CAPÍTULO 4. 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 se elige la distribución a priori que las represente bien?

Existen varias posibilidades.

Distribuciones Conjugadas

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

Ejemplo 31 *Sea θ la probabilidad de que una tirada de una moneda sea cruz. Estimamos $E[\theta] = 0{,}4$ y supongamos que nuestros conocimientos equivalen a una muestra de 100 tiradas.*

Para una distribución a priori $\mathcal{B}(\alpha, \beta)$, se tiene

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

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

Ejemplo 32 *Sea $X|\theta \sim \mathcal{E}(\theta)$ el tiempo entre llegadas en un supermercado. Una distribución a priori conjugada será $\theta \sim \mathcal{G}(\alpha, \beta)$.*

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 implica que $\alpha = 4$ y $\beta = 3$.

Observación 32 *Fijando más momentos, cuantiles etc. implica la necesidad de considerar mixturas de distribuciones conjugadas.*

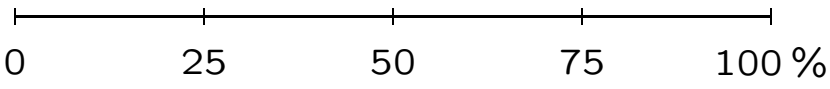
Distribuciones Subjetivas

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

¿Cómo solicitar las predicciones?

Existen varios métodos para solicitar probabilidades o quantiles:

1 Medida directa: escala de probabilidades.

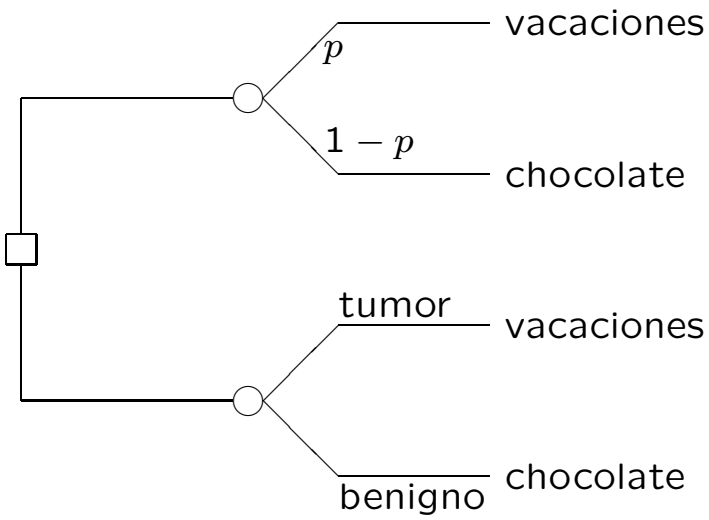


Una escala de probabilidades

Se exige al experto marcar su probabilidad en la escala.

Problemas del efecto del centro (*centring*) y problemas con probabilidades pequeñas.

2 Loterías

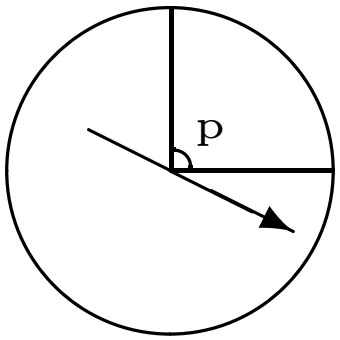


Se quiere solicitar P(tumor).

El experto debe elegir una de los dos loterías. Se varía el valor de p hasta que sea indiferente; $p = p_E$. Luego P(tumor) $= p_E$.

Problemas éticas. El método no es muy natural. Problemas con actitudes al riesgo con algunas loterías.

3 Rueda de fortuna



Se pregunta al experto ?Cuál es más probable? Ganar en la rueda o tener, el paciente, un tumor. Variando el valor de p se llega a un punto de igualdad. Luego

$$P(\text{tumor}) = \frac{p}{2\pi}.$$

4 Otras posibilidades: comparaciones entre parejas, AHP etc. Ver Cooke (1991), Spetzler y Stael von Holstein (1975).

Honestidad y reglas estrictamente 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 se debe definir $R(S, p)$?.

