Pontificia Universidad Católica de Chile
Facultad de Matemática
Departamento de Matemática

# EYP2805

# Métodos Bayesianos

Rubén Soza

Profesora: Valeria Leiva

15 de noviembre de 2017

# Índice

CAPÍTULO 1

# Introducción

## 1.1 Inferencia Bayesiana

De Finnetti muestra que la única forma de representar incertidumbre de manera coherente es a través de una medida de probabilidad, la cual es actualizada a la luz de nuevos datos.

**Teorema 1.1.** *Sea* $\{A_n\}_{n\in\mathbb{N}}$ *una partición de* $\Omega$*. Entonces*

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{n\in\mathbb{N}} P(B|A_n)P(A_n)}$$

De forma práctica, podemos pensar que:

1. $A_i$ : Estados de la naturaleza
2. $P(A_i)$: Opinión de los estados de la naturaleza
3. $B$: Empírico datos.

A partir de los datos se puede reevaluar la opinión de los estados de la naturaleza.

**Ejemplo 1.1.** Consideramos 3 muebles, cada uno con dos cajones que contienen una moneda. El material de dicha moneda sigue la siguiente regla:

1. Mueble 1: Plata-Plata
2. Mueble 2: Plata-Oro
3. Mueble 3: Oro-Oro

Se elige un mueble al azar, y en el cajón, se encontró una moneda de oro. ¿Cuál es la probabilidad de que el otro cajón también tenga una moneda de oro?.

Según lo visto antes, tenemos

1. $A_i$: Mueble $i$, $i = 1, 2, 3$
2. $P(A_i) = \frac{1}{3}$
3. $B$: Moneda de oro en el cajón $j$.

★

El **Paradigma Bayesiano** es realizar inferencias sobre cantidades que no conocemos dado lo que hemos observado. Luego, nuestro objetivo será realizar inferencias a partir de datos, utilizando modelos de probabilidad para cantidades que observamos y para cantidades que deseamos aprender.

El proceso de análisis puede ser idealizado en los siguientes pasos:

1. Establecer un **modelo de probabilidad completo**: distribución conjunta de todas las cantidades observables y no observables.
2. Condicionar en los datos observados: Inferir a partir de la distribución a posteriori de las cantidades no observadas dados los datos.
3. Evaluar el ajuste del modelo planteado en el primer paso.

La idea en general, entonces, es tener una creencia inicial, o probabilidad apriori $P(\theta)$ y la distribución de los datos dado $\theta$, que viene dada por la función de verosimilitud $P(Y|\theta)$. Con esto, nuestra probabilidad a posteriori, es decir nuestra probabilidad actualizada según los datos observados será:

$$P(\theta|Y) = \frac{P(Y|\theta)P(\theta)}{P(Y)}$$

**Ejemplo 1.2** (Notaciones)**.** Se diseña un ensayo clínico para una nueva droga contra el cáncer, y se desea comparar la probabilidad a los 5 años en una población dado el uso de esta droga, con la misma probablidad dado el tratamiento habitual. De aquí, tenemos:

1. Tasas de sobrevida poblacionales a los 5 años, $\theta = (\theta_1, \theta_2)$.
2. Observaciones $Y = (Y_1, \ldots, Y_n)$.
3. Observaciones futuras, $\widetilde{Y}$.

Con esto, planteamos:

1. Modelo completo: $P(\theta, Y) = P(Y|\theta)P(\theta)$.
2. Distribución aposteriori: $P(\theta|Y) = \dfrac{P(Y|\theta)P(\theta)}{P(Y)} \ \alpha \ P(Y|\theta)P(\theta)$.
3. Distribución predictiva aposteriori: $P(\widetilde{Y}|Y) = \displaystyle\int_\Theta P(\widetilde{Y}, \theta|Y)\ d\theta$. Dado que asumimos independencia condicional entre los datos y los datos futuros, podemos establecer:

$$\begin{aligned} P(\widetilde{Y}|Y) &= \int_\Theta P(\widetilde{Y}, \theta|Y)\ d\theta \\ &= \int_\Theta P(\widetilde{Y}|\theta, Y)P(\theta|Y)\ d\theta \\ &= \int_\Theta P(\widetilde{Y}|\theta)P(\theta|Y)\ d\theta \end{aligned}$$

★

**Ejemplo 1.3.** Suponga que en un examen médico relativo a la tuberculosis se sabe que:

1. La probabilidad de que el test resulte positivo en una persona sin tuberculosis es de 0,005.
2. La probabilidad de que el test resulte positivo en una persona con tuberculosis es de 0,98.
3. El $1\,\%$ de las personas de la población de interés tiene tuberculosis.

Con esto se obtiene que, si una vez realizado el examen éste resulta positivo, la probabilidad de tener realmente tuberculosis es sólo 0,165. ¿Por qué, pese a la gran sensitividad y especificidad del examen, la probabilidad de que el resultado del test sea correcto es tan baja?. **Esto se debe al peso de la apriori. Antes de ver los datos, la probabilidad es muy baja, y esto hace dudar de los resultados del test**.

$\star$

**Ejemplo 1.4.** La Hemofilia se produce por un defecto en un gen en el cromosoma X de las personas. Consideramos un matrimonio, tal que la hija tiene a su vez dos hijos, los cuales no padecen hemofilia. Sea $i = 1, 2$. Definimos

$$Y_i = \begin{cases} 1 & \text{si hijo i representa la enfermedad} \\ 0 & \text{si no} \end{cases}$$

Suponemos que el papá no esta afectado, y que la mujer por si misma tiene $50\,\%$ de probabilidad de tener el gen. Aquí, la cantidad de interés desconocida será el estado de la mujer, el cual posee 2 posibles valores: $P(\theta = 1) = 0{,}5, P(\theta = 0) = 0{,}5$ donde

$$\theta = \begin{cases} 1 & \text{si tiene gen} \\ 0 & \text{si no} \end{cases}$$

Lo anterior corresponde a la probablidad apriori. Nuestra verosimilitud, dada independencia condicional es:

$$P(Y_1 = 0, Y_2 = 0|\theta = 1) = 0,25$$
$$P(Y_1 = 0, Y_2 = 0|\theta = 0) = 1$$

Luego, en virtud del teorema de Bayes obtenemos:

$$\begin{aligned} P(\theta = 1|Y_1 = 0, Y_2 = 0) &= \frac{P(Y|\theta = 1)P(\theta = 1)}{P(Y)} \\ &= \frac{P(Y|\theta = 1)P(\theta = 1)}{P(Y|\theta = 1)P(\theta = 1) + P(Y|\theta = 0)P(\theta = 0)} \\ &= 0,2 \end{aligned}$$

Ahora, si agregamos un tercer hijo que tampoco posea hemofilia, obtenemos usando propiedades de probabilidad condicional:

$$\begin{aligned}
P(\theta = 1|Y_1, Y_2, Y_3 = 0) &= \frac{P(\theta = 1|Y_1 = 0, Y_2 = 0)P(Y_3 = 0|\theta = 1, Y_1 = 0, Y_2 = 0)P(Y_1 = 0, Y_2 = 0)}{P(Y_3 = 0|Y_1 = 0, Y_2 = 0)P(Y_1 = 0, Y_2 = 0)} \\
&= \frac{P(\theta = 1|Y_1 = 0, Y_2 = 0)P(Y_3 = 0|\theta = 1, Y_2 = 0, Y_1 = 0)}{P(Y_3|Y_1 = 0, Y_2 = 0)} \\
&= \frac{P(\theta = 1|Y_1 = 0, Y_2 = 0)P(Y_3 = 0|\theta = 1)}{P(Y_3 = 0|Y_1 = 0, Y_2 = 0)} \\
&= \frac{0{,}20 \cdot 0{,}25}{P(Y_3 = 0|\theta = 1)P(\theta = 1|Y_1, Y_2 = 0) + P(Y_3 = 0|\theta = 0)P(\theta = 0|Y_1, Y_2 = 0)} \\
&= \frac{0{,}20 \cdot 0{,}25}{0{,}5 \cdot 0{,}2 + 1 \cdot 0{,}8}
\end{aligned}$$

Luego, la probabilidad calculada anteriormente pasa a ser una nueva probabilidad a priori, mientras que el denominador corresponde a una distribución predictiva a posteriori.

★

Recordemos que al defininir un modelo de probabilidad, tendremos datos observables y no observables. Dicho modelo estará sujeto a reglas de probabilidad.

## 1.2 Principios generales

**Definición 1.1.** El conjunto $\Omega$ se llamará **espacio muestral** si contiene a todos los posibles resultados de un experimento.

**Definición 1.2.** Un **evento** $A \subset \Omega$ corresponde a una colección de posibles resultados de un experimento. Diremos además que dos eventos $A, B$ son mutuamente excluyentes si $A \cap B = \varnothing$. Los eventos $(A_n)_{n\in\mathbb{N}} \subset \Omega$ se dirán una **Partición** de $\Omega$ si son disjuntos dos a dos y además $\Omega = \bigcup_{n\in\mathbb{N}} A_n$.

**Definición 1.3.** Una $\sigma$-álgebra $\mathcal{F}$ corresponde a una colección de subconjuntos de $\Omega$ que satisface

1. $\varnothing \in \mathcal{F}$
2. $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$
3. $A_1, \ldots, A_n \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{n} A_i$

**Ejemplo 1.5.** Sea $\Omega$ un conjunto cualquiera. Los siguientes son ejemplos sencillos de $\sigma$-álgebra de subconjuntos de $\Omega$.

1. $\mathcal{F}_1 = \{\varnothing, \Omega\}$

2. $\mathcal{F}_2 = \{\varnothing, A, A^c, \Omega\}$

$\star$

**Proposición 1.1.** *Sea $\Omega$ un espacio muestral, $\mathcal{F}$ una $\sigma$-álgebra. Si $A_1, \ldots, A_n \in \mathcal{F}$, entonces $\bigcap_{i=1}^{n} A_i \in \mathcal{F}$.*

*Demostración.* Como $A_1, \ldots, A_n \in \mathcal{F}$ y $\mathcal{F}$ es $\sigma$-álgebra se tiene que

$$\begin{aligned} A_1^c, \ldots, A_n^c \in \mathcal{F} &\Rightarrow \bigcup_{i=1}^{n} A_i^c \in \mathcal{F} \\ &\Rightarrow \left( \bigcup_{i=1}^{n} A_i^c \right)^c \in \mathcal{F} \\ &\Rightarrow \bigcap_{i=1}^{n} A_i \in \mathcal{F} \end{aligned}$$

□

**Definición 1.4.** Sea $\Omega$ un espacio muestral, $\mathcal{F}$ un $\sigma$-álgebra definido sobre este. Una función de probabilidad $P$ con dominio en $\mathcal{F}$ es tal que

1. $0 \leqslant P(A) \leqslant 1 \; \forall \; A \in \mathcal{F}$
2. $P(\Omega) = 1$
3. Si $(A_n)_{n\in\mathbb{N}} \subset \mathcal{F}$ son disjuntos dos a dos, entonces $P\left( \bigcup_{n\in\mathbb{N}} A_n \right) = \sum_{n\in\mathbb{N}} P(A_i)$.

La tripleta $(\Omega, \mathcal{F}, P)$ se llama espacio de probabilidad.

**Ejemplo 1.6.** Sea $(\Omega, \mathcal{F}, P)$ un espacio de probabilidad, $B \in \mathcal{F}$ tal que $P(B) > 0$. Se define

$$P_B(A) = \frac{P(A \cap B)}{P(B)} = P(A|B)$$

Veamos que $P_B(A)$ es una función de probabilidad.

1. Como $A \cap B \subset B$ se tiene que $0 \leqslant P(A \cap B) \leqslant P(B)$. Luego, como $P(B) > 0$ tenemos

$$0 \leqslant P_B(A) \leqslant 1$$

2. $P_B(\Omega) = \dfrac{P(\Omega \cap B)}{P(B)} = 1$.

3. Sean $(A_n)_{n\in N}$ disjuntos dos a dos. Si $i \neq j$ tenemos que $A_i \cap B \cap A_j \cap B = (A_i \cap A_j) \cap B = \varnothing$. Además

$$\begin{aligned} P_B\left(\bigcup_{n\in\mathbb{N}} A_n\right) &= \frac{P(\bigcup_{n\in\mathbb{N}} A_n \cap B)}{P(B)} \\ &= \frac{P\left(\bigcup_{n\in\mathbb{N}}(A_n \cap B)\right)}{P(B)} \\ &= \sum_{n\in\mathbb{N}} \frac{P(A_i \cap B)}{P(B)} \\ &= \sum_{n\in\mathbb{N}} P_B(A_i) \end{aligned}$$

probando lo pedido.

★

**Definición 1.5.** Sea $(\Omega, \mathcal{F}, P)$ un espacio de probabilidad. Sean $A, B$ dos eventos en $\mathcal{F}$ tal que $P(B) > 0$. Se define la probabilidad condicional de $A$ dado $B$ como

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

**Proposición 1.2.** *Sea $(\Omega, \mathcal{F}, P)$ un espacio de probabilidad. Sean $A, B, C$ tres eventos en $\mathcal{F}$ tal que $P(B) > 0$. Entonces*

$$P(A \cap B \cap C) = P(A)P(B|A)P(C|A \cap B).$$

*Demostración.* Es directa de la definición de Probabilidad Condicional. □

**Teorema 1.2** (Bayes)**.** *Sea $(\Omega, \mathcal{F}, P)$ un espacio de probabilidad. Sean $A, B$ dos eventos en $\mathcal{F}$ tal que $P(B) > 0$. Entonces*

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

*Demostración.* Por definición de probabilidad condicional tenemos que

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A).$$

Sigue que

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

□

**Definición 1.6.** Sea $(\Omega, \mathcal{F}, P)$ un espacio de probabilidad. Dos eventos $A, B \in \mathcal{F}$ se dicen independientes si $P(A|B) = P(A)$.

**Observación.** Por definición de probabilidad condicional se tiene que

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Luego si $A, B$ son independientes, es decir $P(A|B) = P(A)$, concluimos que

$$P(A \cap B) = P(A)P(B).$$

De una forma análoga se puede concluir que $P(B|A) = P(B)$.

**Definición 1.7.** Los eventos $A, B$ se dicen condicionalmente independientes dado $C$, con $P(C) > 0$ si

$$P(A|B \cap C) = P(A|C).$$

**Observación.** Por definición de probabilidad condicional se tiene que

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Luego si $A, B$ son condicionalmente independientes dado $C$ se tiene

$$\begin{aligned} P(A \cap B|C) &= \frac{P(A \cap B \cap C)}{P(C)} \\ &= \frac{P(A|B \cap C)P(B \cap C)}{P(C)} \\ &= P(A|C)P(B|C). \end{aligned}$$

Por lo tanto $P(A \cap B|C) = P(A|C)P(B|C)$. Finalmente

$$P(A \cap B \cap C) = P(A|C)P(B|C)P(C).$$

**Definición 1.8.** Una **variable aleatoria** $X$ corresponde a una función definida sobre $\Omega$ tal que $X(\omega) \in \mathbb{R}$ con $\omega \in \Omega$.

**Definición 1.9.** Sea $(\Omega, \mathcal{F}, P)$ un espacio de probabilidad. Dada una variable aleatoria $X$, la medida de probabilidad inducida $P^*$ cumple que $\forall A \subset \mathbb{R}$, tal que $X^{-1}(A) \in \mathcal{F}$

$$P^*(A) = P(X^{-1}(A)) = P(\omega : X(\omega) \in A).$$

**Definición 1.10.** Sean $X$ e $Y$ variables aleatorias. La densidad (o función de masa) de $X$ dado $Y = y$

con $P(y) > 0$ vendrá dada por la distribución condicional

$$P(X|Y = y) = \frac{P(X, Y = y)}{P(Y = y)}.$$

**Teorema 1.3** (Bayes generalizado)**.** *Sean $X$ e $Y$ dos variables aleatorias. La densidad (o función de masa) de $X$ dado $Y = y$ con $P(y) > 0$ puede ser expresada como*

$$P(X|Y = y) = \frac{P(Y = y|X)P(X)}{P(Y = y)}.$$

*Demostración.* Es análoga a la demostración del teorema de Bayes. □

**Teorema 1.4.** *Sean $X, Y, Z$ variables aleatorias definidas sobre el espacio de probabilidad $(\Omega, \mathcal{F}, P)$ y sea $X$ condicionalmente independiente de $Y$ dado $Z$. Entonces*

$$P(X|Y, Z) = P(X|Z)$$

$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$

*donde $P()$ corresponde a la función densidad o masa.*

*Demostración.* Ejercicio. □

**Proposición 1.3.** *Sea $X, Y, Z$ variables aleatorias. Si $X$ e $Y$ son independientes dado $Z$, entonces para cualquier función $g, h$ se tiene*

1. $g(x) \perp Y|Z$.
2. $X \perp Y|Z, g(Z)$.
3. $X \perp (Y, g(Z))|Z$.
4. $h(X, Z) \perp g(Y, Z)|Z$.
5. $X \perp Y|Z, h(X, Z), g(X, Z)$.
6. *Si además* $X \perp W|(Y, Z)$ *entonces* $X \perp (Y, W)|Z$.

*donde $X \perp Y$ se lee como $X$ independiente de $Y$.*

CAPÍTULO 2

# Modelo Bayesiano

Consideramos $P(\widetilde{Y}, \theta)$ donde $\widetilde{Y} = (Y_1, \ldots, Y_n)$. Como habiamos visto antes, el paradigma Bayesiano es realizar inferencias sobre la distribución a posteriori de $\theta$. Sabemos que por el teorema de bayes

$$P(\theta|Y) = \frac{P(Y|\theta)P(\theta)}{P(Y)} = \frac{P(\widetilde{Y}|\theta)P(\theta)}{\displaystyle\int_\Theta P(\widetilde{Y}|\theta)P(\theta)\ d\theta}.$$

**Ejemplo 2.1.** Consideremos $\widetilde{Y}|\theta \overset{\text{i.i.d}}{\sim} \text{Ber}(\theta)$ donde $\theta \sim \text{Uniforme}(0,1)$. Por lo visto antes sabemos que

$$P(\theta|\widetilde{Y}) \ \alpha \ P(\theta)\prod_{i=1}^{n}\theta^{y_i}(1-\theta)^{1-y_i}$$

pero como $\theta$ distribuye uniforme obtenemos

$$P(\theta|\widetilde{Y}) \ \alpha \ \theta^{\sum y_i}(1-\theta)^{n-\sum y_i}.$$

Por lo tanto

$$\theta|\widetilde{Y} \sim \text{Beta}\left(\sum y_i + 1, n - \sum y_i + 1\right)$$

.

★

**Ejemplo 2.2.** Consideremos ahora $\Theta = \{\theta_0, \theta_1\}$. Luego

$$P(\theta) = \begin{cases} P(\theta = \theta_0) \\ P(\theta = \theta_1) \end{cases}$$

donde los valores anteriores son conocidos. Además

$$P(\widetilde{Y}|\theta) = \begin{cases} f_0(\widetilde{Y}) \\ f_1(\widetilde{Y}) \end{cases}$$

Con lo anterior tenemos que

$$\frac{P(\theta = \theta_0|\widetilde{Y})}{P(\theta = \theta_1|\widetilde{Y})} = \frac{f_0(\widetilde{Y})}{f_1(\widetilde{Y})}\frac{P(\theta = \theta_0)}{P(\theta = \theta_1)}$$

Donde el primer factor es llamado factor de verosimilitud y el segundo es llamado chance a priori.

