CAPÍTULO 5. LA DISTRIBUCIÓN A PRIORI

Para leer

Gelman et al (1995), Secciones 2.8 – 2.9.
Lee (1997), Secciones 3.2 – 3.3.
Berger (1985), Capítulo 3.

Dado un problema, necesitamos estructurar un modelo probabilístico para nuestras creencias. ¿Cómo podemos elegir la distribución inicial que las represente bien?

Existen varias posibilidades.

Distribuciones Conjugadas

Si existe una familia conjugada, podemos eligir una distribución dentro de esta familia por especificación de, por ejemplo, los primeros momentos de la distribución de $\boldsymbol{\theta}$ o de la distribución predictiva de X.

Ejemplo 42 *Sea θ la probabilidad de que una tirada de una moneda sea cruz. Estimamos $E[\theta] = 0{,}4$ y suponemos que nuestras conocimientos son equivalentes a una muestra de 100 tiradas.*

Suponiendo una distribución a priori $\mathcal{B}(\alpha, \beta)$,

$$\begin{aligned} \frac{\alpha}{\alpha+\beta} &= 0{,}4 \\ \alpha+\beta &= 100 \end{aligned}$$

Resolviendo las ecuaciones, la solución es $\alpha = 40$ y $\beta = 60$ y la distribución a priori es $\mathcal{B}(40, 60)$.

Ejemplo 43 *Sea* $X|\theta \sim \mathcal{E}(\theta)$ *el tiempo entre llegadas en un supermercado. Una distribución a priori conjugada será de forma* $\theta \sim \mathcal{G}(\alpha, \beta)$. *Se estima que* $E[X] = 1$ *y* $V[X] = 2$.

Entonces, suponiendo una distribución a priori gamma:

$$\begin{aligned}
E[X] &= E[E[X|\theta]] \\
&= E[1/\theta] = \frac{\beta}{\alpha - 1} \\
V[X] &= V[E[X|\theta]] + E[V[X|\theta]] \\
&= V[1/\theta] + E[1/\theta^2] \\
&= 2V[1/\theta] + E[1/\theta]^2 \\
&= 2\frac{\beta^2}{(\alpha-1)^2(\alpha-2)} + \left(\frac{\beta}{\alpha-1}\right)^2 \\
&= \frac{\alpha\beta^2}{(\alpha-1)^2(\alpha-2)}
\end{aligned}$$

Resolviendo las ecuaciones, se tiene la solución $\alpha = 4$, $\beta = 3$.

Distribuciones Reales

En muchos problemas reales, se quiere solicitar las distribuciones de expertos.

¿Cómo solicitar las predicciones?

Una posibilidad es dividir Θ en intervalos y solicitar las probabilidades del experto (cuantiles) para cada intervalo. Dados los cuantiles solicitados, se estima la densidad entera mediante suavización.

Para solicitar los cuantiles, se pueden utilizar loterías.

Reglas estríctamente propias

Se quiere solicitar la (verdadera) probabilidad de un experto para una variable Bernouilli S.

Se paga el experto una cantidad $R(S, p)$ donde p es la probabilidad proporcionada por el experto.

¿Cómo definir $R(S, p)$?.

El experto quiere maximizar su sueldo. Si q es su verdadera probabilidad, su sueldo esperado si dice p es

$$qR(1, p) + (1 - q)R(0, p)$$

Una regla (estrictamente) propia es una regla $R(S, p)$ para que el experto máximiza su sueldo esperado si (y sólo si) $p = q$.

Ejemplo 44 $R(S,p) = 1 - |S - p|$

Para el experto

$$\begin{aligned} E[R] &= q\,(1 - |1 - p|) + (1 - q)\,(1 - |0 - p|) \\ &= qp + (1 - q)(1 - p) \\ &= 1 - q + (2q - 1)p \end{aligned}$$

Entonces $R(S,p)$ *no es propia.*

Si $q >$ $(<)$ 0,5, *el experto máximiza su sueldo esperado con* $p = 1$ (0).

Ejemplo 45 $R(S,p) = 1-(S-p)^2$ *es la regla de Brier.*

$$\begin{aligned} E[R] &= q\left(1-(1-p)^2\right)+(1-q)\left(1-p^2\right) \\ &= 1-q+2pq-p^2 \\ \frac{dE}{dp} &= 2q-2p \\ \widehat{p} &= q \end{aligned}$$

R *es una regla estrictamente propia.*

Observación 20 *Existen otras reglas estrictamente propias: la regla logaritmica,*

$$R(S,p) = \log(1-|S-p|)$$

o le regla esférica

$$R(S,p) = \frac{1-|S-p|}{\sqrt{p^2+(1-p)^2}}.$$

Ejemplo 46 *El experto debe proporcionar un estimador puntual de una variable continua X. Sea e su estimador y define la regla*

$$R(x,e) = \begin{cases} a(e-x) & \text{si } e < x \\ b(x-e) & \text{si } e > x \end{cases}$$

El experto minimiza su pérdida esperada si proporciona su $a/(a+b) \times 100\,\%$ cuantil.

