Estadística:
Teoría y Aplicaciones

Felipe Tobar

CENTRO DE MODELAMIENTO MATEMÁTICO

VERSIÓN: 15 DE ENERO DE 2021
Email address: ftobar@dim.uchile.cl

Índice general

Prefacio

Estas notas de clase contienen el material en base al cual se ha dictado el curso de Estadística (MA3402) en el Departamento de Ingeniería Matemática de la Universidad de Chile durante los años 2019 y 2020. 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 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. También agradezco al Centro de Modelamiento Matemático, por ser mi segundo hogar mientras he dictado el curso. Finalmente, he sido muy afortunado de contar con el profesionalismo y entrega de un grupo espectacular de auxiliares: sin Arie Wortsman, Francisco Vásquez y Bruno Moreno ni este documento y ni el curso serían lo que hoy son. ¡Gracias!

Felipe Tobar,
Santiago,
octubre 2020.

Capítulo 1

Introducción

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 diferencias claras pueden ser identificadas en sus objetivos pueden, 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. 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 hecho 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.

2. 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 1.1 (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 1.2 (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 1.4 (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. 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

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 es una función de (las realizaciones de) una variable aleatoria, definida desde el espacio muestral. Es decir, cualquier función *medible* de los datos.

Definición 2.1 (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 2.1: Es muy relevante diferenciar el valor de 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:

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

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 2.2 (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 2.1 (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 2.2 (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 2.3 (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 2.1 (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:

$$p_\theta(x) = g_\theta(T(x))h(x). \tag{1}$$

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 2.1, revisemos un par de ejemplos.

Ejemplo 2.3 (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 2.1. En efecto, consideremos $x = (x_1, \ldots, x_n) \sim$ Bernoulli(θ) y el estadístico $T(x) = \sum x_i$, entonces,

$$\mathbb{P}(X = x) = \underbrace{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}}_{g_\theta(T(x))} \cdot \underbrace{1}_{h(x)} \tag{2}$$

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 2.4 (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 2.1 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 (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

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

Incluyendo estas últimas dos expresiones en la ec. (1), tenemos

$$p_\theta(X = x|T(X) = t) = \frac{\mathbb{1}_{T(x)=t}\cancel{g_\theta(T(x))}h(x)}{\sum_{x';T(x')=t}\cancel{g_\theta(T(x'))}h(x')} = \frac{\mathbb{1}_{T(x)=t}h(x)}{\sum_{x';T(x')=t}h(x')} \tag{3}$$

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

Finalmente, como el lado derecho de la ecuación (3) no depende de θ, se concluye la demostración. ■

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

Teorema 2.2: Un estadístico es suficiente si y sólo si la partición que induce es suficiente.

DEMOSTRACIÓN. Ejercicio. ■

Ejemplo 2.5: Sean $X_1, X_2, X_3 \sim Bernoulli(\theta)$. Sea $T = \sum_i X_i$. En el Cuadro 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 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 2.2: Dos estadísticos pueden generar la misma partición. Por ejemplo T del ejemplo anterior, y $U = 3T$.

Observación 2.3: 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 2.6: Sean X_1, X_2 y $X_3 \sim Bernoulli(\theta)$. Entonces $T = X_1$ no es suficiente. Veamos su partición en el Cuadro 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 2. *Outcomes* y estadísticos (Bernoulli, $T = X_1$)

2. Suficiencia minimal y completitud

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 2.1. 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 2.5 (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 2.7: 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 2.3 (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 $x, y \in \mathcal{X}$:

$$\frac{p_\theta(x)}{p_\theta(y)} \text{ no depende de } \theta \Leftrightarrow T(x) = T(y) \tag{4}$$

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

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

Ejemplo 2.8: 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

$$p_\theta(x) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \frac{e^{-n\theta}\theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}. \tag{5}$$

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

$$\frac{p_\theta(x)}{p_\theta(y)} = \frac{\theta^{\sum_{i=1}^n x_i - \sum_{i=1}^n y_i}}{\prod_{i=1}^n x_i! / \prod_{i=1}^n y_i!}, \tag{6}$$

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 2.3.

DEMOSTRACIÓN DE TEOREMA 2.3. Recordemos que queremos demostrar que la ec. (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

$$p_\theta(x) = \underbrace{p_\theta(x)/p_\theta(x_T)}_{h(x) \text{ indep. } \theta} \underbrace{p_\theta(x_T)}_{q_\theta(T(x))} \tag{7}$$

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

$$\frac{p_\theta(x)}{p_\theta(y)} = \frac{g'_\theta(T'(x))h'(x)}{g'_\theta(T'(y))h'(y)} = \frac{h'(x)}{h'(y)}, \quad \text{pues } T'(x) = T'(y) \tag{8}$$

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'.

■

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).

$$A(\eta) = \log \int_{\mathcal{X}} \exp\left(\sum_{i=1}^{s} \eta_i T_i(x)\right) h(x) \mathrm{d}x, \tag{9}$$

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

Definamos la siguiente función de densidad de probabilidad parametrizada por $\eta \in \{\eta | A(\eta) < \infty\}$

$$p_\eta(x) = \exp\left(\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\right) h(x) \tag{10}$$

donde el hecho que $p_\eta(x)$ integra uno puede claramente verificarse reemplazando la ecuación (9) en (10), con lo cual se puede ver que A definido en (9) es precisamente el logaritmo de la constante de normalización de la densidad definida en la ec. (10).

Observación 2.4: El estadístico T es un estadístico suficiente para ν en la familia exponencial. En efecto, notemos que

$$p_\eta(x) = \underbrace{\exp\left(\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\right)}_{g_\theta(T(x))} \underbrace{h(x)}_{h(x)} \tag{11}$$

consecuentemente, por el Teorema 2.1 (Neyman-Fisher), tenemos que T es un estadístico suficiente para ν.

Muchas de las distribuciones que usalmente consideramos pertenecen a la familia exponencial, por ejemplo, la distribución normal, exponencial, gamma, chicuadrado, 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 2.9 (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 2.10 (El modelo normal pertenece a la familia exponencial)**:** La distribución normal $\mathcal{N}(\mu,\sigma^2)$ tiene densidad:

$$p_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-1}{2\sigma^2}(x-\mu)^2} \tag{12}$$

$$= \frac{1}{\sqrt{2\pi}}\exp\left(\frac{\mu}{\sigma^2}x-\frac{-1}{2\sigma^2}x^2-\left(\frac{\mu^2}{2\sigma^2}+\log\sigma\right)\right) \tag{13}$$

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 2.11: Ejercicio Demuestre que las distribuciones *Poisson* y *Bernoulli* pertenecen a la familia exponencial.

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

$$p_1(x_1) = h_1(x_1) \exp\left(\theta_1 T_1(x_1) - A_1(\theta_1)\right) \tag{14}$$

$$p_2(x_2) = h_2(x_2) \exp\left(\theta_2 T_2(x_2) - A_2(\theta_2)\right) \tag{15}$$

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

$$\begin{aligned} p(x) &= p_1(x_1) p_2(x_2) \\ &= \underbrace{h_1(x_1) h_2(x_2)}_{h(x)} \exp\left(\underbrace{[\theta_1, \theta_2]}_{\theta} \underbrace{\begin{bmatrix} T_1(x_1) \\ T_2(x_2) \end{bmatrix}}_{T(x)} - \underbrace{(A_1(\theta_1) + A_2(\theta_2))}_{A(\theta)} \right), \end{aligned} \tag{16}$$

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

$$Q(\theta) = \exp\left(A(\theta)\right) = \int_{\mathcal{X}} \exp\left(\theta T(x)\right) h(x) \mathrm{d}x, \tag{17}$$

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

$$\begin{aligned} \frac{dA(\theta)}{d\theta} &= Q^{-1}(\theta) \frac{dQ(\theta)}{d\theta} \\ &= \frac{\int_{\mathcal{X}} T(x) \exp\left(\theta T(x)\right) h(x) \mathrm{d}x}{\int_{\mathcal{X}} \exp\left(\theta T(x)\right) h(x) \mathrm{d}x} \\ &= \frac{\int_{\mathcal{X}} T(x) \exp\left(\theta T(x) - A(\theta)\right) h(x) \mathrm{d}x}{\int_{\mathcal{X}} \exp\left(\theta T(x) - A(\theta)\right) h(x) \mathrm{d}x} \qquad \cdot A(\theta)/A(\theta) \\ &= \mathbb{E}\left(T(x)\right) \end{aligned} \tag{18}$$

Ejemplo 2.12: Verificar para derivadas de orden superior (ejercicio)

Observación 2.5: Consideremos el mapa

$$\theta \mapsto \mu = \frac{dA(\theta)}{d\theta} = \int_{\mathcal{X}} T(x) \exp\left(\theta T(x) - A(\theta)\right) h(x) \mathrm{d}x = \mathbb{E}\left(T(x)\right) \tag{19}$$

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?

Capítulo 3

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 3.1: 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 3.1: 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 3.2: 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

$$g(\theta) = [\mu - 2\sigma, \mu + 2\sigma]. \tag{20}$$

Ejemplo 3.1 (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

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

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 3.2 (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 3.2 (Estimador insesgado de la media Gaussiana)**:** El estimador de $g(\theta) = \mu$ descrito en el Ejemplo 3.1 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 3.3: 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

$$\mathbb{V}_\theta(\bar{x}) = \mathbb{V}_\theta\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \underbrace{=}_{\text{i.i.d.}} \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}_\theta(x_i) = \frac{\sigma^2}{n} \tag{21}$$

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

Consideremos ahora el siguiente estimador para la varianza:

$$S_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{22}$$

y notemos que la esperanza de dicho estimador es

$$\begin{aligned}
\mathbb{E}_\theta(S_2) &= \mathbb{E}_\theta\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu + \mu - \bar{x})^2\right) \\
&= \mathbb{E}_\theta\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 + 2\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(\mu-\bar{x}) + \frac{1}{n}\sum_{i=1}^{n}(\mu-\bar{x})^2\right) \\
&= \mathbb{E}_\theta\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 - 2(\mu-\bar{x})^2 + (\mu-\bar{x})^2\right) \\
&= \mathbb{E}_\theta\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 - (\mu-\bar{x})^2\right) \\
&= \mathbb{V}_\theta(x_i) - \mathbb{V}_\theta(\bar{x}) \quad \text{ver ecuación (21)} \\
&= \sigma^2 + \sigma^2/n = \left(\frac{n+1}{n}\right)\sigma^2 \qquad (23)
\end{aligned}$$

Esto quiere decir que el sesgo del estimador en la ecuación (22) 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 (22) por $n/(n+1)$, con lo que el estimador corregido denotado por

$$S_2' = \frac{n}{n+1}S_2 = \frac{1}{n+1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{24}$$

cumple con

$$\mathbb{E}_\theta(S_2') = \left(\frac{n}{n+1}\right)\mathbb{E}_\theta(S_2) \underbrace{=}_{\text{ec.(23)}} \left(\frac{n}{n+1}\right)\left(\frac{n+1}{n}\right)\sigma^2 = \sigma^2 \tag{25}$$

es decir, el estimador S_2' en la ecuación (24) es insesgado.