Finalmente, concluimos que

$$P(\theta = \theta_0|\widetilde{Y}) = \frac{f_0(\widetilde{Y})P(\theta = \theta_0)}{f_0(\widetilde{Y})P(\theta = \theta_0) + f_1(\widetilde{Y})P(\theta = \theta_1)}$$
$$\alpha \quad f_0(\widetilde{Y})P(\theta = \theta_0)$$

★

## 2.1 Distribución predictiva a posteriori

Nos interesa predecir una nueva observación $\widetilde{Y}$ dada una observación $Y$. Dicha distribución se obtiene de la siguiente forma:

$$\begin{aligned} P(\widetilde{Y}|Y) &= \frac{P(\widetilde{Y}, Y)}{P(Y)} \\ &= \frac{\int_\Theta P(\widetilde{Y}, Y, \theta)\ d\theta}{P(Y)} \\ &= \frac{\int_\Theta P(\widetilde{Y}|Y, \theta)P(Y, \theta)\ d\theta}{P(Y)} \\ &= \int_\Theta P(\widetilde{Y}|Y, \theta)\frac{P(Y, \theta)}{P(Y)}\ d\theta \\ &= \int_\Theta P(\widetilde{Y}|Y)P(\theta|Y)\ d\theta \end{aligned}$$

Por lo tanto la distribución predictiva del comportamiento futuro $\widetilde{Y}$ dado $Y$ será

$$P(\widetilde{Y}|Y) = \int_\Theta P(\widetilde{Y}|Y)P(\theta|Y)\ d\theta.$$

**Observación.** En algunos libros se encontrará el concepcto de distribución predictiva. Esto corresponde a

$$P(Y_1, \ldots, Y_n) = \int_\Theta P(Y, \theta)\ d\theta = \int_\Theta P(Y|\theta)P(\theta)\ d\theta.$$

**Ejemplo 2.3** (Clasificación de errores de tipeo)**.** Supongamos que una persona tipea 'RADOM'. ¿Es un error?. Llamaremos $\Theta$ al conjunto de posibles palabras tipeadas. Con esto

$$\Theta = \{RADON, RANDOM, RADOM\}$$

Según Google las probabilidades (no normalizadas) de tipear cada una de las palabras es:

| $\theta$ | $P(\theta)$ |
|---|---|
| RANDOM | $7{,}6 * 10^{-5}$ |
| RADON | $6{,}05 * 10^{-6}$ |
| RADOM | $3{,}12 * 10^{-7}$ |

Utilizando además un modelo de Google se determinan los errores en el tipeo.

| $\theta$ | $P(\text{RADOM}\|\theta)$ |
|---|---|
| RANDOM | $0,00193$ |
| RADON | $0{,}000143$ |
| RADOM | $0{,}975$ |

Finalmente para calcular la posteriori debemos calcular $P(\theta)P(\text{RADOM}|\theta)$ lo cual se ve en la siguiente tabla:

| $\theta$ | $P(\theta)P(\text{RADOM}\|\theta)$ | $P(\theta\|\text{RADOM})$ |
|---|---|---|
| RANDOM | $1{,}47 * 10^{-7}$ | 0,325 |
| RADON | $8{,}65 * 10^{-10}$ | 0,002 |
| RADOM | $3{,}04 * 10^{-7}$ | $\underbrace{0{,}673}_{\text{Normalizados}}$ |

★

## 2.2 Modelo Predictivo.

**Definición 2.1** (Modelo Predictivo)**.** Un modelo de probabilidad predictivo para una secuencia de variables $Y_1, Y_2, \ldots, Y_n$ corresponde a una medida de probabilidad $P$ que especifica matemáticamente la forma de creencia conjunta para cualquier subconjunto de $Y_1, \ldots, Y_n$, es decir, nos otorga la distribución conjunta $P(Y_1, \ldots, Y_n)$.

**Definición 2.2** (Permutabilidad Finita)**.** Una secuencia de variables aleatorias se dice permutable o intercambiable si $P(Y_1, \ldots, Y_n)$ es invariante bajo permutaciones de los argumentos, esto es, si $\pi$ es una permutación de $\{1, \ldots, n\}$ entonces

$$P(Y_1, \ldots, Y_n) = P(Y_{\pi(1)}, \ldots, Y_{\pi(n)}).$$

**Ejemplo 2.4** (Urna de Polya)**.** Considere una urna que inicialmente tiene $r$ bolitas rojas y $v$ verdes. Se realiza la siguiente secuencia de ensayos. En cada etapa se escoge una bolita al azar, se observa su color

y se devuelve junto con $c$ bolitas del mismo color. Supongamos que queremos calcular $P(R_1, V_2, R_3)$ y $P(V_1, R_2, R_3)$. Notemos que

$$\begin{aligned} P(R_1, V_2, R_3) &= P(R_1)P(V_2|R_1)P(R_3|V_2 \cap R_1) \\ &= \frac{r}{r+v} \cdot \frac{v}{r+v+c} \cdot \frac{r+c}{r+v+2c} \end{aligned}$$

mientrás que

$$\begin{aligned} P(V_1, R_2, R_3) &= P(V_1)P(R_2|V_1)P(R_3|V_1 \cap R_R) \\ &= \frac{v}{r+v} \cdot \frac{r}{r+v+c} \cdot \frac{r+c}{r+v+2c} \end{aligned}$$

por lo cual $P(R_1, V_2, R_3) = P(V_1, R_2, R_3)$ y por lo tanto las variables son permutables.

★

**Definición 2.3** (Permutabilidad Infinita)**.** La secuencia infinita de variables aleatorias se dice permutable infinita si toda subsecuencia finita es permutable finita.

**Teorema 2.1** (Teorema de DeFinetti para variables 0-1)**.** *Sea $Y_1, \ldots, Y_n$ una secuencia permutable infinita con valores $\{0, 1\}$ y medida de probabilidad $P$. Entonces existe una distribución $Q$ tal que la función de probabilidad predictiva $P(Y_1, \ldots, Y_n)$ tiene la forma*

$$P(Y_1, \ldots, Y_n) = \int_0^1 \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i}\, dQ(\theta)$$

*donde*

$$Q(\theta) = \lim_{n\to\infty} P\left(\frac{\delta_n}{n} \leqslant \theta\right),\ \delta_n = \sum_{i=1}^{n} y_i,\ \theta = \lim_{n\to\infty} \frac{\delta_n}{n}.$$

*Demostración.* Informalmente podemos concebir el proceso que genera la distribución conjunta de $Y_1, \ldots, Y_n$ como

$$Y_1, \ldots Y_n \overset{\text{i.i.d}}{\sim} \text{Ber}(\theta)$$

$$\theta \sim P(\theta).$$

Luego nuestra función de verosimilitud es

$$P(Y_1, \ldots, Y_n|\theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i}$$

y por lo tanto nuestra distribución predictiva es:

$$P(Y_1,\ldots,Y_n) = \int_0^1 P(Y_1,\ldots,Y_n|\theta)P(\theta)\ d\theta = \int_0^1 \prod_{i=1}^n \theta^{y_i}(1-\theta)^{1-y_i} \underbrace{P(\theta)d\theta}_{dQ(\theta)}.$$

□

**Corolario 2.1.1.** *Bajo las condiciones del teorema anterior tendremos que*

$$P(Y_1+Y_2+\cdots+Y_n = S_n) = \int_0^1 \binom{n}{S_n}\theta^{S_n}(1-\theta)^{n-S_n}\ dQ(\theta).$$

*Demostración.* Sigue del hecho de que suma de variables aleatorias Bernoulli iid es Binomial. □

**Corolario 2.1.2.** *Sea $Y_1,\ldots,Y_n$ una secuecnia infinita permutable de variables 0-1, con medida de probabilidad $P$. La probabilidad condicional de $Y_{m+1},\ldots,Y_n$ dado $Y_1,\ldots,Y_m$, con $1 \leqslant m < n$, tiene la forma*

$$P(Y_{m+1},\ldots,Y_n|Y_1,\ldots,Y_m) = \int_0^1 \prod_{i=1}^n \theta^{y_i}(1-\theta)^{1-y_i}\ dQ(\theta|Y_1,\ldots,Y_m)$$

*donde*

$$dQ(\theta|Y_1,\ldots,Y_m) = \frac{\prod_{i=1}^n \theta^{y_i}(1-\theta)^{1-y_i}\ dQ(\theta)}{\int_0^1 \prod_{i=1}^n \theta^{y_i}(1-\theta)^{1-y_i}\ dQ(\theta)}, \qquad Q(\theta) = \lim_{m\to\infty} P\left(\frac{S_m}{m} \leqslant \theta\right).$$

*Demostración.* Ejercicio. □

**Corolario** (Teorema Representación General)**.** *Asumiendo que las densidades necesarias existen y bajo las condiciones del teorema anterior, la densidad conjunta ó predictiva de $Y_1,\ldots,Y_n$ tiene la forma*

$$P(Y_1,\ldots,Y_n) = \int_0^1 \prod_{i=1}^n p(Y_i = y_i|\theta)\ dQ(\theta)$$

*con $P(\cdot|\theta)$ función densidad.*

**Ejemplo 2.5.** Consideremos el experimento de girar una moneda sobre su eje. La experiencia nos dice que la mayoría de las monedas tenían cierta inclinación a caer cara. Luego, nuestra creencia a priori sería una mezcla de betas:

$$P(\theta) = \omega_1\text{Beta}(\alpha_1,\beta_1) + \omega_2\text{Beta}(\alpha_2,\beta_2) + \omega_3\text{Beta}(\alpha_3,\beta_3)$$

donde $\omega_1+\omega_2+\omega_3 = 1$. Simular una grilla(secuencia) para $\theta$ en $0-1$ con pasos pequeños. En virtud de que

$$P(\theta|\widetilde{Y})\ \ \alpha\ \ P(\widetilde{Y}|\theta)P(\theta)$$

evaluamos la grilla en $P(\theta|\widetilde{Y})$.

$\star$

**Ejemplo 2.6.**

1. Sea la secuencia no necesariamente independiente $Y_1, \ldots, Y_n$ que toma valores 0 y 1. Sea $S_n = Y_1 + \cdots + Y_n$. Represente la probabilidad $P(Y_{n+1} = 1|S_n = s)$ en el caso en que todas las permutaciones de los $Y_i$ tengan igual probabilidad. Notar que

$$\begin{aligned} P(Y_{n+1} = 1|S_n = s) &= \int_0^1 P(Y_{n+1} = 1|\theta)\, dQ(\theta|S_n = s) \\ &= \int_0^1 \theta\, dQ(\theta|S_n = s) \end{aligned}$$

pues el experimento $Y_{n+1}$ es Bernoulli. Por otra parte, en virtud del teorema de Bayes:

$$\begin{aligned} dQ(\theta|S_n = s) &= P(\theta|S_n = s)d\theta \\ &= \frac{P(S_n = s|\theta)P(\theta)}{P(S_n = s)}d\theta \\ &= \frac{\binom{n}{s}\theta^s(1-\theta)^{n-s}P(\theta)}{P(S_n = s)}d\theta \end{aligned}$$

Finalmente, en virtud del Corolario 4.1 concluimos que

$$P(Y_{n+1}|S_n = s) = \int_0^1 \frac{\binom{n}{s}\theta^{s+1}(1-\theta)^{n-s}}{\int_0^1 \binom{n}{s}\theta^s(1-\theta)^{n-s}\, dQ(\theta)} dQ(\theta)$$

2. Discuta: Si una persona siente subjetivamente que un evento ocurrirá con frecuencia relativa $\theta$ en una gran cantidad de ensayos independientes e idénticos, entonces su probabilidad subjetiva de que el evento ocurra en un único intento debe ser $\theta$.

★

## 2.3 Modelos Paramétricos

### 2.3.1. Caso Unidimensional

Supongamos $Y_1, \ldots, Y_n \in \{0, 1\}$ tal que

$$Y_1, \ldots, Y_n|\theta \overset{\text{i.i.d}}{\sim} \text{Ber}(\theta)$$

La verosimilitud viene dada por

$$P(\widetilde{Y}|\theta) = \theta^s(1-\theta)^{n-s}$$

con $s = y_1 + \ldots + y_n$. Supongamos además que $\theta \sim \text{Beta}(\alpha, \beta)$, es decir

$$P(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

con $0 < \theta < 1$, $\alpha, \beta > 0$. Sabemos que

$$\begin{aligned} P(\theta|\widetilde{Y}) &\propto P(\widetilde{Y}|\theta)P(\theta) \\ &\propto \theta^{\alpha+s-1}(1-\theta)^{\beta+n-s-1} \end{aligned}$$

De donde sigue

$$\theta|\widetilde{Y} \sim \text{Beta}(\alpha+s, \beta+n-s)$$

Por tanto $P(\theta)$ y $P(\theta|\widetilde{Y})$ tienen la misma forma funcional. Este comportamiento es característico de un modelo conjugado. Veamos como se comportan la Esperanza y la varianza de $\theta|\widetilde{Y}$:

$$\begin{aligned} \mathbb{E}(\theta|\widetilde{Y}) &= \frac{\alpha+s}{\alpha+\beta+n} \\ &= \frac{n}{\alpha+\beta+n}\cdot\frac{s}{n} + \frac{\alpha+\beta}{\alpha+\beta+n}\cdot \underbrace{\frac{\alpha}{\alpha+\beta}}_{\mathbb{E}(\theta) \text{ apriori}} \end{aligned}$$

Se infiere que la esperanza corresponde a una combinación lineal entre la esperanza apriori y la proporción empírica de éxitos $\frac{s}{n}$. Notar que si aumentamos $n$, la información que me entrega la distribución a priori es irrelevante. Además,

$$\text{Var}(\theta|\widetilde{Y}) = \frac{(\alpha+s)(\beta+n-s)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}$$

Más aún se puede probar, mediante el teorema del límite central, que

$$\frac{\theta - \mathbb{E}(\theta|\widetilde{Y})}{\sqrt{\text{Var}(\theta|\widetilde{Y})}} \mid \widetilde{Y} \xrightarrow[n\to\infty]{D} \text{N}(0,1).$$

**Observación.** Si hubiesemos tomado $Y|\theta \sim \text{Bin}(n, \theta)$ tendremos que la función de verosimilitud será

$$P(Y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$

la cual es proporcional a la vista anteriormente. Luego, en virtud del principio de verosimilitud, ambas

distribuciones nos llevaran a las mismas inferencias. Por otro lado

$$\begin{aligned} P(Y) &= \int_0^1 \binom{n}{y}\theta^y(1-\theta)^{n-y}P(\theta)\ d\theta \\ &= \int_0^1 \binom{n}{y}\theta^y(1-\theta)^{n-y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\ d\theta \\ &= \binom{n}{y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 \theta^{y+\alpha-1}(1-\theta)^{\beta+n-y-1}\ d\theta \\ &= \binom{n}{y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+y)\Gamma(\beta+n-y)}{\Gamma(\alpha+\beta+n)} \end{aligned}$$

Luego

$$P(Y) = \binom{n}{y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+y)\Gamma(\beta+n-y)}{\Gamma(\alpha+\beta+n)}$$

que es llamada distribución Beta-Binomial.

**Definición 2.4.** El principio o axioma de verosimilitud plantea que dentro del sistema de un modelo estadístico, toda la información que proveen los datos en cuanto a los méritos relativos de dos hipótesis está contenida en el cociente de verosimilitud de esas hipótesis sobre los datos, y el cociente de verosimilitud se interpretará como el grado en que los datos soporta una hipótesis frente a la otra. Es decir,

$$P_2(\widetilde{Y}|\theta) = c(\widetilde{Y})P_1(\widetilde{Y}|\theta)$$

En este caso, la inferencia bayesiana genera una posteriori que respeta el principio de verosimilitud.

**Observación.** A veces conviene reparametrizar la distribución beta usando como parámetros

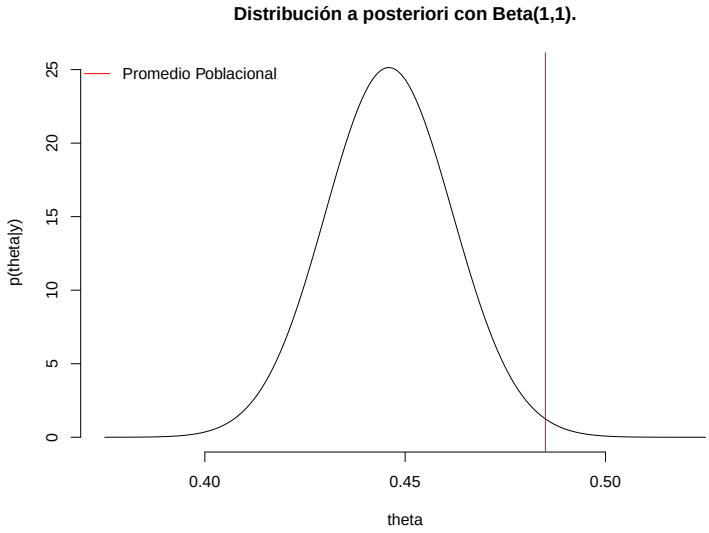
$$\rho = \frac{\alpha}{\alpha+\beta},\ \ n_0 = \alpha+\beta$$

donde $\rho$ es la media de la beta. Además, la varianza nos queda

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{\rho(1-\rho)}{n_o+1}.$$

**Ejemplo 2.7** (BDA3)**.** Se desea estudiar la proporción $\theta$ de nacimientos femeninos entre aquellos casos de embarazos con placenta previa. Se observan 437 de estos nacimientos entre 980 casos de placenta previa, es decir, un $44,6\,\%$. El porcentaje de nacimientos femeninos en la población general es de $48,5\,\%$. Simularemos en R, variando la distribución apriori de Beta:

1. $\theta \sim$ Beta$(1,1) =$ Uniforme$(0,1)$. Con R obtenemos el siguiente gráfico:

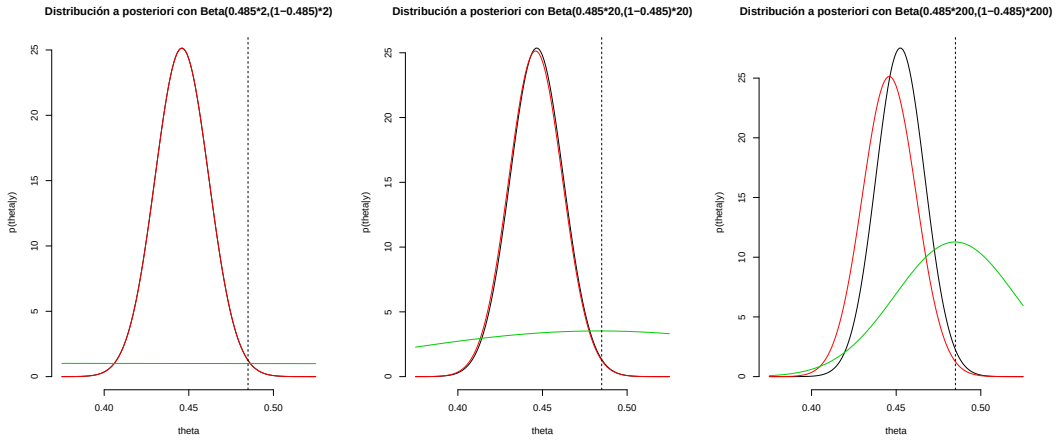



Observamos que la media de la beta obtenida esta muy alejada de lo que pasa en la población. Esto se debe a que la cantidad de casos tomados puede ser muy pequeña en comparación a la población general, además de que la muestra quizás no es tan representativa de la población (Recordemos que a medida que aumenta el $n$ la apriori no tiene mucho peso).

