XS3310 Teoría Estadística

I Semestre 2020

2021-07-05

class: center, middle

¿Qué hemos visto hasta ahora?

Introducción a la Estadística Bayesiana: filosofía, historia y un poco de cálculo.

¿Qué vamos a discutir hoy?

Distribuciones previas

Aclaración de parametrización

Para facilitar algunos cálculos en este tema estaremos usando una parametrización alterna de la Gamma (y por ende de la ji-cuadrado y la Exponencial). Anteriormente si teníamos una $Gamma(\alpha, \beta)$ su función de densidad venía dada por

$$f_X(x) = \frac{x^{\alpha-1}e^{-\frac{x}{\beta}}}{\beta^\alpha \Gamma(\alpha)} \quad si \quad x > 0$$

Con la nueva parametrización que vamos a estar utilizando, la función de densidad vendría dada de la siguiente manera

$$f_X(x) = \frac{\beta^\alpha x^{\alpha-1}e^{-\beta x}}{\Gamma(\alpha)} \quad si \quad x > 0$$

Noten que la única diferencia es que el β de la nueva parametrización, llamémoslo β' por un momento, es el inverso multiplicativo del beta de la parametrización vieja. Es decir, $\beta' = \frac{1}{\beta}$. Por lo tanto, para la nueva parametrización de la Gamma tenemos que $E(X) = \frac{\alpha}{\beta}$ y $Var(X) = \frac{\alpha}{\beta^2}$.

Densidades previas conjugadas y estimadores de Bayes

Distribución previa (distribución a priori)

Suponga que tenemos un modelo estadístico con parámetro θ. Su θ es aleatorio entonces su densidad (antes de observar cualquier muestra) se llama **densidad previa**: π.

Ejemplo: $X_1, \ldots, X_n \sim \text{Exp}(\theta)$ y θ es aleatorio tal que $\theta \sim \Gamma(\overset{\alpha}{1}, \overset{\beta}{2})$ entonces

$$\pi(\theta) = \frac{1}{\Gamma(\alpha)} \beta^{\alpha} \theta^{\alpha-1} e^{\beta\theta} = 2e^{-2\theta}, \quad \theta > 0$$

Ejemplo: Sea θ la probabilidad de obtener cara al tirar una moneda.

En este caso antes de modelar exactamente el θ, lo importante es modelar el tipo de moneda. Es decir, supongamos que tenemos dos opciones

- *Moneda justa:* $\theta = \dfrac{1}{2}$ con probabilidad previa 0.8 ($\pi(\frac{1}{2}) = 0.8$).
- *Moneda con solo una cara:* $\theta = 1$ con probabilidad previa 0.2 ($\pi(1) = 0.2$).

En este ejemplo si tuviéramos 100 monedas con probabilidad previa π entonces 20 tendrían solo una cara y 80 serían monedas normales.

Notas:

- π está definida en Ω (espacio paramétrico).
- π es definida antes de obtener la muestra.

Ejemplo (Componentes eléctricos) Supoga que se quiere conocer el tiempo de vida de cierto componente eléctrico. Sabemos que este tiempo se puede modelar con una distribución exponencial con parámetro θ desconocido. Este parámetro asumimos que tiene una distribución previa Gamma.

Un experto en componentes eléctricos conoce mucho de su área y sabe que el parámetro θ tiene las siguientes características:

$$\mathbb{E}[\theta] = 0.0002, \quad \sqrt{\text{Var}(\theta)} = 0.0001.$$

Como sabemos que la previa π es Gamma, podemos deducir lo siguiente:

$$\mathbb{E}[\theta] = \frac{\alpha}{\beta}, \text{Var}(\theta) = \frac{\alpha}{\beta^2}$$

$$\implies \begin{cases} \dfrac{\alpha}{\beta} = 2 \times 10^{-4} \\ \sqrt{\dfrac{\alpha}{\beta^2}} = 1 \times 10^{-4} \end{cases} \implies \beta = 20000, \alpha = 4$$

Notación:

- $X = (X_1, \ldots, X_n)$: vector que contiene la muestra aleatoria.
- Densidad conjunta de X: $f_\theta(x)$.

- Densidad de X condicional en θ: $f_n(x|\theta)$.

Supuesto: X viene de una muestra aleatoria si y solo si X es condicionalmente independiente dado θ.

Consecuencia:

$$f_n(X|\theta) = f(X_1|\theta) \cdot f(X_2|\theta) \cdots f(X_n|\theta)$$

Ejemplo

Si $X = (X_1, \ldots, X_n)$ es una muestra tal que $X_i \sim \text{Exp}(\theta)$,

$$f_n(X|\theta) = \begin{cases} \prod_{i=1}^n \theta e^{-\theta X_i} & \text{si } X_i > 0 \\ 0 & \text{si no} \end{cases}$$
$$= \begin{cases} \theta^n e^{-\theta \sum_{i=1}^n X_i} & X_i > 0 \\ 0 & \text{si no} \end{cases}$$

Densidad posterior

Definición. Considere un modelo estadístico con parámetro θ y muestra aleatoria $X_1, \ldots, X_n$. La densidad condicional de θ dado $X_1, \ldots, X_n$ se llama *densidad posterior*: $\pi(\theta|X)$

Teorema. Bajo las condiciones anteriores:

$$\pi(\theta|X) = \frac{f(X_1|\theta) \cdots f(X_n|\theta)\pi(\theta)}{g_n(X)}$$

para $\theta \in \Omega$, donde g_n es una constante de normalización.

Prueba:

$$\pi(\theta|X) = \frac{\pi(\theta, X)}{\text{marginal de X}} = \frac{\pi(\theta, X)}{\int \pi(\theta, X)\, d\theta} = \frac{P(X|\theta) \cdot \pi(\theta)}{\int \pi(\theta, X)\, d\theta}$$
$$\frac{f_n(X|\theta) \cdot \pi(\theta)}{g_n(X)} = \frac{f(X_1|\theta) \cdots f(X_n|\theta)\pi(\theta)}{g_n(X)}$$

Del ejemplo anterior,

$$f_n(X|\theta) = \theta^n e^{-\theta y}, y = \sum X_i \text{ (estadístico)}$$

Numerador:

$$f_n(X|\theta)\pi(\theta) = \underbrace{\theta^n e^{-\theta y}}_{f_n(X|\theta)} \cdot \underbrace{\frac{200000^4}{3!}\theta^3 e^{-20000 \cdot \theta}}_{\pi(\theta)} = \frac{20000^4}{3!}\theta^{n+3} e^{(20000+y)\theta}$$

Denominador:

$$g_n(x) = \int_0^{+\infty} \theta^{n+3} e^{-(20000+y)\theta}\, d\theta = \frac{\Gamma(n+4)}{(20000+y)^{n+4}}$$

Entonces la posterior corresponde a

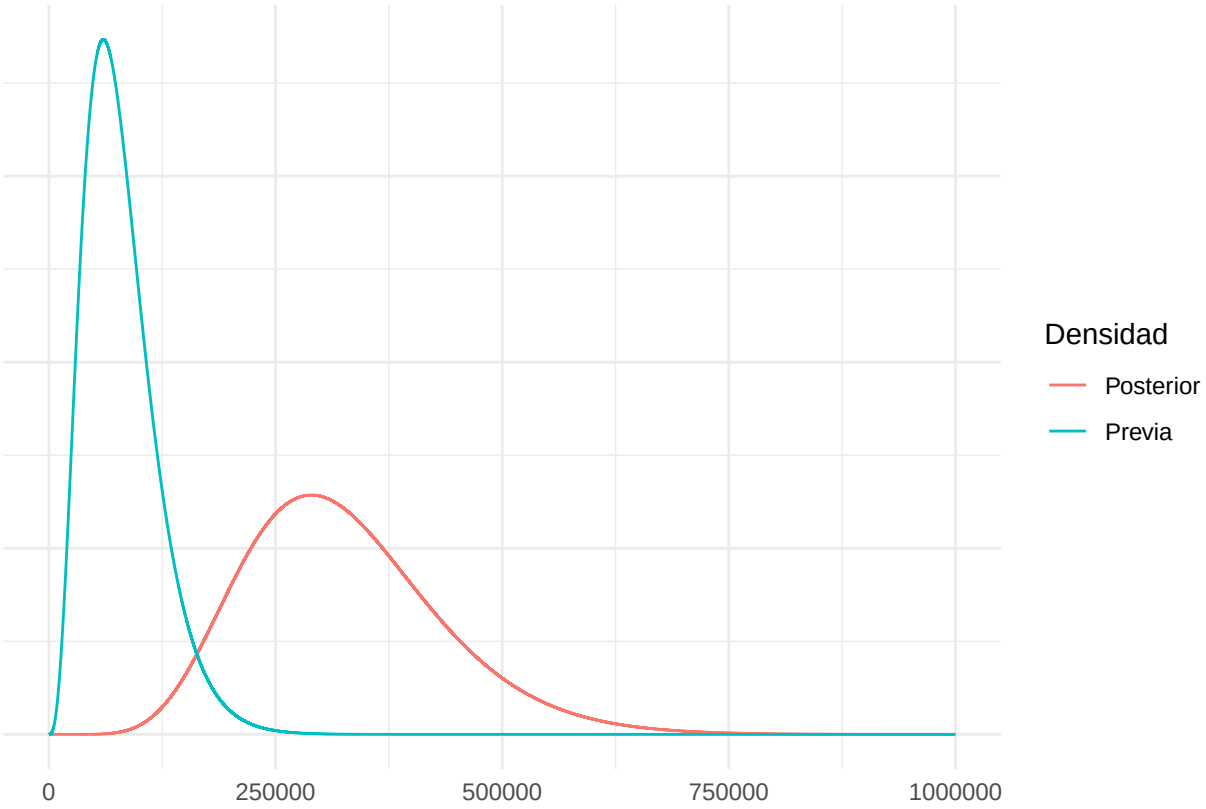
$$\pi(\theta|X) = \frac{\theta^{n+3}e^{-(20000+y)\theta}}{\Gamma(n+4)}(20000+y)^{n+4}$$

que es una $\Gamma(n+4, 20000+y)$.

Con 5 observaciones (horas): 2911, 3403, 3237, 3509, 3118.

$$y = \sum_{i=1}^{5} X_i = 16478, \quad n = 5$$

por lo que $\theta|X \sim \Gamma(9, 36178)$



Es sensible al tamaño de la muestra (una muestra grande implica un efecto de la previa menor).

Hiperparámetros: parámetros de la previa o posterior.

Proceso de modelación de parámetros.

De ahora en adelante vamos a entender un modelo como el conjunto de los datos $X_1, \ldots, X_n$, la función de densidad f y el parámetro de la densidad θ. Estos dos últimos resumen el comportamiento de los datos.

Ahora para identificar este modelo se hace por partes,

1. La información previa $\pi(\theta)$ es la información extra o basado en la experiencia que tengo del mdoelo.
2. Los datos es la información observada. La función de densidad f filtra y mejora la información de la previa.
3. La densidad posterior es la "mezcla" entre la información y los datos observados. Es una versión más informada de la distribución del parámetro.

Función de verosimilitud

Bajo el modelo estadístico anterior a $f_n(X|\theta)$ se le llama **verosimilitud** o **función de verosimilitud**.

Observación. En el caso de una función de verosimilitud, el argumento es θ.

Ejemplo.

Sea θ la proporción de aparatos defectuosos, con $\theta \in [0,1]$

$$X_i = \begin{cases} 0 & \text{falló} \\ 1 & \text{no falló} \end{cases}$$

$\{X_i\}_{i=1}^n$ es una muestra aleatoria y $X_i \sim Ber(\theta)$.

- **Verosimilitud**

$$f_n(X|\theta) = \prod_{i=1}^n f(X_i|\theta) = \begin{cases} \theta^{\sum X_i}(1-\theta)^{n-\sum X_i} & X_i = 0,1 \ \forall i \\ 0 & \text{si no} \end{cases}$$

- **Previa**:

$$\pi(\theta) = 1_{\{0 \leq \theta \leq 1\}}$$

- **Posterior**:

Por el teorema de Bayes,

$$\begin{aligned} \pi(\theta|X) &\propto \theta^y(1-\theta)^{n-y} \cdot 1 \\ &= \theta^{\overbrace{y+1}^{\alpha}-1}(1-\theta)^{\overbrace{n-y+1}^{\beta}-1} \qquad \implies \theta|X \sim \text{Beta}(y+1, n-y+1) \end{aligned}$$

- **Predicción**.

