Estadística Bayesiana

Irving Gómez Méndez

Ejemplo 1

Tenemos una prueba de laboratorio que detecta una cierta enfermedad A. Denotamos el que salga una prueba positiva para la enfermedad como $T = 1$ y negativa como $T = 0$, y que el paciente tenga la enfermedad A como $E = 1$ y $E = 0$ en otro caso. Las características de la prueba son:

$$\mathbb{P}(T = 1|E = 1) = 0.92, \quad \mathbb{P}(T = 0|E = 0) = 0.99,$$

y la prevalencia de la enfermedad (la proporción de personas enfermas en la población en cuestión) es de 0.12. O sea $\mathbb{P}(E = 1) = 0.12$.

Yo voy y me hago la prueba referida (perteneciendo yo a la población en cuestión) y esta sale positiva. ¿Cuál es la probabilidad posterior de que yo tenga la enfermedad A?

$\mathbb{P}(T = 1|E = 1)$ ó $\mathbb{P}(T = 1|E = 0)$ **no** es lo que necesitamos. Más bien queremos saber qué sucede dado o una vez que $T = 1$.

Teorema de Bayes:

$$\mathbb{P}(E = 1|T = 1) = \frac{\mathbb{P}(T = 1|E = 1)\mathbb{P}(E = 1)}{\mathbb{P}(T = 1|E = 1)\mathbb{P}(E = 1) + \mathbb{P}(T = 1|E = 0)\mathbb{P}(E = 0)}$$

$$= 0.9262$$

Pero, en realidad, o tengo la enfermedad o no la tengo; entonces...

¡¿Qué quiere decir: La probabilidad de que tenga la enfermedad es 0.9262?!

¡No hay frecuencias o eventos repetidos: o estoy o no estoy enfermo!

El estadístico Bruno de Finetti comenzaba su libro con la frase *LA PROBABILIDAD NO EXISTE*. Las mayúsculas aparecen en el texto original. ¿A qué se refiere?

La probabilidad es condicional y subjetiva

Uno de los puntos de partida básicos en la Estadística Bayesiana es el concepto de probabilidad (y su definición). Para empezar, pongamos unos ejemplos en que usamos la probabilidad y tratemos de encontrar una definición para esta.

1. ¿Cuál es la probabilidad de que al lanzar una moneda caiga "águila"?
2. ¿Cuál es la probabilidad de que haya un eclipse mañana?
3. ¿Cuál es la probabilidad de que llueva mañana?
4. ¿Cuál es la probabilidad de que esté lloviendo en México?
5. ¿Cuál es la probabilidad de que haya más de 10^9 estrellas en nuestra galaxia?

Se podría pensar que para la primer pregunta la respuesta es 1/2, pero ¿de qué moneda estamos hablando? ¿tiene esta moneda un águila? (¡podría ser extranjera!). La moneda que se está usando, ¿se la dieron de cambio en la tiendita al profesor? o ¿es una moneda que se acaba de sacar del bolsillo un ilusionista?

Para la pregunta 4, yo podría informarme si está lloviendo en México (¿la ciudad o el estado completo o alguna parte del país?) y entonces esta probabilidad sería 0 ó 1; pero, en este instante, ¿qué tanto sabemos del evento "está lloviendo en México"?

Sea A el evento "está lloviendo en México". Una persona que vive a algunos kilómetros de Qaanaaq, sin acceso a internet, no tendría ninguna razón para asignar mayor probabilidad a A que a A^c. Si denotamos por H_1 la información de esta persona, entonces $\mathbb{P}(A|H_1) = \mathbb{P}(A^c|H_1)$.

Una persona que vive en el Bajío posee una información distinta H_2 y podría ser que

$$\mathbb{P}(A|H_2) = \begin{cases} 3/4 & \text{si llueve en el Bajío,} \\ 1/4 & \text{si no llueve en el Bajío.} \end{cases}$$

Por el contrario, si una persona vive en México, poseerá una información H_3, tal que

$$\mathbb{P}(A|H_3) = \begin{cases} 1 & \text{si llueve en México,} \\ 0 & \text{si no llueve en México.} \end{cases}$$

Por lo tanto, la probabilidad de un evento A es una medida del grado de incertidumbre que un individuo (o agente) tiene sobre ese evento. Esto quiere decir que la probabilidad es siempre contextual, dada una serie de supuestos y consideraciones, aun para las probabilidades más simples.

Ejemplo 2

Suponga que una bolsa tiene 4 canicas que pueden ser blancas o negras, pero no sabemos cuántas hay de cada color. Sabemos que hay cinco posibilidades: $\{B, B, B, B\}$, $\{N, B, B, B\}$, $\{N, N, B, B\}$, $\{N, N, N, B\}$, $\{N, N, N, N\}$. Llamemos a estas posibilidades *conjeturas*. Deseamos saber cuál de estas conjeturas es más plausible dada cierta evidencia sobre el contenido de la bolsa.
Como no contamos con más información sobre la plausibilidad de cada conjetura, asignamos una probabilidad de $1/5$ a cada una. Sacamos tres canicas al azar de la bolsa, una a la vez, haciendo la selección con reemplazo y observamos: (N, B, N).

Podemos calcular la probabilidad del evento (N, B, N) bajo cada una de las conjeturas.

Conjetura	Probabilidad Previa	Probabilidad Posterior
$\{B, B, B, B\}$	$1/5$	$\propto 1/5 \times 0$
$\{N, B, B, B\}$	$1/5$	$\propto 1/5 \times 3$
$\{N, N, B, B\}$	$1/5$	$\propto 1/5 \times 8$
$\{N, N, N, B\}$	$1/5$	$\propto 1/5 \times 9$
$\{N, N, N, N\}$	$1/5$	$\propto 1/5 \times 0$

Por lo tanto, la conjetura $\{N, N, N, B\}$ es la más plausible.

La posterior de hoy es la previa del mañana
El futuro no es como era antes

Suponga que sacamos otra canica, y que resulta ser una canica negra. Con esta nueva información podemos actualizar la probabilidad de cada conjetura.

