_{(i)}}(x) = \frac{n!}{(i-1)!(n-i)!}\frac{1}{\theta}\left(\frac{x-\theta}{\theta}\right)^{i-1}\left(1-\left(\frac{x-\theta}{\theta}\right)\right)^{n-i} \quad , \quad x \in [\theta, 2\theta].$$

Se busca estimar el parámetro θ. Para esto se pide lo siguiente:

a) Encuentre un estadístico suficiente para θ.

b) Considere el estimador $\hat{\theta} = \frac{2}{3}X_1$. Muestre que es insesgado.

c) Utilice el Teorema de Rao-Blackwell para encontrar un estimador $\widetilde{\theta}$ de θ con menor MSE que $\hat{\theta}$.

d) Demuestre que $\widetilde{\theta}$ es insesgado.

e) Interprete $\widetilde{\theta}$

Capítulo 6

Construcción de estimadores

6.1. Estimador de Máxima Verosimilitud (EMV)

Informalmente, el estimador de un parámetro es una función de los datos que deseamos que entregue un valor cercano al parámetro. Dada una cantidad desconocida, se hace natural la idea de buscar encontrar una *buena* (y ojalá la *mejor*) función de los datos que nos permita estimarla, pero ¿Qué significa que un estimador sea un buen estimador?

Dado que el parámetro θ es desconocido, calcular la distancia de un estimador $\hat{\theta} = \hat{\theta}(X)$ a este no es posible, pues de lo contrario podríamos simplemente utilizar una función de pérdida como las definidas en el capítulo anterior.

En esta sección, veremos cómo construir estimadores usando directamente la densidad de probabilidad de la VA $X \in \mathcal{X}$, donde aparece el parámetro θ y una colección de datos (o realizaciones del modelo). Para este fin la función de verosimilitud en la definición 8.3 será fundamental. Recordemos que la función de verosimilitud (del parámetro θ dados los datos X) es la densidad de probabilidad de los datos X si el valor del parámetro fuese efactivamente θ. Consecuentemente, la verosimilitud permite encontrar un estimador en base a una métrica clara: cuan probable es cada estimador de haber generado los datos. Esto da las condiciones para determinar un estimador que recibe mucha atención en la literatura estadística:

Definición 6.1 (Estimador de máxima verosimilitud (MV))**.**
Sea una observación x y una función de verosimilitud $L(\theta)$, el estimador de máxima verosimilitud está dado por

$$\theta_{\text{MV}} = \arg\max_{\theta} L(\theta|x) \tag{6.1}$$

Claramente, el estimador de MV puede ser definido con respecto a la verosimilitud o a cualquier función no decreciente de ésta, como también pude no existir o no ser único. En particular, nos enfocaremos en encontrar θ_{MV} mediante la maximización de la log-verosimilitud $l(\theta) = \log L(\theta)$, la cual es usualmente más fácil de optimizar en términos computacionales o analíticos. De hecho, muchas veces incluso ignoraremos constantes de la (log) verosimilitud, pues éstas no cambian el máximo de $L(\theta)$.

Ejemplo 6.1 (Máxima verosimilitud: Bernoulli)**.**
Sea $X_1, \dots X_n \sim \text{Ber}(\theta)$, la verosimilitud de θ está dada por

$$L(\theta) = \prod_{i=1}^{n} \theta_i^x (1-\theta)^{1-x_i}, \tag{6.2}$$

y su log-verosimilitud por $l(\theta) = (\sum_{i=1}^n x_i) \log\theta + (n - \sum_{i=1}^n x_i)\log(1-\theta)$. El estimador de MV puede ser encontrado resolviendo $\frac{\partial l(\theta)}{\partial\theta} = 0$:

$$\begin{aligned}\frac{\partial l(\theta)}{\partial\theta} = 0 \Rightarrow (\sum_{i=1}^n x_i)\theta^{-1} &= (n - \sum_{i=1}^n x_i)(1-\theta)^{-1} \\ \Rightarrow \sum_{i=1}^n x_i(1-\theta) &= (n - \sum_{i=1}^n x_i)\theta \\ \Rightarrow \theta &= \sum_{i=1}^n x_i/n.\end{aligned}$$

Notemos que este estimador de MV ¡es a su vez el EIVUM!

Ejercicio 6.1.
Graficar $l(\theta)$ en el Ejemplo 6.1.

Ejercicio 6.2.
Encuentre el estimador de MV de $\theta = (\mu, \Sigma)$ para la VA $X \sim \mathcal{N}(\mu, \Sigma)$.

Ejemplo 6.2.
Sea la VA $X \sim \text{Uniforme}(\theta)$, es decir, $p(x) = \theta^{-1}\mathbb{1}_{0\le x\le\theta}$. Para calcular la verosimilitud, recordemos en primer lugar que la verosimilitud factoriza de acuerdo a

$$L(\theta) = \prod_{i=1}^{n} p_\theta(x_i) \tag{6.3}$$

y observemos que necesariamente $p_\theta(x_i) = 0$ si $x_i > \theta$. Consecuentemente, $L(\theta) > 0$ solo si θ es mayor que toda las observaciones, en particular, si $\theta \ge \text{máx}\{x_i\}_1^n$.

Además, si efectivamente tenemos $\theta \ge \text{máx}\{x_i\}_1^n$, entonces notemos que $p_\theta(x_i) = 1/\theta$, por lo que la verosimilitud está dada por

$$L(\theta) = \theta^{-n}, \quad \theta \ge \text{máx}\{x_i\}_1^n \tag{6.4}$$

y consecuentemente, el estimador de máxima verosimilitud es $\theta_{\text{MV}} = \text{máx}\{x_i\}_1^n$.

6.2. Propiedades del EMV

6.2.1. Consistencia

La primera propiedad que veremos del EMV es su consistencia. Que un estimador $\hat{\theta}$ sea *consistente* quiere decir que éste tiende (de alguna forma) al parámetro real θ a medida vamos considerando más datos. Recordar en lo siguiente que la KL hace referencia a la divergencia de Kullback-Leibler.

Con la KL, definiremos que un modelo/parámetro es **identificable** si los valores para los parámetros $\theta \neq \theta'$ implican $\mathrm{KL}\left(p_\theta \| p'_\theta\right) > 0$, lo que significa que distintos valores del parámetro dan origen a distintos modelos, intuitivamente, esto significa que la *parametrización* del modelo estadístico no es redundante. Asumiremos desde ahora que los modelos considerados son identificables.

El estimador de MV puede ser obtenido de la maximización de

$$M_n(\theta') = n^{-1}(l_n(\theta') - l_n(\theta)) = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{p_{\theta'}(x_i)}{p_\theta(x_i)}\right), \tag{6.5}$$

donde n es la cantidad de observaciones $\{x_1, \ldots, x_n\}$, θ es el parámetro real y $l_n(\cdot)$ es la log-verosimilitud en base a dichas observaciones. La obtención del EMV desde la maximización de $M_n(\theta')$ en la ecuación (6.5) es posible porque $l_n(\theta)$ es constante para θ', con lo que $l_n(\theta') \propto_\theta M_n(\theta')$.

Entonces, gracias a la ley de los grandes números, tenemos que

$$M_n(\theta') \to \mathbb{E}_\theta\left(\log\left(\frac{p_{\theta'}(x)}{p_\theta(x)}\right)\right) = -\mathbb{E}_\theta\left(\log\left(\frac{p_\theta(x)}{p_{\theta'}(x)}\right)\right) = -\mathrm{KL}\left(p_\theta \| p_{\theta'}\right). \tag{6.6}$$

Consecuentemente, como el objetivo del estimador de MV tiende a la KL negativa, entonces maximizar la verosimilitud es equivalente a minimizar la KL-divergencia entre el modelo real y el modelo generado por el parámetro.

Observación 6.1.
Máxima verosimilitud es (asintóticamente) efectivavamente equivalente a minimizar discrepancias en el espacio de modelos.

Observación 6.2.
Si el modelo obtenido mediante MV tiende efectivamente al modelo real (no tenemos garantías de esto todavía) nuestro supuesto de *identificabilidad* implica que el estimador de MV tiende al parámetro real también. Sin embargo, si el modelo está parametrizado de tal forma que no es identificable, convergencia en el espacio de modelos no implica necesariamente convergencia en los parámetros.

Otra propiedad muy utilizada en la práctica es el **Principio de equivarianza**, el cual establece que si θ_{MV} es el estimador de MV de θ, entonces, $g(\theta_{\mathrm{MV}})$ es el estimador de MV del parámetro transformado $g(\theta)$.

Ejemplo 6.3.

(Cálculo del EMV en Gaussiana: varianza versus precisión versus log-precisión versus cholesky - reparametrisation trick)

6.2.2. Normalidad asintótica

Otra propiedad es la **normalidad asintótica del EMV**, esto significa que el estimador ML (como cantidad aleatoria) es normal en el límite que la cantidad de observaciones tiende a infinito.

Formalmente, si tenemos una colección de VA $X_1, \ldots, X_n \sim p_\theta$ con θ el parámetro real, entonces, la secuencia de estimadores de MV, ${\theta_{\text{MV}}}^{(n)}$ cumple con

$$\sqrt{n}({\theta_{\text{MV}}}^{(n)} - \theta) \to \mathcal{N}(0, (I(\theta))^{-1}), \tag{6.7}$$

lo cual intuitivamente corresponde a que, para n suficientemente grande, el estimador de MV está distribuido de forma normal en torno al parámetro real con varianza $(nI(\theta))^{-1}$. Lo que implica también *eficiencia asintótica*: si n es suficientemente grande, entonces la distribución del estimador es normal y su varianza tiende a cero. Es decir, asintóticamente, el EMV alcanza la Cota de Cramer Rao para la varianza.

6.3. Estimador de Mínimo Cuadrático Ordinario (MCO)

El Estimador de Mínimo Cuadrático Ordinario (MCO) es un tipo de aproximación de mínimos cuadrados mediante una función lineal de los datos. Formalmente, supongamos que se busca predecir $Y_i \in \mathbb{R}$ con $i = 1, ..., n$ donde n representa la cantidad de datos mediante una función lineal de los datos $(X_i) \in R^k$, es decir, mediante una ponderación y suma de k variables. Se asume que $X_{i1} = 1, \forall i$ para considerar un intercepto y que existe un error ϵ_i de estimación para cada dato. Con esto, el modelo lineal viene dado por:

$$Y_i = \sum_{j=1}^{k} \beta_j X_{ij} + \varepsilon_i$$

con $\mathbb{E}(\varepsilon_i) = 0$. En forma matricial, se considera que:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}; X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} \text{ y } \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Recordar que para fila de $X \in \mathbb{R}^{n \times k}$ es una observación. Con esto, el modelo lineal en su

forma matricial viene dado por:

$$Y = X\beta + \varepsilon$$

Definición 6.2 (Estimador de Mínimo Cuadrático Ordinario (MCO))**.**

La idea es buscar una solución del sistema $X\beta = Y$, el cual generalmente no tiene solución, por lo cual se busca una aproximación para β. Bajo la siguiente función de costo:

$$S(\beta) = \sum_{i=1}^{n} |y_i - \sum_{i=1}^{n} X_{ij}\beta_j|^2 = ||Y - X\beta||^2$$

el estimador $\hat{\beta}$ viene dado por:

$$\hat{\beta} = \arg\min_{\beta} S(\beta)$$

el cual tiene como solución, si es que $X^T X$ es invertible, dada por:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Cuando se cumplen las siguientes condiciones:

- Exogenedidad: $\mathbb{E}[\varepsilon|X] = 0$,
- Homocedasticidad: $\mathbb{E}[\varepsilon_i^2|X] = \sigma^2$
- No autocorrelación: $\mathbb{E}[\varepsilon_i\varepsilon_j|X] = 0, \forall i \neq j$.

se entiende a $\hat{\beta}$ como el estimador de mínimos cuadrados ordinarios (MCO, o OLS por su nombre en ingles), en donde además:

$$\mathbb{V}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

6.3.1. Teorema de Gauss-Markov

Propiedad 6.1.

El estimador $\hat{\beta}$ es insesgado, es decir, $\mathbb{E}[\hat{\beta}|X] = \beta$

Teorema 6.1.

][Teorema de Gauss-Markov] Si es que se cumple homocedasticidad y no autocorrelación de los errores para el estimador MCO, se tiene que el estimador $\hat{\beta}$ es eficiente en la clase de estimadores lineales insesgados. Lo anterior se denomina un estimador lineal insesgado óptimo (ELIO), o en su nombre en ingles, best linear unbiased estimator (BLUE). Esta eficiencia es en el sentido de que, sea $\bar{\beta}$ otro estimador lineal e insesgado de β, entonces se tiene que:

$$\mathbb{V}[\hat{\beta}|X] - \mathbb{V}[\bar{\beta}] \leq 0$$

es decir, el estimador MCO $\hat{\beta}$ es el estimador linear insesgado con menor varianza.

6.4. Regresión

La palabra *regresión* fue introducida por Francis Galton (1822-1911), haciendo referencia a que los hijos de personas altas, tendían a ser más bajos que sus padres, fenómeno que denominó **Regresión a la media**. Este mismo fenómeno se puede observar cuando la segunda película de una saga no es tan buena como la primera parte.
La regresión es un método para estudiar la relación entre una variable Y, y otra variable independiente X, denominada característica.

Definición 6.3 (Función de Regresión)**.**
Se define la función de regresión $r(x)$ como:

$$r(x) = \mathbb{E}(Y|X = x) = \int yf(y|x)dx$$

La idea de este método consiste en, dados datos $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$, encontrar una distribución $F_{X,Y}$.

6.4.1. Regresión Lineal Simple

Comencemos viendo el caso unidimensional, es decir $X_i \in \mathbb{R}$. Buscamos ajustar $r(x)$ de forma tal que:

$$r(x) = \beta_0 + \beta_1 x,$$

es decir, de forma que y sea una función lineal (o lineal a fin) de x. Supondremos que hay un ruido ε_i tal que $\mathbb{V}(\varepsilon_i) = \sigma^2$, y es independiente de x.

Definición 6.4.
Se define el modelo de regresión lineal simple como:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

con $\mathbb{E}(\varepsilon_i) = 0$, y $\mathbb{V}(\varepsilon_i) = \sigma^2$.

Buscamos estimar β_0 y β_1 de forma que tengamos una aproximación lineal que sea lo mejor posible. Estas últimas palabras nos hacen preguntarnos ¿Los mejores estimadores con respecto

a qué? La respuesta es, con respecto a la métrica de mínimo cuadrados:

$$J(\hat{\beta}_0, \hat{\beta}_1) = \frac{1}{2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2,$$

donde $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$.

Teorema 6.2.

Los estimadores de mínimos cuadrados son:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}_i)}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Un estimador insesgado de σ^2 es:

$$\hat{\sigma^2} = \frac{1}{n-2} J(\hat{\beta}_0, \hat{\beta}_1)$$

6.4.2. Mínimos Cuadrados y Máxima Verosimilitud

Supongamos ahora que $\varepsilon \sim \mathcal{N}(0, \sigma^2$, es decir, $Y_i|X_i \sim \mathcal{N}(\mu_i, \sigma^2)$, con $\mu_i = \beta_0 + \beta_1 X_i$. Calculemos la verosimilitud:

$$\mathcal{L} = \prod_{i=1}^{n} f_X(X_i) f_{Y|X}(Y_i|X_i) = \prod_{i=1}^{n} f_X(X_i) \prod_{i=1}^{n} f_{Y|X}(Y_i|X_i)$$

Llamemos $\mathcal{L}_1$ a la primera parte de este producto, y $\mathcal{L}_2$ a la segunda parte. Como $\mathcal{L}_1$ no depende de β_0 y β_1, tenemos que para calcular los estimadores de máxima verosimilitud de estos parámetros, nos importa el segundo parámetro. Entonces, considerando la log-verosimilitud de $\mathcal{L}_2$:

$$\mathcal{L}_2 = \prod_{i=1}^{n} f_{Y|X}(Y_i|X_i) \propto \sigma^{-n} exp(\frac{-1}{2\sigma^2}\sum_i (Y_i - \mu_i)^2)$$

$$\implies l = -nlog(\sigma) - \frac{-1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - (\beta_0 + \beta_1 X_i)^2$$

Notemos que al minimizar esto, como el primer término es constante con respecto a β_0 y β_1, tenemos:

Teorema 6.3.

Bajo la hipótesis de normalidad, el estimador de mínimos cuadrados coincide con el estimador de máxima verosimilitud. También se tiene:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)$$

Observación 6.3.

El estimador anterior de $\hat{\sigma}^2$ normalmente se reemplaza por el estimador insesgado de la parte anterior.

Observación 6.4.

Podemos tener observaciones de muchos X_i, pero no incluirlos todos al modelo. Un modelo más reducido tiene dos ventajas: La primera es que puede entregar mejores predicciones que un modelo más grande, y la segunda es que es más simple.
Generalmente, mientras más variables se añaden a la regresión, el sesgo de la predicción disminuye pero aumenta la varianza. Una muestra pequeña genera mucho sesgo; esto se llama *underfitting*. Una muestra muy grande lleva a una alta varianza; esto se llama *overfitting*. Las buenas predicciones vienen de balancear sesgo y varianza.

6.4.3. Regresión Logística

Definición 6.5.

Se llama función logística a la función:

$$f(x) = \frac{e^x}{1+e^x}$$

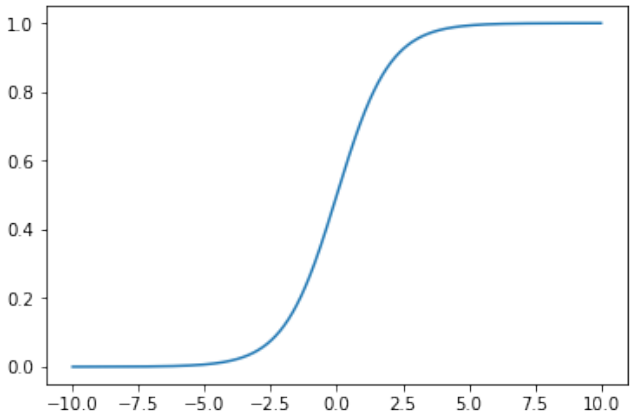

Figura 6.1: Función Logística

La regresión logística es un método de regresión para el caso en que $Y_i \in \{0, 1\}$. El modelo considera:

$$p_i \equiv p_i(\beta) \equiv \mathbb{P}(Y_i = 1|X = x) = f(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}),$$

con f la función logística. De forma equivalente, si definimos $logit(p) = log(\frac{p}{1-p})$:

$$p_i = logit(p_i) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

Como Y_i son variables binarias, se tiene que $Y_i|X_i = x_i \sim Ber(p_i)$. Así, la función de verosimilitud será:

$$\mathcal{L} = \prod_{i=1}^{n} p_i(\beta)^{Y_i}(1 - p_i(\beta))^{1-Y_i}$$

Obtenemos el estimador de β, $\hat{\beta}$ usando método numéricos de optimización.

6.4.4. Sobre Regresión LinealEMV en práctica: tres ejemplos

Regresión lineal y gaussiana

Una aplicación muy popular del estimador de MV es en los modelos de regresión lineal y gaussianos. Consideremos el caso donde se desea modelar la cantidad de pasajeros que mensualmente viajan en una aerolínea, para esto, sabemos de nuestros colaboradores en la división de análisis de datos de la aerolínea que ésta cantidad tiene una tendencia de crecimiento cuadrática en el tiempo y además una componente oscilatoria de frecuencia anual. Estos fenómenos pueden ser explicados por el aumento de la población, los costos decrecientes de la aerolínea y la estacionalidad anual de las actividades económicas.

Asumiendo que la naturaleza de la cantidad de pasajeros es estocástica, podemos usar los supuestos anteriores para modelar la densidad condicional de dicha cantidad (con respecto al tiempo t) mediante una densidad normal parametrizada de acuerdo a

$$X \sim \mathcal{N}\left(\theta_0 + \theta_1 t^2 + \theta_2 \cos(2\pi t/12), \theta_3^2\right), \tag{6.8}$$

donde $\theta_0, \theta_1, \theta_2$ parametrizan la media y θ_3 la varianza.