Demostración

$$\begin{aligned}
E[R(X,e)] &= \int R(x,e)f(x)\,dx \\
&= \int_e^\infty f(x)a(e-x)\,dx + \\
&\quad \int_{-\infty}^e f(x)b(x-e)\,dx \\
&= ae(1-F(e)) - a\int_e^\infty xf(x)\,dx + \\
&\quad b\int_{-\infty}^e xf(x)\,dx - beF(e) \\
\frac{dE}{de} &= -aef(e) + a(1-F(e)) + aef(e) + \\
&\quad bef(e) - bef(e) - bF(e) \\
&= a(1-F(e)) - bF(e) \\
\left.\frac{dE}{de}\right|_{\widehat{e}} &= 0 \Rightarrow \\
0 &= a(1-F(\widehat{e})) - bF(\widehat{e}) \\
F(\widehat{e}) &= \frac{a}{a+b}
\end{aligned}$$

Observamos que el segundo derivado es

$$\frac{d^2E}{de^2} = -af(e) - bf(e) < 0$$

y entonces el punto $\widehat{e}$ maximiza la ganancia esperada (o minimiza la pérdida esperada).

Problemas con predicciones subjetivas

Existen muchos problemas en solicitar distribuciones subjetivas reales. Típicamente, predicciones humanas son sesgadas o (peor) incoherentes.

Ejemplo 47 *"Federico tiene 35 años, es inteligente pero poco imaginativo, compulsivo y aburrido. En la escuela era muy abil en mátematicas pero con poco talento en los artes"*

Ordenar las siguientes frases por probabilidad (1 = más probable, 8 menos probable).

1. *Federico es un médico que juega a las cartas como pasatiempos*

2. *Es arcitecto.*

3. *Es contable.*

4. *Toca un instrumento en un grupo Jazz.*

5. *Lee Marca.*

6. *Le gusta el senderismo.*

7. *Es contable y toca un instrumento Jazz.*

8. *Es periodista.*

En este ejemplo, la mayoría de la gente clasifican el 3 (contable) como más probable. Por cierto, la descripción es muy representativa de los contables.

Bastantes personas también ponen $P(7) > P(4)$. *Pero*

$$P(\textit{contable} \cap \textit{jazz}) < P(\textit{jazz})$$

Este problema ilustra la falacia de la tasa base. En calcular $P(3|\textit{info})$ *y* $P(5|\textit{info})$ *se ignoran las frecuencias base* $P(3)$ *y* $P(5)$.

Evaluación de predicciones subjetivas

Algunos criterios son los siguientes:

- Honestidad: se quiere que el experto diga sus verdaderas opiniones.

- Coherencia: las predicciones deben cumplir las leyes de probabilidad.

- Consistencia: si el experto no padece de información nueva, sus predicciones no deben cambiar.

- Precisión: debe llover un 50 % de los veces cuando el experto dice $P(\text{llover}) = 0{,}5$.

- Información: si, en Madrid, llueve aproximadamente 50 dias al año, un experto que dice $P(\text{llover mañana}) = 50/364$ no es informativo.

Medidas de precisión y información

Supongamos que el experto proporciona sus probabilidades $\mathbf{p}$ para sucesos $X_1, \ldots, X_n$. Despues de ver los datos $\mathbf{x}$, se puede evaluar sus predicciones.

Consideramos la regla de Brier

$$R(X, p) = 1 - (S - p)^2.$$

Se puede calcular el estadístico

$$R(\mathbf{x}, \mathbf{p}) = \frac{1}{n} \sum_{i=1}^{n} R(x_i, p_i)$$

que es una medida de la calidad media de las predicciones.

Se puede dividir la medida en dos partes: una medida de precisión y una medida de información. (Murphy 1973).

$$
\begin{aligned}
R(\mathbf{x},\mathbf{p}) &= \frac{1}{n}\sum_{i=1}^{n} R(x_i,p_i) \\
&= 1-\frac{1}{n}\sum_{j=1}^{k}\left(n_j r_j(1-p_j)^2 + n_j(1-r_j)p_j^2\right) \\
&= 1-C(\mathbf{x},\mathbf{p})-I(\mathbf{x},\mathbf{p}) \qquad \text{donde} \\
C(\mathbf{x},\mathbf{p}) &= \frac{1}{n}\sum_{j=1}^{k} n_j(r_j-p_j)^2 \\
I(\mathbf{x},\mathbf{p}) &= \frac{1}{n}\sum_{j=1}^{k} n_j r_j(1-r_j)
\end{aligned}
$$

donde el experto utilizó la probabilidad p_j un número n_j veces y una frecuencia de r_j sucesos ocurrieron, $j = 1,\ldots,k$.

C es una medida de precisión:

- $0 \leq C \leq 1$
- $C = 0$ si y sólo si $r_j = p_j$ para $j = 1, \ldots, k$.
- Para un experto preciso, cuando $n \to \infty$, $C \to 0$.
- C es grande si las frecuencias observadas r_i son muy distintas a las probabilidades del experto p_i.