Supongamos que el experto quiere maximizar su sueldo esperado. Si q es su verdadera probabilidad, su sueldo esperado cuando dice una probabilidad p es

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

Una regla (estrictamente) propia o (*strictly*) *proper scoring rule*, (Savage 1971) es una regla $R(S, p)$ para que el experto máximiza su sueldo esperado si (y sólo si) $p = q$.

Ejemplo 33 $R(S,p) = 1 - |S - p|$

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 34 *$R(S,p) = 1-(S-p)^2$ es la regla de Brier (1950).*

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

R es una regla estrictamente propia.

Observación 33 *Existen otras reglas estrictamente propias: la regla logaritmica,*

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

o le regla esférica

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

Ver Winkler (1986).

Se han definido reglas propias para la solicitación de distribuciones de variables continuas (Buehler 1971) y cuantiles (Matheson y Winkler 1976).

Ejemplo 35 *El experto debe proporcionar un estimador puntual de una variable continua X. Sea e su estimador y define la regla*

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

Sea $f(x)$ la distribución del experto para X:

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

El experto minimiza su perdida esperada si proporciona su $b/(a+b)\times 100\,\%$ cuantil. Ver Raiffa y Schlaifer (1961).

Observación 34 *A menudo, la utilidad de dinero no es lineal. Ver Kadane y Winkler (1988) para extensiones a este caso.*

Problemas con predicciones subjetivas

Existen muchos problemas en solicitar distribuciones subjetivas reales. Los humanos usan métodos heurísticos para medir su incertidumbre que típicamente inducen sesgos o incoherencias.

- sesgos motivacionales.

- sesgos cognitivos:
  - disponibilidad
  - anclaje y ajuste
  - representatividad
  - la falacia de la tasa base

Disponibilidad

Ejemplo 36 *(Russo y Shoemaker 1989)*

Para cada pareja dice cúal de las dos opciones causa más muertes anuales.

- *cáncer del estómago o accidentes de coches*

- *tísis o incendios y fuego*

Causa de muerte	*Elección*	*Total anual (USA /1000)*	*# reportajes en periódicos*
cáncer	*14 %*	*95*	*46*
accidente	*86 %*	*1*	*137*
tísis	*23 %*	*4*	*0*
incendios	*77 %*	*5*	*0*

En este ejemplo la gente tiene más información sobre accidentes que sobre cáncer y entonces esta opción es más disponible.

Ver Tversky y Kahneman (1973) para más ejemplos.

Anclaje y ajuste

La gente usa una información disponible para hacer una estimación inicial (por ejemplo una media) y luego ajusten su estimación inicial.

Ejemplo 37 *(Kahneman y Tversky 1974)*

Se quería solicitar una estimación del porcentaje de paises africanos dentro de los paises del ONU. En el grupo 1 (2) se preguntó ¿piensas que el porcentaje es más o menos que 65 % (10 %)? Luego se pidió una estimación del porcentaje a ambos grupos.

La estimación promedia del grupo 1 fue un 45 % y del grupo 2 un 25 %.

Un anclaje aleatorio y irrelevante ha influido las estimaciones de cada grupo.

Representatividad

Ejemplo 38 *"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 y 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 eligen la opción 3. Además, mucha gente dice que la opción 7 es más probable que la opción 4.

Este último es imposible ya que para dos sucesos A y B,

$$P(A \cap B) \leq \text{mín}\{P(A), P(B)\}.$$

Este problema ilustra tanto el uso de la heurística de representatividad como la falacia de la tasa base (*base rate fallacy*).

Ver Kahneman et al (1982) por más ejemplos.

La falacia de la tasa baja

Ejemplo 39 *(Tversky y Kahneman 1980):*

Un taxi atropelló a un peaton en la ciudad de Darlington. En Darlington hay sólo dos empresas que operan un servicio de taxis. La primera empresa tiene taxis verdes y la otra utiliza taxis azules. Se sabe que:

1. *Un 85 % de los taxis en Darlington son verdes y los demás son azules.*

2. *Un testigo dice que el taxi era azul.*

3. *En varias pruebas hechas en las mismas condiciones de la noche del accidente, el testigo podía identificar los dos colores correctamente un 80 % de las veces.*

Estimar la probabilidad de que el taxi fuera azul.