2. Ahora veremos 3 formas diferentes simultaneamente de una distribución Beta:

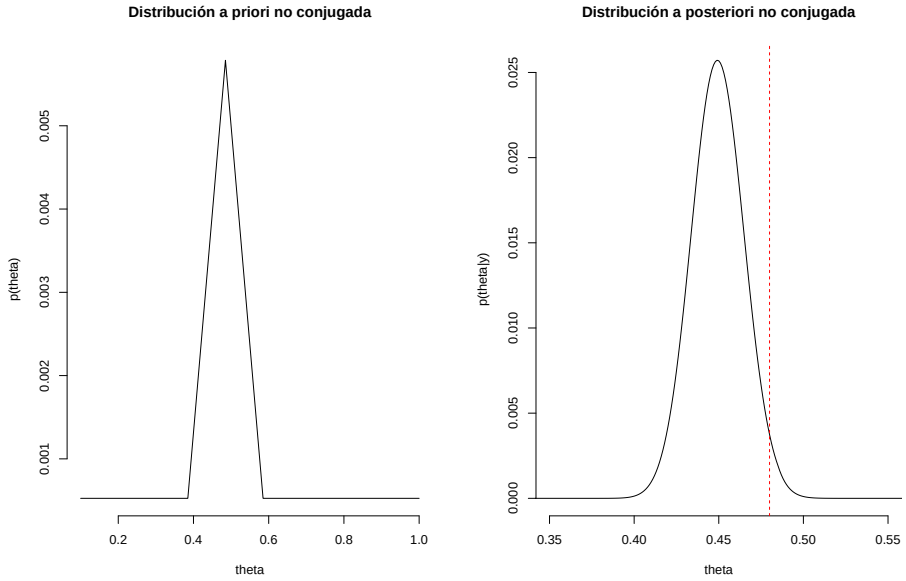
$$\theta \sim \text{Beta}(0{,}485 * 2, (1 - 0{,}485) * 2)$$
$$\theta \sim \text{Beta}(0{,}485 * 20, (1 - 0{,}485) * 20)$$
$$\theta \sim \text{Beta}(0{,}485 * 200, (1 - 0{,}485) * 200)$$

Graficando en R obtenemos:




Nuevamente observamos que la media esta alejada de lo que pasa en la población. Esto se puede deber a que la muestra obtenida no es representativa de la población. **Preguntar!**

3. Ahora utilizaremos una distribución apriori no conjugada para $\theta$. Gráficamente tenemos:




Observamos que la distribución a priori se concentra entre $0{,}4$ y $0{,}6$. Al obtener la aposteriori observamos que esta tiende a concentrarse, al igual que con la Beta, entre $0{,}4$ y $0{,}5$.

★

**Poisson-Gamma**

Consideremos el siguiente modelo

$$Y_1, \ldots, Y_n \overset{\text{i.i.d}}{\sim} \text{Poisson}(\theta),\ \theta > 0$$
$$\theta \sim \text{Gamma}(\alpha, \beta)$$

Luego

$$\left.\begin{array}{l} P(\widetilde{Y}|\theta) \propto \prod_{i=1}^{n} \theta^{y_i} e^{-\theta} \\ P(\theta) \propto \theta^{\alpha-1} e^{-\beta\theta} \end{array}\right\} \Rightarrow P(\theta|\widetilde{Y}) \propto P(\widetilde{Y}|\theta)P(\theta) = \theta^{\alpha+\sum y_i - 1} e^{-(\beta+n)\theta}$$

Por ende

$$\theta|\widetilde{Y} \sim \text{Gamma}(\alpha + \sum y_i, \beta + n)$$

lo que nos dice que este modelo corresponde a un modelo conjugado.

**Observación.** Si consideramos $n = 1$ tenemos que nuestra distribución predictiva es:

$$P(\widetilde{Y}) = P(Y_1) = \int_0^\infty \frac{\theta^{y_1}}{y_1!} e^{-\theta} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}\, d\theta = \frac{\Gamma(\alpha + y_1)\beta^\alpha}{(\beta+1)^{\alpha+y_1}\Gamma(\alpha)y_1!}$$

La cual corresponde a una binomial negativa desplazada.

**Normal**

1. **Normal con $\mu$ desconocido y $\sigma^2$ conocido**: Consideremos el siguiente modelo

$$Y_1, \ldots, Y_n \overset{\text{i.i.d}}{\sim} \text{Normal}(\theta, \sigma^2),\ \sigma^2 \text{ conocido}$$
$$\theta \sim \text{Normal}(\mu_0, \tau_0^2).$$

   Luego $\theta|\widetilde{Y} \sim \text{Normal}(\mu_n, \tau_n^2)$ donde

$$\mu_n = \frac{\frac{\mu_0}{\tau_0^2} + \frac{n\bar{Y}}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}, \qquad \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$

   donde $\dfrac{1}{\tau_n^2}$ es llamada la precisión a posteriori. De lo anterior, vemos que este modelo también es conjugado.

   **Ejercicio.** Determinar la posteriori predictiva $P(\widetilde{Y}|\hat{Y})$.

2. **Normal con $\mu$ conocido y $\sigma^2$ desconocido:** Consideremos $Y_1, \ldots, Y_n \overset{\text{i.i.d}}{\sim} \text{Normal}(\theta, \sigma^2)$. Nos gustaría encontrar una distribución adecuada para $\sigma^2$. Para esto consideremos $W \sim \text{Gamma}(a, b)$ con densidad

$$P_W(w) = \begin{cases} \dfrac{b^a}{\Gamma(a)} w^{a-1} e^{bw} & \text{si } w > 0 \\ 0 & \text{si e.o.c} \end{cases}$$

   Buscamos la distribución de $Z = \dfrac{1}{W}$, la cual se llama Gamma Inversa(GI), la cual tiene densidad

$$P_Z(z) = \begin{cases} \dfrac{b^a}{\Gamma(a)} \dfrac{1}{z^{a+1}} e^{b/z} & \text{si } z > 0 \\ 0 & \text{si e.o.c} \end{cases}$$

Con esto, supongamos $\sigma^2 \sim \text{GI}(a, b)$. Con esta apriori, nuestra aposteriori queda de la forma:

$$\begin{aligned} P(\sigma^2|\widetilde{Y}) \;&\alpha\; P(\widetilde{Y}|\sigma^2)P(\sigma^2) \\ &\alpha\; \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right\}\frac{1}{(\sigma^2)^{a+1}}e^{-b/\sigma^2} \\ &\alpha \left(\frac{1}{\sigma^2}\right)^{n/2+a+1} \exp\left\{-\frac{1}{\sigma^2}\left[\frac{\sum_{i=1}^{n}(y_i-\theta)^2}{2}+b\right]\right\} \end{aligned}$$

Por lo tanto

$$\sigma^2|\widetilde{Y} \sim \text{GI}\left(n/2+a, \frac{\sum_{i=1}^{n}(y_i-\theta)^2}{2}+b)\right)$$

**Observación.** Recordemos que $\chi^2(\nu) \cong \text{Gamma}\left(\frac{\nu}{2}, \frac{1}{2}\right)$ por lo cual

$$\chi^{-2}(\nu) \cong \text{GI}\left(\frac{\nu}{2}, \frac{1}{2}\right).$$

La distribución de $\chi^2$ inversa escalada es $\text{GI}\left(\frac{\nu}{2}, \frac{\nu}{2}S^2\right)$. Si en el modelo anterior asumimos que $\sigma^2 \sim \text{GI}\left(\frac{\nu}{2}, \frac{\nu}{2}S^2\right) \cong \chi^{-2}(\nu_0, \sigma_0^2)$ entonces

$$\sigma^2|\widetilde{Y} \sim \chi^{-2}\left(\nu_0+n, \frac{\nu_0\sigma^2+n}{\nu_0+n}V\right)$$

con $V = \frac{1}{n}\sum_{i=1}^{n}(y_i-\theta)^2$.

## 2.4 Estadísticos de Suficiencia

**Definición 2.5** (Estadístico)**.** Dadas las variables aleatorias $Y_1, \ldots, Y_n$ con conjuntos de valores posibles $\mathcal{Y}_1, \ldots, \mathcal{Y}_n$ respectivamente, un vector aleatorio

$$t_n : \mathcal{Y}_1 \times \ldots \times \mathcal{Y}_n \to \mathbb{R}^{k(n)}$$

con $k(n) \leqslant n$ se denomina estadístico-$k(n)$ dimensional.

**Definición 2.6** (Suficiencia Predictiva)**.** Dada una secuencia de variables aleatorias $Y_1, Y_2, \ldots$ permutables, con medida de probabilidad $P$, con $Y_i$ tomando valores en $\mathcal{Y}_i$. La secuencia de estadísticos $T_1, \ldots, T_n$ definido en $\mathcal{Y}_1 \ldots \times \mathcal{Y}_n$ se dice suficiente predictiva para la secuencia $Y_1, Y_2, \ldots$ si

$\forall n \geqslant 1, r \geqslant 1$ y $\{i_1, \ldots, i_r\} \cap \{1, \ldots, n\} = \varnothing$ se tiene que

$$P(Y_{i_1}, \ldots, Y_{i_r} | Y_1, \ldots Y_n) = P(Y_{i_1}, \ldots, Y_{i_r} | T_n)$$

donde $P(\cdot|\cdot)$ corresponde a la densidad condicional inducida por $P$.

**Definición 2.7** (Suficiencia Paramétrica)**.** Dada una secuencia de variables aleatorias $Y_1, Y_2, \ldots$ permutables, con medida de probabilidad $P$, con $Y_i$ tomando valores en $\mathcal{Y}_i$. La secuencia de estadísticos $T_1, T_2, \ldots$ con $T_j$ definido en $\mathcal{Y}_i \times \ldots \times \mathcal{Y}_j$ se dice suficiente paramétrica si $\forall n \geqslant 1$

$$dQ(\theta | Y_1, \ldots, Y_n) = dQ(\theta | T_n)$$

para cualquier $dQ(\theta)$ que defina un modelo predictivo a través de

$$P(Y_1, \ldots, Y_n) = \int_\Theta \prod_{i=1}^{n} P(Y_i|\theta)\ dQ(\theta).$$

**Proposición 2.1.** *La secuencia $T_1, T_2, \ldots$ es suficiente para una secuencia infinita permutable $Y_1, Y_2, \ldots$ que admite una representación de mezcla ssi $\forall n \geqslant 1$ la densidad conjunta de $Y_1, \ldots, Y_n$ dado $\theta$ tiene la forma*

$$P(Y_1, \ldots, Y_n | \theta) = H_n(t_n, \theta) G(Y_1, \ldots, Y_n)$$

*para algunas funciones $H_n, G$.*

*Demostración.* $\boxed{\Longrightarrow}$

Supongamos que la secuencia $T_1, T_2, \ldots$ es suficiente para $Y_1, Y_2, \ldots$. Luego

$$\begin{aligned} dQ(\theta | Y_1, \ldots, Y_n) &= dQ(\theta | T_n) \\ \frac{P(Y_1, \ldots, Y_n|\theta) dQ(\theta)}{P(Y_1, \ldots, Y_n)} &= \frac{P(T_n|\theta) dQ(\theta)}{P(T_n)} \\ P(Y_1, \ldots, Y_n|\theta) &= \underbrace{\frac{P(T_n|\theta)}{P(T_n)}}_{H_n(T_n, \theta)} \underbrace{P(Y_1, \ldots, Y_n)}_{G(Y_1, \ldots, Y_n)} \end{aligned}$$

Probando lo pedido.

$\boxed{\Longleftarrow}$

Supongamos que se cumple la factorización

$$P(Y_1, \ldots, Y_n | \theta) = H_n(t_n, \theta) G(Y_1, \ldots, Y_n)$$

Luego

$$\begin{aligned} dQ(\theta|Y_1,\ldots,Y_n) &= \frac{P(Y_1,\ldots,Y_n|\theta)dQ(\theta)}{\int_\Theta P(Y_1,\ldots,Y_n)\ dQ(\theta)} \\ &= \frac{H_n(T_n,\theta)G(Y_1,\ldots,Y_n)dQ(\theta)}{\int_\Theta H_n(T_n,\theta)G(Y_1,\ldots,Y_n)\ dQ(\theta)} \\ &= dQ(\theta|T_n) \end{aligned}$$

Lo que nos dice que la secuencia de estadísticos es suficiente. □

**Teorema 2.2.** *La secuencia $T_1, T_n, \ldots$ es suficiente para la secuencia $Y_1, Y_2, \ldots$ infinita permutable ssi el vector $(Y_1,\ldots,Y_n)$ es independiente de $\theta$ dado $T_n$, $P(\widetilde{Y}|T_n)$.*

**Definición 2.8** (Familia de distribuciones apriori cojugada)**.** Sea $\mathcal{F}$ una familia de distribuciones de muestreo $P(Y|\theta)$, $\mathcal{P}$ una clase de distribuciones a priori para $\theta$. Entonces $\mathcal{P}$ es conjugada para $\mathcal{F}$ si

$$P(\theta|Y) \in \mathcal{P}$$

para todo $P(\cdots|\theta)$ y $P(\cdot) \in \mathcal{P}$.

**Definición 2.9.** Sea $(Y_1,\ldots,Y_n)$ una muestra aleatoria proveniente de una familia exponencial regular(el soporte de la variable aleatoria no depende del parámetro), entonces

$$P(\widetilde{Y}|\theta) = \prod_{i=1}^{n} f(y_i)(g(\theta))^n \exp\{\phi(\theta)^T \sum_{i=1}^{n} h(y_i)\}$$

Entonces la familia de distribuciones conjugadas para $\theta$ es de la forma

$$P(\theta|\xi,\nu)\ \ \alpha\ \ [g(\theta)]^\xi \exp\{\phi(\theta)^T\nu\}, \qquad \theta \in \Theta.$$

Para la verosimilitud de una familia exponencial regular y su priori conjugada, la densidad aposteriori para $\theta$ está dada por

$$P(\theta|\widetilde{Y},\xi,\nu) = P(\theta|\xi+n,\nu+T(y))$$

donde $T(y)$ es el estadístico suficiente para $\theta$. Entonces

$$P(\theta|\widetilde{Y},\xi,\nu)\ \ \alpha\ \ [g(\theta)]^{\xi+n} \exp\{\phi(\theta)^T[\nu+T(y)].$$

**Ejemplo 2.8.** Consideremos la distribución Weibull$(\lambda,k)$, con $k>0$ conocido y función de densidad

$$P(Y|\lambda) = \frac{k}{\lambda}\left(\frac{y}{\lambda}\right)^{k-1} \exp\left\{-\left(\frac{y}{\lambda}\right)^k\right\}$$

con $y, \lambda > 0$. Notemos que esta distribución pertenece a la familia exponencial regular pues

$$P(Y|\lambda) = \frac{k}{\lambda}\left(\frac{y}{\lambda}\right)^{k-1} \exp\left\{-\left(\frac{y}{\lambda}\right)^{k}\right\} = k\frac{1}{\lambda^k} y^{k-1} \exp\left\{-\left(\frac{y}{\lambda}\right)^{k}\right\}$$

donde tomamos

$$\begin{aligned} g(\lambda) &= \frac{1}{\lambda^k} \qquad & \phi(\lambda) &= \frac{-1}{\lambda^k} \\ f(y) &= y^{k-1} \qquad & h(y) &= y^k \end{aligned}$$

Si ahora consideramos $(Y_1, \ldots, Y_n)$ tendremos que el estadístico suficiente es

$$T_n = \sum_{i=1}^{n} y_i^k$$

Luego, nuestra función apriori queda

$$P(\lambda, \tau = (T_0, T_1)) \ \alpha \ \left(\frac{1}{\lambda}\right)^{kT_0} \exp\left\{\frac{-1}{\lambda^k} T_1\right\}$$

Ahora bien, si quisieramos encontrar la distribución específica, tendremos que

$$P(\lambda|\tau = (T_0, T_1)) = K\left(\frac{1}{\lambda}\right)^{kT_0} \exp\left\{\frac{-1}{\lambda^k} T_1\right\}$$

Integrando obtenemos que

$$K = \frac{1}{\int_0^\infty \left(\frac{1}{\lambda}\right)^{kT_0} \exp\left\{\frac{-1}{\lambda^k} T_1\right\}}.$$

En el caso general tendremos que

$$P(\lambda|\tau, \widetilde{Y}) \ \alpha \ \left(\frac{1}{\lambda}\right)^{k(T_0+n)} \exp\left\{\frac{-1}{\lambda^k}(T_1 + T_n(y))\right\}.$$

**Ejercicio:** Calcular distribución predictiva a posteriori $P(\hat{Y}|\widetilde{Y})$.

★

## 2.5 Distribuciones a priori no informativas, vagas y débilmente informativas

Supongamos que

$$P(\theta \in (0,1)) = 1 \Rightarrow P(\theta \in (0,1)|\widetilde{Y}) = 1.$$

Esto representa un ejemplo de priori no informativa, pues refleja el hecho de no tener alguna idea del comportamiento de la distribución de la priori. Jeffreys propone el siguiente razonamiento: Si no sabemos nada de un cierto parámetro $\theta$, entonces no sabemos nada de transfomaciones tales como $\theta^2, \log(\theta), etc$. Consideremos la transformación $\phi = h(\theta)$ donde $\theta \sim P(\theta)$. Entonces

$$\phi \sim P(h^{-1}(\phi)) \left| \frac{\partial}{\partial \phi} h^{-1}(\phi) \right|$$

### 2.5.1. Priori de Jeffreys

Nos gustaría encontrar una priori no informativa. Jeffreys encuentra que la distribución apriori debe ser de la forma

$$P(\theta) \ \alpha \ |J(\theta)|^{\frac{1}{2}}$$

donde $J(\theta)$ corresponde a la matriz de información de Fisher

$$J(\theta) = \mathbb{E}\left( \left[ \frac{\partial \log(P(Y|\theta)}{\partial \theta} \right]^2 |\theta \right).$$

Además bajo ciertas condicones de regularidad

$$J(\theta) = -\mathbb{E}\left( \frac{\partial^2 \log(P(Y|\theta))}{\partial \theta^2} |\theta \right).$$

**Principio de Invarianza**

Si tenemos un criterio para definir una distribución apriori para $\theta$, es decir podemos dar $P(\theta)$, entonces el mismo criterio aplicado a un parámetro $\phi = h(\theta)$, con $h$ función inyectiva, debe entregar

$$P(\phi) = P(\theta) \left| \frac{\partial h^{-1}(\phi)}{\partial \phi} \right|$$

**Definición 2.10** (Distribución Impropia)**.** Decimos que una distribución es impropia si su integral no esta definida, es decir, si es infinita.

