Estadística Bayesiana

J. Andrés Christen

Centro de Investigación en Matemáticas (CIMAT)
Guanajuato, Mexico.
email: *jac at cimat dot mx*, web page: `http://www.cimat.mx/~jac`

CIMAT: Curso Maestría/Doctorado

Curso de **Estadística Bayesiana** CIMAT

Características generales del curso:

- Se estudiaran conceptos básicos de Estadística Bayesiana, dentro de cuatro temas principales, 1) Fundamentación Axiomática, 2) Teoría de decisiones, 3) Inferencia y 4) Métodos computacionales.
- El alumno al final del curso deberá de ser capaz de 1) Entender (hasta cierto punto) un artículo de investigación donde se use un método Bayesiano, 2) Leer la bibliografía respectiva, 3) Continuar su formación para desarrollar temas de investigación (para estudios de maestría o doctorado) en el área de Estadística Bayesiana.

Índice

- **Profesor:** Dr. J. Andrés Christen. e-mail: *jac@cimat.mx*.
- **Nivel:** Maestría y Doctorado.

Bibliografía Básica

Textos

Aun cuando no se sigue ningún texto, los que recomendamos son:

1. J. O. Berger (1985), “Statistical Decision Theory: foundations, concepts and methods”, Second Edition, Springer-Verlag.
2. Bernardo, J. M. y Smith, A. F. M. (1994), “Bayesian Theory”, Wiley: Chichester, UK.
3. M. H. DeGroot (1970), “Optimal statistical decisions”, McGraw–Hill: NY.
4. Para la sección de MCMC se usó C.P. Robert y G. Casella (1999), “Monte Carlo Statistical Methods”, Springer: NY.

Introducción: Ejemplos (3)

"El mundo de los humanos viene determinado por tres elementos: tiempo, espacio y probabilidad." H. Murakami, *La muerte del comendador*, 2019, Tusquets, p. 236.

Lo que sabemos de Bayesiana:
La incertidumbre es cuantificada con una medida de probabilidad

Teorema de Bayes: Modificar la probabilidad con evidencia

ie. Condicionar a los datos.

Tenemos entonces que toda probabilidad es condicional (a las circunstancias, el *agente* que habla etc.) y en realidad deberíamos escribir

$$P(\cdot \mid H), \quad \text{con } H = \text{circunstancias, } \textit{agente} \text{ que habla etc..}$$

Sea $B \in @$ un evento observable, ¿qué es la probabilidad de $A \in @$ siendo que observamos a B?
En realidad estamos hablando del evento $A \mid H, B$, dado H el contexto etc. Por definición esto lo podemos calcular como

$$P(A \mid H, B) = \frac{P(A \cap B \mid H)}{P(B \mid H)}.$$

Ya que también por definición (probabilidad condicional) $P(B \mid H, A) = \frac{P(A \cap B|H)}{P(A|H)}$ entonces

$$P(A \cap B \mid H) = P(B \mid H, A)P(A \mid H).$$

Por lo tanto

$$P(A \mid H, B) = \frac{P(B \mid H, A)P(A \mid H)}{P(B \mid H)} \quad \textbf{¡Teorema de Bayes!}.$$

Si tenemos $A_1, A_2, \ldots, A_n$, con $\Omega = \cup_{i=1}^{n} A_i$ y $A_i \cap Aj =$ (una partición de eventos), entonces, por probabilidad total:

$$P(A_j \mid H, B) = \frac{P(B \mid H, A_j)P(A_j \mid H)}{\sum_{i=1}^{n} P(B \mid H, A_i)P(A_i \mid H)}.$$

Analicemos

$$P(A \mid H, B) = \frac{P(B \mid H, A)P(A \mid H)}{P(B \mid H)}.$$

- $P(A \mid H)$ la llamamos probabilidad *a priori* o inicial ("prior", en Inglés) para A.
- $P(A \mid H, B)$ la llamamos probabilidad *a posteriori* o posterior para A, dado que observamos B.
- $P(B \mid H, A)$ es el "modelo" (observacional)...¿como sería la probabilidad de B si supiéramos A? (y que en inferencia paramétrica es la verosimilitud).
- $P(B \mid H)$ es la constante de normalización, entendiendo a $P(\cdot \mid H, B)$ como una nueva medida y que

$$P(\cdot \mid H, B) \propto P(B \mid H, \cdot)P(\cdot \mid H).$$

Normalmente omitimos condicionar con respecto a H y preferimos dejarlo implícito.

Ejemplo 1

Supóngase que se tiene una prueba o test de sangre para detectar cierta infermedad. El sujeto puede tener o no tener la enfermedad $E = 0, 1$, y el test puede resultar negativo o positivo $T = 0, 1$.

Las características del test son:

$P(T = 1 \mid E = 1) =$		$0.90^{(1)}$	$P(T = 1 \mid E = 0) =$		0.05
$P(T = 0 \mid E = 1) =$		0.10	$P(T = 0 \mid E = 0) =$		$0.95^{(2)}$
	$+$	1.00		$+$	1.00

(1) Sensibilidad. (2) Especificidad.

Yo acudo a hacerme el test, y este sale positivo $T = 1$, ¿qué me dice esto sobre si estoy enfermo o no?

Ejemplo 1

$P(T = 1 \mid E = 1)$ ó $P(T = 1 \mid E = 0)$ ***no*** es lo que necesitamos.

Más bien queremos saber qué sucede *dado* o *una vez que* $T = 1$ (el test me salió positivo): ¿cual es la probabilidad de que yo esté enfermo dado que el test salió positivo?

O sea: $P(E = 1 \mid T = 1) = 1 - P(E = 0 \mid T = 1)$, y (Teorema de Bayes):

$$P(E = 1|T = 1) = \frac{P(T = 1|E = 1)P(E = 1)}{P(T = 1|E = 1)P(E = 1) + P(T = 1|E = 0)P(E = 0)}.$$

¿Qué es $P(E = 1) = 1 - P(E = 0)$?

Ejemplo 1

Las carsacterísticas del test son:

$P(T=1 \mid E=1) = 0.90$	$P(T=1 \mid E=0) = 0.05$
$P(T=0 \mid E=1) = 0.10$	$P(T=0 \mid E=0) = 0.95$
$+ \quad 1.00$	$+ \quad 1.00$

Para calcular la posterior hacemos:

E	0	1	Suma
$P(E)$	$1 - 10^{-6}$	10^{-6}	1.0
$P(T=1 \mid E)$	0.05	0.90	0.95
Prod.	0.04999995	8.9×10^{-6}	0.05000085
$P(E \mid T=1)$	0.999982	0.00001799	1.0

¿Qué podemos concluir?

Ejemplo 2

Supóngase que una particular población de células puede estar en uno de los siguientes tres estados de producción de una cierta proteina. Los estados son A, B y C, de producción baja, media y alta. Se toma una muestra al azar de 20 celulas, dentro de cierta población y se verifica si cada una de estas está en producción de la proteina (el resultado del aparato es si o no: 1 o 0, por cada célula analizada). De esta muestra resultan 12 células en producción (1) y las demás en negativo (0).

Por otro lado sabemos que si la población está en el estado A, solo el 20% de la células producen la proteina, si está en el estado B el 50% de las células la producen y si está en el estado C el 70% la produce. ¿cual es la probabilidad de que la población esté en cada uno de estos estados?

Sea $Y_i = 1$ si la célula i de la muestra está produciendo la proteina y $Y_i = 0$ si no, entonces $Y_i \sim Be(\theta)$ donde θ es la probabilidad de que la célula esté produciendo la proteina. Sea $X = \sum_{i=1}^{20} Y_i$, entonces es fácil ver que $X \sim Bi(20, \theta)$. Ahora, sabemos que $\theta = 0.2, 0.5, 0.7$ solamente, y como no tenemos más información de en qué estado está la población decimos que

$$p_\theta(t) = \frac{1}{3} \text{ si } t = 0.2, 0.5, 0.7$$

y cero en otro caso (esta es la *a priori* para θ).

Por otro lado tenemos que

$$p_{X|\theta}(12 \mid t) = C_{12}^{20} t^{12}(1-t)^8$$

ya que $X \sim Bi(20, \theta)$, este sería el modelo (observacional) para los datos. Entonces

$$p_{\theta|X}(t \mid 12) = \frac{C_{12}^{20} t^{12}(1-t)^{8\,\frac{1}{3}}}{\sum_{h=0.2,0.5,0.7} C_{12}^{20} h^{12}(1-h)^{8\,\frac{1}{3}}} \quad \text{para} \quad t = 0.2, 0.5, 0.7$$

(que sería la *a posteriori* para θ). Los cálculos respectivos los resumimos en la tabla 1.

(Note que $p_{X|\theta}(x \mid t)$, vista como función de t, es la verosimilitud en el sentido usual.)

$\theta = t$	0.2	0.5	0.7	suma
$p_\theta(t)$	0.3333	0.3333	0.3333	1.0000
$p_{X\|\theta}(12 \mid t)$	0.0008	0.1201	0.1143	
Prod.	0.0026	0.0400	0.0380	0.0806
$p_{\theta\|X}(t \mid 12)$	0.0322	0.4963	0.4715	1.0000

Table: Cálculos para la *a posteriori* de θ.

En el argot Bayesiano se suele hacer uso indiscriminado de abusos, a veces hasta peligrosos, de notación, que, sin embargo, resultan en un gran ahorro de tinta y en textos más compactos. Por ejemplo

$$p_{\theta|X}(t \mid x) = \frac{p_{X|\theta}(x \mid t)p_\theta(t)}{\sum_{h \in M_\theta} p_{X|\theta}(x \mid h)p_\theta(h)} \quad \text{para} \quad t \in M_\theta$$

lo escribiríamos como

$$f(\theta \mid X) \propto f(X \mid \theta)f(\theta).$$

Distribución posterior

Supongamos que $X_i \mid p \sim Be(p)$ y estas son independientes y que la incertidumbre acerca de $p \in [0,1]$ la cuantificamos con $f(p)$ y $p \sim Beta(\alpha, \beta)$ *a priori*. Obtenemos $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ y entonces

$$P(p \leq p_0 \mid \boldsymbol{X}) = \frac{P(\boldsymbol{X} \mid p \leq p_0) P(p \leq p_0)}{P(\boldsymbol{X})}.$$

Pero vemos que

$$P(\boldsymbol{X} \mid p \leq p_0) P(p \leq p_0) = P(\boldsymbol{X}, p \leq p_0) = \int_0^{p_0} f(\boldsymbol{X}, p) dp.$$

Ahora $f(\boldsymbol{X}, p) = f(\boldsymbol{X} \mid p) f(p)$ y entonces

$$P(p \leq p_0 \mid \boldsymbol{X}) \propto \int_0^{p_0} f(\boldsymbol{X} \mid p) f(p) dp.$$

Notando que la parte izquierda es la **distribución posterior para** p, tenemos que su **densidad posterior** es

$$f(p \mid \boldsymbol{X}) \propto f(\boldsymbol{X} \mid p) f(p).$$

Pero

$$f(\boldsymbol{X} \mid p) = p^{\sum_{i=1}^n x_i} (1-p)^{n - \sum_{i=1}^n x_i}$$

y

$$f(p) = B(\alpha, \beta)^{-1} p^{\alpha - 1} (1-p)^{\beta - 1} I_{[0,1]}(p),$$

y entonces

$$f(p \mid \boldsymbol{X}) \propto p^{(\alpha + \sum_{i=1}^n x_i) - 1} (1-p)^{(\beta + n - \sum_{i=1}^n x_i) - 1} I_{[0,1]}(p).$$

Por lo tanto

$$p \mid \boldsymbol{X} \sim Beta\left(\alpha + \sum_{i=1}^n X_i, \beta + n - \sum_{i=1}^n X_i\right).$$

Ejemplo 3

A continuación presentamos algunos ejemplos de iniciales y posteriores para muestreo Bernoulli con inicial Beta, con el programa en R.

Tenemos una pareja que ha tenido 5 hijas en 5 embarazos, como sería la distribución posterior de la probabilidad de que la pareja tenga una hija en otro embarazo.

Estimación con muchos parámetros

El objeto principal de la inferencia Bayesiana es la distribución posterior de los parámetros de interés involucrados. Por ejemplo, si tenemos dos parámetros θ_1 y θ_2, y los datos $\boldsymbol{X}$, tenemos que encontrar la distribución posterior

$$\pi(\theta_1, \theta_2 \mid \boldsymbol{X}).$$

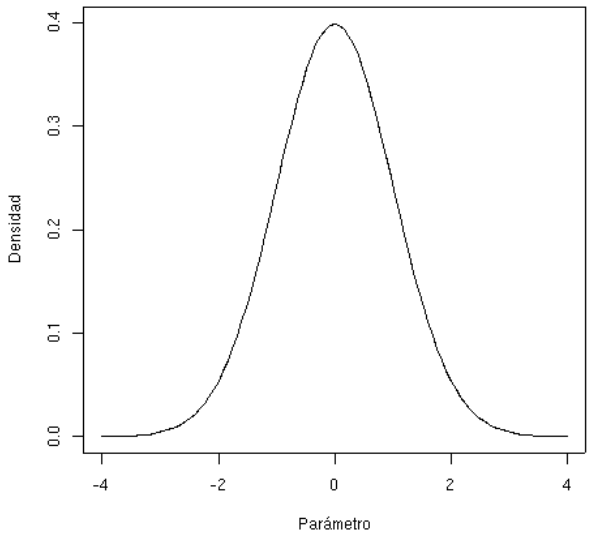
Ahora, ¿qué haríamos si el parámetro de interés es solamente θ_1?

Lo que necesitamos es la posterior de θ_1, y esto por teoría básica de probabilidad es

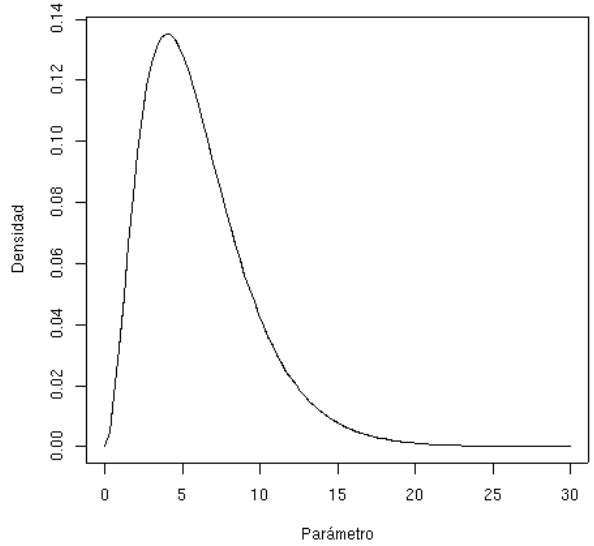
$$\pi(\theta_1 \mid \boldsymbol{X}) = \int \pi(\theta_1, \theta_2 \mid \boldsymbol{X}) d\theta_2.$$

O sea, encontrado la marginal. Esta es la llamada distribución marginal posterior de θ_1 y etc.

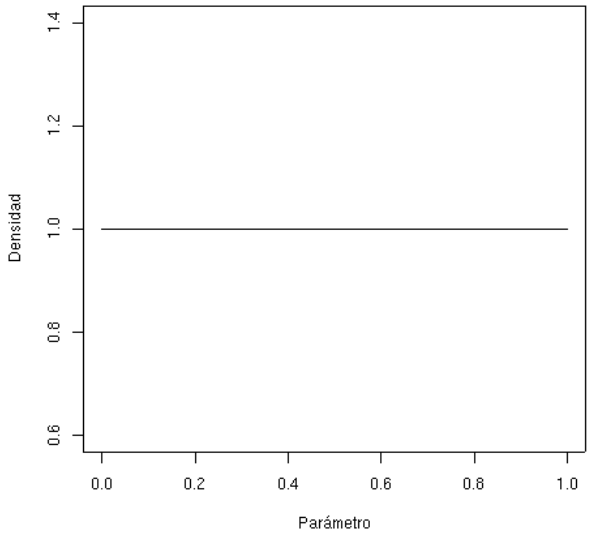
Suponiendo que tenemos ahora la distribución posterior $f(\theta \mid \boldsymbol{X})$, lo único que nos resta hacer es *reportarla* como el resultado de nuestra inferencia: ¿Como la reportaría? (vea los ejemplos en la figura 5).


Densidad
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
Parámetro

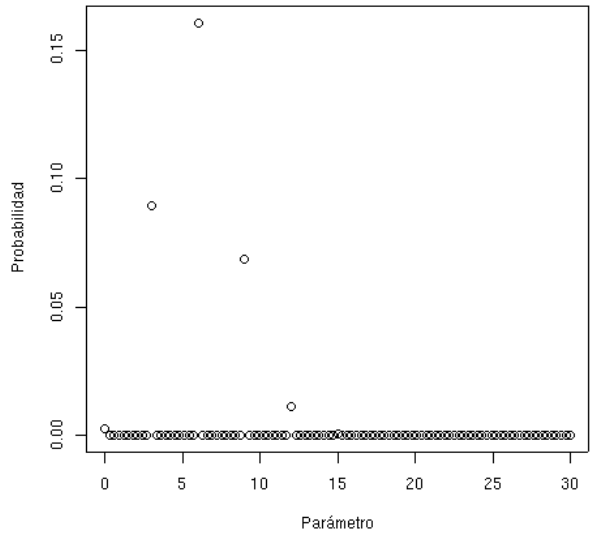
(a)


Densidad
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0
5
10
15
20
25
30
Parámetro

(b)


Densidad
0.6
0.8
1.0
1.2
1.4
0.0
0.2
0.4
0.6
0.8
1.0
Parámetro

(c)


Probabilidad
0.00
0.05
0.10
0.15
0
5
10
15
20
25
30
Parámetro

(d)

El concepto de “estimación” dentro de la estadística Bayesiana, no se entiende más que como **resúmenes de la distribución posterior** con la que se cuenta (desde luego que hay resúmenes buenos y resúmenes malos). Por lo tanto, el concepto de estimación puntual lo podemos ver como resumir una distribución con un solo punto (por absurdo que en esta perspectiva parezca).