Conjetura	Probabilidad Previa	Probabilidad Posterior
$\{B, B, B, B\}$	$\propto 0$	$\propto 0 \times 0 = 0$
$\{N, B, B, B\}$	$\propto 3$	$\propto 3 \times 1 = 3$
$\{N, N, B, B\}$	$\propto 8$	$\propto 8 \times 2 = 16$
$\{N, N, N, B\}$	$\propto 9$	$\propto 9 \times 3 = 27$
$\{N, N, N, N\}$	$\propto 0$	$\propto 0 \times 4 = 0$

Hay que notar que l**a que antes era la probabilidad posterior ahora se ha vuelto nuestra probabilidad previa**.

Hasta ahora la nueva información ha sido de la misma naturaleza que la anterior. Pero en general, la información previa y la nueva pueden ser de distinto tipo. Suponga, por ejemplo, que alguien de la fábrica de canicas le dice que las canicas negras son raras. Pues por cada bolsa del tipo $\{N, N, N, B\}$ hay dos del tipo $\{N, N, B, B\}$ y tres del tipo $\{N, B, B, B\}$. Además, todas las bolsas contienen al menos una canica negra y una blanca.

Con esta información, actualizamos la probabilidad de cada conjetura

Conjetura	Probabilidad Previa	Probabilidad Posterior
$\{B, B, B, B\}$	$\propto 0$	$\propto 0 \times 0 = 0$
$\{N, B, B, B\}$	$\propto 3$	$\propto 3 \times 3 = 9$
$\{N, N, B, B\}$	$\propto 16$	$\propto 16 \times 2 = 32$
$\{N, N, N, B\}$	$\propto 27$	$\propto 27 \times 1 = 27$
$\{N, N, N, N\}$	$\propto 0$	$\propto 0 \times 0 = 0$

Con la información que ahora contamos, resulta que la conjetura $\{N, N, B, B\}$ es ahora más plausible.

Filosofía bayesiana

De una u otra manera, la subjetividad siempre ha estado presente en la actividad científica, comenzando por los supuestos sobre los cuales se decide abordar un determinado problema, típicamente se les denomina *supuestos razonables*, pero son "razonables" de acuerdo a la experiencia e información (subjetiva) particular de quien o quienes estudian un fenómeno en un momento dado.

De acuerdo a Wolpert (1992):

> *[...] la idea de una "objetividad científica" tiene tan solo un valor limitado, ya que el proceso mediante el cual se generan las ideas [o hipótesis] científicas puede ser bastante subjetivo [...] Es una ilusión creer que los científicos no tienen un vínculo emocional con sus convicciones científicas [...] las teorías científicas implican una continua interacción con otros científicos y el conocimiento previamente adquirido [...] así como una explicación que tenga posibilidades de ser aceptada [o al menos considerada seriamente] por el resto de la comunidad científica.*

De acuerdo a Press (2003) la subjetividad es una parte inherente y requerida para la inferencia estadística, y para el *método científico*. Sin embargo, ha sido la manera informal y desorganizada en la que se ha permitido la presencia de la subjetividad, la responsable de diversos errores y malas interpretaciones en la historia de la ciencia. La estadística bayesiana incorpora de manera formal y fundamentada la información subjetiva.

Al hablar de *información subjetiva* nos referimos a toda aquella información previa que se tiene sobre el fenómeno aleatorio de interés, antes de recolectar o realizar nuevas mediciones sobre el mismo, incluyendo: datos históricos, teorías, opiniones y conjeturas de expertos, conclusiones basadas en estudios previos, etc.

Mantras

En resumen:

- La probabilidad de un evento A es una medida del grado de creencia que tiene un individuo en la ocurrencia de A con base en la información H que posee. Por lo tanto, toda probabilidad es condicional (a las circunstancias, el *agente*, etc.)
- Sea $\mathcal{P} = \{p(Y|\theta), \theta \in \Theta\}$ una familia paramétrica de distribuciones. Como toda incertidumbre debe ser medida a través de una probabilidad y contamos con incertidumbre en los parámetros, entonces los parámetros deben ser modelados con una medida de probabilidad (una distribución).

Modelos Uniparamétricos

Consideremos Y una variable aleatoria, cuya verosimilitud, cuando toma el valor y y los parámetros θ toman el valor t, está dada por

$$p_{Y|\theta}(y|t).$$

Por ejemplo, si Y es una variable aleatoria normal con media θ y con varianza unitaria, entonces

$$p_{Y|\theta}(y|t) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(y-t)^2}{2}\right\} \mathbb{1}_{y\in\mathbb{R}}.$$

En el argot bayesiano se suele hacer uso indiscriminado de abusos, a veces hasta peligrosos, de notación, por ejemplo

$$p_{Y|\theta}(y|t)$$

lo escribiríamos como

$$p(Y|\theta).$$

Notación

- Variables aleatorias: Y, θ.
- Espacio muestral: $\mathcal{Y}$, $Y \in \mathcal{Y}$.
- Espacio paramétrico: Θ, $\theta \in \Theta$.
- Muestra aleatoria: $\mathbf{Y} = (Y_1, \ldots, Y_n)$.
- Muestra observada: $\mathbf{y} = (y_1, \ldots, y_n)$.
- Distribución previa o *a priori*: $p(\theta)$.
- Verosimilitud: $p(Y|\theta)$.
- Verosimilitud de la muestra: $p(\mathbf{Y}|\theta)$.
- Evidencia: $p(\mathbf{Y})$.
- Distribución posterior o *a posteriori*: $p(\theta|\mathbf{Y})$.

Distribución posterior

La distribución de probabilidad previa se trata de una distribución basada en experiencia previa (experiencia de especialistas, datos históricos, etc.), antes de obtener datos muestrales nuevos.