2. 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 3.3 (Estadístico completo)**:** Un estadístico $T(X)$ es completo si para toda función f, se tiene que

$$\mathbb{E}_\theta\left(f(T)|\theta\right) = 0, \forall\theta \in \Theta \Rightarrow \mathbb{P}_\theta\left(f(T) = 0\right) = 1, \forall\theta \in \Theta. \tag{26}$$

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 3.4: 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

$$\mathbb{E}_\theta\left(f(T)\right) = \sum_{t=0}^{n} f(t)\binom{n}{t}\theta^t(1-\theta)^{n-t} = (1-\theta)^n \sum_{t=0}^{n} f(t)\binom{n}{t}\left(\frac{\theta}{1-\theta}\right)^t, \tag{27}$$

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 3.5: Consideremos dos estimadores, ϕ_1, ϕ_2 insesgados de θ distintos, es decir,

$$\mathbb{E}\left(\phi_1\right) = \mathbb{E}\left(\phi_2\right) = \theta,\ \mathbb{P}\left(\phi_1 \neq \phi_2\right) > 0. \tag{28}$$

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 3.6 (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

$$\mathbb{E}_\theta\left(f(X)\right) = \int_0^\infty f(x)\theta \exp(-\theta x)\mathrm{d}x = 0 \Rightarrow \int_0^\infty f(x)\exp(-\theta x)\mathrm{d}x = 0 \tag{29}$$

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)

$$\int_0^\infty x\hat{g}(x)\exp(-\theta x)\mathrm{d}x = 0, \forall\theta. \tag{30}$$

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.

3. 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 3.4 (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:

$$L : (\Omega \times \Omega) \to \mathbb{R} \tag{31}$$

$$(\theta \times a) \mapsto L(\theta, a). \tag{32}$$

Ejemplo 3.7 (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 3.8 (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 3.9 (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 3.8 para un estimador ϕ del parámetro θ, está dado por:

$$\begin{aligned} R(\theta, \phi) &= \mathbb{E}\left((\theta - \phi)^2\right) \\ &= \mathbb{E}\left((\theta - \bar{\phi} + \bar{\phi} - \phi)^2\right); \quad \text{denotando } \bar{\phi} = \mathbb{E}(\phi) \\ &= \mathbb{E}\left((\theta - \bar{\phi})^2 + 2(\theta - \bar{\phi})\cancel{(\bar{\phi} - \phi)} + (\bar{\phi} - \phi)^2\right) \\ &= \underbrace{(\theta - \bar{\phi})^2}_{=b_\phi^2 \text{ (sesgo}^2\text{)}} + \underbrace{\mathbb{E}\left((\bar{\phi} - \phi)^2\right)}_{=V_\phi \text{ (varianza)}}. \end{aligned} \tag{33}$$

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.

4. 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 2.2), puede solo mejorar un estimador.

Teorema 3.1 (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,

$$\mathbb{E}_\theta\left((\phi^\star - \theta)^2\right) \leq \mathbb{E}_\theta\left((\phi - \theta)^2\right), \forall\theta, \tag{34}$$

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 3.10: 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,

$$\phi^* = \mathbb{E}_\theta\left(X_1 \middle| \sum_i X_i = t\right). \tag{35}$$

Para calcular esta esperanza condicional, observemos primero que

$$\sum_{j=1}^n \mathbb{E}_\theta\left(X_j \middle| \sum_{i=1}^n X_i = t\right) = \mathbb{E}_\theta\left(\sum_{j=1}^n X_j \middle| \sum_{i=1}^n X_i = t\right) = t, \tag{36}$$

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

$$\phi^* = \frac{t}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{37}$$

Antes de demostrar el Teorema 3.1 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

$$\begin{aligned}
\mathbb{E}_Y \mathbb{E}_{X|Y}(X|Y) &= \int_{\mathcal{Y}} \int_{\mathcal{X}} x \mathrm{d}P(x|y) \mathrm{d}P(y) && \text{def. esperanza} \\
&= \int_{\mathcal{X}} x \int_{\mathcal{Y}} \mathrm{d}P(x|y) \mathrm{d}P(y) && \text{linealidad} \\
&= \int_{\mathcal{X}} x \int_{\mathcal{Y}} \mathrm{d}P(x,y) && \text{def. esperanza condicional} \\
&= \int_{\mathcal{X}} x \mathrm{d}P(x) = \mathbb{E}_X(X). && \text{def. esperanza} \qquad (38)
\end{aligned}$$

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

$$0 \leq \mathbb{V}(X) = \mathbb{E}\left(X^2\right) - \mathbb{E}(X)^2 \Rightarrow \mathbb{E}\left(X^2\right) \geq \mathbb{E}(X)^2. \tag{39}$$

La desigualdad de Jensen es geométricamente intuitiva, como se observa en la Figura 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 (38) y (39), podemos demostrar el teorema anterior.

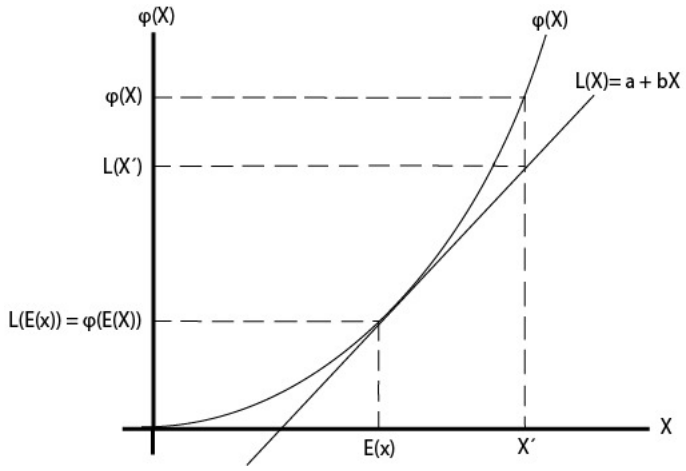

Figura 1. Intuición geométrica de la desigualdad de Jensen

DEMOSTRACIÓN DE TEOREMA 3.1. 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 (39), 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 3.3: Notemos que si el estimador ϕ es insesgado, su *Rao-Blackwellización* ϕ^* también lo es, en efecto

$$\mathbb{E}_\theta\left(\phi^*\right)=\mathbb{E}_\theta\left(\mathbb{E}_\theta\left(\phi|T\right)\right)=\mathbb{E}_\theta\left(\phi\right)=\theta, \tag{40}$$

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

5. Varianza uniformemente mínima

Observemos que, en base al riesgo cuadrático definido en la ecuación (33), si un estimador es insesgado (Definición 3.2) entonces su riesgo cuadrático es únicamente su varianza. Esto motiva la siguiente definición de optimalidad para estimadores insesgados.

Definición 3.5 (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

$$\mathbb{V}_\theta(\phi) \leq \mathbb{V}_\theta(\phi'), \forall \theta \in \Theta. \tag{41}$$

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

Ejemplo 3.11: 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

$$\mathbb{E}_\theta(\phi_1(X)) = \mathbb{E}_\theta(\phi_2(X)) = \mathbb{E}_\theta(\phi_3(X)) = \theta. \tag{42}$$

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

$$\mathbb{V}_\theta(\phi_3(X)) = \mathbb{V}_\theta\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\sum_{i=1}^n \mathbb{V}_\theta(X_i) = \frac{\theta(1-\theta)}{n} \tag{43}$$

pues $\mathbb{V}_\theta(X_i) = \mathbb{E}_\theta\left((\theta - X_i)^2\right) = \mathbb{E}_\theta\left(X_i^2\right) - \theta^2 = (0^2 \cdot (1-\theta) + 1^2 \cdot \theta) - \theta^2 = \theta(1-\theta)$. 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 3.2 (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

$$\phi(T) = \mathbb{E}_\theta \left(\phi'(X) | T \right), \tag{44}$$

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 3.12 (EIVUM para Bernoulli)**:** Recordemos que en el Ejemplo 3.4 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,

$$\mathbb{E}_\theta \left(\phi(T) \right) = \mathbb{E}_\theta \left(T/n \right) = \sum_{i=1}^n \mathbb{E}_\theta \left(X_i \right) / n = \theta, \tag{45}$$

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

Observación 3.4: Lectura personal: Estadístico auxiliar (ancilliary) y teoremas de Basur y de Bahadur.

6. 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

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 $ECM_{\sigma^2}(\hat{\sigma}^2) < ECM_{\sigma^2}(S^2)$

Capítulo 4

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 $\Theta = \theta$.

1. Contexto y definiciones principales

Definición 4.1 (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,

$$p(x, \theta) = p(x|\theta)p(\theta) \tag{46}$$

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 (46), podemos definir:

Definición 4.2 (Distribución marginal)**:** La distribución de X, obtenida mediante la desintegración de parámetro Θ del par (X, Θ), es decir

$$p(x) = \int_\Omega p(x|\theta)p(\theta)\mathrm{d}\theta \tag{47}$$

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 4.3 (Función de verosimilitud)**:** La densidad de probabilidad evaluada en un conjunto de observaciones $\mathcal{D}$ como función del valor del parámetro Θ, es decir

$$L : \Omega \to \mathbb{R} \tag{48}$$

$$\theta \mapsto l(\theta) = L_{\mathcal{D}}(\theta) = p(\mathcal{D}|\theta), \tag{49}$$

recibe el nombre de función de verosimilitud, o en inglés, *likelihood*.

Observación 4.1: La función de verosimilitud no es una densidad de probabilidad, es decir, no es cierto que

$$\int_{\Omega} L(\theta)\mathrm{d}\theta = 1 \tag{50}$$

Observación 4.2: 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,

$$l(\theta) = \log L(\theta) = \log p(\mathcal{D}|\theta). \tag{51}$$

Esta formulación será particularmente útil cuando queramos optimizar la verosimilitud.

Observación 4.3: 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:

$$l_{\mathcal{D}}(\theta) = \sum_{i=1}^{n} l_{X_i}(\theta) \tag{52}$$

Ejemplo 4.1: 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:

$$L(\mu) = p(\mathcal{D}|\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-1}{2\sigma^2}(x_i - \mu)^2\right)$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right). \tag{53}$$

Luego, la log-verosimilitud está dada por:

$$l(\mu) = \log L(\mu) = -\frac{n}{2}\log(2\pi\sigma^2) + \frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2. \tag{54}$$

Ahora estamos en condiciones de definir el elemento central de la inferencia bayesiana, sobre el cual todo el proceso de inferencia toma lugar.

Definición 4.4 (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

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D})|\theta)p(\theta)}{p(\mathcal{D})} \propto p(\mathcal{D}|\theta)p(\theta) \tag{55}$$

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 (55) 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 4.4: El símbolo $\propto$ en la ecuación (55) 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 4.2 (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 4.5: 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 4.6: En la actualización bayesiana en línea, la posterior de la etapa n sirve de prior de la etapa $n+1$.

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 4.5: 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

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \tag{56}$$

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 4.3 (continuación de Ejemplo 4.2)**:** Tarea: Verifique si el Ejemplo 4.2 es en efecto uno de prior conjugado.

Ejemplo 4.4 (Distribución Multinomial)**:** Consideremos una variable aleatoria multinomial $X \sim \text{Mult}\,(n,\theta)$ donde θ pertenece al simplex

$$\{\theta \in [0,1]^k : \theta_1 + \cdots + \theta_k = 1\}. \tag{57}$$

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

$$\theta_{\text{dado}} \sim \text{Mult}\left(100, \left[\tfrac{1}{6}, \tfrac{1}{6}, \tfrac{1}{6}, \tfrac{1}{6}, \tfrac{1}{6}, \tfrac{1}{6}\right]\right). \tag{58}$$

Denotando $X = [x_1, \ldots, x_n]$, observemos que una muestra multinomial $X \sim \text{Mult}(n, \theta)$ cumple con

$$\{x_i\}_{i=1}^k \subset \{0, 1, \ldots, n\}, \quad \sum_{i=1}^k x_i = n. \tag{59}$$

Finalmente, la distribución Multinomial está dada por

$$\text{Mult}(X; n, \theta) = \frac{n!}{x_1! \cdots x_k!} \theta_1^{x_1} \cdots \theta_k^{x_k}, \tag{60}$$

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 4.6 (Distribución de Dirichlet)**:** Consideremos la distribución de Dirichlet

$$\theta \sim \text{Dir}(\theta|\alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^k \theta_i^{\alpha_i - 1}, \tag{61}$$

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 (57).