La respuesta típica es aproximadamente 80 %. No obstante, si A es el suceso "el taxi es azul" y a es el suceso "el testigo dice que es azul", or el teorema de Bayes se tiene

$$\begin{aligned} P(A|a) &= \frac{P(a|A)P(A)}{P(a|A)P(A) + P(a|\bar{A})P(\bar{A})} \\ &= \frac{0{,}8 \times 0{,}15}{0{,}8 \times 0{,}15 + 0{,}2 \times 0{,}85} \\ &= 0{,}41 \end{aligned}$$

Criterios para evaluar predicciones subjetivas

Algunos criterios (Lichtenstein et al 1982) son los siguientes:

- honestidad

  Se quiere que el experto diga sus verdaderas opiniones.

- coherencia

  Las predicciones deben cumplir las leyes de probabilidad.

- consistencia

  Se el experto no padece de información nueva, sus predicciones no deben cambiar.

- precisión (*calibration*)

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

- información

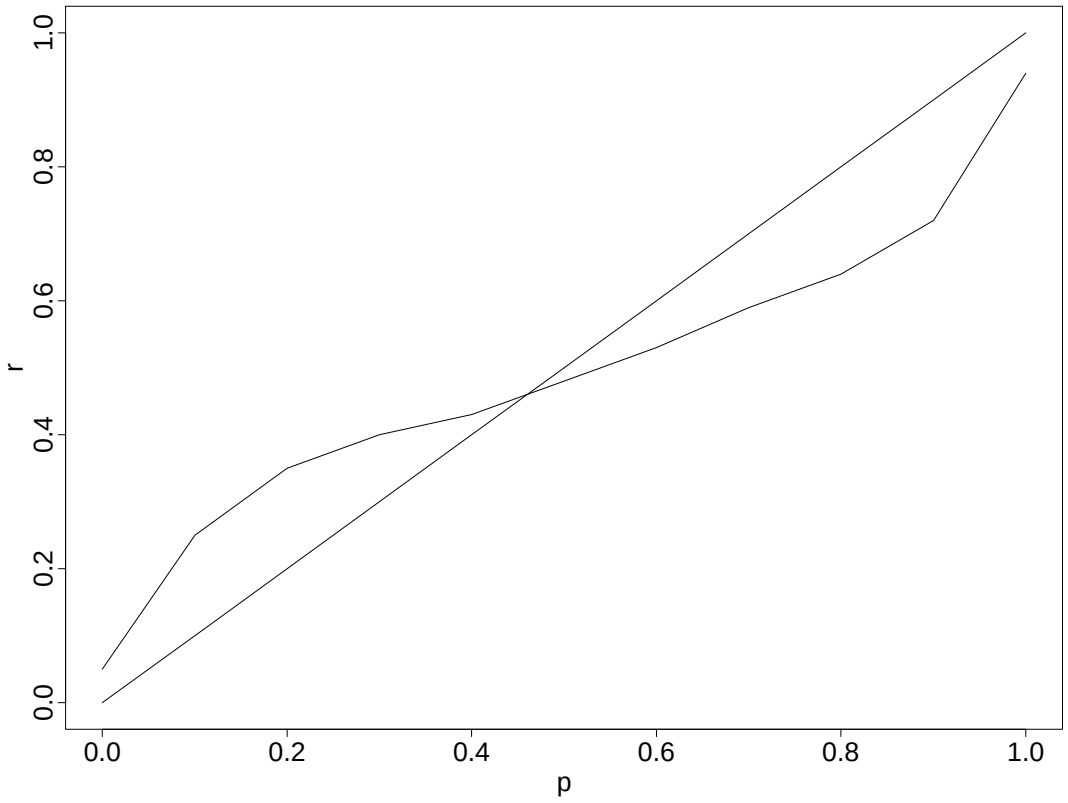
  Si, en Madrid, llueve aproximadamente 50 dias al año, un experto que dice

  $$P(\text{llover mañana}) = 50/364$$

  no es informativo.

La curva de precisión

Supongamos que un previsor dice su probabilidad de lluvia cada día durante un periodo largo. *The calibration curve* es un gráfico de las frecuencias observadas de días lluviosos, r_j, frente a las probabilidades utilizadas, p_j.



