Universidad Rey Juan Carlos

Máster Universitario en Ingeniería de Sistemas de Decisión

Caso Práctico III: Inferencia Estadística Bayesiana

Autor:

Elisa Cascudo

Asignatura:

Modelización y Tratamiento de la Incertidumbre

13 Octubre 2025

Índice general

Índice de figuras

Tarea III

Se ha realizado un estudio sobre los efectos de la exposición a niveles moderados de plomo (encontrado en determinados alimentos) en el desarrollo cognitivo a largo plazo de los niños. Los investigadores analizaron el contenido en plomo en los dientes de leche de los niños, una vez que se les habían caído. De los 29 niños estudiados que tenían un contenido en plomo superior a 22.22 ppm (partes por millón), 22 terminaron la Educación Secundaria y otros 7 no la terminaron. Suponemos una distribución a priori $p \sim Be(1,1)$ *para la proporción de alumnos que terminan la Educación Secundaria.*

1. Escribe la función de verosimilitud, $L(p) = P(X = x \mid p)$*, y la distribución a posteriori,* $f(p \mid x)$ *(salvo constantes multiplicativas).*

Sabemos que la distribución a priori se trata de una distribución Beta

$$\pi(p) = \mathrm{Be}(1,1) = 1, \quad \text{para } p \in [0,1].$$

Dado que se observa un número de éxitos $x = 22$ en $n = 29$ ensayos independientes, y cada niño tiene una probabilidad p de terminar la Educación Secundaria, el modelo adecuado es una distribución binomial:

$$X \mid p \sim \mathrm{Binomial}(n = 29, p)$$

por lo que la función de verosimilitud es, retirando las constantes multiplicativas:

$$L(p) = P(X = x \mid p) = \binom{29}{22} p^{22}(1-p)^7 \propto p^{22}(1-p)^7$$

Según la regla de Bayes, la distribución a posteriori es proporcional al producto de la verosimilitud por la distribución a priori:

$$f(p \mid x) \propto L(p) \cdot \pi(p) = p^{22}(1-p)^7 \cdot 1 = p^{22}(1-p)^7.$$

Por tanto, salvo una constante de normalización, que también consideraremos una constante multiplicativa, la distribución a posteriori es de la forma:

$$f(p \mid x) \propto p^{22}(1-p)^7$$

Notamos que esta es la forma de una distribución Beta:

$$f(p \mid x) = \mathrm{Be}(p; \alpha = 23, \beta = 8),$$

ya que:

$$\mathrm{Beta}(\alpha, \beta) \propto p^{\alpha-1}(1-p)^{\beta-1}.$$

Por lo que la distribución a posteriori es:

$$f(p \mid x) \sim \mathrm{Beta}(23, 8).$$

2. Obtén una muestra aleatoria de tamaño 1000 de la distribución a posteriori de p, y representa su densidad empírica, comparándola en la misma gráfica con la densidad a priori y a posteriori de p.

Este ejercicio se ha resuelto en R empleando el siguiente código:

```
set.seed(123)
x <- rbeta(1000, shape1 = 23, shape2 = 8)

plot(density(x), ylab = "Densidad",
     main = "Representacion densidades simulada, a priori y a posteriori",
     col = "black")

curve(dbeta(x, 1, 1), from = 0, to = 1, col = "blue", add = TRUE)
curve(dbeta(x, 23, 8), from = 0, to = 1, col = "red", add = TRUE)
```

Obteniendo la siguiente representación gráfica:

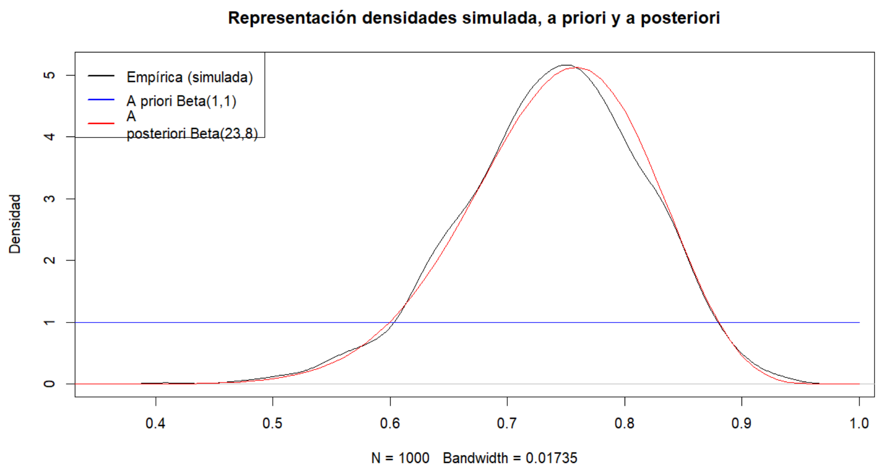

Figura 1: Representación gráfica de la densidad empírica, densidad a priori y a posteriori. Elaboración propia.