Consecuentemente, si nuestras observaciones están dadas por $\{(t_i, x_i)\}_{i=1}^n$ podemos escribir la log-verosimilitud de θ como

$$\begin{aligned} l(\theta) &= \log\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta_3^2}} \exp\left(-\frac{(x_i - \theta_0 - \theta_1 t^2 - \theta_2 \cos(2\pi t/12))^2}{2\theta_3^2}\right)\right) \\ &= \frac{n}{2}\log\left(2\pi\theta_3^2\right) - \frac{1}{2\theta_3^2}\sum_{i=1}^{n}(x_i - \theta_0 - \theta_1 t_i^2 - \theta_2 \cos(2\pi t_i/12))^2 \end{aligned} \tag{6.9}$$

con lo que vemos que θ_{MV} puede ser calculado explícitamente y es función de $\{(t_i, x_i)\}_{i=1}^n$ debido a que la ecuación (6.9) es cuadrática en $[\theta_0, \theta_1, \theta_2]$.

Regresión no lineal: clasificación

La razón por la cual θ_{MV} pudo ser calculado de forma explícita es porque el modelo Gaussiano con media parametrizada de forma lineal resulta en una log-verosimilitud cuadrática, donde el mínimo es único y explícito. Sin embargo, en muchas situaciones el modelo lineal y gaussiano no es el apropiado.

Un ejemplo es esto es problema de evaluación crediticia (*credit scoring*) donde en base a un conjunto de *características* que definen a un cliente, un ejecutivo bancario debe evaluar si otorgarle o no el crédito que el cliente solicita. Para tomar esta decisión, el ejecutivo puede revisar la base de datos del banco e identificar los clientes que en el pasado pagaron o no pagaron sus créditos para determinar el perfil del *pagador* y el del *no-pagador*. Finalmente, un nuevo cliente puede ser *clasificado* como pagador/no-pagador en base su similaridad con cada uno de estos grupos.

Formalmente, denotemos las características del cliente como $t \in \mathbb{R}^N$ y asumamos que el cliente paga su crédito con probabilidad $\sigma(t)$ y no lo paga con probabilidad $1 - \sigma(t)$, la función $\sigma(t)$ a definir. Esto es equivalente a construir la VA X

$$X|t \sim \text{Ber}\left(\sigma(t)\right) \tag{6.10}$$

donde $X = 1$ quiere decir que el cliente paga su crédito y $X = 0$ que no. Una elección usual para la función $\sigma(\cdot)$ es la función logística aplicada a una transformación lineal de t, es decir,

$$\Pr\left(X = 1|t\right) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 t)}}. \tag{6.11}$$

Notemos que este es un clasificador lineal, donde $\theta = [\theta_0, \theta_1]$ define un hiperplano en $\mathbb{R}^N$ en donde los clientes $t \in \{t|0 \leq \theta_0 + \theta_1 t\}$ pagan con probabilidad mayor o igual a $1/2$ y el resto con probabilidad menor o igual a $1/2$. Esto es conocido como **regresión logística**.

Entonces, usando los registros bancarios $\{(x_i, t_i)\}_{i=1}^n$ ¿cuál es el $\theta = [\theta_0, \theta_1]$ de máxima verosimilitud? Para esto notemos que la log-verosimilitud puede ser escrita como

$$\begin{aligned} l(\theta) &= \log \prod_{i=1}^{n} p(x_i|t) \\ &= \sum_{i=1}^{n} x_i \log \sigma(t) + \left(n - \sum_{i=1}^{n} x_i\right) \log(1 - \sigma(t)) \\ &= \sum_{i=1}^{n} x_i \log \frac{1}{1 + e^{-(\theta_0 + \theta_1 t)}} + \left(n - \sum_{i=1}^{n} x_i\right) \log(1 - \frac{1}{1 + e^{-(\theta_0 + \theta_1 t)}}) \end{aligned}$$

Esta expresión no tiene mínimo global y a pesar que podemos calcular su gradiente, no podemos resolver $\partial l(\theta)/\partial\theta = 0$ de forma analítica, por lo que debemos usar métodos de descenso de gradiente.

Variables latentes: *Expectation-Maximisation*

En ciertos escenarios es natural asumir que nuestros datos provienen de una mezcla de modelos, por ejemplo, consideremos la distribución de estaturas en una población, podemos naturalmente modelar esto como una mezcla de distribuciones marginales para las estaturas de hombres y mujeres por separado, es decir,

$$X \sim p\mathcal{N}(X|\mu_H, \Sigma_H) + (1-p)\mathcal{N}(X|\mu_M, \Sigma_M) \tag{6.12}$$

donde la verosimilitud de los parámetros $\theta = [p, \mu_H, \sigma_H, , \mu_M, \sigma_M]$ dado un conjunto de observaciones $\{x_i\}_{i=1}^n$ es

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \left(p\mathcal{N}(X|\mu_H, \Sigma_H) + (1-p)\mathcal{N}(X|\mu_M, \Sigma_M)\right) \\ &= \prod_{i=1}^{n} \left(p\frac{1}{\sqrt{2\pi\Sigma_H^{-1}}} \exp\left(\frac{-(x_i-\mu_H)^2}{2\Sigma_H^2}\right) + (1-p)\frac{1}{\sqrt{2\pi\Sigma_M^{-1}}} \exp\left(\frac{-(x_i-\mu_M)^2}{2\Sigma_M^2}\right)\right). \end{aligned}$$

Optimizar esta expresión con respecto a las 5 componentes de θ es difícil, en particular por la suma en la expresión, lo cual no permite simplificar la expresión mediante la aplicación de $\log(\cdot)$.

Una interpretación de la diferencia de este modelo con respecto a los anteriores es la introducción implícita de una *variable latente* que describe de qué gaussiana fue generada cada observación. Si conociésemos esta variable latente, el problema sería dramáticamente más sencillo. En efecto, asumamos que tenemos a nuestra disposición las observaciones $\{z_i\}_{i=1}^n$ de la VA $\{Z_i\}_{i=1}^n$ las cuales denota de qué modelo es generada cada observación, por ejemplo, $Z_i = 0$ (cf. $Z_i = 1$) denota que el individuo con estatura X_i es hombre (cf. mujer).

En este caso, asumamos por un segundo que estas variables latentes están disponibles y consideremos los **datos completos** $\{(x_i, z_i)\}_{i=1}^n$ para escribir la función de verosimilitud completa mediante

$$\begin{aligned} l(\theta|z_i, x_i) &= \prod_{i=1}^{n} \mathcal{N}(X|\mu_H, \Sigma_H)^{z_i}\mathcal{N}(X|\mu_M, \Sigma_M)^{(1-z_i)} \\ &= \sum_{i=1}^{n} \left(z_i \log \frac{1}{\sqrt{2\pi\Sigma_H^{-1}}} \exp\left(\frac{-(x_i-\mu_H)^2}{2\Sigma_H^2}\right) + (1-z_i)\log\frac{1}{\sqrt{2\pi\Sigma_M^{-1}}} \exp\left(\frac{-(x_i-\mu_M)^2}{2\Sigma_M^2}\right)\right). \end{aligned}$$

Esta función objetivo es mucho más fácil de optimizar, pero no es observable pues la VA Z es desconocida. Una forma de resolver esto es tomando la esperanza condicional de la expresión anterior (con respecto a Z) condicional a los datos y los parámetros *actuales*, para luego maximizar esta expresión c.r.a. θ y comenzar nuevamente. Específicamente, como la

expresión anterior es lineal en z_i basta con tomar su esperanza:

$$\begin{aligned}\mathbb{E}_\theta\left(Z_i|\theta_t, x_i\right) &= 1 \cdot \mathbb{P}\left(Z_i = 1|\theta_t, x_i\right) + 0 \cdot \mathbb{P}\left(Z_i = 0|\theta_t, x_i\right) \\ &= \frac{\mathbb{P}\left(x_i|\theta_t, z_i = 1\right) p(z_i = 1)}{p(x_i|\theta)} \\ &= \frac{\mathbb{P}\left(x_i|\theta_t, z_i = 1\right) p(z_i = 1)}{p(x_i|z = 1, \theta)p(z = 1) + p(x_i|z = 0, \theta)p(z = 0)}\end{aligned}$$

6.5. Método de los Momentos

Sean $X_1, ..., X_n$ n muestras i.i.d. de una variable aleatoria X con función de densidad $p_X(x;\theta)$ o $f_X(x;\theta)$ en el caso de ser discreta o continua respectivamente, dependientes de un parámetro $\theta \in \mathbb{R}^k$ a estimar.

Recordar que los momentos de una variable aleatoria viene dado por $\mu_k = \mathbb{E}[X^k]$ donde k representa el k-momento.

El estimador del método de los momentos es un estimador $\hat{\theta} \in \mathbb{R}^k$ de θ como la solución (si es que existe) de las k ecuaciones simultaneas de los primeros k momentos, esto es:

$$\begin{aligned}\mathbb{E}[X^1] &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ \mathbb{E}[X^2] &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 \\ &\vdots \\ \mathbb{E}[X^k] &= \frac{1}{n}\sum_{i=1}^{n} x_i^k\end{aligned}$$

Ejemplo 6.4.

][Caso Exponencial] Consideremos $X_1, ..., X_n$ n muestras i.i.d. de una variable aleatoria $X \sim exp(\theta)$ donde claramente $k = 1$. Aplicando el método de los momentos, debemos resolver que:

$$\mathbb{E}[X] = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Por lo que

$$\mathbb{E}[X] = \frac{1}{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

y entonces:

$$\hat{\theta} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_i}$$

el cual calza justamente con el estimador EMV de θ.

6.6. Intervalos de Confianza

En distintas situaciones, la estimación puntual de un parámetro puede no ser apropiada e incluso inverosímil, mientras que la distribución posterior puede ser poco interpretable por el público general. En dichos casos, es recomendable identificar un rango donde, con cierta probabilidad, el parámetro real está contenido. Esto motiva la siguiente definición:

Definición 6.6 (Intervalo de confianza)**.**
Un $(1-\alpha)$-intervalo de confianza para el parámetro θ fijo y desconocido, $\alpha \in [0,1]$, es el intervalo aleatorio $(A(X), B(X))$ tal que

$$\mathbb{P}_\theta\left(A(X) \leq \theta \leq B(X)\right) = 1-\alpha, \forall \theta \in \Theta. \tag{6.13}$$

Observación 6.5.
La definición del intervalo de confianza no describe una probabilidad sobre el parámetro θ, pues estamos tomando un enfoque frecuentista (no bayesiano) donde éste es fijo. Por el contrario, lo que es aleatorio en la ecuación (6.13) es el intervalo, no el parámetro. Entonces, si bien es una sutileza, la definición anterior se debe entender como la probabilidad de que "el intervalo (aleatorio) contenga al parámetro (fijo)", y no como la probabilidad de que "el parámetro esté en el intervalo".

Una consecuencia clave de este concepto es que si $I_{1-\alpha}$ es un $(1-\alpha)$-intervalo de confianza, entonces si fuese posible repetir una gran cantidad de veces el ejercicio de recolectar datos X y calcular este intervalo para cada una de estas observaciones, entonces el parámetro θ estaría contenido en el intervalo un $100(1-\alpha)\,\%$ de las veces. Esto es muy diferente de asegurar que para un solo experimento, la probabilidad de que el parámetro θ esté contenido en $I_{1-\alpha}$ es $1-\alpha$, lo cual no es cierto. Los siguientes ejemplos tienen por objetivo ayudar a aclarar este concepto.

Para la construcción de intervalos de confianza es crucial la siguiente definición:

Definición 6.7 (Pivote)**.**
Sea $X = (X_1, ..., X_n)$ una muestra aleatoria simple de una variable aleatoria cualquiera de depende de un parámetro θ, el cual puede ser escalar o un vector. Sea $g(X, \theta)$ una variable aleatoria cuya distribución sea la misma para cualquier valor de θ. A la función g se de llama pivote.

Veamos un ejemplo de la construcción de un pivote para el caso de la variable aleatoria normal.

Ejemplo 6.5 (Intervalo de confianza para la media de la distribución normal)**.**
Consideremos la muestra $X_1, \ldots, X_n \sim \mathcal{N}(\theta, 1)$. Como $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}(\theta, 1/n)$ tenemos que

$$\sqrt{n}(\bar{X} - \theta) \sim \mathcal{N}(0, 1). \tag{6.14}$$

Esta cantidad es un *pivote*. Consecuentemente, podemos identificar directamente un intervalo de confianza para el pivote desde una tabla de valores para la distribución normal de media cero y varianza unitaria. Si $\phi(x)$ denota la distribución Normal, entonces podemos elegir x_1 y x_2 tal que $\phi(x_2) - \phi(x_1) = 1 - \alpha$ con lo que tenemos

$$\mathbb{P}_\theta\left(x_1 \leq \sqrt{n}(\theta - \bar{X}) \leq x_2\right) = 1 - \alpha \Leftrightarrow \mathbb{P}_\theta\left(\bar{X} + x_1/\sqrt{n} \leq \theta \leq \bar{X} + x_2/\sqrt{n}\right) = 1 - \alpha, \tag{6.15}$$

es decir, $(\bar{X} + x_1/\sqrt{n}, \bar{X} + x_2/\sqrt{n})$ es el $(1-\alpha)$-intervalo de confianza para θ.

Eligiendo $\alpha = 0{,}05$ una alternativa es tenemos $x_2 = -x_1 = 1{,}96$, con lo que el intervalo de confianza del 95 % para θ está dado por

$$(\bar{X} - 1{,}96/\sqrt{n}, \bar{X} + 1{,}96/\sqrt{n}). \tag{6.16}$$

El procedimiento estándar para encontrar intervalos de confianza es como el ilustrado en el ejemplo anterior, en donde construimos una cantidad que tiene una distribución que no depende del parámetro desconocido (llamada pivote). Construir un intervalo de confianza para esta cantidad es directo desde las tablas de distribuciones, luego, podemos encontrar el intervalo de confianza para la cantidad deseada, e.g., el parámetro desconocido, mediante transformaciones de la expresión del pivote.

Observación 6.6.
El intervalo de confianza no es único. Por ejemplo, en el caso gaussiano podemos elegir en intervalo centrado en cero o desde $-\infty$. Esta elección dependerá de las aplicación en cuestión: una regla general es elegirlo de forma centrada para densidades que son simétricas, centrado en la moda para distribuciones unimodales, mientas que para densidades con soporte positivo podemos elegirlo desde cero.

Hasta ahora hemos solo definido intervalos de confianza para cantidades escalares, en donde el concepto de intervalo tiene sentido. Para parámetros vectoriales, nos referiremos a *conjuntos de confianza*. Siguiendo la Definición 6.6, un $(1-\alpha)$-conjunto de confianza $S(X)$ es tal que

$$\mathbb{P}_\theta\left(\theta \in S(X)\right) = 1 - \alpha, \forall \theta \in \Theta. \tag{6.17}$$

Ejercicio 6.3.
Considere $X_1, \ldots, X_{50} \sim \mathcal{N}(0, \sigma^2)$, calcule el intervalo de confianza del 99 % para σ.

Ejemplo 6.6 (Intervalo de confianza —aproximado— para Bernoulli)**.**

Consideremos $X_1, \ldots, X_n \sim \text{Ber}(\theta)$ y calculemos un intervalo de confianza para θ. Recordemos que el EMV es $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$ y debido a la normalidad asintótica del EMV, tenemos que para n grande, podemos asumir

$$\hat{\theta} \sim \mathcal{N}\left(\theta, \frac{\theta(1-\theta)}{n}\right), \tag{6.18}$$

donde la varianza $\frac{\theta(1-\theta)}{n} = I_n(\theta)^{-1}$ es la inversa de la información de Fisher.

Podemos entonces considerar el pivote

$$\frac{\sqrt{n}(\hat{\theta}-\theta)}{\sqrt{\theta(1-\theta)}} \sim \mathcal{N}(0,1), \tag{6.19}$$

y calcular el $(1-\alpha)$-intervalo de confianza asumiendo los valores x_1 y x_2 mediante

$$\mathbb{P}_\theta\left(x_1 \leq \frac{\sqrt{n}(\theta-\hat{\theta})}{\sqrt{\theta(1-\theta)}} \leq x_2\right) = 1-\alpha \Leftrightarrow \mathbb{P}_\theta\left(\hat{\theta} + \frac{x_1\sqrt{\theta(1-\theta)}}{\sqrt{n}} \leq \theta \leq \hat{\theta} + \frac{x_2\sqrt{\theta(1-\theta)}}{\sqrt{n}}\right) = 1-\alpha.$$

Sin embargo, los bordes de este intervalo no son conocidos, pues dependen de θ. Una forma de aproximar el intervalo es reemplazar el parámetro por su EMV.

Ejercicio 6.4 (Encuesta de elecciones presidenciales)**.**

Considere una encuesta que ha consultado a 1000 votantes y su candidato ha recibido 200 votos. Use el resultado del ejemplo anterior para determinar el intervalo de confianza del 95 % de la cantidad de votos que su candidato obtendría en la elección presidencial.

Finalmente, revisaremos el siguiente ejemplo, el cual pretende ejemplificar el concepto de que en solo un experimento, la determinación del $(1-\alpha)$-intervalo de confianza no quiere decir que la probabilidad de que el parámetro esté contenido en él es $(1-\alpha)\,\%$.

Ejemplo 6.7 (Intervalo de confianza para una distribución uniforme)**.**

Considere $X_1, X_2 \sim \text{Uniforme}\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$ y observe que

$$\begin{aligned}\mathbb{P}_\theta\left(\text{mín}(X_1,X_2) \leq \theta \leq \text{máx}(X_1,X_2)\right) &= \mathbb{P}_\theta\left(X_1 \leq \theta \leq X_2\right) + \mathbb{P}_\theta\left(X_2 \leq \theta \leq X_1\right)\\ &= \frac{1}{2}\times\frac{1}{2} + \frac{1}{2}\times\frac{1}{2}\\ &= \frac{1}{2}\end{aligned}$$

corresponde al intervalo del 50 %.

Sin embargo, si observáramos $X_1 = x_1$ y $X_2 = x_2$ tal que $|x_1 - x_2| \geq \frac{1}{2}$ entonces necesariamente está contenido en el intervalo $(\text{mín}(X_1,X_2), \text{máx}(X_1,X_2))$ con probabilidad 1. Esto ilustra la idea de que, en un experimento dado, la probabilidad de que θ esté en intervalo de confianza del $(1-\alpha)\,\%$ no es necesariamente $(1-\alpha)\,\%$.

6.7. Ejercicios

1. Se quiere estudiar el comportamiento de un vector bidimensional que tiene sus dos componentes ortogonales, independientes y que siguen una distribución normal. Al realizar las mediciones respectivas de cada componente, se obtiene una Muestra Aleatoria Simple (MAS, cada dato es generado desde una misma distribución y son independientes entre sí (iid)) $U = (U_1, ..., U_n)$ de n observaciones con $U_n \sim \mathcal{N}(0, \sigma^2)$ y una MAS $W = (W_1, ..., W_n)$ de n observaciones con $W_n \sim \mathcal{N}(0, \sigma^2)$. En específico, se busca estudiar el comportamiento de los módulos de los vectores obtenidos. Se obtiene una nueva MAS $X = (X_1, ..., X_n)$ dada por:

$$X_i = \sqrt{U_i^2 - W_i^2}$$

   i. Encuentre la función de densidad de X_1

   a) Encuentre un estadístico suficiente para σ^2.

   b) Encuentre un estadístico suficiente minimal para σ^2.

   c) Encuentre el Estimador de Máxima Verosimilitud de σ^2.

   d) ¿Es el estimador $\widehat{\sigma^2}_{EMV}$ insesgado? Si no lo es, modifíquelo para que así sea.

   e) Encuentre un EIVUM de σ^2.

   f) Encuentre la distribución de $\sqrt{n}(\widehat{\sigma^2}_{MLE} - \sigma^2)$

2. Estudiaremos la varianza σ^2 de una Muestra Aleatoria Simple (MAS) $X = (X_1, ..., X_n)$ donde $X_i \sim \mathcal{N}(\mu, \sigma^2), \forall i = 1, ..., n$. Se considera que μ y σ son parámetros desconocidos.