Luego procedemos a observar los nuevos datos (obtención de evidencia) y combinamos esta información con la distribución previa mediante la Regla de Bayes y obtenemos una distribución de probabilidad posterior:

$$p_{\theta|\mathbf{Y}}(t|\mathbf{y}) = \frac{p_{\mathbf{Y}|\theta}(\mathbf{y}|t)p_\theta(t)}{p_{\mathbf{Y}}(\mathbf{y})} = \frac{p_{\mathbf{Y}|\theta}(\mathbf{y}|t)p_\theta(t)}{\int_\Theta p_{\mathbf{Y}|\theta}(\mathbf{y}|\tilde{t})p_\theta(\tilde{t})d\tilde{t}}$$

Usando los abusos usuales de la Estadística Bayesiana, el resultado anterior lo escribiríamos como:

$$p(\theta|\mathbf{Y}) = \frac{p(\mathbf{Y}|\theta)p(\theta)}{p(\mathbf{Y})} = \frac{p(\mathbf{Y}|\theta)p(\theta)}{\int_\Theta p(\mathbf{Y}|\tilde{\theta})p(\tilde{\theta})d\tilde{\theta}}$$

o también como:

$$p(\theta|\mathbf{Y}) \propto p(\mathbf{Y}|\theta)p(\theta).$$

Beta-Bernoulli

Supongamos que $Y_i|\theta \overset{iid}{\sim} \text{Bernoulli}(\theta)$ y que la incertidumbre acerca de $\theta \in (0,1)$ la cuantificamos mediante $\theta \sim \mathsf{Beta}(\alpha, \beta)$. Obtenemos $\mathbf{y} = (y_1, \ldots, y_n)$, entonces

$$p(\mathbf{Y}|\theta) = \theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i} \prod_{i=1}^n \mathbb{1}_{\{0,1\}}(y_i)$$

y

$$p(\theta) = B(\alpha,\beta)^{-1}\theta^{\alpha-1}(1-\theta)^{\beta-1}\mathbb{1}_{(0,1)}(\theta).$$

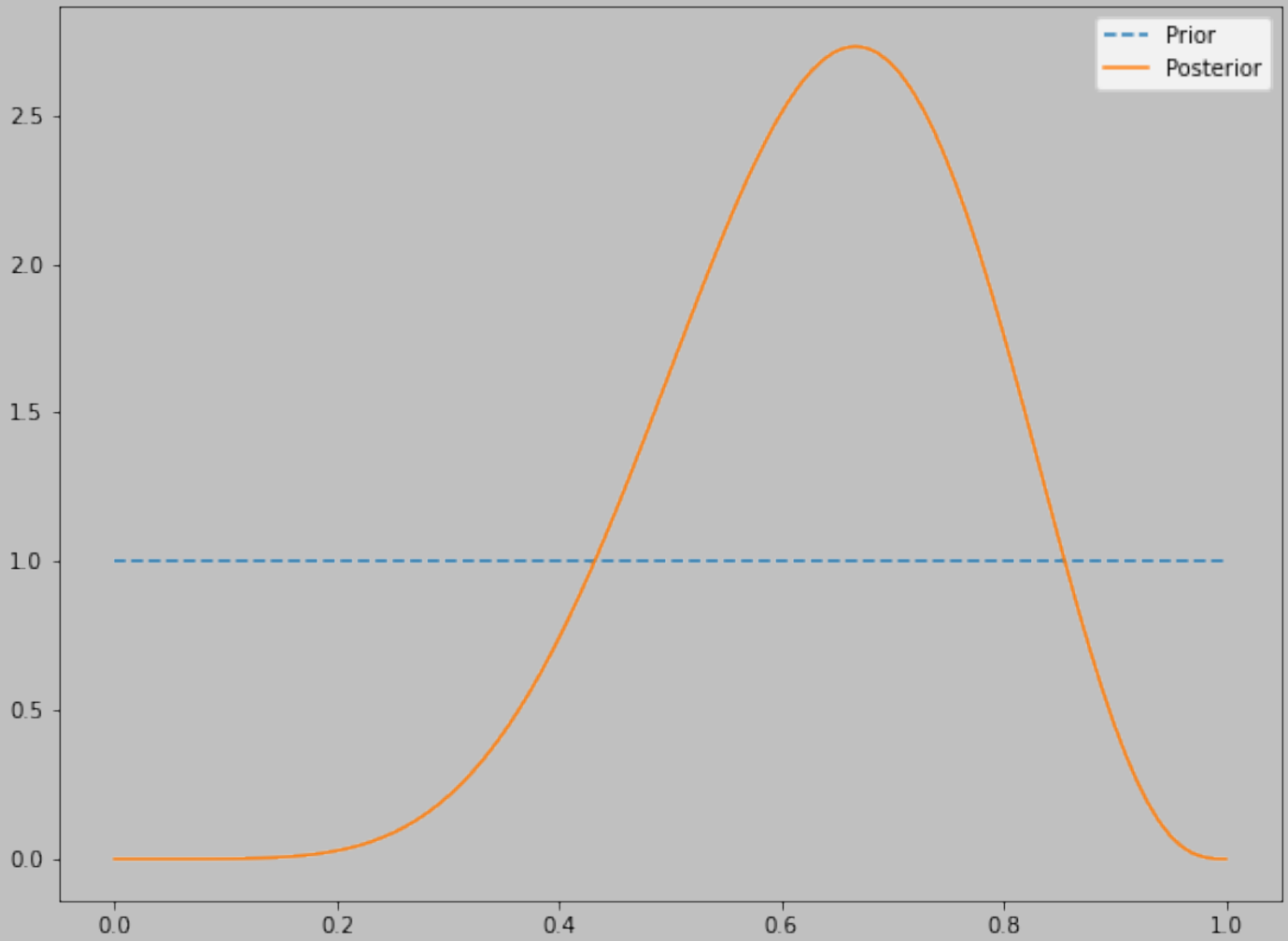
Por lo tanto,

$$p(\theta|\mathbf{Y}) \propto \theta^{\alpha+\sum_{i=1}^n y_i-1}(1-\theta)^{\beta+n-\sum_{i=1}^n y_i-1}\mathbb{1}_{(0,1)}(\theta),$$

es decir $\theta|\mathbf{Y} \sim \mathsf{Beta}\left(\alpha + \sum_{i=1}^n y_i, \beta + n - \sum_{i=1}^n y_i\right)$.