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

Medidas numéricas 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 calcula 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 promedia de las predicciones.

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

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

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, $r_j = 0{,}5$.

Observación 35 *Se puede dividir cualquier regla estrictamente propia en dos partes de manera análoga. (De Groot y Fienberg 1982).*

Ejemplo 40 *Wiper (1987,1990) 12 expertos dicen si 50 afirmaciones son verdaderas o falsas y proporcionan sus probabilidades de que acierten como un porcentaje entre 50 % (= ni idea) y 99 %.*

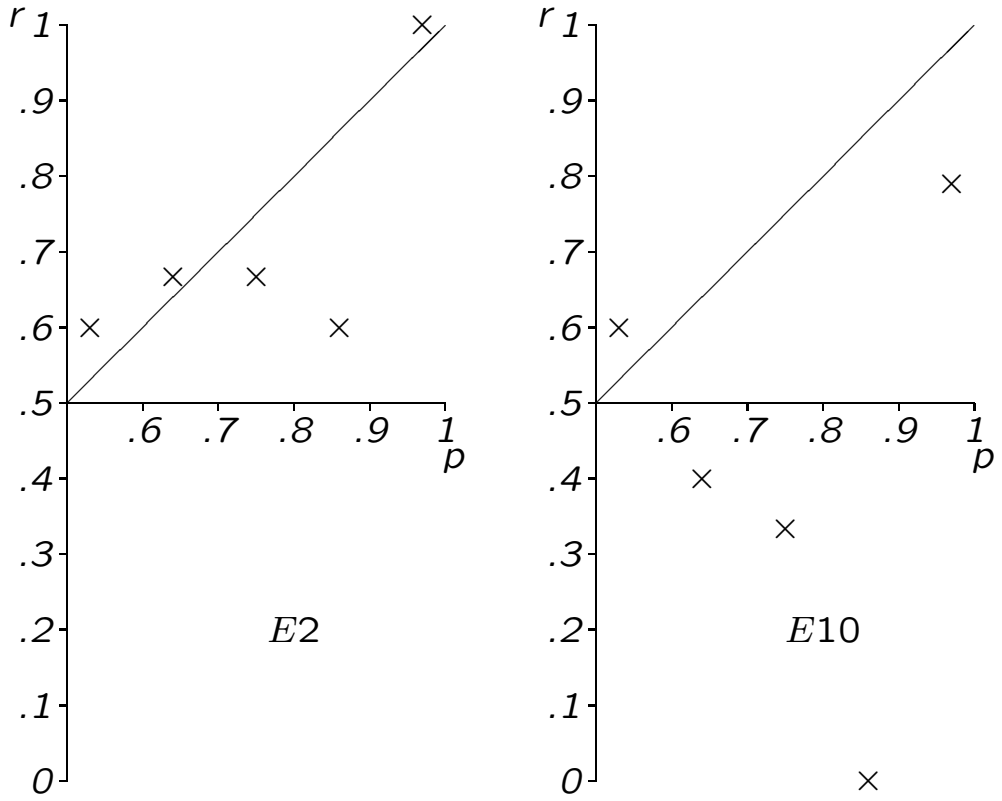
En este caso se agrupan las probabilidades en 5 clases: $\mathbf{p} = \{0{,}53, 0{,}64, 0{,}75, 0{,}86, 0{,}97\}$.

E	p_i	0,53	0,64	0,75	0,86	0,97
2	n_i	25	6	6	5	8
	$n_i r_i$	15	4	4	3	8
3	n_i	25	5	10	5	5
	$n_i r_i$	16	1	3	2	4
10	n_i	10	5	15	1	19
	$n_i r_i$	6	2	5	0	15

En la siguiente tabla, se muestran los valores de las medidas de precisión, información y la regla de Brier.

E	C	I	*Brier*
2	,0093	,1973	,7934
3	,0900	,2132	,6968
10	,1059	,2018	,6923

El experto 2 es mucho más preciso pero menos informativo que los otros dos.



Medidas basadas en contrastes de hipótesis

Existen otras medidas de precisión y información para variables continuas, basadas en p-valores clásicos.

Seguimos Cooke et al (1988) y Cooke (1991).

