CAPÍTULO 2. INTRODUCCIÓN

Para leer

Lee, Capítulo 2, Sección 2.1, Capítulo 3, Sección 3.1.

Gelman et al, Capítulo 1, Secciones 1.1 – 1.7, Capítulo 2, Secciones 2.1 y 2.2.

El problema de inferencia

$$X|\boldsymbol{\theta} \sim f(\cdot)$$

Dada una muestra de datos, $\mathbf{x} = (x_1, \ldots, x_n)^T$, queremos hacer inferencia (contrastes, ...) sobre $\boldsymbol{\theta}$.

Hay dos métodos tipo: inferencia clásica e inferencia Bayesiano (y también inferencia fiducial, inferencia no-paramétrico, ...).

Inferencia clásica

- El concepto de probabilidad está limitado a aquellos sucesos en los que se pueden definir frecuencias relativas,

- $\boldsymbol{\theta}$ es un valor fijo (pero desconocido),

- Estimación usando estimadores de máxima verosimilitud (justificado asintóticamente) o usando estimadores insesgados. Estimadores insesgados pueden ser muy malos.

  Ejemplo 10 *Supongamos que tenemos una observación x de una distribución Poisson: $\mathcal{P}(\lambda)$. Queremos estimar $\phi = e^{-2\lambda}$.*

  Sólo existe un estimador insesgado que es $(-1)^x = \pm 1$. Pero $0 < e^{-2\lambda} \leq 1$ por cualquier valor de λ. No existe ningún estimador insesgado para $\theta = 1/\lambda$.

- ¿Qué es un intervalo de confianza?

  Si $(1, 3)$ es un intervalo de confianza de 95 %, significa que si repetimos el procedimiento muchos veces y calculamos un intervalo de confianza cada vez, 95 % de los intervalos incluiran θ, el verdadero valor del parametro.

  No significa que la probabilidad de que θ sea en el intervalo es 95 %.

  La definición es una consecuencia del punto de vista de probabilidad como frecuencia.

- El método de muestreo es muy importante,

- Problemos con parámetros de molestia (*nuisance parameters*).

Inferencia Bayesiana

- Todos tenemos en nuestras propias probabilidades para cualquier suceso: P(cruz), P(llovera mañana), P(yo nací en 1962).

  Nuestras probabilidades pueden ser diferentes porque son nuestras propias medidas de similitud. La única restricción es que nuestras probabilidades sean coherentes (se cumplen con las reglas de probabilidad).

- $\boldsymbol{\theta}$ es un variable. Tiene una distribución $f(\boldsymbol{\theta})$. Modificamos nuestras creencias sobre $\boldsymbol{\theta}$ usando **el teorema de Bayes**:

$$\begin{aligned} f(\boldsymbol{\theta}|\mathbf{x}) &= \frac{f(\mathbf{x}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{f(\mathbf{x})} \\ &\propto f(\mathbf{x}|\boldsymbol{\theta})f(\boldsymbol{\theta}) = l(\boldsymbol{\theta}|\mathbf{x})f(\boldsymbol{\theta}) \end{aligned}$$

- siendo $l(\boldsymbol{\theta}|\mathbf{x})$ la función de verosimilitud,
- siendo $f(\boldsymbol{\theta})$ la **distribución a priori** (inicial) y
- $f(\boldsymbol{\theta}|\mathbf{x})$ la **distribución a posteriori** (final).

- La estimación es un problema de decisión. En situaciones distintas se eligen estimadores diferentes. Se usa **la teoría de utilidad** para elegir.

- Un intervalo de credibilidad de 95 % para θ es un intervalo en lo que, tenemos una probabilidad del 95 % de que contenga θ.

- El método de muestreo no importa. Sólo los datos son importantes.

- No hay problemas con parámetros de molestia.

  Si $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ donde $\boldsymbol{\theta}_2$ son parámetros perturbadores, se puede expresar

  $$f(\boldsymbol{\theta}) = f(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) f(\boldsymbol{\theta}_2)$$

  y luego

  $$f(\boldsymbol{\theta}_1|\mathbf{x}) = \int f(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \mathbf{x}) f(\boldsymbol{\theta}_2|\mathbf{x}) \, d\boldsymbol{\theta}_2.$$

  (Nota: si $f(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \mathbf{x})$ cambia mucha con valores diferentes de $\boldsymbol{\theta}_2$, se tiene que pensar. Ver Box y Tiao, Sección 1.6.)

