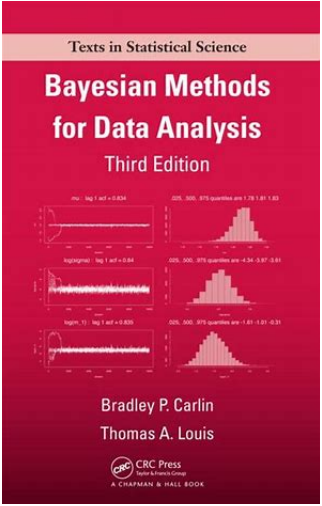
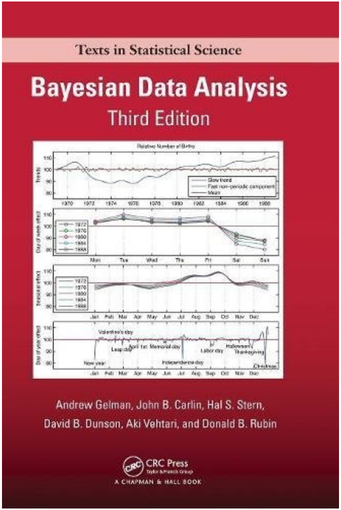
Inference Estadística
Métodos Bayesiano de Inferencia

Gabriel Martos Venturini
gmartos@utdt.edu

UTDT

UNIVERSIDAD TORCUATO DI TELLA


Texts in Statistical Science
Bayesian Methods for Data Analysis
Third Edition
Bradley P. Carlin
Thomas A. Louis
CRC Press
A CHAPMAN & HALL BOOK


Texts in Statistical Science
Bayesian Data Analysis
Third Edition
Andrew Gelman, John B. Carlin, Hal S. Stern,
David B. Dunson, Aki Vehtari, and Donald B. Rubin
CRC Press
A CHAPMAN & HALL BOOK

Agenda

Dos escuelas de inferencia

- Clásica: El parámetro de interés θ es una cantidad fija y desconocida.
  - Los métodos de inferencia están pensados para tener garantías *a largo plazo* (i.e. intervalos de confianza y test de hipótesis).
- Bayesiana:
  - Las "creencias o juicios subjetivos" sobre θ se codifican en términos probabilísticos mediante una distribución (subjetiva) $P(\theta)$ para θ.
  - Con los datos y la verosimilitud corregimos $P(\theta)$ de forma coherente.
  - Con la distribución aposteriori $P(\theta \mid \text{Datos})$ hacemos inferencia para θ.
  - Output: A diferencia del enfoque frecuentista, los métodos Bayesianos permiten hacer interpretaciones probabilísticas de los resultados.

... un poco más de contexto

A. Gelman, C. Rohilla Shalizi (2012)
Philosophy and the practice of Bayesian statistics
British Journal of Mathematical and Statistical Psychology. Vol 66(1), pp. 8–38.

M. J. Bayarri, J. O. Berger (2004)
The Interplay of Bayesian and Frequentist Analysis
Statistical Science. 19(1), pp. 58–80.

B. Efron (1986)
Why Isn't Everyone a Bayesian?
The American Statistician. Vol 40(1), pp. 1–5.

A. Gelman (2008)
Objections to Bayesian statistics.
Bayesian Analysis. Vol 3(3), pp. 445–449.

Elementos de inferencia Bayesiana

- Fórmula de Bayes:

$$P(\theta \mid \texttt{datos}) = \frac{P(\texttt{datos} \mid \theta) P(\theta)}{P(\texttt{datos})}.$$

  - $P(\theta) = \pi(\theta)$ (en gral. $\pi(\theta; \eta)$), donde η es un hyperparámetro).
  - Por $\texttt{datos} = \underline{x}$ entendemos la realización de una muestra aleatoria.
  - $P(\texttt{datos} \mid \theta) = \Pi_{i=1}^{n} f(x_i; \theta) = L(\theta | \underline{x})$ (verosimilitud).
  - $P(\texttt{datos}) = \int L(\theta|\underline{x})\pi(\theta)\mathrm{d}\theta = c(\underline{x})$ (no depende de θ).
  - Por tanto $P(\theta \mid \texttt{datos}) \equiv \underbrace{\pi(\theta \mid \texttt{datos})}_{\text{posterior}} \propto L(\theta|\underline{x})\pi(\theta)$.
    - ★ Con la distribución a posteriori hacemos inferencia para θ.

- En general aproximamos (algún funcional de) $\pi(\theta \mid \texttt{datos})$ utilizando técnicas de integración numérica (Gibbs–sampling, MCMC, etc).