Supuesto: los datos son secuenciales. Calculamos la distribución posterior secuencialmente:

$$\begin{aligned} \pi(\theta|X_1) &\propto \pi(\theta)f(X_1|\theta) \\ \pi(\theta|X_1, X_2) &\propto \pi(\theta)f(X_1, X_2|\theta) \\ &= \pi(\theta)f(X_1|\theta)f(X_2|\theta) \text{ (por independencia condicional)} \\ &= \pi(\theta|X_1)f(X_2|\theta) \\ &\vdots \\ \pi(\theta|X_1, \ldots, X_n) &\propto f(X_n|\theta)\pi(\theta|X_1, \ldots, X_{n-1}) \end{aligned}$$

Bajo independencia condicional no hay diferencia en la posterior si los datos son secuenciales.

Luego,

$$\begin{aligned} g_n(X) &= \int_\Omega f(X_n|\theta)\pi(\theta|X_1, \ldots, X_{n-1})\, d\theta \\ &= P(X_n|X_1, \ldots, X_{n-1}) \text{ (Predicción para } X_n) \end{aligned}$$

Continuando con el ejemplo de los artefactos, $P(X_6 > 3000|X_1, X_2, X_3, X_4, X_5)$. Se necesita calcular $f(X_6|X)$. Dado que

$$\pi(\theta|X) = 2.6 \times 10^{36}\theta^8 e^{-36178\theta}$$

se tiene

$$f(X_6|X) = 2.6 \times 10^{36} \int_0^1 \underbrace{\theta e^{-\theta X_6}}_{\text{Densidad de } X_6} \theta^8 e^{-36178\theta}\, d\theta = \frac{9.55 \times 10^{41}}{(X_6 + 36178)^{10}}$$

Entonces,

$$P(X_6 > 3000) = \int_{3000}^{\infty} \frac{9.55 \times 10^{41}}{(X_6 + 36178)^{10}}\, dX_6 = 0.4882$$

La vida media se calcula como $\frac{1}{2} = P(X_6 > u|X)$.

Familias conjugadas

Definición. Sea $X_1, \ldots, X_n$ i.i.d. condicional dado θ con densidad $f(X|\theta)$. Sea ψ la familia de posibles densidades previas sobre Ω. Si, sin importar los datos, la posterior pertenece a ψ, entonces decimos que ψ es una familia conjugada de previas.

Ejemplos:

- La familia Beta es familia conjugada para muestras según una Bernoulli.
- La familia Gama es familia conjugada para muestras exponenciales.
- Para el caso Poisson, si $X_1, \ldots, X_n \sim Poi(\lambda)$,entonces la familia Gamma es familia conjugada.

La función de densidad de una Poisson es $P(X_i = k) = e^{-\lambda}\frac{\lambda^k}{k!}$. La verosimilitud corresponde a

$$f_n(X|\lambda) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda_i^X}{X_i!} = \frac{e^{-n\lambda^y}}{\prod_{i=1}^n X_i}.$$

La previa de λ está definida por $\pi(\lambda) \propto \lambda^{\alpha-1}e^{-\beta\lambda}$. Por lo tanto, la posterior es

$$\pi(\lambda|X) \propto \lambda^{y+\alpha-1}e^{-(\beta+n)\lambda} \implies \lambda|X \sim \Gamma(y + \alpha, \beta + n)$$

- En el caso normal, si $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$,entonces la familia normal es conjugada si σ^2 es conocido.

Si $\theta \sim N(\mu_0, V_0^2) \implies \theta|X \sim N(\mu_1, V_1^2)$ donde,

$$\mu_1 = \frac{\sigma^2\mu_0 + nV_0^2\bar{X}_n}{\sigma^2 + nV_0^2} = \frac{\sigma^2}{\sigma^2 + nV_0^2}\mu_0 + \frac{nV_0^2}{\sigma^2 + nV_0^2}\bar{X}_n$$

Combina de manera ponderada la previa y la de los datos.

Ejemplo

Considere una verosimilitud Poisson(λ) y una previa

$$\pi(\lambda) = \begin{cases} 2e^{-2\lambda} & \lambda > 0 \\ 0 & \lambda \geq 0 \end{cases} \quad \lambda \sim \Gamma(1,2)$$

Supongamos que es una muestra aleatoria de tamaño n. ¿Cuál es el número de observciones para reducir la varianza, a lo sumo, a 0.01?

Por teorema de Bayes, la posterior $\lambda|x \sim \Gamma(y+1, n+2)$. Luego, la varianza de la Gamma es

$$\frac{\alpha}{\beta^2} = \frac{\sum x_i + 1}{(n+2)^2} \leq 0.01 \implies \frac{1}{(n+2)^2} \leq \frac{\sum x_i + 1}{(n+2)^2} \leq 0.01 \implies 100 \leq (n+2)^2 \implies n \geq 8$$

Teorema. Si $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$ con σ^2 conocido y la previa es $\theta \sim N(\mu_0, V_0^2)$, entonces $\theta|X \sim N(\mu_1, V_1^2)$ donde

$$\mu_1 = \frac{\sigma^2\mu_0 + nV_0^2\bar{X}_n}{\sigma^2 + nV_0^2}, \quad V_1^2 = \frac{\sigma^2 V_0^2}{\sigma^2 + nV_0^2}$$

Prueba:

- **Verosimilitud**:

$$f_n(X|\theta) \propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i\theta)^2\right]$$

Luego,

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\theta)^2 &= \sum_{i=1}^{n}(X_i - \bar{X} + \bar{X} - \theta)^2 \\ &= n(\bar{X}+\theta)^2 + \sum_{i=1}^{n}(X_i-\bar{X})^2 + \underbrace{2\sum_{i=1}^{n}(X_i-\bar{X})(\bar{X}-\theta)}_{=0 \text{ pues } \sum Xi = n\bar{X})}\end{aligned}$$

Entonces

$$f_n(X|\theta) \propto \exp\left[-\frac{n}{2\sigma^2}(\bar{X}-\theta)^2\right].$$

- **Previa**:

$$\pi(\theta) \propto \exp\left[-\frac{1}{2V_0^2}(\theta-\mu_0)^2\right].$$

- **Posterior**:

$$\pi(\theta|X) \propto \exp\left[-\frac{n}{2\sigma^2}(\bar{X}-\theta)^2 - \frac{1}{2V_0^2}(\theta-\mu_0)^2\right].$$

Con μ_1 y V_1^2 definidos anteriormente, se puede comprobar la siguiente identidad:

$$-\frac{n}{\sigma^2}(\bar{X}-\theta)^2-\frac{1}{V_0^2}(\theta-\mu_0)^2=\frac{1}{V_1^2}(\theta-\mu_1)^2+\underbrace{\frac{n}{\sigma^2+nV_0^2}(\bar{X}_n-\mu_0)^2}_{\text{Constante con respecto a }\theta}$$

Por lo tanto,

$$\pi(\theta|X)\propto\exp\left[-\frac{n}{2V_1^2}(\theta-\mu_1)^2\right]$$

Media posterior:

$$\mu_1=\underbrace{\frac{\sigma^2}{\sigma^2+nV_0^2}}_{W_1}\mu_0+\underbrace{\frac{nV_0^2}{\sigma^2+nV_0^2}}_{W_2}\bar{X}_n$$

Afirmaciones:

1) Si V_0^2 y σ^2 son fijos, entonces $W_1 \xrightarrow[n\to\infty]{} 0$ (la importancia de la media empírica crece conforme aumenta n).
2) Si V_0^2 y n son fijos, entonces $W_2 \xrightarrow[\sigma^2\to\infty]{} 0$ (la importancia de la media empírica decrece conforme la muestra es menos precisa).
3) Si σ^2 y n son fijos, entonces $W_2 \xrightarrow[V_0^2\to\infty]{} 1$ (la importancia de la media empírica crece conforma la previa es menos precisa).

Ejemplo (determinación de n)

Sean $X_1,\ldots,X_n\sim N(\theta,1)$ y $\theta\sim N(\mu_0,4)$. Sabemos que

$$V_1^2=\frac{\sigma^2V_0^2}{\sigma^2+nV_0^2}.$$

Buscamos que $V_1\leq 0.01$, entonces

$$\frac{4}{4n+1}\leq 0.01 \implies n\geq 99.75 \text{ (al menos 100 observaciones)}$$

Retomando el ejemplo de la clase anterior

i	θ_i	$\pi(\theta_i)$	$\mathcal{L}(\theta_i \mid x)$	$\mathcal{L}(\theta_i \mid x)\pi(\theta_i)$	$\pi(\theta_i \mid x)$
1	2	0.50	$2.15\cdot10^{-4}$	$10.76\cdot10^{-5}$	0.416
2	2.5	0.25	$3.21\cdot10^{-4}$	$8.03\cdot10^{-5}$	0.311
3	3	0.25	$2.82\cdot10^{-4}$	$7.06\cdot10^{-5}$	0.273

- ¿Cuál fue el aporte Bayesiano al estudio?
- ¿Qué hubiera pasado si hubiéramos asignado probabilidades iguales a previa para cada valor de θ? Esto se llama utilizar una *previa no informativa* pues no está influyendo mucho en los valores que pueda tomar θ.

Distribuciones previas

- Supongamos ahora que θ ya no toma solo ciertos valores, sino que puede tomar cualquier valor mayor a cero. Ya no vamos a poder usar una distribución de probabilidad como previa sino que más bien necesitamos una función de densidad. Podríamos usar una previa no informativa, de manera que $\theta \sim Unif(0, B)$ donde B va a representar un valor arbitrario muy grande.
- Ahora vamos a suponer que tenemos una muestra aleatoria $X_1, X_2, ..., X_n$ tal que $X_j \sim Poisson(\theta)$, por lo que se cumple que $\mathcal{L}(\theta|x) = \frac{\theta^{\sum_{j=1}^{n} x_j} e^{-n\theta}}{\prod_{j=1}^{n} x_j!}$. Para llegar a cuál sería la función de densidad a posteriori podemos utilizar el Teorema de Bayes:

$$\pi(\theta|x) = \frac{\mathcal{L}(\theta|x)\pi(\theta)}{\int_0^{+\infty} \mathcal{L}(\theta|x)\pi(\theta)}$$

Distribuciones previas

Y dada la distribución a previa que escogimos entonces tenemos lo siguiente:

$$\mathcal{L}(\theta|x)\pi(\theta) = \frac{\theta^{\sum_{j=1}^{n} x_j} e^{-n\theta}}{\prod_{j=1}^{n} x_j!} \cdot \frac{1}{B}$$

Este es el numerador de la expresión del Teorema de Bayes, pero por lo expresado anteriormente, para encontrar la posteriori, sabemos que $\pi(\theta|x) \propto \mathcal{L}(\theta|x)\pi(\theta)$.

Esto quiere decir que debemos encontrar el núcleo (o la parte variable) de esta expresión e identificar a qué función de densidad conocida pertenece. Por lo tanto,

$$\pi(\theta|x) \propto \mathcal{L}(\theta|x)\pi(\theta)$$

$$\Rightarrow \pi(\theta|x) \propto \frac{\theta^{\sum x_j} e^{-n\theta}}{\prod x_j!} \cdot \frac{1}{B} \propto \theta^{n\bar{x}} e^{-n\theta}$$

Distribuciones previas

$$\Rightarrow \pi(\theta|x) \propto \frac{\theta^{\sum x_j} e^{-n\theta}}{\prod x_j!} \cdot \frac{1}{B} \propto \theta^{n\bar{x}} e^{-n\theta}$$

Este es el núcleo de una distribución Gamma con $\alpha = n\bar{x} + 1$ y $\beta = n$. Por lo tanto, podemos decir que la distribución a posteriori para θ es una $Gamma(n\bar{x}+1, n)$, denotado como $\theta|x \sim Gamma(n\bar{x}+1, n)$. Nótese como dato curioso que esta es una distribución centrada en $\frac{n\bar{x}+1}{n} = \bar{x} + \frac{1}{n}$, el cual es un valor muy cercano a $\bar{x}$, especialmente con un n muy grande.