Ejemplo 3

Suponga que $\mathbf{Y}|\theta \sim \mathsf{Bernoulli}(\theta)$ y que no contamos con información sobre θ como para preferir un valor sobre otro, así que modelamos $\theta \sim \mathsf{Beta}(1,1)$. Se obtiene una muestra aleatoria simple (i.e. se obtienen variables i.i.d.)
$\mathbf{y} = (1,0,1,1,1,0,1,0,1)$. A continuación se muestra la función de densidad previa y posterior para θ.


Prior
Posterior
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0

Beta-Binomial

Recuerde que si $Y_1 \ldots, Y_n|\theta \overset{iid}{\sim} \mathsf{Bernoulli}(\theta)$, entonces

$$Z = \sum_{i=1}^{n} Y_i|\theta \sim \mathsf{Binomial}(n, \theta).$$

Luego, si

$$\theta \sim \mathsf{Beta}(\alpha, \beta),$$

concluimos que

$$\theta|Z \sim \mathsf{Beta}(\alpha + Z, \beta + n - Z).$$

Principio de indiferencia

Al analizar el modelo Binomial, Laplace consideró la distribución Uniforme como previa, argumentando lo que llamó el *principio de indiferencia* también llamado *principio de razón insuficiente*, que establece que el supuesto de uniformidad es apropiado cuando no se sabe nada sobre θ, i.e.

$$\begin{aligned} \theta &\sim \mathsf{Beta}(1,1), \\ Y|\theta &\sim \mathsf{Binomial}(n,\theta) \\ \Rightarrow \theta|Y &\sim \mathsf{Beta}(Y+1, n-Y+1). \end{aligned}$$

Una de las primeras aplicaciones de Laplace fue el estimar la proporción θ de nacimientos de mujeres en una población. Laplace obtuvo que entre 1745 y 1770 nacieron 241945 niñas y 251527 niños en París, si Y denota el número de nacimiento de niñas

$$\Rightarrow \theta|Y \sim \mathsf{Beta}(241946, 251528)$$

Ejercicio 1

Suponga que una particular población de células puede estar en uno de los siguientes tres estados de producción de una cierta proteína. Los estados son A, B y C, de producción baja, media y alta. Se toma una muestra al azar de 20 células, dentro de cierta población y se verifica si cada una de estas está en producción de la proteína (el resultado del aparato es si o no: 1 o 0, por cada célula analizada). De esta muestra resultan 12 células en producción (1) y las demás en negativo (0).

Por otro lado, sabemos que si la población está en el estado A, se espera que el 20% de la células produzcan la proteína, si está en el estado B se espera que el 50% de las células la produzcan y si está en el estado C se espera que el 70% la produzca.

¿Cuál es la probabilidad de que la población esté en cada uno de estos estados?

Distribuciones predictivas

Muchos son los casos en los que, más que estar interesados en el vector de parámetros θ, lo que queremos es describir el comportamiento de observaciones futuras del fenómeno aleatorio en cuestión, esto es, hacer predicción.

Por lo regular, la Estadística Frecuentista aborda este problema estimando puntualmente a θ con base en la muestra observada y dicho estimador $\hat{\theta}$ es sustituido en $p(Y|\theta)$, es decir, se utiliza $p(Y|\hat{\theta})$.

En Estadística Bayesiana este problema se aborda marginalizando la distribución conjunta de θ y Y.

- Distribución predictiva previa o *a priori*:

$$p(Y) = \int_{\Theta} p(Y|\theta)p(\theta)d\theta$$

Una vez obtenida una muestra $\mathbf{Y}$ se induce una distribución conjunta para Y y θ condicional a la muestra,

$$\begin{aligned} p(Y, \theta|\mathbf{Y}) &= \frac{p(Y, \theta, \mathbf{Y})}{p(\mathbf{Y})} \\ &= p(Y|\theta, \mathbf{Y})\frac{p(\theta, \mathbf{Y})}{p(\mathbf{Y})} \\ &= p(Y|\theta)p(\theta|\mathbf{Y}) \end{aligned}$$

- Distribución predictiva (posterior o *a posteriori*):

$$p(Y|\mathbf{Y}) = \int_{\Theta} p(Y|\theta)p(\theta|\mathbf{Y})d\theta$$

Ley de sucesión de Laplace

Considere el modelo Beta-Binomial, tomando como previa $\theta \sim \mathsf{Beta}(1,1)$, por lo que la posterior de θ está dada por:

$$\theta|Y \sim \mathsf{Beta}(Y+1, n-Y+1).$$

Además, recuerde que si θ es una variable aleatoria con distribución $\mathsf{Beta}(\alpha,\beta)$, entonces

$$\mathbb{E}(\theta) = \frac{\alpha}{\alpha+\beta},$$

por lo tanto

$$\mathbb{E}(\theta|Y) = \frac{Y+1}{n+2}.$$

Suponga que deseamos saber cuál es la probabilidad de que una nueva observación Bernoulli $\tilde{Y}$ tome el valor de 1, es decir $\mathbb{P}(\tilde{Y} = 1|Y)$.

$$\begin{aligned}\mathbb{P}(\tilde{Y} = 1|Y) &= \int_0^1 \mathbb{P}(\tilde{Y} = 1|\theta, Y)p(\theta|Y)d\theta \\ &= \int_0^1 \theta p(\theta|Y)d\theta \\ &= \mathbb{E}(\theta|Y) \\ &= \frac{Y+1}{n+2}.\end{aligned}$$

Si, por ejemplo, se ha realizado un experimento Bernoulli n veces sin éxito ($Y = 0$), la probabilidad de éxito la siguiente vez es de $1/(n+2)$, a pesar de que la probabilidad de éxito estimada por probabilidad clásica es $0/n = 0$.

Determinando los hiperparámetros

Uno de los retos a resolver, cuando se hace uso de la Estadística Bayesiana, es determinar los parámetros de la distribución previa, los cuales son llamados hiperparámetros. Una manera de solucionar este problema es, primero interpretar los hiperparámetros y, a partir de la interpretación dada, determinar los valores más apropiados.