Por ejemplo, podemos usar la esperanza de la distribución posterior. Se puede también tomar como estimador el máximo de la distribución posterior, este se llama el estimador MAP (*Maximum a posteriori*).

Ejemplo 4

Tenemos un tratamiento experimental para una enfermedad, de la cual no se sabe mucho, y este tratamiento se ha usado en 20 ratas con las mismas características, de las cuales 15 se han curado (éxito). El tratamiento estándar tiene una probabilidad de éxito de 50%. Se plantea la siguiente hipótesis: *El tratamiento experimental es mejor que el estándar*.

¿qué podria decir con inferencia Bayesiana al respecto?

Ejemplo 4

Sea $(X_1, X_2, \ldots, X_n) \sim Be(\theta_E)$.

$$H_1: \ \theta_E > \theta_S \ \text{ vs. } \ H_2: \ \theta_E \leq \theta_S.$$

En este caso se sabe que $\theta_S = 0.5$.

Ejemplo 5

Sea $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ con $X_i \sim N(\mu, \sigma)$, σ, desviación estandar, conocida.

1. Sea $\lambda = \sigma^{-2}$, la "precisión" (el inverso de la varianza, en Bayesiana se utiliza solo por conveniencia matemática). Y sea $X_i \sim N(\mu, \lambda)$ (abuso de notación).
2. Inicial para μ: $\mu \sim N(\mu_0, \lambda_0)$.
3. Sea $\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\lambda_n = \lambda_0 + n\lambda$ y $\mu_p = \frac{\lambda_0\mu_0 + n\bar{\boldsymbol{X}}\lambda}{\lambda_0 + n\lambda}$.
4. Entonces, $\mu \mid \boldsymbol{X} \sim N(\mu_p, \lambda_n)$. La distribución Normal es conjugada para muestreo Normal con precisión (varianza) conocida.

Cuentas de inferencia en el modelo Normal y regresión lineal, análisis conjugado

$$\boldsymbol{y} = \boldsymbol{X}\beta + \epsilon,$$

$\boldsymbol{X}$ matriz $n \times p$ de diseño (o regresores), β vector $p \times 1$ de parámetros, $\boldsymbol{y}$ vector $n \times 1$ de observaciones (respuesta) y $\epsilon \sim MVN(0, \sigma I_n)$.
A priori *NIG* (normal gamma inversa).

$$\begin{aligned} f(\beta, \sigma^2) &= f(\beta|\sigma^2) f(\sigma^2) = N(\mu_0, \sigma \boldsymbol{V}_0) IG(a_0, b_0) \\ &\propto \left(\frac{1}{\sigma^2}\right)^{a+p/2+1} \\ &\quad \exp\left[-\frac{1}{\sigma^2}\left\{b_0 + \frac{1}{2}(\beta - \mu_0)' \boldsymbol{V}_0{}^{-1}(\beta - \mu_0)|\right\}\right]. \end{aligned}$$

Cuentas de inferencia en el modelo Normal y regresión lineal, análisis conjugado

$$
\begin{aligned}
f(\beta, \sigma^2) &= f(\beta|\sigma^2) f(\sigma^2) = N(\mu_0, \sigma \boldsymbol{V}_0) IG(a_0, b_0) \\
&\propto \left(\frac{1}{\sigma^2}\right)^{a+p/2+1} \\
&\quad \exp\left[-\frac{1}{\sigma^2}\left\{b_0 + \frac{1}{2}(\beta - \mu_0)' \boldsymbol{V}_0{}^{-1}(\beta - \mu_0)\right\}\right].
\end{aligned}
$$

Verosimilitud:

$$
\begin{aligned}
f(\boldsymbol{X}|\beta, \sigma^2) &\propto \left(\frac{1}{\sigma^2}\right)^{n/2} \\
&\quad \exp\left[-\frac{1}{\sigma^2}\left\{\frac{1}{2}(\boldsymbol{X}\beta - \boldsymbol{y})'(\boldsymbol{X} - \boldsymbol{y})|\right\}\right].
\end{aligned}
$$

Cuentas de inferencia en el modelo Normal y regresión lineal, análisis conjugado

Inicial:

$$
\begin{aligned}
f(\boldsymbol{\beta}, \sigma^2) &= f(\boldsymbol{\beta}|\sigma^2) f(\sigma^2) = N(\boldsymbol{\mu}_0, \sigma \boldsymbol{V}_0) IG(a_0, b_0) \\
&\propto \left(\frac{1}{\sigma^2}\right)^{a+p/2+1} \\
&\quad \exp\left[-\frac{1}{\sigma^2}\left\{b_0 + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_0)' \boldsymbol{V}_0^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu}_0)\right\}\right].
\end{aligned}
$$

Posterior:

$$
f(\boldsymbol{X}|\boldsymbol{\beta}, \sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{a+p/2+n/2-1} \cdot
$$

$$
\exp\left[-\frac{1}{\sigma^2}\left\{b_0 + \frac{1}{2}(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})'(\boldsymbol{X} - \boldsymbol{y}) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_0)' \boldsymbol{V}_0^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu}_0)\right\}\right].
$$

Ejemplo 6

Tenemos una prueba de laboratorio que detecta una cierta enfermedad A. Denotamos el que salga una prueba positiva para la enfermedad como $T = 1$ y negativa como $T = 0$, y que el paciente tenga la enfermedad A como $E = 1$ y $E = 0$ en otro caso. Las características de la prueba son:

$$P(T = 1 \mid E = 0) = 0.01, \quad P(T = 0 \mid E = 1) = 0.08,$$

y la prevalencia de la enfermedad (la proporción de personas enfermas en la población en cuestión) es de 0.12. O sea $P(E = 1) = 0.12$.

Ejemplo 6

Yo voy y me hago la prueba referida (perteneciendo yo a la población en cuestión) y esta sale positiva. ¿Cual es la probabilidad posterior de que yo tenga la enfermedad A?

Las características de la prueba son:

$P(T=1 \mid E=1) =$		0.92	$P(T=1 \mid E=0) =$		0.01
$P(T=0 \mid E=1) =$		0.08	$P(T=0 \mid E=0) =$		0.99
	$+$	1.00		$+$	1.00

con $P(E=1) = 0.12$.

Ejemplo 6

Para calcular la posterior hacemos:

E	0	1	Suma
$P(E)$	0.88	0.12	1.00
$P(T=1 \mid E)$	0.01	0.92	0.93
Prod.	0.0088	0.1104	0.1192
$P(E \mid T=1)$	0.0738	0.9262	1.0

O sea: $P(E = 1 \mid T = 1) = 0.9262$.

Pero, en realidad, **o tengo la enfermedad o no la tengo**; entonces...

¿Que quiere decir:
La probabilidad de que tenga la enfermedad es 0.9262?!

Es *nuestra* probabilidad, dadas las características de la prueba y que esta salió positiva y que *a priori* $P(E = 1) = 0.12$.

No hay frecuencias o eventos repetidos: o estoy o no estoy enfermo!

(Más aun: ¿Porqué la distribución condicional de los parámetros dados los datos (eg $P(E|T = 1)$) nos da la respuesta deseada?)

Probabilidad subjetiva, probabilidad condicional (5)

Entonces, ¿qué quiere decir?

- La probabilidad de que al lanzar una moneda esta caiga en “águila”.
- La probabilidad de que llueva mañana es 0.2 en la ciudad A.
- La probabilidad de que el equipo B de futbol gane su próximo partido.
- La probabilidad de que alguien se muera si se enferma de la enfermedad C.
- La probabilidad de alcanzar inmunidad a la enfermedad C al aplicarse la vacuna D.
- La probabilidad de que la cotización del USD esté por arriba de x pesos al final del año.
- etc.

Probabilidad subjetiva, probabilidad condicional (5)

Entonces, ¿qué quiere decir?

NO sabemos (bien). Pero, son eventos de los que hacemos aseveraciones de incertidumbre, y queremos y necesitamos hablar de ellos, aunque sean inciertos. ESO es hacer estadística/análisis de riesgo/confiabilidad/predicción financiera/etc.

Es más. Tenemos la **necesidad e interés** de cuantificar esta incertidumbre y tomar decisiones bajo incertidumbre, para todo tipo de eventos, complejos, únicos, no repetibles, no observables (solo indirectamente) etc. más allá de lanzar un volado (eventos infinitamente repetibles, iid).

Probabilidad subjetiva, probabilidad condicional (5)

Uno de los puntos de partida básicos en la estadística Bayesiana es el concepto de probabilidad (y su definición). Para empezar, pongamos unos ejemplos en que usamos la probabilidad y tratemos de encontrar una definición suficientemente ámplia para esta

1. ¿Cuál es la probabilidad de que al lanzar una moneda caiga "águila"?
2. ¿Cuál es la probabilidad de que el profesor traiga más de 200 pesos en sus bolsas del pantalón?
3. ¿Cuál es la probabilidad de que llueva mañana?
4. ¿Cuál es la probabilidad de que haya llovido ayer en Durango?
5. ¿Cuál es la probabilidad de que haya más de 10^9 estrellas en nuestra galaxia?

Sin duda, para las preguntas 1–5 nosotros podríamos pensar en que existen probabilidades concretas para su respuesta, aún cuando en la mayoría de los casos es difícil establecer cual es el valor “exacto” de éstas. Se podría pensar que para la primera pregunta la respuesta es $\frac{1}{2}$, pero ¿De que moneda estamos hablando? ¿tiene esta moneda un águila? (podría ser extranjera!). Para la pregunta 4, yo podría informarme si llovió en Durango ayer (la ciudad o el estado completo??) y entonces esta probabilidad sería 0 ó 1; pero, en este instante, ¿que tanto sabemos del evento “ayer llovió en Durango”?

La probabilidad, en un sentido amplio, es entonces una medida de lo que **sabemos** acerca de un evento. Esto quiere decir que la probabilidad es siempre **contextual**, dada una una serie de supuestos y consideraciones, aun para las probabilidades más simples: En el caso de la moneda, para decir "la probabilidad es $\frac{1}{2}$", tendríamos que suponer, explicita o implícitamente, algo como:

- Que, como es el expositor o un colega él/la que está sacando la moneda para ser lanzada, esperemos que ésta sea una moneda común y corriente (de México) y que al caer quedará horizontal y una, y solo una, de sus caras será el grabado de un águila (el escudo nacional).
- Que el evento A de que salga águila es igualmente probable al que salga la otra cara de la moneda, evento B (la moneda es "justa").

Con estos dos supuestos, uno puede entonces *calcular* la probabilidad de A, esto es: Por el primer supuesto tenemos que $A \cup B = \Omega$ y $A \cap B = \emptyset$. Por el segundo supuesto tenemos que $P(A) = P(B) = p$ y por los axiomas de probabilidad tenemos que $P(A) + P(B) = P(\Omega) = 1$ ó $2p = 1$ lo cual implica que $P(A) = \frac{1}{2}$. La tradicional fórmula de "casos favorables entre casos totales", que en una visión simplista se utiliza como una *definición* de probabilidad, ahora la vemos como un *teorema* o *cálculo* para unas ciertas sencillas probabilidades: Si tenemos un conjunto de n eventos *disjuntos y equiprobables* A_i cuya unión es Ω, entonces $P(A_i) = \frac{1}{n}$ y $P\left(\cup_{i=1}^{m} A_i\right) = \frac{m}{n}$.

Sin embargo, en un marco más general no podemos restringirnos a “casos favorables entre casos totales”: Si yo digo que la probabilidad de que llueva mañana es 0.2, es absurdo creer que lo que quiero decir es que de 100 mañanas posibles en 20 lloverá! Más bien es una medida de lo que yo se acerca del la verosimilitud del evento.

Recapitulando

La probabilidad es siempre contextual o *condicional*: depende de quien la asienta, bajo que condiciones y que supuestos etc. Esto es, tendríamos que decir $P(A) = 0.2$ *dado* que X, Y y Z; o en la notación estándar $P(A \mid X, Y, Z)$.

- Deberíamos de siempre poner $P(\cdot \mid H)$, condicional al *agente* que habla, las circunstancias, los supuestos que hace, la sigma-álgebra que seleccione, etc.
- Una vez que observamos datos X estos son fijos y actualizamos estos datos usando la distribución posterior $P(\cdot \mid X)$. (¿Porqué la actualizamos con la distribución posterior?)
- El *agente* que establece la medida de probabilidad debe de seguir ciertas reglas, y en particular algo tenderá que ver dicha medida e interpretación con las apuestas que el *agente* esté dispuesto a aceptar.

Nota acerca de Momios

Las apuestas son culturalmente muy comunes y aceptadas en otros culturas, en especial la anglosajona. Sin embargo, no es necesariamente el caso en otros lugares. Añadido a esto hay varias manera de expresar apuestas, de las cuales las más populares son por ejemplo: 3 a 1, 3:1, +300, -1000, o 1.3, 2.3.

Lo que quiere decir el momio $a : b$ **a favor de** un evento A es que, si el evento resulta ser cierto al apostar 100 al *apostador se le regresa su suma apostada de* 100 y **gana** $100 \cdot a/b$, o sea, al apostar 100 y ganar, el apostar obtiene $100 \cdot (1 + a/b)$.

Nota acerca de Momios

Las apuestas en el RU se presentan como a/b o en Hong Kong como a/b en decimales (eg 1.25) y en Europa y otros países como $(1 + a/b)$ (eg. 2.25).

Para **aceptar** apuestas a favor de A la apuesta $a : b$ con $a/b = \frac{p}{1-p}$, con $P(A) = p$ tiene un valor esperado de, apostando 100 pesos:

Note que el agente es la casa de apuestas, si sucede A gana el agente, y se queda con los 100 pesos, sino tiene que pagar $100 \cdot a/b$, y entonces la apuesta tiene un retorno esperado de

$$100p - 100\frac{p}{1-p}(1-p) = 0, \quad \text{i.e. apuesta equilibrada}$$

Podemos entonces aceptar cualquier apuesta $a/b \leq \frac{p}{1-p}$ para tener un retorno no negativo.

Introducción

La estadística Bayesiana se fundamenta en un marco teórico general para hacer inferencias. Éste se basa en que podemos comparar que tan *verosímiles* son dos eventos cualquiera de nuestro espacio de eventos. Partiendo de esto, se forman una serie de axiomas que dicho ordenamiento debería de tener y de ahí los axiomas usuales de la probabilidad son deducidos como teoremas. La conclusión final es que para cualquier conjunto de eventos es posible que una *agente*[1] establezca una medida de probabilidad que defina las probabilidad de cada evento; no sin esto representar un esfuerzo considerable.

[1] Ponemos *agente* en cursivas pues nos referimos no necesariamente a un individuo. Bien podemos estar hablando de un par de expertos, un panel o, por ejemplo, de una sociedad en su conjunto o al mundo en su totalidad, que en un momento dado y para unas circunstancias específicas, se pongan de acuerdo en una medida de probabilidad para un conjunto de eventos particulares.

Preferencias entre eventos

El formalismo siguiente ha sido establecido de varias maneras diferentes por varios autores. La versión que expondremos aquí es la aparecida en DeGroot (1970, cap. 6). Para esto, empezamos con una cita a DeGroot (1970, p. 70):

> *...suitable probabilities can often be assigned objectively and quickly because of wide agreement on the appropriateness of a specific distribution for a certain type of problem...On the other hand, there are some situations for which it would be very difficult to find even two people who would agree on the appropriateness of any specific distribution.*

Comparación entre eventos

Vamos a partir de que tenemos un espacio medible $(\Omega, @)$ y que para cada dos eventos $A, B \in @$, un *agente* puede decir si A es más, menos o igual de *factible* (*verosímil*, *probable*, **Lo vamos a definir claramente a continuación**) que B. Esto lo escribimos

$$A \succ B, \quad A \prec B, \quad A \sim B.$$

Y si para indicar que A no es más verosímil que B decimos que

$$A \preceq B.$$

Consecuencias

Así como se tiene un ordenamiento entre eventos, vamos a pensar que hay un ordenamiento $\preceq$ (usando el mismo símbolo como abuso de notación) entre una serie de consecuencias $c \in \mathcal{C}$, que podemos enfrentar en el contexto de $(\Omega, @)$. Por ejemplo, ¿preferiría comer comida china o pizza?, ¿prefiere el sistema operativo Linux o MS Windows?.

Vamos a suponer que $\mathcal{C}$ es acotado, esto es, que existen c_* y c^* tales que $c_* \preceq c \preceq c^*$, para todo $c \in \mathcal{C}$.

Por el momento solamente vamos a suponer, y necesitar, la existencia de las consecuencias extremas c_* y c^*.