   Considere el siguiente estimador de la varianza dado por:

$$S^2 := \frac{1}{n-1} \sum_{i=1}^{n} (\overline{X}_n - X_i)^2$$

   donde $\overline{X}_n$ denota al promedio de $X_1, ..., X_n$, es decir $\overline{X}_n = \sum_{i=1}^{n} X_i$.

   i. Demuestre que $\mathbb{E}(S^2) = \sigma^2$

   ii. Calcule la varianza de S^2, para esto encontraremos primero la distribución de $\frac{n-1}{\sigma^2} S^2$, siga los siguientes pasos:

   ii.a. Desarrolle la expresión $W = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2}$ para llegar a:

$$W = \frac{(n-1)}{\sigma^2} S^2 + \frac{n(\overline{X}_n - \mu)^2}{\sigma^2}$$

ii.b. Encuentre las distribuciones asociadas a W y a $\frac{n(\overline{X}_n-\mu)^2}{\sigma^2}$

ii.c. Aplique la función generadora de momentos en ambos lados de la ecuación. Para esto, asuma que S^2 es independiente de $\overline{X}_n$.

ii.d. Encuentre la distribución de $\frac{n-1}{\sigma^2}S^2$

ii.e. Calcule $\mathbb{V}(S^2)$

iii. (Ejercicio) Calcule la varianza de S^2 desarrollando *a mano* (Muy largo).

Ahora se considera otro estimador de la varianza dado por:

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

iv. Muestre que $\hat{\sigma}^2$ cumple que $\mathbb{E}(\hat{\sigma}^2) \neq \sigma^2$

v. Muestre que $\hat{\sigma}^2$ cumple que $\lim_{n\to\infty} \mathbb{E}(\hat{\sigma}^2) = \sigma^2$

vi. Calcule $ECM_{\sigma^2}(S^2)$

vii. Calcule $ECM_{\sigma^2}(\hat{\sigma}^2)$

viii. Verifique que $ECM_{\sigma^2}(\hat{\sigma}^2) < ECM_{\sigma^2}(S^2)$

3.

4. Sea $X = (X_1, ..., X_n)$ una MAS con una función de densidad dada por:

$$f(x|\theta) = \theta x^{\theta-1}, \theta > 0, x \in (0,1)$$

- Encuentre el estimador MLE para θ. Calcule su esperanza y su varianza. ¿Cómo se comporta la varianza asintoticamente?

(vii) ¿Es $\hat{\theta}_{MLE}$ insesgado para θ? Si no lo es, modifíquelo con tal de que sea insesgado.

(viii) Calcule el ECM de $\hat{\theta}_{MLE}$ y del estimador encontrado en la parte anterior. Basado en el criterio ECM, ¿Cúal de los dos estimadores es preferible?

(iX) Denotemos como $\theta^\star$ al estimador encontrado en la parte (vii). ¿$\theta^\star$ alcanza la Cota de Cramer-Rao?

(X) Si el estimador $\theta^\star$ no alcanza la Cota de Cramer Rao, ¿Tenemos que buscar un mejor estimador insesgado para θ?

5. Sea $X = (X_1, ..., X_n)$ una MAS con una función de densidad dada por:

$$f(x|\theta) = \frac{2x}{\theta^2}, \theta > 0, 0 < x \leq \theta$$

(i) Encuentre un estadístico suficiente para θ.

(ii) Encuentre un estadístico minimal suficiente para θ.

(iii) Encuentre el estimador MLE para θ

(iv) Calcule el ECM de $\hat{\theta}_{MLE}$

(v) Si $\hat{\theta}_{MLE}$ es sesgado, modifíquelo para que sea insesgado. Denotémoslo como $\theta^{\star}$

(vi) ¿Es $\theta^{\star}$ el mejor estimador insesgado de θ?

6. Estudios relacionados con el comportamiento de ciertos bichos indican que estos tienden a organizarse al azar, linealmente, en un intervalo de longitud $\theta > 0$, a la derecha de un punto donde se ubica una feromona. Nos gustaría estimar el valor del parámetro θ. Sea $X = (X_1, ..., X_n)$ una muestra aleatoria simple (MAS) de la distancia de n bichos con respecto a la feromona.

   a) Defina el modelo paramétrico correspondiente.

   b) Considere el estimador $\hat{\theta} = 2\bar{X}_n$. ¿Será insesgado? Si no lo es, modifíquelo para que lo sea.

   c) Ahora, considere el estimador $\hat{\theta} = \text{máx}\{X_1, ..., X_n\}$. ¿Será insesgado? Si no lo es, modifíquelo para que lo sea.

   d) Calcule el ECM para cada uno de los estimadores y compárelos.

7. Considere un modelo lineal dado por $y_i = \beta x_i + \epsilon_i$ con $i = 1, ..., n$. Se tiene que ϵ_i son i.i.d. con $\mathbb{V}(\epsilon_i)) = \sigma^2$, $\mathbb{E}(\epsilon_i) = 0$ y que $X_i \perp \epsilon_i$, $\forall i$. Para estimar β se han sugerido los siguientes estimadores:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} \qquad \hat{\beta}_2 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i}$$

   Estudie el sesgo de cada estimador. Calcule la varianza de cada uno de los estimadores y compárelos con la varianza del estimador de MCO.

8. Considere el modelo lineal en forma matricial dado por:

$$Y = X\beta + \epsilon$$

   Con $\beta \in \mathbb{R}^k$, $\mathbb{E}(\epsilon) = 0_{nx1}$ y $\mathbb{V}(\epsilon) = \sigma^2 Id_n$, con σ un parámetro desconocido. Pruebe las siguientes propiedades del modelo:

   - $X^t\epsilon = 0$
   - $Y^tY = \hat{Y}^t\hat{Y} + \hat{\epsilon}^t\hat{\epsilon}$
   - $\hat{\epsilon}^t\hat{\epsilon} = Y^tY - \hat{\beta}X^tY$

9. **Un poco de Geometría** Llamemos E_X al espacio generado por las columnas de X. Llamemos P_X a la matriz que proyecta a un vector de $\mathbb{R}^n$ en el espacio generado por las columnas de X. Llamemos $E_{X^\perp}$ al espacio ortogonal a E_X. Se puede demostrar que la matriz que proyecta en el espacio generado por las columnas de X y la proyección en el ortogonal vienen dadas por:

$$P_X = X(X^tX)^{-1}X^t \qquad P_{X^\perp} = Id - X(X^tX)^{-1}X^t$$

Escriba los siguientes en términos de P_X y $P_{X^\perp}$

- $\hat{\epsilon} = Y - X\hat{\beta}$
- $\hat{Y} = P_X Y$

Se postula el siguiente estimador para la varianza de los errores σ^2_ϵ:

$$\hat{\sigma^2_\epsilon} = \hat{\epsilon}^t\hat{\epsilon}$$

Escriba el estimador en términos de P_X y $P_{X^\perp}$ y estudie su sesgo.

Capítulo 7

Test de Hipótesis

7.1. Teoría de decisiones

En términos generales, la teoría de decisiones estudia las acciones que puede tomar un agente en un escenario dado. En este contexto afloran de forma natural los conceptos de incertidumbre (de aspectos clave del escenario), funciones de pérdida y procedimientos de decisión. En estadística, podemos identificar al menos los siguientes problemas de decisión.

- **Estimación**: Decidir el valor apropiado para un parámetro desconocido usando datos X y un distribución condicional P_θ
- **Test**: Decidir la hipótesis correcta usando datos $X \sim P_\theta$

$$H_0 : P_\theta \in \mathcal{P}_0 \tag{7.1}$$
$$H_1 : P_\theta \notin \mathcal{P}_0 \tag{7.2}$$

- **Ranking**: Elaborar una lista ordenada de ítems, por ejemplo, productos evaluados por una muestra de la población, resultados de eventos deportivos o juegos online.
- **Predicción**: Estimar/decidir el valor de una variable dependiente en base a observaciones de observaciones pasadas.

Como se puede apreciar, la teoría de decisiones presenta un contexto general para abordar una gran cantidad de situaciones. A continuación se describen los elementos básicos de un problema de decisión, en donde, con fines ilustrativos, ponemos como ejemplo su contraparte en el problema de estimación.

- $\Theta = \{\theta\}$ es el espacio de estado, donde la cantidad θ es el *estado del mundo*. En el problema de estimación, donde convenientemente se ha usado la misma notación, θ es el parámetro del modelo
- $\mathcal{A} = \{a\}$ es el espacio de acciones, donde a es la acción a tomar por el estadístico. En estimación, podemos usar una notación simplificada y considerar acción a como la elección del valor a para el parámetro θ.

- $L(\theta, a)$ es la función de pérdida asociada a tomar la decisión a cuando el estado es θ; nótese que $L : \Theta \times \mathcal{A} \to \mathbb{R}$. En el caso de estimación, usualmente consideramos la pérdida cuadrática:
$$L(\theta, a) = (\theta - a)^2. \tag{7.3}$$

Ejemplo 7.1.

(Inversión bajo incertidumbre) Consideremos los estados $\Theta = \{\theta_1, \theta_2\}$, donde θ_1 quiere decir mercado sano y θ_2 quiere decir mercado no sano. Se debe elegir una estrategia de inversión del siguiente conjunto $\mathcal{A} = \{a_1, a_2, a_3, a_4, a_5, a_6\}$, en base a la siguiente función de pérdida $L(\theta, a)$.

$L(\theta, a)$	a_1	a_2	a_3	a_4	a_5
θ_1	-4	-4	-1	2	4
θ_2	4	0	-1	-6	-4

Notemos que

- a_1, a_2 son buenos cuando $\theta = \theta_1$
- a_4, a_5 son buenos cuando $\theta = \theta_2$
- a_3 es medianamente bueno (pérdida negativa) siempre

entonces, ¿cómo elegimos la acción?

Además de los elementos básicos del problema de decisión (estado, acciones y pérdida), en el enfoque estadístico de la teoría de decisiones existen los siguientes elementos:

- $X \sim P_\theta$ es la variable aleatoria, la cual define la distribución condicional, el espacio muestral, la densidad, etc.
- $\delta(X)$ es el procedimiento de la decisión, es decir, el mapa que asocia una observación $X = x$ con la acción a:
$$\delta(\cdot) : \mathcal{X} \to \mathcal{A}. \tag{7.4}$$
- $\mathcal{D} = \{\delta : \mathcal{X} \to \mathcal{A}\}$ es el espacio de decisiones
- $R(\theta, \delta)$ es el riesgo asociado a δ y θ, el cual está definido como el valor esperado de la pérdida incurrida al tomar la acción $\delta(X)$ cuando el parámetro es θ. Es decir,
$$R(\theta, \delta) = \mathbb{E}_\theta \left(L(\theta, \delta(X)) \right). \tag{7.5}$$

Ejemplo 7.2.

Volviendo al contexto del problema de estimación, consideremos el uso de una VA $X \sim \mathcal{N}(\theta, 1), \theta \in \mathbb{R}$, para encontrar el valor del parámetro θ. En el contexto de teoría estadística de decisiones, el espacio de posibles acciones es precisamente en espacio de

parámetros, es decir,

$$\mathcal{A} = \Theta = \mathbb{R}. \tag{7.6}$$

Elegimos además la pérdida cuadrática, $L(\theta, \hat{\theta}(X)) = (\theta - \hat{\theta}(X))^2$, asociada a estimar θ mediante $\hat{\theta}(X)$. Consideremos que el espacio de acciones está dado por versiones escaladas de la observación X, es decir,

$$\mathcal{A} = \{cX, c \in [0,1]\}. \tag{7.7}$$

Con esta forma del estimador, podemos calcular el riesgo asociado mediante

$$R(\theta, \hat{\theta}) = \mathbb{E}_\theta\left((\hat{\theta}(X) - \theta)^2\right) = \mathbb{V}_\theta\left(\hat{\theta}(X)\right) + \mathbb{E}_\theta\left(\hat{\theta}(X) - \theta\right)^2 = c^2 + (c-1)^2\theta^2 \tag{7.8}$$

¿Qué valor de c sugiere elegir?

Observación 7.1.

La elección del parámetro c en el ejemplo anterior no es trivial. Denotando el procedimiento δ_c (que asigna la acción $a = cX$), notemos que δ_1 domina cualquier procedimiento $\delta_c, c \geq 1$, lo que quiere decir que el procedimiento δ_c es **inadmisible** para $c \in [1, \infty]$. Para el resto del intervalo, ningún procedimiento $\delta_c, c \in [0,1]$ domina a otro, lo que quiere decir que son **admisibles**.

7.2. Intuición en un test de hipótesis

El objetivo del análisis estadístico es obtener conclusiones razonables mediante el uso de observaciones, como también aseveraciones precisas sobre la incertidumbre asociada a dichas conclusiones. De forma ilustrativa, consideremos el siguiente escenario hipotético.

En base a estudios preliminares, se sabe que los pesos de los recién nacidos (RN) en Santiago, Chile, distribuyen aproximadamente normal con promedio 3000gr y desviación estándar de 500gr. Creemos que los RNs en Osorno pesan, en promedio, más que los RNs en Santiago. Nos gustaría formalmente aceptar o rechazar esta hipótesis.

Intuitivamente, una forma de evaluar esta hipótesis es tomar una muestra de RNs en Osorno, calcular su peso promedio y verificar si éste es *significativamente mayor* que 3000gr. Asumamos que hemos tenido acceso al peso de 50 RNs nacidos en Osorno, los cuales exhiben un preso promedio de 3200gr. ¿Podemos entonces concluir directamente y decir que efectivamente los RNs de Osorno pesan más que los de Santiago? Si bien esta es una posibilidad, una postura más escéptica podría argumentar que el obtener una población de 50 RNs con peso promedio de 3200gr es perfectamente plausible de una población de RNs distribuidos de acuerdo a $\mathcal{N}(3000, 500^2)$. Entonces, ¿cómo justificamos la plausibilidad de este resultado?

Para esto distingamos entre las dos hipótesis:

- H_1: Los RNs en Osorno pesan en promedio más de 3000gr (esta es la hipótesis alternativa)

- H_0: Los RNs en Osorno pesan en promedio 3000gr (esta es la hipótesis nula)

Para decidir cuál es verdadera, trataremos de *falsificar* H_0. La forma de hacer esto es calcular la probabilidad de obtener el resultado observado bajo el supuesto que H_0 es cierta. En este caso, sabemos que una muestra

$$X = X_1, \ldots, X_{50} \sim \mathcal{N}(3000, 500^2), \tag{7.9}$$

tiene una media que está distribuida de acuerdo a la siguiente densidad

$$\bar{X} = \frac{1}{50}\sum_{i=1}^{50} X_i \sim \mathcal{N}(3000, 500^2/50). \tag{7.10}$$

Entonces, cuál es la probabilidad de que la la distribución anterior haya generado una muestra $\bar{X} \geq 3200$? Para calcular esto, construyamos el **pivote** (también conocido como z-test)

$$z = \frac{\bar{X} - 3000}{500/\sqrt{50}} \sim \mathcal{N}(0, 1), \tag{7.11}$$

con el cual podemos realizar el cálculo:

$$\mathbb{P}\left(\bar{X} \geq 3200\right) = \mathbb{P}\left(z = \frac{\bar{X} - 3000}{500/\sqrt{50}} \geq 2\sqrt{2}\right) = 0{,}0023388674905235884, \tag{7.12}$$

donde el valor de esta probabilidad puede ser calculado usando la función[1] `cdf` de SciPy mediante la siguiente instrucción.

```
from scipy.stats import norm
import numpy as np
print(1-norm.cdf(2*np.sqrt(2)))
```

Concluimos entonces que la probabilidad de que una muestra de 50 RNs exhiban un promedio de peso mayor o igual a 3200gr, bajo el supuesto que H_0 es cierta, es del orden del 0.23 %.

Nos referiremos a esta probabilidad como **p-valor**, el cual nos dice cuán verosímil es obtener la observación dada bajo el supuesto de que la hipótesis nula H_0 es cierta. Mientras más pequeño es el p-valor, entonces más fuerte es la evidencia en contra de H_0. Entonces nos encontramos ante dos posibles explicaciones:

- H_0 es falsa
- hemos obtenido un resultado que solo ocurre una de cada 434 veces.

Nos referiremos a significancia del test α al umbral para el p-valor en el cual se rechaza el test. En general, este umbral es del 1 % o del 5 %, sin embargo esto depende de la aplicación en cuestión. Por ejemplo, si estamos considerando la evaluación de la vacuna para Covid-19 debemos ser muy estrictos. Entonces necesariamente nuestro nivel de significancia debe ser

[1] Acrónimo de *cumulative denstiy function*.

muy bajo, lo que quiere decir que la hipótesis nula requiere mucha evidencia en su contra para ser rechazada.

En un test de hipótesis hay dos tipos de errores posibles: El error de Tipo I en el cual H_0 es rechazada a pesar de que es verdadera, y el error de Tipo II donde H_0 no es rechazada a pesar de que es falsa (lo cual diremos que tiene probabilidad β). Los tipos de errores se definen mediante la siguiente Tabla y Figura

	H_0 es cierto	H_0 no es cierto
se rechaza H_0	*false positive* o error Tipo I (α)	*true positive* ($1-\beta$)
no se rechaza H_0	*true negative* ($1-\alpha$)	*false negative* o error Tipo II (β)

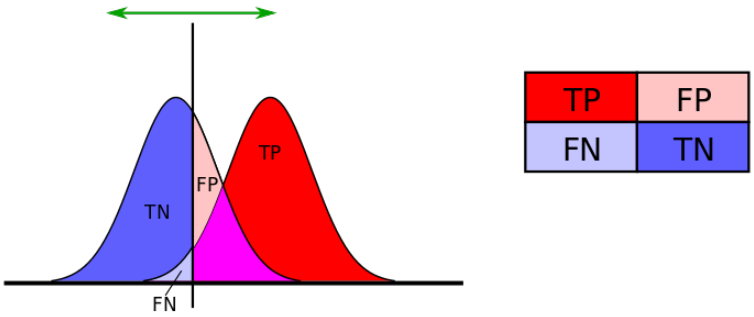

Figura 7.1: Ilustración de tipos de errores (adaptada de Sharpr - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=44059691)

Volviendo a nuestro ejemplo de los recién nacidos, el p-valor del test es del orden de 0.0023, lo cual, si consideramos una significancia del $\alpha = 0{,}01 = 1\,\%$, resulta en el rechazo de H_0. Decimos entonces que **hay suficiente evidencia para rechazar** H_0 **al 1 %** (falsificación de la hipótesis nula), o bien que **rechazamos la hipótesis nula** H_0 **al 1 %** (doble negación). Por el contrario, en el caso que el p-valor fuese mayor que el nivel de significancia del test, entonces no rechazamos H_0 y simplemente decimos que **la evidencia para rechazar** H_0 **no es significativa al 1 %**. Es importante notar que solo podemos **no rechazar** la hipótesis nula, mas no confirmarla.

Test de Hipótesis

En resumen, un test de hipótesis consta de los siguientes pasos:

1. Proponer una hipótesis alternativa H_1
2. Construir una hipótesis nula H_0 (básicamente lo contrario de H_1)
3. Recolectar datos
4. Calcular el pivote (un estadístico de prueba) usando los datos
5. Calcular el p-valor para el pivote
6. Comparar el p-valor con la significancia estadística.
7. Rechazar si corresponde

ADVERTENCIA: Existe la mala costumbre de usar métodos de Test de Hipótesis, incluso cuando no corresponde. Comúnmente, usar estimación e intervalos de confianza son mejores herramientas. Sólo se debe usar Test de Hipótesis cuando se tiene una hipótesis bien definida.