Por ejemplo, considere el modelo Beta-Binomial, en el que la verosimilitud está dada por

$$p(Y|\theta) \propto \theta^a(1-\theta)^b,$$

donde a es el número de éxitos que se han obtenido y b es el número de fracasos.
Por otro lado, la distribución previa está dada por

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

Comparando estas dos expresiones podemos concluir que $\alpha - 1$ puede ser interpretado como el número de éxitos a priori y $\beta - 1$ como el número de fracasos a priori.

Convergencia Normal

Ya sabemos que en el modelo Beta-Binomial se cumple que

$$\theta|Y \sim \mathsf{Beta}(\alpha + Y, \beta + n - Y).$$

Además, recuerde que si θ es una variable aleatoria con distribución $\mathsf{Beta}(\alpha, \beta)$, entonces

$$\mathbb{E}(\theta) = \frac{\alpha}{\alpha + \beta}$$

y

$$\mathbb{V}(\theta) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

Luego, se tiene que

$$\mathbb{E}(\theta|Y) = \frac{\alpha + Y}{\alpha + \beta + n}$$

y

$$\mathbb{V}(\theta|Y) = \frac{(\alpha + Y)(\beta + n - Y)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)}$$

Note que cuando $n \to \infty$, se cumple que $Y \to \infty$ y $n - Y \to \infty$, siempre que $\theta \in (0, 1)$. Por lo que los valores de los hiperparámetros se vuelven despreciables, luego

$$\mathbb{E}(\theta|Y) \approx \frac{Y}{n}$$

y

$$\mathbb{V}(\theta|Y) \approx \frac{1}{n}\frac{Y}{n}\left(1 - \frac{Y}{n}\right).$$

Por otro lado, por el teorema del límite central y el teorema de Slutsky, sabemos que

$$\left(\sqrt{n}\frac{\bar{Y} - \theta}{\sqrt{\bar{Y}(1 - \bar{Y})}}\middle|\theta\right) \xrightarrow[n\to\infty]{L} \mathsf{Normal}(0, 1).$$

De manera análoga, se satisface que

$$\left(\frac{\theta - \mathbb{E}(\theta|Y)}{\sqrt{\mathbb{V}(\theta|Y)}}\Big|Y\right) \xrightarrow[n\to\infty]{L} \mathsf{Normal}(0,1).$$

Es decir, la distribución posterior converge en distribución a una variable aleatoria Normal.

Ejemplo 4

Consideremos una urna que contiene dos monedas: una cargada y la otra equilibrada. Supongamos que la moneda cargada está construida para tener una probabilidad de 3/4 de que salga águila. Una persona tomará una de las dos monedas de la urna (no necesariamente al azar) y echará volados.

En este caso podemos definir el espacio medible como $\Omega = \{\text{“Sale águila”}, \text{“Sale sol”}\}$ y tomar como sigma-álgebra el conjunto potencia de Ω, $\mathcal{P}(\Omega)$. En este espacio definimos la variable aleatoria

$$Y(\omega) = \begin{cases} 1 & \text{si } \omega = \text{“Sale águila”}, \\ 0 & \text{si } \omega = \text{“Sale sol”}. \end{cases}$$

Implícitamente estamos eliminando como posibles resultados eventos como que la moneda caiga vertical, que no sepamos el resultado del volado, etc.

$\Theta = \left\{\frac{3}{4}, \frac{1}{2}\right\}$, $\mathcal{Y} = \{0, 1\}$, $Y|\theta \sim \text{Bernoulli}(\theta)$.

Información previa:
$\mathbb{P}\left(\theta = \frac{3}{4}\right) = \alpha$, $\mathbb{P}(\theta = \frac{1}{2}) = 1 - \alpha$, $\alpha \in (0, 1)$.

$$p(\theta) = \alpha \mathbb{1}_{\{3/4\}}(\theta) + (1 - \alpha)\mathbb{1}_{\{1/2\}}(\theta).$$

Verosimilitud:

$$p(Y|\theta) = \theta^Y (1 - \theta)^{1-Y} \mathbb{1}_{\{0,1\}}(Y).$$

Predictiva previa:

$$
\begin{aligned}
p(Y) &= \sum_{\theta\in\Theta} p(Y|\theta)p(\theta) \\
&= \alpha\left(\frac{3}{4}\right)^{Y}\left(\frac{1}{4}\right)^{1-Y}\mathbb{1}_{\{0,1\}}(Y) + (1-\alpha)\left(\frac{1}{2}\right)\mathbb{1}_{\{0,1\}}(Y) \\
&= \alpha\left[\left(\frac{3}{4}-\frac{1}{2}\right)\mathbb{1}_{\{1\}}(x) + \left(\frac{1}{4}-\frac{1}{2}\right)\mathbb{1}_{\{0\}}(Y)\right] + \frac{1}{2}\mathbb{1}_{\{0,1\}}(Y) \\
&= \frac{\alpha}{4}\left[\mathbb{1}_{\{1\}}(Y) - \mathbb{1}_{\{0\}}(Y)\right] + \frac{1}{2}\mathbb{1}_{\{0,1\}}(Y).
\end{aligned}
$$

Es decir,

$$
p_Y(1) = \frac{1}{2} + \frac{\alpha}{4}, \quad p_Y(0) = \frac{1}{2} - \frac{\alpha}{4}.
$$

Verosimilitud de la muestra: Sean $Y_1, \ldots, Y_n|\theta \overset{iid}{\sim} \text{Bernoulli}(\theta)$, entonces

$$p(\mathbf{Y}|\theta) = \theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i} \prod_{i=1}^n \mathbb{1}_{\{0,1\}}(y_i)$$
$$= \theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i} g(\mathbf{y}).$$

Evidencia:

$$p(\mathbf{Y}) = \sum_{\theta\in\Theta} p(\mathbf{Y}|\theta)p(\theta)$$
$$= \alpha\left(\frac{3}{4}\right)^{\sum_{i=1}^n y_i}\left(\frac{1}{4}\right)^{n-\sum_{i=1}^n y_i} g(\mathbf{y}) + (1-\alpha)\left(\frac{1}{2}\right)^n g(\mathbf{y})$$

$$= \left[\alpha\frac{3^{\sum_{i=1}^n y_i}}{4^n} + (1-\alpha)\frac{1}{2^n}\right] g(\mathbf{y}).$$