**Observación.** En el caso de Joffreys, puede pasar que la distribución encontrada sea propia o impropia.

**Ejemplo 2.9** (Binomial)**.** Consideremos $Y|\theta \sim \text{Bin}(n, \theta)$. Entonces

$$P(Y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

Notemos que

$$\log(P(Y|\theta)) = k + y \log(\theta) + (n-y) \log(1-\theta)$$

Derivando obtenemos

$$\frac{\partial^2 \log(P(Y|\theta))}{\partial \theta^2} = \frac{-y}{\theta^2} - \frac{(n-y)}{(1-\theta)^2}$$

Luego

$$J(\theta) = -\mathbb{E}\left(\frac{-y}{\theta^2} - \frac{(n-y)}{(1-\theta)^2}\right) = \frac{n}{\theta} + \frac{n}{(1-\theta)} = \frac{1}{\theta(1-\theta)}$$

Finalmente

$$P(\theta) \ \alpha \ \theta^{\frac{1}{2}-1}(1-\theta)^{\frac{1}{2}-1}$$

La cual corresponde a una Beta $\left(\frac{1}{2}, \frac{1}{2}\right)$.

★

**Ejemplo 2.10.** Consideremos

$$Y_1, \ldots, Y_n \overset{\text{i.i.d}}{\sim} \text{Normal}(\theta, \sigma^2)$$

- Con $\sigma$ conocido, al calcular nuestra priori de Jeffreys resulta

$$P(\theta) \ \alpha \ 1$$

- Con $\theta$ conocido, al calcular nuestra priori de Jeffreys resulta

$$P(\sigma^2) \ \alpha \ \frac{1}{\sigma^2}$$

  Un caso especial es que

$$P(\log(\sigma^2)) \ \alpha \ 1$$

  Notemos que si nuestra priori para $\sigma^2$ cumple que

$$\sigma^2 \sim \chi^{-2}(\nu_0, \sigma_0^2),$$

  nuestra priori vendrá dada por

$$P(\sigma^2) \ \alpha \ (\sigma^2)^{-\nu_0/2+1} \exp\left\{\frac{-\nu_0\sigma_0^2}{2\sigma^2}\right\}$$

  y tomando $nu_0 = 0$ obtenemos

$$P(\sigma^2) \ \alpha \ \frac{1}{\sigma^2}$$

  la cual coincide con la priori de Jeffreys.

★

### 2.5.2. Modelos Multiparamétricos

**Definición 2.11.** Suponga un modelo de probabilidad multiparamétrico $\theta = (\theta_1, \theta_2)$. Si nuestro interés está en realizar inferencias únicamente sobre uno de ellos, digamos $\theta_1$, diremos que $\theta_2$ corresponde a un parámetro de ruido.

**Observación.** En este caso, debemos integrar la distribución conjunta con respecto a $\theta_2$, es decir obtener

$$P(\theta_1|\widetilde{Y}) = \int_{\Theta} P(\theta_1, \theta_2|\widetilde{Y})\, d\theta_2$$

A menudo es posible generar muestras de $P(\theta_1|\widetilde{Y})$ mediante:

1. $\theta_2^* \sim P(\theta_2|\widetilde{Y})$.
2. $\theta_1^* \sim P(\theta_1|\theta_2^*\widetilde{Y})$.
3. $(\theta_1^*, \theta_2^*) \sim P(\theta_1, \theta_2|\widetilde{Y})$.

Luego de esto, por propiedades de condicionalización se tiene que

$$P(\theta_1, \theta_2|\widetilde{Y}) \;\alpha\; P(\theta_1|\theta_2, \widetilde{Y})P(\theta_2|\widetilde{Y})$$

**Ejemplo 2.11.** Consideremos

$$Y_1, \ldots, Y_n|\mu, \sigma^2 \sim \text{Normal}(\mu, \sigma^2)$$
$$P(\mu, \sigma^2) \;\alpha\; \frac{1}{\sigma^2}$$

Luego

$$P(\widetilde{Y}|\mu, \sigma^2) \;\alpha\; \prod_{i=1}^{n} \frac{1}{\sqrt{\sigma^2}} \exp\left\{-\frac{(y_i - \mu)^2}{2\sigma^2}\right\}.$$

Por ende

$$\begin{aligned} P(\mu, \sigma^2|\widetilde{Y}) \;&\alpha\; P(\widetilde{Y}|\mu, \sigma^2)P(\mu, \sigma^2) \\ &\alpha\; \left(\frac{1}{\sigma^2}\right)^{n/2+1} \exp\left\{-\frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2\sigma^2}\right\} \end{aligned}$$

Ahora bien, notando que

$$\sum_{i=1}^{n}(y_i - \mu)^2 = \sum_{i=1}^{n}(y_i - \bar{y} + \bar{y} - \mu)^2$$

y recordando que $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ se obtiene

$$P(\mu, \sigma^2|\widetilde{Y}) \;\; \alpha \;\; (\sigma^2)^{-(n/2+1)} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)S^2 + n(\bar{Y}-\mu)^2\right]\right\}$$

Con la expresión anterior resulta

$$P(\mu|\sigma^2, \widetilde{Y}) \;\; \alpha \;\; \exp\left\{-\frac{1}{2\sigma^2}[n(\bar{Y}-\mu)^2]\right\}$$

de donde obtenemos que

$$\mu|\sigma^2, \widetilde{Y} \sim \text{Normal}(\bar{Y}, \sigma^2/n)$$

Por otro lado, como

$$P(\sigma^2|\widetilde{Y}) = \frac{P(\mu, \sigma^2|\widetilde{Y})}{P(\mu|\sigma^2, \widetilde{Y})}$$

tendremos

$$P(\sigma^2|\widetilde{Y}) \;\; \alpha \;\; \exp\left\{\frac{1}{2\sigma^2}(n-1)S^2\right\}$$

lo que nos dice que

$$\sigma^2|\widetilde{Y} \sim \chi^{-2}((n-1), S^2)$$

Finalmente, genera muestras de la posteriori, en estos casos, es muy directo pues

$$\sigma^* \sim \chi^{-2}(n-1, S^2)$$
$$\mu^* \sim \text{Normal}(\bar{Y}, \sigma^*/n)$$

Lo que nos dice que

$$(\mu^*, \sigma^*) \sim P(\mu, \sigma^2|\widetilde{Y})$$

Por último, si nos interesa una predicción, es decir, la distribución predictiva a posteriori será

$$P(\hat{Y}|\widetilde{Y}) = \int_{\mathbb{R}^+}\int_{\mathbb{R}} P(\widetilde{Y}|\mu, \sigma^2)P(\mu, \sigma^2|\widetilde{Y})\; d\mu d\sigma^2$$

Luego vía simulación basta

- Generar $(\mu^*, \sigma^*) \sim P(\mu, \sigma^2|\widetilde{Y})$
- Generar $\widetilde{Y} \sim \text{Normal}(\mu^*, \sigma^*)$

★

**Ejercicio.** Deducir que $\hat{Y}|\widetilde{Y}$ tiene distribución $t$-student con $n-1$ grados de libertad, centrada en $\bar{Y}$ y

escalada.

Existe una priori conjugada para el caso de la distribución normal, que es llamada Normal-$\chi^{-2}$ inversa escalada

$$P(\mu, \sigma^2) = P(\mu|\sigma^2)P(\sigma^2)$$

donde

$$\mu|\sigma^2 \sim \text{Normal}(\mu_0, \sigma^2/k_0)$$
$$\sigma^2 \sim \chi^{-2}(\nu_0, \sigma_0)$$

Luego

$$P(\mu, \sigma^2) \ \propto \ (\sigma^2)^{1/2} \exp\left\{-\frac{1}{2\sigma^2}[\nu_0\sigma_0^2 + k_0(\mu - \mu_0)^2]\right\} (\sigma^2)^{-(\nu_0/2+1)}$$

y se anota

$$\mu, \sigma^2 \sim \text{N} - \chi^{-2}(\mu_0, \sigma_0^2/k_0, \nu_0, \sigma_0^2)$$

con $k_0$ : Tamaño muestral a priori.

Se puede probar que para el modelo

$$Y_1, \ldots, Y_n|\mu, \sigma^2 \overset{\text{i.i.d}}{\sim} \text{Normal}(\mu, \sigma^2)$$
$$\mu, \sigma^2 \sim \text{N} - \chi^{-2}(\mu_0, \sigma_0^2/k_0, \nu_0, \sigma_0^2)$$

se tendrá que

$$\mu, \sigma^2|\tilde{Y} \sim N - \chi^{-2}(\mu_n, \sigma_n^2/k_n, \nu_n, \sigma_n^2)$$

con

- $\mu_n = \dfrac{k_0}{k_0 + n}\mu_0 + \dfrac{n}{k_0 + n}\bar{Y}$
- $k_n = k_0 + n$
- $\nu_n\sigma_n^2 = \nu_0\sigma_0^2 + (n-1)S^2 + \dfrac{k_0 n}{k_0 + n}(\bar{Y} - \mu_0)^2$
- $\nu_n = \nu_0 + n$

Lo que corresponde a un modelo conjugado. Notemos ademas que sí $k_0 \to 0, \sigma_0^2 \to 0, \nu_0 \to -1$ tendremos que

$$P(\mu, \sigma^2) \ \propto \ \frac{1}{\sigma^2}$$

la cual corresponde a una priori no informativa.

**Ejemplo 2.12.** En 1882, Simon Newcambe diseño un experimento para medir la velocidad de la luz, en el que midió esta velocidad en una distancia de 74442 metros, obteniendo 66 observaciones aproximadamente. **(Mirar script de R)**

★

**Normal Multivariada**

Supongamos que

$$\widetilde{Y_1}, \ldots, \widetilde{Y_n} | \widetilde{\mu}, \Sigma \overset{\text{i.i.d}}{\sim} \mathrm{N}_d(\widetilde{\mu}, \Sigma)$$

con

$$\begin{aligned} P(\widetilde{Y}|\widetilde{\mu}, \Sigma) \ &\propto \ |\Sigma|^{-n/2} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} (\widetilde{Y_i} - \widetilde{\mu})^T \Sigma^{-1} (\widetilde{Y_i} - \widetilde{\mu}) \right\} \\ &= |\Sigma|^{-n/2} \exp\left\{ -\frac{1}{2} \mathrm{T}_r(\Sigma^{-1} S_0) \right\} \end{aligned}$$

donde $S_0 = \sum_{i=1}^{n} (\widetilde{Y_i} - \widetilde{\mu})^T (\widetilde{Y_i} - \widetilde{\mu})$. Supongamos que $\Sigma$ es conocida, entonces la siguiente es una priori conjugada

$$\widetilde{\mu} \sim \mathrm{N}_d(\mu_0, \Delta_0)$$

Para este caso se prueba

$$\widetilde{\mu} | \widetilde{Y} \sim \mathrm{N}_d(\widetilde{\mu}_n, \Delta_n)$$

donde

- $\widetilde{\mu}_n = (\Delta_0^{-1} + n\Sigma^{-1})^{-1} [\Delta_0^{-1} \widetilde{\mu_0} + n\Sigma^{-1} \bar{\mathbf{Y}}]$.
- $\Delta_n^{-1} = \Delta_0^{-1} + n\Sigma^{-1}$

**Ejemplo 2.13.** En el caso de que $n = 1$ y $\Sigma$ conocido, notemos que

$$\mathbb{E}(\widetilde{Y_1}) = \mathbb{E}(\mathbb{E}(\widetilde{Y_1} | \mu_1, \Sigma)) = \mathbb{E}(\widetilde{\mu}) = \mu_0.$$

Además

$$\mathrm{Var}(\widetilde{Y_1}) = \mathbb{E}(\mathrm{Var}(\widetilde{Y_1} | \widetilde{\mu}, \Sigma)) + \mathrm{Var}(\mathbb{E}(\widetilde{Y_1} | \widetilde{\mu}, \Sigma)) = \Sigma + \Delta_0$$

Luego, si consideramos una priori no informativa

$$P(\widetilde{\mu}) \ \propto \ 1$$

se tendrá

$$\widetilde{\mu} | \widetilde{Y} \sim \mathrm{N}_d(\bar{\mathbf{Y}}, \Sigma / n)$$

★

En el caso en que ambos $\widetilde{\mu}$ y $\Sigma$ sean desconocidos existe una priori conjugada. Esto se basa, en un principio, en la distribución Whishart-inversa

$$\widetilde{\alpha}_1, \ldots, \widetilde{\alpha_\nu} \overset{\text{i.i.d}}{\sim} \text{N}_d(0, S)$$

entonces la distribución de $\sum_{i=1}^{\nu} \widetilde{\alpha}_i \widetilde{\alpha}_i^T$ matriz $d$x$d$ se llama Whishart con $\nu$ grados de libertad y matriz de escala $S$, siempre que $\nu \geqslant d$. En general, decimos que $W_{(\nu-d-1)} \sim \text{Wishart}_\nu(S)$ si

$$P(w) \ \propto \ |w|^{(\nu-d-1)} \exp\left\{-\frac{1}{2}\text{Tr}(S^{-1}W)\right\}$$

con $W$ definida positiva y $\mathbb{E} = \nu S$.

**Observación.** La Whishart es conjugada para la matriz de precisión de una normal multivariada. Si además $W^{-1} \sim \text{Wishart}_\nu(S)$ decimos

$$W \sim \text{Wishart-Inversa}_\nu(S^{-1}) \text{ ó } IW(\nu, S^{-1})$$

y la densidad tiene la forma

$$P(w) \ \propto \ |w|^{-(\nu+d+1)/2} \exp\left\{-\frac{1}{2}\text{Tr}(SW^{-1})\right\}$$

con esperanza $E(W) = (\nu - d - 1)S^{-1}$ si $\nu \geqslant d + 2$.

Si ambos valores son desconocidos, una priori conjugada es

$$\begin{aligned} P(\widetilde{\mu}, \Sigma) &= P(\widetilde{\mu}|\Sigma)P(\Sigma) \\ \widetilde{\mu}|\Sigma &\sim \text{N}_d(\widetilde{\mu_0}, \Sigma|k_0) \\ \Sigma &\sim IW(\nu_0, \Delta_0^{-1}) \end{aligned}$$

Realizando los calculos se tiene que

$$\widetilde{\mu}, \Sigma \overset{\text{i.i.d}}{\sim} \text{Normal-}IW(\widetilde{\mu_n}, \Delta_n|k_n, \nu_n, \Delta_n)$$

con

- $\mu_n = \dfrac{k_0}{k_0+n}\widetilde{\mu_0} + \dfrac{n}{k_0+n}\bar{\mathbf{Y}}$.
- $k_n = k_0 + n$.
- $\nu_n = \nu_0 + n$.
- $\Delta_n = \Delta_0 + S\dfrac{k_0 n}{k_0+n}(\bar{\mathbf{Y}} - \widetilde{\mu}_0)(\bar{\mathbf{Y}} - \widetilde{\mu}_0)^T$ con $S = \sum_{i=1}^{n}(\bar{\mathbf{Y}} - \widetilde{Y}_i)(\bar{\mathbf{Y}} - \widetilde{Y}_i)$.

Una versión no informativa para $(\widetilde{\mu}, \Sigma)$ sería

$$P(\widetilde{\mu}, \Sigma) \ \propto \ |\Sigma|^{-(d+1)/2}$$

Con posteriori

$$\widetilde{\mu}, \Sigma|\widetilde{Y} \ \propto \ \text{Normal} - IW(\widetilde{\bar{\mathbf{Y}}}, S|n, n-1, S)$$

Además

$$\widetilde{\mu}|\widetilde{\Sigma}, \widetilde{Y} \sim \mathrm{N}_d(\widetilde{\bar{\mathbf{Y}}}, \Sigma/n).$$

**Ejercicio.** Encuentre $\Sigma|\widetilde{Y}$.

**Ejemplo 2.14.** Se disponde de 20 animales sometidos a diferentes dosis de cierto compuesto químico. Cinco animales fueron asignados a cada una de las 4 dosis y se observó el número de muertes. Los datos son:

| $\log$(Dosis):$x_i$ | $n_i$ | $y_i$ |
|---|---|---|
| -0,86 | 5 | 0 |
| -0,30 | 5 | 1 |
| -0,05 | 5 | 3 |
| 0,73 | 5 | 5 |

Queremos modelar la probablidad de que se produzca muerte para las diferentes dosis. ¿Cómo se construye el modelo?. Los datos son $(x_i, n_i, y_i)$. Debemos modelar el efecto-dosis sobre el número de muertes. Podemos decir que el número de muertes se comporta como

$$y_i|\theta_1 \sim \text{Bin}(n_i, \theta_i)$$

donde

$$\theta_i: \text{ Probabilidad asociada a la i-ésima dosis.}$$

La priori para $\theta_i$ no tiene por qué ser permutable, ya que se espera diferencias con respecto a la dosis(se debe reflejar a priori). Si no se estaría ignorando una característica importante del problema. Luego, un modelo Dosis-Respuesta podría ser el de regresión logística:

$$\text{Logit}(\theta_i) = \log\left(\frac{\theta_i}{1-\theta_i}\right) = \alpha + \beta x_i \Rightarrow \theta_i = \frac{\exp(\alpha+\beta x_i)}{1+\exp(\alpha+\beta x_i)}.$$

Ahora bien, la función de verosimilitud asociada es:

$$\begin{aligned} P(\widetilde{Y}|\alpha,\beta) &= \prod_{i=1}^{4} P(y_i|\alpha,\beta) \\ &\propto \prod_{i=1}^{4} \left[\frac{\exp(\alpha+\beta x_i)}{1+\exp(\alpha+\beta x_i)}\right]^{y_i} \left(\frac{1}{1+\exp(\alpha+\beta x_i)}\right)^{n_i-y_i} \\ &\propto \exp\left\{\sum_{i=1}^{4}(\alpha+\beta x_i) - n_i \log\left(\sum_{i=1}^{4}[1+\exp(\alpha+\beta x_i)]\right)\right\} \end{aligned}$$

Ahora bien, tomaremos dos prioris, una no informativa y otra vaga. Veamos que sucede:

1. $P(\alpha,\beta) \propto 1$.
2. $(\alpha,\beta) \sim \mathrm{N}_2\left(\begin{pmatrix}0\\0\end{pmatrix}, 10^4\mathcal{I}_2\right)$

Existe un interés específico en el el llamado LD50, que corresponde a la dosis que conduce a una mortalidad del $50\,\%$, esto es, un valor $\bar{x}$ tal que

$$\log\left(\frac{0{,}5}{1-0{,}5}\right) = \alpha+\beta\bar{x} \Rightarrow \bar{x} = -\frac{\alpha}{\beta}$$

De aquí, podemos graficar la distribución mediante los valores de $\alpha$ y $\beta$ obtenidos antes.