Sobre p-valor y región crítica. Otra forma de cuantificar la evidencia en contra de H_0 es mediante la identificación de una región crítica, es decir, un subconjunto de $\mathcal{X}$ en donde, de tomar valores la observación (o el estadístico), su p-valor estaría por debajo del nivel de significancia y consecuentemente H_0 se rechazaría. En el ejemplo anterior, este puede ser calculado usando la función de SciPy `ppf`[2]. Considerando una significancia del 1 % podemos ejecutar

```
1 from scipy.stats import norm
2 print(norm.ppf(0.99))
```

lo cual nos da una región crítica $[2{,}326, \infty)$, la cual contiene a nuestro umbral $2\sqrt{2} = 2{,}82$; concluimos de igual forma y rechazamos H_0 al 1 %.

Observación 7.2.

Un hipótesis de la forma $\{\theta = \theta_0\}$ se dice hipótesis simple. Una hipótesis de la forma $\theta > \theta_0$ o $\theta < \theta_0$ se dice hipótesis compuesta. Un test de la forma :

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta \neq \theta_0$$

se dice bilateral, y un test de la forma:

$$H_0 : \theta \leq \theta_0 \text{ versus } H_1 : \theta > \theta_0$$

se dice unilateral.

Los tests bilaterales son los más comunes. Notemos que en la Figura 7.2 no ocupamos las dos colas de la figura, pues hicimos un test unilateral. Si hubiésemos hecho un test bilateral (*"Los recién nacidos tienen una media distinta a 3000"*), habríamos ocupado las dos colas de la normal.

[2] Acrónimo para *percent point function*.

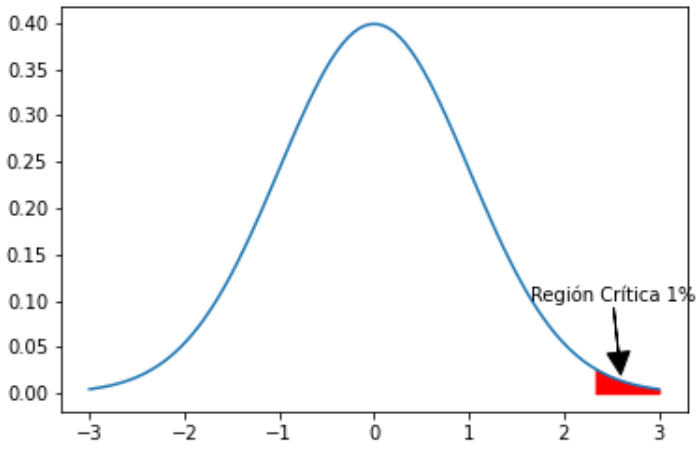

Figura 7.2: La región crítica son aquellos valores que tienen una probabilidad menor al nivel de significancia, en este caso, el 1 %

Falta discusión sobre test simétricos y asimétricos: gráfico ilustrando el uso de p-valor, pivote, significancia y región crítica.

7.3. Rechazo, potencia y nivel

Formalmente, frente a dos hipótesis generales denotadas por

$$H_0 : \theta \in \Theta_0 \tag{7.13}$$
$$H_1 : \theta \in \Theta_1, \tag{7.14}$$

definiremos el problema del test de hipótesis como la búsqueda de una función

$$\phi : \mathcal{X} \to \{0, 1\}, \tag{7.15}$$

donde:

- Si $\phi(x) = 0$ entonces aceptamos H_0 (no rechazamos H_0).
- Si $\phi(x) = 1$ entonces rechazamos H_0, lo cual implícitamente acepta H_1.

En teoría de decisiones, diríamos que ϕ es una regla de decisión.

A continuación, revisamos definiciones que serán de utilidad para analizar y construir tests.

Definición 7.1 (Región crítica de un test)**.**

La región crítica o región de rechazo de un test de hipótesis ϕ se define como

$$R_\phi = \{x \in \mathcal{X} | \phi(x) = 1\} = \phi^{-1}(1). \tag{7.16}$$

Definición 7.2 (Función de probabilidad de rechazo)**.**
Para un test ϕ y cualquier parámetro $\theta \in \Theta$ podemos definir la probabilidad de rechazo mediante

$$\alpha_\phi(\theta) = \mathbb{P}_\theta\left(\phi(x) = 1\right) = \mathbb{P}_\theta\left(x \in R_\phi\right), \forall \theta \in \Theta, \tag{7.17}$$

donde nos gustaría entonces que $\alpha \approx 0$ si $\theta \in \Theta_0$ es cierto y que $\alpha \approx 1$ si $\theta \in \Theta_1$. Luego, usaremos esta función para evaluar la calidad del test.

Definición 7.3 (Potencia de un test)**.**
En el caso que H_1 sea cierta, es decir, $\theta \in \Theta_1$, podemos definir la potencia del test como la probabilidad rechazar H_0 cuando H_1 es efectivamente cierta ($\theta \in \Theta_1$). Es decir,

$$\pi_\phi(\theta) = \mathbb{P}\left(\text{rechazar } H_0 | H_1 \text{ es cierta}\right) = \mathbb{P}_{\theta_1}\left(\phi(x) = 1\right). \tag{7.18}$$

Nos gustaría entonces minimizar $\alpha(\theta)$ cuando H_0 y maximizar $\alpha(\theta)$ cuando H_1, lo cual es equivalente a minimizar la probabilidad de cometer errores de Tipo I y II respectivamente.

Ejemplo 7.3 (Un test absurdo)**.**
Existen tests absurdos, por ejemplo $\phi(x) = 0, \forall x \in \mathcal{X}$. Este test tiene $\alpha(\theta) = 0$ cuando H_0 (lo cual es bueno), pero también tiene potencia nula, es decir, incluso si H_1, no rechaza a H_0.

En general, consideramos más importante prevenir un error de tipo I que uno de tipo II. Es decir, nos protegemos ante el rechazo de H_0 cuando es cierta.

Definición 7.4 (Nivel de un test)**.**
Decimos que un test es de nivel $\alpha \in [0, 1]$ si

$$\alpha_\phi(\theta) \leq \alpha, \forall \theta \in \Theta_0, \tag{7.19}$$

equivalentemente, $\sup_{\theta \in \Theta_0} \alpha_\phi(\theta) \leq \alpha$. Además, denotamos por T_α la clase de todos los tests de nivel α.

Dentro de esta clase, la cual nos restringe únicamente a los test que tienen probabilidad de rechazo acotada superiormente por α para $\theta \in \Theta_0$ (probabilidad de cometer error tipo I), podemos buscar el test de mayor potencia (probabilidad de rechazar H_0 cuando H_1 es cierta). Caracterizamos este test mediante:

Definición 7.5 (Test uniformemente más potente, UMP)**.**
Diremos que $\phi^\star$ es un test UMP (de nivel α) si

$$\pi_{\phi^\star}(\theta) \geq \pi_\phi(\theta), \forall \theta \in \Theta_1. \tag{7.20}$$

7.4. Test de Neyman-Pearson

Consideremos el siguiente problema de test de dos hipótesis simples.

$$H_0 : \theta \in \Theta_0 = \{\theta_0\} \quad \text{v.s.} \quad H_1 : \theta \in \Theta_1 = \{\theta_1\}, \tag{7.21}$$

donde por una notación más simple escribiremos simplemente

$$H_0 : \theta = \theta_0 \quad \text{v.s.} \quad H_1 : \theta = \theta_1, \tag{7.22}$$

y asumiremos que $\mathcal{P} = \{P_\theta \text{ t.q. } \theta \in \Theta\} = \{P_{\theta_0}, P_{\theta_1}\}$ con densidades respectivamente dadas por $p_0(x) = p_{\theta_0}(x)$ y $p_1(x) = p_{\theta_1}(x)$.

Denotamos además la región crítica (donde se rechaza H_0) mediante

$$R^* = \{x \in \mathcal{X} | p_1(x) \geq k p_0(x)\}, \tag{7.23}$$

donde $k \in \mathbb{R}_+$ es una constante a determinar.

Podemos entonces definir el test ϕ^* como el test que tiene el conjunto R^* como región de rechazo, es decir,

$$\phi^*(x) = 1 \Leftrightarrow x \in R^*. \tag{7.24}$$

Finalmente, determinaremos la constante k de tal manera de que

$$\alpha_{\phi^\star}(\theta_0) = \mathbb{P}_{\theta_0}(x \in R^*) = \alpha, \ \alpha \in [0, 1], \tag{7.25}$$

donde, por definición, $\phi^\star \in T_\alpha$. Consecuentemente, de acuerdo al siguiente lema, $\phi^\star$ es el test UMP en T_α.

Lema 7.1 (Neyman-Pearson)**.**
Consideremos un test de hipótesis de la forma

$$H_0 : \theta = \theta_0 \quad \text{v.s.} \quad H_1 : \theta = \theta_1 \tag{7.26}$$

con probabilidad de rechazo dada por

$$\alpha = \mathbb{P}(p_1(X) \geq k p_0(X)). \tag{7.27}$$

Entonces, este test es el UMP de nivel α.

Demostración. Denotemos ϕ^* el test de Neyman-Pearson y R^* su región critica, por definición, $\phi^* \in T_\alpha$. Además, consideremos otro test $\phi \in T_\alpha$ con su propia región crítica R. Recordemos que la probabilidad de los datos estén en la región R es $(\forall \theta)$

$$\mathbb{P}_\theta(R) = \int_R p_\theta(x) \mathrm{d}x. \tag{7.28}$$

Luego, podemos escribir

$$\mathbb{P}_\theta(R) = \mathbb{P}_\theta(R \cap R^*) + \mathbb{P}_\theta(R \cap \bar{R}^*) \tag{7.29}$$
$$\mathbb{P}_\theta(R^*) = \mathbb{P}_\theta(R^* \cap R) + \mathbb{P}_\theta(R^* \cap \bar{R}), \tag{7.30}$$

restando y evaluando para $\theta = \theta_1$, tenemos

$$\begin{aligned}
\mathbb{P}_{\theta_1}(R^*) - \mathbb{P}_{\theta_1}(R) &= \mathbb{P}_{\theta_1}(R^* \cap \bar{R}) - \mathbb{P}_{\theta_1}(R \cap \bar{R}^*) \\
&= \int_{R^* \cap \bar{R}} p_{\theta_1}(x)\mathrm{d}x - \int_{R \cap \bar{R}^*} p_{\theta_1}(x)\mathrm{d}x \\
&\geq k \int_{R^* \cap \bar{R}} p_{\theta_0}(x)\mathrm{d}x - k \int_{R \cap \bar{R}^*} p_{\theta_0}(x)\mathrm{d}x \quad [\text{pues } p_1 \geq k p_0 \text{ en } R^*] \\
&= k\left(\mathbb{P}_{\theta_0}(R^* \cap \bar{R}) - \mathbb{P}_{\theta_0}(R \cap \bar{R}^*)\right) \\
&= k\left(\underbrace{\mathbb{P}_{\theta_0}(R^*)}_{=\alpha} - \underbrace{\mathbb{P}_{\theta_0}(R)}_{\leq \alpha}\right) \quad [\text{primera igualdad de este desarrollo}] \\
&\geq 0
\end{aligned} \tag{7.31}$$

Hemos probado que $\mathbb{P}_{\theta_1}(R^*) \geq \mathbb{P}_{\theta_1}(R)$. Es decir, si $\theta = \theta_1$ entonces $x \in \mathbb{R}^*$ tiene mayor probabilidad que cualquier otra región R. Consecuentemente, el test que tiene a R^* por región crítica es el test UMP.

■

Ejemplo 7.4.

Sea $X_1, \ldots, X_n$ iid $\mathrm{Ber}(\theta), \theta \in \{\theta_0, \theta_1\}$:

$$H_0 : \theta = \theta_0 \quad \text{v.s.} \quad H_1 : \theta = \theta_1. \tag{7.32}$$

Asumamos que $\theta_1 > \theta_0$ y expresemos las densidades de cada hipótesis como

$$p_i(x) = \theta_i^{\sum x_j}(1-\theta_i)^{n-\sum x_j}, \ i = 0, 1. \tag{7.33}$$

Para rechazar H_0 según el test de Neyman-Pearson, es decir, $x \in R^*$ de acuerdo a la ecuación (7.23), el test requiere:

$$\frac{p_1(x)}{p_0(x)} = \frac{\theta_1^{\sum x_j}(1-\theta_1)^{n-\sum x_j}}{\theta_0^{\sum x_j}(1-\theta_0)^{n-\sum x_j}} = \left(\frac{1-\theta_1}{1-\theta_0}\right)^n \left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right)^{\sum x_j} \geq k. \tag{7.34}$$

Como $\theta_1 \geq \theta_0$, la expresión anterior es monótona en $\sum x_j$, consecuentemente, $\sum x_j$ debe ser lo suficientemente grande para rechazar H_0.

Para calcular el valor de k dado un α, tenemos que resolver $\mathbb{P}_{\theta_0}(x \in R^*) = \alpha$, para lo

cual notemos que la ecuación (7.34) es equivalente a

$$\begin{aligned}\left(\frac{1-\theta_1}{1-\theta_0}\right)^n \left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right)^{\sum x_j} \geq k \Leftrightarrow \left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right)^{\sum x_j} &\geq k\left(\frac{1-\theta_0}{1-\theta_1}\right)^n \\ \Leftrightarrow \sum x_j \log\left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right) &\geq n\log\left(k\left(\frac{1-\theta_0}{1-\theta_1}\right)\right) \\ \Leftrightarrow \sum x_j \geq \frac{n\log\left(k\left(\frac{1-\theta_0}{1-\theta_1}\right)\right)}{\log\left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right)} &= k'.\end{aligned} \tag{7.35}$$

Finalmente, como $\sum x_j$ es binomial, podemos resolver directamente para k' (y consecuentemente para k). La región crítica está definida mediante

$$R^* = \left\{(x_1, \ldots, x_1) \text{ t.q. } \sum x_j \geq k'\right\}. \tag{7.36}$$

7.5. Test Parámetricos

7.5.1. Test de razón de verosimilitud

Consideremos un caso más general que los anteriores, donde al menos una de las hipótesis es compuesta, es decir, especifican que el parámetro pertenece a un conjunto en vez de tomar un valor puntual. Es decir,

$$H_0 : \theta \in \Theta_0 \quad \text{v.s.} \quad H_1 : \theta \notin \Theta_0. \tag{7.37}$$

El test de razón de verosimilitud (TRV) indica que se debe rechazar H_0 si

$$\lambda(x_1, \ldots, x_n) = \frac{\sup_{\theta\in\Theta_0} L(\theta)}{\sup_{\theta\in\Theta} L(\theta)} = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})} \leq C, \tag{7.38}$$

donde $\hat{\theta}$ es el EMV y $\hat{\theta}_0$ es el EMV restringido a $\{\theta \in \Theta_0\}$. Claramente, la región de rechazo está dada por

$$R^* = \{x \in \mathcal{X} | \lambda(x) \leq C\}. \tag{7.39}$$

Observación 7.3.

Para el caso de hipótesis simples, es decir, $\Theta = \{\theta_0, \theta_1\}$ y $\Theta_0 = \{\theta_0\}$, entonces el TRV coincide con el test de Neyman-Pearson (TNP). Al igual que en el TNP, en el TRV fijamos C en función del un nivel deseado α.

Observación 7.4.

Notemos que podemos escribir la expresión en la ecuación (7.38) como

$$\lambda(x_1, \ldots, x_n) = \mathbb{1}_{\hat{\theta}\in\Theta_0} + \mathbb{1}_{\hat{\theta}\notin\Theta_0} \frac{\sup_{\theta\in\Theta_0} L(\theta)}{\sup_{\theta\notin\Theta_0} L(\theta)} \tag{7.40}$$

donde el segundo término (de activarse) es estrictamente menor que 1, con lo que el TRV puede enunciarse en función del estadístico

$$\tilde{\lambda}(x_1, \ldots, x_n) = \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \notin \Theta_0} L(\theta)} \leq \tilde{k}. \tag{7.41}$$

Ejemplo 7.5 (TRV Bernoulli)**.**

Sea $X_1, \ldots, X_n \sim \text{Ber}(\theta)$ iid, se quiere resolver

$$H_0 : \theta \leq \theta_0 \quad \text{v.s.} \quad H_1 : \theta > \theta_0, \tag{7.42}$$

donde θ_0 es conocido y sabemos que $p_\theta(x) = \theta^{n\bar{x}}(1-\theta)^{n(1-\bar{x})}$. En la notación de la definición anterior del TRV, podemos identificar

$$\Theta_0 = [0, \theta_0] \quad \& \quad \Theta_1 = (\theta_0, 1] \tag{7.43}$$

calculamos el EMV (restricto e irrestricto) mediante

$$\hat{\theta} = \bar{x} \quad \text{irrestricto} \tag{7.44}$$

$$\hat{\theta}_0 = \bar{x} \quad \text{si } \bar{x} \in \Theta_0, \ \theta_0 \text{ si no.} \tag{7.45}$$

podemos escribir esta última expresión como $\hat{\theta}_0 = \bar{x}\mathbb{1}_{\bar{x} \in \Theta_0} + \theta_0 \mathbb{1}_{\bar{x} \notin \Theta_0}$, entonces

$$\lambda(x) = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})} \tag{7.46}$$

$$= \frac{L(\bar{x})}{L(\bar{x})}\mathbb{1}_{\bar{x} \in \Theta_0} + \frac{L(\theta_0)}{L(\bar{x})}\mathbb{1}_{\bar{x} \notin \Theta_0} \tag{7.47}$$

$$= \mathbb{1}_{\bar{x} \in \Theta_0} + \mathbb{1}_{\bar{x} \notin \Theta_0} \left(\frac{\theta_0}{\bar{x}}\right)^{n\bar{x}} \left(\frac{1-\theta_0}{1-\bar{x}}\right)^{n(1-\bar{x})} \tag{7.48}$$

Donde ahora rechazaremos si $\lambda(x) \leq C$, pero, ¿cómo elegimos C?

Al igual que en TNP, podemos imponer que el test sea de nivel α, es decir,

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta \left(\lambda(x) \leq C\right) = \alpha \tag{7.49}$$

donde recordemos que $\lambda(x)$ es una función decreciente de $\bar{x}$, por lo que la condición $\lambda(x) \leq C$ puede expresarse como $\bar{x} \geq C'$, para algún C'. Esta expresión dependerá de C', que es función de C, de θ_0 y de α; despejamos para C.

Observación 7.5.

En general, los umbrales para los tests de hipótesis son fijados en función del nivel deseado. Consecuentemente, podemos escribir $k = k(\alpha)$ y $C = C(\alpha)$ en TNP y TRV

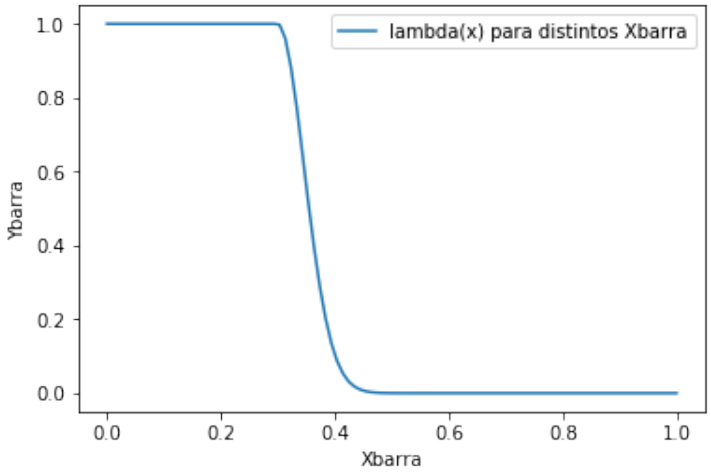

Figura 7.3: $\lambda(x)$ en función de $\bar{x}$

respectivamente.