Posterior:

$$p(\theta|\mathbf{Y}) = \frac{p(\mathbf{Y}|\theta)p(\theta)}{p(\mathbf{Y})}$$

$$= \frac{\theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i}\left[\alpha \mathbb{1}_{\{3/4\}}(\theta) + (1-\alpha)\mathbb{1}_{\{1/2\}}(\theta)\right]}{\alpha\frac{3^{\sum_{i=1}^n y_i}}{4^n} + (1-\alpha)\frac{1}{2^n}}$$

$$= \frac{[2(1-\theta)]^n\left(\frac{\theta}{1-\theta}\right)^{\sum_{i=1}^n y_i}\left[\alpha \mathbb{1}_{\{3/4\}}(\theta) + (1-\alpha)\mathbb{1}_{\{1/2\}}(\theta)\right]}{1-\alpha+\alpha\left(\frac{3^{\sum_{i=1}^n y_i}}{2^n}\right)}.$$

Sea $v = \frac{\alpha}{1-\alpha}\left(\frac{3^{\sum_{i=1}^n y_i}}{2^n}\right)$, entonces

$$p(\theta|\mathbf{Y}) = \frac{[2(1-\theta)]^n \left(\frac{\theta}{1-\theta}\right)^{\sum_{i=1}^n y_i}\left[\alpha \mathbb{1}_{\{3/4\}}(\theta) + (1-\alpha)\mathbb{1}_{\{1/2\}}(\theta)\right]}{(1-\alpha)(1+v)},$$

es decir

$$p_{\theta|\mathbf{Y}}(1/2|\mathbf{y}) = \frac{1}{1+v}$$

y

$$p_{\theta|\mathbf{Y}}(3/4|\mathbf{y}) = 1 - \frac{1}{1+v} = \frac{v}{1+v} = \frac{1}{1+v^{-1}}.$$

Predictiva (posterior)

$$p(Y|\mathbf{Y}) = \sum_{\theta\in\Theta} p(Y|\theta)p(\theta|\mathbf{Y})$$

$$= \left[\frac{1}{2}\left(\frac{1}{v+1}\right) + \frac{3^y}{4}\left(\frac{1}{v^{-1}+1}\right)\right] \mathbb{1}_{\{0,1\}}(y)$$

$$= \left[\frac{2}{4(v+1)} + \frac{3^y v}{4(v+1)}\right] \mathbb{1}_{\{0,1\}}(y)$$

$$= \frac{3^y v + 2}{4(v+1)} \mathbb{1}_{\{0,1\}}(y).$$

Resumen del ejemplo 4

Previa:

$$p(\theta) = \alpha \mathbb{1}_{\{3/4\}}(\theta) + (1-\alpha)\mathbb{1}_{\{1/2\}}(\theta), \quad \alpha \in (0,1)$$

Verosimilitud:

$$p(Y|\theta) = \theta^y (1-\theta)^{(1-y)} \mathbb{1}_{\{0,1\}}(y)$$

Predictiva previa:

$$p_Y(1) = \frac{1}{2} + \frac{\alpha}{4}, \quad p_Y(0) = \frac{1}{2} - \frac{\alpha}{4}$$

Posterior:

$$p_{\theta|\mathbf{Y}}(1/2|\mathbf{Y}) = \frac{1}{1+v}, \quad p_{\theta|\mathbf{Y}}(3/4|\mathbf{Y}) = \frac{1}{1+v^{-1}},$$

$$v = \frac{\alpha}{1-\alpha}\left(\frac{3^{\sum_{i=1}^{n} y_i}}{2^n}\right)$$

Predictiva (posterior):

$$p_{Y|\mathbf{Y}}(1|\mathbf{Y}) = \frac{3v+2}{4(v+1)}, \quad p_{Y|\mathbf{Y}}(0|\mathbf{Y}) = \frac{v+2}{4(v+1)}$$

No hay torta gratis

Cuando se hace Inferencia Estadística Frecuentista, los procedimientos suelen justificarse por el comportamiento asintótico del método. Como consecuencia, su desempeño con muestras pequeñas es cuestionable. Al contrario, la Estadística Bayesiana es válida para cualquier tamaño de muestra. Esto no quiere decir que tener más datos no sea útil, todo lo contrario.

El precio que hay que pagar por este poder es la dependencia en la información previa. Si se tiene una mala previa, los resultados no serán confiables.

Discusión sobre la previa

Históricamente, algunos detractores de la Estadística Bayesiana argumentan sobre la arbitrariedad de la distribución previa. Es cierto que las distribuciones previas son lo suficientemente flexibles para codificar distinto tipo de información. Entonces, si la previa puede ser cualquier cosa, ¿no es posible obtener cualquier respuesta que quieras? De hecho sí.

Sin embargo, si tu objetivo es mentir usando la estadística es una tontería hacerlo a través de la previa, pues dicha mentira estaría rápidamente cubierta por los datos. Es más fácil modificar la inferencia cambiando la verosimilitud.

Tarea 1

- Dejar tarea 1.
- Mostrar ejemplo Beta-Bernoulli.
- Mostrar ejemplo moneda cargada.