**Ejercicio.** Verificar que las posteriori son propias.

En el primer caso se tiene que

$$\begin{aligned} P(\alpha,\beta|\widetilde{Y}) &\propto P(\widetilde{Y}|\alpha,\beta)P(\alpha,\beta) \\ &\propto P(\widetilde{Y}|\alpha,\beta) \end{aligned}$$

Mientras que en el segundo caso

$$P(\alpha,\beta|\widetilde{Y}) \propto P(\widetilde{Y}|\alpha,\beta)\exp\left\{-\frac{\alpha^2}{2\cdot 10^4} - \frac{\beta^2}{2\cdot 10^4}\right\}$$

★

**Primera forma de explorar** $P(\alpha,\beta|\widetilde{Y})$:

- Ubicar donde esta la 'moda' ó el estimador $(\hat{\alpha},\hat{\beta})$ (Usar la función GLM de R).
- Se exploran rangos donde se concentra lo más relevante y se propone una malla de puntos.
- Luego, para cada $\alpha_i,\beta_j$ en una malla bidimensional se evalúa $\log P(\alpha_i,\beta_j|\widetilde{Y})$.

- Para generar muestras de $(\alpha_i^*, \beta_j^*)$ de $P(\alpha, \beta|\widetilde{Y}) \propto P(\widetilde{Y}|\alpha, \beta)P(\alpha, \beta)$ usaremos la representación

$$P(\alpha, \beta|\widetilde{Y}) \propto P(\widetilde{Y}|\alpha, \beta)P(\alpha, \beta)$$

- Con los valores de la malla podemos construir una aproximación de $P(\beta|\widetilde{Y})$ simplemente sumando sobre los $\alpha_i$'s para cada $\beta_j$ fijo (Sumar por filas).
- Generar $\beta_j^*$ de la aproximación de $P(\beta|\widetilde{Y})$ evaluando en la malla correspondiente en $\beta$.
- Fijando $\beta_j^*$ generar muestras para $\alpha_i^*$ de $P(\alpha_i|\beta_j^*, \widetilde{Y})$.

**Segunda forma de explorar** $P(\alpha, \beta|\widetilde{Y})$**:** En este caso, buscaremos aproximaciones normales a la distribución a posteriori. Para esto tenemos dos alternativas:

1. Aproximar tanto la distribución apriori como la verosimilitud a distribuciones normales y combinarlas para formar la distribución a posteriori.
2. Aproximar directamente la distribución a posteriori.

Esta aproximación es apropiada para casos donde la distribución a posteriori es unimodal y relativamente simétrica.

**Observación.** La primera alternativa tiene el problema de que la Normal obtenida puede verse afectada por los diferentes pesos que adquieren la verosimilitud y la priori, lo que puede afectar las conclusiones.

Se puede construir una aproximación Normal a la distribución a posteriori usando la expansión de Taylor:

$$\log(P(\theta|\widetilde{Y}) = \log(P(\hat{\theta}|\widetilde{Y}) + \frac{1}{2}(\theta - \hat{\theta})^T \left[ \frac{\partial^2 \log(P(\theta|\widetilde{Y})}{\partial \theta^2} \right] |_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \dots$$

donde $\hat{\theta}$ es la moda a posteriori, es decir

$$\hat{\theta} = \text{argmax}_{\theta \in \Theta} P(\theta|\widetilde{Y})$$

Llamando $I(\theta)$ a la información observada

$$I(\theta) = -\frac{\partial^2 \log(P(\theta)|\widetilde{Y})}{\partial \theta^2}$$

Luego, de la expansión de Taylor obtenemos

$$P(\theta|\widetilde{Y}) \sim \text{Normal}\left(\hat{\theta}, I(\hat{\theta})^{-1}\right).$$

En R se puede usar

- nlm. $\rightarrow \hat{\theta}$.
- optim. $\rightarrow I(\hat{\theta})$.

**Ejemplo 2.15.** Veamos en R el segundo método aplicado. Sabemos que para $n$ grande,

$$P(\theta|Y) \dot{\sim} \mathrm{N}(\hat{\theta}, I^{-1}(\hat{\theta})$$

donde $I$ corresponde a la información de Fisher observada

$$I(\theta) = -\frac{\partial^2 \log(P(\theta|Y))}{\partial\theta\partial\theta^T}$$

Consideremos

$$Y_1, \ldots, Y_n|\mu, \sigma^2 \overset{\text{i.i.d}}{\sim} \mathrm{N}(\mu, \sigma^2)$$

con ambos parámetros desconocidos. Sea

$$P(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$$

es decir

$$P(\mu, \phi) \propto 1$$

con $\phi = \log(\sigma)$. De aquí

$$\log(P(\mu, \log(\sigma)|\widetilde{Y}) \propto -n\log(\sigma) - \frac{1}{2\sigma^2}[(n-1)S^2 + n(\bar{Y} - \mu)^2]$$

★

**Para el laboratorio, se preguntará aproximación Normal de la priori y la verosimilitud.**

### 2.5.3. Cota en los gráficos de contornos.

Notemos que

$$\log(P(\theta|\widetilde{Y}) = K - \frac{1}{2}\chi^2_d$$

donde $K$ es una constante, $d$ la dimensión del espacio paramétrico. Entonces aproximadamente el $95\,\%$ de la masa de la distribución se encuentra en valores de $\theta$ tal que

$$P(\theta|\widetilde{Y}) \geqslant \exp\{-\chi^2_{d,0,95}/2\}P(\hat{\theta}|\widetilde{Y})$$

Para $d = 2$ se tiene

$$P(\theta|\widetilde{Y}) \geqslant \exp\{-5,99/2\}P(\hat{\theta}|\widetilde{Y}) \approx 0,05P(\hat{\theta}|\widetilde{Y})$$

## 2.6 Métodos MCMC(Markov-chain, Monte Carlo).

1. Corresponden a métodos de simulación que extraen valores desde distribuciones aproximadas de la distribución objetivo.

2. Estructura de muestreo secuencial y la distribución de la última observación unicamente depende del valor de la observación inmediatamente anterior, es decir

$$P(\theta_{t+1}|\theta_1, \ldots, \theta_t) = P(\theta_{t+1}|\theta_t).$$

   Nuestro objetivo será crear un proceso de Markov cuya distribución estacionaria sea $P(\theta|Y)$.

**Ejemplo 2.16.** Espacio de estados de la cadena $S = \{1, 2, 3, 4\}$ y las probabilidades de pasar de un estado a otro llamadas probabilidades de transición.

★

**Definición 2.12.** Diremos que una cadena de Markov es:

- Irreducible si es posible llegar a cualquier estado desde cualquier otro.
- Aperiódica si el retorno a cada paso puede ocurrir en cualquier número de pasos y no únicamente en múltiplos de $k$ pasos, para algún $k \neq 1$.
- Recurrente positiva si el valor esperado del número de pasos para volver de un estado a sí mismo es finito para todos los estados.

**Observación.** Si la cadena es irreducible, aperiódica y recurrente positiva, $\theta_t$ converge a la distribución estacionaria.

### 2.6.1. Metrópolis-Hostings

Dado un cierto valor $r$, digamos $\widetilde{\theta}_{old}$, se propone un candidato $\widetilde{\theta}^*$ de acuerdo a $q(\widetilde{\theta}_{old}, \widetilde{\theta}^*)$, donde $q(\cdot, \cdot)$ es la distribución de propuestas, que en general es una distribución de probabilidad. Este candidato se acepta con probabilidad

$$\alpha = \text{mín}\left\{1, \frac{\pi(\widetilde{\theta}^*)q(\widetilde{\theta}_{old}, \widetilde{\theta}^*)}{\pi(\widetilde{\theta}_{old})q(\widetilde{\theta}^*, \widetilde{\theta}_{old})}\right\}$$

donde $\pi(\theta)$ es nuestra distribución objetivo. En la práctica generamos $U \sim \text{U}(0, 1)$. Si

$$U < \alpha \Rightarrow \text{ Acepto } \widetilde{\theta}^*$$

el caso contrario, me quedo con $\widetilde{\theta}_{old}$. Por tanto

$$\widetilde{\theta}_{new} = \begin{cases} \widetilde{\theta}^* & \text{si acepto} \\ \widetilde{\theta}_{old} & \text{si rechazo} \end{cases}.$$

**Ejemplo 2.17.**

$$q(\widetilde{\theta}_{old}, \widetilde{\theta}^*) \cong \text{N}(\widetilde{\theta}^*; \widetilde{\theta}_{old}, \Sigma)$$

donde $\widetilde{\theta}^*$ corresponde al argumento, $\widetilde{\theta}_{old}$ la media y $\Sigma$ la varianza(que tan lejos voy a llegar).

★

De lo anterior, la cadena de Markov se genera

- Tomar $\widetilde{\theta}^{(0)}$ inicial fijo (i = 1).
- Definir $\widetilde{\theta}_{old} = \widetilde{\theta}^{(i-1)}$, para luego redefinir $i = i + 1$ y repetir el algoritmo hasta que se cumpla el criterio deparado.

**Observación.** 1. Si $q(\widetilde{\theta}_{old}, \widetilde{\theta}^*)$ es tal que

$$q(\widetilde{\theta}_{old}, \widetilde{\theta}^*) = q(\theta^*)$$

entonces el valor propuesto es independiente del valor actual. Esta variación es llamada Metrópolis-Independence Sampler.

2. Si $q(\widetilde{\theta}_{old}, \widetilde{\theta}^*) = q(\widetilde{\theta}^*, \widetilde{\theta}_{old})$ es decir, $q(\cdot, \cdot)$ es simétrico en los argumentos, entonces

$$\alpha = \text{mín}\left\{1, \frac{\pi(\widetilde{\theta}^*)}{\pi(\widetilde{\theta}_{old})}\right\}$$

Supongamos que $\widetilde{\theta} = (\theta_1, \theta_2, \theta_3)$. Nos gustaría que las probabilidades

$$P(\theta_1|\theta_2, \theta_3, \widetilde{Y}), \qquad P(\theta_2|\theta_1, \theta_3, \widetilde{Y}), \qquad P(\theta_3|\theta_1, \theta_2, \widetilde{Y})$$

tuviesen formas cerradas, para que de alguna forma $\alpha = 1$. Gelman y Smith proponen una versión especial del M-A, llamado muestreo de Gibbs, el cual tiene la siguiente formulación:

Si $\widetilde{\theta} = (\theta_1, \ldots, \theta_d)$ entonces realizamos

- Tomar $\widetilde{\theta}^{(0)} = (\widetilde{\theta}_1^{(0)}, \ldots, \widetilde{\theta}_d^{(0)})$ y definir $i = 1$.

- Generar $\widetilde{\theta}^{(i)}$ usando el siguiente procedimiento:

$$\begin{aligned}
\theta_1^{(i)} &\sim P(\theta_1|\theta_2^{(i-1)},\ldots,\theta_d^{(i-1)},\widetilde{Y})\\
\theta_2^{(i)} &\sim P(\theta_2|\theta_2^{(i)},\ldots,\theta_d^{(i-1)},\widetilde{Y})\\
\theta_3^{(i)} &\sim P(\theta_3|\theta_1^{(i)},\theta_2^{(i)},\ldots,\theta_d^{(i-1)},\widetilde{Y})\\
\vdots \quad &\quad \vdots\\
\theta_d^{(i)} &\sim P(\theta_d|\theta_1^{(i)},\theta_2^{(i)},\ldots,\theta_{d-1}^{(i)},\widetilde{Y})
\end{aligned}$$

  Y luego redefinir $i = i + 1$.

**Observación.** Si alguna(varias o todas) de las distribuciones condicionales son difíciles de muestrar, se puede uar una M-H para generar candidatos para esa coordenada en particular (distribución). Esto nos dice que deberemos usar Metrópolis dentro de Gibbs.

**Ejemplo 2.18.** 1. Consideremos

$$Y_i|\alpha,\beta,\sigma^2 \overset{\text{i.i.d}}{\sim} \text{N}(\alpha+\beta t_i,\sigma^2)$$

donde $\alpha,\beta,\sigma^2$ son independientes y además

$$\begin{aligned}
\alpha &\sim \text{N}(\alpha_0,T_a^2)\\
\beta &\sim \text{N}(\beta_0,T_b^2)\\
\sigma^2 &\sim \chi^{-2}(V_0,\sigma_0^2
\end{aligned}$$

donde $\alpha_0,\beta_0,T_a^2,T_b^2,V_0,\sigma_0^2>0$ son conocidos. De aquí

$$\begin{aligned}
P(\alpha,\beta,\sigma^2|Y) &\propto P(Y|\alpha,\beta,\sigma^2)P(\alpha,\beta,\sigma^2)\\
&\propto P(Y|\alpha,\beta,\sigma^2)P(\alpha)P(\beta)P(\sigma^2)
\end{aligned}$$

Para realizar Metrópolis-Hostings, es necesario conocer las distribuciones aposteriori para cada parámetro. En este caso, dichas distribuciones son conocidas:

$$\begin{aligned}
\alpha|\beta,\sigma^2,Y &\sim \text{N}(\alpha_n,T_{a_n}^2)\\
\beta|\alpha,\sigma^2,Y &\sim \text{N}(\beta_n,T_{b_n}^2)\\
\sigma^2|\alpha,\beta,Y &\sim \chi^{-2}(\nu_n,\sigma_n^2)
\end{aligned}$$

$$\alpha_n = \frac{\alpha_0 \big/ T_a^2 + \sum y_i \big/ \sigma^2 - \beta \frac{\sum t_i}{\sigma^2}}{1 \big/ T_a^2 + n \big/ \sigma^2}, \qquad \beta_n = \frac{\beta_0 \big/ T_b^2 - \alpha \sum t_i \big/ \sigma^2 + \sum y_i t_i \big/ \sigma^2}{1 \big/ T_b^2 + \sum t_i \big/ \sigma^2}$$

$$T_{a_n}^2 = \frac{1}{\frac{1}{T_a^2} + \frac{n}{\sigma^2}}, \qquad T_{b_n} = \frac{1}{\frac{1}{T_b^2} + \frac{\sum t_i}{\sigma^2}}.$$

2. Volviendo al ejemplo del Bioensayo, podemos tomar

$$q(\widetilde{\theta}) = \mathrm{N}_2(\hat{\theta}, I(\hat{\theta})^{-1}$$

con punto de partida

$$(\alpha^{(0)}, \beta^{(0)}) = (0{,}85, 7{,}75).$$

★

### 2.6.2. Estudio de Convergencia

- Vamos a generar varias cadenas, utilizando diferentes puntos de partida cada vez.
- Para cada una de ellos, eliminar la primera mitad de los valores generados (Quema de datos).

En BPA 3, $\hat{R}$ corresponde al estadístico de Gelman y Rubin. Este nos indica en que momento una cadena alcanza la convergencia. De hecho, si $\hat{R} \approx 1$ diremos que la cadena ha convergido. Llamaremos $\underline{m}$ al número de cadenas retenidas, y a $\underline{n}$ al número de observaciones dentro de cada una de ellas. Consideremos $\varphi_{ij}$ con $i = 1, \ldots, n$ y $j = 1, \ldots, m$ las cadenas correspondientes al parámetro de interés. Calcularemos la siguiente expresión:

$$B = \frac{n}{m-1} \sum_{j=1}^{n} (\bar{\varphi}_{\cdot j} - \bar{\varphi}_{\cdot\cdot})^2$$

con

$$\bar{\varphi}_{\cdot j} = \frac{1}{n} \sum_{i=1}^{n} \varphi_{ij}, \qquad \bar{\varphi}_{\cdot\cdot} = \frac{1}{m} \sum_{j=1}^{m} \bar{\varphi}_{\cdot j}.$$

Consideramos entonces

$$W = \frac{1}{m} \sum_{j=1}^{m} S_j^2$$

donde $S_j^2 = \dfrac{1}{n-1} \displaystyle\sum_{i=1}^{n} (\varphi_{ij} - \bar{\varphi}_{\cdot j})^2$. Notemos que $W$ se puede interpretar como la varianza dentro de las cadenas y $B$ la varianza entre las cadenas. Finalmente, el estadístico $\hat{R}$ vendrá dado por

$$\hat{R} = \sqrt{\frac{\hat{\mathrm{Var}}(\phi|Y)}{W}}$$

con $\hat{\text{Var}}(\phi|Y) = \frac{n-1}{n}W + \frac{1}{m}B$. Si $\hat{R}$ es mucho mayor a 1, se sugiere continuar la simulación.

## 2.7 Modelos Jerárquicos.

En esta sección estudiaremos problemas del siguiente estilo:

**Ejemplo 2.19.** Suponga $J$ estudios llevados a cabo en $J$ hospitales diferentes, donde se desea inferir sobre $\theta$, la probabilidad de los pacientes a sobrevivir luego de un tratamiento. Los hispitales, ¿Comparten un único parámetro $\theta$?.

★

El modelo que hemosw estudiado tiene la siguiente representación

$$\begin{aligned} Y_1, \ldots, Y_n|\theta &\sim P(Y|\theta) \\ \theta &\sim P(\theta) \end{aligned}$$

donde cada $Y_i$ posee como parámetro el mismo $\theta$. En modelos jerárquicos, la representación es:

$$\begin{array}{ccc} Y_1 & \cdots & Y_n \\ \downarrow & \cdots & \downarrow \\ \theta_1 & \cdots & \theta_n \end{array}$$

Bajo el modelo jerárquico, la idea es permitir parámetros individuales para las distintas observaciones. La pregunta que nos asalta es que distribución a priori escoger, es decir, $P(\theta_1, \ldots, \theta_n) = ?$. Lo natural es asumir un modelo permutable o intercambiable para $\theta_1, \ldots, \theta_n$. En dicho caso, por el teorema de DeFinetti propone:

$$P(\theta_1, \ldots, \theta_n) = \int_\Theta \prod_{i=1}^{n} P(\theta_i|\phi)P(\phi)\, d\phi$$

donde $\phi$ es un hiperparámetro. Luego, obtenemos la siguiente representación

$$\begin{aligned} \theta_1, \ldots, \theta_n &\overset{\text{i.i.d}}{\sim} P(\theta|\phi) \\ \phi &\sim P(\phi) \end{aligned}$$

Luego, el modelo jerárquico se representa como

$$\begin{aligned} Y_1, \ldots, Y_n | \theta_1, \ldots, \theta_n &\sim P(Y_i|\theta_i) \\ \theta_1, \ldots, \theta_n &\overset{\text{i.i.d}}{\sim} P(\theta|\phi) \\ \phi &\sim P(\phi) \end{aligned}$$

donde $Y_i|\theta_i$ son independientes. Gráficamente se tendrá

$$\begin{array}{ccccc} Y_1 & \cdots & Y_n \\ \downarrow & \cdots & \downarrow \\ \theta_1 & \cdots & \theta_n \\ \searrow & & \swarrow \\ & \phi & \end{array}$$

Lo cual implica que el modelo completo se puede expresar como

$$P(Y, \theta, \phi) = P(Y|\theta, \phi)P(\theta|\phi)P(\phi) = \left(\prod_{i=1}^{n} P(y_i|\theta_i)\right)\left(\prod_{i=1}^{n} P(\theta_i|\phi)\right)P(\phi).$$

**Ejemplo 2.20.** En un experimento llevado a cabo en un laboratorio, se observó la presencia de tumores en roedores. Al finalizar el período, 4 de 14 habían desarrollado un tumor. El Modeolo Bayesiano Clásico plantea

$$a$$

La pregunta que asalta es como escojo $\alpha$ y $\beta$. Notemos que por nuestro supuesto se tiene que

$$\mathbb{E}(\theta) = \frac{\alpha}{\alpha+\beta} \qquad \text{Var}(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Una manera es mirar información de experimentos semejantes. Supongamos que tenemos información de 70 experimentos anteriores, realizados en 70 laboratorios distintos. En ellos se obtuvo valores $\hat{\theta}_j = \frac{y_j}{n_j},\ j = 1, \ldots, 70$. De ellos sabemos que

$$\frac{1}{70}\sum_{j=1}^{70} \hat{\theta}_j = 0{,}136, \qquad \text{SD}(\hat{\theta}_j) = 0{,}103$$