I mide la información

- $0 \leq I \leq 0{,}25$.
- $I = 0$ si para cualquier p_j, la frecuencia $r_j = 0$ o 1.
- $I = 0{,}25$ si para cualquier p_j, $r_j = 0{,}5$.

Ejemplo 48 *Tres expertos dan sus probabilidades de que pierda Getafe durante los primeros 29 partidos de la temporada 2004-2005 con los siguientes resultados:*

Día	**1**	**2**	**3**	4	**5**	6	**7**	**8**	**9**	10
E_1	,3	,9	,7	,6	,8	,5	,4	,8	,2	,3
E_2	,9	,8	,8	,7	,6	,9	,8	,7	,2	,5
E_3	,4	,5	,5	,4	,4	,4	,5	,5	,4	,4
Día	11	12	**13**	14	**15**	16	17	18	19	20
E_1	,6	,2	,2	,5	,4	,5	,4	,3	,2	,2
E_2	,5	,3	,8	,3	,7	,2	,1	,4	,2	,5
E_3	,5	,5	,5	,4	,4	,4	,5	,5	,4	,5
Día	**21**	22	**23**	24	25	26	27	28	**29**	
E_1	,4	,2	,4	,3	,4	,6	,5	,6	,8	
E_2	,3	,2	,8	,7	,6	,9	,8	,7	,2	
E_3	,9	,4	,8	,5	,4	,5	,4	,4	,6	

Los días cuando perdió el Getafe están marcados en negrito.

Se tabulan las frecuencias relativas a cada probabilidad utilizada.

Para E_1 *se tiene*

p_j	,1	,2	,3	,4	,5	,6	,7	,8	,9
n_j	0	6	4	6	4	4	1	3	1
r_j	0	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{2}{3}$	0	0	1	1	1

y para E_2*,*

p_i	,1	,2	,3	,4	,5	,6	,7	,8	,9
n_i	1	5	3	1	3	2	5	6	3
r_i	0	$\frac{2}{5}$	$\frac{1}{3}$	0	0	$\frac{1}{2}$	$\frac{2}{5}$	$\frac{5}{6}$	$\frac{1}{3}$

y para E_3*,*

p_i	,1	,2	,3	,4	,5	,6	,7	,8	,9
n_i	0	0	0	14	12	1	0	1	1
r_i	0	0	0	$\frac{2}{7}$	$\frac{5}{12}$	1	0	1	1

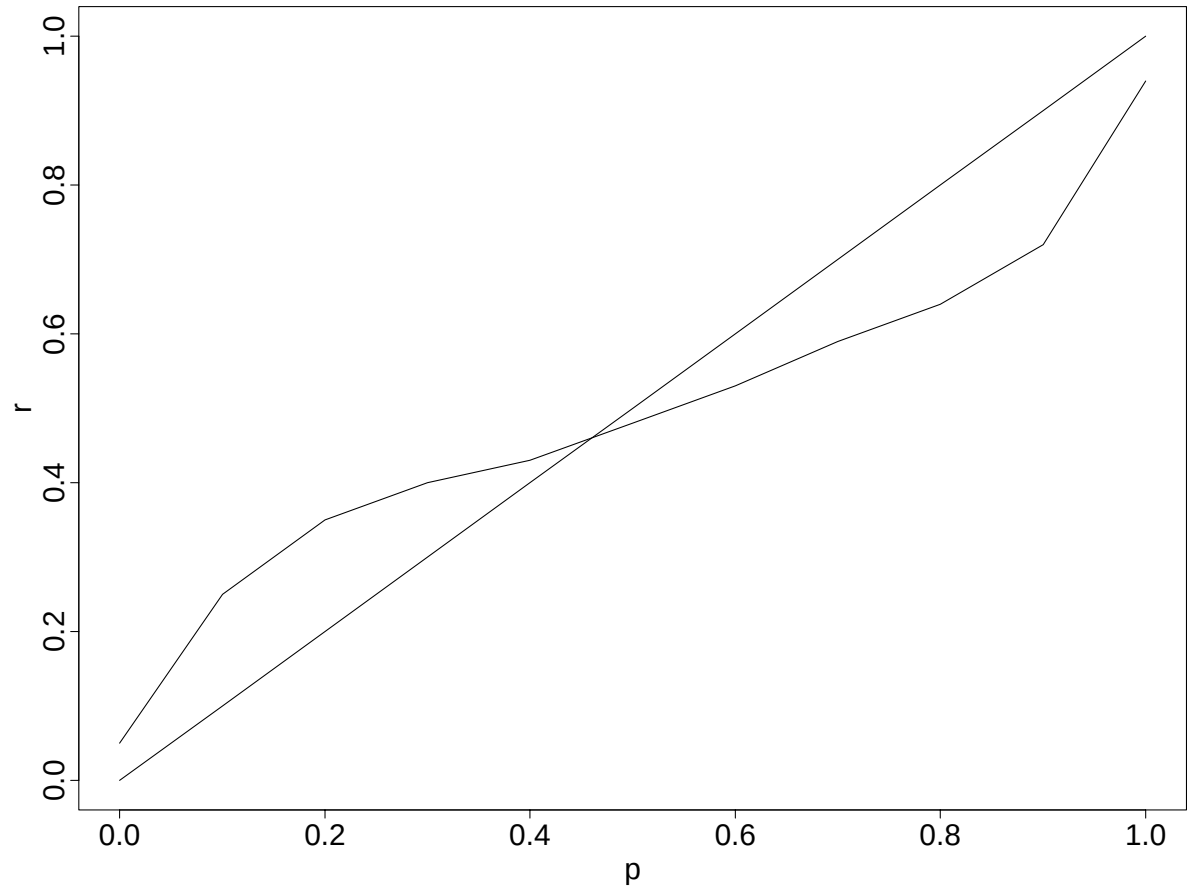
Ahora se calculan las medidas de precisión e información para cada experto.