Nosotros ya sabíamos que $\bar{x}$ es el estimador de máxima verosimilitud para θ, por lo tanto si no tenemos mucha información sobre θ a previa entonces tendría sentido basar nuestro conocimiento posterior alrededor de su estimador máximo verosímil. Esto nuevamente representa como el análisis Bayesiano con previas no informativas es muy similar al análisis frecuentista.

Distribuciones previas

Siguiendo con el mismo ejemplo, supongamos que tenemos más información sobre nuestro parámetro θ, ¿cómo podríamos simularla? Sabemos que θ, al ser la media de una población Poisson, debe ser mayor a cero. Ya utilizamos la distribución Uniforme como una previa no informativa, por lo que sería contraproducente usarla cuando tenemos más información. Podríamos utilizar una distribución Gamma (o alguna de sus variaciones) ya que esta es para valores mayores que cero y además se pueden escoger parámetros α y β de manera que satisfagan nuestro conocimiento inicial sobre θ.

Este α y β llevan el nombre de **hiperparámetros** y podríamos definirlos como los parámetros de la distribución de un parámetro. Siguiendo el ejemplo podemos decir que inicialmente $\theta \sim Gamma(\alpha, \beta)$. Es decir,

$$\pi(\theta) = \frac{\beta^{\alpha}\theta^{\alpha-1}e^{-\beta\theta}}{\Gamma(\alpha)}$$

Distribuciones previas

Podemos proceder a encontrar la distribución posterior:

$$\pi(\theta|x) = \frac{\mathcal{L}(\theta|x)\pi(\theta)}{\int_0^{+\infty}\mathcal{L}(\theta|x)\pi(\theta)} \propto \mathcal{L}(\theta|x)\pi(\theta) = \frac{\theta^{n\bar{x}}e^{-n\theta}}{\prod x_j!} \cdot \frac{\beta^{\alpha}\theta^{\alpha-1}e^{-\beta\theta}}{\Gamma(\alpha)}$$

$$\propto \theta^{n\bar{x}}e^{-n\theta} \cdot \theta^{\alpha-1}e^{-\beta\theta} = \theta^{n\bar{x}+\alpha-1}e^{-\theta(n+\beta)}$$

Este es el núcleo de una Gamma con parámetros $n\bar{x} + \alpha$ y $n + \beta$. Por lo tanto concluimos que $\theta|x \sim Gamma(n\bar{x} + \alpha, n + \beta)$. Solo con propósitos de comparación podemos observar que esta distribución tiene media $\frac{n\bar{x}+\alpha}{n+\beta}$ la cual la podemos reescribir como $\frac{n\bar{x}+\beta\left(\frac{\alpha}{\beta}\right)}{n+\beta}$. Podemos ver que esta tiene la forma de un promedio ponderado de la media muestral $\bar{x}$ y de la media a previa $\frac{\alpha}{\beta}$. Entre mayor sea el tamaño de la muestra, mayor influencia van a tener los datos sobre la información posterior, mientras que si el β es más grande, entonces mayor influencia va a tener la previa sobre la posteriori.

Distribuciones previas

Función de densidad a previa Gamma con $\alpha = 4$ y $\beta = 2$ (azul) y a posteriori (roja), utilizando $n = 10$ y $\bar{x} = 3.5$ (línea punteada) para una muestra aleatoria de una población Poisson(θ).

Distribuciones previas

En la figura anterior podemos observar la diferencia entre la función de densidad a previa (utilizando $\alpha = 4$ y $\beta = 2$) y la función de densidad a posteriori si hubiésemos obtenido una muestra de tamaño 10 cuyo promedio fuera de 3.5. Podemos ver que antes de observar los datos la previa nos indicaba que la media era de 2. Cuando observamos los datos la media muestral era de 3.5 por lo que nuestro conocimiento a posteriori se "modificó" para representar esta nueva información. Noten como la función de densidad a posteriori tiene una media más cercana al valor observado en la muestra y como valores alrededor de 3.5 se volvieron más probables que antes.

https://andreavargasmontero.github.io/apps_shiny_bayes/

Distribuciones previas

En estadística Bayesiana es posible modelar los hiperparámetros de la distribución de un parámetro. Por ejemplo, yo podría trata el α de este ejemplo como un parámetro desconocido al cual le puedo modelar la incertidumbre por medio de una distribución a previa, supongamos que por medio de otra Gamma con hiperparámetros α' y β'. Esto ocasiona que tengamos distintas etapas de previas en nuestro modelo:

$$X_1, X_2, ..., X_n \text{ t.q. } X_j \sim Poisson(\theta)$$
$$\theta|\alpha \sim Gamma(\alpha, \beta)$$
$$\alpha \sim Gamma(\alpha', \beta')$$

Distribuciones previas

Este tipo de modelo se denomina un **modelo Bayesiano jerárquico** y es el tipo de modelo más utilizando en la práctica pues tiene varias ventajas sobre algunos análisis frecuentistas. Sin embargo presentan un problema y es que en muy pocas ocasiones se puede llegar a un modelo a posteriori conocido por lo que se necesitan de simulaciones numéricas para poder hacer el análisis Bayesiano.

Aunque existen muchos métodos hoy en día el más popular sigue siendo el método de cadenas de Markov Monte Carlo (MCMC) mediante muestreo de Gibbs. Dependiendo de la complejidad del modelo este requiere de mucha potencia computacional, por lo que estos modelos no eran muy utilizados en los inicios de la Estadística Bayesiana. Fue hasta la década de los 90s, donde la población en general tuvo mayor acceso a computadoras más poderosas, donde las técnicas Bayesianas empezaron a cobrar una mayor relevancia.

Hasta el momento hemos hecho la selección de la previa un poco intuitivamente, sin embargo en la práctica la selección de la previa puede deberse a varios factores.

Selección de distribuciones previas

Previas informativas

Una forma de seleccionar una previa es utilizando una previa informativa. Claramente, para poder hacer uso de esta debemos tener bastante conocimiento sobre el fenómeno de interés que estamos estudiando (como en el ejemplo que hicimos al inicio sobre la proporción de desempleo en el país). Para ello es imperativo tener información de varios expertos para poder llegar a construir el modelo estadístico que mejor represente esa información. Es importante destacar algo del uso de previas muy informativas; entre más informativa sea la previa que estamos utilizando entonces más datos se necesitan para tratar de observar algo nuevo. Imaginen que están casi 100% seguros sobre cuánto debería ser el valor de cierto parámetro desconocido. Para que ustedes piensen cambien de opinión entonces van a requerir de una gran cantidad de evidencia que apunte a lo contrario. Eso es lo que pasa si se usa una previa muy informativa; la posterior se va a parecer mucho a esta y solo podrá cambiar si existen muchos datos que apuntan a lo contrario.

Selección de distribuciones previas

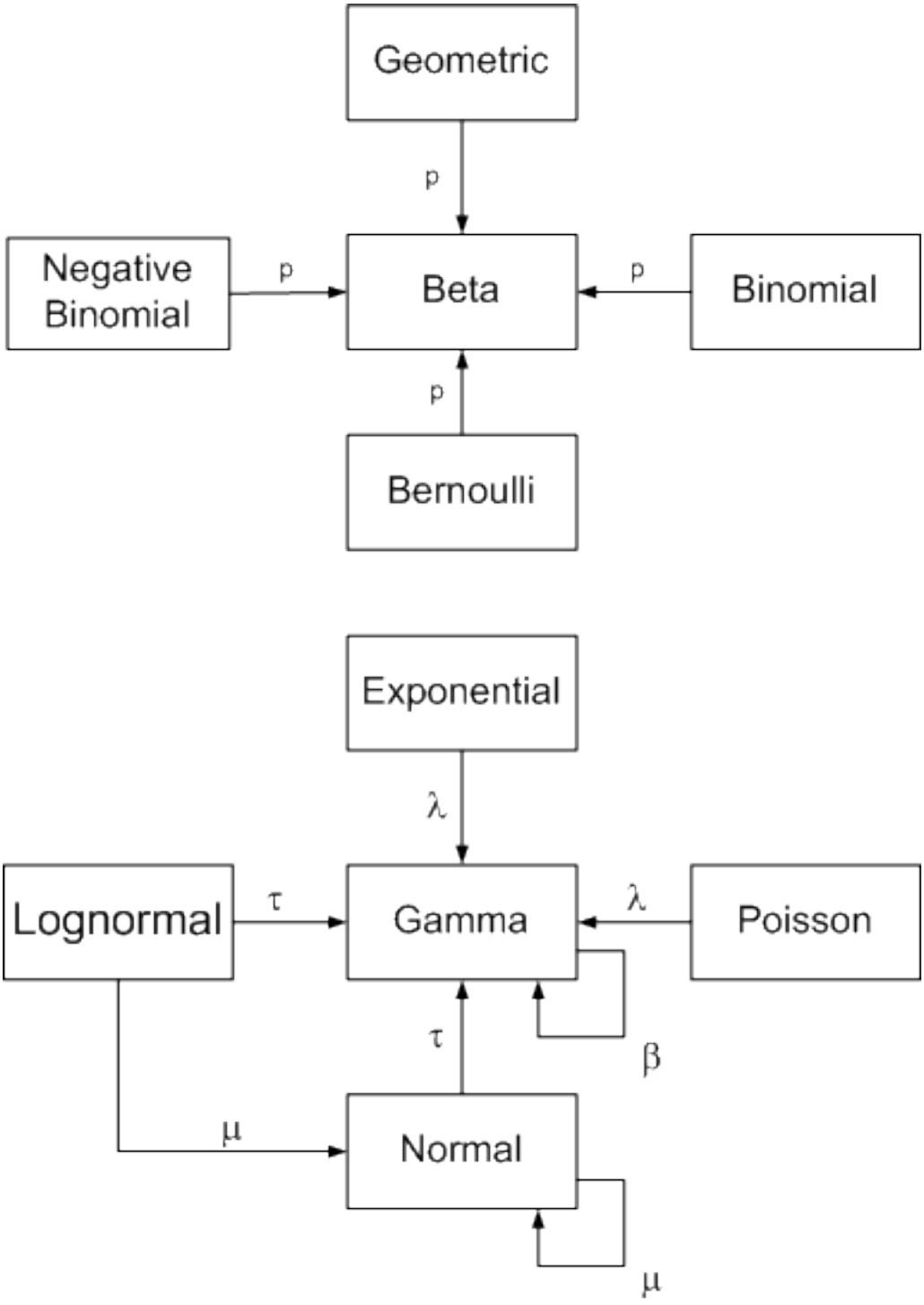
Previas conjugadas

Un tipo de previas informativas que son sumamente convenientes de utilizar son las **previas conjugadas**. Se dice que una previa es conjugada si la distribución de la posteriori pertenece a la misma familia que la distribución de la previa. El ejemplo anterior donde utilizamos una previa Gamma y terminamos con una posteriori Gamma es un ejemplo de una previa conjugada. Este tipo de previas son muy convenientes pues nos aseguran que vamos a tener una distribución a posteriori conocida, lo único que cambiaría son los parámetros de la distribución. Sin embargo, esto no quiere decir que la Gamma siempre vaya a ser una previa conjugada; esto solo va a pasar si los datos son Poisson. Por lo tanto, una parte importante que permite que la previa sea o no sea conjugada es la distribución de la población.

Distribuciones conjugadas conocidas

Distribución	previa conjugada
Bernoulli(p)	$p \sim Beta(\alpha, \beta)$
Binomial(n, p)	$p \sim Beta(\alpha, \beta)$
Binomial Negativa(n, p)	$p \sim Beta(\alpha, \beta)$
Poisson(λ)	$\lambda \sim Gamma(\alpha, \beta)$
Exponencial(θ)	$\theta \sim Gamma(\alpha, \beta)$
Normal(μ, σ^2)	$\mu \sim N(\mu_0, \sigma_0^2)$ y $\sigma^2 \sim GammaInversa(\alpha, \beta)$

Distribuciones conjugadas conocidas



https://en.wikipedia.org/wiki/Conjugate_prior

https://www.johndcook.com/blog/conjugate_prior_diagram/

Previas no informativas

Este tipo de previa es posiblemente el más utilizado en la práctica. Con modelos relativamente simples utilizar una previa no informativa por lo general brinda resultados muy similares a los resultados frecuentistas, mientras que con modelos jerárquicos más complejos los resultados sí pueden ser más distintos. No obstante, la mayoría del tiempo en que se quiere hacer inferencia sobre parámetros desconocidos no se tiene mucha información al respecto, aparte de un posible rango en donde se puedan encontrar; por esto es más atractivo utilizar una previa no informativa. Diremos que una previa es no informativa si le da la libertad a los datos de encontrar los mejores valores de los parámetros.