Críticas a la Teoría Bayesiana

- θ no tiene que ser variable.

  θ puede ser fijo pero la distribución $f(\theta)$ muestra los conocimientos de θ. Los conocimientos y luego las creencias cambian con datos.

  En unas situaciones, los métodos Bayesianos son incontrovertidos. La distribución a priori es objetiva.

  Ejemplo 11 *Hay ratones de dos colores, negros y marrones. Los negros son de dos clases genéticas; homocigóticos (BB) y heterocigóticos (Bb), y los marrones son de una clase (bb). Se sabe por la teoría genética que las probabilidades de crías de las diferentes clases asociadas con padres diferentes son:*

Ratones	*BB (neg)*	*Bb (neg)*	*bb (marr)*
BB y bb	0	1	0
Bb y bb	0	$\frac{1}{2}$	$\frac{1}{2}$
Bb y Bb	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Tenemos un ratón negro cuyos padres son de clase Bb.

Define $\theta = 0$ *si el ratón es de clase BB y* $\theta = 1$ *si es de clase Bb. Entonces*

$$P(\theta = 0) = \frac{\frac{1}{4}}{\frac{1}{4} + \frac{1}{2}} = \frac{1}{3}$$

y $P(\theta = 1) = \frac{2}{3}$.

Ahora supongamos que se reciben los datos de que el ratón se procrea con otra ratón marrón y que crían siete ratoncitos negros. La verosimilitud es

$$
\begin{aligned}
P(7negros|\theta = 0) &= 1 \\
P(7negros|\theta = 1) &= \left(\frac{1}{2}\right)^7
\end{aligned}
$$

Usando el teorema de Bayes, podemos calcular las probabilidades a posteriori.

$$
\begin{aligned}
P(\theta = 0|\mathbf{x}) &= \frac{l(\theta = 0|\mathbf{x})P(\theta = 0)}{l(0|\mathbf{x})P(\theta = 0) + l(1|\mathbf{x})P(\theta = 1)} \\
&= \frac{1 \times \frac{1}{3}}{1 \times \frac{1}{3} + \left(\frac{1}{2}\right)^7 \times \frac{2}{3}} \\
&= \frac{64}{65}
\end{aligned}
$$

Igualmente, $P(\theta = 1|\mathbf{x}) = \frac{1}{65}$.

- Falta de objetividad. ¿Cómo se puede elegir la distribución a priori?

  - El modelo no es objetivo.

  - A menudo, un análisis clásico corresponde a un análisis Bayesiano con una distribución a priori no informativa.

  - Los aspectos subjetivos son explícitos en un análisis Bayesiano. Los bayesianos tienen que justificar sus elecciones.

  - Es imprescendible hacer **análisis de sensibilidad**. Si los resultados cambian cuando se cambia la distribución inicial, implica que la verosimilitud no da mucha información y la elección de la distribución inicial es fundamental.

El principio de verosimilitud

es otra justificación de los métodos bayesianos. Dice que:

para hacer inferencia sobre θ*, después de haber visto* $\mathbf{x}$*, toda la información pertinente está contenida en la función de verosimilitud* $l(\boldsymbol{\theta}|\mathbf{x})$*. Además, dos funciones de verosimilitud tienen la misma información sobre* $\boldsymbol{\theta}$ *si son proporcionales.*

Está claro que los métodos bayesianos cumplen con el principio de verosimilitud. Si $l_1(\boldsymbol{\theta}|\mathbf{x}) \propto l_2(\boldsymbol{\theta}|\mathbf{x})$ entonces, dada una distribución a priori $f(\boldsymbol{\theta})$,

$$f_1(\boldsymbol{\theta}|\mathbf{x}) \propto f(\boldsymbol{\theta}) l_1(\boldsymbol{\theta}|\mathbf{x}) \propto f(\boldsymbol{\theta}) l_2(\boldsymbol{\theta}|\mathbf{x}) \propto f_2(\boldsymbol{\theta}|\mathbf{x})$$

No obstante, se demuestra que contrastes clásicos de significación al nivel fijo (por ejemplo $\alpha =$,05) e intervalos de confianza no cumplen con este principio.