Supongamos que el experto proporciona su mediana y un intervalo de 90 % para una sucesión de variables $X_1, \ldots, X_n$. Teoricamente, si el experto es preciso, aproximadamente 5 % de los veces, el valor de la variable será menor que su primer cuantil, 45 % de los veces caerán entre el primer cuantil y la mediana etc.

Si el experto es preciso, la distribución teórica de las frecuencias en cada clase es

$$(p_1, \ldots, p_4) = (0{,}05, 0{,}45, 0{,}45, 0{,}05)$$

Comparamos las frecuencias, $(f_1, \ldots, f_4)$,en la muestra, con las frecuencias teóricas.

Se contrasta $H_0 : f = p$ frente a la alternativa $H_1 : f \neq p$ mediante un contraste ji-cuadrado.

Sea $S = 2\sum_{i=1}^{4} n_i f_i \log \frac{f_i}{p_i}$ donde n_i es el número de valores en cada clase y $\sum_{i=1}^{4} n_i = n$. Luego, bajo H_0, se tiene $S \sim \chi^2_3$.

Observación 36 *Se usa el mismo método si el experto proporciona probabilidades para sucesos. Cooke (1991) desarolla una teoría de reglas propias basada en combinar el p-valor con una medida de información.*

Si el experto proporciona sus distribuciones enteras, se emplea un contraste Kolmogorov Smirnov. Ver Wiper et al (1994).

Ejemplo 41 *Retomamos el Ejemplo 40.*

Se tienen 5 (independientes) frecuencias teóricas $\mathbf{p} = \{0{,}53, 0{,}64, 0{,}75, 0{,}86, 0{,}97\}$. *Para cada* p_i *hay dos clases; éxitos y fracasos.*

El estadístico ji-cuadrado es

$$S = 2 \sum_{i=1}^{5} \left(n_i r_i \log \frac{r_i}{p_i} + n_i (1 - r_i) \log \frac{1 - r_i}{1 - p_i} \right).$$

Bajo $H_0 : r = p$, *se tiene* $S \sim \chi^2_5$.

La tabla muestra los p valores para cada experto.

E	2	3	10
p	,7	,00	,00

Sólo parece que el experto 2 es razonablemente 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 1783, Laplace 1812) 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 37 *Si el soporte de θ es infinito, la distribución a priori será* **impropia***:*

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

Lo importante es que exista la distribución a posteriori.

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

Ejemplo 42 *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.

La verosimilitud trasladada por datos (Data Translated Likelihood)

Este idea, (Box y Tiao, 1973) proporciona un método para eligir la escala de medida apropriada del parámetro para definir una distribución a priori uniforme.

Definición 6 *Sea θ unidimensional. Se dice que la verosimilitud $l(\theta|\mathbf{x})$ está trasladada por datos (TPD) si*

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

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

Observación 38 *Es una generalización de un parámetro de posición (location parameter). Se dice que θ es un parámetro de posición si:*

$$f(x|\theta) = g(x - \theta).$$

Ejemplo 43 *$X|\mu \sim \mathcal{N}\left(\mu, \sigma^2\right)$ donde σ^2 es conocido. En este caso,*

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

y la verosimilitud está TPD.

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

Ejemplo 44 *$X|p \sim \mathcal{BI}(n, p)$. Entonces*

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

no está TPD.

Para una verosimilitud TPD, distintas valores de los datos proporcionan una verosimilitud de la misma forma funcional salvo por un cambio en la posición. $\Rightarrow$ la función de los datos es determinar la posición de la verosimilitud.

Dada una distribución a priori uniforme para θ, la forma funcional de la distribución a posteriori es igual para dos muestras distintas, salvo por cambios en la estimación ($t(\mathbf{x})$) de la posición de θ. La inferencia representa sólo la determinación de la posición de θ, implicando que la elección de la distribución a priori uniforme es razonable si la verosimilitud está TPD.

Si no se puede expresar la verosimilitud como en (6), 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 45 $X|\phi \sim \mathcal{N}(\mu, 1/\phi)$ *(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 TPD en términos de $\psi = \log\phi$*. Observamos que una distribución uniforme para* ψ *implica que la distribución para* ϕ *es*

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

como se ha utilizado en el capítulo anterior.