experto	*precisión*	*información*	*Brier*
E_1	,1105	,1178	,7717
E_2	,0880	,1747	,7372
E_3	,0164	,1991	,7845

El mejor experto con respeto a la regla de Brier es E_3. También es el experto más preciso. No obstante es el menos informativo. El más informativo y también el menos preciso es E_1.

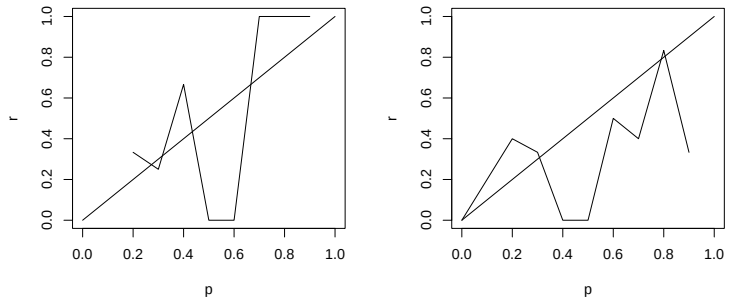
La curva de precisión

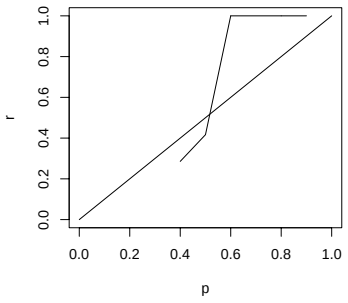
La curva de precisión es un gráfico de las frecuencias observadas r_j frente a las probabilidades utilizadas p_j.



Para un experto preciso, la curva aproxima a la recta de 45 grados.

Ejemplo 49 *Sacamos curvas de precisión para los expertos del Ejemplo 48*





Ninguno de los 3 parece muy preciso.

Distribuciones a priori no informativas

De vez en cuando no se quiere poner información en la distribución a priori, porque

- no se sabe "nada" sobre el problema,
- se quiere ser objetivo.

En estas situaciones se tienen que elegir distribuciones a priori no informativas.

Pero hay muchas posibilidades. ¿Cúal es lo más útil?

El principio de razón insuficiente

Este principio (Bayes 1763, Laplace 1814) dice que si no hay información para diferenciar entre valores diferentes de θ, se debe dar la misma probabilidad a todos los valores. El principio implica una distribución a priori uniforme para θ.

Observación 21 *Si el soporte de θ es infinito, la distribución a priori será* **impropia***:*

$$f(\theta) \propto 1.$$

Este problema no es tan importante. Lo importante es que exista la distribución a posteriori.

Falta de invarianza

Una crítica más importante es que la distribución uniforme no es invariante en caso de transformación.

Ejemplo 50 *Si se define* $\phi = \log \theta$, *dada una distribución uniforme para* θ, *la distribución de* ϕ *es*

$$f(\phi) \propto e^{\phi},$$

que no es uniforme.

Este problema implica que se debe estar seguro sobre la escala de la variable en la que una distribución a priori uniforme es natural.

La verosimilitud traslada por datos (Data Translated Likelihood)

Esta idea proporciona un método para eligir la escala de medida del parámetro apropriada para definir una distribución a priori uniforme.

Definición 9 *Sea θ unidimensional. Se dice que la verosimilitud $l(\theta|\mathbf{x})$ está trasladada por datos si*

$$l(\theta|\mathbf{x}) = g(\theta - t(\mathbf{x}))$$

para alguna función (estadístico suficiente) $t(\cdot)$.

Ejemplo 51 *$X|\mu \sim \mathcal{N}\left(\mu, \sigma^2\right)$ donde σ^2 es conocido. En este caso,*

$$l(\mu|\mathbf{x}) \propto \exp\left(-\frac{1}{2}(\mu - \bar{x})^2\frac{\sigma}{n}\right)$$

y la verosimilitud está trasladada por datos.

Se ha visto anteriormente que dada una distribución a priori uniforme para μ, la media a posteriori coincide con la media muestral.