Ejemplo 12 $\theta = P(cruz)$ *por una moneda. Se quiere hacer el contraste* $H_0 : \theta = 1/2$ *contra la alternativa* $H_1 : \theta > 1/2$ *al nivel de significación de 5 %.*

Supongamos que se tira la moneda y que salen 9 cruces y 3 caras.

No hay datos suficientes para determinar la función de verosimilitud. Hay dos posibilidades:

1. *Se ha fijado el numero de tiradas de la moneda en doce. Luego* $X = \#$ *cruces*$|\theta \sim \mathcal{BI}(12, \theta)$ *y*

$$l(\theta|x = 9) = \left(\begin{array}{c} 12 \\ 9 \end{array} \right) \theta^9 (1-\theta)^3$$

Entonces, el p-valor es

$$p_1 = \frac{1}{2}^{12} \left[\left(\begin{array}{c} 12 \\ 9 \end{array} \right) + \cdots + \left(\begin{array}{c} 12 \\ 12 \end{array} \right) \right] \approx ,075$$

y no se rechaza la hipótesis nula.

2. *Se ha decidido recordar el numero de cruces* X *hasta que salga la tercera cara. Luego* $X|\theta \sim \mathcal{BN}(3, \theta)$ *y*

$$l(\theta|x) = \left(\begin{array}{c} 11 \\ 9 \end{array} \right) \theta^9 (1-\theta)^3$$

Entonces, el p valor es

$$\begin{aligned} p_2 &= \binom{11}{9}\theta^9(1-\theta)^3 + \\ &\quad \binom{12}{10}\theta^{10}(1-\theta)^3 + \cdots \\ &= \text{,0325} \end{aligned}$$

Entonces se rechaza la hipótesis nula.

(Nota: el p valor sería diferente todavía si se hubiera visto la secuencia completa de caras y cruces.)

La razón es que para hacer un contraste de significación, se tiene que especificar el espacio muestral. Es diferente en los casos diferentes que hemos visto:

1. $\Omega = \{(u,d) : u + d = 12\}$

2. $\Omega = \{(u,d) : d = 3\}$

La inferencia clásica depende del espacio muestral y luego de los datos que pudieramos haber visto pero no hemos visto de verdad.

(Nota: no se tiene que creer en el principio, pero hay otros principios equivalentes al principio de verosimilitud que también son creíbles. Ver Berger, Smith & Bernardo).

El principio de suficiencia

Definición 3 *Un estadístico* $\mathbf{t}$ *es suficiente para* $\boldsymbol{\theta}$ *si*

$$l(\boldsymbol{\theta}|\mathbf{x}) = h(\mathbf{t}, \boldsymbol{\theta})g(\mathbf{x}).$$

El **Principio de Suficiencia** dice que si existe un estadístico suficiente, $\mathbf{t}$, dadas dos muestras, $\mathbf{x}_1$, $\mathbf{x}_2$, que cumplen $\mathbf{t}(\mathbf{x}_1) = \mathbf{t}(\mathbf{x}_2)$, las conclusiones basadas en $\mathbf{x}_1$ y $\mathbf{x}_2$ deben ser iguales.

Observación 10 *Todos los métodos estandar de inferencia estadística cumplen el principio de suficiencia.*

El Principio de Condicionalidad

Suponiendo que se puede hacer uno de dos experimentos E_1 y E_2 sobre $\boldsymbol{\theta}$ y que se elige uno con probabilidad 0,5, entonces la inferencia sobre $\boldsymbol{\theta}$ debe depender sólo del resultado del experimento seleccionado.

Teorema 4 *El principio de verosimilitud = el principio de suficiencia + el principio de condicionalidad.*

Demostración Parcial

Demostramos que suficiencia más condicionalidad $\Rightarrow$ verosimilitud. Ver Lee (1997) para una demostración completa.

Sea $EV(E, \mathbf{x})$ la información sobre $\boldsymbol{\theta}$ proveniente del experimento E.

Definimos el experimento $E^\star = E_1$ con probabilidad 0,5 o E_2 con probabilidad 0,5. Los resultados del experimento $E^\star$ son el número del experimento y la observación $(i, \mathbf{x}_i)$.