Vamos ahora a permitir la construcción de consecuencias "compuestas" (a los elementos de $\mathcal{C}$ las llamaremos consecuencias "simples"). Tomemos un evento $E \in \mathcal{E}$ y dos consecuencias (simples o compuestas) c_1 y c_2 y hagamos una "lotería" en el que, si resulta cierto el evento E, enfrentas la consecuencia c_1 y si resulta E' enfrentas la consecuencia c_2. Denotamos esta nueva consecuencia como

$$\{c_1 \mid E, c_2 \mid E'\}.$$

Interpretación de *Factible* o *Verosimil*

Supuesto

Para $A, B \in @$*, tenemos que*

$$\{c^* \mid A, c_* \mid A'\} \preceq \{c^* \mid B, c_* \mid B'\}$$

si y solo si $A \preceq B$.

Entonces, en este contexto, *verosímil*, *factible* o *probable* quiere decir: Dadas las consecuencias c_* y c^*, en el contexto de $(\Omega, @)$, al *agente* se le ofrece la lotería $\{c^* \mid A, c_* \mid A'\}$ o la lotería $\{c^* \mid B, c_* \mid B'\}$.

A **es menos factible o verosimil o *probable* que** B **para el *agente*** ($A \prec B$), significa que este (el *agente*) siempre preferiría la lotería $\{c^* \mid B, c_* \mid B'\}$ sobre la lotería $\{c^* \mid A, c_* \mid A'\}$. Si $A \sim B$ a el *agente* le son indiferentes las loterias.

Interpretación de *Factible* o *Verosimil*

Esto nos da la interpretación de lo que quiere decir factibilidad y eventualmente nos dará la interpretación (definición bayesiana) de la probabilidad.

Eventualmente vamos a **demostrar que** existe una única medida de Probabilidad en $(\Omega, @)$ tal que

$$A \preceq B \text{ si y solo si } P(A) \leq P(B);$$

poniendo ciertas restricciones, axiomas, sobre el ordenamiento de factibilidades $\preceq$.

Axiomas de preferencias

Vamos a establecer una serie de axiomas que la relación $\preceq$ debería de seguir; o al menos para una *agente* coherente (racional).

Axioma

Para cualesquiera dos eventos $A, B \in \mathcal{Q}$, exactamente una de las tres condiciones siguientes es válida: $A \succ B$, $A \prec B$, $A \sim B$.

Axioma

Si $A_1, A_2, B_1, B_2 \in \mathcal{Q}$*, son cuatro eventos tales que* $A_1 A_2 = B_1 B_2 = \emptyset$ *y* $A_i \preceq B_i, i = 1, 2$*, entonces* $A_1 \cup A_2 \preceq B_1 \cup B_2$*. Más aún, si* $A_i \prec B_i$ *para algún* i *entonces* $A_1 \cup A_2 \prec B_1 \cup B_2$*.*

La interpretación de estos axiomas es sencilla y es creíble que cualquier *agente* racional tendría que seguirlos.

Con estos dos sencillos axiomas podemos probar una serie de resultados. En especial, la transitividad de la relación $\preceq$. Primero demostramos el siguiente lema

Lema

Suponga que $A, B, D \in @$ son eventos tales que $AD = BD = \emptyset$. Entonces $A \preceq B$ si y solo si $A \cup D \preceq B \cup D$.

Demostración.

Es una consecuencia del axioma 2 suponiendo $A \preceq B$ y tomando $A = A_1$, $B = B_1$ y $A_2 = B_2 = D$, y suponiendo $A \succ B$ y tomando $B = A_1$ etc. $\square$

Teorema

Sean $A, B, D \in @$ tres eventos tales que $A \preceq B$ y $B \preceq D$, entonces $A \preceq D$.

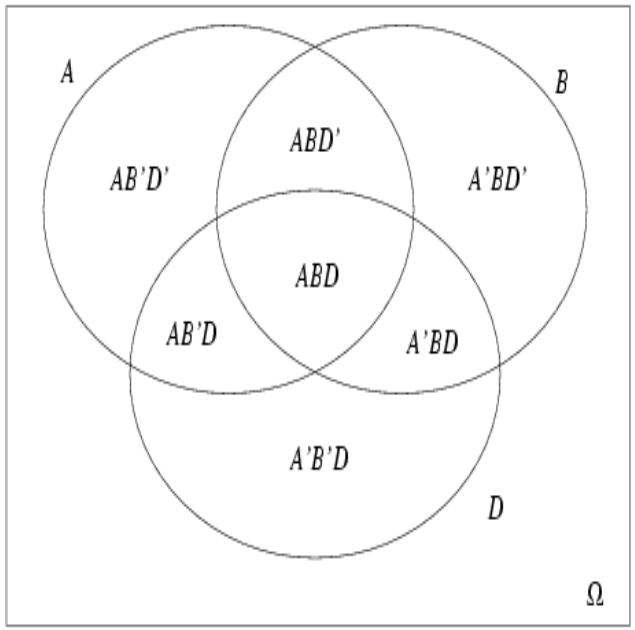


Figure: Diagrama auxiliar para la demostración del Teorema 2, C' denota el complemento de C.

Demostración.

Usando la figura 2 vemos que, dado que $A \preceq B$ se sigue que

$$AB'D' \cup AB'D \preceq A'BD' \cup A'BD.$$

Y dado que $B \preceq D$ se sigue que

$$ABD' \cup A'BD' \preceq AB'D \cup A'B'D.$$

Como los elementos de la izquierda son disjuntos así como los elementos de la derecha entre si, se sigue del axioma 2 que

$$AB'D' \cup AB'D \cup ABD' \cup A'BD' \preceq A'BD' \cup A'BD \cup AB'D \cup A'B'D.$$

Ahora eliminamos el evento común $AB'D \cup A'BD'$ en los dos lados usando el lema 1 y tenemos $AB'D' \cup ABD' \preceq A'BD \cup A'B'D$ de lo cual se concluye que $A \preceq D$ agregando los conjuntos necesarios usando el lema 1. □

CIMAT

Tenemos otros dos teoremas cuyas demostraciones son elementales

Teorema

Si A_i son n eventos disjuntos, al igual que B_i tales que $A_i \preceq B_i$ entonces $\cup_{i=1}^n A_i \preceq \cup_{i=1}^n B_i$. Si para 2 i tenemos que $A_i \prec B_i$, entonces $\cup_{i=1}^n A_i \prec \cup_{i=1}^n B_i$.

Teorema

Para cualesquiera dos eventos $A, B \in @$, $A \preceq B$ si y solo si $A' \succeq B'$.

Exprese $A = B'A \cup AB$, $A' = B'A' \cup A'B$ etc.

El siguiente axioma, aun cuando perfectamente claro, es necesario pues no se puede deducir de los dos anteriores:

Axioma

Si $A \in \mathscr{A}$ es un evento, entonces $\emptyset \preceq A$. Más aún, $\emptyset \preceq \Omega$.

Usando este axioma y los demás resultados, podemos probar que

Teorema

Si $A, B \in \mathscr{A}$ son dos eventos tales que $A \subset B$, entonces $A \preceq B$. En particular $\emptyset \preceq A \preceq \Omega$.

El siguiente axioma podría ser evitado si quisiéramos trabajar solamente con @ finitas. Sin embargo, por razones de conveniencia matemática este es introducido y, desde luego, es intuitivamente razonable:

Axioma

Si $A_1 \supset A_2 \supset \cdots$ *es una secuencia decreciente de eventos en* @ *y* $B \in$ @ *es otro evento tal que* $A_i \succeq B$ *para toda* i*, entonces* $\cap_{i=1}^{\infty} A_i \succeq B$.

(Si B es una "cota" inferior para la verosimilitud de los A_i's, entonces el límite inferior de los A_i's no es menos verosímil que B.)

El siguiente resultado es un teorema que es recíproco de el axioma 4:

Teorema

Si $A_1 \subset A_2 \subset \cdots$ es una secuencia creciente de eventos en @ y $B \in$ @ es otro evento tal que $A_i \preceq B$ para toda i, entonces $\cup_{i=1}^{\infty} A_i \preceq B$.

Ahora, usando el axioma 4 podemos generalizar el teorema 3 para familias numerables.

Teorema

Si A_i son eventos disjuntos al igual que B_i tales que $A_i \preceq B_i$ entonces $\cup_{i=1}^{\infty} A_i \preceq \cup_{i=1}^{\infty} B_i$. Si para algún i tenemos que $A_i \prec B_i$, entonces $\cup_{i=1}^{\infty} A_i \prec \cup_{i=1}^{\infty} B_i$.

El experimento auxiliar

El problem ahora es que con los 4 axiomas aun no podemos definir unívocamente una medida de probabilidad en @. Por ejemplo, imaginemos dos eventos A y A', estos junto con $\emptyset$ y Ω forman una σ–álgebra, y al establecer que, por ejemplo, $A \prec A'$, tendríamos una relación de verosimilitud acorde con los axiomas anteriores. Sin embargo, hay una infinidad de medidas de probabilidad que concordarían con $\preceq$.

Definición

P medida en @ coincide con $\preceq$ si para todo $A, B \in$ @, $A \preceq B$ si y solo si $P(A) \leq P(B)$.

En ejemplo de arriba podemos poner, por ejemplo, $P(A) = 0.3$ o $P(A) = 0.2$, etc.

Lo que tenemos que hacer es agregar a @ una serie de eventos auxiliares, independientes de los originales (elementales, como una ruleta en un círculo), tales que para toda $0 \leq p \leq 1$ exista un $B \in @$ con probabilidad p. Entonces lo único que necesitamos encontrar es un tal B tal que $A \sim B$, para encontrar la probabilidad de A. Esto en otras palabras, es que vamos a comparar la verosimilitud de nuestros eventos con aquella de eventos auxiliares, de los cuales está establecida su probabilidad, y así encontrar la probabilidad de cualquier evento.

Usando un poco de teoría de la medida es muy fácil establecer el último axioma. Sea λ la medida de Lebesgue y $\mathcal{B}$ los Borelianos en el $[0, 1]$:

Axioma

Existe una variable aleatoria X en $(\Omega, @)$, con $0 \leq X(\omega) \leq 1$, para todo $\omega \in \Omega$, tal que para cualesquiera $I_1, I_2 \in \mathcal{B}$, $\{X \in I_1\} \preceq \{X \in I_2\}$ si y solo si $\lambda(I_1) \leq \lambda(I_2)$.

Teoremas de probabilidad

Con los 5 axiomas anteriores ahora crearemos una medida de probabilidad en @. Partimos primero generando una función que le asigne un número entre 0 y 1 a cualquier evento. Sea $G(I) = \{X \in I\}$ (X la v.a. referida en el axioma 5).

Teorema

Si $A \in$ @ es cualquier evento, existe un único número $a^* \in [0,1]$ tal que $A \sim G([0, a^*])$.

Demostración.

Considere $a^* = \inf\{a : G([0,a]) \succeq A\}$. Acercándose a a^* por una secuencia decreciente demuestre que $G([0,a^*]) \succeq A$ y acercándose a a^* mediante una secuencia creciente demuestre que $G([0,a^*]) \preceq A$. Tenga cuidado con los casos $a^* = 0, 1$. Ahora, si $a_1 \neq a^*$ entonces necesariamente $G([0,a_1])$ no es equivalente a $G([0,a^*])$ y entonces A no puede ser equivalente a los dos al mismo tiempo. □

Sea ahora $P(A)$, con $A \in \mathscr{A}$ el real tal que $A \sim G([0, P(A)])$. Demostraremos que esta función concuerda con $\preceq$ y que es una medida de probabilidad para $\mathscr{A}$.

Teorema

Sean $A, B \in \mathscr{A}$ dos eventos. $A \preceq B$ si y solo si $P(A) \leq P(B)$.

Demostración.

Es claro que $A \preceq B$ si y solo si $G([0, P(A)]) \preceq G([0, P(B)])$. □

Es claro que $P(\emptyset) = 0$, $P(\Omega) = 1$ y que $P(A) \in [0, 1]$ para todo $A \in \Omega$. Es entonces solo necesario demostrar la σ–aditividad de P.

Teorema

Para $A, B \in$ @ con $AB = \emptyset$, tenemos que $P(A \cup B) = P(A) + P(B)$.

Demostración.

Note que si suponemos que $B \prec G((P(A), P(A \cup B)])$ vemos que

$$A \cup B \prec G([0, P(A)]) \cup G((P(A), P(A \cup B)]) \sim G([0, P(A \cup B)])$$

lo cual es una contradicción, y similarmente si supusiéramos que $B \succ G((P(A), P(A \cup B)])$. Por lo tanto $B \sim G((P(A), P(A \cup B)])$ o $B \sim G((0, P(A \cup B) - P(A)])$ y, por lo tanto,

$$P(A \cup B) - P(A) = P(B).$$

Teorema

P es una medida de probabilidad en $(\Omega, @)$.

Demostración.

Para demostrar la σ–aditividad de P primero se generaliza, por inducción el resultado anterior para n conjuntos. Después establecemos que

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{n} P(A_i) + P(\cup_{i=n+1}^{\infty} A_i)$$

notando que $\cup_{i=n+1}^{\infty} A_i$ es una sucesión decreciente de conjuntos. Sea $b = \lim_{n\to\infty} P(\cup_{i=n+1}^{\infty} A_i)$ (existe el límite por tratarse de una sucesión decreciente acotada), necesariamente $\cup_{i=n+1}^{\infty} A_i \succeq G([0, b])$, pero entonces

$$\emptyset = \cap_{n=1}^{\infty} \cup_{i=n+1}^{\infty} A_i \succeq G([0, b])$$

y por lo tanto $\emptyset \sim G([0, b])$ y $b = 0$. □

Falta demostrar que P así definida es la única medida que concuerda con la relación $\preceq$. Acercándonos por arriba y por abajo para demostrar que cualquier medi da P' que concuerde con $\preceq$ debe de cumplir que $P'(G([0, a])) = a$. Entonces para todo $A \in \mathcal{Q}$,
$P'(A) = P'(G([0, P(A)])) = P(A)$.

Preferencias Condicionales

¿Como cambian mis preferencias si ahora yo observo el evento D?

En este sentido, al observar el evento D nuestra relación de preferencia $\preceq$ cambia a otra, digamos $\preceq_D$. El meollo del asunto es establecer cual es la relación entre $\preceq$ y $\preceq_D$. Se propone el siguiente

axioma (que de hecho define lo que queremos decir con "dado D"):

Axioma

$A \preceq_D B$ si y solo si $AD \preceq BD$.

Es fácil ver que si $\preceq$ cumple con los 5 axiomas de arriba, entonces $\preceq_D$ también. En particular la v.a. uniforme X del axioma 5 es la *misma* (c.s.) en ambos casos. (Se usa la misma ruleta.)

Preferencias condicionales

Hemos demostrado que existe una única medida de probabilidad P en $(\Omega, @)$ que concuerda con $\preceq$; pero entonces también demostramos que existe una única medida P_D que concuerda con $\preceq_D$. ¿cual es la relación entre P y P_D?

Teorema

Si $P(D) \neq 0$,

$$P_D(A) = P(A \mid D) = \frac{P(AD)}{P(D)}.$$

Demostración.

$P(A \mid D)$ coincide con $\preceq_D$ porque $P(A \mid D) \leq P(B \mid D)$ si y solo si $P(AD) \leq P(BD)$, si y solo si $AD \preceq BD$, y si y solo si $A \preceq_D B$. Entonces $P_D(A) = P(A \mid D)$. □

Teoría de utilidad y decisiones (2)

En esta sección presentaremos los conceptos básicos de la teoría de decisiones bajo incertidumbre. Es muy común que la razón ulterior de un problema estadístico no sea la inferencia en sí sobre alguna medida de probabilidad P, sino que sea en realidad la toma de una decisión basada en esa medida.

En una cierta perspectiva, todo problema estadístico es en realidad, en un sentido amplio, un problema de decisión.

La metodología Bayesiana incorpora este concepto desde su fundación. Es entonces la metodología Bayesiana no solo un marco teórico general para hacer inferencia sino para tomar decisiones bajo incertidumbre.

Utilidad

Hay varios recuentos de la teoría de la utilidad, probablemente el más moderno y compacto es el presentado por Bernardo y Smith (1994, Cap. 2). Sin embargo, no daremos un recuento completo de esta teoría por cuestión de tiempo.

Daremos aquí una versión simplificada de Bernardo y Smith (1994, Cap. 2), utilizando las ideas de DeGroot explicadas arriba. Se trata de una versión "campechana" de ambas teorías; ciertamente un poco informal, pero que esperamos que exprese las ideas fundamentales del tema.

Empezamos primero con una partición finita $\mathcal{E}$ de @, que serán nuestros eventos relevantes. Tenemos también un conjunto finito de acciones $\mathcal{A}$ y un conjunto finito de consecuencias $\mathcal{C}$. La estructura es como sigue:

Para cada posible acción a_i que deseemos tomar, alguno de los eventos E_j aleatorios surgirá y tendremos que enfrentar la consecuencia c_{ij}. Tenemos entonces que

$$\mathcal{E} = \{E_j : j \in J\} \quad \mathcal{A} = \{a_i : i \in I\},$$

$$\mathcal{C} = \{c_{ij} : i \in I,\ j \in J\}.$$

Axioma

Existe una relación de preferencia $\preceq$ en $\mathcal{C}$ tal que para $c_1, c_2 \in \mathcal{C}$, una y solo una de estas tres relaciones es cierta:

$$c_1 \prec c_2, c_1 \sim c_2, c_1 \succ c_2.$$

Es claro lo que quiere decir la relación de preferencia entre consecuencias.