3. Calcular la media a posteriori y la desviación típica a posteriori.

Para calcular la media y la desviación típica de la distribución a posteriori se han utilizado las definiciones de la esperanza y la varianza para una distribución continua.

Dadas una función de densidad $f(p)$ y una variable continua $p \in [0, 1]$, la esperanza y la varianza se definen como:

$$\mathbb{E}[p] = \int_0^1 p \cdot f(p)\, dp$$

$$\mathrm{Var}(p) = \int_0^1 (p - \mathbb{E}[p])^2 \cdot f(p)\, dp$$

$$\text{Desviación típica} = \sqrt{\mathrm{Var}(p)}$$

Los cálculos se han realizado en R. A continuación se muestra el código empleado:

```r
results <- binogcp(22, 29, density = "beta", params = c(1, 1), n.pi = 1e4)

p1 <- results$pi
p2 <- results$posterior

dens <- p1 * p2
post.mean <- sintegral(p1, dens)$value

dens <- (p1 - post.mean)^2 * p2
post.var <- sintegral(p1, dens)$value
post.sd <- sqrt(post.var)
```

Los resultados obtenidos han sido:

- Media a posteriori: $\hat{p} = 0{,}7419$
- Desviación típica a posteriori: $\hat{\sigma}_p = 0{,}0774$

4. Calcular un intervalo de probabilidad al 90 % para p.

Para determinar un intervalo de probabilidad del 90 % para el parámetro p, buscamos un intervalo $[a, b]$ tal que:

$$P(p \in [a, b] \mid \text{datos}) = 0{,}90$$

En el caso de una distribución continua posterior Beta que hemos simulado, una forma de construir este intervalo es eliminando un 5 % de probabilidad de cada cola de la distribución, lo que corresponde a los percentiles 5 % y 95 %.

El código R utilizado para este cálculo es el siguiente:

```
cdf <- sintegral(p1, p2, n.pts = length(p1))$cdf

lcb <- cdf$x[with(cdf, which.max(x[y <= 0.05]))]
ucb <- cdf$x[with(cdf, which.max(x[y <= 0.95]))]
```

Los valores obtenidos para los límites del intervalo han sido:

- Límite inferior del 90 %: lcb = 0,606
- Límite superior del 90 %: ucb = 0,859

Por tanto, el intervalo de probabilidad al 90 % para p es:

$$p \in [0{,}606,\ 0{,}859]$$

5. Contrastar la hipótesis $H_0 : p \leq 0{,}4$ frente a $H_1 : p > 0{,}4$.

Para contrastar las hipótesis bayesianas:

$$H_0 : p \leq 0{,}4 \quad \text{frente a} \quad H_1 : p > 0{,}4,$$

evaluamos directamente la probabilidad a posteriori de cada una de ellas utilizando la distribución posterior simulada de p. Dado que la distribución posterior es continua, estas probabilidades se obtienen mediante la integración numérica de la densidad posterior $f(p \mid \text{datos})$ en los intervalos correspondientes:

$$P(H_0 \mid \text{datos}) = P(p \leq 0{,}4 \mid \text{datos}) = \int_0^{0{,}4} f(p)\, dp,$$

$$P(H_1 \mid \text{datos}) = P(p > 0{,}4 \mid \text{datos}) = \int_{0{,}4}^{1} f(p)\, dp.$$

Estas probabilidades se han calculado en R utilizando dos métodos distintos:

1. A partir de la función de distribución acumulada (CDF) aproximada.
2. Mediante integración numérica directa usando la función `sintegral()`.

```
#CDF
ph0 <- cdf$y[with(cdf, which.max(y[x <= 0.4]))]
ph1 <- 1 - ph0

#Sintegral
p0 <- 0.4
x0 <- p1[p1 <= p0]
```

```
x1 <- p1[p1 > p0]
P0 <- sintegral(x0, p2[p1 <= p0])$value
P1 <- sintegral(x1, p2[p1 > p0])$value
```

Los resultados obtenidos con ambos métodos:

- $P(H_0 \mid \text{datos}) = 0{,}00005$
- $P(H_1 \mid \text{datos}) = 0{,}99995$

Dado que la probabilidad a posteriori de H_1 es prácticamente 1, existe una evidencia abrumadora a favor de la hipótesis alternativa $H_1 : p > 0{,}4$, y en contra de la hipótesis nula H_0. Por tanto, los datos apoyan claramente que la proporción p de niños que terminan la educación secundaria es mayor que 0,4.