Por mucho tiempo se tuvo la idea de que las previas no informativas eran exclusivamente las uniformes, pues asignaban probabilidades iguales a cualquier parámetro. Sin embargo, hay quienes criticaron esto, como Fisher, diciendo que no es posible que la uniforme sea siempre no informativa.

Previas no informativas

El argumento de Fisher era algo así:

> Supongamos que tenemos un parámetro desconocido θ que representa la probabilidad de éxito de un experimento Bernoulli. Supongamos que no sabemos nada de θ entonces decimos que $\theta \sim Unif(0,1)$. Ahora, si no sabemos nada de θ entonces tampoco sabemos nada de $\lambda = -\ln(\theta)$, por lo que también podríamos preferir una previa uniforme para λ. Sería lógico pensar que, mediante las técnicas de transformaciones, la transformación aplicada a la previa de θ llegue a la previa de λ. Sin embargo hay un problema lógico pues si aplicamos las técnicas de transformación a la previa de θ no vamos a llegar a la misma previa de θ. Esto sugiere que una distribución uniforme no es un buen ejemplo de una previa no informativa.

Previas no informativas

Un tipo de previa no informativa es la **previa de Jeffreys**. Estas hacen uso de la Información de Fisher, previamente utilizada en el Tema 1. Por lo tanto, si tenemos una muestra aleatoria $X_1, X_2, ..., X_n$ de una población con función de densidad $f_X(x|\theta)$ entonces la información de Fisher se define como:

$$I(\theta) = -E\left[\frac{\partial^2 \ln(f_X(x|\theta))}{\partial \theta^2}\right]$$

Por lo tanto, la previa de Jeffreys se define como:

$$\pi_J(\theta) = c\sqrt{I(\theta)}$$

Donde c es una constante positiva mayor a cero que asegura que esta función de densidad integra a uno. Como c es constante entonces podemos decir que $\pi_J(\theta) \propto \sqrt{I(\theta)}$, por lo que muchas veces no nos importa el valor de c para encontrar la distribución a posteriori a partir de la previa de Jeffreys.

Previas no informativas

Ejemplo: Sea $X_1, X_2, ..., X_n$ una muestra aleatoria tal que $X_j \sim N(\mu, 1)$. Encuentre la previa de Jeffreys para μ.

Solución: Recordemos la función de densidad de una Normal:

$$f_X(x|\mu) = \sqrt{2\pi}e^{-\frac{(x-\mu)^2}{2}}$$

Por lo que el logaritmo natural de esta sería:

$$\ln(f_X(x|\mu)) = \frac{1}{2}\ln(2\pi) - \frac{(x-\mu)^2}{2}$$

Derivamos dos veces con respecto a μ:

$$\frac{\partial \ln(f_X(x|\mu))}{\partial \mu} = x - \mu$$

$$\Rightarrow \frac{\partial^2 \ln(f_X(x|\mu))}{\partial \mu^2} = -1$$

Previas no informativas

Finalmente obtenemos

$$I(\mu) = 1$$

Por lo tanto, podemos concluir que $\pi_J(\theta) \propto 1$, es decir, es proporcional a una constante. Por lo tanto, la previa de Jeffreys para estimar a μ sería una distribución Uniforme, escogida en un rango bastante amplio.

Hay un par de puntos importantes de destacar cuando se usa la previa de Jeffreys. El primero de ellos es que no siempre se va a llegar a una función de densidad propia (es decir, una función de densidad que integre a 1 en su dominio). Por lo general se ignora este problema si la posteriori si es propia. Por lo tanto, siempre que se vaya a utilizar una previa impropia hay que revisar que la posteriori sea propia, si no los resultados no tendrían sentido. El segundo punto es un poco más filosófico y trata con el hecho de que las previas se deben elegir antes de ver los datos. La previa de Jeffreys usa la distribución de los datos para encontrar una previa lo que contradice lo que muchos dicen sobre el planteamiento de la previa.

Densidades previas impropias

Definición. Sea π una función positiva cuyo dominio está en Ω. Suponga que $\int \pi(\theta)\, d\theta = \infty$. Entonces decimos que π es una **densidad impropia**.

Ejemplo: $\theta \sim \text{Unif}(\mathbb{R})$, $\lambda \sim \text{Unif}(0, \infty)$.

Una técnica para seleccionar distribuciones impropia es sustituir los hiperparámetros previos por 0.

Ejemplo:

Se presenta el número de soldados prusianos muertos por una patada de caballo (280 conteros, unidades de combate en 20 años).

Unidades	Ocurrencias
144	0
91	1
32	2
11	3
2	4

- Muestra de Poisson: $X_1 = 0, X_2 = 1, X_3 = 1, \ldots, X_{280} = 0 \sim \text{Poi}(\lambda)$.
- Previa: $\lambda \sim \Gamma(\alpha, \beta)$.
- Posterior: $\lambda|X \sim \Gamma(y + \alpha, n + \beta) = \Gamma(196 + \alpha, 280 + \beta)$.

Sustituyendo, $\alpha = \beta = 0$

$$\pi(\lambda) = \frac{1}{\Gamma(\alpha)}\beta^{\alpha}\lambda^{\alpha-1}e^{\beta\lambda}$$

$$\propto \lambda^{\alpha-1}e^{-\lambda\beta}$$

$$= \frac{1}{\lambda}$$

donde $\int_0^{\infty} \frac{1}{\lambda} d\lambda = \infty$.

Por teorema de Bayes,

$$\theta|X \sim \Gamma(196, 280)$$

Funciones de pérdida

Definición. Sean $X_1, \ldots, X_n$ datos observables cuyo modelo está indexado por $\theta \in \Omega$. Un estimador de θ es cualquier estadístico $\delta(X_1, \ldots, X_n)$.

Notación:

- Estimador $\rightarrow \delta(X_1, \ldots, X_n)$.
- Estimación o estimado: $\delta(X_1, \ldots, X_n)(\omega) = \delta(\overbrace{x_1, \ldots, x_n}^{datos})$

Definición. Una **función de pérdida** es una función de dos variables:

$$L(\theta, a), \quad \theta \in \Omega$$

con a un número real.

Interpretación: es lo que pierde un analista cuando el parámetro es θ y el estimador es a.

Asuma que θ tiene una previa. La pérdida esperada es

$$\mathbb{E}[L(\theta, a)] = \int_{\Omega} L(\theta, a)\pi(\theta)\, d\theta$$

la cual es una función de a, que a su vez es función de $X_1, \ldots, X_n$. Asuma que a se selecciona el minimizar esta esperanza. A ese estimador $a = \delta^*(X_1, \ldots, X_n)$ se le llama **estimador bayesiano**, si ponderamos los parámetros con respecto a la posterior.

$$\mathbb{E}[L(\theta, \delta^*)|X] = \int_\Omega L(\theta, a)\pi(\theta)\, d\theta = \min_a \mathbb{E}[L(\theta|a)X].$$

Función de pérdida cuadrática

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

En el caso en que θ es real y $\mathbb{E}[\theta|X]$ es finita, entonces

$$\delta^*(X_1, \ldots, X_n) = \mathbb{E}[\theta|X] \text{ cuando } L(\theta, a) = (\theta - a)^2.$$

Ejemplo: $X_1, \ldots, X_n \sim \text{Ber}(\theta)$, $\theta \sim \text{Beta}(\alpha, \beta) \implies \theta|X \sim \text{Beta}(\alpha + y, \beta + n - y)$.

El estimador de θ es

$$\delta^*(X_1, \ldots, X_n) = \frac{\alpha + y}{\alpha + \beta + n} = \overbrace{\frac{\alpha}{\alpha + \beta}}^{\text{Esperanza previa}} \cdot \frac{\alpha + \beta}{\alpha + \beta + n} + \overbrace{\frac{y}{n}}^{\bar{X}} \cdot \frac{n}{\alpha + \beta + n}.$$

Función de pérdida absoluta

$$L(\theta, a) = |\theta - a|$$

La pérdida esperada es

$$f(a) = \mathbb{E}[L(\theta, a)|X] = \int_{-\infty}^{+\infty} |\theta - a|\pi(\theta|X)\, d\theta = \int_a^{+\infty} (\theta - a)\pi(\theta|X)\, d\theta + \int_{-\infty}^{a} (a - \theta)\pi(\theta|X)\, d\theta$$

Usando el teorema fundamental del cálculo,

$$F_\pi(a|X) = \int_{-\infty}^{\hat{a}} \pi(\theta|X)\, d\theta = \frac{1}{2} \Leftrightarrow \hat{a} = \operatorname*{argmin}_a f(a)$$

La **mediana** es el punto de $X_{0.5}$ tal que $F(X_{0.5}) = \frac{1}{2}$.

Corolario. Bajo la función de pérdida absoluta, el estimador bayesiano es la mediana posterior.

Ejemplo: Bernoulli.

$$\frac{1}{\text{Beta}(\alpha + y, \beta + n - y)} \int_{-\infty}^{X_{0.5}} \theta^{\alpha+y-1}(1 - \theta)^{\beta+n-y-1}\, d\theta = \frac{1}{2}$$

Resuelva para $X_{0.5}$.

Otras funciones de pérdida

- $L(\theta, a) = |\theta - a|^k$, $k \neq 1, 2$, $0 < k < 1$.
- $L(\theta, a) = \lambda(\theta)|\theta - a|^2$ ($\lambda(\theta)$ penaliza la magnitud del parámetro).
- $L(\theta, a) = \begin{cases} 3(\theta - a)^2 & \theta \leq a \text{ (sobreestima)} \\ (\theta - a)^2 & \theta \geq a \text{ (subestima)} \end{cases}$

Efecto de muestras grandes

Ejemplo: ítemes malos (proporción: θ), $\theta \in [0,1]$. Función de pérdida cuadrática. El tamaño de muestra son $n = 100$ ítemes, de los cuales $y = 10$ están malos.

$$X_1, \ldots, X_n \sim \text{Ber}(\theta)$$

- Primer previa. $\alpha = \beta = 1$ (Beta). El estimador bayesiano corresponde a

$$\mathbb{E}[\theta|X] = \frac{\alpha + y}{\alpha + \beta + n} = \frac{1+10}{2+100} = 0.108$$

- Segunda previa. $\alpha = 1, \beta = 2 \implies \pi(\theta) = 2e^{-2\theta}, \theta > 0$.

$$\mathbb{E}[\theta|X] = \frac{1+10}{1+2+100} = \frac{11}{103} = 0.107$$

La media es $\bar{X}_n = \dfrac{10}{100} = 0.1$.

Consistencia

Definición. Un estimador de θ $\delta(X_1, \ldots, X_n)$ es consistente si

$$\delta(X_1, \ldots, X_n) \xrightarrow[n\to\infty]{\mathbb{P}} \theta.$$

Bajo pérdida cuadrática, $\mathbb{E}[\theta|X] = W_1\mathbb{E}[\theta] + X_2\bar{X}_n = \delta^*$. Sabemos, por ley de grandes números, que $\bar{X}_n \xrightarrow[n\to\infty]{\mathbb{P}} \theta$. Además, $W_1 \xrightarrow[n\to\infty]{} 0$ y $W_2 \xrightarrow[n\to\infty]{} 1$.

En los ejemplos que hemos analizado

$$\delta^* \xrightarrow[n\to\infty]{\mathbb{P}} \theta$$

Teorema. Bajo condiciones generales, los estimadores bayesianos son consistentes.