Recordemos la construcción de consecuencias “compuestas” (a los elementos de $\mathcal{C}$ las llamaremos consecuencias “simples”). Tomemos un evento $E \in \mathcal{E}$ y dos consecuencias (simples o compuestas) c_1 y c_2 y hagamos una “lotería” en el que, si resulta cierto el evento E, enfrentas la consecuencia c_1 y si resulta E' enfrentas la consecuencia c_2. Denotamos esta nueva consecuencia como:

$$\{c_1 \mid E, c_2 \mid E'\}.$$

Ejemplo: “Si llueve mañana te doy 100 pesos, sino te doy 20 pesos”.

Ejemplo: “Si tiro una moneda y sale águila te doy 100 pesos, sino te doy 0 pesos”. (¿Cómo compararía esta consecuencia compuesta con la consecuencia simple “te doy 50 pesos”?)

Notamos ahora que es posible extender nuestro ordenamiento $\preceq$ en $\mathcal{C}$ a consecuencias compuestas. Sabemos que, por el supuesto inicial:

Para $A, B \in \mathcal{E}$, tenemos que $\{c^* \mid A, c_* \mid A'\} \preceq \{c^* \mid B, c_* \mid B'\}$ si y solo si $A \preceq B$ si y solo si $P(A) \leq P(B)$.

Este supuesto solamente indica que debe de existir una coherencia entre las preferencias y las probabilidad de los eventos.

Axioma

Para toda consecuencia c, existe d tal que

$$c \sim \{c^* \mid G([0,d]), c_* \mid G([0,d])'\}.$$

Definimos ahora la función de utilidad de cualquier consecuencia c como el número $u(c) = d$ de arriba. *Por construcción, podemos construir un evento auxiliar* $G([0,d])$ *independiente de los eventos relevantes* $\mathcal{E}$.

Considere ahora consecuencias (simples o compuestas) c_1, c_2 y definamos A_1 y A_2 tales que $c_i \sim \{c^* \mid A_i, c_* \mid A_i'\}$. Entonces tenemos que $c_1 \preceq c_2$ si y solo si

$$\{c^* \mid A_1, c_* \mid A_1'\} \preceq \{c^* \mid A_2, c_* \mid A_2'\}$$

y esto si y solo si $P(A_1) \leq P(A_2)$ y esto si y sólo si $u(c_1) \leq u(c_2)$.

Esto es, la función u (de utilidad) *coincide* con nuestras preferencias entre las consecuencias. Nuestras preferencias entre consecuencias son medidas con una función en el $[0, 1]$.

El problema aun continua pues aun no sabemos cómo decidir cual de las acciones en $\mathcal{A}$ es la que debemos de tomar. El problema se basa en saber para cualquiera dos acciones a_i y a_k cual decidir entre ellas. **Note, sin embargo, que decidir entre a_i y a_k es equivalente a establecer la relación de preferencia entre las siguientes consecuencias:**

$$c(i) = \{c_{ij} \mid E_j : j \in J\} \text{ ó } c(k) = \{c_{kj} \mid E_j : j \in J\}.$$

El problema ahora se reduce a calcular $u(c(i))$ y $u(c(k))$ y decidirnos por la que tenga mayor utilidad.

Consideremos la consecuencia $c = \{c_1 \mid A, c_2 \mid A'\}$ y calculemos $u(c)$. Será fácil generalizar el resultado para cuando tenemos J eventos.

Sea S_1, S_2 eventos auxiliares tales que

$$c_i \sim \{c^* \mid S_i, c_* \mid S_i'\}.$$

Notamos que (vea la figura 3)

$$c \sim \{\{c^* \mid S_1, c_* \mid S_1'\} \mid A, \{c^* \mid S_2, c_* \mid S_2'\} \mid A'\} \sim \{c^* \mid H, c_* \mid H'\}$$

con $H = (A \cap S_1) \cup (A' \cap S_2)$.

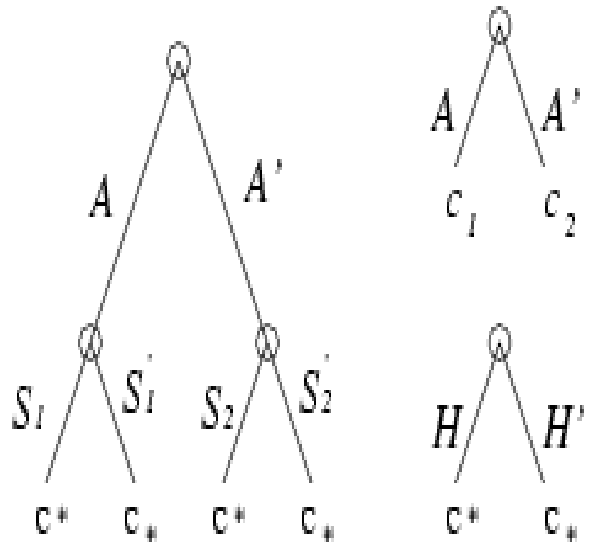


Figure: Árboles de decisión equivalentes a c, $H = (A \cap S_1) \cup (A' \cap S_2)$

Pero entonces tenemos que $u(c) = P(H)$, esto es, *por la independencia de los eventos auxiliares* S_i,

$$u(c) = P(A \cap S_1) + P(A' \cap S_2) = P(A)P(S_1) + P(A')P(S_2)$$

o sea

$$u(c) = P(A)u(c_1) + P(A')u(c_2).$$

Esto es, la utilidad esperada de $\{c_1 \mid A, c_2 \mid A'\}$.

Para decidir entonces cual acción tomar necesitamos calcular las utilidades esperadas de a_i, lo cual es equivalente a calcular la utilidad de $\{c_{ij} \mid E_j : j \in J\}$

$$u^*(a_i) = \sum_{j=1}^{J} P(E_j) u(c_{ij})$$

y decidirnos por aquella acción que maximice la utilidad esperada. (Este es el paradigma Bayesiano.)
Una notación usual, y más compacta, es poner

$$u(a_i, E_j) = u(c_{ij})$$

y evitarnos los c_{ij}'s. También se suele usar en algunos contextos una función de pérdida en vez de utilidad, donde $L(a_i, E_j) = -u(a_i, E_j)$.

Para cualquier $a \in \mathcal{A}$ calculamos

$$u^*(a) = \sum_{j=1}^{J} P(E_j) u(a, E_j)$$

y nos decidimos por a^* tal que

$$u^*(a^*) = \max_{a \in \mathcal{A}} u^*(a).$$

Existen las extensiones obvias para el caso numerable o no numerable de eventos relevantes.

Usando la figura 4 es fácil demostrar que

$$u^*(a) = \sum_{j=1}^{J} P(E_j)u(a, E_j)$$

dado que la acción a es equivalente a la consecuencia $\{c^* \mid H, c_* \mid H'\}$ donde

$$H = \cup_{i=1}^{n} E_i \cap S_i$$

y por lo tanto $u^*(a) = P(H)$. Pero

$$P(H) = \sum_{i=1}^{n} P(E_i \cap S_i) = \sum_{i=1}^{n} P(E_i)P(S_i)$$

es decir

$$P(H) = \sum_{i=1}^{n} P(E_i)u(a, E_j).$$

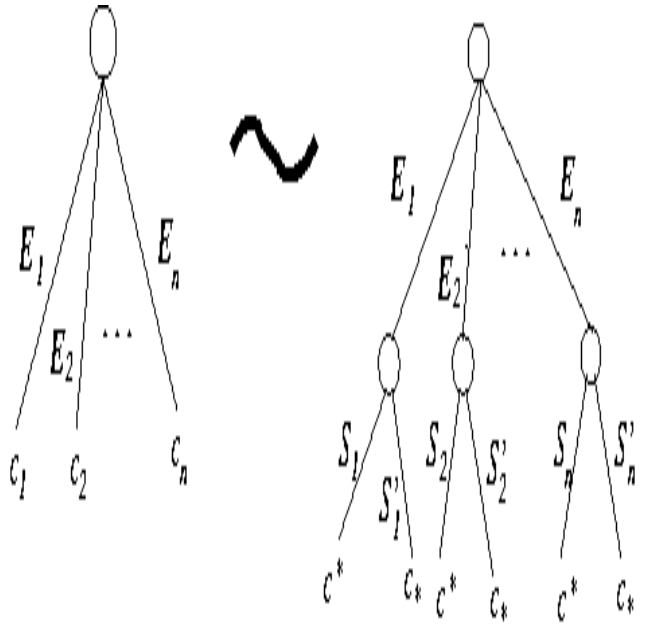


Figure: Árbol de decisión equivalente a *a*.

Ahora, sea X una v.a. en el $[0,1]$ para nuestro espacio de probabilidad, y sea $u(a,x)$ una función de utilidad para X. Con esto vamos a querer decir que existe una función de utilidad $u'(a,E)$ donde

$$u(a,x) = \lim_{\epsilon \to 0} u'(a, \{x-\epsilon < X \leq x+\epsilon\}).$$

Entonces, si tenemos una partición finita
$I = \{0 = I_0 < I_1 < \ldots < I_n = 1$ del $[0,1]$, vemos que

$$u'(a)^* = \sum_{i=0}^{n} u'(a, \{I_{i-1} < X \leq I_i\}) P(\{I_{i-1} < X \leq I_i\}).$$

Pero esto es igual a

$$u'(a)^* = \sum_{i=1}^{n} u'(a, \{l_{i-1} < X \leq l_i\})(F_X(l_i) - F_X(l_{i-1})).$$

De aquí podemos ver que cuando la norma de la partición I tienda a cero vamos a tener que

$$u^*(a) = \int_0^1 u(a, x) dF_X(x)$$

que es la utilidad esperada de la acción a.

Otra vez, la acción que hay que tomar es a^* tal que $u^*(a^*) = \max_{a \in \mathcal{A}} u^*(a)$. Este es el **paradigma Bayesiano:**

Maximizar la utilidad esperada

La inferencia Bayesiana (16)

Como vimos al principio del curso, toda probabilidad es condicional (a las circunstancias, la *agente* que habla etc.) y en realidad deberíamos escribir

$$P(\cdot \mid H), \quad \text{con } H = \text{circunstancias, } \textit{agente} \text{ que habla etc..}$$

Ahora, sea $B \in @$ un evento observable, ¿qué es la probabilidad de $A \in @$ siendo que observamos a (dado) B? Sabemos que esta

probabilidad es $P_B(\cdot) = P(\cdot \mid B)$.

En otras palabras

$$P(A \mid H, B) = \frac{P(A \cap B \mid H)}{P(B \mid H)},$$

o también como

$$P(A \mid H, B) = \frac{P(B \mid H, A)P(A \mid H)}{P(B \mid H)}$$

que es el teorema de Bayes. Note ahora que pasamos de la medida $P(\cdot \mid H)$ a la medida $P(\cdot \mid H, B)$ al observar B.

Analicemos

$$P(A \mid H, B) = \frac{P(B \mid H, A)P(A \mid H)}{P(B \mid H)}.$$

- $P(A \mid H)$ la llamamos probabilidad *a priori* o inicial ("prior", en Inglés) para A.
- $P(A \mid H, B)$ la llamamos probabilidad *a posteriori* o posterior para A, dado que observamos B.
- $P(B \mid H, A)$ es el "modelo" (observacional)...¿como sería la probabilidad de B si supiéramos A?

- $P(B \mid H)$ es la constante de normalización, entendiendo a $P(\cdot \mid H, B)$ como una nueva medida y que

$$P(\cdot \mid H, B) \propto P(B \mid H, \cdot)P(\cdot \mid H).$$

Normalmente omitimos condicionar con respecto a H y dejarlo implícito.

Pruebas de hipótesis

Sea $\theta \in \Theta$ nuestro parámetro de interés y sean

$$H_1 : \theta \in \Theta_1, \quad H_2 : \theta \in \Theta_2$$

dos hipótesis, donde Θ_1 y Θ_2 forman una partición de Θ. En términos de inferencia Bayesiana, dado un modelo $f(X \mid \theta)$ y unas observaciones $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$, ¿qué puede significar "probar" estas hipótesis?

Sea $f(\theta)$ una *a priori* para θ. Calculamos entonces

$$P(H_i) = \int_{\Theta_i} f(\theta \mid \boldsymbol{X}) d\theta$$

y "preferimos", o "los datos apoyan", a H_1 si $P(H_1) > P(H_2)$ (equivalentemente para H_2).
Más aún, podemos tener más de dos hipótesis

$$H_i : \quad \theta \in \Theta_i,$$

y solamente hay que calcular la probabilidad posterior de cada una de estas.

Considerando

$$H_1 : \theta \in \Theta_1, \quad H_2 : \theta \in \Theta_2$$

¿qué haríamos si el problema es *decidirse* por alguna de estas dos hipótesis? Esto es, hay que tomar la decisión, y enfrentar las consecuencias, de que si θ está en Θ_1 o en Θ_2.

La función de pérdida la podemos resumir como:

L	Θ_1	Θ_2
H_1	a	b
H_2	c	d

¿Qué valores razonables daría para a, b, c y d? ¿Qué regla se puede dar para la decisión?

Suponga ahora que θ es una parámetro "continuo", es decir, su *a priori* y su *a posteriori* son absolutamente continuas con respecto a la medida de Lebesgue. ¿Qué sucede si

$$H_1 : \theta = \theta_0, \quad H_2 : \theta \in \Theta - \{\theta_0\}?$$

¡Desde luego que es un absurdo establecer una hipótesis que *a priori* tiene probabilidad cero!

Como se ha intentado resolver el problema, si es que por alguna extraña razón se insiste en investigar sobre la hipótesis H_1, es estableciendo una distribución inicial para θ que tenga un punto de masa θ_0 usando una distribución mixta. (Tarea: Paradoja de Lindley.)

Distribuciones posterior y predictiva

Supongamos que $X_i \mid p \sim Be(p)$ independientes y que la incertidumbre acerca de $p \in [0,1]$ la cuantificamos con $f(p)$ y $p \sim Beta(\alpha, \beta)$ *a priori*. Obtenemos $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ y entonces

$$P(p \leq p_0 \mid \boldsymbol{X}) = \frac{P(\boldsymbol{X} \mid p \leq p_0) P(p \leq p_0)}{P(\boldsymbol{X})}.$$

Pero vemos que

$$P(\boldsymbol{X} \mid p \leq p_0) P(p \leq p_0) = P(\boldsymbol{X}, p \leq p_0) = \int_0^{p_0} f(\boldsymbol{X}, p) dp.$$

Ahora $f(\boldsymbol{X}, p) = f(\boldsymbol{X} \mid p) f(p)$ y entonces

$$P(p \leq p_0 \mid \boldsymbol{X}) \propto \int_0^{p_0} f(\boldsymbol{X} \mid p) f(p) dp.$$

Notando que la parte izquierda es la **distribución posterior para** p, tenemos que su **densidad posterior** es

$$f(p \mid \boldsymbol{X}) \propto f(\boldsymbol{X} \mid p) f(p).$$

Pero

$$f(\boldsymbol{X} \mid p) = p^{\sum_{i=1}^n X_i} (1-p)^{n - \sum_{i=1}^n X_i}$$

y

$$f(p) = B(\alpha, \beta)^{-1} p^{\alpha - 1} (1-p)^{\beta - 1},$$

y entonces

$$f(p \mid \boldsymbol{X}) \propto p^{(\alpha + \sum_{i=1}^n X_i) - 1} (1-p)^{(\beta + n - \sum_{i=1}^n X_i) - 1}.$$

Por lo tanto

$$p \mid \boldsymbol{X} \sim Beta\left(\alpha + \sum_{i=1}^{n} X_i, \beta + n - \sum_{i=1}^{n} X_i\right).$$

Si por ejemplo $X_i \sim N(\theta, \sigma^2)$ (independientes) con σ conocida, y $\theta \sim N(\mu_0, \sigma_0^2)$ *a priori*, entonces

$$f(\theta \mid \boldsymbol{X}) = \frac{f(\theta, \boldsymbol{X})}{f(\boldsymbol{X})}$$

ó

$$f(\theta \mid \boldsymbol{X}) = \frac{f(\boldsymbol{X} \mid \theta) f(\theta)}{\int_{-\infty}^{\infty} f(\boldsymbol{X} \mid \theta) f(\theta) d\theta}$$

o, simplemente

$$f(\theta \mid \boldsymbol{X}) \propto f(\boldsymbol{X} \mid \theta) f(\theta).$$

Esto es:

La posterior es proporcional a la verosimilitud multiplicada por la inicial.

También, si ambas son continuas (tienen una densidad), por la definición de densidad condicional:

$$f(\theta \mid \boldsymbol{X}) = \frac{f(\theta, \boldsymbol{X})}{f(\boldsymbol{X})} = \frac{f(\boldsymbol{X}|\theta) f(\theta)}{f(\boldsymbol{X})}$$

- Demuestre que si $X_i \sim F_\theta$ y $t(\boldsymbol{X})$ (con $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$) es una estadística suficiente (en el sentido usual) entonces

$$f(\theta \mid \boldsymbol{X}) = f(\theta \mid t(\boldsymbol{X})).$$

- Sea $f^{f(\theta)}(\theta \mid \boldsymbol{X})$ la posterior de θ dado $\boldsymbol{X}$; usando como inicial $f(\theta)$, y sea $\boldsymbol{X}'$ observaciones adicionales a $\boldsymbol{X}$: Demuestre que

$$f^{f(\theta)}(\theta \mid \boldsymbol{X}, \boldsymbol{X}') = f^{f^{f(\theta)}(\theta|\boldsymbol{X})}(\theta \mid \boldsymbol{X}').$$

Predicción

The future ain't what it used to be. Yogi Berra quotes (American professional Baseball Player and Manager. b.1925)[2]

Suponga que $X_i \sim F_\theta$ y que tenemos una inicial $f(\theta)$ para θ y una muestra independiente $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$. Ahora suponga que vamos a observar una nueva variable independiente $Y \sim F_\theta$; ¿cómo podemos predecir dicha variable?