Condicionalidad implica que

$$EV[E^\star, (i, \mathbf{x})] = EV[E_i, \mathbf{x}_i].$$

Elegimos dos valores $\mathbf{x}_1^0$, $\mathbf{x}_2^0$ donde

$$l(\theta|\mathbf{x}_1^0) = cl(\theta|\mathbf{x}_2^0) \; \forall \, \theta$$

Definimos T como

$$T(i, \mathbf{x}_i) = \begin{cases} (1, \mathbf{x}_1^0) & \text{si } i = 2, \mathbf{x}_2 = \mathbf{x}_2^0 \\ (i, \mathbf{x}_i) & \text{sino} \end{cases}$$

Se quiere demostrar que T es suficiente para $\boldsymbol{\theta}$.

Si $t \neq (1, \mathbf{x}_1^0)$ entonces

$$P(\mathbf{X}^\star = (i, \mathbf{x}_i) | T = t, \boldsymbol{\theta}) = I_{t=(i,\mathbf{x}_i)}.$$

Si $t = (1, \mathbf{x}_1^0)$ entonces

$$\begin{aligned} P(\mathbf{X}^\star = (1, \mathbf{x}_1^0) | T = (1, \mathbf{x}_1^0), \boldsymbol{\theta}) &= \frac{0{,}5 c l(\boldsymbol{\theta}|\mathbf{x}_2^0)}{0{,}5 l(\boldsymbol{\theta}|\mathbf{x}_1^0) + 0{,}5 c l(\boldsymbol{\theta}|\mathbf{x}_2^0)} \\ &= \frac{c}{1+c} \ \forall \boldsymbol{\theta}. \end{aligned}$$

En ambos casos, la probabilidad no depende de $\boldsymbol{\theta}$ y entonces T es suficiente para $\boldsymbol{\theta}$.

Ahora, suficiencia implica que

$$EV(E^\star, (1, \mathbf{x}_1)) = EV(E^\star, (2, \mathbf{x}_2))$$

es decir el principio de verosimilitud.

Observación 11 *Igualmente, se puede demostrar que verosimilitud + suficiencia $\Rightarrow$ condicionalidad o que verosimilitud + condicionalidad $\Rightarrow$ suficiencia.*

Análisis Bayesiana simple del Ejemplo 12.

Necesitamos recordar la definición de la distribución beta.

Definición 4 *θ tiene una distribución beta con parámetros α, β ($\theta \sim \mathcal{B}(\alpha, \beta)$) si*

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

por $0 < \theta < 1$ donde $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.

Propiedades

- Si α, β son números enteros, tenemos

$$B(\alpha, \beta) = \frac{(\alpha - 1)!(\beta - 1)!}{(\alpha + \beta - 1)!}$$

- $E[\theta] = \frac{\alpha}{\alpha+\beta}$

- $V[\theta] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

- $\phi = 1 - \theta \sim \mathcal{B}(\beta, \alpha)$.

Ejemplo 13 *Para ilustrarlo, se supone una distribución inicial uniforme* $\theta \sim \mathcal{U}(0,1)$*. (Es poco real, significa que no se sabe mucho de la moneda. Sería mejor una distribución Beta simétrica, por ejemplo* $\mathcal{B}(5,5)$*)*

Luego la distribución a posteriori es

$$\begin{aligned} f(\theta|x) &\propto \binom{12}{9} \theta^9 (1-\theta)^3 \\ &\propto \theta^9 (1-\theta)^3 \\ &\propto \theta^{10-1} (1-\theta)^{4-1} \end{aligned}$$

que significa que $\theta|\mathbf{x} \sim \mathcal{B}(10,4)$*.*

Se puede mostrar que $P(\theta \leq 1/2|\mathbf{x}) \approx$,046*.*

(Nota: no es un contraste formal. Estudiaremos métodos formales más adelante.)

La media final es entre la media inicial y la EMV

Ejemplo 14 *retomando el Ejemplo 13 también se puede calcular la media a posteriori de* θ.

$$E[\theta|x] = \frac{10}{10+4} = \frac{5}{7}$$

de las propiedades de la distribución beta.

Además

$$\frac{5}{7} = \frac{1}{7} \times \frac{1}{2} + \frac{6}{7} \times \frac{9}{12}$$

que implica que

$$E[\theta|x] = \frac{1}{7}E[\theta] + \frac{6}{7}\widehat{\theta}$$

donde $E[\theta] = 1/(1+1) = 1/2$ *es la media inicial y* $\widehat{\theta} = 9/12$ *es el estimador de maxima verosimilitud de* θ.