Inferencia con $\pi(\theta \mid \texttt{datos})$

- Idealmente reportaríamos toda la distribución: $\pi(\theta \mid \texttt{datos})$.
- Sin embargo, suele ser más adecuado reportar:
  - Alguna medida de centralidad:
    - ★ $E(\theta \mid \texttt{datos}) = \int_\Theta \theta \pi(\theta \mid \texttt{datos}) \mathrm{d}\theta$.
    - ★ *Valor más probable* de $\theta \mid \texttt{datos}$ (MAP): $\arg\max_{\theta \in \Theta} \pi(\theta \mid \texttt{datos})$.
  - Medidas de variabilidad (cuantiles o sd) relativos a la distribución de la variable aleatoria $\theta \mid \texttt{datos}$ (cuantifican la incertidumbre), etc.
- Veamos algunos ejemplos de cómo construir $\pi(\theta \mid \texttt{datos})$.

Ejemplo I: $X|\theta \sim \text{Bern}(\theta)$ y $\theta \sim \text{Beta}(\alpha, \beta)$

$$\theta \sim \text{Beta}(\underbrace{\alpha, \beta}_{\boldsymbol{\eta}}) \Rightarrow f(\theta; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}, \text{ para } 0 \leq \theta \leq 1.$$

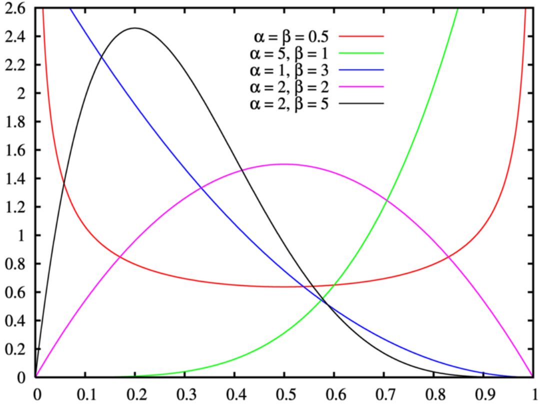


Figure: $E_\eta(\theta) = \alpha/(\alpha+\beta)$ y $\text{Mo}_\eta(\theta) = (\alpha-1)/(\alpha+\beta-2)$ si $\beta, \alpha > 1$.

Ejemplo I: $\theta \sim \text{Beta}(\alpha, \beta)$ y $X|\theta \sim \text{Bern}(\theta)$

- Prior $\theta \sim \pi(\theta; \boldsymbol{\eta} = (\alpha, \beta)) = \text{Beta}(\alpha, \beta)$.
- $\{X_1, \ldots, X_n\} \overset{iid}{\sim} \text{Bern}(\theta)$ y llamemos $t = \sum_{i=1}^{n} x_i$ (datos):

$$L(\theta \mid t) = \theta^t (1 - \theta)^{n-t}.$$

- Distribución a–posteriori: $\pi(\theta \mid t) \propto \theta^{\alpha+t-1}(1-\theta)^{n+\beta-t-1}$.

$$\theta \mid t \sim \text{Beta}(\alpha', \beta'),$$

  con $\alpha' = \alpha + t$ y $\beta' = \beta + n - t$.

- Veamos como luce $\pi(\theta \mid t)$ cuando $n = 10$, $t = 8$–sampling de Bern($\theta = 0.7$)– y elegimos los hiperparámetros $\alpha = 2$ y $\beta = 3$.
- Puedes replicar la gráfica con las líneas de código en R.

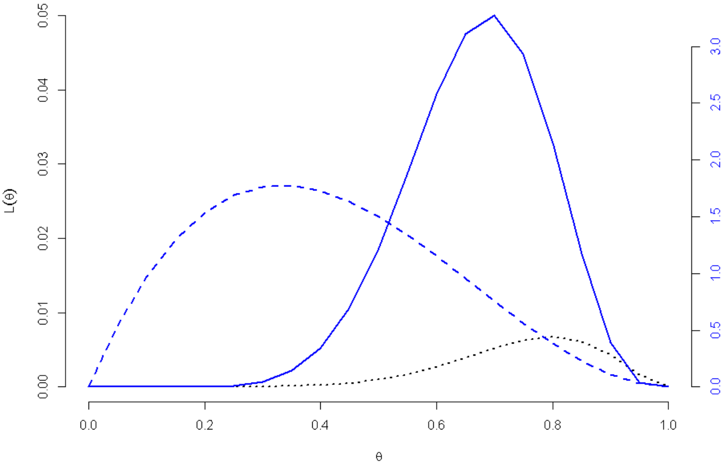


Figure: Verosimilitud $(\cdots)$, Prior (- - -) y Posterior (—).

- La verosimilitud "acomoda" $\pi(\theta)$ a la evidencia.
- ¿Media y varianza a posteriori?
- ¿Qué ocurre con $\pi(\theta \mid t)$ a medida que n crece?