Análisis Conjugado

Cuando la distribución posterior sigue la misma forma paramétrica que la distribución previa, se dice que se tiene un modelo conjugado.
Formalmente, podemos definir un modelo conjugado de la siguiente manera:
Si $p(Y|\theta) \in \mathcal{F}$ y $\mathcal{P}$ es una familia de distribuciones previas para θ, entonces $\mathcal{P}$ es conjugada para $\mathcal{F}$ si

$$p(\theta|\mathbf{Y}) \in \mathcal{P} \text{ para toda } p(\cdot|\theta) \in \mathcal{F} \text{ y } p(\cdot) \in \mathcal{P}$$

Modelos conjugados para la familia exponencial

Cuando la verosimilitud de los datos sigue una distribución perteneciente a la familia exponencial de distribuciones es posible obtener la forma de las previas conjugadas.
Sea $p(Y|\theta) \in \mathcal{F}$, donde $\mathcal{F}$ es la familia exponencial de distribuciones. Entonces, por el teorema de la factorización de Fisher, $p(Y|\theta)$ acepta una parametrización de la forma:

$$p(Y|\theta) = f(y)g(\theta)\exp\left\{\phi^T(\theta)u(y)\right\},$$

donde $\phi(\theta)$ y $u(y)$ son, en general de la misma dimensión que θ y $\phi(\theta)$ es llamado el parámetro natural de la familia $\mathcal{F}$.

Si $Y_1, \ldots, Y_n$ son variables aleatorias independientes e idénticamente distribuidas, entonces

$$p(\mathbf{Y}|\theta) = \left(\prod_{i=1}^{n} f(y_i)\right) g^n(\theta) \exp\left\{\phi^T(\theta) \sum_{i=1}^{n} u(y_i)\right\}$$
$$\propto g^n(\theta) e^{\phi^T(\theta) t(\mathbf{y})},$$

donde $t(\mathbf{y}) = \sum_{i=1}^{n} u(y_i)$ es llamada la estadística suficiente. Si consideramos la distribución previa de la forma:

$$p(\theta) \propto g(\theta)^{\eta} e^{\phi^T(\theta)\nu},$$

entonces

$$p(\theta|\mathbf{Y}) \propto g(\theta)^{n+\eta} e^{\phi^T(\theta)(\nu + t(\mathbf{y}))}$$

Ejercicio 2, Distribución Normal con media desconocida

Sean $Y_1, \ldots, Y_n | \theta, \sigma^2 \overset{iid}{\sim} \text{Normal}(\theta, \sigma^2)$. Suponga que σ^2 es conocida y considere la distribución previa para $\theta \sim \text{Normal}(\mu_0, \tau_0^2)$. Obtenga la distribución posterior de θ.

En Estadística Bayesiana es común parametrizar la distribución Normal en términos de $\lambda = \sigma^{-2}$. A λ se le conoce como la "precisión".

Es decir, una distribución previa Normal es conjugada para la media de la verosimilitud normal con varianza conocida.
Podríamos haber llegado a esta conclusión usando nuestro resultado anterior.
Si $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathsf{Normal}(\theta, \sigma^2)$ con σ^2 conocida, entonces

$$
\begin{aligned}
p(\mathbf{Y}|\theta) &= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n} y_i^2 - 2\theta\sum_{i=1}^{n} y_i + n\theta^2\right]\right\} \\
&\propto \exp\left\{-\frac{n\theta^2}{2\sigma^2}\right\}\exp\left\{\frac{\theta}{\sigma^2}\sum_{i=1}^{n} y_i\right\}
\end{aligned}
$$

La estadística suficiente es entonces $t(\mathbf{y}) = \frac{1}{\sigma^2}\sum_{i=1}^{n} y_i$ y

$$p(\mathbf{Y}|\theta) \propto \exp\left\{-\frac{1}{2\sigma^2}\theta^2\right\}^n \exp\left\{\theta t(\mathbf{y})\right\}.$$

Por lo tanto, la previa conjugada toma la forma:

$$\begin{aligned} p(\theta) &\propto \exp\left\{-\frac{1}{2\sigma^2}\theta^2\right\}^\eta \exp\left\{\theta\nu\right\} \\ &= \exp\left\{-\frac{\eta}{2\sigma^2}\theta^2 + \theta\nu\right\} \\ &\propto \exp\left\{A\theta^2 + B\theta + C\right\} \\ &= \exp\left\{-\frac{1}{2\tau_0^2}(\theta - \mu_0)^2\right\} \end{aligned}$$

Distribución Normal con varianza desconocida

Considere $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathsf{Normal}(\theta, \sigma^2)$ con θ conocida, entonces

$$p(\mathbf{Y}|\sigma^2) \propto \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right\}$$
$$= (\sigma^2)^{-n/2} \exp\left\{-\frac{n}{2\sigma^2}t(\mathbf{y})\right\},$$

donde $t(\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i-\theta)^2$. Luego, la previa conjugada para σ^2 tiene la forma:

$$p(\sigma^2) \propto (\sigma^2)^{-\alpha+1} e^{-\beta/\sigma^2}.$$

Es decir, σ^2 sigue una distribución Inversa Gama con hiperparámetros α, β.

Ejercicio

Usando el resultado sobre el modelo conjugado para la familia exponencial demuestre los siguientes modelos conjugados, y calcule la distribución posterior:

- La distribución Beta para la verosimilitud Binomial.
- La distribución Gama para la verosimilitud Exponencial, $p(Y|\theta) = \theta e^{-\theta y} \mathbb{1}_{(0,\infty)}(y)$.
- La distribución Gama para la verosimilitud Poisson.

Distribuciones predictivas

Con las familias conjugadas, el conocer la forma de la distribución previa y la distribución posterior puede ser usado para hallar distribuciones marginales, como la predictiva previa, $p(Y)$, usando la fórmula:

$$p(Y) = \frac{p(Y|\theta)p(\theta)}{p(\theta|Y)}.$$

Considere el modelo Gama-Poisson, demuestre que la predictiva previa tiene la forma:

$$p(Y) = \binom{\alpha + y - 1}{y} \left(\frac{\beta}{\beta+1}\right)^{\alpha} \left(\frac{1}{\beta+1}\right)^{y} \mathbb{1}_{\{0,1,\ldots\}}(y).$$

Es decir Y sigue una distribución Binomial Negativa con parámetros α y β.