Ejemplo 52 *$X|p \sim \mathcal{BI}(n,p)$. Entonces*

$$l(p|x) \propto p^x(1-p)^{n-x}$$

no está trasladada por datos.

Para una verosimilitud así, distintas valores de los datos proporcionan una verosimilitud de la misma forma funcional salvo por un cambio en la posición. Implica que la función de los datos es determinar la posición de la verosimilitud. Si se emplea una distribución a priori uniforme para θ, la forma funcional de la distribución a posteriori es igual para muestras distintas, salvo por cambios en la estimación ($t(\mathbf{x})$) de la posición de θ. Así, la inferencia representa sólo la determinación de la posición de θ, lo que implica que la elección de la distribución a priori uniforme es razonable si la verosimilitud está trasladada por datos.

Si no se puede expresar la verosimilitud como en (9), puede que exista una transformación $\psi = \psi(\theta)$ para que

$$l(\theta|\mathbf{x}) = g(\psi(\theta) - t(\mathbf{x}))$$

y en este caso, es natural elegir una distribución a priori uniforme para ψ.

Ejemplo 53 $X|\phi \sim \mathcal{N}(\mu, 1/\phi)$ *(μ conocido).*

La verosimilitud es

$$\begin{aligned} l(\phi|\mathbf{x}) &\propto \phi^{n/2} \exp\left(-\frac{\phi}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \\ &\propto s^n \phi^{n/2} \exp\left(-\frac{1}{2}ns^2\phi\right) \\ &\quad \textit{donde } s^2 = \tfrac{1}{n}\textstyle\sum_{i=1}^{n}(x_i-\mu)^2 \\ &= \exp\left(\frac{n}{2}(\log\phi + \log s^2) - \frac{n}{2}\exp(\log\phi + \log s^2)\right) \end{aligned}$$

La verosimilitud está trasladada por datos en términos de la transformación $\psi = \log \phi$. Observamos que una distribución uniforme para ψ implica la distribución $f(\phi) \propto \frac{1}{\phi}$ utilizada en el capítulo anterior.

Ejemplo 54 $X|\lambda \sim \mathcal{E}(\lambda)$.

$$\begin{aligned} l(\lambda|\mathbf{x}) &= \lambda^n \exp(-n\lambda\bar{x}) \\ &= \exp(n \log \lambda - n\lambda\bar{x}) \\ &\propto \exp\left(n(\log \lambda + \log \bar{x}) - ne^{\log \lambda + \log \bar{x}}\right) \end{aligned}$$

y la distribución a priori natural para λ es $f(\lambda) \propto \frac{1}{\lambda}$.

Observamos que en este caso, la distribución a posteriori es $\lambda|\mathbf{x} \sim \mathcal{G}(n, n\bar{x})$, cuando la media a posteriori, $\frac{1}{\bar{x}}$, coincide con el EMV.

¿Qué hacer si la verosimilitud no está trasladada por datos?

Esencialmente, sólo las familias normal y log-gamma están de la forma adecuada. Para otras distribuciones, se puede suponer el uso de una transformación normal cuando la verosimilitud está aproximadamente trasladada por datos. Ver Box y Tiao (1973).

Ejemplo 55 *Retomamos el ejemple 52.* $X \sim \mathcal{BI}(n, \theta)$ *y en este caso, definiendo* $Z = \sin^{-1}\sqrt{X/n}$, *se puede demostrar que*

$$Z|\psi \approx \mathcal{N}\left(\psi, \frac{1}{4n}\right)$$

donde $\psi = \sin^{-1}\sqrt{\theta}$.

Se puede concluir que la distribución a priori natural para ψ es (aproximadamente) uniforme, lo que implica que la distribución a priori par θ sería

$$f(\theta) \propto \theta^{1/2}(1-\theta)^{1/2},$$

es decir que $\theta \sim \mathcal{B}(1/2, 1/2)$.

No parece muy natural pero ...

Distribuciones a priori de Jeffreys

Jeffreys introdujó una distribución a priori con una propiedad de invarianza.

Sea θ unidimensional.

Definición 10 *La distribución a priori de Jeffreys es*

$$f(\theta) \propto \sqrt{I(\theta)}$$

donde $I(\theta) = -E_X\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right]$ *es la información esperada de Fisher.*

Si $\phi = \phi(\theta)$,

$$f(\phi) = f(\theta)|\phi'(\theta)|^{-1}$$

y se demuestra que

$$\sqrt{I(\phi)} = \sqrt{I(\theta)}\left|\frac{d\theta}{d\phi}\right|.$$

Si se elige $f(\theta) \propto \sqrt{I(\theta)}$, entonces $f(\phi) \propto \sqrt{I(\phi)}$.