Ejemplo II: $\theta \sim \text{Gamma}(\alpha, \beta)$ y $X|\theta \sim \text{Exp}(\theta)$

Agenda

Modelos conjugados

Definition

Si $X|\theta \sim f(x;\theta) \in \mathcal{F}_\theta$, decimos que el conjunto de distribuciones a priori $\mathcal{P}_\theta$ es conjugada para los modelos en $\mathcal{F}_\theta$ cuando ocurre que:

$$\text{Para toda } f(x;\theta) \in \mathcal{F}_\theta \text{ y } \pi(\theta) \in \mathcal{P}_\theta \text{ luego } \pi(\theta \,|\, \underline{x}) \in \mathcal{P}_\theta.$$

- Informalmente: $\pi(\theta)$ y $\pi(\theta \,|\, \underline{x})$ siguen el mismo modelo estadístico.
- Ejemplos:
  - Modelo Beta–Bernoulli y Gamma–Exponencial (slides anteriores).
  - Modelo Normal–Normal (próxima slide).
- Remark: Como $\pi(\theta|\underline{x}) \propto L(\theta|\underline{x})\pi(\theta) \in \mathcal{P}_\theta$, podemos identificar de manera rápida la constante de proporcionalidad $c(\mathbf{x}) = P(\text{datos})$.

Modelos conjugados en familias exponenciales

- Si $X|\theta \sim f(x;\theta) \in \mathcal{F}_\theta = \{f(x) \equiv h(x)c(\theta)\exp\big(w(\theta)t(x)\big)\}$ luego:

$$L(\theta \,|\, \underline{x}) \propto c(\theta)^n \exp\Big(w(\theta)\sum_{i=1}^{n} t(x_i)\Big).$$

- Por tanto si especificamos $\pi(\theta;\eta,\nu) \propto c(\theta)^\eta \exp\big(w(\theta)\nu\big)$, tendremos:

$$\pi(\theta \,|\, \underline{x}) \propto c(\theta)^{\overbrace{(n+\eta)}^{\eta'}} \exp\Big(w(\theta)\underbrace{\big(\sum_{i=1}^{n} t(x_i) + \nu\big)}_{\nu'}\Big).$$

  - $\mathcal{P}_\theta = \{\pi(\theta;\eta,\nu) \propto c(\theta)^\eta \exp\big(w(\theta)\nu\big)\}$ es conjugado para $\mathcal{F}_\theta$.

- Esto facilita bastante las cuentas, porque si conoces la estructura general de $\pi(\theta \,|\, \underline{x})$ puedes en muchos casos determinar el valor exacto de la constante de proporcionalidad sin tener que calcular una integral.

Ejemplo III: Modelo Normal–Normal (σ_0^2 conocida)

- El modelo para los datos es $X|\theta \sim N(\theta, \sigma_0^2)$.
- $\pi(\theta; \boldsymbol{\eta}) = N(\mu_0, \tau_0^2)$, donde $\boldsymbol{\eta} = (\mu_0, \tau_0^2)$ son hiperparámetros.
- Se puede demostrar (tarea G5) que:

$$\pi(\theta \,|\, \underline{x}) \propto \underbrace{\exp\{-\frac{1}{2\sigma_n^2}(\theta - \mu_n)^2\}}_{L(\theta \,|\, \underline{x})\pi(\theta;\boldsymbol{\eta})},$$

donde:

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma_0^2}\bar{x}_n}{\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2}} \text{ y } \sigma_n^2 = \left(\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2}\right)^{-1}.$$

- $\pi(\theta \,|\, \underline{x}) = N(\mu_n, \sigma_n^2)$ (¿cuánto vale $P(\text{datos})$?).
- ¿A donde convergen la media y varianza a posteriori?

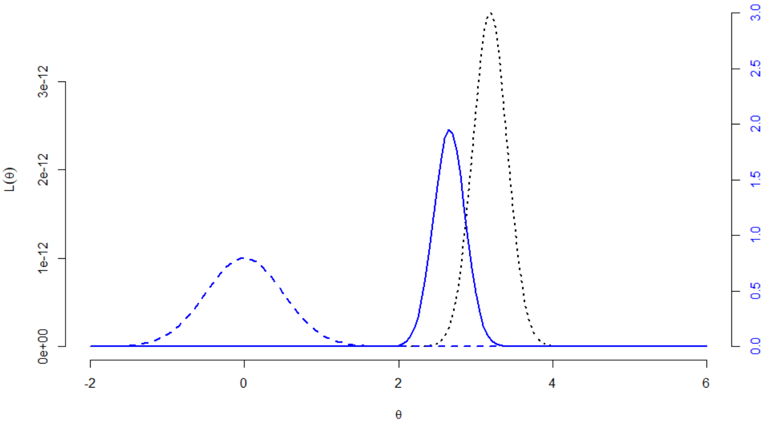


Figure: Verosimilitud $(\cdots)$, Prior (- - -) y Posterior (—).