[2] La frase se la han achacado a muchos, como Felipe Gonzalez, Presidente del gobierno Español 1982-1996, etc.

Predicción

Lo que requerimos es $f(Y \mid \boldsymbol{X})$, y esto lo podemos calcular como:

$$f(Y \mid \boldsymbol{X}) = \int f(Y \mid \theta, \boldsymbol{X}) \pi(\theta \mid \boldsymbol{X}) d\theta,$$

pero como Y es una nueva observación *condicionalmente* independiente de $\boldsymbol{X}$ tenemos que

$$f(Y \mid \theta, \boldsymbol{X}) = f(Y \mid \theta)$$

que representa simplemente nuestro modelo F_θ. Entonces

$$f(Y \mid \boldsymbol{X}) = \int f(Y \mid \theta) \pi(\theta \mid \boldsymbol{X}) d\theta.$$

Es importante notar que esta fórmula solo se aplica para muestreo independiente.

Ejemplo

Imagínese que en un barrio visitamos m casas al azar cada día y X_i es el número de nuestros artículos que vendemos. Imagínese también que ya llevamos n días, $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$, independientes entre si (y los demás supuestos generosos para poder resolver el problema...¡es un problema clásico de salón donde las cosas son demasiado bellas!). Por nuestra experiencia sabemos que de cada diez casas que visitamos esperamos vender entre 0 y 4 artículos, siendo un promedio como de 1.

Si visitar una casa nos cuesta a pesos, cada artículo lo vendemos a b_1 pesos y lo compramos a b_2 pesos, ¿nos conviene el día de mañana salir a vender en dicho barrio?

Suponga que $X_i \sim Bi(m, \theta)$ independientes.

¿qué *a priori* conviene, es útil, se puede usar en este caso?

¿como se resuelve el problema?

Suponga también que *a priori* $\theta \sim$ *Beta*(α, β). ¿Cómo se distribuye $\theta \mid \boldsymbol{X}$?

Si observamos Y del mismo modelo, ¿Cómo se distribuye $Y \mid \boldsymbol{X}$? ¿Qué valor esperamos para θ?...¿y para Y?

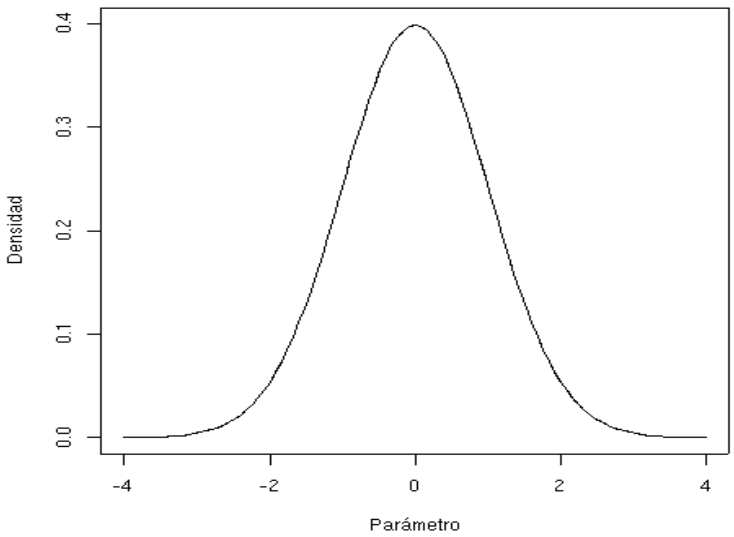
Estimación puntual y de intervalo

El objeto principal de la inferencia Bayesiana es la distribución posterior de los parámetros de interés involucrados. Por ejemplo, si tenemos dos parámetros θ_1 y θ_2, y los datos $\boldsymbol{X}$, tenemos que encontrar la distribución posterior

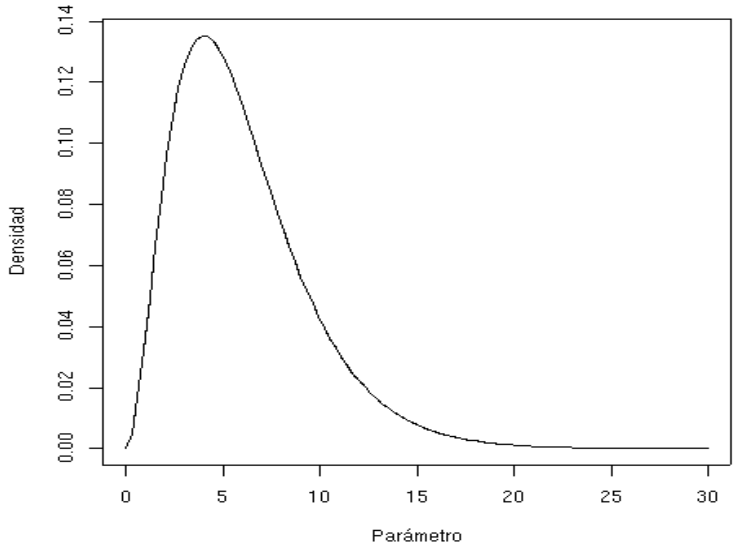
$$f(\theta_1, \theta_2 \mid \boldsymbol{X}).$$

Ahora, ¿qué haríamos si el parámetro de interés es solamente θ_1?

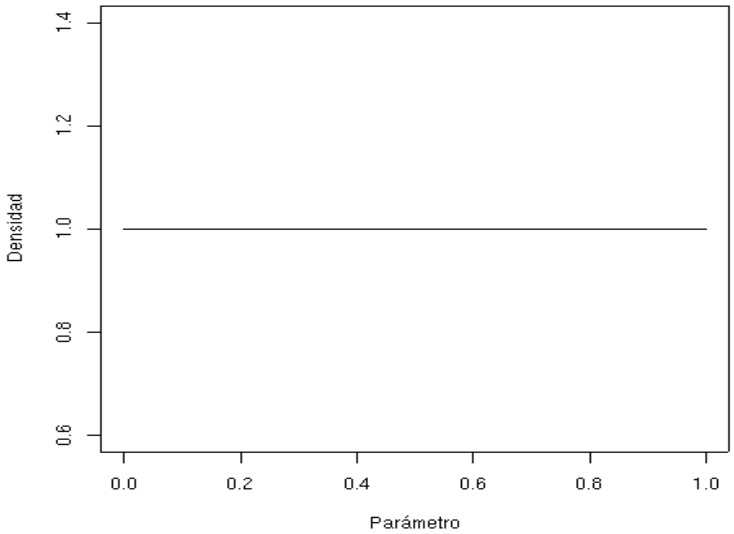
Suponiendo que tenemos ahora la distribución posterior $f(\theta \mid \boldsymbol{X})$, lo único que nos resta hacer es *reportarla* como el resultado de nuestra inferencia: ¿Como la reportaría? (vea los ejemplos en la figura 5).


Densidad
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
Parámetro

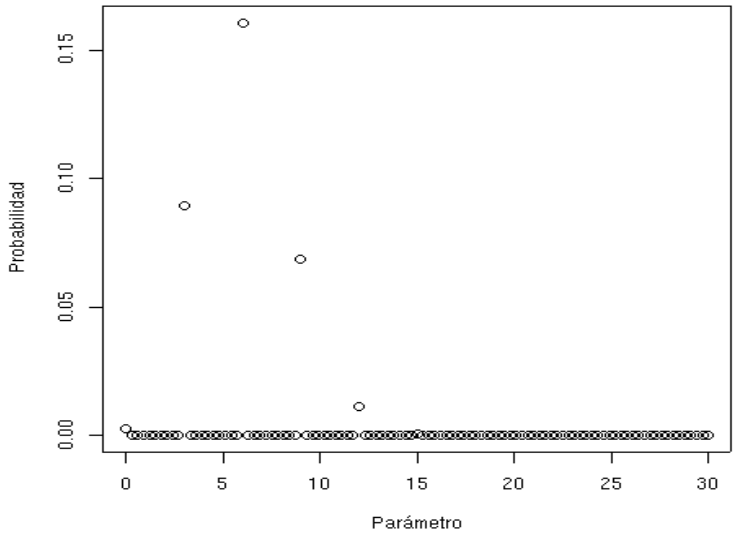
(a)


Densidad
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0
5
10
15
20
25
30
Parámetro

(b)


Densidad
0.6
0.8
1.0
1.2
1.4
0.0
0.2
0.4
0.6
0.8
1.0
Parámetro

(c)


Probabilidad
0.00
0.05
0.10
0.15
0
5
10
15
20
25
30
Parámetro

(d)

El concepto de “estimación” dentro de la estadística Bayesiana, no se entiende más que como resúmenes de la posterior con la que se cuenta[3]. Por lo tanto, el concepto de estimación puntual lo podemos ver como resumir una distribución con un solo punto (por absurdo que en esta perspectiva parezca).

Si se trata de decidir sobre un punto del espacio paramétrico Θ podemos entonces plantearnos el problema como uno de decisión. Nuestro espacio de acciones es $\mathcal{A} = \Theta$ y necesitaríamos una función de utilidad (o pérdida):

$$u(a, \theta), \quad a \in \mathcal{A}, \quad \theta \in \Theta.$$

El estimador puntual sería entonces

$$\hat{a} \text{ tal que } u^*(\hat{a}) = \sup_{a \in \mathcal{A}} u^*(a).$$

[3] Desde luego que hay resúmenes buenos y resúmenes malos.

Ejemplo

Podemos usar la distancia cuadrática del estimador propuesto a a el valor de θ como función de pérdida:

$$L(a, \theta) = k(a - \theta)^2.$$

En este caso $L^*(a) = E[k(a - \theta)^2]$ y sabemos que esto es minimizado cuando $a = E[\theta]$.

Es común que en problemas concretos se diga que "es peor la sobre estimación que la sub estimación". ¿Como se podría abordar este problema en esta perspectiva? ¿qué pasa cuando θ es discreta?

Note que la pérdida de arriba la podemos expresar como:

$$L(a, \theta) = g(a - \theta);$$

donde $g(x) = kx^2$ ¿qué otras opciones propone para g?

Se puede tomar como estimador el máximo de la distribución posterior, este se llama el estimador MAP (*Maximum a posteriori*).

Pérdida no simétrica

Sea la siguiente pérdida linea (L_1) no simétrica

$$l(a,\theta) = \begin{cases} l_1(\theta - a) & \text{if } a < \theta \text{ (sub-estimar)} \\ l_2(a - \theta) & \text{if } a \geq \theta \text{ (sobre-estimar)} \end{cases}$$

con $l_1, l_2 > 0$. Note que

$$l^*(a) = \int_{-\infty}^{a} l(a,\theta)\pi(\theta|x)d\theta + \int_{a}^{\infty} l(a,\theta)\pi(\theta|x)d\theta$$

or

$$l^*(a) = \int_{-\infty}^{a} l_2(a-\theta)\pi(\theta|x)d\theta + \int_{a}^{\infty} l_1(\theta - a)\pi(\theta|x)d\theta.$$

Pérdida no simétrica

Regla integral de Leibniz:

$$\frac{d}{dx}\int_{-\infty}^{x} f(x,t)\mu(dt) = f(x,x) + \int_{-\infty}^{x} \frac{\partial}{\partial x} f(x,t)\mu(dt)$$

Entonces, notando que $l(a,a) = 0$:

$$\frac{d}{da} l^*(a) = \int_{-\infty}^{a} l_2 \pi(\theta|x) d\theta - \int_{a}^{\infty} l_1 \pi(\theta|x) d\theta.$$

Igualando a cero, con $p(a) = \int_{-\infty}^{a} \pi(\theta|x) d\theta$

$$p(a^*) = \frac{1}{\frac{l_2}{l_1} + 1}.$$

Entonces, a^* es el cuantil de probabilidad $p(a^*)$. Solamente depende de $\frac{l_2}{l_1}$ y si $l_1 = l_2$ a^* es la mediana. Sino, es un cuantil para abajo o para arriba.

Estimación por Intervalo

En relación con la estimación por intervalo, aquí el problema es resumir una distribución con un conjunto en el espacio paramétrico Θ. Una función de pérdida razonable (para θ continua) es la siguiente:

$$L(A, \theta) = \lambda(A) + kI_{A^c}(\theta)$$

donde $\lambda(A)$ es la medida de Lebesgue de A y $I_{A^c}(\theta)$ es la función indicadora. Esto es, penalizamos con respecto al tamaño de A y agregamos una penalización 0-1 dependiendo de si θ está o no en el conjunto seleccionado.

Vemos que

$$L^*(A) = E_{\theta|x}[L(A,\theta)] = \lambda(A) + kP_{\theta|x}(A^c).$$

De aquí vemos que $L^*(A) \geq 0$ y que $L^*(A)$ deben tener un ínfimo. Usando el teorema de descomposición de Hahn es fácil probar que $L^*(A)$ debe tener un mínimo A^*.

(Tomemos la medida con signo $\mu(E) = \lambda(E) - kP_{\theta|x}(E)$; el ínfimo de esta es igual a A^*. Por el teorema de descomposición de Hahn existe una partición A, B de Ω tal que $\mu(E \cap A), -\mu(E \cap B) \geq 0$. Como $\mu(E) = \mu(E \cap A) + \mu(E \cap B)$, el valor más pequeño posible de μ sería cuando $\mu(E \cap A) = 0$ y $\mu(E \cap B)$ es lo más negativo posible. Dicho valor se alcanza con $E = B$, o sea $A^* = B$.)

Sea $P_{\theta|x}(A^*) = 1 - \alpha$ y sea $B \in \mathcal{A}$ tal que $P_{\theta|x}(B) = 1 - \alpha$, entonces $L^*(B) = \lambda(B) + k\alpha$, y entonces $\lambda(B) \geq \lambda(A^*)$. Entonces A^* es un conjunto con probabilidad $1 - \alpha$ y con longitud (medida de Lebesgue) mínima y entonces es de la forma $A(a) = \{t \in \mathcal{R} : f_{\theta|\boldsymbol{x}}(t \mid x) > a\}$.

Teorema

Utilizando la función de pérdida

$$L(A, \theta) = \lambda(A) + kI_{A^c}(\theta)$$

tenemos que

$$A^* = \{t \in \mathcal{R} : f_{\theta|\boldsymbol{x}}(t \mid x) > k^{-1}\}$$

excepto por conjunto de probabilidad cero.

Suponemos primero que $f_{\theta|\boldsymbol{x}}(t \mid x)$ es absolutamente continua, con una sola moda. Sea $g(a) = L(A(a))$, como vimos arriba lo que tenemos que hacer es encontrar el mínimo de g. Como g tiene una sola moda entonces funciones $\theta_1(a)$ y $\theta_2(a)$ tales que

$$g(a) = \theta_2(a) - \theta_1(a) + k\left(\int_{-\infty}^{\theta_1(a)} f_{\theta|\boldsymbol{x}}(t \mid x)dt + \int_{\theta_2(a)}^{\infty} f_{\theta|\boldsymbol{x}}(t \mid x)dt\right).$$

Note que $f_{\theta|\boldsymbol{x}}(\theta_i(a) \mid x) = a$. Al derivar g obtenemos:

$$g'(a) = \theta_2'(a) - \theta_1'(a) + ak\left(\theta_2'(a) - \theta_1'(a)\right).$$

Igualando a cero obtenemos $\left(\theta_2'(a) - \theta_1'(a)\right)(1 - ak) = 0$ y entonces $a = k^{-1}$ si $\theta_2'(a) \neq \theta_1'(a)$. Si $\theta_2'(a) = \theta_1'(a)$ para toda a, entonces $\theta_2(a) - C = \theta_1(a) + C$ y la función debe ser simétrica alrededor de la moda. La expresión para g se puede modificar usando solo $\theta_1(a)$ y se llega al mismo resultado. Así se puede generalizar el resultado para múltiples modas.

Definición

Los conjuntos de la forma A^* son llamados conjuntos de máxima densidad posterior o HPD (Hieghest Posterior Density), vea la figura 6.

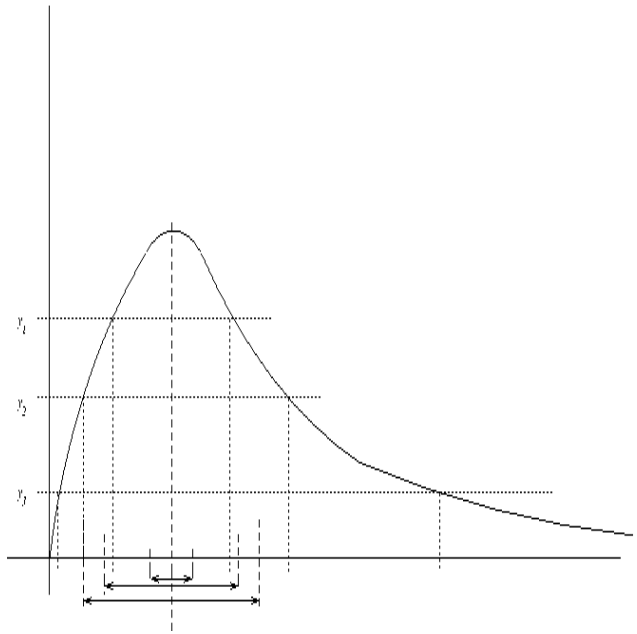


Figure: Intervalos (conjuntos) de máxima densidad posterior (HPD's). Note que no necesariamente son de la forma $\theta_0 \pm \sigma$

Análisis conjugado

En esta sección estudiaremos ejemplos del uso del análisis conjugado en Bayesiana. La idea principal es muy simple. Si tenemos para la inicial de θ, $\theta \sim F_{\alpha_0}$, entonces para la posterior de θ, $\theta \sim F_{\alpha_p}$.