Estimador. Si $X_1, \ldots, X_n$ es una muestra en un modelo indexado por θ, $\theta \in \Omega$ (k-dimensiones), sea

$$h : \Omega \to H \subset \mathbb{R}^d.$$

Sea $\psi = h(\theta)$. Un **estimador** de ψ es un estadístico $\delta^*(X_1, \ldots, X_n) \in H$. A $\delta^*(X_1, \ldots, X_n)$ estimador de ψ se puede evaluar y construir estimadores nuevos.

Ejemplo. $X_1, \ldots, X_n \sim \text{Exp}(\theta)$, $\theta|X \sim \Gamma(\alpha, \beta) = \Gamma(4, 8.6)$. La característica de interés es $\psi = \dfrac{1}{\theta}$, el valor esperado del tiempo de fallo.

Es estimador se calcula de la siguiente manera:

$$\begin{aligned}\delta^*(x) = \mathbb{E}[\psi|x] &= \int_0^\infty \frac{1}{\theta}\pi(\theta|x)\, d\theta \\ &= \int_0^\infty \frac{1}{\theta}\frac{8.6^4}{\Gamma(4)}\theta^3 e^{-8.6\theta}\, d\theta \\ &= \frac{8.6^4}{6}\underbrace{\int_0^\infty \theta^2 e^{-8.6\theta}\, d\theta}_{\frac{\Gamma(3)}{8.6^3}} \\ &= \frac{8.6^4}{6}\frac{2}{8.6^3} = 2.867 \text{ unidades de tiempo.}\end{aligned}$$

Por otro lado, vea que $\mathbb{E}(\theta|X) = \frac{4}{8.6}$. El estimador *plug-in* correspondería a

$$\frac{1}{\mathbb{E}(\theta|X)} = \frac{8.6}{4} = 2.15.$$

Laboratorio

Lo primero es cargar los paquetes necesarios que usaremos en todo el curso

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

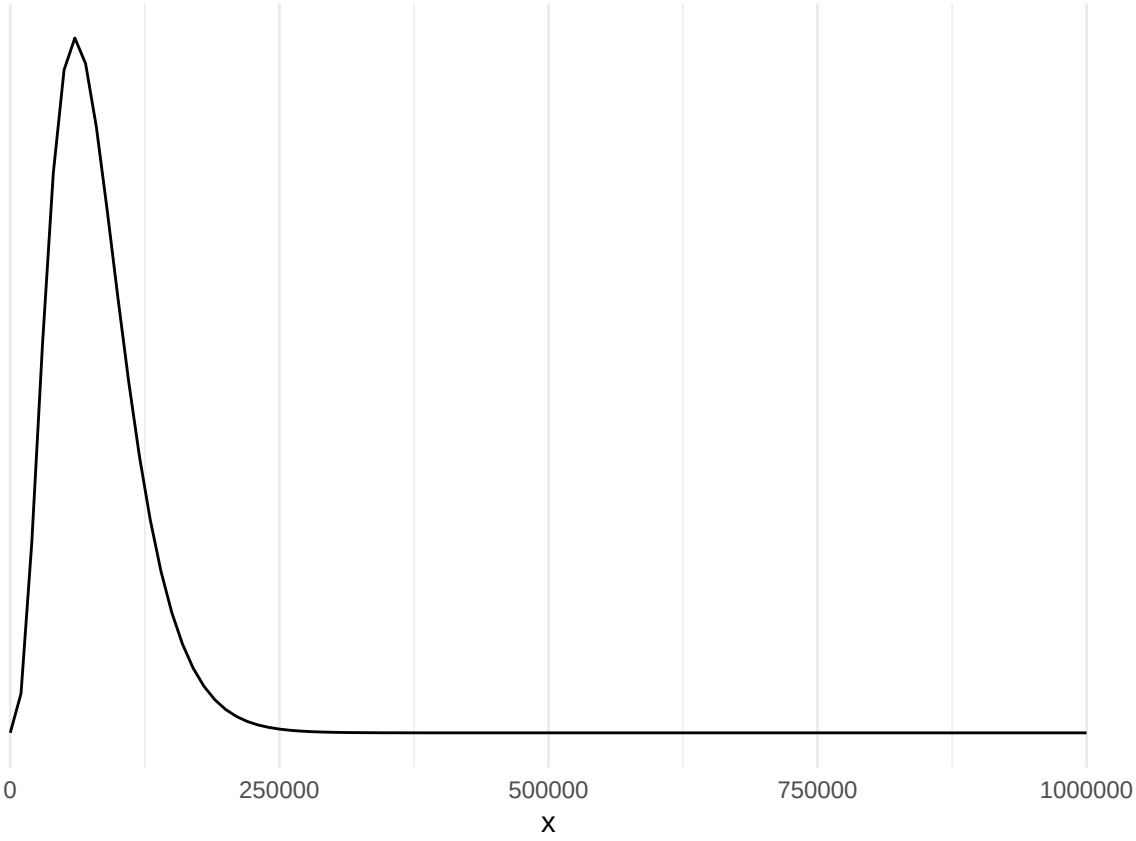
```
## v tibble  3.1.2     v dplyr   1.0.7
## v readr   1.4.0     v stringr 1.4.0
## v purrr   0.3.4     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Distribución previa

En nuestro ejemplo se tenía que $\mathbb{E}[\theta] = 0.0002$ y $\text{Var}(\theta) = 0.001$. Suponiendo que θ es gamma se puede resolver el sistema de ecuaciones obtenemos que $\beta = 20000$ y $\alpha = 4$.

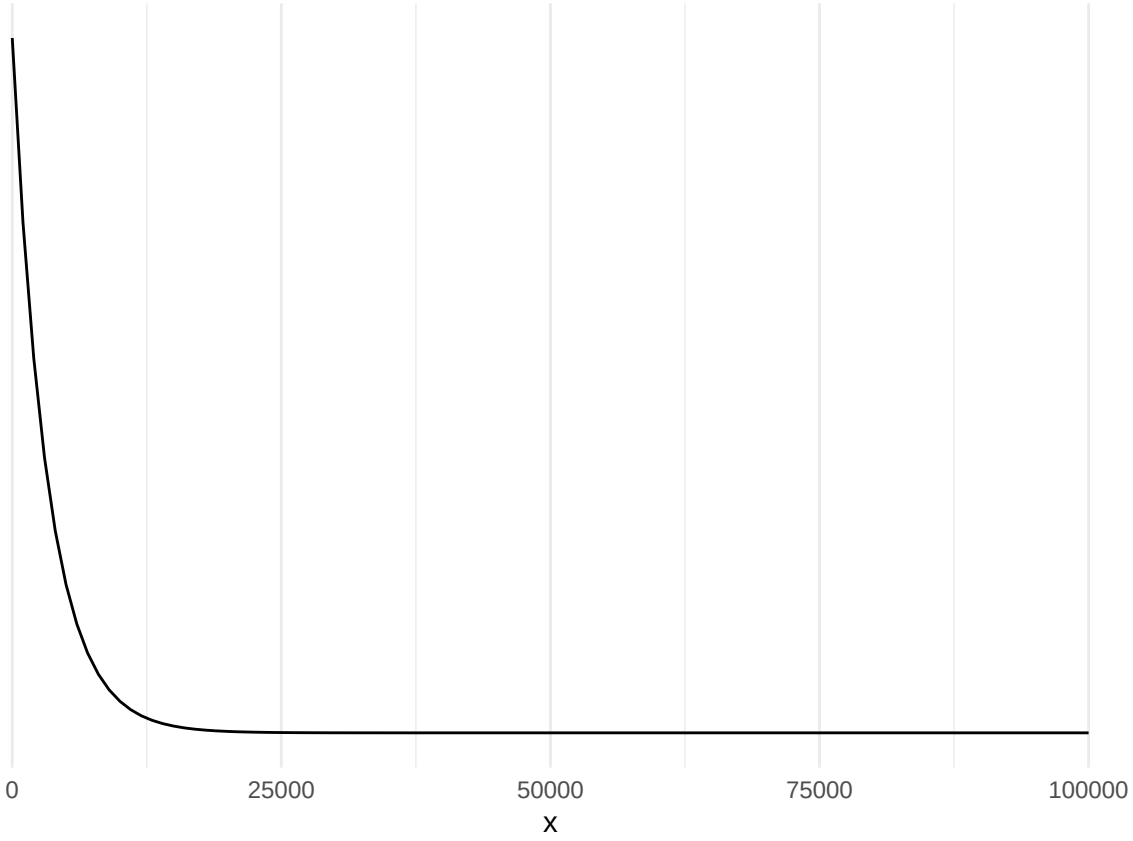
```
alpha_previa <- 4
beta_previa <- 20000

ggplot(data = data.frame(x = c(0, 1e6)), aes(x)) +
  stat_function(fun = dgamma, args = list(shape = alpha_previa, scale = beta_previa)) +
  ylab("") +
  scale_y_continuous(breaks = NULL) +
  theme_minimal()
```



Distribución conjunta

Asumiendo que tenemos algunos datos $X_1, ..., X_n$, asumimos que estos son exponencial recordando que $\mathbb{E}[X] = 1/\theta$, entonces una aproximación de esta densidad es

```
x <- c(2911, 3403, 3237, 3509, 3118)

theta <- 1 / mean(x)

ggplot(data = data.frame(x = c(0, 1e5)), aes(x)) +
  stat_function(fun = dexp, args = list(rate = theta)) +
  ylab("") +
  scale_y_continuous(breaks = NULL) +
  theme_minimal()
```



Distribución posterior

Según los contenidos del curso, se puede estimar los parámetros de la densidad posterior de la forma

```
(y <- sum(x))
```

```
## [1] 16178
```

```
(n <- length(x))
```

```
## [1] 5
```

```
(alpha_posterior <- n + alpha_previa)
```

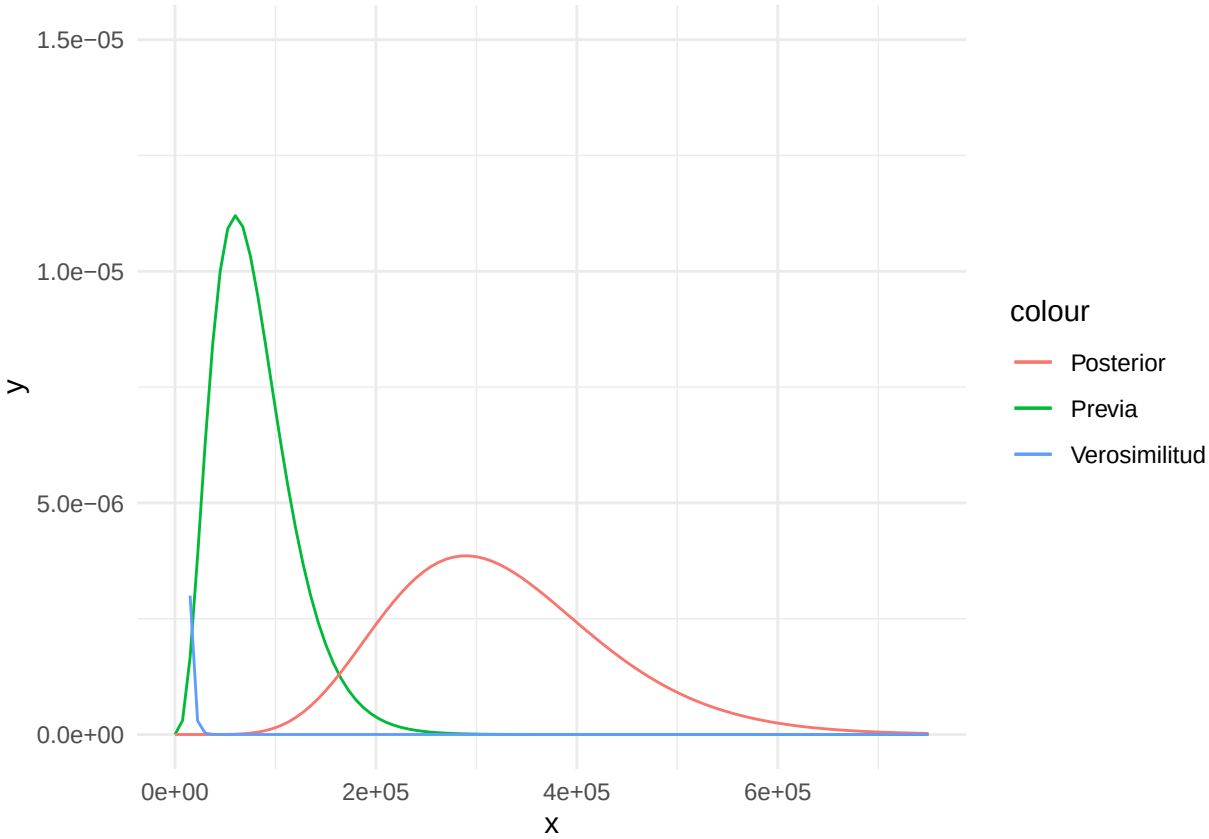
```
## [1] 9
```