En el caso $k = 3$, la distribución de Dirichlet puede ser graficada en el simplex de 2 dimensiones. La Figura 1 presenta tres gráficos para distintos valores del parámetro de concentración.

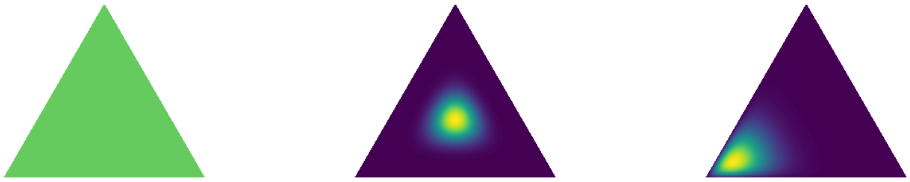

Figura 1. Distribuciones Dirichlet para $k = 3$ con parámetros de concentración α (desde izquierda a derecha) dado por $[1,1,1]$, $[10,10,10]$ y $[10,2,2]$.

.

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

$$\begin{aligned} p(\theta|x) &= \frac{\text{Mult}(x;n,\theta)\,\text{Dir}(\theta;\alpha)}{p(x)} \\ &= \frac{n!}{x_1!\cdots x_k! p(x) B(\alpha)} \prod_{i=1}^{k} \theta_i^{x_i+\alpha_i-1} \\ &= \frac{1}{B(\alpha')} \prod_{i=1}^{k} \theta_i^{\alpha_i'-1} \end{aligned} \tag{62}$$

donde $\alpha' = (\alpha_1', \ldots, \alpha_k') = (\alpha_1' + x_1, \ldots, \alpha_k' + x_k)$ es el nuevo parámetro de concentración.

Ejemplo 4.5: 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 (62). La Figura 2 muestra 50 muestras de la distribución posterior para distintas cantidades de observaciones entre 0 y 10^5.

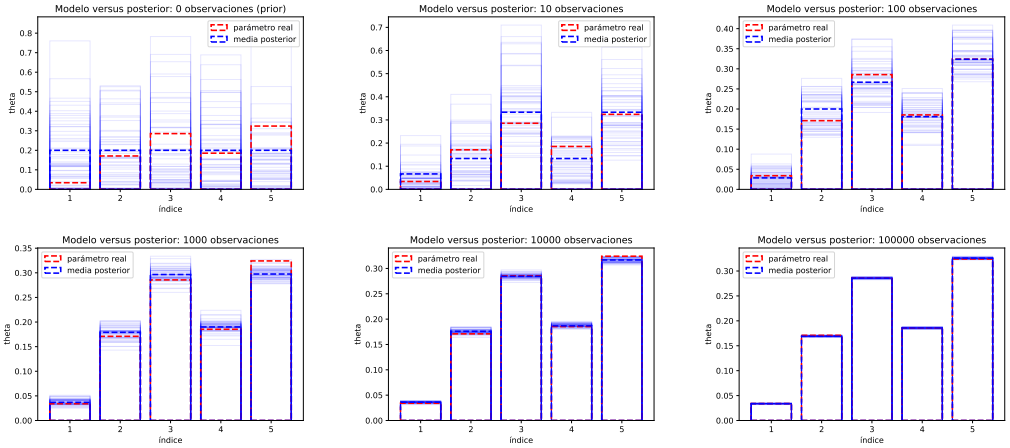

Figura 2. Concentración de la distribución posterior en torno al parámetro real para un modelo $X \sim \text{Mult}(\theta)$ y una distribución a priori Dirichlet $\theta \sim \text{Dir}(\alpha)$. Se considera desde 0 hasta 10^5 observaciones y cada gráfico (desde izquierda-arriba hasta derecha-abajo) 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 4.6: 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

$$p(\mu|\mathcal{D}) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right) \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right) \tag{63}$$

$$\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right), \tag{64}$$

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. (64), obtenemos

$$p(\mu|\mathcal{D}) \propto \exp\left(-\frac{1}{2\sigma_n^2}(\mu - \mu_n)^2\right), \tag{65}$$

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

$$\int_{\mathbb{R}} \exp\left(-\frac{1}{2\sigma_n^2}(\mu-\mu_n)^2\right) \mathrm{d}\mu = (2\pi\sigma_n^2)^{n/2}. \tag{66}$$

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):

$$p(\mu|\mathcal{D}) = \mathcal{N}(\mu;\mu_n,\sigma_n^2) = \frac{1}{(2\pi\sigma_n^2)^{N/2}} \exp\left(-\frac{1}{2\sigma_n^2}(\mu-\mu_n)^2\right), \tag{67}$$

donde la media y la varianza están dadas respectivamente por

$$\mu_n = \frac{1}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}}\left(\frac{1}{\sigma_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{x}\right), \quad \text{donde } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{68}$$

$$\sigma_n = \left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1}. \tag{69}$$

Observación 4.7: 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. (68) y (69) 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. (69) 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 4.7: Modelo gaussiano (μ conocido). Ahora procedemos con el siguiente prior para la varianza, llamado Gamma-inverso:

$$p(\sigma^2) = \text{inv-}\Gamma(\sigma^2;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)(\sigma^2)^{\alpha+1}} \exp(-\beta/\sigma^2) \tag{70}$$

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:

$$p(\sigma^2|\mathcal{D}) \propto \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right) \frac{\beta^\alpha}{\Gamma(\alpha)(\sigma^2)^{\alpha+1}} \exp(-\beta/\sigma^2) \tag{71}$$

$$\propto \frac{1}{(\sigma^2)^{N/2+\alpha+1}} \exp\left(-\frac{1}{\sigma^2}\left(\frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^2+\beta\right)\right)$$

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 4.7 (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 4.8: No es necesario usar la constante de normalización en las densidades a priori Gaussianas (o ninguna otra en realidad).

Observación 4.9: 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

$$p(\mu|x) \propto p(x|\mu)\cdot 1 = \mathcal{N}(x;\mu,1) = \mathcal{N}(\mu;x,1). \tag{72}$$

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.

3. Máxima Verosimilitud

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 4.3 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 4.8 (Estimador de máxima verosimilitud (MV))**:** Sea una observación x y una función de verosimilitud $L(\theta)$, el estimador de máxima verosimilitud está dado por

$$\theta_{\text{MV}} = \arg\max_{\theta} L(\theta|x) \tag{73}$$

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

$$L(\theta) = \prod_{i=1}^{n} \theta_i^x (1-\theta)^{1-x_i}, \tag{74}$$

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 4.1: Graficar $l(\theta)$ en el Ejemplo 4.8.

Ejercicio 4.2: Encuentre el estimador de MV de $\theta = (\mu, \Sigma)$ para la VA $X \sim \mathcal{N}(\mu, \Sigma)$.

Ejemplo 4.9: 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

$$L(\theta) = \prod_{i=1}^{n} p_\theta(x_i) \tag{75}$$

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

$$L(\theta) = \theta^{-n}, \quad \theta \ge \text{máx}\{x_i\}_1^n \tag{76}$$

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

4. EMV en práctica: tres ejemplos

4.1. 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

$$X \sim \mathcal{N}\left(\theta_0 + \theta_1 t^2 + \theta_2 \cos(2\pi t/12), \theta_3^2\right), \tag{77}$$

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

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

$$= \frac{n}{2}\log\left(2\pi\theta_3^2\right) - \frac{1}{2\theta_3^2}\sum_{i=1}^n (x_i - \theta_0 - \theta_1 t_i^2 - \theta_2 \cos(2\pi t_i/12))^2 \tag{78}$$

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 (78) es cuadrática en $[\theta_0, \theta_1, \theta_2]$.

4.2. 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

$$X|t \sim \text{Ber}\left(\sigma(t)\right) \tag{79}$$

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,

$$\Pr\left(X = 1|t\right) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 t)}}. \tag{80}$$

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.

4.3. 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,

$$X \sim p\mathcal{N}(X|\mu_H, \Sigma_H) + (1-p)\mathcal{N}(X|\mu_M, \Sigma_M) \tag{81}$$

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

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

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:

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

5. Propiedades del EMV

5.1. 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. Definamos en primer lugar la siguiente *divergencia*.

Definición 4.9 (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

$$\mathrm{KL}\left(f\|g\right)=\int_{\mathcal{X}} f(x)\log\left(\frac{f(x)}{g(x)}\right)\mathrm{d}x. \tag{82}$$

Observación 4.10: La divergencia KL es siempre positiva $\forall f,g$ (desigualdad de Gibbs):

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

Adempás, 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 $g(x)=f(x)$.

Observación 4.11: Intuición y gráfico de la KL en relación al soporte.

Observación 4.12: Intuición de la KL desde la teoría de la información (entropía).

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.

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

$$M_n(\theta') = n^{-1}(l_n(\theta') - l_n(\theta)) = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{p_{\theta'}(x_i)}{p_\theta(x_i)}\right), \tag{83}$$

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 (83) 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

$$M_n(\theta') \to \mathbb{E}_\theta\left(\log\left(\frac{p_{\theta'}(x)}{p_\theta(x)}\right)\right) = -\mathbb{E}_\theta\left(\log\left(\frac{p_\theta(x)}{p_{\theta'}(x)}\right)\right) = -\mathrm{KL}\left(p_\theta \| p_{\theta'}\right). \tag{84}$$

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

Observación 4.14: 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 4.10: (Cálculo del EMV en Gaussiana: varianza versus precisión versus log-precisión versus cholesky - reparametrisation trick)

5.2. 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.

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,

$$S_\theta(X) = \frac{\partial \log p_\theta(X)}{\partial \theta}. \tag{85}$$

Observación 4.15: 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

(86)

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

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 4.10 (Información de Fisher)**:** La cantidad denotada mediante

$$I(\theta) = \mathbb{E}_\theta\left(\left(\frac{\partial \log p_\theta(X)}{\partial \theta}\right)^2\right) = -\mathbb{E}_\theta\left(\frac{\partial^2 \log p_\theta(X)}{\partial \theta^2}\right), \tag{87}$$

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

$$\mathbb{V}_\theta\left(S_\theta(X)\right) = \mathbb{E}_\theta\left(S_\theta(X)^2\right) - \cancel{\mathbb{E}_\theta\left(S_\theta(X)\right)^2} = \mathbb{E}_\theta\left(\left(\frac{\partial \log p_\theta(X)}{\partial\theta}\right)^2\right). \tag{88}$$

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 4.3 (Cálculo de la información de Fisher para Bernoulli)**:** Consideremos $X \sim \text{Ber}\,(\theta)$, entonces,

$$\begin{aligned} I(\theta) &= -\mathbb{E}_\theta\left(\frac{\partial^2}{\partial\theta^2}\log\left(\theta^X(1-\theta)^{1-X}\right)\right) \\ &= -\mathbb{E}_\theta\left(\frac{\partial^2}{\partial\theta^2}X\log\theta + \frac{\partial^2}{\partial\theta^2}(1-X)\log\left(1-\theta\right)\right) \\ &= \mathbb{E}_\theta\left(X\theta^{-2} + (1-X)(1-\theta)^{-2}\right) \\ &= \theta^{-1} + (1-\theta)^{-1} \\ &= \frac{1}{\theta(1-\theta)}. \end{aligned} \tag{89}$$

Ejercicio 4.4 (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

$$S_\theta(X_1,\ldots,X_n) = \frac{\partial \log \prod_{i=1}^n p_\theta(X_i)}{\partial\theta} = \sum_{i=1}^n \frac{\partial \log p_\theta(X_i)}{\partial\theta} = \sum_{i=1}^n S_\theta(X_i). \tag{90}$$

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

$$I_n(\theta) = \mathbb{V}_\theta\left(\sum_{i=1}^n S_\theta(X_i)\right) = nI(\theta). \tag{91}$$

Observación 4.16: 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.