Modelo Poisson con exposición

En varias aplicaciones, es conveniente extender el modelo Poisson a la forma:

$$Y|x,\theta \sim \text{Poisson}(x\theta).$$

En epidemiología, el parámetro θ es llamada la tasa y x es llamada la exposición.
Considere que se cuenta con una muestra $Y_i|x_i,\theta \overset{iid}{\sim} \text{Poisson}(x_i\theta)$. Demuestre que la distribución gama es conjugada y que la posterior está dada por

$$\theta|\mathbf{Y} \sim \text{Gama}\left(\alpha + \sum_{i=1}^{n} y_i, \beta + \sum_{i=1}^{n} x_i\right).$$

Mostrar ejemplos de asma y de cáncer de riñón.

Análisis de Referencia

A veces se buscan distribuciones previas que tengan poco impacto en la distribución posterior. Dichas distribuciones son llamadas distribuciones previas de referencia, las cuales son descritas como vagas, constantes, difusas o no informativas.

Por ejemplo, en el caso de la distribución Normal con media θ y varianza conocida σ^2, al considerar una previa normal con media μ_0 y varianza τ_0^2, obtuvimos:

$$\theta|\mathbf{Y} \sim \mathsf{Normal}(\mu_n, \tau_n^2),$$

donde

$$\mu_n = \frac{\frac{1}{\tau_0}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0} + \frac{n}{\sigma^2}} \quad \text{y} \quad \frac{1}{\tau_n^2} = \frac{1}{\tau_0} + \frac{n}{\sigma^2}.$$

Distribuciones impropias

Si hacemos $\tau_0 \to \infty$, entonces

$$\theta|\mathbf{Y} \sim \text{Normal}\left(\bar{y}, \frac{\sigma^2}{n}\right),$$

note que si $\tau_0 \to \infty$, entonces $p(\theta) \propto \mathbb{1}_{\mathbb{R}}(\theta)$. Lo que va en orden con el principio de razón insuficiente.

Note que la función $\mathbb{1}_{\mathbb{R}}(\theta)$ no posee integral finita, y por lo tanto no hay manera de normalizarla para que integre 1. Por lo tanto, no hay manera de obtener una densidad y no determina una distribución en sentido estricto. A este tipo de "densidades" que no poseen integral finita se les conoce como densidades impropias.

Retomando el caso binomial, habíamos propuesto $\theta \sim \text{Uniforme}(0, 1)$. Pero suponga que estamos interesados en $\phi = -\log \theta$, si $\theta \sim \text{Uniforme}(0, 1)$, entonces $\phi \sim \text{Exponencial}(1)$. Lo que viola el principio de razón insuficiente.

Esta ambigüedad en la que no está claro qué debe ser uniforme puede conducir a importantes contradicciones.

Mostrar paradoja de Bertrand

Función score

Suponga que $Y \sim p(\cdot|\theta_0)$, definimos la función score como:

$$sc(\theta) = \frac{d}{d\theta} \log p(y|\theta).$$

$$\begin{aligned}\mathbb{E}_{Y|\theta_0}[sc(\theta)] &= \int_{\mathcal{Y}} \left[\frac{d}{d\theta} \log p(y|\theta)\right] p(y|\theta_0) dy \\ &= \int_{\mathcal{Y}} \frac{p(y|\theta_0)}{p(y|\theta)} \frac{d}{d\theta} p(y|\theta) dy.\end{aligned}$$

Entonces

$$\mathbb{E}_{Y|\theta_0}[sc(\theta_0)] = \int_{\mathcal{Y}} \frac{d}{d\theta} p(x|\theta)\Big|_{\theta=\theta_0} dy$$

Condiciones de regularidad

Si se pueden intercambiar las operaciones de integración y derivación, entonces

$$\begin{aligned}\mathbb{E}_{Y|\theta_0}[sc(\theta_0)] &= \frac{d}{d\theta}\left[\int_{\mathcal{Y}} p(y|\theta)dy\right]_{\theta=\theta_0} \\ &= \frac{d}{d\theta}1\Big|_{\theta=\theta_0} \\ &= 0.\end{aligned}$$

Información (esperada) de Fisher

Por otro lado, definimos la información esperada de Fisher por unidad muestral como:

$$\begin{aligned}\mathcal{I}_{\theta_0}(\theta) &= \mathbb{E}_{Y|\theta_0}[sc^2(\theta)] \\ &= \mathbb{E}_{Y|\theta_0}\left[\frac{1}{p^2(Y|\theta)}\left(\frac{d}{d\theta}p(Y|\theta)\right)^2\right].\end{aligned}$$

Note que

$$\begin{aligned}\frac{d^2}{d\theta^2}\log p(Y|\theta) &= \frac{d}{d\theta}\left[\frac{1}{p(Y|\theta)}\frac{d}{d\theta}p(Y|\theta)\right] \\ &= -\frac{1}{p^2(Y|\theta)}\left(\frac{d}{d\theta}p(Y|\theta)\right)^2 + \frac{1}{p(Y|\theta)}\frac{d^2}{d\theta^2}p(Y|\theta).\end{aligned}$$

Entonces,

$$-\mathbb{E}_{Y|\theta_0}\left[\frac{d^2}{d\theta^2}\log p(Y|\theta)\right] = \mathcal{I}_{\theta_0}(\theta) - \mathbb{E}_{Y|\theta_0}\left[\frac{1}{p(Y|\theta)}\frac{d^2}{d\theta^2}p(Y|\theta)\right].$$

Al evaluar en θ_0 y suponiendo que se pueden intercambiar las operaciones de integración y derivación, obtenemos que

$$\mathcal{I}_{\theta_0}(\theta_0) = -\mathbb{E}_{Y|\theta_0}\left[\frac{d^2}{d\theta^2}\log p(Y|\theta)\Big|_{\theta=\theta_0}\right].$$

Por lo tanto, bajo condiciones de regularidad

$$\mathbb{E}_{Y|\theta_0}[sc(\theta_0)] = 0, \quad \mathbb{V}_{Y|\theta_0}[sc(\theta_0)] = \mathcal{I}_{\theta_0}(\theta_0)$$

y

$$J(\theta_0) \equiv \mathcal{I}_{\theta_0}(\theta_0) = -\mathbb{E}_{Y|\theta_0}\left[\frac{d^2}{d\theta^2}\log p(Y|\theta)\Big|_{\theta=\theta_0}\right].$$