6. En una siguiente fase del estudio, se reciben datos de otros 10 niños que han presentado un contenido en plomo de más de 22.22 ppm. Calcular la probabilidad predictiva de que al menos 9 de ellos terminen la Educación Secundaria.

La función de densidad de probabilidad de la distribución Beta-Binomial es:

$$P(X = k \mid n, \alpha, \beta) = \binom{n}{k} \cdot \frac{B(k+\alpha,\ n-k+\beta)}{B(\alpha, \beta)}$$

donde:

- k es el número de éxitos observados,
- n es el número de ensayos (aquí, niños adicionales),
- $\alpha = 23,\ \beta = 8$ son los parámetros de la distribución **a posteriori** de p,
- $B(\cdot, \cdot)$ es la función beta.

Queremos calcular:

$$P(X \geq 9) = P(X = 9) + P(X = 10)$$

Para ello, se ha empleado en siguiente código en R:

```
predict <- function(k, n, a, b){
  choose(n, k) * beta(k + a, n - k + b) / beta(a, b)
}

probmas9 <- predict(9, 10, 23, 8) + predict(10, 10, 23, 8)
```

El resultado obtenido ha sido:

$$P(X \geq 9) = 0{,}2664$$

Por lo que la probabilidad de que al menos 9 de los próximos 10 niños aprueben la Educación Secundaria es del 26.64 %.

Tarea IV

Supongamos que estamos interesados en estimar el total anual de precipitaciones en forma de nieve (en cm), μ, en un pueblo del norte de la Península. Suponemos que los totales anuales recogidos $y_1, \ldots, y_n$ provienen de una población que sigue una distribución normal de media μ y varianza conocida $\sigma^2 = 10^2$ cm^2.

1. Antes de recoger los datos, suponemos que tenemos las siguientes creencias a priori acerca de la probabilidad de que la media de las precipitaciones totales anuales, μ, tome los siguientes valores:

μ	20	30	40	50	60	70
$g(\mu)$	0.1	0.15	0.25	0.25	0.15	0.1

Representa gráficamente esta distribución a priori.

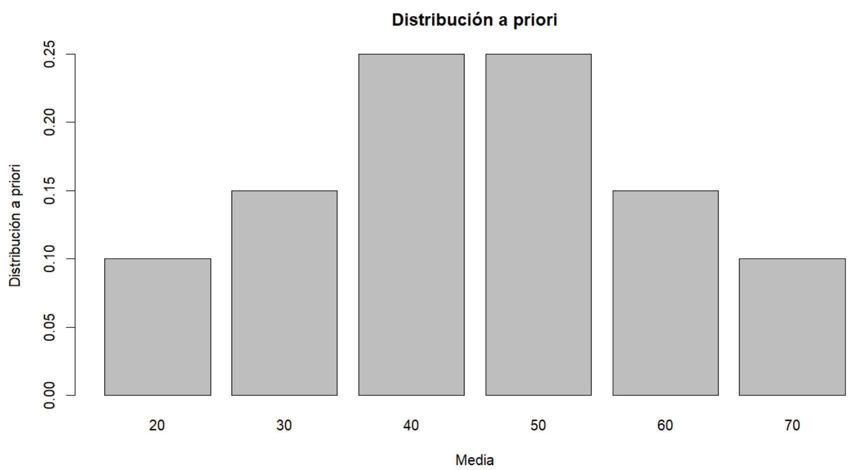

Figura 2: Representación gráfica de la distribución a priori normal discreta. Elaboración propia.

2. Se tiene registro de las precipitaciones de nieve en cm, y, de los últimos 12 años:

$$y = 38.6, 42.4, 57.5, 40.5, 51.7, 67.1, 33.4, 60.9, 64.1, 40.1, 40.7, 6.4$$

La función de verosimilitud es para el modelo normal-normal:

$$\mathcal{L}(\mu) = \left(\frac{1}{\sqrt{2\pi\sigma}}\right)^n \prod_{i=1}^{n} \exp\left[-\frac{1}{2}\left(\frac{y_i - \mu}{\sigma}\right)^2\right].$$

Calcula el valor de la verosimilitud para todos los posibles valores de μ reflejados en la distribución a priori.