Demostración

En primer lugar demostramos dos resultados útiles.

$$\begin{aligned} E_X\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right] &= 0 \\ I(\theta) &= E_X\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right] \end{aligned}$$

Por definición:

$$\begin{aligned} E_X\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right] &= \int \frac{\partial}{\partial\theta}\log f(x|\theta) f(x|\theta)\, dx \\ &= \int \frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)} f(x|\theta)\, dx \\ &= \int \frac{\partial}{\partial\theta} f(x|\theta)\, dx \\ &= \frac{\partial}{\partial\theta}\int f(x|\theta)\, dx \\ &= \frac{\partial}{\partial\theta} 1 = 0. \end{aligned}$$

$$
\begin{aligned}
I(\theta) &= -E_X\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right] \\
&= -\int \frac{\partial^2}{\partial\theta^2}\log f(x|\theta) f(x|\theta)\, dx \\
&= -\int \frac{\partial}{\partial\theta}\left(\frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)}\right) f(x|\theta)\, dx \\
&= -\int \left(\frac{\frac{\partial^2}{\partial^2\theta}f(x|\theta)}{f(x|\theta)}\right) f(x|\theta)\, dx + \\
&\quad \left(\frac{\left(\frac{\partial}{\partial\theta}f(x|\theta)\right)^2}{f(x|\theta)^2}\right) f(x|\theta)\, dx \\
&= -\int \frac{\partial^2}{\partial\theta^2} f(x|\theta)\, dx + \int \left(\frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)}\right)^2 f(x|\theta)\, dx \\
&= -\frac{\partial^2}{\partial\theta^2}\int f(x|\theta)\, dx + \\
&\quad \int \left(\frac{\partial}{\partial\theta}\log f(x|\theta)\right)^2 f(x|\theta)\, dx \\
&= -\frac{\partial^2}{\partial\theta^2} 1 + E_X\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right] \\
&= E_X\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right]
\end{aligned}
$$

Ahora, consideramos la transformación $\phi = \phi(\theta)$. Luego,

$$\frac{\partial}{\partial \phi} \log f(X|\phi) = \frac{\partial}{\partial \theta} \log f(X|\theta) \frac{\partial \theta}{\partial \phi}.$$

Cuadrando ambos lados y tomando esperanzas tenemos

$$\begin{aligned} E_X\left[\left(\frac{\partial}{\partial \phi} \log f(X|\phi)\right)^2\right] &= E_X\left[\left(\frac{\partial}{\partial \theta} \log f(X|\theta) \frac{\partial \theta}{\partial \phi}\right)^2\right] \\ I(\phi) &= E_X\left[\left(\frac{\partial}{\partial \theta} \log f(X|\theta)\right)^2\right] \left(\frac{\partial \theta}{\partial \phi}\right)^2 \\ &= I(\theta) \left(\frac{\partial \theta}{\partial \phi}\right)^2 \end{aligned}$$

Entonces, si elegimos la densidad a priori $f(\theta) \propto \sqrt{I(\theta)}$ y transformamos $\phi = \phi(\theta)$, por la regla de cambio de variables tenemos $f(\phi) \propto \sqrt{I(\phi)}$.

$\diamond$

La distribución a priori de Jeffreys es invariante en el sentido de que la inferencia no depende de la escala elegida para el parámetro.

Ejemplo 56 $X|\theta \sim \mathcal{BI}(n, \theta)$.

$$\begin{aligned}
\log f(X|\theta) &= c + X \log\theta + \\
&\quad +(n-X)\log(1-\theta) \\
\frac{d}{d\theta}\log f(X|\theta) &= \frac{X}{\theta} - \frac{(n-X)}{(1-\theta)} \\
\frac{d^2}{d\theta^2}\log f(X|\theta) &= -\frac{X}{\theta^2} - \frac{(n-X)}{(1-\theta)^2} \\
E\left[\frac{d^2}{d\theta^2}\log f(X|\theta)\right] &= -n\left(\frac{1}{\theta} + \frac{1}{1-\theta)}\right) \\
I''(\theta) &\propto \frac{1}{\theta(1-\theta)}
\end{aligned}$$

Entonces, la distribución a priori de Jeffreys es $f(\theta) \propto \sqrt{\frac{1}{\theta(1-\theta)}}$ *o* $\theta \sim \mathcal{B}(1/2, 1/2)$. *Esta distribución es exactamente la distribución calculada anteriormente en el Ejemplo 55.*

Ejemplo 57 *$X|\mu \sim \mathcal{N}(\mu, \sigma^2)$, con σ^2 conocido.*

$$\begin{aligned} \log f(X|\mu) &= c - \frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2 \\ \frac{d^2}{d\mu^2}\log f(X|\mu) &= -\frac{1}{\sigma^2} \end{aligned}$$

Se pone $f(\mu) \propto 1$, una distribución uniforme.