Esto es, la inicial y la posterior están en la misma familia paramétrica y lo único necesario para calcular la posterior es establecer los parámetros α_p como función de los datos y de α_0.

Ejemplo

Sea $X_i \sim Be(\theta)$ y $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ independientes. Entonces

$$f(\boldsymbol{X} \mid \theta) = \prod_{i=1}^{n} \theta^{X_i}(1-\theta)^{1-X_i}$$

o

$$f(\boldsymbol{X} \mid \theta) = \theta^{\sum_{i=1}^{n} X_i}(1-\theta)^{n-\sum_{i=1}^{n} X_i}.$$

Esto sugiere que si establecemos $\theta \sim Beta(\alpha_0, \beta_0)$ tendremos una *a priori* conjugada. Vemos pues que

$$f(\theta \mid \boldsymbol{X}) \propto \theta^{\alpha_0+\sum_{i=1}^{n} X_i-1}(1-\theta)^{\beta_0+n-\sum_{i=1}^{n} X_i-1}.$$

Esto es $\theta \sim Beta(\alpha_p, \beta_p)$ con

$$\alpha_p = \alpha_0 + \sum_{i=1}^{n} X_i, \quad \beta_p = \beta_0 + n - \sum_{i=1}^{n} X_i.$$

Ejemplos

- $X_i \sim Exp(\theta)$. $\theta \sim Ga(\alpha, \beta)$, $f(X_i \mid \theta) = \theta e^{-\theta X_i}$ y
  $f(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}$.

- $X_i \sim Po(\theta)$. $\theta \sim Ga(\alpha, \beta)$, $f(X_i \mid \theta) = \frac{\{\theta\}^{X_i}}{X_i!} e^{-\theta}$ y
  $f(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}$.

- $X_i \sim U(0, \theta)$. $\theta \sim Pareto(\alpha, \beta)$. $f(X_i \mid \theta) = I_{[0,\theta]}(X_i)\theta^{-1}$ y
  $f(\theta) = \alpha\beta^\alpha \theta^{-(\alpha+1)} I_{[\beta,\infty)}(\theta)$.

Note que si $f(\theta \mid \alpha_i)$ son conjugadas dado el modelo $f(\boldsymbol{X} \mid \theta)$ entonces

$$f(\theta) = \sum_{i=1}^{k} w_i f(\theta \mid \alpha_i)$$

con $\sum_{i=1}^{k} w_i = 1$, es también una conjugada para el mismo modelo.

Esto nos da una gran facilidad pues con mezclas de conjugadas podemos generar una gran diversidad de distribuciones (iniciales). Es el caso, por ejemplo de muestreo Bernoulli, tenemos que cualquier distribución en el $[0, 1]$ puede ser aproximada arbitrariamente por una mezcla de Betas. Aún cuando las conjugadas puedan parecer restrictivas, al utilizar mezclas de estas obtenemos una gran flexibilidad para definir *a priori*es.

Es importante señalar que no existe *la* familia conjugada, para un cierto modelo, pues esta no es única. Y que, por otro lado, el proponer trabajar dentro de una familia conjugada es solo una conveniente práctica común y de ninguna manera un procedimiento estructural o fundamental del análisis Beyasiano.

Ejemplos, análisis NO conjugado

Por ejemplo, datos normales (varianza conocida), pero se sabe que la media es mayor que cero aun cuando cercana a cero. Se puede usar una Gama.

Se tiene un tratamiento clínico que fué probado en dos poblaciones A y B, con respuestas 1, “éxito”, 0 “fracaso”. Se sabe, sin embargo, que la población A tiene una condición mas grave que la población B y que por lo tanto es más fácil que en la población B cualquier tratamiento tenga éxito. Por otra parte sabemos que el tratamiento estandard tiene una probabilidad de éxito de p_A para la población A y p_B para la población B. ¿como haría el análisis estadístico (muestra aleatoria)? ¿qué tratamiento se debe de aplicar?

Intercambiabilidad y visión de de Finetti (2)

Digamos que tenemos una secuencia $X_1, X_2, \ldots$ de variables aleatorias y que el índice de estas es irrelevante, en el sentido que para cualquier subconjunto finito

$$p(X_1, X_2, \ldots, X_n) = p(X_{\sigma(1)}, X_{\sigma(2)}, \ldots, X_{\sigma(n)}),$$

donde σ es una permutación. Se dice entonces que la secuencia de variables aleatorias es infinitamente intercambiable.

Digamos que tenemos la respuesta a una dosis administrada a individuos de la misma edad sexo etc. Posiblemente se pueda pensar que la secuencia es intercambiable.

Ahora, si tenemos varios tipos de dosis, hombres y mujeres, varias edades etc. es probable que en ese caso no haya Intercambiabilidad. Sin embargo, para subconjuntos de estas variables (covariables) sí tengamos Intercambiabilidad.

Teorema

Sea $X_1, X_2, \ldots$ una secuencia infinita mente intercambiable de variables con $X_i = 0, 1$. Entonces existe una medida Q en $[0, 1]$ tal que

$$p(S_n = s) = \int_0^1 C_s^n \theta^s (1-\theta)^{n-s} Q(d\theta)$$

donde $S_n = \sum_{i=1}^n X_i$. (Teorema de representación de De Finetti, 1931.)

Hay muchos teoremas de este tipo, mucho más generales.

1. Podemos pensar a las X_i's *como si* fuesen independientes dado el parámetro θ.
2. θ tiene una distribución de probabilidad que puede ser interpretada como una inicial.
3. Podemos interpretar a $C_s^n \theta^s (1-\theta)^{n-s}$ como la distribución de S_n *condicionada* a θ, siendo $p(S_n = s)$ obtenida por probabilidad total ($p(S_n = s) = \int_0^1 p(S_n = s \mid \theta) dQ(\theta)$).
4. Entonces $S_n \mid \theta \sim Bi(n, \theta)$ y por lo tanto

$$X_i \mid \theta \sim Be(\theta)$$

Esto es, *las X_i son condicionalmente independientes*, dada su probabilidad (¡común!) de éxito, solo por asumir que son intercambiables.

Si graficamos (n, S_n) obtenemos una serie de trayectorias que empiezan en $(0, 0)$. Si tomamos hasta $n = N$ la intercambiabilidad nos dice que todas las trayectorias que terminan en un mismo punto $S_N = s$ tienen la misma probabilidad.

Vamos a clacular la probabilidad condicional $P(S_n = h \mid S_N = H)$; esto sería como sacar, sin repocisión, N bolas de una urna con H bolas blancas (1) y $N - H$ bolas negras (0), sin repocisión. Esto sería como tomar una muestra de una distribución hipergeométrica, y desde luego $S_N = H$.

O sea, si $X_1, X_2, \ldots, X_N$ y sabemos que $S_N = H$, entonces si $S_n = \sum_{i=1}^{n} X_i = h$ $(n \leq N)$

$$P(X_1, X_2, \ldots, X_n \mid S_N = H) = P(S_n = h \mid S_N = H) = p(h \mid N, H, n),$$

donde $p(h \mid N, H.n)$ es la función de probabilidad hipergeométrica con N bolas y H bolas blancas (1) y n intentos. (Tenemos H 1's para poner en total, de un total de N o's y 1's, de las cuales sacamos h 1's en n intentos.)

Usando probabilidad total tenemos que

$$P(S_n = h) = \sum_{H} P(S_n = h \mid S_N = H) P(S_N = H).$$

ó

$$P(S_n = h) = \sum_{H} p(h \mid N, H, n) P(S_N = H).$$

Ahora, para un proceso que no termine lo podemos ir aproximando con procesos a N pasos, pero estos son mezclas de hipergeométricos. Si tomo $F_N(\theta) = P(S_N \leq N\theta) = P(S_N/N \leq \theta)$ la mezcla que necesito es

$$P(S_n = h) = \int_0^1 p(h \mid N, N\theta, n) dF_N(\theta).$$

Ahora $p(h \mid N, N\theta, n)$ tiende a $Bi(n, \theta = N\theta/N)$, cuando $N \to \infty$ y $F_N(\theta) \to Q(\theta) = \lim_{N\to\infty} P(S_n/n \leq \theta)$, cuando N tiende a infinito. Esta demostración es un poco informal pero ilustra el procedimiento general. Es la que aparece en De Finetti (1970), *Theory of Probability, Volumen 2*, p.217–218.

Teorema General de representación

Teorema

Sea $X_1, X_2, \ldots$ una secuencia infinita mente intercambiable de variables aleatorias reales con medida conjunta P, entonces existe una medida Q en el especio de distribuciones de los reales tal que:

$$P(X_1, X_2, \ldots, X_n) = \int \prod_{i=1}^{n} F(X_i) dQ(F)$$

con $Q(F) = \lim_{n\to\infty} P(F_n = F)$, y F_n es la función de distribución empírica definida por $X_1, X_2, \ldots, X_n$.

No hay más condiciones de regularidad: Teo 3.1 (p. 20) de Aldous D.J. (1985) Exchangeability and related topics. In: Hennequin P.L. (eds) École d'Été de Probabilités de Saint-Flour XIII — 1983. Lecture Notes in Mathematics, vol 1117. Springer, Berlin, Heidelberg

Teoría de la Probabilidad de de Finetti

De Finetti presenta un tratamiento un poco diferente sobre la teoría de la probabilidad condicional (o subjetivista). El fundamento es el siguiente.

Sea X una variable aleatoria que representa una ganancia (o pérdida si es negativa). ¿Con qué pérdida segura se puede equiparar a X? A esta pérdida De Finetti le llama $P(X)$. Cuando X es solo un evento (ganancia 0,1) entonces hablamos de probabilidad y sino de previsión (esperanza; a De Finetti no le parece bueno el término de esperanza, pero es lo mismo).

De Finetti establece las propiedades de $P(X)$ justificadas en términos de pérdidas para llegar a los axiomas de probabilidad usuales. (Vea De Finetti (1970), *Theory of Probability, Volumen 1*, Sección 3.1.4, p.72–75.)

Sin embargo, algo muy particular en de De Finetti y es que no acepta la σ–aditividad y trabaja siempre con particiones finitas de eventos (vea sección 3.11 del mismo libro).

Análisis asintótico (1)

En el análisis Bayesiano tenemos teoremas que nos hablan de la convergencia de las distribuciones posteriores según crece el tamaño de muestra n. Presentamos ahora uno muy sencillo

Teorema

Sea $\boldsymbol{X}_n = (X_1, X_2, \ldots, X_n)$ *independientes de un modelo paramétrico* $p(x \mid \theta)$, *donde* $\Theta = \{\theta_1, \theta_2, \ldots\}$ *e inicial* $p(\theta_i) = p_i$. *Suponga que el valor "real" de* θ *es* $\theta_r \in \Theta$, *o sea que* $X_i \sim p(x|\theta_r)$, *y que*

$$\int p(x \mid \theta_r) \log \frac{p(x \mid \theta_r)}{p(x \mid \theta_i)} dx > 0,$$

$i \neq r$, *entonces*

$$\lim_{n\to\infty} p(\theta_r \mid \boldsymbol{X}_n) = 1, \quad \lim_{n\to\infty} p(\theta_i \mid \boldsymbol{X}_n) = 0.$$

Tenemos que

$$
\begin{aligned}
p(\theta_i \mid \boldsymbol{X}_n) &= \frac{p_i\{p(\boldsymbol{X}_n \mid \theta_i)/p(\boldsymbol{X}_n \mid \theta_r)\}}{\sum_{j=n}^{\infty} p_j\{p(\boldsymbol{X}_n \mid \theta_j)/p(\boldsymbol{X}_n \mid \theta_r)\}} \\
&= \frac{\exp\{\log p_i + S_i\}}{\sum_{j=n}^{\infty} \exp\{\log p_j + S_j\}},
\end{aligned}
$$

donde

$$
S_i = \sum_{j=1}^{n} \log \frac{p(X_j \mid \theta_i)}{p(X_j \mid \theta_r)}.
$$

Dadas las θ_i's, la suma

$$S_i = \sum_{j=1}^{n} \log \frac{p(X_j \mid \theta_i)}{p(X_j \mid \theta_r)}.$$

representa una suma de términos independientes e idénticamente distribuidos y por la ley de los grandes números tenemos que

$$\lim_{n\to\infty} \frac{S_i}{n} = \int p(x \mid \theta_r) \log \frac{p(x \mid \theta_i)}{p(x \mid \theta_r)} dx.$$

El lado derecho de esta expresión es negativa para $i \neq r$ y cero para $i = r$. Entonces $S_r \to 0$ y $S_i \to -\infty$, para $i \neq r$.

Note que el teorema anterior es válido para cualquier distribución inicial de θ, siempre y cuando $p_r \neq 0$ (el soporte incluya al valor real). Los teoremas asintóticos nos dicen, en general, que las posteriores tenderán a distribuciones cada vez más concentradas independientemente de la inicial usada.

Como consecuencia, dos usuarios con diferentes opiniones iniciales, sus posteriores coincidirán, aproximadamente, después de un tamaño de muestra grande. “La posterior es consistente y robusta”.

Como ejemplo de los teoremas asintóticos tenemos que, hablando en general para parámetros continuos

$$f(\theta \mid \boldsymbol{X}_n) \propto f(\theta) \prod_{i=1}^{n} f(X_i \mid \theta) = \exp\{\log f(\theta) + \log f(\boldsymbol{X}_n \mid \theta)\}.$$

Usando $\log f(\theta) = \log f(m_0) - \frac{1}{2}(\theta - m_0)^2 h_0$ y
$\log f(\boldsymbol{X}_n \mid \theta) = \log f(\boldsymbol{X}_n \mid \hat{\theta}_n) - \frac{1}{2}(\theta - \hat{\theta}_n)^2 h(\hat{\theta}_n)$, donde m_0 es el máximo de la *a priori* y $\hat{\theta}_n$ es el máximo de la verosimilitud, obtenemos

$$f(\theta \mid \boldsymbol{X}_n) \propto \exp\{-\frac{1}{2}(\theta - m_0)^2 h_0 - \frac{1}{2}(\theta - \hat{\theta}_n)^2 h(\hat{\theta}_n)\} + R_n,$$

o

$$f(\theta \mid \boldsymbol{X}_n) \propto \exp\{-\frac{1}{2}(\theta - m_n)^2 h_n\} + R_n$$

donde $h_n = h_0 + h(\hat{\theta}_n)$ y $m_n = \frac{h_0 m_0 + h(\hat{\theta}_n)\hat{\theta}_n}{h_n}$.

Usando condiciones de regularidad para la aproximación de Taylor podemos ver que la posterior se va a aproximar a una normal, teniendo como media el estimador máximo verosímil (pues en general $h(\hat{\theta}_n)$ tiende a ∞).

Hay muchos teoremas de este tipo. Uno general existe para la familia exponencial usando como *a priori* una conjugada canónica (ver Bernardo y Smith, 1994, p.293).

Se garantiza de que si el modelo verdadero f_0 está el soporte de "Kullback-Leiber" de la apriori, entonces la posterior es consistente y "converge" a f_0.

f_0 está en el soporte K-L de la a priori si
$\pi\left(\{g : \int f_0 log(f_0/g)\lambda < \epsilon\}\right) > 0$ para todo $\epsilon > 0$. En el caso paramétrico queda $\pi\left(\{\theta : \int f(x|\theta_0) log(f(x|\theta_0)/f(x|\theta))\lambda(dx) < \epsilon\}\right) > 0$.

Vea Teorema 4.4.2, Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*, Springer: New York.

Aproximaciones numéricas (no MCMC) (2)

Como hemos visto, el objeto principal en el análisis Bayesiano (paramétrico) es obtener la posterior

$$f(\theta \mid \boldsymbol{X}) = \frac{f(\boldsymbol{X} \mid \theta) f(\theta)}{\int f(\boldsymbol{X} \mid \theta) f(\theta) d\theta}.$$

Esto presenta un problema básico: ¿cómo evaluar la integral del dividendo? o ¿cómo encontrar la constante de proporcionalidad?
A esto se reduce la parte operativa del análisis Bayesiano, después de haber definido un modelo y una *a priori*. Más allá de esto, uno puede interesarse en marginales de la posterior, que involucrarían más integraciones de la posterior.

Cuadratura

Se pueden intentar métodos tradicionales de cuadratura (análisis numérico), pero no los vamos a estudiar aquí.