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

$$\mathbb{E}_\theta\left(\hat\theta(X)-\theta\right) = \int_{\mathcal{X}} (\hat\theta(X)-\theta)p_\theta(X)\mathrm{d}x = 0. \tag{92}$$

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

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

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 4.11 (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

$$\mathbb{V}_\theta\left(\theta'\right) \geq (nI(\theta))^{-1}, \quad \forall\theta \in \Theta \tag{93}$$

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

Ahora podemos finalmente volver al concepto de normalidad asintótica. 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

$$\sqrt{n}({\theta_{\text{MV}}}^{(n)} - \theta) \to \mathcal{N}(0, (I(\theta))^{-1}), \tag{94}$$

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.

6. Estimación y predicción

6.1. 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, definimos el siguiente estimador bayesiano:

Definición 4.12 (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).

Observación 4.17: 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 4.18: 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 4.11 (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?

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 4.13 (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

$$R(\hat{\theta}) = \int_{\Omega} L(\theta, \hat{\theta}) p(\theta|\mathcal{D}) \mathrm{d}\theta. \tag{95}$$

Definición 4.14 (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:

$$\theta_{\text{Bayes}} = \arg\min_{\Omega} R(\theta). \tag{96}$$

Ejemplo 4.12: 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 $\theta_{\text{Bayes}} = \mathbb{E}(\theta|\mathcal{D})$. De forma similar,la función de costo *lineal* $L_1(\theta, \hat{\theta}) = |\theta - \hat{\theta}|_1$ 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

$$R(\hat{\theta}) = \mathbb{P}\left(\theta \neq \hat{\theta}|\mathcal{D}\right) = 1 - \mathbb{P}\left(\theta = \hat{\theta}|\mathcal{D}\right), \tag{97}$$

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?

6.2. 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 4.15 (Posterior predictiva)**:** Para un conjunto de datos $\mathcal{D}$ y un parámetro θ, la densidad posterior predictiva está dada por

$$p(x|\mathcal{D}) = \int_{\Omega} p(x|\theta)p(\theta|\mathcal{D})\mathrm{d}\theta = \mathbb{E}\left(p(x|\theta)|\mathcal{D}\right), \tag{98}$$

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 4.19: 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.

7. 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

$$p(\eta) = \text{Uniforme}\,(c, d) = \frac{1}{d-c}. \tag{99}$$

Observemos que la elección uniforme del parámetro θ en el intervalo $[a, b]$ es equivalente a elegir η según

$$\tilde{p}(\eta) = p(\theta) \left|\frac{d\theta}{d\eta}\right| = \frac{1}{b-a} \left|\frac{d \log \eta}{d\eta}\right| = \frac{1}{\eta(b-a)}, \tag{100}$$

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 4.20: ¿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,

$$p(\theta) \propto (I(\theta))^{1/2}, \tag{101}$$

donde recordemos que la información de Fisher está dada por

$$I(\theta) = -\mathbb{E}_{\theta}\left(\frac{\partial^2}{\partial\theta^2} \log p(X|\theta)\right) = \mathbb{E}_{\theta}\left(\left(\frac{\partial}{\partial\theta} \log p(X|\theta)\right)^2\right). \tag{102}$$

Además, si $X_1, \ldots, X_n$ son iid, entonces $I(\theta) = nI_1(\theta)$ y el prior de Jeffreys puede ser expresado como

$$p(\theta) \propto I_1(\theta)^{1/2}. \tag{103}$$

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

$$X \sim p(x|\theta),\ \theta \in \Omega \quad \& \quad X \sim q(x|\eta),\ \eta \in \Gamma, \tag{104}$$

donde $\eta = h(\theta)$. Las informaciones de Fisher para ambos modelos, denotadas respectivamente $I_p(\theta)$ e $I_q(\theta)$, están relacionadas mediante

$$\begin{aligned} I_p(\theta) &= \int_{\mathcal{X}} \left(\frac{\partial}{\partial\theta}\log p(x|\theta)\right)^2 p(x|\theta)\mathrm{d}x \\ &= \int_{\mathcal{X}} \left(\frac{\partial}{\partial\theta}\log q(x|h(\theta))\right)^2 q(x|h(\theta))\mathrm{d}x \\ &= \int_{\mathcal{X}} \left(\frac{\partial}{\partial\eta}\log q(x|\eta)h'(\theta)\right)^2 q(x|\eta)\mathrm{d}x \\ &= \left(h'(\theta)\right)^2 I_q(\eta). \end{aligned} \tag{105}$$

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

$$p(\theta) = p_J(\eta)\left|\frac{d\eta}{d\theta}\right| = \sqrt{I_q(\eta)}\,\left|h'(\theta)\right| = \sqrt{I_p(\theta)} = p_J(\theta). \tag{106}$$

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 $f(x|\theta_2)$.

Capítulo 5

Más Sobre Estimadores

1. 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 5.1 (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

$$\mathbb{P}_\theta\left(A(X) \leq \theta \leq B(X)\right) = 1-\alpha, \forall \theta \in \Theta. \tag{107}$$

Observación 5.1: 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 (107) 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.

Ejemplo 5.1 (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

$$\sqrt{n}(\bar{X} - \theta) \sim \mathcal{N}(0, 1). \tag{108}$$

Esta cantidad se llama *pivote* y es una función (de la VA y del parámetro) cuya distribución no depende del parámetro. 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

(109)

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

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

$$(\bar{X} - 1{,}96/\sqrt{n}, \bar{X} + 1{,}96/\sqrt{n}). \tag{110}$$

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 5.2: 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 5.1, un $(1-\alpha)$-conjunto de confianza $S(X)$ es tal que

$$\mathbb{P}_\theta\left(\theta \in S(X)\right) = 1 - \alpha, \forall \theta \in \Theta. \tag{111}$$

Ejercicio 5.1: Considere $X_1, \ldots, X_{50} \sim \mathcal{N}(0, \sigma^2)$, calcule el intervalo de confianza del 99 % para σ.

Ejemplo 5.2 (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

$$\hat{\theta} \sim \mathcal{N}\left(\theta, \frac{\theta(1-\theta)}{n}\right), \tag{112}$$

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

$$\frac{\sqrt{n}(\hat{\theta} - \theta)}{\sqrt{\theta(1-\theta)}} \sim \mathcal{N}(0, 1), \tag{113}$$

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 5.2 (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 5.3 (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 $(1 - \alpha)$ %.

2. Intervalos de Confianza Bayesianos

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. Para diferenciar estos intervalos con el caso frecuentista, veamos la siguiente definición.

Definición 5.2: Sea π el prior del parámetro θ, un conjunto C_x se dice α-creíble si la posterior correspondiente 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 5.3: 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 5.4: 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 5.3: Si los intervalos de confianza y credibilidad para a la media de la gaussiana son el mismo, ¿cuál es la diferencia?

Ejemplo 5.5: Encuentre el 85 %-intervalo creíble de λ en $x_1, \ldots, x_n \sim \exp(\lambda)$ cuando el prior es uniforme. Tenemos

$$p(\lambda | x_1, \ldots, x_n) \propto \lambda^n e^{-\lambda \sum_i x_i}, \tag{114}$$

con lo que concluimos que

$$p(\lambda | x_1, \ldots, x_n) = \Gamma(n+1, \sum_i x_i). \tag{115}$$

Debemos ahora encontrar a, b tal que

$$\int_a^b \frac{(\sum_i x_i)^{n+1}}{\Gamma(n+1)} \lambda^n e^{-\lambda \sum_i x_i} \mathrm{d}\lambda = 0{,}85 \tag{116}$$

donde tenemos las 3 opciones mencionadas arriba.

Ejercicio 5.3: Encuentre el intervalo creíble para θ en el Ejemplo 5.3

Capítulo 6

Test de Hipótesis

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 incertibumbre (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 $X \sim P_\theta$

$$H_0 : P_\theta \in \mathcal{P}_0 \tag{117}$$

$$H_1 : P_\theta \notin \mathcal{P}_0 \tag{118}$$

- **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:

$$L(\theta, a) = (\theta - a)^2. \tag{119}$$

Ejemplo 6.1: (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:

$$\delta(\cdot) : \mathcal{X} \to \mathcal{A}. \tag{120}$$

- $\mathcal{D} = \{\delta : \mathcal{X} \to \mathcal{A}\}$ 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,

$$R(\theta, \delta) = \mathbb{E}_\theta \left(L(\theta, \delta(X)) \right). \tag{121}$$

Ejemplo 6.2: 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,

$$\mathcal{A} = \Theta = \mathbb{R}. \tag{122}$$

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,

$$\mathcal{A} = \{cX, c \in [0, 1]\}. \tag{123}$$

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

$$R(\theta, \hat{\theta}) = \mathbb{E}_\theta \left((\hat{\theta}(X) - \theta)^2 \right) = \mathbb{V}_\theta \left(\hat{\theta}(X) \right) + \mathbb{E}_\theta \left(\hat{\theta}(X) - \theta \right)^2 = c^2 + (c-1)^2 \theta^2 \tag{124}$$

¿Qué valor de c sugiere elegir?

Observación 6.1: 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**.

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

$$X = X_1, \ldots, X_{50} \sim \mathcal{N}(3000, 500^2), \tag{125}$$

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

$$\bar{X} = \frac{1}{50} \sum_{i=1}^{50} X_i \sim \mathcal{N}(3000, 500^2/50). \tag{126}$$

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)

$$z = \frac{\bar{X} - 3000}{500/\sqrt{50}} \sim \mathcal{N}(0, 1), \tag{127}$$

con el cual podemos realizar el cálculo:

$$\mathbb{P}\left(\bar{X} \geq 3200\right) = \mathbb{P}\left(z = \frac{\bar{X} - 3000}{500/\sqrt{50}} \geq 2\sqrt{2}\right) = 0{,}0023388674905235884, \tag{128}$$

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

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

[1] Acrónimo de *cumulative denstiy function*.

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 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)

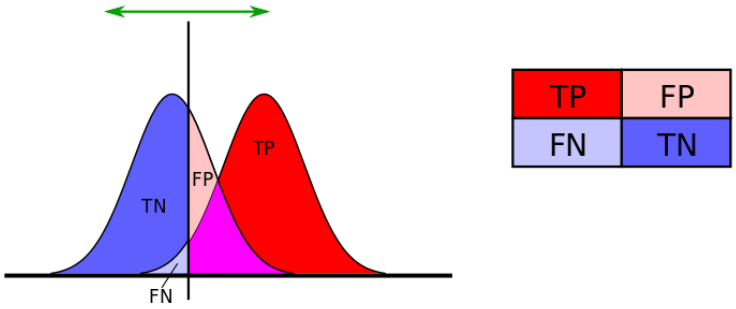

Figura 1. Ilustración de tipos de errores (adaptada de Sharpr - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=44059691)

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 %.