Ejemplo 46 $X|\lambda \sim \mathcal{E}(\lambda)$.

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

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.*

Problemas y Extensiones

- ¿Qué hacer si la verosimilitud no está TPD?

  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 47 *Retomamos el ejemple 44.* $X \sim \mathcal{BI}(n, \theta)$ *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 ...

- ¿Cómo extender a situaciones multivariables?

  Ver Box y Tiao (1973), Kass (1990).

Parámetros de escala

Se dice que $X|\theta$ es una densidad de escala (*scale density*) si

$$f(x|\theta) = \frac{1}{\theta} g\left(\frac{x}{\theta}\right) \quad \text{para } \theta > 0$$

En este caso θ es un parámetro de escala.

Ejemplo 48 *$X|\sigma \sim \mathcal{N}(0, \sigma^2)$ es una densidad de escala con parámetro σ. $X|\theta \sim \mathcal{E}(1/\theta)$ es una densidad de escala.*

Si queremos una distribución a priori no informativa para esta situación, supongamos que, en lugar de observar X, observamos $Y = cX$ (un cambio de escala).

Sea $\phi = c\theta$ y entonces,

$$f(y|\phi) = \frac{1}{\phi} g\left(\frac{y}{\phi}\right).$$

Las dos densidades son de la misma forma y los espacios muestrales y paramétricos son iguales $\Rightarrow$ parece natural que θ y ϕ tienen la misma densidad.

Pero, por la formula para transformación de distribuciones,

$$\begin{aligned} f_\phi(\phi) &= f_\theta(\theta^{-1}(\phi)) \left|\frac{d\theta}{d\phi}\right| \\ &= f_\theta(\phi/c)c^{-1} \end{aligned}$$

y entonces, si $f_\phi(\cdot) = f_\theta(\cdot) = f(\cdot)$, se tiene

$$f(\phi) = f(\phi/c)c^{-1} \forall \phi$$

Fijando $\phi = c$, se tiene $f(c) = f(1)/c$. Este resultado debe ser verdad para todo $c > 0$, es decir

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

Observación 39 *Es una densidad impropia:*

$$\int_0^\infty \frac{1}{\theta}\, d\theta = \infty.$$

Ejemplo 49 *Sea* $X|\sigma \sim \mathcal{N}(\mu, \sigma^2)$*. Entonces, la distribución no informativa natural es* $f(\sigma) = 1/\sigma$*.*

Sea $\phi = 1/\sigma^2$*. La distribución implicada para* ϕ *es*

$$f(\phi) \propto \sqrt{\phi} \left| -\frac{1}{2\phi^{3/2}} \right| \propto \frac{1}{\phi}$$

Anteriormente en el Capítulo 3, se ha utilizado esta distribución para reproducir los resultados clásicos.

Distribuciones a priori de Jeffreys

Jeffreys (1946) introdujó una distribución a priori con una propiedad de invarianza.

Sea θ unidimensional.

Definición 7 *La distribución a priori de Jeffreys es*

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

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

Es consistente con respeto a transformaciones como demuestra el siguiente teorema

Teorema 10 *Sea* $f(\theta) \propto \sqrt{I(\theta)}$ *la distribución a priori de Jeffreys para* θ*. Luego si* $\phi = \phi(\theta)$*,*

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

donde $I(\phi) = -E_X\left[\frac{d^2}{d\phi^2}\log f(X|\phi)\right]$.

Demostración

En primer lugar se demuestran dos lemas 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, sea $\phi = \phi(\theta)$. *Entonces,*

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

Cuadrando ambos lados y tomando esperanzas se tiene

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

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

$\diamond$

Ejemplo 50 $X|\theta \sim \mathcal{BI}(n, \theta)$.

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

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 47.

Ejemplo 51 *$X|\mu \sim \mathcal{N}(\mu, \sigma^2)$, con σ^2 conocido.*

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

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

Ejemplo 52 *Supongamos ahora que μ es conocido y σ^2 desconocido. Pongamos $\tau = \sigma^2$.*

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

que implica que la distribución a priori de Jeffreys para $\tau = \sigma^2$ es $f(\tau) \propto \frac{1}{\tau}$.