```
(beta_posterior <- beta_previa + y)
```

```
## [1] 36178
```

```r
ggplot(data = data.frame(x = c(0, 7.5e5)), aes(x)) +
  stat_function(
    fun = dgamma,
    args = list(shape = alpha_previa, scale = beta_previa),
    aes(color = "Previa")
  ) +
  stat_function(
    fun = dgamma,
    args = list(shape = alpha_posterior, scale = beta_posterior),
    aes(color = "Posterior")
  ) +
  stat_function(
    fun = dexp,
    args = list(rate = theta),
    aes(color = "Verosimilitud")
  ) +
  ylim(0, 1.5e-5) +
  theme_minimal()
```

```
## Warning: Removed 2 row(s) containing missing values (geom_path).
```



Agregando nuevos datos

Si tenemos un 6to dato, y queremos ver cual es su distribución posterior. Lo primero es estimar la densidad posterior de este 6to dato, pero asumiendo que la previa es la densidad que obtuvimos en el caso anterior.

Suponga que $X_6 = 3000$

```
(alpha_previa <- alpha_posterior)
```

```
## [1] 9
```

```
(beta_previa <- beta_posterior)
```

```
## [1] 36178
```

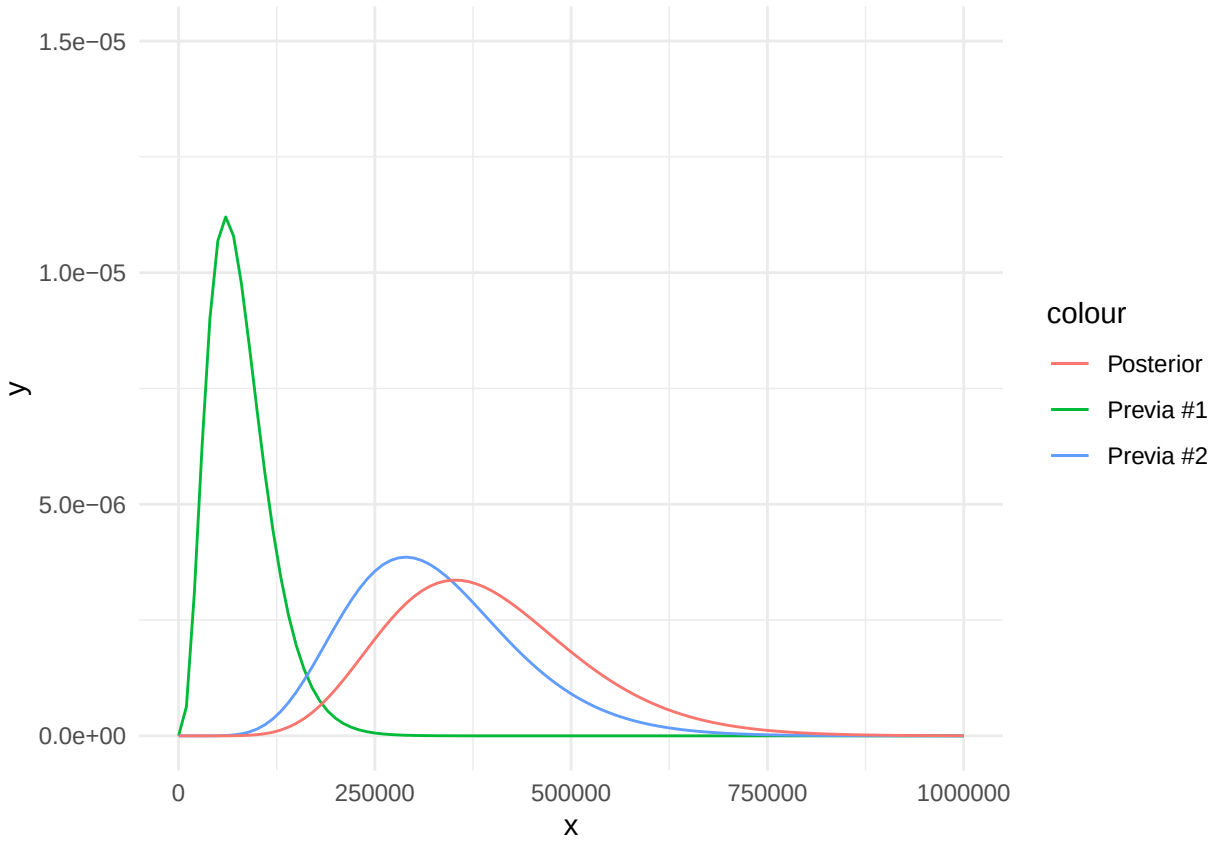
```
(alpha_posterior <- alpha_previa + 1)
```

```
## [1] 10
```

```
(beta_posterior <- beta_previa + 3000)
```

```
## [1] 39178
```

```
ggplot(data = data.frame(x = c(0, 1e6)), aes(x)) +
  stat_function(
    fun = dgamma,
    args = list(shape = 4, scale = 20000),
    aes(color = "Previa #1")
  ) +
  stat_function(
    fun = dgamma,
    args = list(shape = alpha_previa, scale = beta_previa),
    aes(color = "Previa #2")
  ) +
  stat_function(
    fun = dgamma,
    args = list(shape = alpha_posterior, scale = beta_posterior),
    aes(color = "Posterior")
  ) +
  ylim(0, 1.5e-5) +
  theme_minimal()
```



Familias conjugadas normales

Si tenemos pocos datos, la información previa es la que "prevalece".

```
x <- rnorm(n = 3, mean = 10, sd = 1)

(mu <- mean(x))
```

```
## [1] 10.36872
```

```
(sigma <- sd(x))
```

```
## [1] 1.681538
```

```
(n <- length(x))
```

```
## [1] 3
```

```
(mu_previa <- 0)
```

```
## [1] 0
```

```r
(sigma_previa <- 1)
```

```
## [1] 1
```

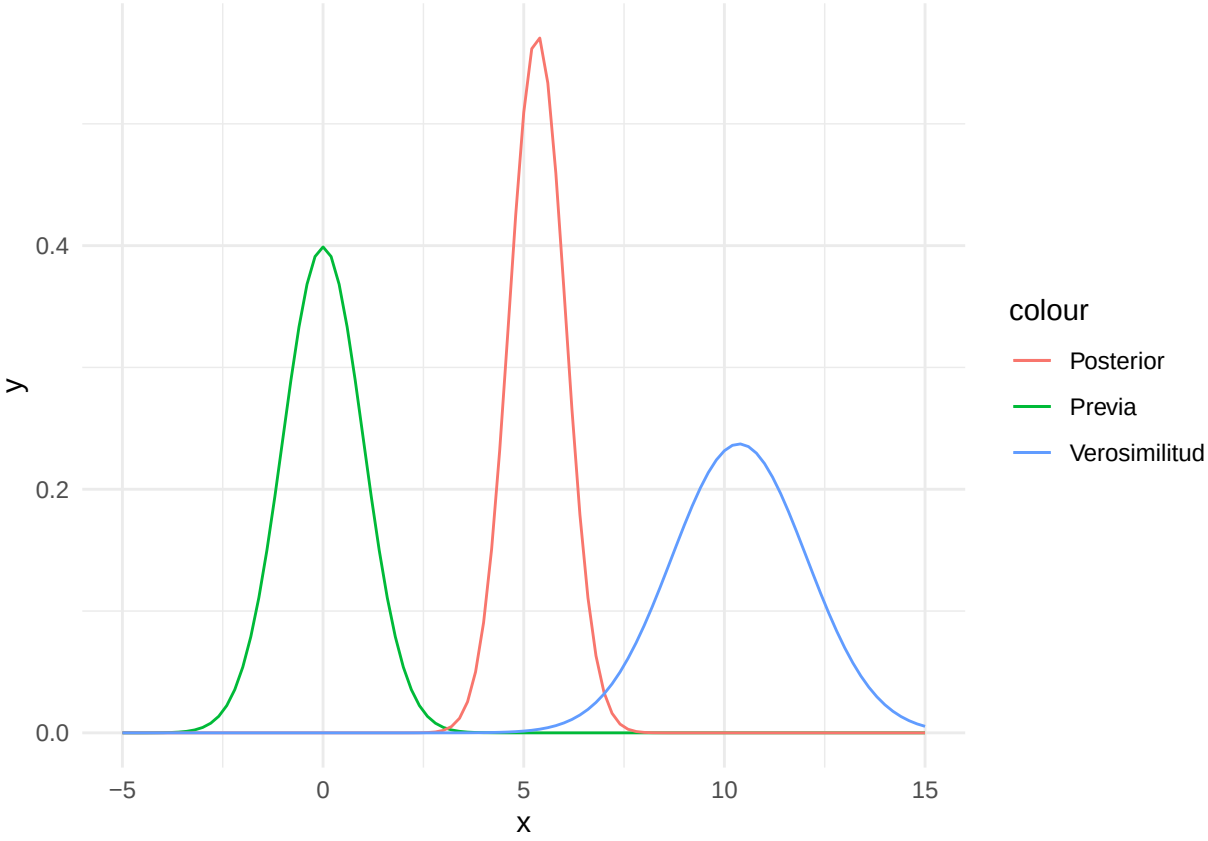
```r
(mu_posterior <- ((sigma^2) / (sigma^2 + n * sigma_previa^2)) * mu_previa + ((n * sigma_previa^2) / (sig
```

```
## [1] 5.337759
```

```r
(sigma2_posterior <- (sigma^2 * sigma_previa^2) / (sigma^2 + n * sigma_previa^2))
```

```
## [1] 0.4852057
```

```r
ggplot(data = data.frame(x = c(-5, 15)), aes(x)) +
  stat_function(
    fun = dnorm,
    args = list(mean = mu_previa, sd = sigma_previa),
    aes(color = "Previa")
  ) +
  stat_function(
    fun = dnorm,
    args = list(mean = mu_posterior, sd = sqrt(sigma2_posterior)),
    aes(color = "Posterior")
  ) +
  stat_function(
    fun = dnorm,
    args = list(mean = mu, sd = sigma),
    aes(color = "Verosimilitud")
  ) +
  theme_minimal()
```



Con más datos, la distribución se ajusta a esto y le quita importancia a la información previa.