Aproximación de Laplace

Un punto importante en el análisis Bayesiano es encontrar momentos *a posteriori* como $E[g(\theta) \mid \boldsymbol{X}]$. Esto se calcularía como

$$E[g(\theta) \mid \boldsymbol{X}] = \frac{\int g(\theta) f(\boldsymbol{X} \mid \theta) f(\theta) d\theta}{\int f(\boldsymbol{X} \mid \theta) f(\theta) d\theta}$$

lo que también podemos escribir como

$$E[g(\theta) \mid \boldsymbol{X}] = \frac{\int \exp\{-nh^*(\theta)\}d\theta}{\int \exp\{-nh(\theta)\}d\theta}.$$

con

$$-nh(\theta) = \log f(\theta) + \log f(\boldsymbol{X} \mid \theta)$$

y

$$-nh^*(\theta) = \log g(\theta) + \log f(\theta) + \log f(\boldsymbol{X} \mid \theta).$$

Usando $\hat{\theta}$ como el máximo de $-h(\theta)$ y θ^* como el máximo de $-h^*(\theta)$, y $\hat{\sigma}$ y σ^* como el valor de las segundas derivas en los máximos elevadas a la $-\frac{1}{2}$, tenemos que

$$-nh(\theta) \approx -nh(\hat{\theta}) - \frac{n}{2\hat{\sigma}^2}(\theta - \hat{\theta})^2$$

y por lo tanto

$$\int \exp\{-nh(\theta)\}d\theta \approx \sqrt{2\pi}\hat{\sigma}^{-1/2} \exp\{-nh(\hat{\theta})\}.$$

Y equivalentemente para $-nh^*(\theta)$. Entonces podríamos aproximar $E[g(\theta) \mid \boldsymbol{X}]$ con

$$\frac{\sigma^*}{\hat{\sigma}} \exp\{-n[h^*(\theta^*) - h(\hat{\theta})]\}.$$

Remuestreo relevante (importance sampling)

Considerando la integral de una función $f(x)$ y usando una densidad $g(x)$ vemos que

$$\int f(x)dx = \int \frac{f(x)}{g(x)} g(x)dx = E_G\left[\frac{f(x)}{g(x)}\right]$$

donde G es la distribución de una v.a. con densidad g. Vemos ahora un proceso para estimar $\int f(x)dx$. Si simulamos x_i con distribución G tenemos que

$$\int f(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} \frac{f(x_i)}{g(x_i)}.$$

Si tomamos f como $p(\boldsymbol{X} \mid \theta)p(\theta)$ y a g como $p(\theta)$ estimaríamos la constante de normalización con

$$\frac{1}{n}\sum_{i=1}^{n} p(\boldsymbol{X} \mid \theta_i).$$

Muestreo–Remuestreo relevante (SIR)

La idea es simple. Suponga que queremos simular de una densidad $f(\theta)$ que solo está determinada proporcionalmente. Suponga además que tenemos otra densidad $g(\theta)$ y que existe un valor M tal que

$$f(\theta) \leq Mg(\theta).$$

Entonces podemos simular un punto dentro de la gráfica de f simulando un valor θ_i con densidad g y otro $y = uMg(\theta_i)$, donde $u \sim U(0,1)$. Si $y \leq f(\theta_i)$ entonces (θ_i, y) es una simulación uniforme de un punto dentro de la gráfica de f y por lo tanto θ_i se distribuye con densidad f (normalizada). Vea la figura 7.

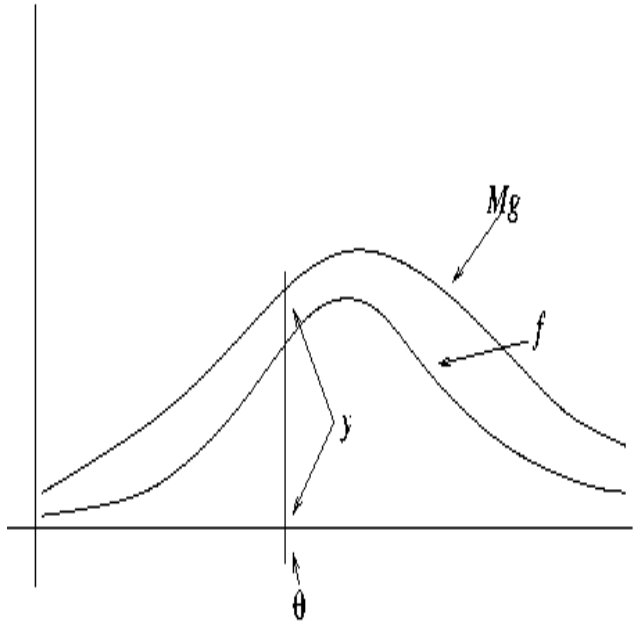


Figure: Simulación por el método de rechazo.

Usando $f_{\boldsymbol{X}}(\theta) = p(\boldsymbol{X} \mid \theta)p(\theta)$, $g(\theta) = p(\theta)$, $M = p(\boldsymbol{X} \mid \hat{\theta})$ y $\hat{\theta}$ el máximo de la verosimilitud, una versión en términos de remuestreo para el teorema de Bayes quedaría como:

Partiendo de una muestra θ_i de la *a priori* $p(\theta)$, aceptar *a posteriori* θ_i si $uMp(\theta_i) \leq f_{\boldsymbol{X}}(\theta_i)$, o sea, aceptar θ_i con probabilidad

$$\frac{p(\boldsymbol{X} \mid \theta_i)}{p(\boldsymbol{X} \mid \hat{\theta})}.$$

(Particle filtering.)

Análisis de referencia (5)

En esta última sección del capítulo de inferencia discutimos un aspecto práctico en el análisis Bayesiano: ¿Qué'e hacer cuando hay poca o nula información disponible acerca de un parámetro? ¿como establecer la *a priori* en ese caso? ¿como establecer una *a priori* que represente una información "vaga" acerca de un parámetro?

Lo que ganamos con los datos

Para empezar bien, lo que tenemos que hacer es definir con claridad lo que entendemos por “vago” o no-informativo. Supongamos que tenemos un modelo observacional $f(\boldsymbol{X}_n \mid \theta)$ (n observaciones independientes) con el parámetro θ. Suponga también que tenemos una función de utilidad $u(\theta, a)$ y una *a priori* $f(\theta)$. Sean a_0^* y a_n^* las decisiones óptimas *a priori* y *a posteriori*.

Lo que ganamos al observar $\boldsymbol{X}_n$ indudablemente sería $u^*_{\boldsymbol{X}_n}(a^*_n) - u^*_0(a^*_0)$. Ahora, haciendo este análisis previamente, antes de obtener los datos, lo que *esperamos* de utilidad de obtener n datos, usando $f(\theta)$ como inicial, es

$$\delta(n, f(\theta)) = \int f(\boldsymbol{X}_n) u^*_{\boldsymbol{X}_n}(a^*_n) d\boldsymbol{X}_n - u^*_0(a^*_0)$$

o

$$\delta(n, f(\theta)) = \int f(\boldsymbol{X}_n) \int u(\theta, a^*_n) f(\theta \mid \boldsymbol{X}_n) d\theta d\boldsymbol{X}_n - \int u(\theta, a^*_0) f(\theta) d\theta,$$

donde $f(\boldsymbol{X}_n) = \int f(\boldsymbol{X}_n \mid \theta) f(\theta) d\theta$ es la predictiva *a priori* (muestreo independiente).

Ahora, el punto es fijarnos en $\delta(n, f(\theta))$ también como función de $f(\theta)$. Suponga que hacemos tender n a infinito, entonces $\delta(\infty, f(\theta))$ sería la influencia de la *a priori* con respecto a información absoluta. Si lo que queremos es que la información *a posteriori* sea, en la medida de lo posible, no influenciada por la *a priori*, lo que buscamos es una *a priori* f^* tal que

$$\delta(\infty, f^*(\theta)) = \sup_{f \in \mathcal{D}} \delta(\infty, f(\theta));$$

esto es, que maximice la información contenida en los datos.

Tres problemas son aparentes aquí:

1. Calcular $\delta(\infty, f^*(\theta))$.
2. Establecer $\mathcal{D}$ y calcular el supremo.
3. Posiblemente f^* no sea densidad!

Ejemplos

- $X_i \sim N(\theta, \lambda)$, precisión λ conocida (considerar primero que $\theta \sim N(\mu_0, \lambda_0)$ *a priori*). Estimar θ con pérdida cuadrática.
- $X_i \sim Be(\theta)$ y $\theta = \{\theta_1, \theta_2\}$. El espacio de acciones es decidirse por θ_1 o θ_2. (Empiece por pérdida 0–1.)

Lo que ganamos con los datos en un ámbito de inferencia

Supongamos que el caso en el que nos encontramos es una de inferencia, ¿como podemos continuar en este caso? Lo primero es establecer cuanto ganamos con los datos. Ahora, sabemos que lo que hacemos en este caso es reportar la distribución posterior correspondiente. ¿Como evaluar la utilidad de reportar dicha distribución?

En un cierto sentido podemos pensar esto como una decisión: decidir qué distribución reportar como la distribución actual para los parámetros. Solo que en este caso **sabemos cual debería ser la decisión a tomar** para ser coherentes: esta debe de ser siempre la posterior.

Tomado el espacio de acciones como el de las densidades en el espacio paramétrico Θ, tendríamos una función de utilidad $u(p(\cdot), \theta)$, y la utilidad esperada de reportar la densidad $p(\theta)$ sería

$$u^*_{\boldsymbol{X}_n}(p(\cdot)) = \int u(p(\cdot), \theta) f(\theta \mid \boldsymbol{X}_n) d\theta,$$

y $p^*_{\boldsymbol{X}_n}(\cdot)$ sería la decisión óptima en este caso.

Ahora, se dice que u es propia (honesta) si

$$p^*_{\boldsymbol{X}_n}(\theta) = f(\theta \mid \boldsymbol{X}_n) \text{ c.s.}$$

Por otro lado se dice que u es local si

$$u(p(\cdot), \theta) = u_\theta(p(\theta)).$$

Teorema

Si $u(p(\cdot),\theta)$ es una utilidad propia, local y diferenciable (en un cierto sentido funcional), entonces

$$u(p(\cdot),\theta) = A\log\{p(\theta)\} + B(\theta)$$

donde $A > 0$ es una constante y $B(\theta)$ es una función arbitraria de θ; a u se le llama utilidad logarítmica.

Tenemos que la utilidad de observar n datos, teniendo como *a priori* $f(\theta)$ es

$$\begin{aligned}\delta(n, f(\cdot)) &= \int f(\boldsymbol{X}_n)\int u(f(\theta \mid \boldsymbol{X}_n),\theta) f(\theta \mid \boldsymbol{X}_n)d\theta d\boldsymbol{X}_n - \\ &\quad \int u(f(\theta),\theta)f(\theta)d\theta.\end{aligned}$$

Notando que $\int f(\boldsymbol{X}_n) f(\theta \mid \boldsymbol{X}_n) d\boldsymbol{X}_n = f(\theta)$ vemos que

$$\delta(n, f(\cdot)) \propto \int f(\boldsymbol{X}_n) \int f(\theta \mid \boldsymbol{X}_n) \log \frac{f(\theta \mid \boldsymbol{X}_n)}{f(\theta)} d\theta d\boldsymbol{X}_n.$$

Esta es la divergencia de Kullback-Liebler de $f(\theta \mid \boldsymbol{X}_n)$ con $f(\theta)$. Esto es: qué tanta información ganamos al pasar de la *a priori* a la *a posteriori*.

Notamos entonces que

$$\delta(\infty, f(\cdot)) = \lim_{n\to\infty} \delta(n, f(\cdot))$$

es la ganancia de una muestra infinita usando como *a priori* $f(\cdot)$. Buscamos entonces f^* que maximice $\delta(\infty, f(\cdot))$ (minimice la influencia de $f(\cdot)$). Desgraciadamente, para parámetros continuos, lo usual es que $\delta(\infty, f(\cdot)) = \infty$ para toda $f(\cdot)$ relevante.

Sin embargo, una alternativa es considerar las $p_k(\cdot)$'s que maximicen a $\delta(k, p(\cdot))$ y luego tomar el límite de estas. Para motivar una definición formal vemos que una expresión alternativa para $\delta(k, p(\cdot))$ es

$$\delta(k, p(\cdot)) = \int p(\theta) \log \frac{f_k(\theta)}{p(\theta)} d\theta,$$

donde

$$f_k(\theta) = \exp\left\{ \int f(\boldsymbol{X}_k \mid \theta) \log f(\theta \mid \boldsymbol{X}_k) d\boldsymbol{X}_k \right\}.$$

La *a priori* que maximice $\delta(k, p(\cdot))$ está sujeta a $\int p(\theta) d\theta = 1$ y por lo tanto ha de ser un extremo del funcional

$$F(p(\cdot)) = \int p(\theta) \log \frac{f_k(\theta)}{p(\theta)} d\theta + \lambda \left\{ \int p(\theta) d\theta - 1 \right\}.$$

El funcional es de la forma $F(p(\cdot)) = \int g(p(\theta))d\theta$. Usando teoría de operadores vemos que, por las características de g una $p(\cdot)$ que maximice a F ha de cumplir con

$$\frac{\partial}{\partial \epsilon} F(p(\cdot) + \epsilon\tau(\cdot)) = 0 \text{ para } \epsilon = 0 \text{ y para toda } \tau.$$

Esto, después de alguna álgebra, nos lleva a que

$$\int \tau(\theta)(\log f_k(\theta) - \log p(\theta) + \lambda)d\theta = 0 \text{ para toda } \tau,$$

lo que implica que el extremo $p_k(\theta)$ debe cumplir con $\log f_k(\theta) - \log p_k(\theta) + \lambda = 0$ y por lo tanto $p_k(\theta) \propto f_k(\theta)$.

Note, sin embargo, que $f_k(\theta)$ depende de la *a priori*. La idea aquí es utilizar una aproximación asintótica de la *a posteriori* $f^*(\theta \mid \boldsymbol{X}_k)$ que ya no dependa de la *a priori* usada y definir

$$p_k^*(\theta) = \exp\left\{ \int f(\boldsymbol{X}_k \mid \theta) \log f^*(\theta \mid \boldsymbol{X}_k) d\boldsymbol{X}_k \right\}.$$

Esta secuencia de "*a priori*'es" definirá una secuencia de posteriores, para una muestra dada $\boldsymbol{X}$,

$$p_k(\theta \mid \boldsymbol{X}) \propto f(\boldsymbol{X} \mid \theta) p_k^*(\theta)$$

la cual tendrá el mismo límite que si hubiésemos usado la secuencia exacta $p_k(\theta)$. De aquí sale la definición de *a priori* de referencia por Bernardo (1979).

Definición

Sea X una observación con el modelo $p(X \mid \theta)$ y $\boldsymbol{X} = \{X_1, X_2, \ldots, X_k\}$ n observaciones independientes de este. Definimos

$$f_k^*(\theta) = \exp\left\{ \int p(\boldsymbol{X} \mid \theta) \log p^*(\theta \mid \boldsymbol{X}) d\boldsymbol{X} \right\},$$

donde $p^*(\theta \mid \boldsymbol{X})$ es una aproximación asintótica a la a posteriori de θ.

*La **distribución de referencia posterior** de θ dado X se define como la distribución $\pi(\theta \mid X)$ tal que*

$$E[\delta(\pi_k(\theta \mid X), \pi(\theta \mid X))] \to 0,$$

cuando $k \to \infty$, suponiendo que dicho límite exista donde $\delta(g, h) = \int g(\theta) \log \frac{g(\theta)}{h(\theta)} d\theta$, $\pi_k(\theta \mid X) = c(X) p(X \mid \theta) f_k^*(\theta)$ y $c(\boldsymbol{X})$ es la constante de normalización. Cualquier función positiva $\pi(\theta)$ tal que $\pi(\theta \mid X) \propto p(X \mid \theta)\pi(\theta)$ la llamamos a priori de referencia para θ para el modelo en cuestión.

Hay muchos resultados relacionados con esta definición. Entre estos se demuestra que la *a priori* de referencia no depende del tamaño de muestra. También, que si nos fijamos en un estimador suficiente, la *a priori* que resulta es la misma.
Una característica muy importante es que si estipulamos el modelo en términos de una transformación uno a uno de θ, por ejemplo $\phi = g(\theta)$ entonces las *a posteriori* de referencia π_θ y π_ϕ tienen la relación

$$\pi_\phi(t) = \pi_\theta(g^{-1}(t)) \left| \frac{dg^{-1}(\phi)}{d\phi} \right|_{\phi=t}$$

(para variables continuas). Esta característica es muy deseable y otros métodos para encontrar *a priori* de referencia no la cumplen.

Para parámetros con soporte finito, la *a priori* de referencia es una constante (la misma probabilidad para todos los valores del soporte). Hacer ejercicio.

Teorema

Bajo los supuestos de la definición 20, si $p^*(\theta \mid \boldsymbol{X})$ *es una aproximación normal asintótica con con precisión* $kh(\hat{\theta}_k)$ *donde* $\hat{\theta}_k$ *es un estimador consistente de* θ*, entonces la a priori de referencia es de la forma*

$$\pi(\theta) \propto \{h(\theta)\}^{1/2}.$$

Esta *a priori* es conocida como la *a priori* de Jeffreys.

Para el caso de muestreo Bernoulli podemos ver que una aproximación normal a la posterior, usando $\hat{\theta}_k = \frac{1}{k}\sum_{i=1}^{k} X_k$, es $N(\hat{\theta}_k, nh(\hat{\theta}_k))$ donde $h(\theta) = \theta^{-1}(1-\theta)^{-1}$ (precisión). Por lo tanto

$$\pi(\theta) \propto \{h(\theta)\}^{1/2} = \theta^{-1/2}(1-\theta)^{-1/2},$$

o sea, $\theta \sim Beta(1/2, 1/2)$, que es una distribución propia. Vemos que la distribución posterior, para $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ es

$$\theta \mid \boldsymbol{X} \sim Beta(1/2 + \sum_{i=1}^{n} X_i, 1/2 + n - \sum_{i=1}^{n} X_i),$$

la cual es propia, aún cuando no se hayan observado ningún fracaso (o éxito).