- $\tau_0 = 1/2$, $\sigma_0 = 1$, $\mu_0 = 0$, $n = 20$, $\bar{x}_n = 3.2$ (samp. de $X \sim N(3, 1)$).

$$E(\theta|\underline{x}) = 2.65 \text{ y } V(\theta|\underline{x}) = 0.204$$

- ¿Cómo elegimos (los valores de los parámetros de) la prior?

Agenda

No informativas

- Siendo $\Theta = \{\theta_1, \ldots, \theta_k\}$, en un contexto no informativo:

$$\pi(\theta_i) = \frac{1}{k}, \text{ para } i = 1, \ldots, k.$$

- Ya que no *favorece* ningún valor de $\theta \in \Theta$ en particular.

- En este caso se tiene que:

$$P(\text{datos}) = c(\underline{x}) = \sum_{i=1}^{k} L(\theta_i|\underline{x})\frac{1}{k} = \overline{L}(\underline{x}).$$

- Luego $\pi(\theta \mid \underline{x}) = L(\theta|\underline{x})/k\overline{L}(\underline{x}) \propto L(\theta|\underline{x})$ para $\theta \in \Theta$.
  - Equivalencias entre inferencia clásica y Bayesiana:

$$\theta_{\text{MAP}} \equiv \operatorname{argmax} \pi(\theta \mid \underline{x}) = \operatorname{argmax} L(\theta|\underline{x}) = \widehat{\theta}_n.$$

No informativas (espacio de parámetros continuo)

- Si $\Theta = [a, b]$, podemos hacer:

$$\pi(\theta) = \frac{1}{b-a} \text{ y } c(\underline{x}) = \int_a^b L(\theta|\underline{x})/(b-a)\mathrm{d}\theta = \overline{L}(\underline{x}).$$

  - (idem al caso anterior).

- Cuando $\Theta = (-\infty, \infty)$ y elegimos $\pi(\theta) = \varepsilon > 0$, resulta que:

$$\int_\Theta \pi(\theta)\mathrm{d}\theta = \infty \quad \text{(prior impropia)}.$$

- La inferencia (que surge a partir de construir $\pi(\theta \,|\, \underline{x})$) será posible en este contexto siempre que $\int_\Theta L(\theta|\underline{x})\,\mathrm{d}\theta = C_L < \infty$:

$$\pi(\theta \,|\, \underline{x}) = \frac{L(\theta|\underline{x}) \times \cancel{\varepsilon}}{\int_\Theta L(\theta|\underline{x}) \times \cancel{\varepsilon}\,\mathrm{d}\theta} = \frac{L(\theta|\underline{x})}{C_L} \propto L(\theta|\underline{x}).$$

  - Notar que $\pi(\theta \,|\, \underline{x})$ es una densidad respecto de θ.

Invariantes a reparametrizaciones

- Si $X|\theta \sim f(x;\theta)$, la Jeffreys Prior (JP) se define como:

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

- Si $\gamma = g(\theta)$ (siendo g una transformación monótona), luego:

$$\pi(\gamma) \propto \underbrace{i(\theta)^{1/2}}_{\pi(\theta)} / |g'(\theta)| \quad \text{(cambio de variable).}$$

- Esto tiene como consecuencia que:

$$P(a \leq \theta \leq b \,|\, \underline{x}) = \int_a^b \pi(\theta \,|\, \underline{x}) \mathrm{d}\theta = \int_{g(a)}^{g(b)} \pi(\gamma \,|\, \underline{x}) \mathrm{d}\gamma = P(g(a) \leq \gamma \leq g(b) \,|\, \underline{x}).$$

- Si dos investigadores utilizan los mismos datos y modelo, aún parametrizando la verosimilitud de formas diferentes, con la JP la inferencia que hagan con los modelos es equivalente.

Agenda

Estimación puntual

- Reportamos un valor puntual del parámetro de acuerdo a $\pi(\theta \mid \underline{x})$.
- Habitualmente se reporta $\theta_{\text{MAP}} = \text{argmax}\, \pi(\theta \mid \underline{x})$.
  - $\pi(\theta \mid \underline{x})$ simétrica: Media, mediana y moda aposteriori coinciden.
  - También reportamos varianza/sd a posteriori (incertidumbre).
  - Si $\pi(\theta)$ es no informativa $\Rightarrow \theta_{\text{MAP}} = \widehat{\theta}_n$ (el EMV).
- Si consideramos: $\min_{c \in \Theta} E(L(\theta, c)|\underline{x})$.
  - Al riesgo cuadrático medio le corresponde $c^* = E(\theta|\underline{x})$.
  - Al riesgo absoluto le corresponde $c^* =$ mediana a posteriri.
- Ejemplos en código `R`.
  - Modelos Bernulli–Beta y Normal–Normal.