Luego, podemos estimar mediante el método de los momentos, es decir,

$$\mathbb{E}(\theta) = \frac{\alpha}{\alpha+\beta} = 0{,}136 \qquad \text{Var}(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = 0{,}103.$$

De aquí, resolviendo el sistema obtenemos que

$$(\alpha, \beta) = (1{,}4, 8{,}6)$$

y por ende

$$\theta|Y \sim \text{Beta}(5{,}4, 18, 6)$$

Por otro lado, el modelo jerárquico considera un modelo para todos estos experimentos conjuntamente. Esto implica

$$\begin{aligned} Y_j|\theta_j &\overset{\text{i.i.d}}{\sim} \text{Bin}(n_j, \theta_j) \\ \theta_1, \ldots, \theta_n|\alpha, \beta &\overset{\text{i.i.d}}{\sim} \text{Beta}(\alpha, \beta) \\ (\alpha, \beta) &\sim P(\alpha, \beta) \end{aligned}$$

Luego, el modelo completo vendrá dado por

$$P(\widetilde{Y}, \widetilde{\theta}, \alpha, \beta) = P(\widetilde{Y}|\widetilde{\theta})P(\widetilde{\theta}|\alpha, \beta)P(\alpha, \beta) \propto \prod_{j=1}^{n} \theta_j^{y_j}(1-\theta_j)^{n_j - y_j} \cdot \prod_{j=1}^{n} \theta_j^{\alpha-1}(1-\theta_j)^{\beta-1} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot P(\alpha, \beta).$$

Luego, en virtud del teorema de Bayes se tendrá

$$P(\widetilde{\theta}, \alpha, \beta|\widetilde{Y}) \propto P(\alpha, \beta) \prod_{j=1}^{n} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_j^{\alpha+y_j-1}(1-\theta_j)^{\beta+n_j-y_j-1}$$

Por ende se tendrá

$$P(\widetilde{\theta}|\alpha, \beta, \widetilde{Y}) \propto \prod_{j=1}^{n} \theta_j^{\alpha+y_j-1}(1-\theta_j)^{\beta+n_j-y_j-1}$$

Dado que $\alpha, \beta$ con $\theta_j$ son independientes se tendrá que

$$\theta_j|\alpha, \beta, \widetilde{Y} \sim \text{Beta}(\alpha + y_j, \beta + n_j - y_j)$$

Por lo tanto, una a posteriori de nuestros hiperparámetros sería

$$P(\alpha, \beta|\widetilde{Y}) = \frac{P(\widetilde{\theta}, \alpha, \beta|\widetilde{Y})}{P(\widetilde{\theta}|\alpha, \beta, \widetilde{Y})}$$

★

Para simular un modelo jerárquico realizaremos los siguientes pasos:

1. Muestrear valores $\phi^{(s)} \sim P(\phi|Y), s = 1, \ldots, S$.

2. Para cada valor de $s$ muestrear $\theta_j^{(s)} \sim P(\theta_j|\phi^{(s)}, \widetilde{Y}), j = 1, \ldots, n$.

3. Muestrear $\widetilde{Y} \sim P(\widetilde{Y}|\widetilde{\theta})$, donde $\widetilde{\theta}$ puede corresponder a un valor de $\theta$ ya existente, o a una nueva observación proveniente de $P(\widetilde{\theta}|\phi)$.

**Ejemplo 2.21** (Ejemplo tumores en las ratas)**.** Consideremos el siguiente modelo

$$\begin{aligned} Y_i|\theta_i &\overset{\text{i.i.d}}{\sim} \text{Bin}(n_i, \theta_i) \\ \theta_1, \ldots, \theta_n &\overset{\text{i.i.d}}{\sim} \text{Beta}(\alpha, \beta) \\ (\alpha, \beta) &\sim P(\alpha, \beta) \end{aligned}$$

El modelo completo esta dado por

$$P(Y, \theta, \alpha, \beta) \propto \prod_{i=1}^{n} \theta_i^{y_i}(1-\theta_i)^{n_i - y_i} \times \prod_{i=1}^{n} \theta_i^{\alpha-1}(1-\theta_i)^{\beta-1} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \times P(\alpha, \beta).$$

La posteriori conjunta será

$$P(\theta, \alpha, \beta|Y) \propto P(\alpha, \beta) \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \prod_{i=1}^{n} \theta_i^{\alpha+y_i-1}(1-\theta_i)^{\beta+n_i-y_i-1}$$

y las posterioris para los hiperparámetros son

$$\begin{aligned} P(\alpha, \beta|Y) &= \frac{P(\theta, \alpha, \beta|Y)}{P(\theta|\alpha, \beta, Y)} \\ &\propto \prod_{i=1}^{n} \left[ \frac{\Gamma(\alpha, \beta)\Gamma(\alpha+y_i)\Gamma(\beta+\eta_i-y_i)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+\eta_i)} \right] P(\alpha, \beta) \end{aligned}$$

¿Qué pasa con la idea de proponer una priori plana es $\left(\log\left(\frac{\alpha}{\alpha+\beta}\right), \log(\alpha+\beta)\right)$ En este caso, tendremos una posteriori impropia, lo cual no es útil. Por esto, en BDA sugieren considerar una priori plana

$$P\left(\frac{\alpha}{\alpha+\beta}, (\alpha+\beta)^{-1/2}\right) \propto 1 \cong \text{U}(0, 1)$$

Luego, usando Jacobiano se tendrá que

$$P(\alpha, \beta) \propto (\alpha+\beta)^{-5/2}.$$

Por otro lado, una Priori vaga será

- $(\log(\alpha), \log(\beta)) \sim \mathrm{N}_2\left(\begin{pmatrix}0\\0\end{pmatrix}, V\right)$. Por ejemplo podríamos tomar

$$V = \begin{pmatrix} 10{,}000 & 0 \\ 0 & 10{,}000 \end{pmatrix}$$

- $(\alpha, \beta) \sim P(\alpha, \beta)$. Si asumimos que $\alpha$ y $\beta$ son independientes, podemos tomar para $\varepsilon > 0$ pequeño

$$\alpha \sim \text{Gamma}(\varepsilon, \varepsilon)$$
$$\beta \sim \text{Gamma}(\varepsilon, \varepsilon)$$

★

**Ejemplo 2.22.** Datos de fallas en bombas de agua en plantas nucleares. Se registra $y_i$ : # fallas observadas, $t_i$ # tiempo de observaciones. El modelo planteado para esto es

$$\begin{aligned} Y_i|\lambda_i &\sim \text{Poisson}(\lambda_i t_i) \\ \lambda_1, \ldots, \lambda_n|\beta &\sim \text{Gamma}(\alpha, \beta),\ \alpha \text{ conocido} \\ \beta &\sim \text{Gamma}(\gamma, \delta),\ \gamma, \delta \text{ conocidos.} \end{aligned}$$

★

## 2.7 Winbugs u Openbugs

Winbugs es un programa basado en un lenguaje de programación utilizado para generar y monitorear muestras de la distribución a posteriori.

Bugs utiliza representaciones gráficas para obtener las distribuciones condicionales a posteriori necesarias para el Gibbs Sampling. Winbugs utiliza que

$$P(\theta_i|\theta_j, y) \propto P(\theta_j|\theta_j^p, y) P(\theta_j^h|\theta, y)$$

donde $\theta_j^p$ son los padres y $\theta_j^h$ hijos.

Recordemos que si $\theta = (\theta_1, \ldots, \theta_d)$ en la iteración $t$, el algoritmo realiza "d"pasos

$$\begin{array}{lll} \theta_1^{t-1} \longrightarrow \theta_1^t & & \\ \vdots\vdots & & \vdots \\ \theta_d^{t-1} \longrightarrow & & \theta_d^t \end{array}$$

donde cada actualización, se realiza muestreando un valor $\theta_j^t$ a partir de $P(\theta_j|\theta_{-j}^{t-1}, y)$, es decir

$$\begin{aligned}
\theta_1^* &\sim P(\theta_1|\theta_2, \ldots, \theta_d, y) \\
\theta_2^* &\sim \theta_3, \ldots, \theta_d, y) \\
&\vdots \qquad\qquad\qquad\qquad\qquad\qquad \vdots
\end{aligned}$$

es decir, la función propuestas corresponde a:

$$q_{jt}(\theta^*|\theta^{t-1}) = \begin{cases} P(\theta_j^*|\theta_{-j}^{t-1}, y) & \text{si } \theta_{-j}^* = \theta_{-j}^{t-1} \\ 0 & \text{si } no \end{cases}.$$

La razón

$$\alpha = \frac{P(\theta^*|y) \diagup q_{jt}(\theta^*|\theta^{t-1})}{P(\theta^{t-1} \diagup q_{jt}(\theta^{t-1}\theta^*)}$$

corresponde a la probabilidad de aceptación del valor propuesto, que por convención vale 1. Esto quiere decir que el valor propuesto siempre es aceptado.

**Ejemplo 2.23.** 1. Modelo Binomial-Normal. Considere el modelo

$$\begin{aligned}
Y|\pi &\sim \text{Bin}(n, \pi) \\
\pi &= \frac{e^\theta}{1 - e^\theta} \\
\theta &\sim \text{Normal}(0, 10000)
\end{aligned}$$

Suponga que se observa $y = 399, n = 845$. En el programa $\theta \sim \text{Normal}(0, 10000)$ Normal parametrizada de términos de su media y la precisión $\Gamma = \frac{1}{\sigma^2}$.

2. Modelo Normal - Gamma. Considere el modelo

$$\begin{aligned}
y_i|\theta, \tau &\sim \text{N}(\theta, 1 \diagup \tau) \\
\theta &\sim \text{N}(0, 100) \\
\tau &\sim \text{Gamma}(0{,}01, 0{,}01)
\end{aligned}$$

★

CAPÍTULO 3

# Evaluación de Modelos

Para evaluar un modelo, seguiremos la siguiente secuencia de análisis:

1. Especificación de la distribución conjunta de observables y no observables.
2. Obtención de la distribución a posteriori de los parámetros.
3. Evaluación del ajuste y pertinencia del modelo.

Una revisión efectiva debe incluir lo siguiente:

1. Revisión de la estructura jerárquica del modelo.
2. Revisión de las distribuciones apriori e hiperpriori.
3. Revisión de la distribución de muestreo(Verosimilitud).
4. Revisión de variables eplicativas o excluidas del modelo.

## 3.1 Análisis de Sensibilidad

Nos gustaría estudiar la sensibilidad del modelo, esto es, responder las siguientes preguntas:

- ¿Cómo cambian las inferencias a posteriori al utilizar otros modelos razonables?.
- ¿Tienen las deficiencias del modelo un efecto sustantivo sobre las inferencias de interés?

Recordemos que: "No existe modelos verdaderos, solo existen modelos útiles".

## 3.2 Validación Externa

1. Utilizar los datos ya recolectados y revisar si ellos resultan razonables bajo la distribución predictiva a posteriori.
2. Simular muestras a partir de la distribución predictiva a posteriori. Luego comparar su comportamiento con los valores de las observaciones(Buscar ciertas diferencias).

En BDA se propone el procedimiento de réplica de lo datos: Consideremos $Y_{rep}$ como una réplica de los datos. Esta réplica corresponde a valores que pudiesen haber sido observados, o ser observables en el futuro. Compararemos dichos valores con $\hat{Y}$, los cuales son valores futuros de $Y$. Notemos que

$$P(Y_{rep}|Y) = \int_{\Theta} P(Y_{rep}|\theta)P(\theta|Y)\ d\theta$$

Luego $Y_{rep} \sim P(\hat{Y}|\theta^{(s)}$, donde $\theta^{(s)}$ es el valor actual de $\theta$.

### 3.2.1. Cantidades de Testeo

Definimos una medida de discrepancia entre el modelo y los datos como un escalar $T(y, \theta)$. Su objetivo es comparar el comportamiento de las observaciones y sus réplicas.

**Ejemplo 3.1.**

$$T(Y, \theta) = \text{mín}\{y_i\} \qquad T(Y, \theta) = \text{máx}\{y_i\} - \text{mín}\{y_i\}$$

★

**Observación.** Recordemos que el valor-$p$, dado el estadístico $T(Y)$ corresponde a

$$\text{valor-}p = P_\theta(T(Y_{rep}) \geqslant T(Y)).$$

En el caso de la estadística Bayesiana, dada la cantidad de testeo $T(Y, \theta)$ el valor-$p$ vendrá dado por

$$\text{valor-}p = P_\theta(T(Y_{rep}, \theta) \geqslant T(Y, \theta)|Y))$$

**Ejemplo 3.2.** En 1882, Simon Newcombe diseñó un experimento para medir la velocidad de la luz. En ellos midió esta velocidad en una distancia de $7,442$ metros obteniendo 66 observaciones. **Véase Script de R** Concluimos que la distribución sugerida no es correcta pues el mínimo observado no corresponde con las obtenidas con la repeticiones. Por esto, otras distribuciones de muestreo se tendrá que:

- Distribución asimétrica.
- Distribución simétrica con colas largas.

★

CAPÍTULO 4

# Regresión lineal

Consideremos las variables respuestas $y_1, \ldots, y_n$ explicadas por los vectores predictores o covariables $(x_1, \ldots, x_n)$. Asumiremos el supuesto de permutabilidad, es decir, que el conjunto de vectores $(y_1, x_1), \ldots, (y_n, x_n)$ son intercambiables. Nuestro interés será estudiar $Y|X, \theta$, para lo cual consideraremos un modelo conjunto en la forma

$$P(\widetilde{Y}, \widetilde{X}, \theta, \phi).$$

Bajo el supuesto de independencia condicional se tendrá que

$$P(\widetilde{Y}, \widetilde{X}, \theta, \phi) = P(\widetilde{Y}|\widetilde{X}, \theta)P(\widetilde{X}|\theta)P(\theta)P(\phi)$$

Por lo cual, la distribución a posteriori será

$$P(\theta, \phi|\widetilde{Y}, \widetilde{X}) = P(\theta|\widetilde{Y}, \widetilde{X})P(\phi|\widetilde{X})$$

lo que implica que podemos tratar $P(\theta|\widetilde{Y}, \widetilde{X})$ separadamente de $P(\phi|\widetilde{?})$.

### 4.0.1. Selección de Predictores

La selección de la matriz de diseño asociada a la regresión sufre esencialmente los mismos problemas que en el enfoque clásico. La formulación del modelo clásico de regresión es

$$Y|X, \beta, \sigma^2 \sim \mathrm{N}_n(X\beta, \sigma^2 I)$$

donde

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \qquad X_{n\times k} = \begin{pmatrix} 1 & x_1 & \cdots & x_k \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n_1} & \cdots & x_{n_k} \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

Por ejemplo, en el caso de los modelos ANOVA, $X$ tiene la forma

$$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 1 & 0 & \cdots & \vdots \\ 0 & 1 & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

Aquí, $\beta_1, \ldots, \beta_k$ representaran el efecto de los tratamientos.

Para el estudio, tenemos dos posibilidades para elegir la priori:

- $P(\beta, \sigma^2) \propto \sigma^{-2}$, que es una priori no informativa estándar.
- Si hay información a priori, entonces se pueden seleccionar distribuciones de la forma

$$\begin{aligned} \beta &\sim \mathrm{N}_n(\beta_0, V_0) \\ \sigma^2 &\sim \chi^{-2}(\nu_0, \sigma_0^2) \end{aligned}$$

Con el fin de estudiar lo anterior, se puede utilizar Gibbs Sampling.

Bajo la priori no informativa, el análisis se basa en que

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

Luego, si llamamos $V_p = (X^T X)^{-1}$ se tendrá que

$$\begin{aligned} \beta|\sigma^2, Y &\sim \mathrm{N}(\hat{\beta}, \sigma^2 V_p) \\ \sigma^2|Y &\sim \chi^{-2}(n-k, S^2) \end{aligned}$$

donde $S^2 = \dfrac{SCE}{n-k}$. Luego, recordando que

$$P(\beta, \sigma^2|Y) = P(\beta|\sigma^2, Y) P(\sigma^2|Y)$$

podremos generar muestras de $P(\beta, \sigma^2|Y)$ mediante

1. Generar $(\sigma^2)^* \sim \chi^{-2}(n-k, S^2)$.
2. Generar $\beta^* \sim \mathrm{N}(\hat{\beta}, (\sigma^2)^* V_p)$.

## 4.1 Modelo de regresión lineal simple

### 4.1.1. Función de verosimilitud

Considerar las observaciones permutables dados los valores $x_i$

$$y_i|\beta_0, \beta_1, \sigma^2 \sim \mathrm{N}(\beta_0 + \beta_1 x_i, \sigma^2)$$

Si hay distribución a priori, se puede seleccionar distribuciones de la forma

$$\beta \sim \mathrm{N}_k(\tilde{\beta}_0, V_0)$$
$$\sigma^2 \sim \chi^{-2}(\nu_0, \sigma_0^2)$$

En este caos para simular utiliamos Gibbs Sampling.

**Distribución a priori no informativa**

Considere el parámetro $(\beta_0, \beta_1, \sigma^2)$ con función de distribución a priori

$$P(\beta_0, \beta_1, \sigma^2) \propto \frac{1}{\sigma^2}$$

con $\beta_0, \beta_1 \in \mathbb{R}$ y $\sigma^2 \in \mathbb{R}^+$. En este caso, las condicionales a posteriori quedan:

$$\beta_0|\beta_1, \sigma^2, y \sim \mathrm{N}(\bar{y} - \beta_1 \bar{x}, {}^{\sigma^2}\!/_n)$$
$$\beta_1|\beta_0, \sigma^2, y \sim \mathrm{N}\left(\frac{\sum_{i=1}^n x_i y_i - \beta_0 \sum_{i=1}^n x_i}{\frac{\sigma^2}{\sum_{i=1}^n x_i^2}}\right)$$
$$\sigma^2|\beta_0, \beta_1, y \sim \chi^{-2}\left(n, \sum_{i=1}^n \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{n}\right)$$

Luego, tenemos distribuciones condicionales a posterioris conocidas, por lo cual podemos utilizar nuevamente Gibbs Sampling.

## 4.1 Modelo de regresión lineal multiple

Suponga que las variables predictoras $x_1, \ldots, x_{k-1}$, la matriz de diseño $X_{(n\times k)}$ y el vector de observaciones $Y$. Se tendrá entonces que

$$Y|X, \beta, \sigma^2 \sim \mathrm{N}_n(X\beta, \sigma^2 I)$$

donde

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \qquad X_{n\times k} = \begin{pmatrix} 1 & x_1 & \cdots & x_k \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n_1} & \cdots & x_{n_k} \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

Consideremos la siguiente priori no informativa

$$P(\tilde{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}$$

Luego tendremos que

- $\tilde{\beta}|\sigma^2, y \sim \mathrm{N}(\hat{\beta}, V_\beta \sigma^2)$ donde

$$\hat{\beta} = (X^T X)^{-1} X^T Y, \qquad V_\beta = (X^T X)^{-1}$$

- $\sigma^2|y \sim \chi^{-2}(n-k, S^2)$ donde

$$S^2 = \frac{1}{n-k}(Y - X\hat{\beta})^T (Y - X\hat{\beta})$$

**Observación.** Si el tamaño muestral es grande, la inversión de la matriz $(X^T X)$ puede ser computacionalmente muy exigente. El siguiente procedimiento es eficiente para realiar esta inversión

- Encontrar la descomposición QR de la matriz $X^T X$. En este caso se tendrá que

$$X^T X = QR$$

  donde $Q_{n\times k}$ es matriz ortogonal y $R_{k\times k}$ es una matriz triangular superior.