7.5.2. Test de Proporciones

Sea $\hat{p}_1$ una proporción y $\hat{p}_2$ otra proporción a contrastar, donde cada proporción denota la cantidad de éxitos sobre los casos totales de un evento en dos poblaciones distintas, cada una con n_1 y n_2 casos totales respectivamente. Lo que busca este test es comparar, dependiendo de la hipótesis alternativa, si es que ambas proporciones son distintas o si es que una es mayor o menor a otra. El test se plantea como sigue:

$$H_0 : \hat{p}_1 = \hat{p}_2 \quad H_1 : \hat{p}_1 \neq \hat{p}_2, \quad \hat{p}_1 < \hat{p}_2, \quad \hat{p}_1 > \hat{p}_2$$

Donde se debe escoger sobre una de las hipótesis alternativas H_1. El estadístico se plantea como:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p_0(1-p_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

donde p_0 denota a la proporción de éxitos de la muestra de $n = n_1 + n_2$ datos.

Para poder utilizar este test, generalmente se utiliza la siguiente regla:

$$\text{mín}\left\{n\hat{p}, n(1-\hat{p})\right\} \geq n$$

Al final, lo único que se pide es que z distribuya normal, lo cual por el Teorema Central del Límite (TCL), cuando n sea lo suficientemente grande se tiene lo pedido.

7.5.3. Test de Wald

Este test nos permite evaluar si un parámetro θ toma no un valor θ_0 dado. Consideremos un parámetro escalar y $\hat{\theta}$ un estimador asintóticamente normal, es decir,

$$\frac{\hat{\theta} - \theta_0}{\text{ee}} \sim \mathcal{N}(0,1), \tag{7.50}$$

cuando el número de observaciones tiende a infinito y $\text{ee} = \sqrt{\mathbb{V}\left(\hat{\theta}\right)}$ es conocido como el *error estándar* y puede ser calculado muestralmente o desde p_{θ_0}. Entonces, el test de Wald de tamaño α para las hipótesis

$$H_0 : \theta = \theta_0 \quad \text{v.s.} \quad H_1 : \theta \neq \theta_0, \tag{7.51}$$

indica rechazar H_0 cuando el pivote $W = \frac{\hat{\theta} - \theta_0}{\text{ee}}$ cumple con

$$|W| \geq z_{\alpha/2}, \tag{7.52}$$

donde $z_{\alpha/2} = \Phi(1 - \alpha/2)$, es decir, $\mathbb{P}\left(Z \geq z_{\alpha/2}\right) = \alpha/2,\ Z \sim \mathcal{N}(0,1)$.

Observación 7.6.

Notemos que, asintóticamente, el nivel del test de Wald de tamaño α, es α. En efecto,

$$\mathbb{P}_{\theta_0}\left(|W| \geq z_{\alpha/2}\right) = \mathbb{P}_{\theta_0}\left(\left|\frac{\hat{\theta} - \theta_0}{\text{ee}}\right| \geq z_{\alpha/2}\right) \to \mathbb{P}_{\theta_0}\left(|Z| \geq z_{\alpha/2}\right) = \alpha \tag{7.53}$$

donde, hemos usado que $Z \sim \mathcal{N}(0,1)$.

Ejemplo 7.6.

Consideremos dos conjuntos de VAs $X_1, \ldots, X_n$ y $Y_1, \ldots, Y_m$, con medias respectivas μ_1 y μ_2. Se requiere evaluar las hipótesis

$$H_0 : \mu_x = \mu_y \quad \text{v.s.} \quad H_1 : \mu_x \neq \mu_y, \tag{7.54}$$

lo cual está dentro del alcance del test de Wald denotando $\delta = \mu_x - \mu_y$ e identificando las hipótesis

$$H_0 : \delta = 0 \quad \text{v.s.} \quad H_1 : \delta \neq 0. \tag{7.55}$$

Utilicemos el estimador no-paramétrico 'plug in' de δ dado por $\hat{\delta} = \bar{X} - \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{m}\sum_{i=1}^{m} y_i$. Además, la varianza de este estimador está dada por $v = \frac{1}{n}s_x^2 + \frac{1}{m}s_y^2$ (por el CLT), con lo que el estadístico de Wald es

$$W = \frac{\hat{\delta} - 0}{\text{ee}} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{1}{n}s_x^2 + \frac{1}{m}s_y^2}}, \tag{7.56}$$

y rechazamos rechazamos H_0 si

$$|\bar{X} - \bar{Y}| \geq z_{\alpha/2}\sqrt{\frac{1}{n}s_x^2 + \frac{1}{m}s_y^2}, \tag{7.57}$$

donde recordemos que el lado derecho decae como $1/\sqrt{n}$.

Observación 7.7.

Notemos que el test de Wald de tamaño α rechaza $H_0 : \theta = \theta_0$ (v.s. $H_1 : \theta \neq \theta_0$) si y solo si

$$\theta_0 \notin (\hat{\theta} - \text{ee}z_{\alpha/2}, \hat{\theta} + \text{ee}z_{\alpha/2}), \tag{7.58}$$

es decir, realizar el test de Wald es equivalente a calcular el α intervalo de confianza para el parámetro θ_0 asumiendo normalidad.

7.5.4. Test T de Student

Se le denomina Test T de Student a cualquier test de hipótesis donde bajo la hipótesis nula se tenga que el estadístico a utilizar distribuya como una distribución t-student. Veremos solo un caso de aplicación, pero existen muchos.

One-Sample T-Test

Se tiene una MAS $X_1, ..., X_n$ de una variable aleatoria con media μ y varianza σ^2, donde ambos parámetros son desconocidos. Se tiene que se desea saber si es que μ es igual a un valor μ_0. Para esto, planteamos el siguiente test de hipótesis:

$$H_0 : \mu = \mu_0 \quad H_1 : \mu \neq \mu_0$$

el cual corresponde a un test bilateral. El estadístico a utilizar viene dado por:

$$t = \frac{\overline{X}_n - \mu_0}{S/\sqrt{n}}$$

donde $\overline{X}_n$ representa al promedio muestral y S^2 a la varianza muestral insesgada. Se tiene que bajo la hipótesis nula, t distribuye según una t-student con $n-1$ grados de libertad. En el caso de rechazar el test, se tiene que $\mu \neq \mu_0$.

7.6. Tests no paramétricos y de bondad de ajuste

7.6.1. Test de Mann-Whitney

Este es un test no paramétrico donde se contrasta si es que dos poblaciones son no-distintas o no distinguibles.

Formalmente, consideremos a X una observación de la población 1 y a Y una observación de la población 2, el test se plantea como sigue:

$$H_0 : \mathbb{P}(X > Y) = \mathbb{P}(X < Y) \quad H_1 : \mathbb{P}(X > Y) \neq \mathbb{P}(X < Y)$$

es decir, la hipótesis alternativa es que la probabilidad de que la observación X de una población sea mayor a otra observación Y de otra población sea distinta que la probabilidad de Y sea mayor a X. Consideremos a $X_1, ..., X_n$ muestras iid de X y $Y_1, ..., Y_m$ muestras iid de Y, ambas muestras independiente entre ellas. El estadístico de Mann-Whitney, el estadístico U, se define como:

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} S(X_i, Y_j)$$

donde:

$$S(X,Y) = \begin{cases} 1, & si \quad Y < X \\ \frac{1}{2}, & si \quad Y = X \\ 0, & si \quad Y > X \end{cases}$$

7.6.2. Test de Kruskal-Wallis

Este test es la alternativa no paramétrica del test ANOVA y la generalización del Test de Mann-Whitney para más de dos grupos. El test prueba si un grupo de datos provienen de la misma población. Este test contrasta si las muestras están equidistribuidas. Bajo algunas simplificaciones, puede considerarse que el test compara las medianas.

Supongamos que se tiene una población $k \in [1 : K]$ con n_k observaciones denotadas como $\{x_{k,j}\}_{j=1}^{j=n_k}$. Se denota a $F_k(x)$ la distribución continua de $x_{k,j}$, con $j \in [1 : n_k]$. El test se plantea como sigue:

$$H_0 : F_1(x) = F_2(x) = ... = F_k(x) \tag{7.59}$$
$$H_1 : \exists i, k \in [1 : K], i \neq k : F_i(x) \neq F_k(x) \tag{7.60}$$

La hipótesis nula entonces es que las K distribuciones son iguales y la alternativa es que al menos una de las distribuciones es distinta a otra. Se ordenan las $N = \sum_{k=1}^{K} n_k$ observaciones de menor a mayor, a lo anterior se denomina rankear a los datos. Sea R_{kj} el ranking del dato x_{kj}, se define a R_k como la suma de los rankings de k, es decir, $R_k = \sum_{j=1}^{n_k} R_{kj}$ y se denota al

promedio como $\bar{R}_k = \frac{1}{n_k} R_k$. El promedio total de los rankings es $\bar{R} = \frac{N+1}{2}$.

Con esto, se plantea el estadístico H como:

$$H = \frac{12}{N(N+1)} \sum_{k=1}^{K} n_k (\bar{R}_k - \bar{R})^2$$

la distribución de este estadístico es aproximadamente una $*Chi - Square*$ con $k-1$ grados de libertad. Si se considera un nivel de confianza de $\alpha \in (0,1)$ se rechaza la hipótesis nula si es que:

$$H > \mathcal{X}^2_{k-1,1-\alpha}$$

donde $\mathcal{X}^2_{k-1,1-\alpha}$ es el $1-\alpha$ cuantil de una $Chi - Square$ de $k-1$ grados de libertad. El test de Kruskal-Wallis mide el grado en el cual la media observada de los rankings $\bar{R}_k$ difiere del valor esperado $\bar{R}$. Si esta diferencia es grande, entonces se rechaza la hipótesis nula H_0

Condiciones

1. No es necesario que las muestras tengan una distribución normal por grupos.
2. Todos los grupos deben tener las mismas varianzas, esto se llama homocedasticidad. Esto puede testearse con los test de Levene o Barttlet.
3. Las distribuciones de los grupos deben ser las mismas, esto hace referencia a las curvas de densidad empírica, donde todas deben mostrar la misma tendencia.

7.6.3. Test $\mathcal{X}^2$

Se le denomina Test $\mathcal{X}^2$ a cualquier test de hipótesis donde bajo la hipótesis nula se tenga que el estadístico a utilizar distribuya como una distribución $\mathcal{X}^2$. Veremos solo un caso de aplicación, pero existen muchos.

Prueba $\mathcal{X}^2$ de Pearson

Por completar

7.6.4. Test de Kolmogorov-Smirnov

Ahora consideramos otro enfoque, basado en una estrategia muy distinta, al problema de test de hipótesis anterior para distribuciones no paramétricas. Buscamos determinar si la distribución F de una VA es igual a una distribución de referencia F_0 o no, es decir:

$$H_0 : F = F_0 \quad \text{v.s.} \quad H_1 : F \neq F_0. \tag{7.61}$$

Consideremos además la distribución empírica dada por

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x_j \leq x}, \tag{7.62}$$

la cual realmente es una distribución (discontinua y constante por tramos).

Sabemos que, debido a la ley de los grandes números,

$$F_n(x) \to \mathbb{E}\left(\mathbb{1}_{X\leq x}\right) = \mathbb{P}\left(X \leq x\right) = F(x), \tag{7.63}$$

además, por el teorema de Glivenko-Cantelli, tenemos

$$\sup_x |F_n(x) - F(x)| \to 0 \quad \text{c.s.} \tag{7.64}$$

Lo anterior nos permite definir el estadístico $D_n = \sup_x |F_n(x) - F_0(x)|$ y la región crítica

$$R = \{x | D_n \geq k_\alpha\}, \tag{7.65}$$

donde k_α se elige imponiendo $\mathbb{P}_{\theta_0}\left(D_n \geq k_\alpha\right) = \alpha$.

Observación 7.8.

El test de Kolmogorov-Smirnoff sirve tanto para verificar si una VA sigue una distribución dada o si bien dos distribuciones siguen la misma distribución (desconocida).

7.6.5. Test de Wilcoxon

Este es otro testo no paramétrico para verificar si dos VAs siguen la misma distribución. Consideremos las observaciones

$$X_1, \ldots, X_n \sim F, \quad Y_1, \ldots, Y_n \sim G, \tag{7.66}$$

donde F y G son dos distribuciones, de las cuales solo sabemos que son continuas.

Consideremos el siguiente problema de test de hipótesis:

$$H_0 : F = G \quad \text{v.s.} \quad H_1 : F \neq G. \tag{7.67}$$

El test de Wilcoxon se enfoca en este escenario pero solo es sensible a diferentes *localizaciones*, es decir, si G es una versión desplazada de F.

Antes de ver el test de Wilcoxon, notemos que si nos interesase detectar estas desviaciones, entonces podríamos considerar un test que rechace H_0 si $|\bar{X} - \bar{Y}| \geq K$. Esto es exactamente lo que hace el TRV en el problema

$$H_0 : \mu = \eta \quad \text{v.s.} \quad H_1 : \mu \neq \eta, \tag{7.68}$$

cuando $X \sim \mathcal{N}(\mu, \sigma^2)$, $Y \sim \mathcal{N}(\eta, \sigma^2)$.

Sin embargo, en el caso general (cuando no sabemos nada de F) obtener la ley de $|\bar{X} - \bar{Y}|$ bajo H_0 no es trivial, lo cual es necesario para $\mathbb{P}_{\theta_0}(|\bar{X} - \bar{Y}| \geq K) = \alpha$. En esta situación, el test de Wilcoxon propone considerar la siguiente observación conjunta

$$(z_1, \ldots, z_{m+n}) = (x_1, \ldots, x_n, y_1, \ldots, y_m), \tag{7.69}$$

para luego considerar la secuencia ordenada de valores z_i dados por

$$\min_i\{z_i\} = z_{(1)} \leq z_{(2)} \leq \cdots \leq z_{(n+m)} = \max_i\{z_i\}. \tag{7.70}$$

Ahora podemos definir el concepto de *rango* como la posición en el orden anterior, es decir, donde el rango de $z_{(1)}$ es 1, el rango de $z_{(2)}$ es dos y así sucesivamente.

dibujo de bolas negras y blancas.

Denotando el rango de x_i como R_i, podemos construir el estadístico

$$W = \sum_{i=1}^{n} R_i, \tag{7.71}$$

esta cantidad debe intuitivamente interpretarse como el promedio de los rangos (es decir de la posiciones) que toman las observaciones de la variable X, por rechazamos H_0 si W es muy pequeño o muy grande, es decir, si las muestras de X no quedan *mezcladas* con las de Y.

Esto es posible por que la distribución de W bajo H_0 puede ser calculada y de hecho no depende de F, esto es porque (bajo H_0) los elementos de z_i son iid, con lo que todas las posible permutaciones del los valores z_i tienen la misma probabilidad dada por $\binom{n+m}{n}$.

Observación 7.9 (¿Cómo obtenemos la región crítica R para este test?)**.**
Podemos proceder de forma iterativa: Asumimos H_0, consideramos $R = \emptyset$ y agregamos las configuraciones de bolitas que tienen el menor y mayor valor de W, luego seguimos con las siguientes configuraciones hasta acumular una probabilidad $\mathbb{P}_{\theta_0}(W \in \mathbb{R}) = \alpha$.

7.7. Tests de Homogeneidad

Los tests de homogeneidad tienen como hipótesis nula que la varianza de distintas muestras asociadas a distintos grupos excluyentes son iguales.

7.7.1. Test de Levene

En este test, se consideran muestras asociadas a k grupos excluyentes. Lo que se busca es saber si es que estos k grupos tienen las mismas varianzas muestrales. Se considera lo siguiente:

- N_i es la cantidad de muestras del grupo i,

- N es el número total de muestras,
- Y_{ij} es el valor de la muestra j del grupo i,
- $\overline{Y}_i$ es el promedio muestral del grupo i,
- se considera a $Z_{ij} = |Y_{ij} - \overline{Y}_i|$,
- $Z_{i\cdot} = \frac{1}{N_i} \sum_{j=1}^{N_i} Z_{ij}$
- $Z_{\cdot\cdot} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{N_i} Z_{ij}$

y el estadístico W dado por:

$$W = \frac{(N-k)}{(k-1)} \frac{\sum_{i=1}^{k} N_i (Z_{i\cdot} - Z_{\cdot\cdot})^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} (Z_{ij} - Z_{i\cdot})^2}$$

se tiene que W sigue una distribución F con $k-1$ y $N-k$ grados de libertad.

7.8. Ejercicios

1. Se consideran dos variables aleatorias U y W independientes ambas con distribución $\mathcal{N}(0, \sigma^2)$. En un proceso extraño, se tiene una MAS que posee la información transformada dada por: $X_i = \sqrt{U^2 + W^2}$

   i Encuentre la función de densidad de X.

   ii Se considera la MAS i.i.d. dada por $X = (X_1, ..., X_n)$. Derive el EMV de σ^2.

   iii ¿Cuál es la distribución asintotica del estimador EMV $\hat{\sigma}^2$?

   iv Considere el test dado por:

   $$H_0 : \sigma^2 = \sigma_0^2 \qquad H_1 : \sigma^2 = \sigma_1^2$$

   con $\sigma_1^2 > \sigma_0^2$. Desarrolle el TNP de nivel $\alpha \in (0, 1)$.

   v Considere el test dado por:

   $$H_0 : \sigma^2 = 2 \qquad H_1 : \sigma^2 > 2$$

   Se tiene una muestra de tamaño $n = 100$ con $\sum_{i:1}^{n} x_i^2 = 470$. ¿Qué se concluye del test a un nivel de $\alpha = 0{,}1$?

vi Calcule la potencia del test unilateral dado en el item (iv). Con una muestra de tamaño $n = 100$, $\sigma_0^2 = 2$ $\alpha = 0{,}1$, ¿Cuál es la potencia?.

vii Considere el test bilateral dado por:

$$H_0 : \sigma^2 = \sigma_0^2 \qquad H_1 : \sigma^2 \neq \sigma_1^2$$

Desarrolle el TNP de nivel $\alpha \in (0, 1)$.

viii Determine la potencia del test bilateral encontrado en el item (vii) con n=100, $\sigma_0^2 = 2$ y $\alpha = 0{,}1$.

iX Muestre que el test bilateral es insesgado y consistente.

2. Considere $X_1, ..., X_n$ iid con $X_i \sim \mathcal{N}(\mu, \sigma^2)$, σ conocido. Considere el test simple dado por:

$$H_0 : \mu = \mu_0 \quad \text{v/s} \quad H_1 : \mu = \mu_1$$

Donde $\mu_0 < \mu_1$. Desarrolle el TNP de nivel $\alpha \in (0, 1)$

3. Sean $X_1, ..., X_n$ v.a. iid exponenciales de parámetro λ e $Y_1, ..., Y_m$ v.a. iid exponenciales de parámetro $\delta\lambda$ (ambos grupos son mutuamente independientes). El objetivo es desarrollar el siguiente TRV:

$$H_0 : \delta \geq 1 \quad \text{vs} \quad H_1 : \delta > 1$$

a) Especifique el modelo para las n+m observaciones indicando el espacio de parámetro Θ, la función de verosimilitud $L(x, y, \theta)$ y los conjuntos Θ_0 y Θ_1.

b) Calcule lo siguiente:

$$\sup_{\theta \in \Theta} L(x, y, \theta), \qquad \sup_{\theta \in \Theta_0} L(x, y, \theta)$$

c) Desarrolle el TRV para H_0 vs H_1 y muestre que la región crítica depende del cociente entre $\bar{X}_n$ y $\bar{Y}_n$. Indique los pasos a seguir para imponer la condición de que el test resultante tenga tamaño *alpha*.

.