Predicción

Dados los datos $\mathbf{x} = (x_1, \ldots, x_n)^T$, supongamos que se quiere predecir el valor de X_{n+1}. Luego, se calcula **la distribución predictiva**

$$f(x_{n+1}|\mathbf{x}) = \int f(x_{n+1}|\boldsymbol{\theta}, \mathbf{x}) f(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta}.$$

En nuestra situación (los $X_i|\boldsymbol{\theta}$ son **intercambiables** $\approx$ independientes) ésta formula simplifica a

$$f(x_{n+1}|\mathbf{x}) = \int f(x_{n+1}|\boldsymbol{\theta}) f(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta}.$$

Ejemplo 15 *Volviendo al Ejemplo 13, se quiere predecir el numero de cruces en diez tiradas mas de la misma moneda.*

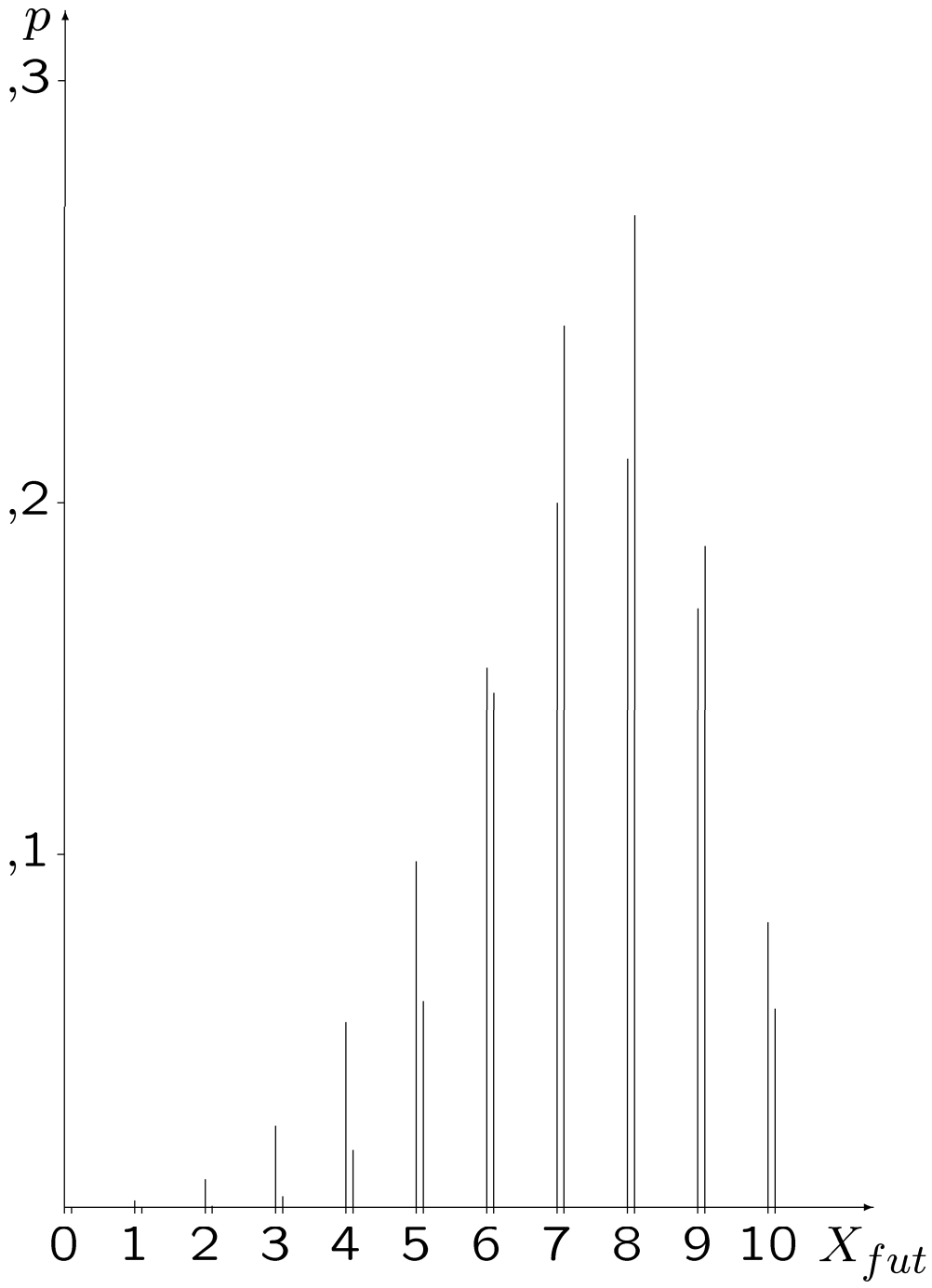
$$X_{fut}|\theta \sim \mathcal{BI}(10, \theta)$$