Intervalos de confianza (credible sets)

- Un intervalo de confianza de nivel $(1-\alpha)$ para θ, es un subconjunto $\mathsf{IC} \subset \Theta$ que verifica (utiliza un sumatorio en el caso discreto):

$$P(\theta \in \mathsf{IC} \,|\, \underline{x}) = \int_{\mathsf{IC}} \pi(\theta \,|\, \underline{x}) \, \mathrm{d}\theta \geq 1 - \alpha.$$

  - Interpretación probabilística: Condicional en la evidencia (dado los datos), la probabilidad de que $\theta \in \mathsf{IC}$ es como mínimo $1-\alpha$.
  - En general existen infinitos Credible Sets para θ.
  - Ejemplos: Beta–Bernoulli (código en `R`).

- Highest posterior density (HPD): Es el intervalo de confianza $1-\alpha$ de menor longitud entre todos los intervalos de confianza de dicho nivel.

$$\max \nu_\alpha : \mathsf{HPD} = \{\theta \in \Theta : \pi(\theta \,|\, \underline{x}) \geq \nu_\alpha\} \text{ y } P(\theta \in \mathsf{HPD} \,|\, \underline{x}) \geq 1 - \alpha.$$

- Ver ejemplo en la siguiente slide.

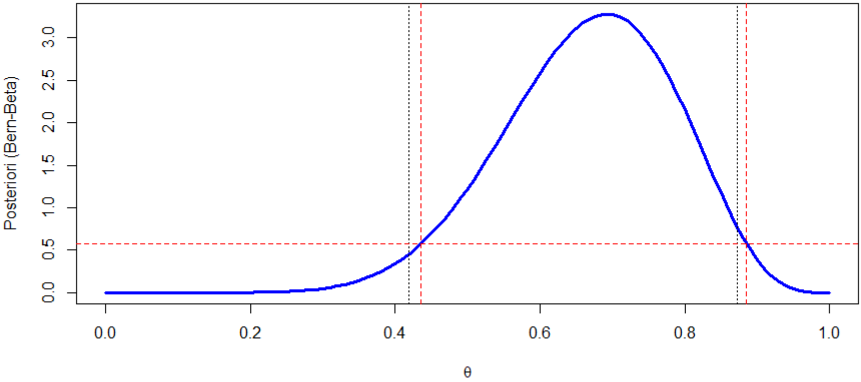


Figure: IC (- - -) vs HPD (—): Modelo Beta–Bernouilli ($\alpha = 0.05$).

- IC $= [0.418, 0.872]$ (longitud: 0.453).
- HPD $= [0.436, 0.885]$ (longitud: 0.449).
- (similar para el modelo normal–normal).

Credible sets vs HPD

- Las regiones HPD tienen una longitud menor (mayor precisión).
- El conjunto HPD no es necesariamente compacto.

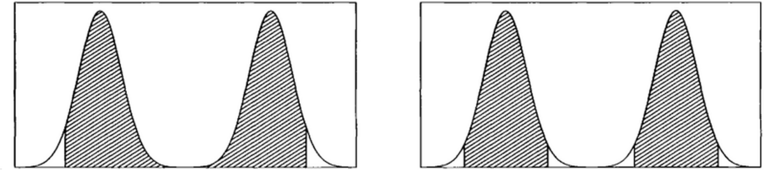


Figure 2.2 *Hypothetical posterior density for which the 95% central interval and 95% highest posterior density region dramatically differ: (a) central posterior interval, (b) highest posterior density region.*

- Si la posterior es continua y unimodal, HPD es único y compacto.

Testeando hipótesis

- $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$. Calculamos probabilidades a posteriori:

$$P(\underbrace{\theta \in \Theta_i}_{H_i \text{ sea cierta}} |\, \textbf{datos}) \equiv \int_{\Theta_i} \pi(\theta|\textbf{datos})\mathrm{d}\theta, \quad i = 0, 1.$$

- Y también probabilidades a priori:

$$P(\theta \in \Theta_i) = \int_{\Theta_i} \pi(\theta)\mathrm{d}\theta, \quad i = 0, 1.$$

- Juzgamos utilizando el Factor Bayesiano (FB):

$$\frac{P(\theta \in \Theta_1|\textbf{datos})}{P(\theta \in \Theta_0|\textbf{datos})} = \underbrace{\frac{P(\textbf{datos}|\theta \in \Theta_1)}{P(\textbf{datos}|\theta \in \Theta_0)}}_{\text{FB}(H_1;H_0)} \frac{P(\theta \in \Theta_1)}{P(\theta \in \Theta_0)}.$$

- FB cuantifica como ajustar el ratio de probabilidades a priori entre H_0 y H_1 al tomar en cuenta la evidencia empírica (los datos).