Capítulo 8

Enfoque bayesiano

En esta sección complementaremos el enfoque visto hasta ahora en cuanto a la incorporación de un modelo para la incertidumbre asociada al parámetro θ. En el paradigma bayesiano, consideraremos que el parámetro es una variable aleatoria, es decir, Θ, la cual para una realización particular tomar el valor $\Theta = \theta$. Esto nos permite información *a priori* sobre la estimación a realizar, lo que permite, en muchos casos, ayudar a la inferencia. Una diferencia conceptual entre ambos enfoques, es que la estadística frecuentista evita la subjetividad, mientras que la estadística bayesiana se basa en la convicción del(a) investigador(a), para emitir juicios sobre una hipótesis.

En este capítulo, se estudiará la estadística bayesiana y se introducirán los mismos conceptos vistos anteriormente desde el punto de vista bayesiano.

8.1. Contexto y definiciones principales

Definición 8.1 (Distribución a priori)**.**
La información, sesgos y cualquier otra característica conocida de Θ codificadas mediante la propia ley de probabilidad de esta VA, la cual tiene densidad $p(\theta)$, nos referimos a esta como la *densidad a priori* o simplemente *prior*.

Con esta definición, podemos ver que la densidad conjunta de las VAs X, Θ pueden ser expresadas combinando la densidad a priori con el modelo visto en las secciones anteriores, es decir,

$$p(x, \theta) = p(x|\theta)p(\theta) \tag{8.1}$$

donde hemos escrito $p(x|\theta)$ en vez de $p_\theta(x)$ para hacer explícito que ahora consideramos el parámetro como una variable aleatoria.

Adicionalmente, con la distribución conjunta en la ecuación (8.1), podemos definir:

Definición 8.2 (Distribución marginal)**.**
La distribución de X, obtenida mediante la desintegración de parámetro Θ del par (X, Θ), es decir

$$p(x) = \int_\Omega p(x|\theta)p(\theta)\mathrm{d}\theta \tag{8.2}$$

es conocida como distribución marginal de X.

Consideremos ahora que tenemos un conjunto de observaciones denotado por $\mathcal{D}$, de un modelo estadístico con parámetro Θ, entonces podemos definir

Definición 8.3 (Función de verosimilitud)**.**
La densidad de probabilidad evaluada en un conjunto de observaciones $\mathcal{D}$ como función del valor del parámetro Θ, es decir

$$L : \Omega \to \mathbb{R} \tag{8.3}$$
$$\theta \mapsto l(\theta) = L_{\mathcal{D}}(\theta) = p(\mathcal{D}|\theta), \tag{8.4}$$

recibe el nombre de función de verosimilitud, o en inglés, *likelihood*.

Observación 8.1.
La función de verosimilitud no es una densidad de probabilidad, es decir, no es cierto que

$$\int_{\Omega} L(\theta)\mathrm{d}\theta = 1 \tag{8.5}$$

Observación 8.2.
Dado que la función función de verosimilitud usualmente adquiere una forma exponencial (como por ejemplo en el caso de la familia exponencial), hay ocasiones en donde es conveniente usar la *log-verosmilitud*, esto es,

$$l(\theta) = \log L(\theta) = \log p(\mathcal{D}|\theta). \tag{8.6}$$

Esta formulación será particularmente útil cuando queramos optimizar la verosimilitud.

Observación 8.3.
En general (pero no siempre) asumimos observaciones $\mathcal{D} = X_1, \ldots, X_n$, $X_i \sim p(x|\theta)$, que son i.i.d. En cuyo caso, la verosimilitud factoriza de la forma $L_{\mathcal{D}}(\theta) = \prod_{i=1}^{n} L_{X_i}(\theta)$, con lo cual la log-verosimilitud toma la forma:

$$l_{\mathcal{D}}(\theta) = \sum_{i=1}^{n} l_{X_i}(\theta) \tag{8.7}$$

Ejemplo 8.1.
Considere los datos $\mathcal{D} = \{x_1, ..., x_n\}$, donde x_i es la observación de una VA $X_i \sim \mathcal{N}(\mu, \sigma^2)$ iid con σ^2 conocido. La función de verosimilitud de μ está dada por:

$$L(\mu) = p(\mathcal{D}|\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-1}{2\sigma^2}(x_i - \mu)^2\right)$$
$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right). \tag{8.8}$$

Luego, la log-verosimilitud está dada por:

$$l(\mu) = \log L(\mu) = -\frac{n}{2}\log(2\pi\sigma^2) + \frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2. \tag{8.9}$$

Ahora estamos en condiciones de definir el elemento central de la inferencia bayesiana, sobre el cual todo el proceso de inferencia toma lugar.

Definición 8.4 (Distribución posterior)**.**

Dado el conjunto de observación $\mathcal{D}$ la distribución *posterior* del parámetro, es decir, considerando la inforamción reportada por los datos $\mathcal{D}$, está dada por el teorema de Bayes mediante

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D})|\theta)p(\theta)}{p(\mathcal{D})} \propto p(\mathcal{D}|\theta)p(\theta) \tag{8.10}$$

donde:

- $p(\theta)$ es el prior del parámetro.
- $p(\theta|\mathcal{D})$ es la posterior del parámetro.
- $p(\mathcal{D}|\theta)$ es la verosimilitud
- $p(\mathcal{D}) = \int \Omega p(\mathcal{D}|\theta)p(\theta)\mathrm{d}\theta$ es la densidad marginal de los datos

La *transición* de prior a posterior puede ser interpretada como el proceso de incorporar la evidencia de los datos (a través de la función de verosimilitud) para reducir la incertidumbre con respecto del valor del parámetro Θ. De la ecuación (8.10) podemos ver que este proceso, a veces referido como *actualización bayesiana*, equivale a multiplicar por la verosimilitud, para luego normalizar, garantizando que $p(\theta|\mathcal{D})$ es en efecto una densidad de probabilidad.

Observación 8.4.

El símbolo $\propto$ en la ecuación (8.10) es usado para indicar que el lado izquierdo es igual al lado derecho salvo una constante de proporcionalidad que depende de $\mathcal{D}$ y no de θ. Con esto, cuando estemos calculando la posterior, solo nos enfocaremos en *una versión proporcional*, pues luego la densidad posterior se puede encontrar mediante la normalización de esta última.

Ejemplo 8.2 (Posterior modelo Bernoulli)**.**

Sea θ la probabilidad de obtener cara al lanzar una moneda, y sean $X_1, ..X_n$ n resultados obtenidos al lanzar la moneda. Si no sabemos nada de θ antes del experimento, hace sentido tomar su prior como una distribución que de igual probabilidad a todo espacio de parámetros, es decir: $\theta \sim Unif(0,1)$. Notemos que el prior encapsula la infromación que tenemos antes del experimento. Modelamos $X_1, ..X_n \sim Bernoulli(\theta)$. Entonces:

$$p(X_1, ..X_n|\theta) = \prod_{i=1}^{n} \theta^{X_i}(1-\theta)^{1-X_i} = \theta^{\sum_{i=1}^n X_i}(1-\theta)^{n-\sum_{i=1}^n X_i}$$

Notemos que en este caso, podemos calcular la distribución $p(X_1, .., X_n)$:

$$p(X_1, .., X_n) = \int_0^1 \theta^{\sum_{i=1}^n X_i}(1-\theta)^{n-\sum_{i=1}^n X_i} d\theta = B\left(\sum_{i=1}^{n} X_i + 1, n - \sum_{i=1}^{n} X_i + 1\right),$$

donde $B(x, y)$ es la función beta:

$$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}.$$

Sea $s = \sum_{i=1}^n X_i$. Entonces la distribución a posteriori será:

$$p(\theta|X_1, ..X_n) = \frac{p(X_1, ..X_n|\theta)}{p(X_1, .., X_n)} = \frac{1}{B(s+1, n-s+1)}\theta^s(1-\theta)^{n-s}.$$

Usualmente, es experimentos reales, los datos $x_1, ..., x_n$ son recibidos de forma secuencial, es decir, *en línea*. De esta forma, es relevante notar que en primer lugar se observa x_1 primero, luego x_2, y así sucesivamente.

Consecuentemente, si se asume el prior para el parámetro θ dado por $p(\theta)$, es posible hacer la actualización bayesiana *en línea* (o de forma adaptativa o continual), lo cual implica una corrección del modelo cada vez que se observan más datos.
Luego de observar x_1, la posterior $p(\theta|x_1)$ puede ser calculada como:

$$p(\theta|x_1) \propto p(x_1|\theta)p(\theta).$$

Luego, al observar x_2, usamos el hecho que X_1 y X_2 son condicionalmente independientes dado θ y obtenemos:

$$p(\theta|x_1, x_2) \propto p(x_2|\theta)p(\theta|x_1) \propto p(x_1|\theta)p(x_2|\theta)p(\theta).$$

Con lo que para el caso general tenemos que

$$p(\theta|x_1, ..x_n) \propto p(x_n|\theta)p(\theta|x_1, ..x_{n-1}) \propto p(\theta)\prod_{i=1}^{n} p(\theta|x_i).$$

Observación 8.5.
Cuando las observaciones $\mathcal{D}$ son condicionalmente independientes dado el parámetro θ, entonces, la posterior $p(\theta|\mathcal{D})$ factoriza en las verosimilitudes de cada uno de los datos.

Observación 8.6.
En la actualización bayesiana en línea, la posterior de la etapa n sirve de prior de la etapa $n+1$.

8.2. Priors Conjugados

La actualización bayesiana puede resultar en una posterior solo conocida de forma proporcional (cuando no es posible calcular la distribución marginal $p(x)$) o bien en una distribución que no pertenece a una familia conocida. Una herramienta que asegurar el cálculo de las distribuciones posteriores (incluyendo la constante de normalización) y que esta adopta una forma conocida es a través del uso de **priors conjugados**.

Definición 8.5.
Sea un modelo con verosimilitud $p(x|\theta)$ y un prior sobre θ con densidad $p(\theta)$. Decimos que $p(\theta)$ es conjugado con la verosimiltud $p(x|\theta)$ si la posterior

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \tag{8.11}$$

pertenece a la*misma familia* que el prior $p(\theta)$. Donde pertenecer a la misma familia quiere decir que ambas tienen una densidad de probabilidad definida por la misma forma funcional, e.g., $f_\lambda(\theta)$ pero con distintos valores para el *parámetro* λ, el cual es un *hiperparámetro* del modelo.

Ejemplo 8.3 (continuación de Ejemplo 8.2)**.**
Tarea: Verifique si el Ejemplo 8.2 es en efecto uno de prior conjugado.

Ejemplo 8.4 (Distribución Multinomial)**.**
Consideremos una variable aleatoria multinomial $X \sim \text{Mult}\,(n, \theta)$ donde θ pertenece al simplex

$$\{\theta \in [0,1]^k : \theta_1 + \cdots + \theta_k = 1\}. \tag{8.12}$$

La distribución multinomial genera vectores $X \in \mathbb{N}^k$ cuya $i-$ésima componente modela la cantidad de veces que ocurre el evento i dentro de k eventos en n intentos. Por ejemplo, si lanzamos un dado balanceado 100 veces, el vector que contiene el conteo de veces que obtenemos cada cara puede modelarse como

$$\theta_{\text{dado}} \sim \text{Mult}\left(100, \left[\tfrac{1}{6}, \tfrac{1}{6}, \tfrac{1}{6}, \tfrac{1}{6}, \tfrac{1}{6}, \tfrac{1}{6}\right]\right). \tag{8.13}$$

Denotando $X = [x_1, \ldots, x_n]$, observemos que una muestra multinomial $X \sim \text{Mult}\,(n, \theta)$

cumple con

$$\{x_i\}_{i=1}^k \subset \{0, 1, \ldots, n\}, \quad \sum_{i=1}^k x_i = n. \tag{8.14}$$

Finalmente, la distribución Multinomial está dada por

$$\text{Mult}\,(X; n, \theta) = \frac{n!}{x_1! \cdots x_k!} \theta_1^{x_1} \cdots \theta_k^{x_k}, \tag{8.15}$$

y es la generalización de las distribuciones:

- Bernoulli cuando $k = 2$ y $n = 1$; pues $\text{Ber}\,(X; \theta) = \theta^x (1-\theta)^{1-x}$
- Categórica (o *multinoulli*): cuando $n = 1$; pues $\text{Cat}\,(X; \theta) = \theta_1^{x_1} \cdots \theta_k^{x_k}$
- Binomial: cuando $k = 2$; pues $\text{Bin}\,(X; n, \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$

Observemos que el parámetro θ en la distribución multinomial (y las otras tres) es precisamente una distribución de probabilidad (discreta). Es decir, el construir un prior $p(\theta)$ implica definir una distribución sobre distribuciones discretas.

Definición 8.6 (Distribución de Dirichlet)**.**
Consideremos la distribución de Dirichlet

$$\theta \sim \text{Dir}\,(\theta|\alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^k \theta_i^{\alpha_i - 1}, \tag{8.16}$$

donde $\alpha = (\alpha_1, \ldots, \alpha_k)$ es el parámetro de concentración y la constante de normalización está dada por $B(\alpha) = \prod_{i=1}^k \Gamma(\alpha_i) / \Gamma(\sum_{i=1}^k \alpha_i)$. El soporte de esta distribución es el simplex presentado en la ecuación (8.12).

En el caso $k = 3$, la distribución de Dirichlet puede ser graficada en el simplex de 2 dimensiones. La Figura 8.1 presenta tres gráficos para distintos valores del parámetro de concentración.

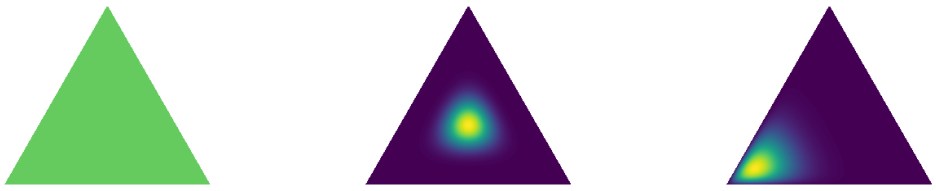

Figura 8.1: Distribuciones Dirichlet para $k = 3$ con parámetros de concentración α (desde izquierda a derecha) dado por $[1, 1, 1]$, $[10, 10, 10]$ y $[10, 2, 2]$.

.

Veamos a continuación que la distribución de Dirichlet es conjugada al modelo Multinomial, y consecuentemente para Bernoulli, Categórica y Binomial. En efecto, si $\theta \sim \text{Dir}\,(\theta; \alpha)$ y $X \sim \text{Mult}\,(X; n, \theta)$, entonces

$$
\begin{aligned}
p(\theta|x) &= \frac{\text{Mult}\,(x;n,\theta)\,\text{Dir}\,(\theta;\alpha)}{p(x)} \\
&= \frac{n!}{x_1!\cdots x_k! p(x) B(\alpha)} \prod_{i=1}^{k} \theta_i^{x_i+\alpha_i-1} \\
&= \frac{1}{B(\alpha')} \prod_{i=1}^{k} \theta_i^{\alpha'_i-1}
\end{aligned} \tag{8.17}
$$

donde $\alpha' = (\alpha'_1, \ldots, \alpha'_k) = (\alpha'_1 + x_1, \ldots, \alpha'_k + x_k)$ es el nuevo parámetro de concentración.

Ejemplo 8.5.

Consideremos $\alpha = [1,2,3,4,5]$ y generemos una muestra de $\theta \sim \text{Dir}\,(\theta|\alpha)$. El siguiente código genera, grafica e imprime esta muestra.

```
import numpy as np
alpha = np.array([1,2,3,4,5])
theta = np.random.dirichlet(alpha)
plt.bar(np.arange(5)+1, theta);
print(f'theta = {theta}')
```

En nuestro caso, obtuvimos los parámetros $\theta = [0{,}034, 0{,}171, 0{,}286, 0{,}185, 0{,}324]$.

Ahora, usaremos un prior Dirichlet sobre θ con $\alpha_p = [1,1,1,1,1]$ para calcular la posterior de acuerdo a la ecuación (8.17). La Figura 8.2 muestra 50 muestras de la distribución posterior para distintas cantidades de observaciones entre 0 y 10^5.

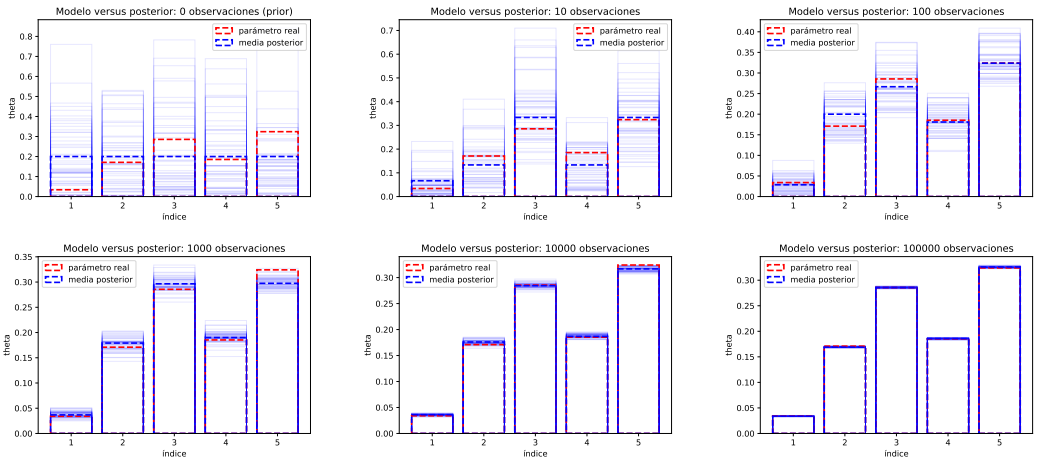

Figura 8.2: Concentración de la distribución posterior en torno al parámetro real para un modelo $X \sim \text{Mult}\,(\theta)$ y una distribución a priori Dirichlet $\theta \sim \text{Dir}\,(\alpha)$. Se considera desde 0 hasta 10^5 observaciones y cada gráfico (desde izquierda-arriba hasta derecha-abajo) muestra el parámetro real (linea roja quebrada), la media posterior (línea azul quebrada) y 50 muestras de la posterior (azul claro). Observe cómo la distribución a priori (línea azul quebrada en la primera figura) pierde importancia a medida que el número de observaciones aumenta.

Ejemplo 8.6.

Modelo gaussiano (σ^2 conocido). Consideremos el prior sobre la media $p(\mu) = \mathcal{N}(\mu_0, \sigma_0^2)$, con lo que la posterior está dada por

$$p(\mu|\mathcal{D}) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right) \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right) \tag{8.18}$$

$$\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right), \tag{8.19}$$

donde la proporcionalidad viene de ignorar la constante $p(\mathcal{D})$ en la primera línea e ignorar todas las contantes que no dependen de μ en la segunda línea. Recordemos que estas constantes para μ incluyen a la varianza de x, σ^2, por lo que ignorar esta cantidad es solo posible debido a que estamos considerando el caso en que σ^2 es conocido. Completando la forma cuadrática para μ dentro de la exponencial en la ec. (8.19), obtenemos

$$p(\mu|\mathcal{D}) \propto \exp\left(-\frac{1}{2\sigma_n^2}(\mu-\mu_n)^2\right), \tag{8.20}$$

donde (ya definiremos μ_n y σ_n^2 en breve) como $p(\mu|\mathcal{D})$ debe integrar uno, la única densidad de probabilidad proporcional al lado derecho de la ecuación anterior es la Gaussiana de media μ_n y varianza σ_n^2. Es decir, la constante de proporcionalidad necesaria para la igualdad en la expresión anterior es

$$\int_{\mathbb{R}} \exp\left(-\frac{1}{2\sigma_n^2}(\mu-\mu_n)^2\right) \mathrm{d}\mu = (2\pi\sigma_n^2)^{n/2}. \tag{8.21}$$

Consecuentemente, confirmamos que el prior elegido era efectivamente conjugado con la verosimilitud gaussiana, con lo que la posterior está dada por la siguiente densidad (gaussiana):

$$p(\mu|\mathcal{D}) = \mathcal{N}(\mu; \mu_n, \sigma_n^2) = \frac{1}{(2\pi\sigma_n^2)^{N/2}} \exp\left(-\frac{1}{2\sigma_n^2}(\mu-\mu_n)^2\right), \tag{8.22}$$

donde la media y la varianza están dadas respectivamente por

$$\mu_n = \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} \left(\frac{1}{\sigma_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{x}\right), \quad \text{donde } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{8.23}$$

$$\sigma_n = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}. \tag{8.24}$$

Observación 8.7.