[2] Acrónimo para *percent point function*.

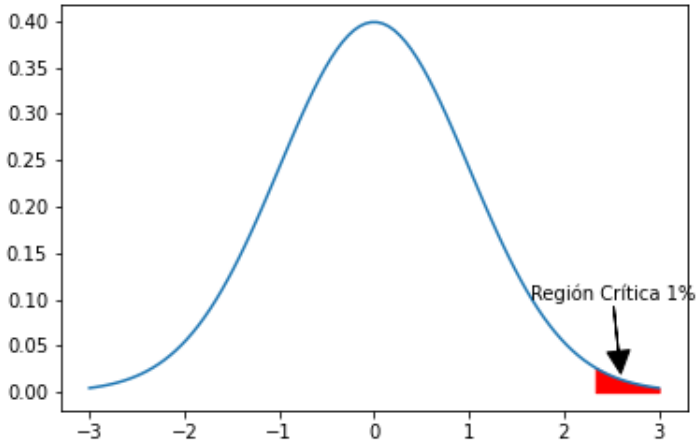

Figura 2. La región crítica son aquellos valores que tienen una probabilidad menor al nivel de significancia, en este caso, el 1 %

Observación 6.2: 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 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.

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

3. Rechazo, potencia y nivel

Formalmente, frente a dos hipótesis generales denotadas por

$$H_0 : \theta \in \Theta_0 \tag{129}$$

$$H_1 : \theta \in \Theta_1, \tag{130}$$

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

$$\phi : \mathcal{X} \to \{0, 1\}, \tag{131}$$

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 6.1 (Región crítica de un test)**:** La región crítica o región de rechazo de un test de hipótesis ϕ se define como

$$R_\phi = \{x \in \mathcal{X} | \phi(x) = 1\} = \phi^{-1}(1). \tag{132}$$

Definición 6.2 (Función de probabilidad de rechazo)**:** Para un test ϕ y cualquier parámetro $\theta \in \Theta$ podemos definir la probabilidad de rechazo mediante

$$\alpha_\phi(\theta) = \mathbb{P}_\theta(\phi(x) = 1) = \mathbb{P}_\theta\left(x \in R_\phi\right), \forall \theta \in \Theta, \tag{133}$$

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 6.3 (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,

$$\pi_\phi(\theta) = \mathbb{P}(\text{rechazar } H_0 | H_1 \text{ es cierta}) = \mathbb{P}_{\theta_1}(\phi(x) = 1). \tag{134}$$

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 6.3 (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 6.4 (Nivel de un test)**:** Decimos que un test es de nivel $\alpha \in [0,1]$ si

$$\alpha_\phi(\theta) \leq \alpha, \forall \theta \in \Theta_0, \tag{135}$$

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

$$\pi_{\phi^\star}(\theta) \geq \pi_\phi(\theta), \forall \theta \in \Theta_1. \tag{136}$$

4. Test de Neyman-Pearson

Consideremos el siguiente problema de test de dos hipótesis simples.

$$H_0 : \theta \in \Theta_0 = \{\theta_0\} \quad \text{v.s.} \quad H_1 : \theta \in \Theta_1 = \{\theta_1\}, \tag{137}$$

donde por una notación más simple escribiremos simplemente

$$H_0 : \theta = \theta_0 \quad \text{v.s.} \quad H_1 : \theta = \theta_1, \tag{138}$$

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

$$R^* = \{x \in \mathcal{X} | p_1(x) \geq k p_0(x)\}, \tag{139}$$

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,

$$\phi^*(x) = 1 \Leftrightarrow x \in R^*. \tag{140}$$

Finalmente, determinaremos la constante k de tal manera de que

$$\alpha_{\phi^\star}(\theta_0) = \mathbb{P}_{\theta_0}(x \in R^*) = \alpha, \ \alpha \in [0,1], \tag{141}$$

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

Lema 6.1 (Neyman-Pearson)**:** Consideremos un test de hipótesis de la forma

$$H_0 : \theta = \theta_0 \quad \text{v.s.} \quad H_1 : \theta = \theta_1 \tag{142}$$

con probabilidad de rechazo dada por

$$\alpha = \mathbb{P}(p_1(X) \geq k p_0(X)). \tag{143}$$

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 ($\forall \theta$)

$$\mathbb{P}_\theta(R) = \int_R p_\theta(x)\mathrm{d}x. \tag{144}$$

Luego, podemos escribir

$$\mathbb{P}_\theta(R) = \mathbb{P}_\theta(R \cap R^*) + \mathbb{P}_\theta(R \cap \bar{R}^*) \tag{145}$$

$$\mathbb{P}_\theta(R^*) = \mathbb{P}_\theta(R^* \cap R) + \mathbb{P}_\theta(R^* \cap \bar{R}), \tag{146}$$

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

$$\begin{aligned}
\mathbb{P}_{\theta_1}(R^*) - \mathbb{P}_{\theta_1}(R) &= \mathbb{P}_{\theta_1}(R^* \cap \bar{R}) - \mathbb{P}_{\theta_1}(R \cap \bar{R}^*) \\
&= \int_{R^* \cap \bar{R}} p_{\theta_1}(x)\mathrm{d}x - \int_{R \cap \bar{R}^*} p_{\theta_1}(x)\mathrm{d}x \\
&\geq k \int_{R^* \cap \bar{R}} p_{\theta_0}(x)\mathrm{d}x - k \int_{R \cap \bar{R}^*} p_{\theta_0}(x)\mathrm{d}x \quad [\text{pues } p_1 \geq k p_0 \text{ en } R^*] \\
&= k\left(\mathbb{P}_{\theta_0}(R^* \cap \bar{R}) - \mathbb{P}_{\theta_0}(R \cap \bar{R}^*)\right) \\
&= k\left(\underbrace{\mathbb{P}_{\theta_0}(R^*)}_{=\alpha} - \underbrace{\mathbb{P}_{\theta_0}(R)}_{\leq \alpha}\right) \quad [\text{primera igualdad de este desarrollo}] \\
&\geq 0
\end{aligned} \tag{147}$$

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

$$H_0 : \theta = \theta_0 \quad \text{v.s.} \quad H_1 : \theta = \theta_1. \tag{148}$$

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

$$p_i(x) = \theta_i^{\sum x_j}(1-\theta_i)^{n-\sum x_j},\ i = 0, 1. \tag{149}$$

Para rechazar H_0 según el test de Neyman-Pearson, es decir, $x \in R^*$ de acuerdo a la ecuación (139), el test requiere:

$$\frac{p_1(x)}{p_0(x)} = \frac{\theta_1^{\sum x_j}(1-\theta_1)^{n-\sum x_j}}{\theta_0^{\sum x_j}(1-\theta_0)^{n-\sum x_j}} = \left(\frac{1-\theta_1}{1-\theta_0}\right)^n \left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right)^{\sum x_j} \geq k. \tag{150}$$

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 (150) es equivalente a

(151)

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

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

$$R^* = \left\{(x_1, \ldots, x_1) \text{ t.q. } \sum x_j \geq k'\right\}. \tag{152}$$

5. 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,

$$\frac{\hat{\theta} - \theta_0}{\text{ee}} \sim \mathcal{N}(0, 1), \tag{153}$$

cuando el número de observaciones tiende a infinito y 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

$$H_0 : \theta = \theta_0 \quad \text{v.s.} \quad H_1 : \theta \neq \theta_0, \tag{154}$$

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

$$|W| \geq z_{\alpha/2}, \tag{155}$$

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

$$\mathbb{P}_{\theta_0}\left(|W| \geq z_{\alpha/2}\right) = \mathbb{P}_{\theta_0}\left(\left|\frac{\hat{\theta}-\theta_0}{\text{ee}}\right| \geq z_{\alpha/2}\right) \to \mathbb{P}_{\theta_0}\left(|Z| \geq z_{\alpha/2}\right) = \alpha \tag{156}$$

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

Ejemplo 6.5: 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

$$H_0 : \mu_x = \mu_y \quad \text{v.s.} \quad H_1 : \mu_x \neq \mu_y, \tag{157}$$

lo cual está dentro del alcance del test de Wald denotando $\delta = \mu_x - \mu_y$ e identificando las hipótesis

$$H_0 : \delta = 0 \quad \text{v.s.} \quad H_1 : \delta \neq 0. \tag{158}$$

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

$$W = \frac{\hat{\delta}-0}{\text{ee}} = \frac{\bar{X}-\bar{Y}}{\sqrt{\frac{1}{n}s_x^2 + \frac{1}{m}s_y^2}}, \tag{159}$$

y rechazamos rechazamos H_0 si

$$|\bar{X}-\bar{Y}| \geq z_{\alpha/2}\sqrt{\frac{1}{n}s_x^2 + \frac{1}{m}s_y^2}, \tag{160}$$

donde recordemos que el lado derecho decae como $1/\sqrt{n}$.

Observación 6.4: 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

$$\theta_0 \notin (\hat{\theta} - \text{ee} z_{\alpha/2}, \hat{\theta} + \text{ee} z_{\alpha/2}), \tag{161}$$

es decir, realizar el test de Wald es equivalente a calcular el α intervalo de confianza para el parámetro θ_0 asumiendo normalidad.

6. 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,

$$H_0 : \theta \in \Theta_0 \quad \text{v.s.} \quad H_1 : \theta \notin \Theta_0. \tag{162}$$

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

$$\lambda(x_1, \ldots, x_n) = \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta} L(\theta)} = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})} \leq C, \tag{163}$$

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

$$R^* = \{x \in \mathcal{X} | \lambda(x) \leq C\}. \tag{164}$$

Observación 6.5: 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 6.6: Notemos que podemos escribir la expresión en la ecuación (163) como

$$\lambda(x_1, \ldots, x_n) = \mathbb{1}_{\hat{\theta} \in \Theta_0} + \mathbb{1}_{\hat{\theta} \notin \Theta_0} \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \notin \Theta_0} L(\theta)} \tag{165}$$

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

$$\tilde{\lambda}(x_1, \ldots, x_n) = \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \notin \Theta_0} L(\theta)} \leq \tilde{k}. \tag{166}$$

Ejemplo 6.6 (TRV Bernoulli)**:** Sea $X_1, \ldots, X_n \sim \text{Ber}(\theta)$ iid, se quiere resolver

$$H_0 : \theta \leq \theta_0 \quad \text{v.s.} \quad H_1 : \theta > \theta_0, \tag{167}$$

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

$$\Theta_0 = [0, \theta_0] \quad \& \quad \Theta_1 = (\theta_0, 1] \tag{168}$$

calculamos el EMV (restricto e irrestricto) mediante

$$\hat{\theta} = \bar{x} \quad \text{irrestricto} \tag{169}$$

$$\hat{\theta}_0 = \bar{x} \quad \text{si } \bar{x} \in \Theta_0,\ \theta_0 \text{ si no.} \tag{170}$$

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

$$\lambda(x) = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})} \tag{171}$$

$$= \frac{L(\bar{x})}{L(\bar{x})}\mathbb{1}_{\bar{x}\in\Theta_0} + \frac{L(\theta_0)}{L(\bar{x})}\mathbb{1}_{\bar{x}\notin\Theta_0} \tag{172}$$

$$= \mathbb{1}_{\bar{x}\in\Theta_0} + \mathbb{1}_{\bar{x}\notin\Theta_0}\left(\frac{\theta_0}{\bar{x}}\right)^{n\bar{x}}\left(\frac{1-\theta_0}{1-\bar{x}}\right)^{n(1-\bar{x})} \tag{173}$$

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

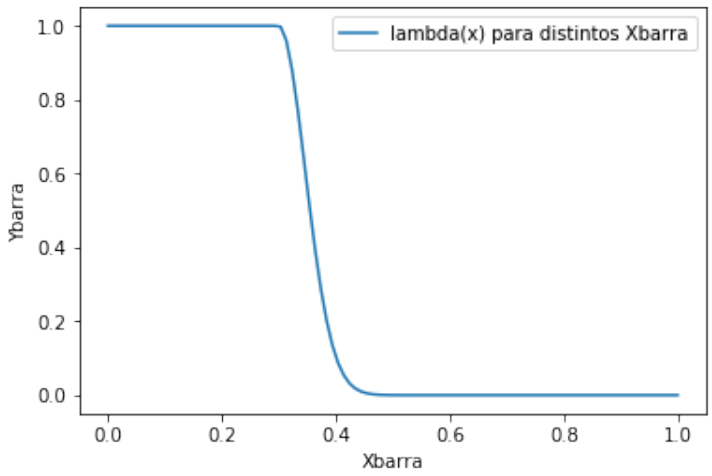

Figura 3. $\lambda(x)$ en función de $\bar{x}$

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

$$\sup_{\theta\in\Theta_0} \mathbb{P}_\theta\left(\lambda(x) \leq C\right) = \alpha \tag{174}$$

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 6.7: 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 respectivamente.