- Resolver $R\hat{\beta} = Q^T Y$, es decir,

$$\hat{\beta} = R^{-1} Q^T Y.$$

## 4.1 Predicción

Sea $\tilde{X}$ la matriz para las nuevas combinaciones de variables de interés, digamos $m \times k$. Luego

$$Y \sim \mathrm{N}_n(X\beta^*, (\sigma^2)^* I_n)$$

Utilizaremos el estadístico de chequeo

$$T(Y_{rep}, \beta, \sigma^2) = (Y_r ep - X\beta)^T (Y_{rep} - X\beta).$$

CAPÍTULO 5

# Estimador de Bayes

## 5.1 Contraste entre enfoque Clásico y Bayesiano

### 5.1.1. Enfoque clásico

Recordemos que un estadístico $T$ es suficiente para $\theta$ si $\widetilde{Y}|T$ no depende de $\theta$. Algunas propiedades asociadas a esta definición son:

1. $T$ es suficiente $\forall\ \theta$ si y solo sí $\exists\ h, g$ tal que

$$P(Y_1, \ldots, Y_n|\theta) = h(T(\widetilde{Y}), \theta)g(Y_1, \ldots, Y_n)$$

2. Las transformaciones inyectivas preservan la suficiencia.

3. El EMV de $\theta$, $\hat{\theta}$, se define como

$$\hat{\theta} = \arg\max_{\theta \in \Theta} P(\widetilde{Y}|\theta).$$

Luego, si $T$ es un estadístico suficiente para $\theta$, entonces en virtud del teorema de factorización se tendrá que

$$\hat{\theta} = \arg\max_{\theta \in \Theta} h(T(\widetilde{Y}), \theta).$$

### 5.1.2. Enfoque Bayesiano

Sea $T$ un estadístico suficiente para $\theta$. Entonces se tendrá que

$$P(\theta|Y) \propto h(T(\widetilde{Y}, \theta)P(\theta)$$

por lo cual la distribución a posteriori dependerá de $\widetilde{Y}$ solo a partir de $t(\widetilde{Y})$. Luego, si en cierto resumen(Media,Mediana,Cuantil,etc) de la posteriori, $P(\theta|\widetilde{Y})$ es un estadístico suficiente, entonces también será suficiente minimal.

## 5.2 Estimación

**Definición 5.1.** Un estimador de un cierto parámetro $\theta \in \Theta$ es una función

$$\delta : \widetilde{Y} \Longrightarrow \Theta.$$

Para medir la cercanía de un estimador $\delta$ con respecto a un parámetro $\theta$, vamos a usar una función que mide el costo o la perdida asociado al error de estimación, $L(\theta, \delta(Y))$. Suponiendo que todas las cantidades a continuación son finitas, se definen:

1. Pérdida a posteriori esperada de $\delta(Y)$: $\mathbb{E}(L(\theta, \delta(Y)|\widetilde{Y}) = \int_{\Theta} L(\theta, \delta(Y))P(\theta|\widetilde{Y})\ d\theta$.
2. Riesgo frecuentista: $R(\theta, \delta(Y)) = \int_{\widetilde{Y}} L(\theta, \delta(Y))P(\widetilde{y}|\theta)\ d\widetilde{y}$.
3. Riesgo integrado: $r(P(\theta, \delta(Y)) = \int_{\Theta} R(\theta, \delta(Y))P(\theta)\ d\theta$.

**Definición 5.2.** El estimador de Bayes de $\theta$ para la priori $P(\theta)$ y función de perdida $L(\theta, \delta)$ es $\delta_B(Y)$ tal que

$$r(P(\theta), \delta_B(Y)) \leqslant r(P(\theta), \delta(Y)),\ \forall \delta(Y).$$

El valor $r(P(\theta), \delta_B(Y))$ se llama Riesgo Bayesiano.

**Teorema 5.1.** *El estimador de Bayes de $\theta$ se puede obtener seleccionando el valor de $\delta(Y)$ que minimiza la pérdida esperada a posteriori.*

*Demostración.* Notemos que en virtud del teorema de Fubini se tiene que

$$\begin{aligned} r(P(\theta), \delta(Y)) &= \int_{\Theta}\int_{\widetilde{Y}} L(\theta, \delta(Y))P(Y|\theta)P(\theta)\ d\widetilde{Y}d\theta \\ &= \int_{\Theta}\int_{\widetilde{Y}} L(\theta, \delta(Y))P(\theta|\widetilde{Y})P(\widetilde{Y})d\widetilde{Y}d\theta \\ &= \int_{\widetilde{Y}} P(\widetilde{Y}\left[\int_{\Theta} L(\theta, \delta(Y))P(\theta|\widetilde{Y})\ d\theta\right] d\widetilde{Y} \\ &= \int_{widetildeY} \mathbb{E}(L(\theta, \delta(Y)))P(\widetilde{Y})\ d\widetilde{y} \end{aligned}$$

Luego, para minimizar el riesgo integrado, basta escoger $\delta(Y)$ que minimice la función de pérdida a posteriori esperada. □

**Ejemplo 5.1** (Funciones de Pérdida)**.**

1. Consideremos la siguiente función de pérdida, la que llamaremos pérdida cuadrática,

$$L(\theta, \delta) = (\theta - \delta)^2.$$

De aquí, la pérdida a posteriori esperada será

$$\mathbb{E}(L(\theta, \delta)|\widetilde{Y}) = \mathbb{E}(\theta^2 - 2\theta\delta + \delta^2|\widetilde{Y}).$$

Luego, el estimador de Bayes se obtendrá derivando según $\delta$ la expresión anterior. De hecho,

$$\delta_B = \mathbb{E}(\theta|\widetilde{Y})$$

Por ende, si

$$\begin{aligned} Y_1, \ldots, Y_n &\overset{\text{i.i.d}}{\sim} \text{Ber}(\theta) \\ \theta &\sim \text{U}(0,1) \end{aligned}$$

En este caso, sabemos que

$$P(\theta|\widetilde{Y}) \sim \text{Beta}\left(\sum y_1 +, n - \sum y_i + 1\right)$$

Por lo tanto, el estimador de Bayes será

$$\delta_B = \mathbb{E}(\theta|\widetilde{Y}) = \frac{\sum y_1 + 1}{n+2}$$

2. Consideremos la función de pérdida

$$L(\theta, \delta) = \begin{cases} k_1(\theta - \delta) & \text{si } \theta > \delta \\ k_2(\delta - \theta) & \text{si no} \end{cases}$$

Luego

$$\begin{aligned} \mathbb{E}(L(\theta,\delta)|\widetilde{Y}) &= k_1 \int_\delta^\infty (\theta - \delta) P(\theta|\widetilde{Y}\ d\theta + k_2 \int_{-\infty}^\delta (\delta - \theta) P(\theta|\widetilde{Y})\ d\theta \\ &= k_1 \int_\delta^\infty \theta P(\theta|\widetilde{Y})\ d\theta - \delta k_1 \int_\delta^\infty P(\theta|\widetilde{Y})\ d\theta + k_2 \int_{-\infty}^\delta \delta P(\theta|\widetilde{Y})\ d\theta - k_2 \int_{-\infty}^\delta \theta P(\theta|\widetilde{Y})\ d\theta \end{aligned}$$

Luego, derivando según $\delta$ e igualando a 0 obtenemos

$$P(\theta < \delta|\widetilde{Y}) = \frac{k_1}{k_1 + k_2}.$$

3. Consideremos la función de pérdida

$$L(\theta, \delta) = \begin{cases} 0 & \text{si } |\theta - \delta| < b \\ 1 & \text{si no} \end{cases}$$

La pérdida esperada a posteriori vendrá dada por

$$\mathbb{E}(L(\theta,\delta)|\widetilde{Y}) = P(|\theta-\delta| \geqslant b|\widetilde{Y})$$

De aquí

$$\begin{aligned}\delta_B(\widetilde{Y}) &= \arg\min_{\delta} P(|\theta-\delta| \geqslant B|\widetilde{Y}) \\ &= \arg\max_{\delta} P(\delta - b < \theta < b + \delta|\widetilde{Y})\end{aligned}$$

Por lo tanto, el valor que máximiza dicha probabilidad a posteriori es la moda. En el caso de multimodalidad, entonces se escoge aquella moda local que obtenga mayor valor de $P(\theta|\widetilde{Y})$. Este estimador recibe el nombre de estimador MAP (máximo a posteriori).

★

**Ejercicio 5.1.** Defina la función de pérdida $L(\theta,\delta) = w(\theta)(\theta-\delta)^2$, donde $w(\theta) > 0$ es una función de peso.

1. Muestre que el estimador de Bayes de $\theta$ para un modelo con verosimilitud $P(Y|\theta)$ y priori $P(\theta)$ es:

$$\delta_B(Y) = \frac{\mathbb{E}(\theta w(\theta)|Y)}{\mathbb{E}(w(\theta)|Y)} = \frac{\displaystyle\int_\Theta \theta w(\theta)P(Y|\theta)P(\theta)\ d\theta}{\displaystyle\int_\Theta w(\theta)P(Y|\theta)P(\theta)\ d\theta}$$

*Demostración.* Debemos encontrar $\delta^*$ que minimiza $\mathbb{E}(L(\theta,\delta)|Y)$. Notemos que

$$\begin{aligned}\mathbb{E}(L(\theta,\delta)|Y) &= \mathbb{E}(w(\theta)\theta^2 - w(\theta)2\theta\delta + w(\theta)\delta^2|Y) \\ &= \mathbb{E}(\theta^2 w(\theta)|Y) - 2\delta\mathbb{E}(\theta w(\theta)|Y) + \delta^2\mathbb{E}(w(\theta)|Y)\end{aligned}$$

Luego, derivando según $\delta$, igualando a 0 y aplicando el teorema de Bayes obtenemos

$$\delta^* = \frac{\mathbb{E}(\theta w(\theta)|Y)}{\mathbb{E}(w(\theta)|Y)} = \frac{\displaystyle\int_\Theta \theta w(\theta)P(\theta|Y)\ d\theta}{\displaystyle\int_\Theta w(\theta)P(\theta|Y)\ d\theta} = \frac{\displaystyle\int_\Theta \theta w(\theta)P(Y|\theta)P(\theta)\ d\theta}{\displaystyle\int_\Theta w(\theta)P(Y|\theta)P(\theta)\ d\theta}$$

lo que prueba lo pedido. □

2. Para el caso en que $Y = (y_1,\ldots,y_n)$ con $y_1,\ldots,y_n|\theta \sim \text{Ber}(\theta)$ y $P(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$ para $0 < \theta < 1$, con $a$ y $b$ conocidos. Obtenga los estimadores Bayesianos correspondientes a

   *a)* $w(\theta) = 1$.

*b)* $w(\theta) = \theta^2$.

*c)* $w(\theta) = \dfrac{1}{\theta}$.

*d)* $w(\theta) = \theta^{-1/2}(1-\theta)^{-1/2}$.

*Solución.* Propuesto. □

**Ejercicio 5.2.** En algunos casos no es razonable suponer una función de pérdida simétrica con respacto a sobreestimación o subestimación. Suponga un problema donde se desa estimar el parámetro $\theta$ en base a su distribución a posteriori dada la muesta $y_1, \ldots, y_n$ con verosimilitud $P(Y|\theta)$, bajo la función de pérdida asimétrica

$$L(\Delta) = \beta e^{\alpha\Delta} - \gamma\Delta - \beta$$

donde $\Delta = \hat{\theta} - \theta$ y $\alpha, \gamma \neq 0, \beta > 0$ constantes.

1. Para que la función de pérdida sea razonable, se requiere que esta tenga un mínimo en $\Delta = 0$. Encuentre una relación entre los parámetros $\alpha, \beta, \gamma$ para que esto ocurra.

   *Solución.* Derivando según $\Delta$ se tiene que

   $$\frac{\partial L}{\partial \Delta} = \alpha\beta e^{\alpha\Delta} - \gamma$$

   Luego, como queremos encontrar el mínimo igualamos a 0 y despejamos $\Delta$, obteniendo que

   $$\Delta = \frac{1}{\alpha}\log\left(\frac{\gamma}{\alpha\beta}\right)$$

   Por lo cual es necesario que

   $$\gamma = \alpha\beta.$$

   Para verificar que efectivamente es un mínimo, derivamos según $\Delta$, reemplazamos con $\Delta = 0$ y obtenemos

   $$\frac{\partial^2 L}{\partial \Delta^2} = \alpha^2\beta > 0$$

   por hipótesis. Por lo tanto, en $\Delta = 0$ encontramos un mínimo. □

2. Bajo la restricción anterior, encuentre el valor esperado de la función de pérdida $L$.

   *Demostración.* Bajo la restricción $\gamma = \alpha\beta$ nuestra función de perdida será

   $$L(\Delta) = \beta(e^{\alpha\Delta} - \alpha\Delta - 1).$$

Debemos tomar esperanza con respecto a la distribución de $\theta|Y$. En virtud de que $\Delta = \hat{\theta} - \theta$ se tendrña que

$$\begin{aligned}\mathbb{E}(L(\Delta)|\widetilde{Y}) &= \mathbb{E}(\beta(e^{\alpha\Delta} - \alpha\Delta - 1)|Y) \\ &= \mathbb{E}(\beta(e^{\alpha(\hat{\theta}-\theta)} - \alpha(\hat{\theta} - \theta) - 1)|Y) \\ &= \beta e^{\alpha\hat{\theta}}\mathbb{E}(e^{-\alpha\theta}|Y) - \alpha\beta\hat{\theta} + \alpha\beta\mathbb{E}(\theta|Y) - \beta\end{aligned}$$

Encontrando lo pedido. □

3. Encuentre el estimador de Bayes bajo la pérdida objetivo.

4. Suponga que la distribución a priori para $\theta$ corresponde a una distribución Normal, con media y varianza conocida, y que la función verosimilitud corresponde a una distribución Normal$(\theta, \sigma^2)$ con $\sigma^2$ conocido. Encuentre el estimador de Bayes para $\theta$ bajo la función de pérdida objetivo.

CAPÍTULO 6

# Test de Hipótesis Bayesiano

El procedimiento de test de hipótesis es similar al método científico. En él, el científico formula una teoría y luego la contrasta con la observación. En nuestro caso, se plantean teorías acerca del valor del parámetro.

Para el contraste, se toma una muestra de la población y compara la observación con la teoría. Si las observaciones discrepan fuertemente con la teoría, el científico probablemente rechazará la hipótesis. Si no, concluye que la teoría es probablemtente correcta, o que la muestra no detecta la diferencia entre el valor real y el planteado en la hipótesis.

## 6.1 Punto de vista clásico

Nos interesa contrastar las hipótesis

$$H_0 : \theta \in \Theta_0, \qquad H_1 : \theta \in \Theta_1$$

con $\Theta_0 \cap \Theta_1 = \varnothing$, $\Theta_0 \cup \Theta_1 = \Theta$ en base a una muestra $Y_1, \ldots, Y_n$ con verosimilitud $P(Y|\theta)$. Para esto, nos gustaría establecer una cierta región de Rechazo

$$R = \left\{ Y \in \mathcal{Y} \;\middle|\; T(Y) > c \right\}$$

**Ejemplo 6.1.** Consideremos $Y_1, \ldots, Y_n|\theta \sim \mathrm{N}(\theta, 1)$. Queremos contrastar

$$H_0 : \theta \leqslant \theta_0, \qquad H_1 : \theta > \theta_0$$

Se puede usar una región de rechazo $\left\{ Y \;\middle|\; \bar{Y} > c \right\}$.

★

Un contraste, prueba o test para dichas hipótesis sería una función de la muestra de la siguiente forma:

$$\delta = \begin{cases} 1 & \text{si } T(Y) \in R \\ 0 & \text{si } T(Y) \notin R \end{cases}$$

donde $R$ es la región de rechazo.

**Definición 6.1.** Una hipótesis es una afirmación sobre un parámetro poblacional. Existen dos tipos:

1. **Hipótesis simple**:

$$H_0 : \theta = \theta_0 \qquad \text{VS} \qquad H_1 : \theta = \theta_1.$$

2. **Hipótesis compuestas**:

$$H_0 : \theta \leqslant \theta_0 \qquad \text{VS} \qquad H_1 : \theta > \theta_0.$$

**Definición 6.2.** Dos hipótesis complementarias en un problema de testeo de hipótesis se denominan hipótesis nula, $H_0$, e hipótesis alternativa $H_1$. En general escribimos

$$H_0 : \theta \in \Theta_0 \qquad \text{VS} \qquad \theta \in \Theta_0^c.$$

**Definición 6.3.** El subconjunto del espacio muestral para el cual $H_0$ es rechazado se denomina región de rechao o crítica. Su complemento se denomina región de aceptación.

A la hora de realizar un test, podemos cometer dos tipos de errores a la hora de tomar una decisión:

- $H_0$ sea verdadera y la rechacemos, lo que denominamos **Error Tipo I**.
- $H_1$ sea verdadera y no rechazamos $H_0$, lo que corresponde a un **Error de tipo II**.

## 6.2 Test de razón de verosimilitud

Consideremos la hipótesis simple

$$H_0 : \theta = \theta_0 \qquad \text{VS} \qquad H_1 : \theta = \theta_1.$$

La razón de verosimilitud vendrá dada por

$$\lambda(\widetilde{X}) = \frac{L(\theta_0, \widetilde{X})}{L(\theta_1, \widetilde{X})}$$

En este caso, valores $\lambda(\widetilde{X}) > 1$ indican evidencia a favor de $H_0$, mientras que valores $\lambda(\widetilde{X}) < 1$ indican evidencia a favor de $H_1$.

**Definición 6.4.** Un test de máxima verosimilitud es cualquier test cuya región de rechazo es de la forma

$$R = \left\{ \widetilde{X} \mid \lambda(\widetilde{X}) \leqslant c \right\}$$

donde $0 \leqslant c \leqslant 1$.

**Observación 6.4.1.** Debemos encontrar $c$ tal que

$$P(\widetilde{X} \in X|\theta_0) \leqslant \alpha$$

es decir, un valor de $c$ que nos garantice una probabilidad de error tipo I de a lo más $\alpha$.