La actualización bayesiana transforma los parámetros del prior de μ desde μ_0 y σ_0^2 hacia μ_n y σ_n^2 en las ecs. (8.23) y (8.24) respectivamente. Notemos que los parámetros de la posterior son combinaciones (interpretables por lo demás) entre los parámetros del prior y los datos, en efecto, la μ_n es el promedio ponderado entre μ_0 (que es nuestro

candidato para μ antes de ver datos) con factor σ_0^{-2} y el promedio de los datos $\bar{x}$ con factor $(\sigma^2/n)^{-1}$, que a su vez es el estimador de máxima verosimilitud. Es importante también notar que estos factores son las varianzas inversas—i.e., precisión—de μ_0 y de $\bar{x}$. Finalmente, observemos que σ_n es la *suma paralela* de las varianzas, pues si expresamos la ec. (8.24) en términos de *precisiones*, vemos que la precisión inicial σ_0^2 aumenta un término σ^2 con cada dato que vemos; lo cual tiene sentido pues con más información es la precisión la que debe aumentar y no la incertidumbre (en este caso representada por la varianza).

Ejemplo 8.7.

Modelo gaussiano (μ conocido). Ahora procedemos con el siguiente prior para la varianza, llamado Gamma-inverso:

$$p(\sigma^2) = \text{inv-}\Gamma(\sigma^2; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)(\sigma^2)^{\alpha+1}} \exp(-\beta/\sigma^2) \tag{8.25}$$

esta densidad recibe dicho nombre pues es equivalente a modelar la precisión, definida como el recíproco de la varianza $1/\sigma^2$, mediante la distribución Gamma. Los hiperparámetros α y β son conocidos como parámetros de forma y de tasa (o precisión) respectivamente.

Con este prior, la posterior de la varianza toma la forma:

$$\begin{aligned} p(\sigma^2|\mathcal{D}) &\propto \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right) \frac{\beta^\alpha}{\Gamma(\alpha)(\sigma^2)^{\alpha+1}} \exp(-\beta/\sigma^2) \\ &\propto \frac{1}{(\sigma^2)^{N/2+\alpha+1}} \exp\left(-\frac{1}{\sigma^2}\left(\frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^2 + \beta\right)\right) \end{aligned} \tag{8.26}$$

donde nuevamente la proporcionalidad ha sido mantenida debido a la remoción de las constantes. Esta última expresión es proporcional a una distribución Gamma inversa con hiperparámetros α y β ajustados en base a los datos observados.

Hay ocasiones en las que el conocimiento a priori sobre el parámetro no puede ser convenientemente expresado mediante una densidad de probabilidad pero sí una densidad que no necesariamente integra uno o incluso es (Lebesgue) integrable. Para reflejar esta idea, se usan priors impropios.

Definición 8.7 (Prior impropia)**.**

Una distribución a priori impropia es una distribución que no es necesariamente de probabilidad (i.e., no integra 1), pero que de todas formas puede ser utilizada como distribución a priori en el contexto de inferencia bayesiana, pues la distribución posterior correspondiente si es una distribución de probabilidad apropiada.

Observación 8.8.

No es necesario usar la constante de normalización en las densidades a priori Gaussianas (o ninguna otra en realidad).

Observación 8.9.

Veamos que un prior impropio puede incluso tener integral infinita, en el caso de la distribución normal $X \sim \mathcal{N}(X; \mu, 1)$, $\mu \in \mathbb{R}$, podemos elegir $p(\mu) \propto 1$ y escribir

$$p(\mu|x) \propto p(x|\mu) \cdot 1 = \mathcal{N}(x; \mu, 1) = \mathcal{N}(\mu; x, 1). \tag{8.27}$$

Considerar distribuciones uniformes impropias como priors no informativas parece tener sentido, pues intuitivamente no estamos dando preferencia (mayor probabilidad a priori) a ningún valor del parámetro por sobre otro. Sin embargo, este procedimiento sufre de una desventaja conceptual.

8.3. Estimación y predicción

8.3.1. Estimadores bayesianos

Si bien ya hemos estudiado el rol del prior en la inferencia bayesiana, hasta ahora no lo hemos considerado en la construcción de estimadores. En particular, el EMV no incorpora conocimiento a priori del parámetro. Con el objetivo de incorporar este conocimiento a priori en el cálculo de estimadores puntales, consideramos que en el caso general, podemos considerar otros estimadores puntuales a través de una función de pérdida asociada a estimar el parámetro θ mediante el estimador $\hat{\theta}$ dada por $L(\theta, \hat{\theta})$. Con esto podemos definir los conceptos de riesgo y estimador bayesiano.

Definición 8.8 (Riesgo bayesiano)**.**

Para una función de pérdida $L(\theta, \hat{\theta})$ y un conjunto de observaciones $\mathcal{D}$, el riesgo bayesiano es la esperanza posterior de dicha función de pérdida, es decir

$$R(\hat{\theta}) = \int_{\Omega} L(\theta, \hat{\theta}) p(\theta|\mathcal{D}) \mathrm{d}\theta. \tag{8.28}$$

Definición 8.9 (Estimador bayesiano)**.**

Dado un conjunto de datos $\mathcal{D}$ y un riesgo bayesiano $R(\theta)$, un estimador bayesiano es uno que minimiza el riesgo bayesiano:

$$\theta_{\text{Bayes}}(\mathcal{D}) = \arg\min_{\Omega} R(\theta). \tag{8.29}$$

donde es implícito que $R(\cdot)$ se define con $\mathcal{D}$.

A continuación, se definirán estimadores bayesianos con distintas funciones de costo o de riesgo.

Definición 8.10 (Bayes' Least-Squares (BLS))**.**

El caso estándar es la función de pérdida cuadrática $L_2(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ la cual resulta en el estimador dado por la media posterior $\theta_{\text{Bayes}}(\mathcal{D}) = \mathbb{E}(\theta|\mathcal{D})$

Definición 8.11 (Minimum absolute-error (MAE))**.**
De forma similar,la función de costo $L_1(\theta, \hat{\theta}) = |\theta - \hat{\theta}|_1$ resulta en el estimador dado por la mediana posterior.

Encontrar una función de pérdida para el máximo a posteriori es menos directo. Consideremos en primer lugar el caso $\theta \in \Omega$ discreto y la pérdida "0-1"

$$L_{0\text{-}1}(\theta, \hat{\theta}) = \begin{cases} 0 & \text{si } \theta = \hat{\theta}, \\ 1 & \text{si no.} \end{cases}$$

El riesgo de Bayes asociado a $L_{0\text{-}1}(\theta, \hat{\theta})$ (en el caso discreto) toma la forma

$$R(\hat{\theta}) = \mathbb{P}\left(\theta \neq \hat{\theta} | \mathcal{D}\right) = 1 - \mathbb{P}\left(\theta = \hat{\theta} | \mathcal{D}\right), \tag{8.30}$$

lo cual es minimizado eligiendo $\hat{\theta}$ tal que $\mathbb{P}\left(\theta = \hat{\theta} | \mathcal{D}\right)$ es máximo, es decir, el MAP. ¿por qué no es posible proceder de esta forma para el caso continuo? ¿cuál es la función de costo asociada al MAP en el caso continuo?

Definición 8.12 (Estimador máximo a posteriori)**.**
Sea $\theta \in \Theta$ un parámetro con distribución a posteriori $p(\theta|D)$ definida en todo Θ. Entonces nos referiremos a su estimación puntual dada por:

$$\theta_{MAP} = \arg\max_{\Theta} p(\theta|D),$$

como el estimador *máximo a posteriori* (MAP). Se utiliza la siguiente función de costo:

$$C(a, b) = \begin{cases} 1, & |a - b| > 0 \\ 0, \sim & \end{cases}$$

Observación 8.10.
Es posible encontrar el MAP solo teniendo acceso a una versión *proporcional* a la distribución posterior, un escenario usual en inferencia bayesiana, o también mediante la maximización del logaritmo de ésta última. En efecto,

$$\theta_{MAP} = \arg\max_{\theta \in \Theta} p(\theta|\mathcal{D}) = \arg\max_{\theta \in \Theta} p(\mathcal{D}|\theta)p(\theta) = \arg\max_{\theta \in \Theta} \left(\underbrace{\log p(\mathcal{D}|\theta)}_{l(\theta)} + \log p(\theta) \right),$$

donde hemos encontrado la maximización de la función de log-verosimilitud, pero ahora junto al log-prior.

Observación 8.11.
Es relevante notar que el estimador MAP es una *modificación* del EMV, pues ambos

comparten una parte de la misma función objetivo (verosimilitud) con la diferencia que el MAP además incluye el término *log-prior*. Esto puede entenderse como una regularización de la solución del problema de MV, en donde el término adicional puede representar las propiedades del estimador más allá de que las pueden ser exclusivamente revelada por los datos.

Ejemplo 8.8 (Máximo a posterior para el modelo gaussiano)**.**

En particular, para el modelo lineal y gaussiano que hemos considerado hasta ahora, podemos calcular θ_{MAP} para un prior Gaussiano de media cero y varianza σ_θ^2. Éste está dado por (asumimos la varianza del ruido σ_ϵ^2 conocida):

$$\begin{aligned}
\theta^\star_{\text{MAP}} &= \text{argmax } p(Y|\theta, X)p(\theta) \\
\text{[ind., def.]} &= \text{argmax } \prod_{i=1}^{N} \mathcal{N}(y_i; \theta^\top x_i, \sigma_\epsilon^2)\mathcal{N}(\theta; 0, \sigma_\theta^2) \\
&= \text{argmax } \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_\epsilon} \exp\left(\frac{-1}{2\sigma_\epsilon^2}(y_i - \theta^\top x_i)^2\right) \frac{1}{(\sqrt{2\pi}\sigma_\theta)^{M+1}} \exp\left(\frac{-||\theta||^2}{2\sigma_\theta^2}\right) \\
&= \text{argmax } \frac{1}{\sqrt{2\pi}\sigma_\epsilon} \frac{1}{(\sqrt{2\pi}\sigma_\theta)^{M+1}} \exp\left(\sum_{i=1}^{N} \frac{-1}{2\sigma_\epsilon^2}(y_i - \theta^\top x_i)^2 - \frac{||\theta||^2}{2\sigma_\theta^2}\right) \\
\text{[log.]} &= \text{argmin } \sum_{i=1}^{N} (y_i - \theta^\top x_i)^2 + \frac{\sigma_\epsilon^2}{\sigma_\theta^2}||\theta|^2.
\end{aligned}$$

Podemos ver que eligiendo un prior uniforme o de normal de varianza muy amplia, el MAP es equivalente al EMV. ¿qué significa esto? ¿qué comportamiento differente de EMV promueve el MAP en este caso?

8.3.2. Posterior predictiva

En la inferencia bayesiana las predicciones ocupan un rol relevante, pues luego de realizar inferencia sobre un modelo estadístico, en general estamos interesados estudiar cómo serán los siguientes datos genearados por el modelo. Para esto definiremos la predicción bayesiana de la forma

Definición 8.13 (Posterior predictiva)**.**

Para un conjunto de datos $\mathcal{D}$ y un parámetro θ, la densidad posterior predictiva está dada por

$$p(x|\mathcal{D}) = \int_\Omega p(x|\theta)p(\theta|\mathcal{D})\mathrm{d}\theta = \mathbb{E}\left(p(x|\theta)|\mathcal{D}\right), \tag{8.31}$$

es decir, el valor esperado del modelo estadístico con respecto a la ley posterior del parámetro (modelo).

Podemos ahora considerar la posterior predictiva como nuestro modelo *aprendido* y generar datos de él, donde nos encontramos frente al mismo dilema de un estimador puntual como en el caso anterior: es posible considerar muestras aleatorias, la media, la mediana o algún

intervalo.

Observación 8.12.

La posterior predictiva es distinta (en general) a la predicción *plug-in*, en donde consideramos en modelo estadístico $p_{\hat{\theta}}$ en base a un estimador (puntual) cualquiera $\hat{\theta}$. Desde esa perspectiva, la posterior predictiva equivale a considerar estimadores y modelos puntuales pero integrar todos ellos con respecto a la ley posterior.

8.4. El prior de Jeffreys

Consideremos $X \sim p(x|\theta)$, $\theta \in [a,b]$, en donde elegimos el prior *no informativo* uniforme dado por

$$p(\theta) = \text{Uniforme}\,(a,b) = \frac{1}{b-a}.$$

Consideremos ahora un modelo *reparametrizado* $\eta = e^{\theta} \in [c,d]$, donde el modelo es expresado como $X \sim q(x|\eta) = p(x|\theta)$. El prior uniforme para el nuevo parámetro es

$$p(\eta) = \text{Uniforme}\,(c,d) = \frac{1}{d-c}. \tag{8.32}$$

Observemos que la elección uniforme del parámetro θ en el intervalo $[a,b]$ es equivalente a elegir η según

$$\tilde{p}(\eta) = p(\theta)\left|\frac{d\theta}{d\eta}\right| = \frac{1}{b-a}\left|\frac{d\log\eta}{d\eta}\right| = \frac{1}{\eta(b-a)}, \tag{8.33}$$

es decir, la distribución sobre η inducida por $p(\theta)$. Esta distribución por supuesto no es equivalente a elegir η uniformemente en el intervalo $[c,d]$.

Observación 8.13.

¿Es un prior uniforme realmente no informativo si luego de elegir otra parametrización este ya no es uniforme? ¿Es posible construir un prior no informativo?

Una forma de construir un prior que es invariante ante reparametrizaciones es mediante la metodología propuesta por Harold Jeffreys (1946), el que sugiere elegir un prior proporcional a la raíz cuadrada del determinante de la información de Fisher, es decir,

$$p(\theta) \propto (I(\theta))^{1/2}, \tag{8.34}$$

donde recordemos que la información de Fisher está dada por

$$I(\theta) = -\mathbb{E}_\theta\left(\frac{\partial^2}{\partial\theta^2}\log p(X|\theta)\right) = \mathbb{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log p(X|\theta)\right)^2\right). \tag{8.35}$$

Además, si $X_1,\ldots,X_n$ son iid, entonces $I(\theta) = nI_1(\theta)$ y el prior de Jeffreys puede ser expresado como

$$p(\theta) \propto I_1(\theta)^{1/2}. \tag{8.36}$$

Observemos que si $\int_\Omega \sqrt{I(\theta)}\mathrm{d}\theta$ es finito, entonces la constante de proporcionalidad es precisamente esta cantidad. Sin embargo, si esta cantidad es infinita el prior de Jeffreys aún es un prior válido pero impropio, siempre y cuando las posteriores respectivas sí sean propias.

Veamos ahora que el prior de Jeffreys es invariante bajo reparametrizaciones. Consideremos los modelos relacionados mediante reparametrización dados por

$$X \sim p(x|\theta),\ \theta \in \Omega \quad \& \quad X \sim q(x|\eta),\ \eta \in \Gamma, \tag{8.37}$$

donde $\eta = h(\theta)$. Las informaciones de Fisher para ambos modelos, denotadas respectivamente $I_p(\theta)$ e $I_q(\theta)$, están relacionadas mediante

$$\begin{aligned} I_p(\theta) &= \int_{\mathcal{X}} \left(\frac{\partial}{\partial \theta} \log p(x|\theta) \right)^2 p(x|\theta) \mathrm{d}x \\ &= \int_{\mathcal{X}} \left(\frac{\partial}{\partial \theta} \log q(x|h(\theta)) \right)^2 q(x|h(\theta)) \mathrm{d}x \\ &= \int_{\mathcal{X}} \left(\frac{\partial}{\partial \eta} \log q(x|\eta) h'(\theta) \right)^2 q(x|\eta) \mathrm{d}x \\ &= \left(h'(\theta)\right)^2 I_q(\eta). \end{aligned} \tag{8.38}$$

Observemos ahora que el prior en θ, $p(\theta)$, inducido por el prior de Jeffreys en η, $p_J(\eta)$, es efectivamente el prior de Jeffreys en θ, $p_J(\theta)$. En efecto, debido al cambio de variable tenemos

$$p(\theta) = p_J(\eta) \left| \frac{d\eta}{d\theta} \right| = \sqrt{I_q(\eta)} \left| h'(\theta) \right| = \sqrt{I_p(\theta)} = p_J(\theta). \tag{8.39}$$

Como ya mencionamos, la construcción del Prior de Jeffreys surge con la idea de usar un prior que sea invariante bajo transformaciones monótonas y que sea no informativo. ¿Pero cómo se logra esto últmo? Resulta ser que el prior de Jeffreys es el prior uniforme sobre el espacio de parámetros Θ, pero no con la métrica euclidiana. Intuitivamente, la topología que se debe considerar es aquella que calcula la distancia entre dos parámetros θ_1 y θ_2 como la divergencia de Kulback-Liebler entre sus distribuciones asociadas $f(x|\theta_1)$ y $f(x|\theta_2)$.

8.5. Intervalos de Credibilidad

En el concepto (frecuentista) de intervalo de confianza, la *aleatoriedad* ocurre antes que veamos los datos, pues recordemos que los supuestos de este paradigmas son que el parámetro es fijo y desconocido, y la generación de datos es aleatoria. Con esto, el hecho de encontrar un intervalo de confianza del, e.g., 95 % quiere decir que existe un 95 % de probabilidad de que un intervalo de confianza observado en el futuro (en realidad lo observado son los datos y el intervalo es función de éstos) contengan al parámetro. Esto es contraintuitivo, pues nos gustaría que la aleatoriedad ocurriera *después de observar los datos*, es decir, dado los datos x cual es la probabilidad de que el parámetro está dentro de un intervalo dado?

El paradigma bayesiano permite enunciar lo anterior y propone una noción de intervalos de confianza más natural que el enfoque frecuentista, pues para $C_x \subset \Omega$, la expresión $\mathbb{P}(\theta \in)$ tiene un significado, incluso condicional a x. En este caso, le llamamos intervalos de credibilidad a los intervalos de confianza bayesianos. Para diferenciar estos intervalos con el caso frecuentista, veamos la siguiente definición.

Definición 8.14.

Sea π el prior del parámetro θ, un conjunto C_x se dice α-creíble si la posterior corres-

pondiente al prior π cumple con

$$\mathbb{P}(\theta \in C_x|x) = \int_{C_x} p(\theta|x) = 1 - \alpha.$$

Notemos que al igual que para el caso frecuentista, esta región no está únicamente determinada, pues puede ser centrada, no convexa, concentrada el en origen, etc. Para esto, podemos considerar los siguientes criterios:

- Elegir el intervalo más pequeño, es decir, el que minimiza el volumen de las regiones α-creíbles. Esto motiva la siguiente definición.

  Definición 8.15.
  Una región de Alta Densidad Posterior (HPD por su sigla en inglés) denotada mediante
  $$C_\alpha = \{\theta : p(\theta|x) \geq k_\alpha\},$$
  donde k_α es la cota más grande tal que:
  $$\mathbb{P}(\theta \in C_\alpha|x) = 1 - \alpha.$$

  Observe que para las distribución unimodales, la moda (el máximo a posteriori) estará incluido en este intervalo.

- Elegir un intervalo tal que la probabilidad de estar a la izquierda es igual a la probabilidad de esta a la derecha. Este intervalo incluye a la mediana y es llamado **intervalo de igual colas**

- Asumir que la media existe y elegir el intervalo centrado en ésta.