7. 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:

$$H_0 : F = F_0 \quad \text{v.s.} \quad H_1 : F \neq F_0. \tag{175}$$

Consideremos además la distribución empírica dada por

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{x_j \leq x}, \tag{176}$$

la cual realmente es una distribución (discontinua y constante por tramos).

Sabemos que, debido a la ley de los grandes números,

$$F_n(x) \to \mathbb{E}\left(\mathbb{1}_{X \leq x}\right) = \mathbb{P}\left(X \leq x\right) = F(x), \tag{177}$$

además, por el teorema de Glivenko-Cantelli, tenemos

$$\sup_x |F_n(x) - F(x)| \to 0 \quad \text{c.s.} \tag{178}$$

Lo anterior nos permite definir el estadístico $D_n = \sup_x |F_n(x) - F_0(x)|$ y la región crítica

$$R = \{x | D_n \geq k_\alpha\}, \tag{179}$$

donde k_α se elige imponiendo $\mathbb{P}_{\theta_0}\left(D_n \geq k_\alpha\right) = \alpha$.

Observación 6.8: 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).

8. Test de Wilcoxon

Este es otro testo no paramétrico para verificar si dos VAs siguen la misma distribución. Consideremos las observaciones

$$X_1, \dots, X_n \sim F, \quad Y_1, \dots, Y_n \sim G, \tag{180}$$

donde F y G son dos distribuciones, de las cuales solo sabemos que son continuas.

Consideremos el siguiente problema de test de hipótesis:

$$H_0 : F = G \quad \text{v.s.} \quad H_1 : F \neq G. \tag{181}$$

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

$$H_0 : \mu = \eta \quad \text{v.s.} \quad H_1 : \mu \neq \eta, \tag{182}$$

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

$$(z_1, \dots, z_{m+n}) = (x_1, \dots, x_n, y_1, \dots, y_m), \tag{183}$$

para luego considerar la secuencia ordenada de valores z_i dados por

$$\min_i\{z_i\} = z_{(1)} \leq z_{(2)} \leq \cdots \leq z_{(n+m)} = \max_i\{z_i\}. \tag{184}$$

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

$$W = \sum_{i=1}^{n} R_i, \tag{185}$$

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 6.9 (¿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 $\mathbb{P}_{\theta_0}(W \in \mathbb{R}) = \alpha$.

9. 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 6.7: 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**.

Capítulo 7

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 7.1 (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 $F_{X,Y}$.

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 7.2: 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 7.1: 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:

$$\hat{\sigma^2} = \frac{1}{n-2}J(\hat{\beta_0}, \hat{\beta_1})$$

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 7.2: 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 7.1: El estimador anterior de $\hat{\sigma}^2$ normalmente se reemplaza por el estimador insesgado de la parte anterior.

Lo anterior se puede extender al caso en que $X \in R^k$ de la siguiente forma. Suponemos:

$$Y_i = \sum_{j=1}^{k} \beta_j X_i j + \varepsilon_i,$$

con $\mathbb{E}(\varepsilon_i) = 0$. Denotemos:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ \cdot \\ Y_n \end{pmatrix} ; X = \begin{pmatrix} X_1 \\ X_2 \\ \cdot \\ \cdot \\ \cdot \\ X_n \end{pmatrix}$$

Cada fila de la matrix en $\mathbb{R}^{n\times k}$ X es una observación. Sean:

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \cdot \\ \cdot \\ \cdot \\ \beta_n \end{pmatrix} \text{ y } \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{pmatrix}$$

Entonces:

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

Teorema 7.3: Si la matriz $X^T X$ es invertible:

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

Observación 7.2: 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.

3. Regresión Logística

Definición 7.3: Se llama función logística a la función:

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

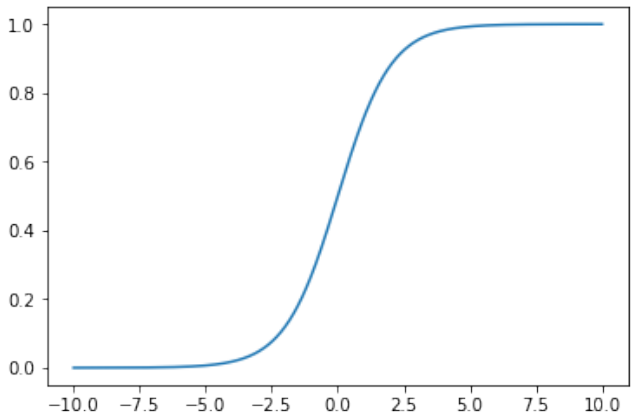

Figura 1. Función Logística

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.

Capítulo 8

Introducción a Series de Tiempo

Las series de tiempo son un tipo de datos, que están indexados por el tiempo. Si bien esto puede parecer natural, es importante notar que a este punto, normalmente hemos trabado con datos que podemos re-indexar sin ningún problema. Las series de tiempo, al estar indexadas por el tiempo le dan un orden a los datos. Explicado de otra forma, no podemos intercambiar índices de forma aleatoria y modelar los datos con la misma distribución.
Una propiedad bastante particular de las series de tiempo es que los datos çrudos.ªportan muy poca información. Como consecuencia, graficar o resumir los datos no aporta mucho al análisis de las series de tiempo.
La literatura sobre series de tiempo es bastante extensa, por lo que nos concentraremos sólo en conceptos básicos de estas.
Sea $(X_t)_{t\geq 0}$ una serie de tiempo. El comportamiento estocástico está determinado por las densidades:

$$p(X_{t_1}, X_{t_2}, ..., X_{t_m}),\ m \in \mathbb{N}$$

(

Definición 8.1: Una serie de tiempo $(X_t)_t \geq 0$ se dice estrictamente estacionaria si cumple invarianza bajo traslaciones temporales:

$$p(X_{t_1+\tau}, X_{t_2+\tau}, ..., X_{t_m}+\tau) = p(X_{t_1}, X_{t_2}, ..., X_{t_m}) \forall \tau, \forall m, \forall \{t_1, ..., t_m\}$$

En otras palabras, desde un punto de vista distribucional, una serie de tiempo es invariante bajo shifts. Dado que la definición es para todo m, incluyendo $m = 1$, la media y la varianza para estos procesos son constantes.
A veces, esta propiedad resulta ser muy fuerte. Por ello podemos "debilitarlaün poco con la siguiente definición:

Definición 8.2: Una serie de tiempo es estacionaria de segundo orden si su media es constante y su covarianza entre dos valores de tiempo sólo depende de

la diferencia de estos valores. Es decir:

$$\mathbb{E}(X_t) = \mu \forall t$$
$$Cov(X_t, X_{t+\tau} = \gamma(\tau) \forall \tau$$

La función $\gamma(\tau)$ se conoce como función de auto-covarianza.

La intuición detrás de las series de tiempo estacionarias es que la distribución de los datos no depende de t, por lo que el conocimiento que se tenga sobre el tiempo no nos dirá nada sobre la distribución. Esto nos permite considerar que las series de tiempo son estables en tiempo, por lo que no habrá mayores cambios en las tendencias en el tiempo.
Es importante mencionar que no todas las series de tiempo son estacionarias. Un buen ejemplo de esto es el clima. Imaginemos que tenemos una serie de tiempo con la temperatura en Santiago. Si la miramos en invierno, las temperatura serán considerablemente más bajas que 6 meses después en verano, por lo que el tiempo si cambia la distribución de los datos.

Definición 8.3 (Operador Lag)**:** El operador Lag $L()$ se define como el operador que hace un shift en un incremento de tiempo:

$$L(X_t) = X_{t-1}$$

Aplicándolo de forma recursiva:

$$L^0(X_t) = X_t; L(X_t) = X_{t-1}; L^2(X_t) = X_{t-2}; ...; L^n(X_t) = X_{t-n}$$

La inversa de estos operadores está bien definida:

$$L^{-n}(X_t) = X_{t+n}$$

1. Modelos AR

Definición 8.4: Una serie de tiempo $(X_t)_{t\geq 0}$ sigue un modelo autorregresivo de orden p si:

$$X_t = \mu + \phi_1(X_{t-1} - \mu) + ... + \phi_p(X_{t-p} - \mu) + \epsilon_t$$

con $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$. Si definimos:

$$\phi(L) = (1 - \phi_1 L + ... - \phi_p L^p),$$

con L el operador Lag, podemos caracterizar los modelos autorregresivos por:

$$\phi(L) \cdot (X_t - \mu) = \epsilon_t$$

Consideremos $\phi(z)$, reemplazando L por una variable compleja, y sean $\lambda_1, ..., \lambda_p$ las raíces de la ecuación $\phi(z) = 0$. Entonces:

$$\phi(Z) = (1 - \frac{1}{\lambda_1}L) \cdots (1 - \frac{1}{\lambda_p}L)$$

La ecuación $\phi(z) = 0$ se llamará la ecuación característica. El siguiente lema no será demostrado pues requiere mayor profundidad en series de tiempo.

Lema 8.1: Una serie de tiempo $(X_t)_{t\geq 0}$ que sigue un modelo AR es estacionaria de segundo orden si y sólo si todas las raíces de su ecuación característica están fuera del círculo unitario. Es decir, $|\lambda_j| > 1 \forall j \in \{1, ..., p\}$.

Ejemplo 8.1: Sea $(X_t)_{t\geq 0}$ una serie de tiempo que sigue un modelo autorregresivo de orden 1, es decir:

$$X_t = c + \phi X_{t-1} + \epsilon_t \forall t$$

con $\epsilon_t \sim \mathcal{N}(0, \sigma^2) \forall t$. Definimos $\theta = [c, \phi, \sigma^2]$.