Observación 40 *Si se transforma* $\nu = \log \tau$, *entonces* $f(\nu) \propto 1$. *La distribución a priori de Jeffreys es uniforme en el logaritmo de* τ.

Observación 41 *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 53 *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*.*

No obstante, este resultado no pasa siempre.

Ejemplo 54 *Volviendo al Ejemplo 50, 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.

Ejemplo 55 *Una alternativa es la distribución a priori de Haldane (1948),*

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

La distribución a posteriori en este caso es $\theta|\mathbf{x} \sim \mathcal{B}(x, n-x)$ y ahora, la media a posteriori es igual al EMV.

Una interpretación de los parametrós 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 equivalen a no tener ninguna información.

Extensión a parámetros multivariables

Se puede generalizar la distribución de Jeffreys a situaciones multivariables. Se aplica la Definición 7 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 56 *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 57 *Volviendo al Ejemplo 56, tenemos las distribuciónes de Jeffreys:*

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

y

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

como vimos en los Ejemplos 51 y 52.

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

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

Esta es la distribución que se utiliza habitualmente.

Máxima Entropía

La idea (Jaynes 1968,1983) es buscar la distribución a priori menos informativa en la presencia de información parcial.

Sea θ univariable y discreta. Si $P(\theta)$ es cualquier distribución, se define

$$e(P) = -\sum_{i \in \Theta} P(\theta_i) \log P(\theta_i)$$

la **entropía** de la distribución.

Si $P(\theta = \theta_i) = 1$ para algún valor $\theta_i \in \Theta$, entonces, $e(P) = 0$. No hay incertidumbre.

Al contrario, si $P(\theta_i) = 1/|\Theta|$, es decir una distribución uniforme, entonces

$$e(P) = -\sum_{i \in \Theta} \frac{1}{|\Theta|} \log \frac{1}{|\Theta|} = \log |\Theta|,$$

la máxima entropía.

Distribuciones de máxima entropía (maxent) son de mínima información.

Supongamos que tenemos información parcial sobre θ de forma

$$E[g_k(\theta)] = \sum_{i \in \Theta} P(\theta_i) g_k(\theta_i) = \mu_k$$

para $k = 1, \ldots, m$.

Observación 42 *Incluye restricciones de momentos:* $g_1(\theta) = \theta$ *y* $g_k(\theta) = (\theta - \mu_1)^k$. *También incluye restricciones de cuantiles:*

$$g_k(\theta) = I_{(-\infty, z_k]} \Rightarrow E[g_k(\theta)] = P(\theta \leq z_k).$$

Se puede demostrar que la solución que maximiza la entropía dadas las restricciones es

$$P(\theta_i) = \frac{\exp\left(\sum_{k=1}^m \lambda_k g_k(\theta_i)\right)}{\sum_{j \in \Theta} \exp\left(\sum_{k=1}^m \lambda_k g_k(\theta_j)\right)}$$

donde se determinan las constantes λ_k usando las restricciones.

Ejemplo 58 *Sea* $X|N \sim \mathcal{BI}(N, 1/2)$. *Se sabe que* $N \geq 1$ *y se estima que* $E[N] = 10$.

Intentamos hallar la distribución maxent para N.

$$\begin{aligned} P(N = n) &= \frac{\exp(\lambda_1 n)}{\sum_{j=1}^{\infty} \exp(\lambda_1 j)} \\ &= \exp(\lambda_1 n) \frac{1 - \exp(\lambda_1)}{\exp(\lambda_1)} \\ &= (1 - e^{\lambda_1}) \exp(\lambda_1 (n-1)) \end{aligned}$$

Es decir que $N-1$ *tiene una densidad geométrica con parámetro* $1 - e^{\lambda_1}$ *y entonces se tiene*

$$E[N] = 1 + \frac{e^{\lambda_1}}{1 - e^{\lambda_1}}$$

y fijando $E[N] = 10$, *se tiene* $e^{\lambda_1} = \frac{9}{10}$, *es decir que la distribución a priori maxent para* N *es* $N - 1 \sim \mathcal{GE}(9/10)$.

Extensión al caso continuo

Los métodos pueden extenderse al caso continuo pero es más complicado porque la definición de la entropía

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

depende de la medida base μ.