y entonces (usando $B(a, b)$ para representar la función beta)

$$
\begin{aligned}
f(x_{fut}|\mathbf{x}) &= \int_0^1 \binom{10}{x_{fut}} \theta^{x_{fut}}(1-\theta)^{10-x_{fut}} \times \\
&\quad \times \frac{1}{B(10,4)} \theta^{10-1}(1-\theta)^{4-1}\, d\theta \\
&= \binom{10}{x_{fut}} \frac{1}{B(10,4)} \times \\
&\quad \times \int_0^1 \theta^{10+x_{fut}-1}(1-\theta)^{14-x_{fut}-1}\, d\theta \\
&= \binom{10}{x_{fut}} \frac{B(10+x_{fut}, 14-x_{fut})}{B(10,4)}
\end{aligned}
$$

la llamada **distribución beta-binomial.**

El diagrama ilustra las probabilidades predictivas de X_{fut} y las probabilidades calculadas usando la EMV ($\mathcal{BI}(10,,75)$).


p
,3
,2
,1
0 1 2 3 4 5 6 7 8 9 10 X_{fut}

La Media y Varianza Predictiva

Se puede evaluar la media de $X_{fut}|\mathbf{x}$ sin tener que evaluar la distribución predictiva. Solo se tiene que recordar una regla de probabilidad.

$$E[Z] = E[E[Z|Y]]$$

por algunas variables Z y Y.

Ejemplo 16 *Retomando el Ejemplo 15, tenemos* $E[X_{fut}|\theta] = 10\theta$.

$E[\theta|\mathbf{x}] = \frac{5}{7}$, *luego*

$$E[X_{fut}|\mathbf{x}] = 10 \times \frac{5}{7} \approx 7{,}141$$

Para evaluar la varianza predictiva, se usa la fórmula

$$V[Z] = E[V[Z|Y]] + V[E[Z|Y]]$$

Evaluamos ambos partes.

$$
\begin{aligned}
V[X_{fut}|\theta] &= 10\theta(1-\theta) \\
E[V[X_{fut}|\theta]|\mathbf{x}] &= 10\left(E[\theta|\mathbf{x}] - E[\theta^2|\mathbf{x}]\right) \\
&= 10\left(E[\theta|\mathbf{x}] - E[\theta|\mathbf{x}]^2 - V[\theta|\mathbf{x}]\right) \\
&= 10\left(\frac{10}{10+4} - \left(\frac{10}{10+4}\right)^2 - \right. \\
&\quad \left. - \frac{10\times 4}{(10+4)^2(10+4+1)}\right) \\
&= \frac{40}{21} \\
E[X_{fut}|\theta] &= 10\theta \\
V[E[X_{fut}|\theta]|\mathbf{x}] &= 100V[\theta|\mathbf{x}] \\
&= 100\frac{10\times 4}{(10+4)^2(10+4+1)} \\
&= \frac{200}{147}
\end{aligned}
$$

Entonces tenemos $V[X_{fut}|\mathbf{x}] = \frac{40}{21} + \frac{200}{147} \approx 3{,}3$.

La verosimilitud escalada

De vez en cuando (si θ es unidimensional), se quiere ver la influencia de la distribución inicial y de la verosimilitud en la distribución final. Para hacer eso, es útil calcular la verosimilitud escalada