La ecuación característica del modelo es:

$$(1 - \phi z) = 0,$$

con raíz $\lambda = 1/\phi$. Luego el modelo AR(1) es estacionario de segundo orden si $|\phi| < 1$. Además:

$$\mathbb{E}(X_t) = \mu$$
$$\mathbb{V}(X_t) = \frac{\sigma^2}{1 - \phi}$$

2. Estimación en modelos AR

En un modelo de series de tiempo AR, la construcción de la densidad conjunta de observaciones $y_1, ..y_n$ no se puede calcular de la forma usual, al igual que su log-verosimilitud, pues estas no son observaciones i.i.d.

En estos casos, el enfoque más usando consiste en factor izar la densidad conjunta en una serie de densidades condicionales y la densidad de un conjunto de valores iniciales. Para ver esto, consideremos la densidad conjunta de dos observaciones y_1 e y_2 de una serie de tiempo AR. Tenemos:

$$f(X_1; X_2; \theta) = f(X_2|X_1, \theta) f(X_1; \theta)$$

Luego para tres observaciones:

$$f(X_1; X_2; X_3; \theta) = f(X_3|X_2; X_1; \theta) f(X_2|X_1, \theta) f(X_1; \theta)$$

De forma inductiva:

$$f(X_T;...;X_1;\theta) = (\prod_{t=p+1}^{T} f(X_t|I_{t-1};\theta))f(X_p;..;X_1;\theta)$$

donde $I_t = \{X_t;...;X_1\}$ denota la información a tiempo t y $X_1,..,X_p$ los valores iniciales. Así, la función de log-verosimilitud es:

$$l(\theta|X) = \sum_{p+1}^{T} ln(f(X_t|I_{t-1})) + ln(f(X_p;..;X_1;\theta))$$

Ejemplo 8.2 (Estimación en el modelo AR(1))**:** Consideremos el modelo estacionario de segundo orden:

$$X_t = c + \phi X_{t-1} + \epsilon_t \forall t$$

con $\epsilon_t \sim \mathcal{N}(0,\sigma^2)\forall t$. Definimos $\theta = [c,\phi,\sigma^2]$ Luego:

$$X_t|I_{t-1} \sim \mathcal{N}(c + \phi X_{t-1};\sigma^2)$$

Entonces, condicional a X_{t-1}:

$$f(X_t|X_{t-1};\theta) = (2\pi\sigma^2)^{\frac{1}{2}} exp(\frac{-1}{2\sigma^2}(X_t - c - \phi X_{t-1})^2); t = 1,...,T$$

Como la serie es estacionaria:

$$\mathbb{E}(X_1) = \mu = \frac{c}{1-\phi}$$

$$\mathbb{V}(X_1) = \frac{\sigma^2}{1-\phi^2}$$

Con esto:

$$X_1 \sim \mathcal{N}(\frac{c}{1-\phi};\frac{\sigma^2}{1-\phi^2})$$

$$f(X_1;\theta) = (2\pi\frac{\sigma^2}{1-\phi^2})^{\frac{1}{2}} exp(-\frac{1-\phi^2}{2\sigma^2}(X_t - \frac{c}{1-\phi})^2)$$

Teniendo esta distribución, podemos obtener la log-verosimilitud, usando la factorización en densidades condicionales, obteniendo:

$$logL(X_t|\theta) = \frac{log(1-\phi^2)}{2} - \frac{Tlog(2\pi\sigma^2)}{2} + \frac{(\phi^2-1)(X_1 - \frac{c}{1-\phi})^2}{2\sigma^2} + \sum_{i=2}^{T} \frac{-(X_t - c - \varphi X_{t-1})^2}{2\sigma^2}$$

Usando métodos numéricos, podemos obtener una expresión para el estimador de máxima verosimilitud.

Notemos que, dado conocimiento experto, también podemos encontrar el estimador máximo a posteriori.

Capítulo 9

Markov Chain Monte Carlo

Como hemos visto en los capítulos anteriores, con frecuencia el enfoque bayesiano implica calcular integrales que no son calculables con métodos convencionales, como el denominador cuando ocupamos Bayes, la marginalización de variables y el cálculo de esperanzas. También es necesario optimizar funciones que es muy difícil optimizar explícitamente, por ejemplo, al momento de maximizar la distribución posterior de un parámetro.
Dados estos problemas, se hace natural buscar herramientas que nos permitan aproximar estas cantidades usando métodos numéricos, y con frecuencia, un computador. La herramienta que estudiaremos en esta sección, juega un rol fundamental en la integración, optimización, y también en la simulación de fenómenos físicos.

1. El principio de Monte Carlo

La idea detrás de las simulaciones de Monte Carlo es extraer una muestra de observaciones i.i.d, $\mathcal{D} = \{x_i\}_{i=1}^N$ de una densidad objetivo $p(x)$ desconocida. Las N observaciones pueden usarse para aproximar la densidad objetivo de la forma:

$$p_N(x) = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}(x)$$

con $\delta_{x_i}(x)$ denota la delta de Dirac centrada en x_i. De esta forma, si buscamos aproximar la esperanza de f, $I(f) = \int f(x)p(x)$, lo haremos mediante las sumas:

$$I_N(f) = \frac{1}{N}\sum_{i=1}^{N} f(x_i) \underset{N\to\infty}{\overset{c.s}{\to}} I(f) = \int f(x)p(x)$$

La convergencia anterior se cumple por la Ley de los Grandes Números Fuerte, e implica que $I_N(f)$ es un estimador insesgado.

2. Sampling

3. Algoritmos de Montecarlo

Capítulo 10

Inferencia Causal

En este capítulo nos enfocaremos en métodos matemáticos que nos permitan modelar causas. Los métodos aquí presentados son bastante recientes, e intentan responder la pregunta *"¿Cuándo X causa Y?"*.
Estas dudas surgen de forma natural en un curso de estadística. Siempre se enseñan herramientas que permiten ver cuándo dos variables están asociadas, por ejemplo, con su correlación. Sin embargo, siempre se dice que *.Asociación no implica causa"*. Es así como la estadística clásica respondió muchas veces la pregunta "¿Qué **no** es X causa Y?", pero nunca "¿Qué es X causa Y?". Esto fue así hasta que Judea Pearl introdujo la inferencia causal a finales del sigo pasado. Este trabajo logró que le dieran a Pearl un premio Turing, considerado el premio Nobel de ciencias de la computación, en el año 2011.
Informalmente, diremos que X causa Y si un cambio en el valor de X cambia la distribución de Y. Podemos notar que cuando X causa Y, X e Y están asociadas, pero la recíproca no es cierta.

1. El modelo contrafactual

Sea X una variable binaria, donde $X = 1$ si X "fue expuestaz $X = 0$, si X "no fue expuesta". En este ámbito, la palabra .expuesta"puede referirse, por ejemplo, a que X se expuso a un tratamiento médico, o a que X realizó determinada acción. Por otro lado, sea Y la variable resultado, por ejemplo, si hay o no una enfermedad.

Definición 10.1: Introducimos las variables aleatorias C_0 y C_1, llamadas resultados potenciales como:

- C_0 es el resultado si $X = 0$
- C_1 es el resultado si $X = 1$

Así:

$$Y = \begin{cases} C_0, & \text{Si X =0} \\ C_1, & \text{Si } X = 1 \end{cases}$$

Esto se puede expresar como:

$$Y = C_X$$

Lo anterior se llama la **Ecuación de Consistencia**.

Observación 10.1:
1. Podemos pensar en C_0 y C_1 como variables escondidas que tienen toda la información relevante de un sujeto.
2. Si $X = 0$, no observamos C_1. Luego decimos que C_1 es un contrafactual de $X = 0$, pues es el resultado que hubiésemos obtenido si, contra los hechos (*counter the facts*), $X = 1$.

Definición 10.2: Definimos el efecto causal promedio de X sobre Y por:

$$\theta = \mathbb{E}(C_1) - \mathbb{E}(C_0)$$

Observación 10.2:
- θ es el promedio si $X = 1$ para todos, menos el promedio si $X = 0$ para todos los sujetos.
- θ mide el efecto causal de X.

Definición 10.3: Se define la asociación de X e Y por:

$$\alpha = \mathbb{E}(Y|X = 1) - \mathbb{E}(Y|X = 0)$$

Teorema 10.1: En general, $\theta \neq \alpha$. (Asociación no implica causa).

Ejemplo 10.1: Queremos analizar el efecto que tiene una vitamina sobre una condición. Luego:

$$X = \begin{cases} 1, & \text{Si la persona toma la vitamina} \\ 0, & \text{Si no} \end{cases}$$

Por otra parte:

$$Y = \begin{cases} 1, & \text{Si la persona está saludable} \\ 0, & \text{Si no} \end{cases}$$

Observamos:

X	Y	C_0	C_1
0	0	0	0*
0	0	0	0*
0	0	0	0*
0	0	0	0*
1	1	1*	1
1	1	1*	1
1	1	1*	1
1	1	1*	1

Los asteriscos hacen referencia a cantidades no observadas. Como $C_0 = C_1$ para cada sujeto, tendremos:

$$\theta = \mathbb{E}(C_1) - \mathbb{E}(C_0) = \frac{1}{8}\sum_{i=1}^{8} C_{1,i} - \frac{1}{8}\sum_{i=1}^{8} C_{0,i} = 0$$

Luego el efecto causal promedio es 0. Sin embargo,

$$\alpha = 1$$

Luego $\theta \neq \alpha$. Sólo podemos concluir que la gente sana ($Y = 1$) tiende a tomar la vitamina, mientras que la gente no sana, no.

En la mayoría de los casos, se hace difícil estimar θ. Luego, se hace natural preguntarse cuando es posible dar un estimador para θ. La respuesta es, cuando se hace una asignación aleatoria al tratamiento:

Teorema 10.2: Supongamos que asignamos el tratamiento al azar a los sujetos, y que $\mathbb{P}(X = 0) > 0$ y $\mathbb{P}(X = 1) > 0$. Entonces $\alpha = \theta$, y por lo tanto cualquier estimador consistente de α es un estimador consistente de θ. En particular:

$$\hat{\theta} = \hat{\mathbb{E}}(Y|X = 1) - \hat{\mathbb{E}}(Y|X = 0) = \frac{1}{n_1}\sum_{i=1}^{n} Y_i X_i - \frac{1}{n_0}\sum_{i=1}^{n} Y_i(1 - X_i),$$

con $n_1 = \sum_{i=1}^{n} X_i$ y $n_2 = \sum_{i=1}^{n} 1 - X_i$.

Dem: Como el tratamiento es asignado de forma aleatoria, tenemos que:

$$X \perp\!\!\!\perp C_0, C_1$$

Con esto:

$\theta = \mathbb{E}(C_1) - \mathbb{E}(C_0) = \mathbb{E}(C_1|X = 1) - \mathbb{E}(C_0|X = 0) = \mathbb{E}(Y|X = 1) - \mathbb{E}(Y|X = 0) = \alpha$■

2. El modelo contrafactual: Generalización

En la sección anterior se estudió que sucedía en el caso binario, es decir, el sujeto podía estar expuesto o no expuesto. Sin embargo, esta es una sobre simplificación de la realidad. Por ejemplo, si se desea estudiar el efecto de un medicamento sobre una enfermedad, no sólo será importante para el modelo si un sujeto se expuso o no a un medicamento, sino también cuál fue la dosis que tomaron.
Sea $x \in \mathcal{X}$. El vector (C_0, C_1) pasará a ser la *función contrafactual* $C(x)$. Luego, en el ejemplo del medicamento x será la dosis que tomó un sujeto, y $C(x)$ será el resultado que habría tenido un sujeto si hubiese recibido una dosis x.
Con lo anterior, la ecuación de consistencia se transforma en:

$$Y = C(X)$$

Por otro lado, cambiamos el efecto causal promedio de X sobre Y por la función de regresión causal:

$$\theta(x) = \mathbb{E}(C(X)).$$

Por otra parte la asociación se mide por la función de regresión:

$$r(x) = \mathbb{E}(Y|X = x)$$

Teorema 10.3: En general, $\theta(x) \neq r(x)$. Sin embargo, si X es asignado al azar (por ejemplo, por una prueba controlada aleatorizada), entonces $\theta(x) = r(x)$.