Sobre el método Bayesiano para los test

- La decisión sobre H_0 o H_1 se toma evaluando:

$$\text{FB}(H_1; H_0) = \frac{P(\theta \in \Theta_1|\textbf{datos})}{P(\theta \in \Theta_0|\textbf{datos})} \times \frac{P(\theta \in \Theta_0)}{P(\theta \in \Theta_1)}.$$

- Kass and Raftery (1995) identifican 3 situaciones para juzgar la evidencia empírica (los datos) en favor de H_1:
  1. $\text{FB}(H_1; H_0) < 3$ evidencia débil,
  2. $\text{FB}(H_1; H_0) \in [3, 20]$ evidencia positiva y,
  3. $\text{FB}(H_1; H_0) > 20$ evidencia fuerte.

- Es un procedimiento simétrico: $\text{FB}(H_1; H_0) = 1/\text{FB}(H_0; H_1)$.
  - Podemos concluir que 'aceptamos' H_0 sin mayores inconvenientes.

Preferencias de consumo (BMDA–§ 2.11).

- Estudio de mercado en donde se valora el grado de preferencia respecto de una versión premium de una marca de hamburguesas.
- Antes de lanzar el producto, se selecciona al azar a 16 consumidores habituales de la marca. El mismo chef prepara dos tipos de hamburguesas para cada individuo: Una con el producto standard y la otra con la versión premium.
- El chef elige con una moneda justa que hamburguesa entregar primero a cada consumidor (el experimento es doblemente ciego, el chef tampoco sabe a que grupo pertenece cada hamburguesa).
- Después de llevar a cabo el experimento, 13 de los 16 consumidores habituales revelan preferir la hamburguesa premium a la estandard.
- Analicemos estos datos en el contexto de inferencia Bayesiano.

- $X_i = 1$ preferencia del cliente i en favor de la hamburguesa premium.
  - $X|\theta \sim \text{Bern}(\theta)$, donde $\theta = P(X = 1)$ (preferencia premium en la pob).
- Luego $P(\text{Datos}|\theta) = L(\theta) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}$.
- Como prior elegimos una Beta (el modelo Bern–Beta es conjugado):

$$\pi(\theta; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

  - ¿Cómo elegimos los hiperparámetros α, β?
  - Tres estrategias diferentes y análisis de sensibilidad posterior.
- Posterior (datos: $t = \sum_{i=1}^n x_i$):

$$\pi(\theta|t) = \frac{L(\theta)\pi(\theta; \alpha, \beta)}{c(t, \alpha, \beta)} = \text{Beta}(\alpha + t, \beta + n - t).$$

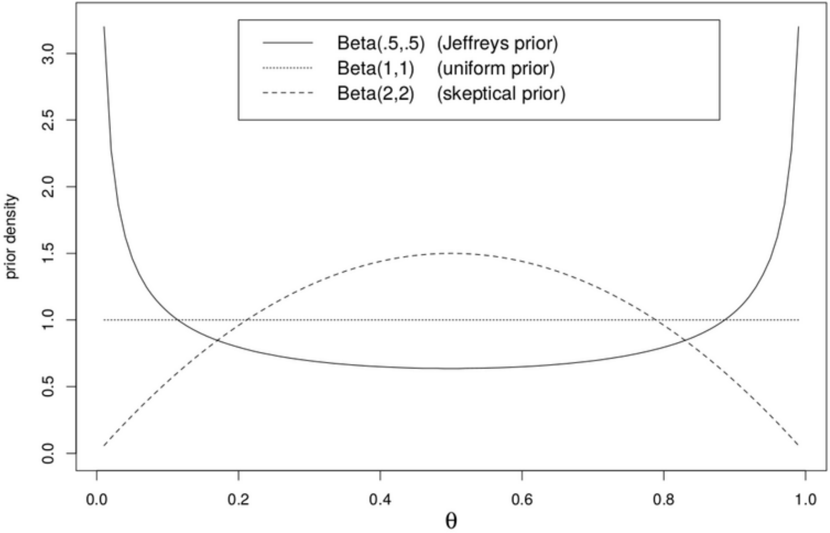


Figure 2.13 *Prior distributions used in analyzing the consumer preference data.*