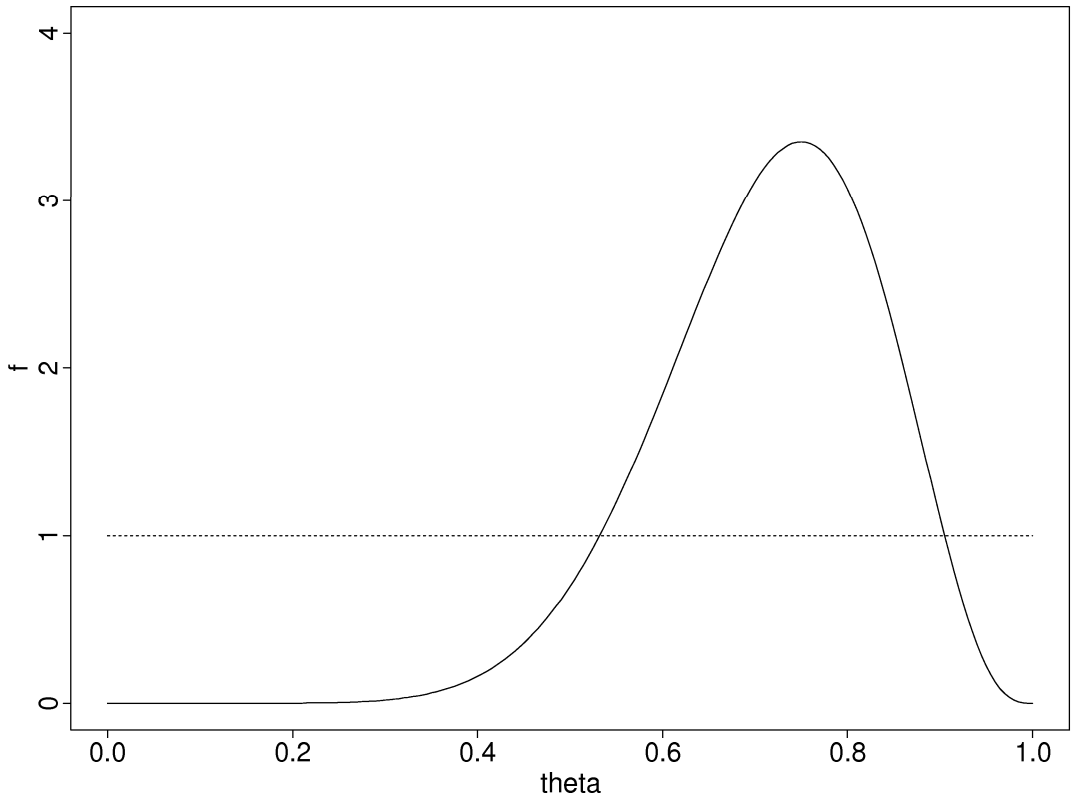
$$\frac{l(\theta|\mathbf{x})}{\int l(\theta|\mathbf{x})\ d\theta}$$

Nota: no es cierto que exista.

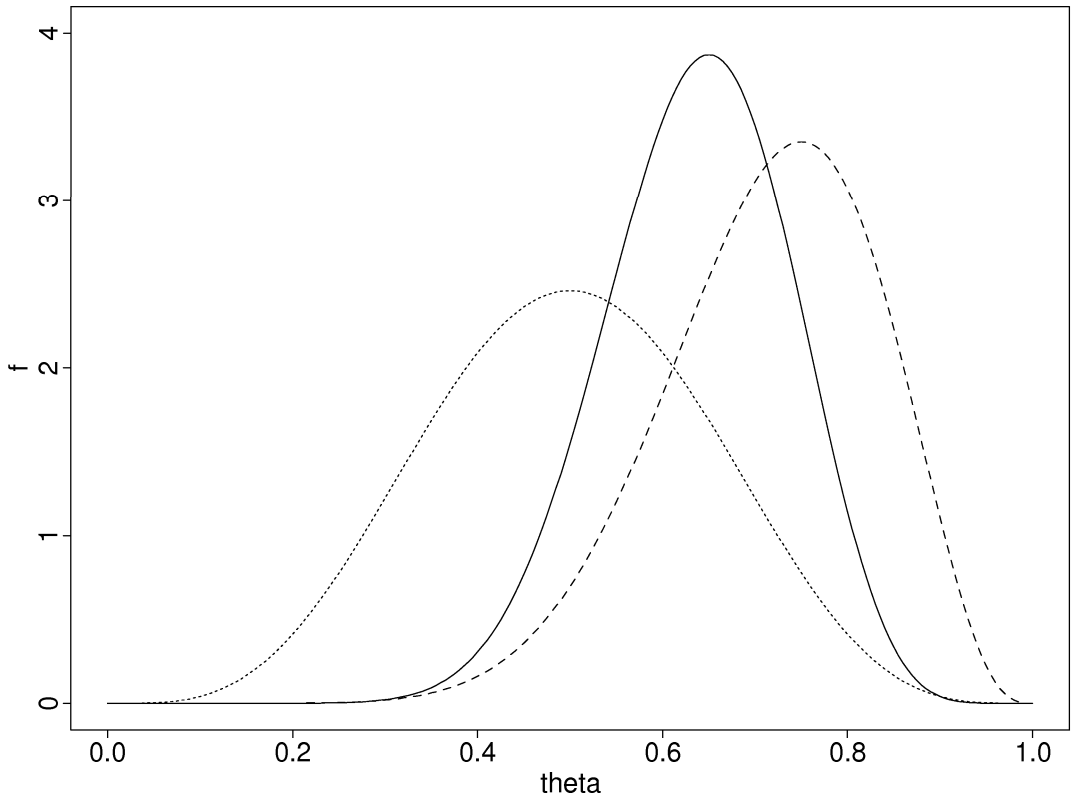
Dada la verosimilitud escalada, se puede hacer un diagrama mostrando la distribución a priori, la distribución a posteriori y la verosimilitud escalada para ver la relación entre las tres.