Lamentablemente, dadas las dificultades que podría significar, la exposición de X no suele ser al azar. Luego, ¿Cómo separamos aquello que podemos controlar para estudiar causalidad, y aquello que no?

3. Confounders

El problema que presenta que la exposición de X no sea aleatoria es que genera que $C(X)$ no sea independiente de X. Pero, ¿Qué pasaría si pudiésemos separar los sujetos en grupos de forma que X y $C(X)$ fueran independientes dentro de los grupos?.
Si lo pensamos, lo anterior no es muy difícil. Por ejemplo si estudiamos el efecto de un tratamiento médico, podríamos separar a las personas en distintos grupos según su edad, género, hábitos, etc. Dentro de un mismo grupo, se hace razonable que X y $C(X)$ sean independientes. Las variables que usamos para asignar los distintos grupos se llaman **confounders** o **factores de confusión**. Si denotamos

por Z a todas estas variables, entonces:

$$\{C(x) : x \in \mathcal{X}\} \perp\!\!\!\perp X|Z.$$

Teorema 10.4: Si $\{C(x) : x \in \mathcal{X}\} \perp\!\!\!\perp X|Z$, entonces:

$$\theta(x) = \int \mathbb{E}(Y|X = x, Z = z)dF_Z(z)dz \tag{186}$$

Si $\hat{r}(x, z)$ es un estimador consistente de la funcióon de regresión $\mathbb{E}(Y|X = x, Z = z)$, entonces un estimador consistent de $\theta(x)$ es:

$$\hat{\theta}(x) = \frac{1}{n}\sum_{i=1}^{n} \hat{r}(x, Z_i) \tag{187}$$

Observación 10.3:

1. Si $r(x, z) = \beta_0 + \beta_1 x + \beta_2 z$ entonces:
$$\hat{\theta}(x) = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 z,$$
donde $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ son los estimadores de MCO.
2. Los epidemiólogos suelen llamar a la expresión de la ecuación 187 efecto ajustado de tratamiento.
3. La selección de factores de confusión **requiere** conocimiento experto. No es posible asegurarse de que no haya confounders que no conozcamos.

Esta última observación tiene un efecto muy potente en la "filosofía"de la inteligencia artificial y en particular en el aprendizaje de máquinas según Judea Pearl. No es posible hacer que una máquina entienda relaciones causales sin un modelo con conocimiento humano experto. Pasar a ese nivel de ïnteligenciarequiere que un humano haya modelado el fenómeno estudiado.

4. DAGs

Definición 10.4: Sean X, Y, Z variables aleatorias. X e Y se dicen condicionalmente independientes dado Z si:

$$f_{X,Y|Z}(x, y|z) = f_{X|Z}(x|z)f_{Y|Z}(y|z)\forall x, y, z$$

Esta relación se denota $X \perp\!\!\!\perp Y|Z$.

Definición 10.5: Un grafo dirigido G es un par (V, E), donde V corresponde a los vértices, y E a pares ordenados de vértices.

Si se considera cada vértice en un grafo dirigido como una variable aleatoria, entonces esta se convierte en una forma muy conveniente de indicar independencia de las distintas variables. Para comprender este método, se hacen necesarias algunas definiciones y conceptos de Teoría de Grafos.

Definición 10.6:

- Si $\exists e \in E$ tal que $e = (X, Y)$, X e Y se dicen adyacentes.
- Si $\exists e \in E$ tal que $e = (X, Y)$, es decir el grafo tiene la configuración $X \to Y$, se dice que X es padre de Y, y que Y es hijo de X. Se denota por π_X o $\pi(X)$ al conjunto de padres de X.
- Un camino dirigido entre dos variables aleatorias es un conjunto de flechas apuntando en la misma dirección que va dese una variable a la otra.
- X es ancestro de Y si existe un X, Y-camino dirigido. En este caso, Y se dice descendiente de X.

Esto nos permite graficar distintos tipos de relaciones causales e independencia entre variables.

Insertar Dibujo

Definición 10.7: Una configuración de la forma $X \to Y \leftarrow Z$ se llama **collider**. Cualquier otra configuración se llama **no-collider**.

Un camino dirigido se dice ciclo si empieza hay termina en el mismo vértice. Un grafo dirigido se dice acíclico si no contiene ciclos. De ahora en adelante, sólo trabajaremos con grafos dirigidos acíclicos o **DAGs** por su sigla en inglés (Directed Acyclic Graph).

Si bien hoy en día el uso de DAGs como herramienta es bastante más común que hace un par de décadas, se debe tener en cuenta que fue muy difícil que la escuela clásica de la estadística lograra aceptarlos como una herramienta útil. Es más, no fueron los matemáticos ni los estadísticos los qu comenzaron usando esta herramienta, sino los epidemiólogos.

Definición 10.8: Sea $G = (V, E)$ un DAG, y sea $\mathbb{P}$ una distribución para V con función de probabilidad f. Se dice que G representa a $\mathbb{P}$ si:

$$f(v) = \prod_{i=1}^{k} f(x_i|\pi_i)$$

con π_i padres de x_i. Se denota por $M(G)$ al conjunto de distribuciones representadas por G.

Ejemplo 10.2: **[insertar dibujo]** Sobrepeso $\rightarrow$ Enfermedades al Corazón $\leftarrow$ Fumar $\rightarrow$ Tos

Tenemos que las enfermedades al corazón son un collider en el camino Sobrepeso $\rightarrow$ Enfermedades al Corazón $\leftarrow$ Fumar. Además:

$$f(\text{Sobrepeso}, \text{Enfermedades al Corazón}, \text{Fumar}, \text{Tos}) =$$

$$f(\text{Sobrepeso})f(\text{Fumar})f(\text{Enfermedades al Corazón}|\text{Sobrepeso}, \text{Fumar})f(\text{Tos}|\text{Fumar})$$

Teorema 10.5: Dado $G = (V, E)$ un DAG, $\mathbb{P} \in M(G)$ si y sólo si $\forall W$ variable aleatoria, $W \perp\!\!\!\perp \widetilde{W}|\pi_W$, con $\widetilde{W}$ todas las variables excepto padres y descendientes de W.

5. d-Separación

Capítulo 11

Apéndice

1. Convergencia de Variables Aleatorias

Sea $(X_n)_{n\in\mathbb{N}}$ una sucesión de variables aleatorias con $F_{X_n}(x)$ su función de distribución. Sea además X variable aleatoria con distribución $F_X(x)$.

Definición 11.1 (Convergencia)**:** Diremos que X_n converge en **distribución** a X ssi

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x) \Leftrightarrow \lim_{n\to\infty} Pr(X_n \leq x) = Pr(X \leq x)$$

Diremos además que X_n converge en **probabilidad** a X ssi para todo $\epsilon > 0$

$$\lim_{n\to\infty} Pr(|X_n - X| > \epsilon) = 0$$

Finalmente, X_n converge **casi seguramente** a X ssi

$$Pr(\lim_{n\to\infty} X_n = X) = 1$$

Con estas definiciones en mente, enunciaremos a continuación dos teoremas fundamentales para la justificación de muchos resultados del curso de Estadística

2. Ley de los Grandes Números (LGN)

Consideremos un ensayo de Bernoulli como lo podría ser el lanzamiento de una moneda cargada con probabilidad p de obtener cara y $1-p$ de obtener sello, supongamos que no conocemos el valor de p pero nos gustaría estimarlo lo más precisamente posible. Un acercamiento natural al problema sería realizar n lanzamientos y promediar los resultados con el fin de recuperar el parámetro p,

además, la intuición nos dice que entre más lanzamientos realicemos, más cercana debería ser nuestra estimación de p al valor real.

La ley de los grandes números, se encarga de describir el comportamiento del promedio de una sucesión de variables aleatorias a medida que aumentamos la cantidad de repeticiones del evento. Veamos las dos formas del teorema:

Teorema 11.1 (Ley débil de los grandes números)**:** Sea $(X_1, \dots X_n)$ muestra aleatoria i.i.d provenientes de una distribución X con valor esperado μ y varianza σ^2 finita. Entonces el promedio definido como

$$\overline{X_n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

converge en probabilidad a μ, es decir,

$$\overline{X_n} \xrightarrow[n]{P} \mathbb{E}(X) = \mu$$

Teorema 11.2 (Ley fuerte de los grandes números)**:** Sea $(X_1, \dots X_n)$ muestra aleatoria i.i.d provenientes de una distribución X con valor esperado μ y varianza σ^2 finita. Entonces

$$\overline{X_n} \xrightarrow[n]{c.s.} \mathbb{E}(X)$$

Es decir, el promedio muestral converge casi seguramente a la media de la distribución X

Observación 11.1: Esta última definición, nos indica que si realizamos un evento una cantidad indefinida de veces y esperamos lo suficiente, eventualmente el promedio converge a la media de la distribución salvo que alguna de las muestras se tomen de un evento con probabilidad 0. Notar además que la ley fuerte implica la débil pero no al revés.

Ejemplo 11.1: De nuestro ejemplo del ensayo de Bernoulli, utilizando la ley de los grandes números, obtendríamos que nuestro promedio eventualmente converge a la esperanza de la variable aleatoria Bernoulli que es justamente p.

3. Teorema Central del Límite (TCL)

Otro teorema fundamental de convergencia es el llamado teorema central del límite y es la herramienta que nos permite justificar la elección de la distribución normal como un modelo de aproximación de muchas variables aleatorias. Veamos una vez más el ensayo Bernoulli de nuestra moneda, supongamos la lanzamos una cantidad n suficientemente grande, el teorema central del límite nos

permite aproximar el promedio de los lanzamientos como una variable aleatoria normal con media igual al valor esperado de nuestro ensayo y con varianza igual a la varianza de una distribución Bernoulli sobre el número n de lanzamientos. Veamos la definición formal del teorema

Teorema 11.3 (Teorema central del límite (TCL))**:** Sea $(X_1 \dots X_n)$ variables aleatorias i.i.d con media μ y varianza σ^2 ambas finitas. Definamos $\overline{X_n}$ como el promedio de estas n variables aleatorias y definamos

$$Z_n = \frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}}$$

Entonces Z_n converge en **distribución** a $\mathcal{N}(0,1)$, es decir

$$Z_n \xrightarrow[n]{D} Z \sim \mathcal{N}(0,1)$$

Observación 11.2: Usando propiedades de la distribución normal, lo anterior es equivalente a decir que para n suficientemente grande

$$\overline{X_n} \sim \mathcal{N}\left(\mathbb{E}(X), \frac{\mathbb{V}ar(X)}{n}\right)$$

Observación 11.3: Existe otra versión del TCL más general en que no se pide que las variables aleatorias sean identicamente distribuidas pero si independientes y se le conoce como Lyapunov TCL

Ejemplo 11.2: Siguiendo el teorema y nuestro ejemplo anterior, podemos concluir que para el caso de n lanzamientos de una variable aleatoria Bernoulli de media p y varianza $p(1-p)$ que

$$\overline{X_n} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$