Una posibilidad (Jaynes 1968) es definir

$$e(f) = -\int f(\theta) \log \frac{f(\theta)}{f_0(\theta)} \, d\theta$$

donde $f_0(\theta)$ es la distribución a priori de Jeffreys para θ. Entonces, dadas las restricciones $E[g_k(\theta)] = \lambda_k$ la solución es

$$f(\theta) = \frac{f_0(\theta) \exp\left(\sum_{k=1}^m \lambda_k g_k(\theta)\right)}{\int f_0(\theta) \exp\left(\sum_{k=1}^m \lambda_k g_k(\theta)\right) \, d\theta}$$

análoga a la solución en el caso discreto.

Un problema

Es posible de que no exista una solución.

Ejemplo 59 *Sea* $X|\mu \sim \mathcal{N}(\mu, 1)$. *Impongamos la restricción* $E[\mu] = c$.

Se ha visto en el Capítulo 3 que la distribución a priori de Jeffreys es $f(\mu) \propto 1$. *Luego, la distribución maxent es*

$$f(\mu) = \frac{\exp(\lambda_1 \mu)}{\int_{-\infty}^{\infty} \exp(\lambda_1 \mu) d\mu}$$

y no existe ninguna densidad maxent para μ.

Observación 43 *Se tienen los mismos problemas si el soporte de* θ *es infinito y sólo se fijan algunas cuantiles de* θ.

Otras posibilidades

- distribuciones a priori de referencia (Bernardo 1979).

  Sea $X|\boldsymbol{\theta} \sim f(\cdot|\boldsymbol{\theta})$ y supongamos una muestra $\mathbf{x}^{(n)} = (x_1, \ldots, x_n)$.

  La cantidad de información sobre $\boldsymbol{\theta}$ proporcionada por muestras repetidas de tamaño n depende de la distribución a priori $\pi(\boldsymbol{\theta})$ elegida, es decir

$$I(\mathcal{X}^n, \pi) = \int_{\boldsymbol{\Theta}} \int_{\mathcal{X}^n} \pi(\boldsymbol{\theta}) f(\mathbf{x}^{(n)}|\boldsymbol{\theta}) \log \frac{\pi(\boldsymbol{\theta}|\mathbf{x}^{(n)})}{\pi(\boldsymbol{\theta})} d\mathbf{x}^{(n)} d\boldsymbol{\theta}$$

  Cuando $n \to \infty$, se acerca a tener la información perfecta sobre $\boldsymbol{\theta}$ y luego $I(\mathcal{X}^n, \pi)$ se acerca a la información faltante sobre $\boldsymbol{\theta}$ que corresponde a la distribución a priori $\pi(\boldsymbol{\theta})$.

  La distribución a priori de referencia es la distribución que maximiza esta información faltante.

El enfoque de Bernardo es el enfoque más popular en problemas multivariantes.

- distribuciones de Haar. Ver Berger (1985).

  Basadas en consideraciones de simetría y invarianza en grupos.

- distribuciones jerárquicos

  Se supone que la distribución a priori tiene una estructura jerárquica.

Ejemplo 60 *Modelo de ADEVA.*

$$\begin{aligned} Y_{ij} &= \theta_i + \epsilon_{ij} \\ \theta_i &\sim \mathcal{N}(\mu, V) \\ \mu &\sim \textit{no informativa} \end{aligned}$$

La distribución a priori de θ_i está definida en dos etapas.

- Otros. Ver Yang y Berger (1997), Kass y Wassermann (1996).

Problemas con distribuciones no informativas

- posibilidades de distribuciones a posteriori impropias.

  Ejemplo 61 *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 62 *Suponiendo que vamos a generar datos de una binomial $X|\theta \sim \mathcal{BI}(n,\theta)$, se ha visto en el Ejemplo 50 que la distribución a priori de Jeffreys es $\theta \sim \mathcal{B}(1/2,1/2)$.*

Supongamos ahora que se generan datos de una distribución 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 4 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 automático de distribuciones de Jeffreys no cumple con el principio de verosimilitud.

- Paradojas de marginalización.

  Ver Dawid et al (1973).