```
x <- rnorm(n = 100, mean = 10, sd = 1)

(mu <- mean(x))
```

```
## [1] 10.04396
```

```
(sigma <- sd(x))
```

```
## [1] 1.109477
```

```
(n <- length(x))
```

```
## [1] 100
```

```
(mu_previa <- 0)
```

```
## [1] 0
```

```
(sigma_previa <- 1)
```

```
## [1] 1
```

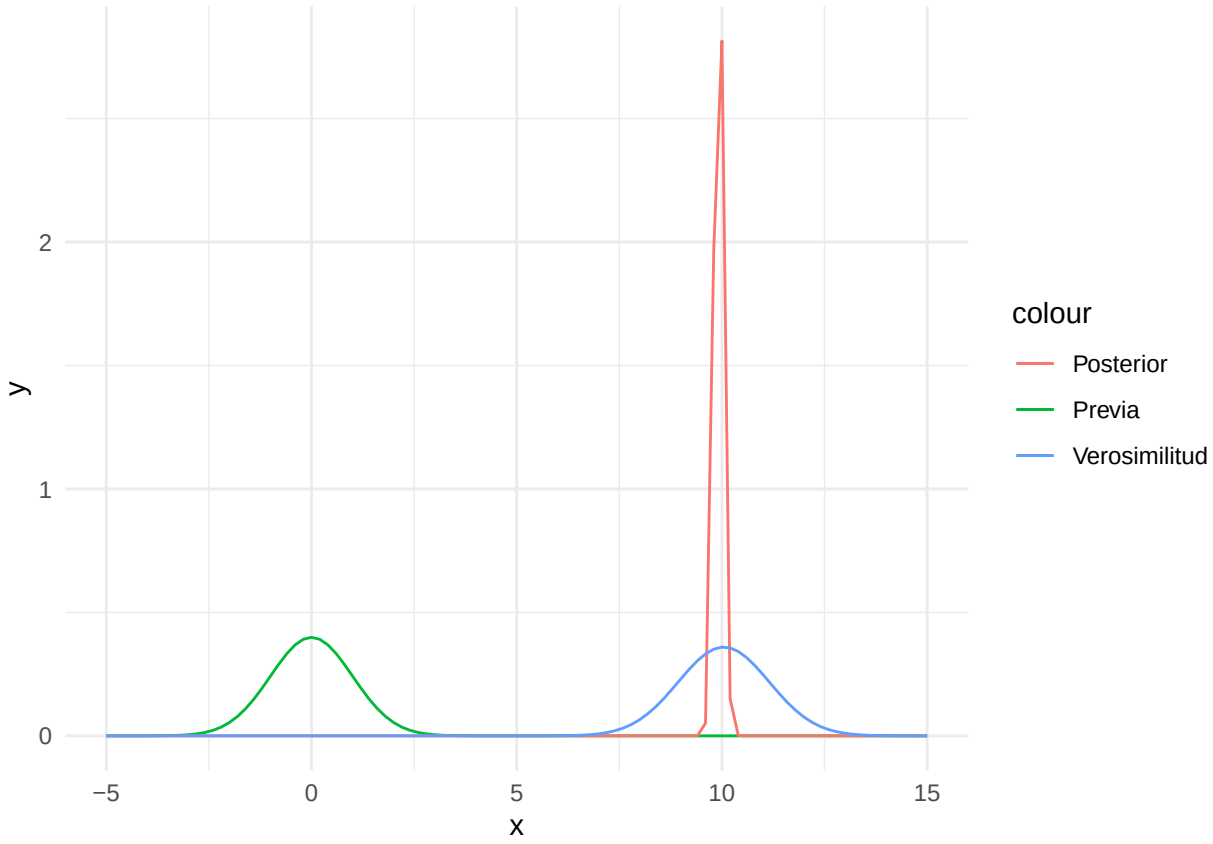
```r
(mu_posterior <- ((sigma^2) /
  (sigma^2 + n * sigma_previa^2)) * mu_previa +
  ((n * sigma_previa^2) /
    (sigma^2 + n * sigma_previa^2)) * mu)
```

```
## [1] 9.921828
```

```r
(sigma2_posterior <- (sigma^2 * sigma_previa^2) / (sigma^2 + n * sigma_previa^2))
```

```
## [1] 0.01215972
```

```r
ggplot(data = data.frame(x = c(-5, 15)), aes(x)) +
  stat_function(
    fun = dnorm,
    args = list(mean = mu_previa, sd = sigma_previa),
    aes(color = "Previa")
  ) +
  stat_function(
    fun = dnorm,
    args = list(mean = mu_posterior, sd = sqrt(sigma2_posterior)),
    aes(color = "Posterior")
  ) +
  stat_function(
    fun = dnorm,
    args = list(mean = mu, sd = sigma),
    aes(color = "Verosimilitud")
  ) +
  theme_minimal()
```



Si los datos por si solo son muy variable, la posterior tiende a parecerse a la distribución previa en lugar que a la verosimilitud.

```
x <- rnorm(n = 10, mean = 10, sd = 5)

(mu <- mean(x))
```

```
## [1] 9.97915
```

```
(sigma <- sd(x))
```

```
## [1] 4.743485
```

```
(n <- length(x))
```

```
## [1] 10
```

```
(mu_previa <- 0)
```

```
## [1] 0
```

```
(sigma_previa <- 1)
```

```
## [1] 1
```

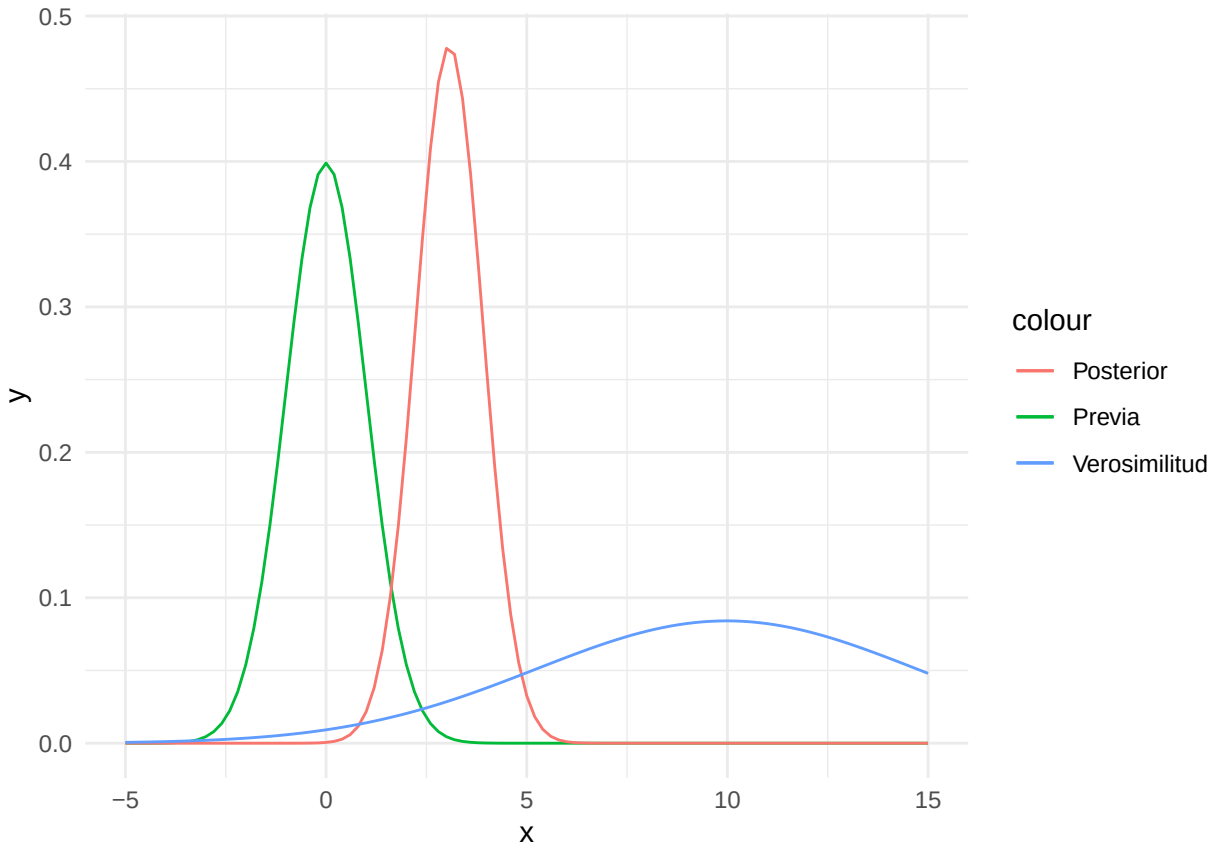
```r
mu_posterior <- ((sigma^2) /
  (sigma^2 + n * sigma_previa^2)) * mu_previa +
  ((n * sigma_previa^2) /
    (sigma^2 + n * sigma_previa^2)) * mu

mu_posterior
```

```
## [1] 3.070446
```

```r
(sigma2_posterior <- (sigma^2 * sigma_previa^2) / (sigma^2 + n * sigma_previa^2))
```

```
## [1] 0.6923138
```

```r
ggplot(data = data.frame(x = c(-5, 15)), aes(x)) +
  stat_function(
    fun = dnorm,
    args = list(mean = mu_previa, sd = sigma_previa),
    aes(color = "Previa")
  ) +
  stat_function(
    fun = dnorm,
    args = list(mean = mu_posterior, sd = sqrt(sigma2_posterior)),
    aes(color = "Posterior")
  ) +
  stat_function(
    fun = dnorm,
    args = list(mean = mu, sd = sigma),
    aes(color = "Verosimilitud")
  ) +
  theme_minimal()
```



Funciones de pérdida

Lo más importante acá es que dependiendo de la función de pérdida podemos construir una estimador para θ. En el caso de los componentes electrónicos recordemos que la posterior nos daba

```
alpha <- 9
beta <- 36178
```

- **Pérdida cuadrática:** Recoremos que la media de una gamma es α/β entonces

```
(theta <- alpha / beta)
```

```
## [1] 0.00024877
```

Y por lo tanto el tiempo promedio del componente electrónico es $1/\theta$=4019.7777778.

- **Pérdidad absoluta:** La distribución Gamma no tiene una forma cerrada para la mediana, por que se puede aproximar así,

```
m <- rgamma(n = 1000, scale = beta, shape = alpha)
(theta <- median(m))
```

```
## [1] 312333.5
```

Y por lo tanto el tiempo promedio del componente electrónico es $1/\theta$=3.201706 $\times 10^{-6}$.

OJO: En este caso la pérdida cuadrática ajusta mejor ya que la distribución que la pérdida absoluta ya que la distribución NO es simétrica. En el caso simétrico los resultados serían muy similares.

Caso concreto

Suponga que se que quiere averiguar si los estudiantes de cierto colegio duermen más de 8 horas o menos de 8 horas.

Para esto primero cargaremos el siguiente paquete,

```
library(LearnBayes)
```

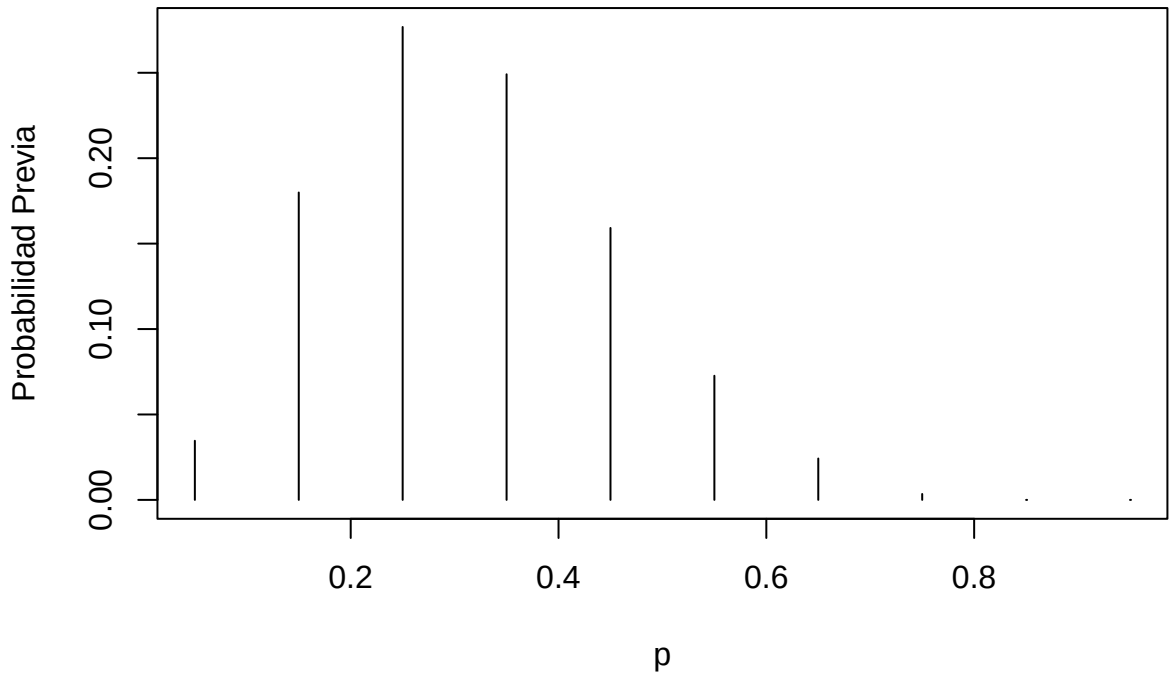
Suponga que se hace una encuesta a 27 estudiantes y se encuentra que 11 dicen que duermen más de 8 horas diarias y el resto no. Nuestro objetivo es encontrar inferencias sobre la proporción p de estudiantes que duermen al menos 8 horas diarias. El modelo más adecuado es

$$f(x|p) \propto p^s(1-p)^f$$

donde s es la cantidad de estudiantes que duermen más de 8 horas y f los que duermen menos de 8 horas.

Una primera aproximación para la previa es usar una distribución discreta. En este caso, el investigador asigna una probabilidad a cierta cantidad de horas de sueño, según su experiencia. Así, por ejemplo:

```
p <- seq(0.05, 0.95, by = 0.1)
prior <- c(1, 5.2, 8, 7.2, 4.6, 2.1, 0.7, 0.1, 0, 0)
prior <- prior / sum(prior)
plot(p, prior, type = "h", ylab = "Probabilidad Previa")
```



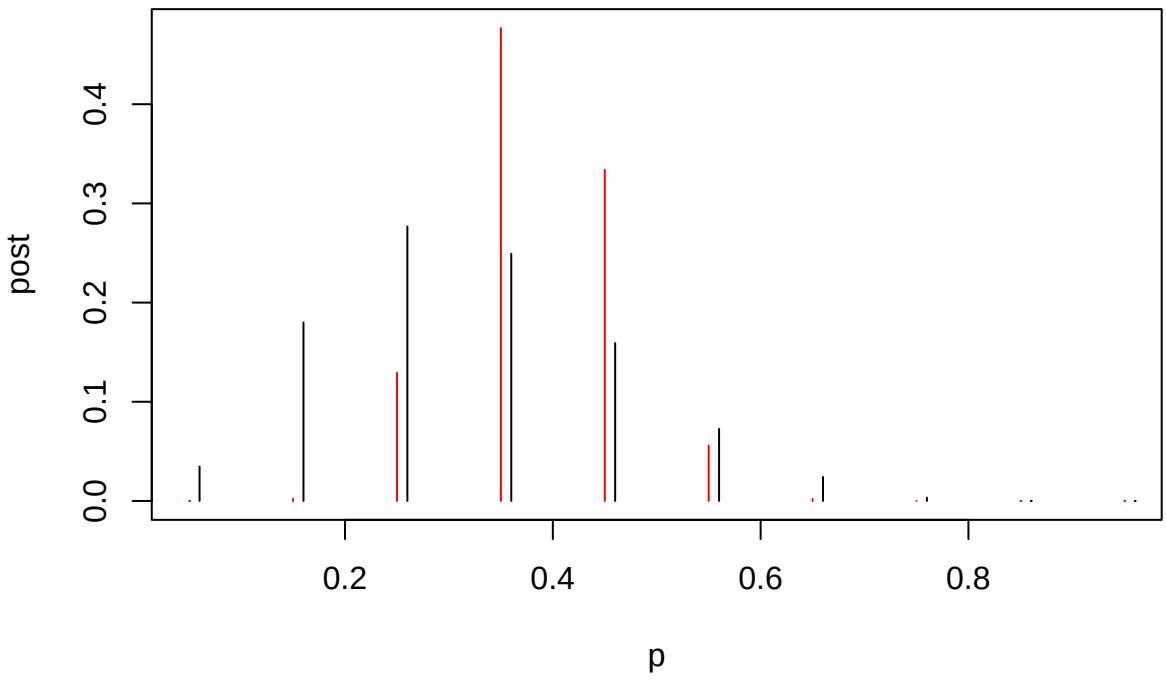
El paquete `LearnBayes` tiene la función `pdisc` que estima la distribución posterior para una previa discreta binomial. Recuerde que el valor 11 representa la cantidad de estudiantes con más de 8 horas de sueño y 16 lo que no duermen esa cantidad.

```
data <- c(11, 16)
post <- pdisc(p, prior, data)
round(cbind(p, prior, post), 2)
```

```
##          p prior post
##  [1,] 0.05  0.03 0.00
##  [2,] 0.15  0.18 0.00
##  [3,] 0.25  0.28 0.13
##  [4,] 0.35  0.25 0.48
##  [5,] 0.45  0.16 0.33
##  [6,] 0.55  0.07 0.06
##  [7,] 0.65  0.02 0.00
##  [8,] 0.75  0.00 0.00
##  [9,] 0.85  0.00 0.00
## [10,] 0.95  0.00 0.00
```

Y podemos ver la diferencia entre la previa (negro) y la posterior (roja),

```
plot(p, post, type = "h", col = "red")
lines(p + 0.01, prior, type = "h")
```



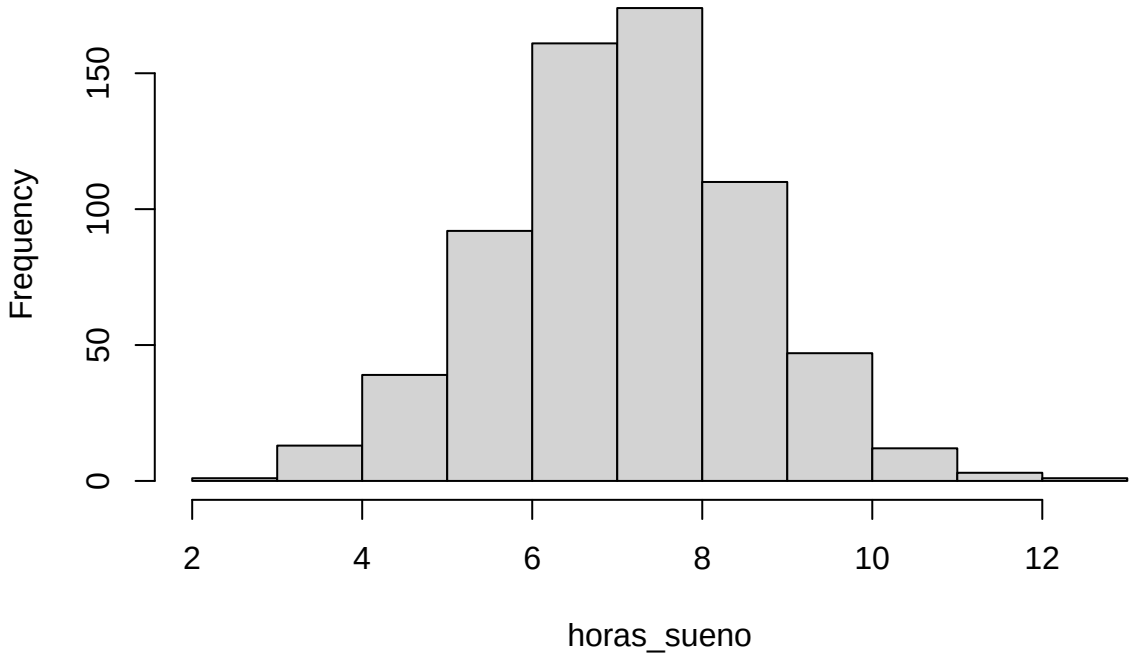
¿Qué se puede deducir de estos resultados?

Ejercicio: Suponga que se tiene la base de datos `studentdata`. Realice los cálculos anteriores con esos datos,

```
data("studentdata")
horas_sueno <- studentdata$WakeUp - studentdata$ToSleep
horas_sueno <- na.omit(horas_sueno)
summary(horas_sueno)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.500   6.500   7.500   7.385   8.500  12.500
```

```
hist(horas_sueno, main = "")
```



Ahora supongamos que se tiene quiere ajustar una previa continua a este modelo. Para esto usaremos una distribución Beta con parámetros α y β, de la forma

$$pi(p|\alpha,\beta) \propto p^{1-\alpha}(1-p)^{1-\beta}.$$

El ajuste de los paramétros de la Beta depende mucho de la información previa que se tenga del modelo. Una forma fácil de estimarlo es a través de cuantiles con los cuales se puede reescribir estos parámetros. Para una explicación detallada revisar https://stats.stackexchange.com/a/237849

En particular, suponga que se cree que el 50% de las observaciones la proporción será menor que 0.3 y que el 90% será menor que 0.5.

Para esto ajustaremos los siguientes parámetros

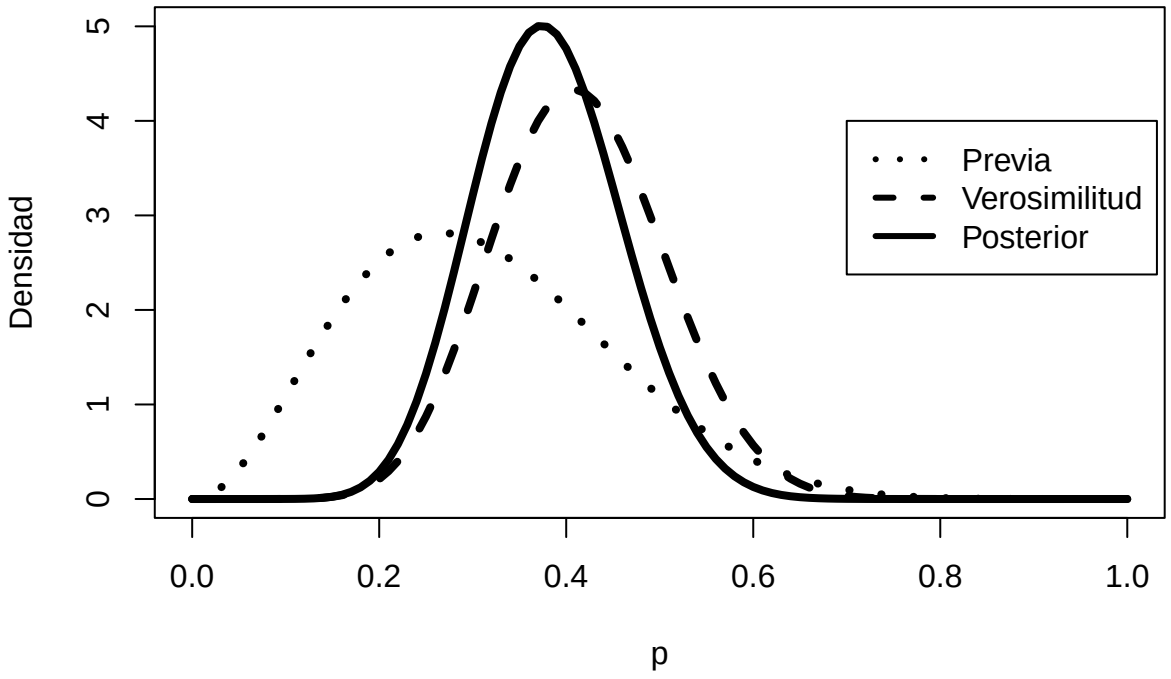
```
quantile2 <- list(p = .9, x = .5)
quantile1 <- list(p = .5, x = .3)
(ab <- beta.select(quantile1, quantile2))
```

```
## [1] 3.26 7.19
```

```r
a <- ab[1]
b <- ab[2]
s <- 11
f <- 16
```

En este caso se obtendra la distribución posterior Beta con paramétros $\alpha + s$ y $\beta + f$,

```r
curve(dbeta(x, a + s, b + f),
  from = 0, to = 1,
  xlab = "p", ylab = "Densidad", lty = 1, lwd = 4
)
curve(dbeta(x, s + 1, f + 1), add = TRUE, lty = 2, lwd = 4)
curve(dbeta(x, a, b), add = TRUE, lty = 3, lwd = 4)
legend(.7, 4, c("Previa", "Verosimilitud", "Posterior"),
  lty = c(3, 2, 1), lwd = c(3, 3, 3)
)
```



Práctica

Pueden iniciar con estos ejercicios https://mediacionvirtual.ucr.ac.cr/pluginfile.php/521945/mod_resource/content/0/Ejercicios%201.pdf

Además de los que están al final del handout de Mauricio.

---

class: center, middle

¿Qué discutimos hoy?

Distribuciones previas o a previa.

¿Qué nos falta para terminar el curso?

Estadística Bayesiana: inferencia (estimación puntual, intervalos de credibilidad y factores de Bayes).