Ejemplo 58 *Supongamos ahora que μ es conocido y σ^2 desconocido. Pongamos $\tau = \sigma^2$.*

$$\begin{aligned} \log f(X|\tau) &\propto -\frac{1}{2}\log\tau - \frac{(X-\mu)^2}{2\tau} \\ \frac{d}{d\tau}\log f(X|\tau) &= -\frac{1}{2\tau} + \frac{(X-\mu)^2}{2\tau^2} \\ \frac{d^2}{d\tau^2}\log f(X|\tau) &= \frac{1}{2\tau^2} - \frac{(X-\mu)^2}{\tau^3} \\ -E\left[\frac{d^2}{d\theta^2}\log f(X|\tau)\right] &= -\frac{1}{2\tau^2} + \frac{\tau}{\tau^3} \\ &= \frac{1}{2\tau^2} \end{aligned}$$

La distribución a priori de Jeffreys es $f(\tau) \propto \frac{1}{\tau}$.

Observación 22 *$\tau = \sigma^2$ y entonces*

$$\frac{d\tau}{d\sigma} = 2\sigma.$$

Luego

$$f(\sigma) \propto \frac{1}{\sigma^2}|2\sigma| \propto \frac{1}{\sigma}.$$

Observación 23 *Si se transforma $\nu = \log\tau$, entonces*

$$\frac{d\nu}{d\tau} = \frac{1}{\tau} \Rightarrow \frac{d\tau}{d\nu} = e^{\nu}.$$

Luego

$$f(\nu) \propto \frac{1}{e^{\nu}} e^{\nu} \propto 1.$$

La distribución a priori de Jeffreys es uniforme en el logaritmo de τ.

Observación 24 *Si ϕ es la precisión, $\phi = 1/\tau$ y la distribución de Jeffreys para ϕ es*

$$f(\phi) \propto 1/\phi.$$

Estimadores a posteriori, la distribución de Jeffreys y la EMV

En muchos casos, la media a posteriori de $\theta|\mathbf{x}$ es igual al EMV cuando se ha utilizado una distribución a priori de Jeffreys.

Ejemplo 59 *Para datos normales,*

$$\bar{X}|\mu \sim \mathcal{N}(\mu, \sigma^2/n).$$

Entonces, dada la distribución a priori de Jeffreys $f(\mu) \propto 1$*, la distribución a posteriori será*

$$\mu|\mathbf{x} \sim \mathcal{N}(\bar{x}, \sigma^2/n)$$

con media a posteriori igual a la EMV de μ*.*

El resultado no pasa siempre.

Ejemplo 60 *Volviendo al Ejemplo 56, dada una muestra binomial con x caras y $n-x$ cruces, la distribución a posteriori dada la distribución a priori de Jeffreys es*

$$\begin{aligned} f(\theta|\mathbf{x}) &\propto \theta^{x-1/2}(1-\theta)^{n-x-1/2} \\ \theta|\mathbf{x} &\sim \mathcal{B}(x+1/2, n-x+1/2) \end{aligned}$$

Entonces, la media a posteriori es

$$E[\theta|\mathbf{x}] = \frac{x+1/2}{n+1} \neq \frac{x}{n},$$

el EMV de θ.

Es interesante preguntar ?qué distribución a priori proporciona la media a posteriori igual al estimador máximo verosimil?

La distribución de Haldane

La distribución a priori de Haldane es

$$f(\theta) \propto \frac{1}{\theta(1-\theta)}.$$

La distribución a posteriori en este caso es

$$\begin{aligned} f(\theta|\mathbf{x}) &\propto \theta^{x-1}(1-\theta)^{n-x-1} \\ \theta|\mathbf{x} &\sim \mathcal{B}(x, n-x). \end{aligned}$$

Ahora, la media a posteriori es

$$E[\theta|\mathbf{x}] = \bar{x}$$

igual al EMV.

Una interpretación de los parámetros de la distribución a priori $\mathcal{B}(\alpha, \beta)$ es como el número de cruces y caras en un experimento equivalente. La distribución de Haldane corresponde al caso de $\mathcal{B}(0, 0)$. Los conocimientos a priori son equivalentes a no tener ninguna información.

Extensión a parámetros multivariables

Se puede generalizar la distribución de Jeffreys a situaciones multivariables. Se aplica la Definición 10 con la información definida como

$$I(\boldsymbol{\theta}) = \left| E_X \left[\frac{d^2}{d\boldsymbol{\theta}^2} \log f(X|\boldsymbol{\theta}) \right] \right|$$

el determinante de la matriz de información de Fisher esperada.

Ejemplo 61 *Sea* $X|\mu, \sigma^2 \sim \mathcal{N}(\mu, \sigma^2)$. *Entonces, escribiendo* $\tau = \sigma^2$,

$$\log f(X|\mu, \tau) = c - \frac{1}{2}\log(\tau) - \frac{1}{2\tau}(X-\mu)^2$$

y derivando, se tiene

$$\begin{aligned} \frac{\partial^2}{\partial \mu^2} \log f &= -\frac{1}{\tau} \\ \frac{\partial^2}{\partial \mu \partial \tau} \log f &= -\left(\frac{X-\mu}{\tau^2}\right) \\ \frac{\partial^2}{(\partial \tau^2)^2} &= \frac{1}{2\tau^2} - \frac{(X-\mu)^2}{\tau^3} \end{aligned}$$

Si $\mathbf{J}$ *es la matriz de información de Fisher,*

$$\mathbf{J} = \begin{pmatrix} -\frac{1}{\tau} & -\left(\frac{X-\mu}{\tau^2}\right) \\ -\left(\frac{X-\mu}{\tau^2}\right) & \frac{1}{2\tau^2} - \frac{(X-\mu)^2}{\tau^3} \end{pmatrix}$$

y tomando esperanzas,

$$E[\mathbf{J}] = \begin{pmatrix} -\frac{1}{\tau} & 0 \\ 0 & -\frac{1}{2\tau^2} \end{pmatrix}$$

y luego $I(\mu, \sigma^2) = |E[\mathbf{J}]| \propto \frac{1}{\tau^3}$.