Posterior

$$\pi(\theta|t = 13) = \text{Beta}(\alpha + 13, \beta + 3).$$

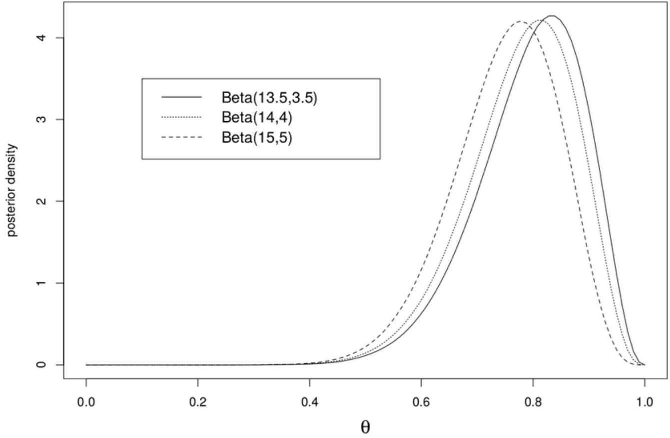


Figure 2.14 *Posterior distributions arising from the three prior distributions chosen for the consumer preference data.*

- Comparativa inferencia en los 3 escenarios:

Prior	$E(\theta\|t)$	MAP	$V(\theta\|t)$	$q_{0.025}$	$q_{0.5}$	$q_{0.975}$	$P(\theta > 0.6\|t)$
Jeffrey	0.79	0.83	0.03	0.58	0.81	0.94	0.964
No informativa	0.77	0.81	0.03	0.56	0.79	0.93	0.954
Escéptica	0.75	0.77	0.02	0.54	0.76	0.91	0.930

- Aún bajo escenarios muy diferentes para las priors, la inferencia a posteriori resulta bastante coherente con la hipótesis de preferencia respecto del nuevo producto.

- Los 3 intervalos a posteriori al 95% no incluyen a $\theta = 0.5$ (indiferencia de los consumidores habituales respecto del nuevo producto).

- Testeamos hipótesis.

- Supongamos que los dueños de la cadena se preguntan si los datos evidencian un incremento sustancial de la preferencia. Traducido a lenguaje Bayesiano la pregunta podría ser: Es ¿$\theta \geq 0.6$? (arbitrario).
  - Planteemos $H_0 : \theta \leq 0.6$ y $H_1 : \theta > 0.6$.
  - Asumiendo una prior uniforme tenemos que $P(H_0) = 0.6$ y $P(H_1) = 0.4$
  - $P(\underbrace{\theta \leq 0.6|t = 13}_{H_0|t}) \approx 0.046$ ($\pi(\theta|t = 13) \sim \text{Beta}(14, 14)$) y,
  - $P(\underbrace{\theta > 0.6|t = 13}_{H_1|t}) \approx 0.954$ (idem).

$$\text{FB}(H_1, H_0) = (0.954/0.046) \times (0.6/0.4) = 31.11$$

  - El ratio de probabilidades a posteriori es 31 veces más grande que el ratio de probabilidades a priori: Hay evidencia fuerte ($t = 13$) en favor de la preferencia por las hamburguesas premium ($H_1 : \theta > 0.6$).

Extensiones

- En la práctica los modelos Bayesianos incluyen varios niveles de jerarquías en su especificación, haciendo que muchas veces resulte imposible utilizar modelos conjugados.
  - Aproximaciones asintóticas de la posteriori.
  - Métodos numéricos de integración (MCMC) para construir los cuantiles aproximados de $\pi(\theta|\text{datos})$.
  - Otros métodos rápidos de integración (aproximaciones de Laplace).
  - Estas extensiones se discuten de forma introductoria en la bibliografía propuesta al comienzo de las diapositivas.

Agenda

- El modelo de regresión lineal es una de las herramientas más utilizadas en Estadística y Econometría.
- Los objetivos de este último apartado son:
  - Contrastar los procedimientos de inferencia clásicos contra los Bayesianos en el contexto del modelo lineal.
  - Implementar los modelos y hacer inferencia en el entorno R.
  - Caso de estudio: "Hedonic prices and the demand for clean air" J. Economics and Management 5, 81–102.

El modelo lineal (normal)

- El modelo presupone que
  - La esperanza de Y (endógena) es una función lineal respecto de un conjunto de variables exógenas (tratamientos y controles):

$$E(Y|X_1 = x_1, \ldots, X_p = x_p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

  - El modelo normal estandard asume: $\mathbf{y}|\mathbf{X} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$, esto es: Independencia + homoscedasticidad. Dada una muestra de tamaño n:

$$\ell(\boldsymbol{\beta}, \sigma^2|\mathbf{X}) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)}{2\sigma^2}$$

- EMV: $\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$ y $\widehat{\sigma}^2 = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)/n$.

- Inferencia en el contexto frecuentista:

$$\widehat{\boldsymbol{\beta}} \sim N_{p+1}\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\right)$$

Inferencia con el modelo

- Intervalos para los parámetros:

$$\text{IC}_{1-\alpha}(\beta_i) = \{\widehat{\beta}_i \pm t_{n-p-1,\alpha/2}\text{se}(\widehat{\beta}_i)\}, \quad i = 0, 1, \ldots, p.$$

donde $\text{se}(\widehat{\beta}_i) = \widehat{\sigma}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}}$ y $\widehat{\sigma} = \text{RSS}/(n-p-1)$.