Para calcular la verosimilitud para cada valor de μ propuesto en la distribución a priori, se ha empleado el siguiente código en R:

```
verosimilitud <- function(mu){
  prod(dnorm(y, mean=mu, sd=sigma.y))
}

verosimilitud.mu <- sapply(mu, verosimilitud)
vero.data <- data.frame(mu = mu, Verosimilitud = verosimilitud.mu)
```

Los resultados obtenidos para cada valor de μ son los siguientes:

μ	Verosimilitud
20	$8{,}91 \times 10^{-41}$
30	$3{,}32 \times 10^{-30}$
40	$7{,}58 \times 10^{-25}$
50	$1{,}06 \times 10^{-24}$
60	$9{,}19 \times 10^{-30}$
70	$4{,}88 \times 10^{-40}$

3. Calcular la distribución a posteriori para μ.

La distribución a posteriori para μ se calcula utilizando la regla de Bayes. Dado que la distribución a priori discreta de μ está dada por $g(\mu)$ y la función de verosimilitud por $\mathcal{L}(\mu)$, la distribución a posteriori, salvo una constante de normalización, es proporcional al producto de ambas:

$$f(\mu \mid y) \propto \mathcal{L}(\mu) \times g(\mu).$$

Para obtener mayor coherencia en la distribución a posteriori propiamente dicha, se normaliza este producto dividiendo entre la suma total.

El código en R utilizado para calcular la distribución a posteriori es el siguiente:

```
dist.unnorm <- verosimilitud.mu * prob.mu
dist.posterior <- dist.unnorm / sum(dist.unnorm)
dist.postdata <- data.frame(Media = mu, Distribucion = dist.posterior)
```

Los resultados obtenidos para la distribución a posteriori son:

Media μ	Distribución a posteriori
20	$1{,}95 \times 10^{-17}$
30	$1{,}09 \times 10^{-6}$
40	$4{,}16 \times 10^{-1}$
50	$5{,}84 \times 10^{-1}$
60	$3{,}03 \times 10^{-6}$
70	$1{,}07 \times 10^{-16}$

4. Calcula un intervalo de probabilidad al 80 % para μ.

Para calcular un intervalo de probabilidad al 80 % para μ, se busca el rango de valores de μ que contienen el 80 % de la probabilidad total según la distribución a posteriori calculada previamente.

El código en R utilizado para llegar al resultado ha sido:

```
prob = 0.8
mat <- as.matrix(dist.postdata)
discint(mat, prob)
```

El resultado obtenido es el intervalo $[40, 50]$, es decir, el 80 % de la probabilidad a posteriori para μ se encuentra entre 40 y 50.

5. Repetir los apartados 3–4 cuando la distribución a priori se toma como $\mu \sim \mathcal{N}(\mu_0 = 60,\ \sigma_0^2 = 5^2)$.

En esta pregunta se repiten los cálculos de los apartados anteriores, pero suponiendo una nueva distribución a priori continua para μ, concretamente una normal:

$$\mu \sim \mathcal{N}(60,\ 5^2)$$

Dado que los datos también siguen una distribución normal con varianza conocida, el modelo es conjugado, y la distribución a posteriori también será normal. Sus parámetros se obtienen mediante las siguientes expresiones analíticas:

$$\sigma^2_{\text{post}} = \left(\frac{n}{\sigma_y^2} + \frac{1}{\sigma_0^2}\right)^{-1} \quad \text{y} \quad \mu_{\text{post}} = \sigma^2_{\text{post}}\left(\frac{n\bar{y}}{\sigma_y^2} + \frac{\mu_0}{\sigma_0^2}\right)$$

Donde:

- $\bar{y}$ es la media muestral de los datos,
- $\sigma_y = 10$, varianza conocida de la distribución de los datos,
- $n = 12$, número de observaciones.

El código en R empleado para estos cálculos es:

```
mu0 = 60
sigma0 = 5
sigma.y = 10
y <- c(38.6, 42.4, 57.5, 40.5, 51.7, 67.1, 33.4, 60.9, 64.1, 40.1, 40.7, 6.4)
n.mu = 1e4
n <- length(y)
y_bar <- mean(y)

results <- normgcp(y, sigma.y, density = "normal", params = c(mu0, sigma0), n.mu)
sigma_post <- sqrt(1 / (n / sigma.y^2 + 1 / sigma0^2))
mu_post <- sigma_post^2 * (n * y_bar / sigma.y^2 + mu0 / sigma0^2)

p1 <- results$mu
p2 <- results$posterior
cdf <- sintegral(p1, p2, n.pts = length(p1))$cdf
lcb <- cdf$x[with(cdf, which.max(x[y <= 0.1]))]
ucb <- cdf$x[with(cdf, which.max(x[y <= 0.9]))]
```

Los resultados para esta nueva distribución normal a priori son:

- Media a posteriori: $\mu_{\text{post}} = 48{,}96$
- Desviación típica a posteriori: $\sigma_{\text{post}} = 2{,}5$
- Intervalo de probabilidad al 80 %: $[45{,}82,\ 52{,}17]$