Entonces, la distribución a priori de Jeffreys es

$$f(\mu, \tau) \propto \sqrt{\frac{1}{\tau^3}}.$$

Existen otras posibilidades en situaciones multivariables. Una de las más usadas es suponer una distribución a priori en la que los parámetros sean independientes y usar el producto de las distribuciones de Jeffreys para cada parámetro: $f(\boldsymbol{\theta}) = \prod f(\theta_i)$.

Ejemplo 62 *Volviendo al Ejemplo 61, tenemos las distribuciónes de Jeffreys:*

$$f(\mu) \propto 1$$

y

$$f(\tau) \propto \frac{1}{\tau}$$

como vimos en los Ejemplos 57 y 58.

Una distribución razonable para (μ, τ) *es*

$$f(\mu, \tau) \propto \frac{1}{\tau}.$$

Esta es la distribución que se utiliza habitualmente.

Otras posibilidades

- máxima entropía.

  Sea θ univariable y discreta. Si $f(\theta)$ es una densidad, se define

  $$e(f) = -\sum_{\Theta} f(\theta_i) \log f(\theta_i)$$

  la **entropía** de la distribución.

  Distribuciones de máxima entropía son de mínima información.

  Sin otras restricciones, la distribución de máxima entropía es la distribución uniforme. Con restricciones de momentos por ejemplo $E[g_i(\theta)] = m_i$ tenemos

  $$f(\theta) \propto \exp\left(\sum \lambda_i g_i(\theta)\right)$$

  donde se determinan las constantes usando las restricciones.

Los métodos pueden extenderse al caso continuo pero la definición de la entropía

$$e(f) = -\int f \log f \, d\mu$$

depende de la medida μ.

- distribuciones a priori de referencia.

  Basada en minimizar la información esperada de un experimento.

- distribuciones de Haar.

  Basadas en consideraciones de simetría.

Problemas con distribuciones no informativas

- posibilidades de distribuciones a posteriori impropias.

  Ejemplo 63 *Vamos a lanzar una moneda con* $\theta = P(cruz)$*. Utilizamos la distribución inicial de Haldane:*

  $$f(\theta) \propto \frac{1}{\theta(1-\theta)}$$

  equivalente a $\mathcal{B}(0,0)$ *que es impropia.*

  Si se observan n *cruces en* n *tiradas, la distribución a posteriori será*

  $$\theta|x \sim \mathcal{B}(n,0)$$

  que también es una distribución impropia.

- el principio de verosimilitud.

Ejemplo 64 *Suponiendo que vamos a generar datos de una binomial* $X|\theta \sim \mathcal{BI}(n,\theta)$, *hemos visto en el Ejemplo 56 que la distribución a priori de Jeffreys es* $\theta \sim \mathcal{B}(1/2,1/2)$.

Supongamos ahora que vamos a generar datos de una binomial negativa. Entonces

$$\begin{aligned}
\log f(X|\theta) &= c + r\log\theta + \\
&\quad + X\log(1-\theta) \\
\frac{\partial \log f(X|\theta)}{\partial\theta} &= \frac{r}{\theta} - \frac{X}{1-\theta} \\
\frac{\partial^2 \log f(X|\theta)}{\partial\theta^2} &= -\frac{r}{\theta^2} - \frac{X}{(1-\theta)^2} \\
-E\left[\frac{\partial^2 \log f(X|\theta)}{\partial\theta^2}\right] &= \frac{r}{\theta^2} + \frac{r}{\theta(1-\theta)} \\
&= \frac{r}{\theta^2(1-\theta)}
\end{aligned}$$

Entonces la distribución de Jeffreys es

$$f(\theta) \propto \frac{1}{\theta(1-\theta)^{1/2}}.$$

Pero, volviendo al Ejemplo 12 si tenemos la información que hemos observado 9 cruces en 12 tiradas, necesitamos saber el diseño del experimento (binomial o binomial negativa) que nos proporcionó estos datos antes de definir la distribución inicial. La distribución a posteriori de θ *será* $\mathcal{B}(9{,}5, 3{,}5)$ *suponiendo datos binomiales y* $\mathcal{B}(9, 3{,}5)$ *para datos binomiales negativos.*

Por supuesto, el uso de distribuciones de Jeffreys no cumple con el principio de verosimilitud.