**Lema 6.1** (Neyman - Pearson)**.** *Suponga que el test de de máxima verosimilitud con nivel de significancia $\alpha$ se rechaza cuando*

$$\lambda(\widetilde{X}) < c_\alpha$$

*entonces cualquier otro test con nivel de significancia $\alpha^* \leqslant \alpha$ tiene menor o igual potencia que el test de máxima verosimilitud.*

**Definición 6.5** (Test de razón de verosimilitud Generalizado)**.** Supongamos que queremos contrastar 2 hipótesis compuestas de tipo

$$H_0 : \theta \in \Theta_0 \qquad \text{VS} \qquad H_1 : \theta \in \Theta_1$$

Podemos generalizar el TRV planteando el siguiente estadístico

$$\lambda^*(\widetilde{X}) = \frac{\sup_{\theta\in\Theta_0} L(\theta, \widetilde{X})}{\sup_{\theta\in\Theta_1} L(\theta, \widetilde{X})}$$

**Observación 6.5.1.** Valores pequeños de $\lambda^*$ nos indican rechazar $H_0$. Por razones técnicas, consideraremos

$$\lambda^*(\widetilde{X}) = \frac{\text{máx}_{\theta\in\Theta_0} L(\theta, \widetilde{X})}{\text{máx}_{\theta\in\Theta_1} L(\theta, \widetilde{X})}.$$

En este caso, la región de rechazo vendrá dada por

$$R = \left\{ \widetilde{X} \mid \lambda^*(\widetilde{X}) \leqslant c \right\}$$

donde $0 \leqslant c \leqslant 1$.

## 6.3 Método para evaluar Test

Sean $\alpha$ y $1-\beta$ las probabilidades de error tipo I y II respectivamente. En este caso, si $R$ es la región de rechazo, tendremos que

$$\alpha = P(\widetilde{X} \in R|H_0), \qquad 1-\beta = P(\widetilde{X} \notin R|H_1)$$

**Definición 6.6.** La función potencia de un test con región de rechazo $R$ corresponde a la función de $\theta$ definida por

$$\beta(\theta) = P(\widetilde{X} \in R|\theta)$$

**Ejercicio 6.1.** Supongamos que $X \sim \text{Bin}(5, \theta)$. Nos interesa contrastar

$$H_0 : \theta \leqslant \frac{1}{2} \qquad \text{VS} \qquad H_1 : \theta > \frac{1}{2}$$

Consideremos el test que rechaza $H_0$ si y solo sí observamos 5 exitos. Estudie la función de potencia del test y comparela con la función de potencia del test que rechaza $H_0$ cuando $X = 3, 4, 5$.

**Definición 6.7.** Sea $\alpha \in [0, 1]$. Diremos que un test con función potencia $\beta(\theta)$ es:

- De tamaño $\alpha$ si
$$\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha.$$
- De nivel $\alpha$ si
$$\sup_{\theta \in \Theta_0} \beta(\theta) \leqslant \alpha.$$

**Definición 6.8** (Test UMP)**.** Sea $\rho$ una clase de test para las hipótesis

$$H_0 : \theta \in \Theta_0 \qquad \text{VS} \qquad H_1 : \theta \in \Theta_0^c.$$

Diremos que un test en la clase $\rho$, con función de potencia $\beta(\theta)$ es uniformemente más potente en la clase $\rho$ si

$$\beta(\theta) \geqslant \beta'(\theta)$$

para todo $\theta \in \Theta_0^C$ y toda $\beta'(\theta)$ función potencia de un test en la clase $\rho$.

**Lema 6.2.** *Considere las hipótesis*

$$H_0 : \theta = \theta_0 \qquad \textit{VS} \qquad H_1 : \theta = \theta_1$$

*donde la función densidad de la muestra corresponde a $P_\theta$. Un test con región de rechazo*

$$R = \left\{ \widetilde{X} \mid \frac{P_{\theta_0}(\widetilde{x})}{P_{\theta_1}(\widetilde{x})} < K \right\}$$

*con $\alpha = P(\widetilde{X} \in R|\theta = \theta_0)$ es UMP de nivel $\alpha$.*

*Demostración.* Propuesto. □

**Observación 6.2.1.** Generalmente, existe test UMP cuando las hipótesis son unilaterales y las funciones de densidad involucradas poseen la propiedad de razón de verosimilitud monótona.

**Definición 6.9.** Una familia de funciones de densidad $\{g_\theta(t) \mid \theta \in \Theta\}$ para una variable aleatoria univariada $T$ y un parámetro unidimensional, $\theta$, tiene razón de verosimilitud monótona, si $\forall \theta_2 > \theta_1$ la función

$$\frac{g_{\theta_2}(t)}{g_{\theta_1}(t)}$$

es no decreciente en el conjunto $\{t \mid g_{\theta_1}(t) > 0, g_{\theta_2}(t) > 0\}$.

**Teorema 6.1** (Karlin-Rubin)**.** *Suponga las hipótesis*

$$H_0 : \theta \leqslant \theta_0 \qquad VS \qquad H_1 : \theta > \theta_0.$$

*Suponga un estadístico suficiente para $\theta$ y que la familia de distribuciones $\{g_\theta(t) \mid \theta \in \Theta\}$ para $T$ tiene razón de verosimilitud monótona. Entonces $\forall t$ el test con región de rechazo $R = \{T > t_0\}$ es UMP con nivel $\alpha = P(T > t_0 | \theta = \theta_0)$.*

*Demostración.* Propuesto. □

**Definición 6.10.** El valor-$p$de un valor muestral $\widetilde{X}$ corresponde al menor valor de $\alpha$ tal que $\widetilde{X}$ nos llevará a rechazar $H_0$.

## 6.4 Teoría de Decisión

En teoría de decisiones interesa combinar la información de los datos con otras características relevantes del problema para tomar la mejor decisión sobre una cantidad desconocida. Las características serán las consecuencias de la decisión y la información apriori(que proviene del investigador), mientras que las consecuensias de la decisión se llamarán perdidas.

### 6.4.1. Elementos de la teoría de decisión

1. Las cantidades desconocidas, denotadas por $\theta$, se llaman estados de la naturaleza.
2. El espacio de todos los estados de la naturaleza se denota $\Theta$.
3. En experimentos, las cantidades desconocidas son los parámetros $\theta \in \Theta$.
4. La función de pérdida, denotada por $L(\theta, d)$ es un elemento clave de la teoría de decisiones.

**Procedimiento del test Bayesiano**

En test de hipótesis

$$H_0 : \theta \in \Theta_0 \qquad \text{VS} \qquad H_1 : \theta \in \Theta.$$

Las acciones de interés son $d_0$ y $d_1$, donde $d_i$ denota la aceptación de las hipótesis $i$. Describiremos la decisión de Bayes si se considera la función de pérdida $0-1$. Supongamos que $\Theta = \{\theta_0, \theta_1\}$ y consideramos la priori

$$P(\theta = \theta_0) = \pi_0, \qquad P(\theta = \theta_1) = \pi_1$$

con $\pi_0 + \pi_1 = 1$. Nuestro modelo muestral vendrá dado por

$$P(Y_1, \ldots, Y_n|\theta_0) = f_0(Y), \qquad P(Y_1, \ldots, Y_n|\theta = \theta_1) = f_1(Y).$$

Tendremos dos posibles acciones

- $d_0$: Describir que $H_0$ es cierto.
- $d_1$: Describir que $H_1$ es cierto.

Luego, nuestra función de perdida o costo asociado vendrá dada por:

| $L(\theta, d)$ | $d_0$ | $d_1$ |
|---|---|---|
| $\theta = \theta_0$ | $0$ | $w_0$ |
| $\theta = \theta_1$ | $w_1$ | $0$ |

donde $w_0$ y $w_1$ son los parámetros de costos asociados a elegir $\theta_i$ dado que $\theta_j$ era cierto, con $i \neq j$.

Vamos a seleccionar aquella decisión que minimiza la périda esperada a posteriori. Por un lado tenemos

$$\mathbb{E}(L(\theta, d)|Y) = \begin{cases} w_1 P(\theta = \theta_1|Y) & \text{si } d = d_0 \\ w_0 P(\theta = \theta_0|Y) & \text{si } d = d_1 \end{cases}$$

Luego, nuestra decisión es $d_0$ si

$$\mathbb{E}(L(\theta, d_0)|Y) < \mathbb{E}(L(\theta, d_1)|Y)$$

y en caso contrari será $d_1$. Por lo tanto, nuestra decisión es $H_0$ si y solo sí

$$\frac{P(\theta = \theta_0|Y)}{P(\theta = \theta_1|Y)} > \frac{w_1}{w_0} \iff \frac{\pi_0 f_0(Y)}{\pi_1 f_1(Y)} > \frac{w_1}{w_0}$$

**Definición 6.11.** Diremos que

1. $\dfrac{P(\theta = \theta_0|Y)}{P(\theta = \theta_1|Y)}$ será la chance a posteriori de $H_0$ versus $H_1$.

2. $\frac{\pi_0}{\pi_1}$ las chances a priori de $H_0$ versus $H_1$.
3. $\frac{f_0(Y)}{f_1(Y)} := B_{01}$ es el factor de Bayes en favor de $H_0$ y en contra de $H_1$.

**Observación 6.11.1.** Si $\pi_0 = \pi_1$, entonces las chances a posteriori coinciden con el factor de Bayes $B_{01}$. Si además $w_0 = w_1$, la decisión es optar por $H_0$ si $B_{01} > 1$.

Ahora bien, si tenemos el modelo

$$\begin{aligned} Y_1, \ldots, Y_n|\theta &\sim P(Y|\theta) \\ \theta &\sim P(\theta) \end{aligned}$$

El problema es decidir en

$$H_0 : \theta \in \Theta_0 \qquad \text{VS} \qquad H_1 : \theta \in \Theta_1$$

Bajo la misma función de perdida, obtendremos que la decisión óptima es $d_0$ si

$$w_1 P(\theta \in \Theta_1|Y) < w_0 P(\theta \in \Theta_0).$$

**Caso unilateral**

Consideremos el modelo

$$\begin{aligned} Y_1, \ldots, Y_n|\theta &\overset{\text{i.i.d}}{\sim} P(\tilde{Y}|\theta) \\ \theta &\sim P(\theta) \end{aligned}$$

Las hipótesis de interés son

$$H_0 : \theta \leqslant \theta_0 \qquad \text{VS} \qquad \theta > \theta_0$$

Supongamos que

$$\pi_0 = P(\theta \leqslant \theta_0), \qquad \pi_1 = P(\theta > \theta_0)$$

**Teorema 6.2.** *Supongamos que $P(Y|\theta)$ tiene razón de verosimilitud monótona(RVM) en el estadístico $T(Y)$. Para las hipótesis anteriores, consideramos la función de pérdida*

| $L(\theta, d)$ | $d_0$ | $d_1$ |
|---|---|---|
| $\theta \leqslant \theta_0$ | $0$ | $w_0$ |
| $\theta > \theta_1$ | $w_1$ | $0$ |

*Entonces el procedimiento del test que minimiza la pérdida esperada a posteriori es optar por $H_1$ si $T(Y) \geqslant C$ para alguna constante $c$.*

**Ejemplo 6.2.** Considere el modelo $X_1, \ldots, X_n|\theta \sim \text{Beta}(\theta, 1)$, $\theta > 0$.

1. Suponiendo que la distribución a priori es $\theta \sim$ Gamma$(a, b)$, $a, b > 0$ conocidos. Identifique la distribución a posteriori y encuentre el estimador de Bayes de $\theta$ bajo pérdida cuadrática.

   *Solución.* Se tiene que

   $$\begin{aligned} P(\theta|X) &\propto P(X|\theta)P(\theta) \\ &\propto \prod_{i=1}^{n} P(x_i, \theta)\theta^{a-1}e^{-b\theta} \\ &\propto \prod_{i=1}^{n} \frac{\Gamma(\theta+1)}{\Gamma(\theta)\Gamma(1)} x_i^{\theta-1}\theta^{a-1}e^{-b\theta} \end{aligned}$$

   Por ende

   $$\theta|X \sim \text{Gamma}\left(a+n, b-\sum_{i=1}^{n}\log(x_i)\right)$$

   Luego, dado que tenemos pérdida cuadrática obtenemos que

   $$\delta_B = \mathbb{E}(\theta|X) = \frac{a+n}{b-\sum_{i=1}^{n} x_i}$$

   □

2. Identifique un estadístico $T(X)$, con $X = (x_1, \ldots, x_n)$ tal que se tenga RVM en $T(X)$.

   *Solución.* Supongamos que $\theta_1 > \theta_2$. Nos gustaría que la razón de verosimilitud sea monótona. En efecto, esto se cumple pues

   $$\frac{P(X|\theta_1)}{P(X|\theta_2)} = \left(\frac{\theta_1}{\theta_2}\right)^n \exp\left\{\sum_{i=1}^{n}\log(x_i)(\theta_1-\theta_2)\right\}$$

   Y esta función es decreciente para $T(X) = \sum_{i=1}^{n}\log(x_i) < 0$ para $\theta_1 > \theta_2$. □

3. Para las hipótesis

   $$H_0 : \theta \leqslant \theta_0 \qquad \text{VS} \qquad H_1 : \theta > \theta_0$$

   obtenga el test de Bayes correspondiente a la función de pérdida usual, dados los parámetros de penalización $w_0 > 0, w_1 > 0$.

   *Demostración.* Consideramos

   $$H_0 : \theta \leqslant \theta_0 \qquad \text{VS} \qquad H_1 : \theta > \theta_0$$

con función de pérdida

| $L(\theta, d)$ | $d_0$ | $d_1$ |
|---|---|---|
| $\theta = \theta_0$ | $0$ | $w_0$ |
| $\theta = \theta_1$ | $w_1$ | $0$ |

Sabemos que obtamos por $H_1$ si

$$\frac{P(\theta \leqslant \theta_0|X)}{P(\theta > \theta_0|X)} < \frac{w_1}{w_0}$$

Luego, usando que $P(\theta > \theta_0|X) = 1 - P(\theta \leqslant \theta_0|X)$ obtenemos que

$$P(\theta \leqslant \theta_0|X) < \frac{w_1}{w_0 + w_1}$$

De aquí

$$\theta_0 \leqslant \Gamma^{-1}_{\left(\frac{w_1}{w_0+w_1}; a+n, b-\sum_{i=1}^n \log(x_i)\right)}$$

donde $\Gamma^{-1}$ es el cuantil de la Gamma. El problema es que estos valores no estan tabulados. Por esto, recordamos que si $T \sim \text{Gamma}(n, \theta)$ entonces $2\theta T \sim \chi^2_{2n}$. Con esto, preferimos $H_1$ si

$$P\left(2\left(b - \sum_{i=1}^n \log(x_i)\right)\theta \leqslant 2\left(b - \sum_{i=1}^n \log(x_i)\right)\theta_0|X\right) < \frac{w_1}{w_0 + w_1}$$

de donde obtenemos que

$$\begin{aligned} 2\left(b - \sum_{i=1}^n \log(x_i)\right)\theta_0 &< \chi^2_{2(a+n), w_1/w_0+w_1} \\ \sum_{i=1}^n \log(x_i) &> \frac{2\theta_0 b - \chi^2_{2(a+n), w_1/w_0+w_1}}{2\theta_0} \end{aligned}$$

De donde obtenemos que

$$C = \frac{2\theta_0 b - \chi^2_{2(a+n), w_1/w_0+w_1}}{2\theta_0}$$

□

★

**Ejemplo 6.3** (Enfoque clásico)**.** Consideremos el modelo

$$X_1, \ldots, X_n|\theta \overset{\text{i.i.d}}{\sim} \text{N}(0, \theta)$$

Nos interesa

$$H_0 : \theta = \theta_0 \qquad H_1 : \theta = \theta_1$$

con $\theta_1 > \theta_0$. Desde el punto de vista clásico, el test más óptimo esta dado por el lema de Neyman-Pearson de nivel $\alpha$. Rechazamos $H_0$ si

$$\frac{P(X|\theta=\theta_1)}{P(X|\theta=\theta_0)} > K$$
$$\left(\frac{\theta_0}{\theta_1}\right)^{n/2} \exp\left\{\sum_{i=1}^{n} y_i^2 \left(\frac{1}{2\theta_0} - -\frac{1}{2\theta_1}\right)\right\} > K$$
$$\sum_{i=1}^{n} y_i^2 > K'$$

Luego, $K'$ debe ser tal que

$$P\left(\sum_{i=1}^{n} y_i^2 > K'|\theta=\theta_0\right) = \alpha$$

Desde el punto de vista Bayesiando usamos el mismo modelo muestral y además asignamos probabilidades a priori para $H_0$ y $H_1$. Entonves

$$P(\theta=\theta_0) = \pi_0 \qquad P(\theta=\theta_1) = \pi_1$$

Suponiendo parámetros de costo o penalización $w_0$ y $w_1$. La decisión óptima es $H_0$ si

$$\frac{P(\theta=\theta_0|X)}{P(\theta=\theta_1)} \geqslant \frac{w_1}{w_0}$$

y $H_1$ si

$$\frac{P(\theta=\theta_0|X)}{P(\theta=\theta_1|X)} < \frac{w_1}{w_0},1$$

★

### 6.4.2. Intervalos de Credibilidad

Los intervalos de credibilidad pueden ser interpretados como los intervalos dentro de los cuales se encuentra el parámetro con una cierta probabilidad. Los intervalos de credibilidad pueden ser definidos como:

1. **Intervalos de credibilidad centrales a posteriori de** $100(1-\alpha)\,\%$: Consiste en dos cuantiles debado y por encima de los cuales se tiene exactamente $100(\alpha/2)\,\%$ de la probabilidad a posteriori.

   **Observación 6.1.** Este coincide con el intervalo de confianza clásico si la distribución de los datos es simétrica y unimodal.

2. **Intervalos de máxima densidad a posteriori (HPD)**: Es la región que contiene el $100(1-\alpha)\,\%$

de la probablidad a posteriori. Además, la densidad dentro de la región nunca e smenor a la de cualquier otro punto fuera de la misma.

**Definición 6.12.** Un conjunto $100(1-\alpha)\,\%$ a HPD para una densida a posteriori $P(\theta|Y)$ es

$$C(k) = \left\{\theta \in \Theta \;\middle|\; P(\theta|Y) \geqslant k\right\}$$

donde $k$ es el mayor valor tal que

$$\int_{\theta \in C(k)} P(\theta|Y)\; d\theta = 1-\alpha.$$

**Observación 6.12.1.** En general, las condiciones para encontrar el HPD son

- $\displaystyle\int_a^b P(\theta|X)\; d\theta = 1-\alpha.$
- $P(a|X) = P(b|X)$.

**Ejemplo 6.4.** Supongamos que $N_1, \ldots, N_k$ tienen una distribución multinomila

$$P(n_1, \ldots, n_k|\widetilde{P}) = \frac{n!}{n_1! \cdots n_k!} \prod_{i=1}^{k} p_i^{n_i}$$

Consideramos

$$P(\widetilde{P}|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} p_i^{\alpha_i - 1}$$

Dirichlet. Queremos construir un intervalo de credibilidada HPD $100(1-\alpha)\,\%$ para $p_1$. Sabemos que

$$P|n_1, \ldots, n_k, \alpha \propto \prod_{i=1}^{k} p_i^{n_i + \alpha_i - 1}$$

Definiendo $\alpha_i' = n_i + \alpha_i$, utilizaremos que

$$P_1|n_1, \ldots, n_k, \alpha \sim \text{Beta}\left(\alpha_1', \sum_{i=2}^{k} \alpha_i'\right) \cong \text{Beta}\left(n_1 + \alpha_1, n - n_1 + \sum_{i=2}^{k} \alpha_i\right)$$

★