Ejemplo 17 *Veamos unos diagramas de las distribuciones iniciales y finales y la verosimilitud escalada usando distribuciones iniciales diferentes.*

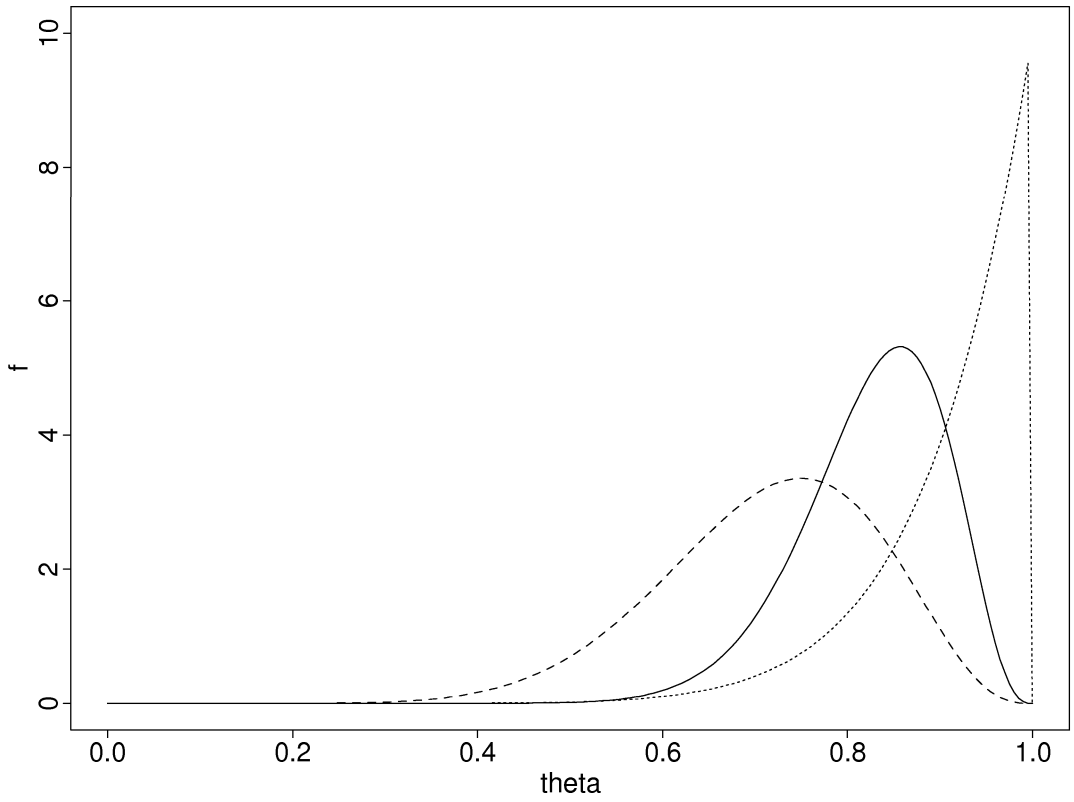
1) con distribución a priori uniforme.



2) con distribución a priori $\mathcal{B}(5,5)$.



3) con distribución a priori $\mathcal{B}(10, 1)$.



4) con distribución a priori $\mathcal{B}(1, 10)$.