Ejemplo 8.9.
Considere el prior $\theta \sim \pi(\theta) = \mathcal{N}(0, \tau^2)$ y una verosimilitud tal que la posterior de θ es una normal $\mathcal{N}(\mu(x), \omega^{-2})$, con $\omega^{-2} = \tau^{-2} + \sigma^{-2}$ y $\mu(x) = \dfrac{\tau^2 x}{\tau^2 + \sigma^2}$. Luego:

$$C_\alpha = [\mu(x) - k_\alpha \omega^{-1}, \mu(x) + k_\alpha \omega^{-1}],$$

con k_α el $\alpha/2$-intil de $\mathcal{N}(0, 1)$. En particular, si $\tau \to \infty$, $\pi(\theta)$ converge a la medida de Lebesgue en $\mathbb{R}$ y:
$$C_\alpha = [x - k_\alpha \sigma, x + k_\alpha \sigma],$$
que corresponde al intervalo de confianza clásico para una normal.

Observación 8.14.
Si los intervalos de confianza y credibilidad para a la media de la gaussiana son el mismo, ¿cuál es la diferencia?

Ejemplo 8.10.
Encuentre el 85 %-intervalo creíble de λ en $x_1, \ldots, x_n \sim \exp(\lambda)$ cuando el prior es

uniforme. Tenemos

$$p(\lambda|x_1,\ldots,x_n) \propto \lambda^n e^{-\lambda\sum_i x_i}, \tag{8.40}$$

con lo que concluimos que

$$p(\lambda|x_1,\ldots,x_n) = \Gamma(n+1,\sum_i x_i). \tag{8.41}$$

Debemos ahora encontrar a, b tal que

$$\int_a^b \frac{(\sum_i x_i)^{n+1}}{\Gamma(n+1)}\lambda^n e^{-\lambda\sum_i x_i}\mathrm{d}\lambda = 0{,}85 \tag{8.42}$$

donde tenemos las 3 opciones mencionadas arriba.

Ejercicio 8.1.

Encuentre el intervalo creíble para θ en el Ejemplo 6.7

8.6. Test de Hipótesis Bayesiano

Hasta ahora sólo hemos visto los distintos test de hipótesis desde una perspectiva frecuentista. En todos estos test, había una relación asimétrica entre dos hipótesis: la hipótesis nula H_0 y la hipótesis alternativa H_1. Un proceso de desición se lleva acabo, y luego, en base a los datos observados, la hipótesis nula se va a rechazar a favor de H_1, o se aceptará.
En el test de hipótesis Bayesiano, puede haber más de dos hipótesis en consideración, y no deben tener, necesariamente, una relación asimétrica. Para simplificar el análisis, consideremos dos hipótesis: H_1 y H_2.
Sabemos que en algún momento tendremos datos X, sin embargo, aún no los tenemos. Nos interesa calcular las distribuciones posteriores $P(H_1|X)$ y $P(H_2|X)$. Usando Bayes:

$$P(H_1|X) = \frac{P(X|H_1)P(H_1)}{P(X)} \; ; \; P(H_2|X) = 1 - P(H_1|X).$$

Por probabilidades totales:

$$P(X) = P(X|H_1)P(H_1) + P(X|H_2)P(H_2).$$

Ejemplo 8.11.

Consideremos un lanzamiento de una moneda, y las hipótesis: H_1 ="La moneda está cargada"($\theta = \frac{1}{2}$) y H_2="La moneda no está cargada". Entonces, si θ es la probabilidad de que salga cara (C):

$$P(\theta|H_1) = 1_{\theta=0,5}$$

Esto es una distribución a priori. Por otra parte, la hipótesis 2 es la que indica que la moneda está cargada. Consideremos que si la moneda está cargada, θ puede valer 1/3

o 2/3 de forma igualmente probable:

$$P(\theta|H_2) = 0{,}5 * 1_{\theta=\frac{1}{3}} + 0{,}5 * 1_{\theta=\frac{2}{3}}$$

Por último, necesitamos las probabilidades $P(H_1)$ y $P(H_2)$. Consideremos (como se suele hacer) que $P(H_1) = P(H_2) = 0{,}5$. Supongamos que al lanzar la moneda obtenemos la secuencia: CCSCSC. Entonces:

$$P(X|H_1) = \binom{6}{4}(\frac{1}{2})^4(\frac{1}{2})^2 = \binom{6}{4}0{,}0156$$

$$P(X|H_2) = P(X|\theta = 1/3)P(\theta = 1/3) + P(X|\theta = 2/3)P(\theta = 2/3) = \binom{6}{4}0{,}0137$$

Con los dos cálculos anteriores:

$$P(X) = \binom{6}{4}0{,}0156P(H_1) + \binom{6}{4}0{,}0137P(H_2) = \binom{6}{4}0{,}01465$$

Entonces:

$$P(H_1|X) = \frac{\binom{6}{4}0{,}0156P(H_1)}{\binom{6}{4}0{,}01465} = 0{,}53$$

Luego pasamos de $P(H_1) = 0{,}5$ a $P(H_1|X) = 0{,}53$, es decir, actualizamos nuestras creencias y ahora pensamos que es más probable que la moneda no esté cargada.

En el test bayesiano, el ratio entre las verosimilitudes se llama **factor de bayes**.

8.7. Evaluación de Modelos

Criterio de información de Akaike (AIC)

Sea $\mathcal{D} = (x_i)_{i=1}^N$ un conjunto de observaciones generadas por una distribución desconocida perteneciente a una familia paramétrica cuyos parámetros están en $\Theta \subset \mathbb{R}^d$. Bajo este modelo, se puede utilizar el estimador de máxima verosimilitud:

$$\hat{\theta} = \arg\max_{\theta\in\Theta} L(\theta|\mathcal{D}) = \arg\max_{\theta\in\Theta} l(\theta|\mathcal{D}) \tag{8.43}$$

Una forma de evaluar el desempeño real de este estimador es mediante el *riesgo de predicción*, el cual se ve reflejado en la log-verosimilitud de $\hat{\theta}$ sobre todas las posibles observaciones: $\mathbb{E}\left(()\, l(\hat{\theta}|x)\right)$. Dado que solo se cuenta con una cantidad finita de muestras, solo es posible obtener un riesgo empírico. El criterio de información de Akaike (AIC) busca ajustar este riesgo para obtener un estimador asintóticamente insesgado del riesgo real. Para esto, se tienen las siguientes definiciones para el estimador de máxima verosimilitud $\hat{\theta}$:

- **Riesgo empírico:** $R_{\mathcal{D}}(\hat{\theta}) = -\hat{l}$, donde $\hat{l} = l(\hat{\theta}|\mathcal{D})$ es la log-verosimilitud del EMV empírico.

- **Riesgo real:** $R(\hat{\theta}) = -\mathbb{E}\left(()\, N \cdot l_0(\hat{\theta})\right)$, donde $l_0(\theta) = \mathbb{E}\left(()\, l(\theta|x)\right)$ corresponde a la log-verosimilitud de θ sobre todo el espacio muestral. Notar que se multiplica por N ya que en el riesgo empírico no se normalizó por N.

Para poder obtener el AIC se analizará el sesgo asintótico del riesgo empírico con respecto al riesgo real. Para esto, se utilizarán aproximaciones sobre ambos riesgos, asumiendo que a medida que N crece, el EMV empírico tiende al EMV global (por LGN), por lo que el residuo de Taylor tenderá a 0.

Sea $\theta_0 = \arg\max_{\theta\in\Theta} l_0(\theta)$ el EMV sobre todo el espacio muestral. Utilizando una aproximación de Taylor de segundo orden sobre l_0 alrededor de θ_0:

$$l_0(\hat{\theta}) \approx l_0(\theta_0) + (\hat{\theta}-\theta_0)^\top \nabla l_0(\theta_0) + \frac{1}{2}(\hat{\theta}-\theta_0)^\top H_{l_0}(\theta_0)(\hat{\theta}-\theta_0) \tag{8.44}$$

$$= l_0(\theta_0) + \frac{1}{2}(\hat{\theta}-\theta_0)^\top H_{l_0}(\theta_0)(\hat{\theta}-\theta_0) \tag{8.45}$$

Donde se usó que $\nabla l_0(\theta_0) = 0$ ya que θ_0 es un punto crítico de l_0. De esta forma, se tiene una aproximación de segundo orden para el riesgo real:

$$R(\hat{\theta}) \approx -N \cdot l_0(\theta_0) - \frac{N}{2}\mathbb{E}\left((\hat{\theta}-\theta_0)^\top H_{l_0}(\theta_0)(\hat{\theta}-\theta_0)\right)$$

Por otra parte, realizando una expansión de Taylor de segundo orden sobre $\hat{l}$ alrededor de θ_0:

$$\hat{l} = \sum_{i=1}^{N} l(\hat{\theta}|x_i) \approx \sum_{i=1}^{N} l(\theta_0|x_i) + (\hat{\theta}-\theta_0)^\top \sum_{i=1}^{N} \nabla l(\theta_0|x_i) + \frac{1}{2}(\hat{\theta}-\theta_0)^\top \sum_{i=1}^{N} H_l(\theta_0|x_i)(\hat{\theta}-\theta_0) \tag{8.46}$$

Usando el hecho de que $\hat{\theta}$ es punto crítico de $l(\cdot|\mathcal{D})$:

$$\sum_{i=1}^{N} \nabla l(\theta_0|x_i) = \sum_{i=1}^{N} \nabla\left(l(\theta_0|x_i) - l(\hat{\theta}|x_i)\right) \approx \left(\sum_{i=1}^{N} \nabla\nabla l(\theta_0|x_i)\right)(\theta_0-\hat{\theta}) \approx N\mathbb{E}(H_l(\theta_0|x_i))(\theta_0-\hat{\theta}). \tag{8.47}$$

Luego, sustituyendo en $\hat{l}$ y notando que $\sum_{i=1}^{N} H_l(\theta_0|x_i) \approx N\mathbb{E}(H_l(\theta_0|x))$:

$$\hat{l} \approx \sum_{i=1}^{N} l(\theta_0|x_i) + N(\hat{\theta}-\theta_0)^\top \mathbb{E}(H_l(\theta_0|x))(\theta_0-\hat{\theta}) + \frac{N}{2}(\hat{\theta}-\theta_0)^\top \mathbb{E}(H_l(\theta_0|x))(\hat{\theta}-\theta_0) \quad (8.48)$$

$$\implies \mathbb{E}(R_{\mathcal{D}}(\hat{\theta})) = -Nl_0(\theta_0) + \frac{N}{2}\mathbb{E}\left((\hat{\theta}-\theta_0)^\top H_{l_0}(\theta_0)(\hat{\theta}-\theta_0)\right). \quad (8.49)$$

De este modo, el sesgo del riesgo empírico como estimador del riesgo real es:

$$\mathbb{E}(R_{\mathcal{D}}(\hat{\theta})) - R(\hat{\theta}) = -N\mathbb{E}\left((\hat{\theta}-\theta_0)^\top H_{l_0}(\theta_0)(\hat{\theta}-\theta_0)\right).$$

Por otra parte, dado que $\sqrt{N}\left(\hat{\theta}-\theta_0\right) \approx \mathcal{N}\left(0, H_{l_0}(\theta_0)^{-1}\right)$, la forma cuadrática anterior puede ser aproximada por una distribución de Pearson: $N(\hat{\theta}-\theta_0)^\top H_{l_0}(\theta_0)(\hat{\theta}-\theta_0) \approx \mathcal{X}_d^2$, donde $\mathbb{E}(\mathcal{X}_d^2) = d$. De este modo,

$$\mathbb{E}(R_{\mathcal{D}}(\hat{\theta})) - R(\hat{\theta}) \approx -d. \quad (8.50)$$

Por lo que corrigiendo $R_{\mathcal{D}}(\hat{\theta})$ se obtiene un estimador asintóticamente insesgado del riesgo real: $R_{\mathcal{D}}(\hat{\theta}) + d$. De esta forma, se tiene la siguiente definición:

Definición 8.16 (AIC)**.**

Sea M un modelo estadístico d-paramétrico y $\mathcal{D} = (x_i)_{i=1}^N$ un conjunto de observaciones. El AIC del modelo (aproximado por $\mathcal{D}$) se define como

$$AIC(M, \mathcal{D}) := 2d - 2\log\left(\hat{L}(\mathcal{D})\right), \quad (8.51)$$

donde $\hat{L}(\mathcal{D})$ corresponde a la verosimilitud del EMV asociado a $\mathcal{D}$, es decir:

$$\hat{L}(\mathcal{D}) = \prod_{i=1}^{N} p(x_i|\hat{\theta}), \text{ para } \hat{\theta} = \arg\max_{\theta\in\Theta} L(\theta|\mathcal{D}). \quad (8.52)$$

Observación 8.15.

El AIC corresponde al estimador asintóticamente insesgado del riesgo real multiplicado por 2. Esta ponderación es realizada por motivos históricos (Model selection and multimodel inference, Burnham & Anderson).

De acuerdo a la derivación anterior, el AIC es una medida relativa de la pérdida de información de un modelo de acuerdo a un conjunto de entrenamiento $\mathcal{D}$. De esta forma, para un conjunto de posibles modelos, se debe elegir el modelo que presente el menor valor AIC ya que será el que minimice el riesgo de predicción.

Como se puede ver en la definición, el criterio de Akaike no se basa únicamente en la verosimilitud del modelo sino que agrega una penalización de acuerdo a la cantidad de parámetros, evitando elegir un modelo sobreajustado a los datos.

Observación 8.16.

Una de las hipótesis de AIC es que el espacio muestral es infinito ya que se asume que el error de Taylor es despreciable. Para una cantidad finita de datos (N), se puede realizar una corrección del estimador dada por:

$$AICc(M, \mathcal{D}) := AIC(M, \mathcal{D}) + \frac{2d(d+1)}{N-d-1}. \tag{8.53}$$

Es importante notar que cuando $N \to \infty$ se recupera el AIC original.

Criterio de información bayesiano (BIC)

Otro enfoque para la selección de modelos corresponde al criterio de información bayesiano (o criterio de Schwarz). Dada una familia de modelos $\mathcal{M}$, se define un prior $p(m)$ para cada modelo $m \in \mathcal{M}$. Además, se define un prior $p(\theta|m)$ sobre los parámetros de cada modelo. El criterio de información bayesiano (BIC) elige al mejor modelo de acuerdo a la posterior $p(m|\mathcal{D})$, la cual viene dada de acuerdo al teorema de Bayes:

$$p(m|\mathcal{D}) = \frac{p(\mathcal{D}|m)p(m)}{p(\mathcal{D})} \propto p(\mathcal{D}|m)p(m). \tag{8.54}$$

De forma similar al criterio de Akaike, se puede calcular la verosimilitud del modelo $p(\mathcal{D}|m)$ mediante aproximaciones de Taylor, probando que es independiente del prior. La derivación de $p(\mathcal{D}|m)$ lleva a la siguiente definición:

Definición 8.17 (BIC)**.**

Sea M un modelo estadístico d-paramétrico y $\mathcal{D} = (x_i)_{i=1}^N$ un conjunto de observaciones. El BIC del modelo (aproximado por $\mathcal{D}$) se define como

$$BIC(M, \mathcal{D}) := d \cdot \log(N) - 2\log\left(\hat{L}(\mathcal{D})\right) \tag{8.55}$$

Donde nuevamente $\hat{L}(\mathcal{D})$ corresponde a la verosimilitud del EMV asociado a $\mathcal{D}$.

En este caso, se vuelve a elegir el modelo que presente el menor BIC. Se observa que, al igual que AIC, BIC contiene una penalización sobre el número de parámetros por lo que también evita el sobreajuste a los datos.

Observación 8.17 (Stone (1977) - Shao (1997))**.**

Para una familia de modelos, minimizar el AIC es asintóticamente equivalente a realizar LOOCV. Por otra parte, minimizar el BIC es asintóticamente equivalente a realizar leave p out cross validation para

$$p = \left\lfloor N \left(1 - \frac{1}{\log(N) - 1}\right)\right\rfloor \tag{8.56}$$

AIC y BIC para la regresión lineal

Al igual que en máxima verosimilitud, se puede considerar un modelo generativo para la regresión lineal de la forma $y = c^\top x + \epsilon$, donde $\epsilon \sim \mathcal{N}(0, \sigma^2)$ y por lo tanto, $y|x \sim \mathcal{N}(y; c^\top x, \sigma^2)$. Sean $\hat{c}$ y $\hat{\sigma}^2$ los EMV del modelo (calculados en el capítulo de regresión), entonces la log-verosimilitud máxima viene dada por:

$$\hat{l}(\mathcal{D}) = \frac{-N}{2}\log(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{N}(y_i - \hat{c}^\top x_i) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}N\hat{\sigma}^2 \quad (8.57)$$

$$= C(N) - \frac{N}{2}\log(\hat{\sigma}^2) = C(N) - \frac{N}{2}\log\left(\frac{1}{N}\text{RSS}(\mathcal{D})\right) \quad (8.58)$$

Donde $C(N) = -\frac{N}{2}\log(2\pi) - N$ y $\text{RSS}(\mathcal{D})$ corresponde a la suma de cuadrados residuales: $\text{RSS}(\mathcal{D}) := \sum_{i=1}^{N}\left(y_i - c^\top x_i\right)^2$. Dado que $C(N)$ es una constante independiente del modelo, puede ser omitida en la comparación de modelos, por lo tanto:

- $AIC = 2d - N\log(\frac{1}{N}\text{RSS}(\mathcal{D}))$
- $BIC = d\log(N) - N\log(\frac{1}{N}\text{RSS}(\mathcal{D}))$

Si bien existen otros métodos de selección de modelo (DIC, WAIC, entre otros), estos tienen una formulación más compleja que se escapa del alcance de este curso ya que se requieren herramientas adicionales como MCMC para el cálculo de distribuciones posteriores.

8.8. Ejercicios

1. Sea $X = (X_1, ..., X_n)$ una MAS con una distribución normal dada por $X_i \sim \mathcal{N}(\mu, \sigma^2)$ en donde μ es desconocido y σ^2 es conocido. Se supone una densidad a priori para μ dada por:

   $$f(\mu) \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

   Donde μ_0 y σ_0^2 son conocidos.

   (i) Calcule la densidad a posteriori del modelo. Verifique que se cumple el fenómeno de conjugación.

   (ii) Calcule el máximo a posteriori del modelo.

2. Sea $X = (X_1, ..., X_n)$ una MAS iid. Se tienen las distribuciones:

   $$\mathbb{P}(x_i|\mu) = Bernoulli(x_i|\mu) = \mu^{x_i}(1-\mu)^{1-x_i}$$

   $$\mathbb{P}(\mu) = Beta(\mu|\alpha, \beta) = \frac{1}{B(\alpha, \beta)}\mu^{\alpha-1}(1-\mu)^{\beta-1}$$

Encuentre el máximo a posteriori de μ.

3. Sea $X = (X_1, ..., X_n)$ una MAS iid. Se tienen las distribuciones:

$$\mathbb{P}(x_i|\mu) \sim \mathcal{N}(\mu, \sigma^2)$$

$$\mathbb{P}(\sigma^2|\alpha, \beta) = \text{Inverse-Gamma}(\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\sigma^{-\alpha-1}\exp\left(\frac{-\beta}{\sigma^2}\right)$$

Encuentre el máximo a posteriori de σ^2.

4. Sea $X = (X_1, ..., X_n)$ una MAS iid. Se tienen las distribuciones:

$$\mathbb{P}(x_i|\mu) \sim \mathcal{N}(\mu, \sigma_0^2)$$

$$\mathbb{P}(x|\nu, \tau) = \text{Scaled Inverse Chi-Squared}(\nu, \tau) = \frac{(\tau^2\nu/2)^{\nu/2}\exp\left(\frac{-\nu\tau^2}{2x}\right)}{\Gamma(\nu/2)x^{1+\nu/2}}$$

Planteé un modelo bayesiano para σ^2 considerando μ conocido.