- Elipses de confianza:

$$\text{RC}(\boldsymbol{\beta}) = \{\boldsymbol{\beta} : (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\mathbf{X}^T\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq (p+1)S^2 f_\alpha\}$$

donde $S^2 = n\widehat{\sigma}^2/(n-p-1)$ y f_α verifica que $P(F_{p,n-p-1} > f_\alpha) = \alpha$.

- Test individuales $H_0 : \beta_i = 0$ vs $H_1 : \beta_i \neq 0$ para $i = 1, \ldots, p$. Resolvemos con $t = \widehat{\beta}_i/\text{se}(\widehat{\beta}_i) \sim t_{n-p-1}$. Hipótesis global

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \text{ vs } H_1 : \text{al menos una pendiente } \neq 0,$$

resolvemos con $F = (n-p-1)(\text{SCT - SCR})/(p\text{SCR}) \sim F_{p,n-p-1}$.

El modelo lineal normal Bayesiano

- El modelo se constituye asumiendo que:

$$\begin{aligned} \mathbf{y}|\beta, \mathbf{X} &\sim N(\mathbf{X}\beta, \mathbf{C}_1), \\ \beta|\boldsymbol{\eta}, \mathbf{C}_2 &\sim N(\boldsymbol{\eta}, \mathbf{C}_2). \end{aligned}$$

- $\mathbf{y}$ es un vector aleatorio de dimensión n.
- $\mathbf{X}$ es una matriz de diseño de dimensión $n \times (p+1)$.
- $\boldsymbol{\eta}$ es un vector de hiperparámetros de dimensión $p+1$.
  - En general, en la práctica elegimos $\boldsymbol{\eta} = \mathbf{0}$.
- $\mathbf{C}_1$ y $\mathbf{C}_2$ son matrices de covarianza conocidas de dimensiones $n \times n$ y $(p+1) \times (p+1)$ respectivamente.
- Asumimos los hiperparámetros conocidos (Carlin y Louis § 4.1).

El modelo lineal normal Bayesiano

- Bajo los supuestos anteriores se tiene que

$$\beta \mid \underbrace{\mathbf{y}, \mathbf{X}}_{\text{datos}} \sim N(\mathbf{D}\gamma, \mathbf{D}),$$

- $\mathbf{D}^{-1} = \mathbf{X}^T \mathbf{C}_1^{-1} \mathbf{X} + \mathbf{C}_2^{-1}$.
- $\gamma = \mathbf{X}^T \mathbf{C}_1^{-1} \mathbf{y} + \mathbf{C}_2^{-1} \boldsymbol{\eta}$.
- $E(\beta|\text{datos}) = \mathbf{D}\gamma$ (estimación puntual).
- $V(\beta|\text{datos}) = \mathbf{D}$ (varianza a posteriori).
- ¿Cuándo recuperamos el modelo clásico de regresión?

$$\underbrace{\mathbf{C}_2^{-1} \to \mathbf{0}}_{\text{Prior no inf.}} \text{ y } \mathbf{C}_1 = \sigma^2 \mathbf{I} \Rightarrow \beta|\mathbf{y}, \mathbf{X} \sim N(\underbrace{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}}_{\widehat{\beta}}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

- Para σ^2 se suele asumir una prior Exponencial o Gamma–Inversa.

Stan en R

- Stan: Librería en $C+$ para fitear modelos Bayesianos.
- rstan y rstanarm: Librería en R para llamar a las funciones de Stan.

```
library(rstan)
library(rstanarm)

stan_glm(formula,data, family=gaussian)
```

- Priors por defecto:
  - $\beta_0 \sim N(\bar{y}, 2.5 s_y)$.
  - $\beta_i \sim N(0, 2.5 s_y / s_{x_i})$ para $i = 1, \ldots, p$.
  - $\sigma^2 \sim \text{Exp}(1/s_y)$.
  - Más detalles en: http://mc-stan.org/rstanarm/articles/priors.html

Discusión final

- La inferencia Clásica y la Bayesiana tienen puntos de contacto bastante claros, pero se diferencian de forma sustancial en cuanto a cómo interpretamos los resultados del análisis de los datos.
- En este curso sólo discutimos (brevemente) los modelos conjugados.
- En general los modelos estadísticos Bayesianos tienen una estructura jerárquica compleja y para ellos no existen soluciones cerradas para $\pi(\theta|\text{datos})$. En estos contextos, solemos recurrir a métodos de aproximación asintóticos y métodos numéricos (MCMC, método de Laplace, etc) para aproximar $\pi(\theta|\text{datos})$ y hacer inferencia. En la bibliografía sugerida al principio de estas slides encontrarán una buena introducción a estas técnicas.