Inferencia comparada (2)

La inferencia frecuentista (clásica) y la inferencia Bayesiana no son totalmente compatibles. Si bien, en ambos casos, se intenta proveer inferencias acerca de lo incierto (parámetros) partiendo de una muestra, la estadística Bayesiana mantiene a la muestra observada como fija y la frecuentista considera múltiples posibles escenarios, tomando a la muestra como una de tantas posibles que pudieron ocurrir.

Como puntos de comparación podemos anotar:

- La Bayesiana intenta crear una teoría para hacer inferencia, la frecuentista intenta dar lineas de acción ante un problema de inferencia.
- La frecuentista considera a la probabilidad como algo medible, la Bayesiano no necesariamente.
- La frecuentista se (auto)promulga como “objetiva” (no depende de quien la aplica), mientras que la Bayesiana se le ve como subjetiva (??) y se demuestra como internamente consistente (“coherente”).

Críticas comunes a la inferencia Bayesiana

La críticas más comunes a la Estadística Bayesiana las ubicamos en los siguientes puntos:

- Falta de objetividad.
- Solo la probabilidad, en el sentido frecuentista, existe.
- ¿Como establecer la *a priori*?
- Es más difícil o, a veces, no es factible hacer los cálculos.
- La Estadística clásica es más conocida y fácil de entender.

Crítica (Bayesiana) a la inferencia frecuentista

Hemos ya, a lo largo del curso, establecido múltiples diferencies e, implícitamente hemos hecho varias críticas a la Estadística frecuentista. El problema fundamental es que la Estadística frecuentista no sigue el principio de verosimilitud:

Definición

Toda la información relevante acerca de una muestra $\boldsymbol{X}$ *está contenida en la verosimilitud. Concretamente: la información contenida en una muestra acerca del mismo parámetro es la misma si las verosimilitudes correspondientes son proporcionales.*

Note que la estadística Bayesiana sigue (sin excepción) el principio de verosimilitud, pues la información relevante de una muestra está contenida en la posterior correspondiente, que depende de la muestra solo a mediante de la verosimilitud.

Ejemplo: (Lindley y Phillips, Berger, 1985, p.28) Tenemos una moneda y estamos interesados en la probabilidad θ de que al tirarla caiga en águila. Se hace un experimento, con ensayos independientes, y resultan 9 águilas y 3 soles,.
Note que con la información anterior no especificamos el procedimiento que se siguió: ¿Se lanzó 12 veces la moneda de manera independiente o se lanzó la moneda hasta que se observaron 3 soles? En el primer caso, el número de éxitos es $Bi(12, \theta)$ y en el segundo caso $BN(3, \theta)$. Las verosimilitudes respectivas son

$$(220)\theta^9(1-\theta)^3$$

y

$$(55)\theta^9(1-\theta)^3.$$

Sin embargo, estas verosimilitudes son proporcionales, y la información es la misma: en general, en Bayesiana, el tiempo de paro es irrelevante para la inferencia.

Violar el principio de verosimilitud lleva a cosas muy absurdas:
Ejemplo:(Pratt, Berger, 1985, p.30) ¿Se usó un voltímetro de 100v o de 1000v?

Un experimentador mide unos voltajes, obteniendo un voltaje máximo de 98. El experimentador acude con un Estadístico (frecuentista) y este evalúa que la muestra se puede ver como de una distribución Normal procediendo a su análisis.

Casualmente, al pasar por el laboratorio del experimentador, el estadístico nota que el voltímetro solo mide hasta 100 voltios.

El estadístico ahora se preocupa y tiene que cambiar de análisis por tratarse de una muestra censurada.

Días después el experimentador le informa que el tiene un voltímetro que mide hasta 1,000 voltios. El estadístico se relaja y desecha el análisis de muestras censuradas.
Lamentablemente, el experimentador luego se acuerda que el día que realizó el experimento no vino el trabajador sindicalizado que tiene la llave donde se guarda el voltímetro de 1,000 voltios...

el Estadístico entra en crisis otra vez! (Por la posibilidad de un hecho en el pasado, que no ocurrió.)

Ejemplo:(Berger y Wolper, 1988, *The Likelihood Principle*, p.5) Este es un ejemplo, un cuanto artificial, pero nos indica en que tipo de absurdos podemos caer si no seguimos el principio de verosimilitud: Suponga que vamos a observar X_1 y X_2 y que estas son independientes, y que $P(X_i = \theta - 1) = P(X_i = \theta + 1) = \frac{1}{2}$. Aquí $-\infty < \theta < \infty$ es el parámetro desconocido de interés. Es fácil ver que un intervalo de 75% de confianza de tamaño mínimo es

$$C(X_1, X_2) = \begin{cases} \frac{1}{2}(X_1 + X_2) & \text{si } X_1 \neq X_2 \\ X_1 - 1 & \text{si } X_1 = X_2. \end{cases}$$

Esto es, si repetimos la muestra de X_1 y X_2 muchas veces, θ pertenecerá a $C(X_1, X_2)$ el 0.75 de la veces.

Sin embargo, note que si $X_1 \neq X_2$ estamos **totalmente seguros** que $\theta = \frac{1}{2}(X_1 + X_2)$, mientras que si $X_1 = X_2$, es igual de factible que $\theta = X_1 - 1$ ó $\theta = X_1 + 1$ (suponiendo que no sabemos nada más acerca de θ).

Lo que concluimos es que la aseveración de que "$C(X_1, X_2)$ es un intervalo de 75% de confianza" es solamente *pre experimental*. Una vez vista la evidencia, o se está seguro del valor de θ o se tiene 50% de incertidumbre o confianza. Reportar 75% de confianza *pos experimental* es verdaderamente un absurdo.

¿Como se haría un análisis Bayesiano en este caso?

Otro ejemplo es el siguiente. Tengo dos instrumentos para medir X, uno sin error y otro con error. Tiro una moneda para decidir qué instrumento uso y resulta que escojo el instrumento sin error y la medición fue x...a sabiendas del resultado de tirar la moneda, ¿tiene sentido considerar el posible hecho, que no ocurrió, de que pudo la moneda haber decidido por el otro instrumento y entonces tener una medición de X con error?:

Las estadísticas no Bayesianas, que evalúan la incertidumbre en términos de ideas de muestras repetidas (violan el principio de verosimilitud y) no tienen un concepto de precisión pre y pos experimental.

¿Discusión?

MCMC (8)

Ya hemos visto que la práctica de la inferencia Bayesiana tiene como obstáculo la integración de la posterior para encontrar la constante de normalización y, en su caso, las marginales requeridas de la posterior conjunta. La solución que se ha implementado desde 1990 (aun cuando en el área de física estadística ya se conocía desde los 1970's) es la Simulación de cadenas de Markov ó *Markov Chain Monte Carlo* (MCMC).

La idea fundamental en MCMC es la siguiente:

1. Formamos una cadena de Markov $X^{(1)}, X^{(2)}, \ldots$ que tenga como distribución estacionaria a la posterior de interés.
2. Dejando correr la cadena un número grande de veces obtendremos entonces una simulación de la posterior.
3. Con muchas simulaciones de la posterior hacemos aproximaciones a distribuciones marginales o momentos de la posterior.

En los tres puntos anteriores se resumen los tres problemas fundamentales en MCMC, siendo el tercero el de menor peso, comparativamente a los otros dos:

1. Diseño de la cadena de Markov.
2. Análisis de convergencia de la cadena.
3. Manejo de las simulaciones de la posterior.

Esta teoría se aplica para simular de cualquier distribución y no solo de una posterior. Le llamaremos a la posterior la *distribución objetivo* $f(X)$ (conforme a la notación anterior, una posterior típicamente la denotábamos como $f(\theta \mid \boldsymbol{X})$; en esta sección quitaremos, a nivel notacional, la dependencia en los datos e identificaremos a X con θ, los parámetros).

Teoría

El algoritmo más general es el siguiente, llamado Metropolis–Hastings:

Definición

Dado $x^{(t)}$

1. *generar* $Y_t \sim q(y \mid x^{(t)})$, *donde* $q(\cdot \mid \cdot)$ *es, en principio, una distribución arbitraria, conocida como instrumental o propuesta.*
2. $$x^{(t+1)} = \begin{cases} Y_t & \text{con probabilidad } \rho(x^{(t)}, Y_t) \\ x^{(t)} & \text{con probabilidad } 1 - \rho(x^{(t)}, Y_t), \end{cases}$$

   donde

$$\rho(x, y) = \min \left\{ 1, \frac{f(y)}{f(x)} \frac{q(x \mid y)}{q(y \mid x)} \right\}.$$

Evidentemente que el algoritmo forma una cadena de Markov, pues $x^{(t+1)}$ solamente depende de $x^{(t)}$. El punto crucial en este algoritmo es que la distribución objetivo $f(x)$ solo es necesario conocerla salvo una constante, pues solo razones de esta $\frac{f(y)}{f(x)}$ son requeridas.

El diseño de la cadena entonces depende de la distribución instrumental que se utilice. La distribución instrumental define entonces, implícitamente, un kernel de transición $K(x, y)$, esto es, la probabilidad (o densidad) de pasar de x a y ($K(x, A)$ es la medida de probabilidad de pasar de x a un medible A).

Sea $\mathcal{E}$ el soporte de f.

Definición

Una cadena de Markov con Kernel de transición K se dice que cumple balance detallado con respecto a la función f si

$$K(y,x)f(y) = K(x,y)f(x)$$

para cualquiera (x,y).

Teorema

Si K cumple con el balance detallado con respecto a la función f entonces f es una densidad invariante de la cadena (y la cadena es reversible).

Demostración.

$$\int K(y,B)f(y)dy = \int \int_B K(y,x)f(y)dxdy = \int \int_B K(x,y)f(x)dxdy =$$

$$\int_B \int K(x,y)dyf(x)dx = \int_B f(x)dx =$$

Por lo tanto:

$$\int K(y,B)f(y)dy = f(B).$$

□

Teorema

Usando en el algoritmo de Metropolis–Hastings una distribución propuesta $q(\cdot \mid x)$, definida para toda $x \in \mathcal{E}$ y que sea reversible (esto es, que si $q(y \mid x) > 0$ entonces $q(x \mid y) > 0$) su Kernel de transición cumple con balance detallado con la función objetivo f y por lo tanto esta es una distribución estacionaria de la cadena.

Demostración.

El Kernel de transición es

$$K(x,y) = \rho(x,y)q(y \mid x) + (1 - r(x))\delta_x(y),$$

donde $r(x) = \int \rho(x,y)q(y \mid x)dy$ (la probabilidad de moverse ie. aceptar la propuesta) y $\delta_x(y)$ la delta de Dirac en x. Es claro que

$$\rho(x,y)q(y \mid x)f(x) = \rho(y,x)q(x \mid y)f(y)$$

y

$$(1 - r(x))\delta_x(y)f(x) = (1 - r(y))\delta_y(x)f(y).$$

□

En general vamos a tomar una combinación de Kerneles de Metropolis–Hastings K_i, para combinar varias propuestas q_i, $i = 1, 2, \ldots, n$. Por ejemplo, un Kernel puede ser que mueva solo uno o un conjunto de las variables, otro otro subconjunto y puede haber kenerles que muevan todas las componentes de $\mathcal{E}$. La combinación es la siguiente:

$$K(x, y) = \sum_{i=0}^{n} p_i K_i(x, y)$$

donde $\sum_{i=0}^{n} p_i = 1$, $p_i > 0$ y $K_0(x, x) = 1$ (o sea, con probabilidad p_0 no hacemos nada). Especificar las p_i's es parte del diseño el MCMC.

Note que como los K_i's cumplen con el balance detallado con respecto a f, entonces K también y f es una distribución estacionaria de la cadena generada (K_0 cumple con balance detallado de manera trivial).

Para que f sea distribución límite de la cadena necesitamos que:

1. K sea aperiódico. Una condición suficiente para aperiodicidad es que $K(x,x) > 0$, para todo $x \in \mathcal{E}$. Esto siempre se cumple ya que incluimos el kernel K_0.
2. Necesitamos que K sea f–irreducible. Esto es, que para todo A medible tal que $f(A) > 0$ y para todo $x \in \mathcal{E}$ existe n tal que

$$K^n(x,A) > 0.$$

(Qué desde cualquier punto del soporte de f alcancemos un conjunto cualquiera no f–nulo en un número finito de pasos.)

En dado caso f es la distribución límite de la cadena. Esto es: si $h \in L^1(f)$ entonces

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} h(X^{(t)}) = \int h(x) f(x) dx \quad c.s.c.r \quad f$$

y

$$\lim_{n\to\infty} \left\| \int K^n(x, \cdot)\mu(dx) - f \right\|_{VT} = 0$$

para cualquier distribución inicial μ.

(La norma de variación total, VT, de medidas es: $\sup_A |\mu_1(A) - \mu_2(A)|$.)

Existen muchos casos particulares de Metropolis–Hastings, que pueden ser usados para definir cada kernel K_i. El resultado se conoce como un kernel híbrido.

Propuesta independiente

Tomamos $q(y \mid x) = q(y)$, independiente de donde estamos.

Ejemplo: Función objetivo:

$$f(x) \propto x^{\alpha-1} e^{-\beta x^2},$$

con $x > 0, \alpha, \beta > 0$. Usar una gamma como propuesta.

Algoritmo Metropolis

Tomamos $q(x \mid y) = q(y \mid x)$, propuesta simétrica. Note que en este caso

$$\rho(x, y) = \min\left\{1, \frac{f(y)}{f(x)}\right\}.$$

Como caso particular tenemos una caminata aleatoria. Esto es, $q(y \mid x) = g(y - x)$ donde g es una densidad simétrica en el cero (eg. Normal con media cero).

Ejemplo: Simular de una Normal usando una uniforme.

Kernel Gibbs

En este caso la propuesta está dada por la distribución objetivo en si. Tomamos un bloque x_1 de x (una componente o un conjunto de componentes de x) y nos fijamos en $f(x_1 \mid x_{-1})$, donde x_{-1} son las componentes que restan de x_1. Note que

$$f(x_1 \mid x_{-1}) \propto f(x),$$

y posiblemente al reducir la dimensión, la distribución condicional de x_1 tenga una forma conocida. A esta distribución se le conoce como condicional total o *full conditional*. Tomamos entonces

$$q(y_1 \mid x) = f(y_1 \mid x_{-1})$$

y tomando (dejando fijos a) $y_{-1} = x_{-1}$. Note que en este caso $\rho(x, y) = 1$ y las propuestas siempre son aceptadas.

Ejemplo: Simular del modelo auto exponencial. La función objetivo es

$$f(x_1, x_2, x_3) \propto \exp\{-(x_1 + x_2 + x_3 + ax_1x_2 + bx_2x_3 + cx_3x_1)\}$$

con $a, b, c > 0$ conocidas.

Ejemplo: Ejemplo en epidemiología (Basado en Robert y Casella, p.300).

Se tienen m rebaños de vacas, en establos separados físicamente los cuales pueden estar afectados por una enfermedad que se presenta en la etapa endémica (A) o en la etapa epidémica (B). De estos rebaños se tomó una muestra de tamaño $n = 100$ siendo X_i el número de vacas infectadas con la enfermedad en la muestra tomada del rebaño i. Se sabe que la etapa endémica es cuando menos del 5% de las vacas están infectadas y la etapa epidémica es cuando mas del 15% de las vacas están infectadas. Lo que se quiere decidir es si se trata a cada rebaño tratando cada vaca con una vacuna que cuesta \$2,000 pesos (en la etapa epidémica de la enfermedad lo más probable es que todas las vacas del rebaño queden infectadas). Se calcula que la enfermedad de una vaca acumula \$15,000 pesos en pérdidas.

El modelo al que se llega es el siguiente:

$$X_i \sim Po(\lambda_A^{z_i} \lambda_B^{1-z_i})$$

donde $\lambda_A \sim Ga(\alpha_1, \beta_1)$, $\lambda_B \sim Ga(\alpha_2, \beta_2)$, $z_i \sim Be(p_i)$. Si $z_i = 1$ lo interpretamos como que el rebaño *i* se encuentra en la etapa endémica (Poisson con parámetro λ_A) y si $z_i = 0$ lo interpretamos como que el rebaño *i* se encuentra en la etapa epidémica (Poisson con parámetro λ_B).

Usar Gibbs sampling para simular de la posterior. ¿Como se tomaría la decisión de tratar o no al rebaño?

Datos futuros o faltantes

Cuando tenemos datos faltantes y, lo que necesitamos es $f(x, y)$ para después marginalizar y quedarnos con la predictiva $f(y)$. Lo único que tenemos entonces que hacer es incluir a los datos faltantes como parámetros.

Ejemplo: Tomar datos faltantes en el ejemplo de epidemiología.

Completación

La idea aquí es tomar como función objetivo una densidad g más grande que f tal que

$$f(x) = \int g(x, y)dy.$$

Esto pude ser útil si, por ejemplo, usar Gibbs es más sencillo en g que directamente en f.

Ejemplo: Simular de una Normal truncada.

Inferencia Bayesiana moderna, ejemplos.

- Ejemplo de Farmacocinética, 7.1.22, p.300, Robert y Casella.
- Datos futuros en curvas de acumulación (acumulación a periodos iguales), Christen y Nakamura.
- Datos censurados, 7.1.8, p.291, Robert y Casella.

Criterios de convergencia en MCMC

No vamos a comentar nada sobre este importante tema en este curso.

Modelos jerárquicos y modelación gráfica

Temas selectos

Se sigue como seminario, en expocisiones de alumnos y maestro.