Análisis Estadístico **Grado en Ingeniería Matemática, UAX**

Tema 6: Inferencia bayesiana

Salvador Robles *curso 2020/2021*

Resumen: *Estudiamos la estimación puntual, la estimación por intervalo y el contraste de hipótesis desde el punto de vista de la inferencia bayesiana (a modo de introducción en el tema).*

Índice

6.1. Introducción a la inferencia bayesiana

6.1.1. Inferencia bayesiana vs. inferencia frecuentista

Hasta ahora hemos estado considerando siempre una interpretación frecuentista de la probabilidad, donde la idea intuitiva es que una medida de probabilidad mide la frecuencia relativa de un suceso favorable en el límite en el que repitiésemos el experimento un número infinito (o muy grande) de veces. En algunos casos esta interpretación se puede implementar sin dificultades. Por ejemplo, podemos lanzar un dado tantas veces como queramos y estudiar la frecuencia relativa de los posibles resultados. Pero en otros solo se puede plantear de forma hipotética ya que hay experimentos aleatorios que no son repetibles, bien porque porque la realización del experimento modifica las condiciones del próximo experimento o bien porque las características de la población bajo estudio van cambiando con el tiempo y la realización del mismo experimento nunca cumple las mismas condiciones de partida.

Por ejemplo, supongamos que monitorizamos la opinión en la sociedad sobre un cierto caso de corrupción política o sobre cualquier otro asunto de actualidad a través de una variable aleatoria. Las encuestas (muestras) que podamos tomar a lo largo del tiempo no pueden considerarse repeticiones del mismo proceso aleatorio porque las condiciones de partida necesariamente cambian: la información que va apareciendo en los medios de comunicación va modificando la opinión que las personas tienen sobre dicho asunto, e incluso la difusión de los resultados de una primera encuesta puede (y suele de hecho hacerlo) modificar su opinión

[0]Estas notas constituyen los apuntes informales para el seguimiento de la asignatura Análisis Estadístico del grado de Ingeniería matemática. No pretenden (y no lo hacen) sustituir al material bibliográfico recomendado para la asignatura. Deben considerarse por tanto como una guía de estudio de la asignatura que hay que trabajar conjuntamente con la bibliografía y con las hojas de ejercicios propuestos. Además, son una primera versión de los mismos así que pueden contener pequeños errores tipográficos o de otro tipo, cuya corrección se irá incorporando en sucesivas versiones.

y por tanto las condiciones para una segunda encuesta. Otro ejemplo es la predicción del tiempo. Cuando los meteorólogos pronostican para un cierto día una probabilidad de lluvia del 80 % no están diciendo que tenemos que considerar una cantidad muy grande de días exactamente iguales y comprobar que en el 80 % de ellos lloverá y en el resto hará un día soleado. En estos casos la probabilidad es más una medida de la creencia del meteorólogo, basada eso sí en ciertos modelos matemáticos, de cómo de plausible es la realización de un resultado particular, en este caso la lluvia.

Por otro lado, la inferencia estadística que realizamos en la aproximación frecuentista de la estadística, o *estadística clásica*, sobre los valores de un parámetro desconocido θ no tienen en cuenta los datos de las muestras. Estas solo se utilizan para confirmar o descartar hipótesis o planteamientos sobre dicho parámetro pues no intervienen en el cálculo de las distribuciones muestrales, y las inferencias realizadas sobre el valor de θ se basan en la distribución muestral, que es en realidad una distribución teórica sobre todos los posibles resultados que se pueden obtener en una muestra asumiendo que las v.a. que las representan siguen la distribución poblacional. El valor obtenido en la muestra se utiliza después para comprobar si cae en el rango de lo que se esperaría a partir de las hipótesis de partida. Como hemos visto en los temas anteriores, obtenemos y utilizamos intervalos de confianza sobre lo que esperamos que debiera producirse al obtener la muestra, y los datos obtenidos en una muestra no modifican de ninguna forma la distribución muestral. En ese sentido, si una vez realizada una muestra pretendo llevar a cabo la toma de una segunda muestra, la distribución muestral es exactamente la misma (recordemos que la distribución muestral es una distribución teórica).

Este planteamiento no parece acomodarse bien al estudio de fenómenos aleatorios que van evolucionando con el tiempo. En estos casos, los valores que obtenemos en una muestra concreta tomada en un determinado momento deberían hacernos reflexionar sobre cómo se modifican las probabilidades del experimento aleatorio para los momentos posteriores. Por ejemplo, supongamos que la evolución de los mercados de valores viene dada por una cierta variable aleatoria. Los valores que las acciones toman en un determinado día (o en una serie de ellos) pueden modificar (y de hecho en general lo hacen) las probabilidades de los distintos resultados que pueden darse en los días siguientes. En el desarrollo frecuentista de la inferencia estadística que hemos visto no es fácil calcular de forma exhaustiva (al menos no de forma directa) la distribución de probabilidad que pueden tomar los valores de una muestra en el futuro condicionada a los valores concretos que la muestra ha tomado en algún momento anterior.

Para abordar algunos de estos inconvenientes o limitaciones se ha desarrollado lo que se conoce como inferencia o estadística *bayesiana*, cuyo nombre hace referencia, como veremos, al uso del concepto de probabilidad condicionada y sobre todo del teorema de Bayes. En la inferencia estadística frecuentista los parámetros sobre los que realizamos la inferencia son valores fijos, aunque desconocidos. En la inferencia bayesiana, por el contrario, el valor de estos parámetros va a estar representado por una variable aleatoria. Es decir, su valor no es fijo sino que es un valor aleatorio que dependerá de los procesos subyacentes al problema general.

Desde este punto de vista, si consideramos que el parámetro θ es una variable aleatoria tendrá entonces asociada una cierta distribución de probabilidad, a la que llamaremos *distribución a priori*, y que será subjetiva en el sentido de que la elección concreta de dicha distribución no vendrá dada por los datos de la muestra sino por nuestras creencias o intuiciones sobre los resultados que esperaríamos obtener o que creemos más o menos probables. En la práctica esta arbitrariedad no es absoluta ya que de alguna forma las distribuciones a priori vendrán sugeridas por la experiencia que tengamos sobre el suceso concreto o sobre sucesos relacionados o análogos. Pero en teoría, sobre todo si no tenemos información previa del fenómeno, podemos utilizar cualquier distribución a priori. Este hecho parece poco riguroso y es en efecto el punto débil de la inferencia bayesiana. Sin embargo, hay que tener en cuenta dos cosas:

Por un lado, esta información subjetiva que incorporamos al introducir la distribución a priori de los parámetros puede convertirse en objetiva[1] si como hemos dicho anteriormente nos basamos para hacerlo en expe-

[1] Más o menos objetiva. La objetividad absoluta no existe.

riencias previas o análogas. Desde esta perspectiva, el punto débil puede convertirse en el punto fuerte si introducir la distribución a priori significa incorporar información relevante sobre el parámetro en cuestión que de otra manera (en los desarrollos frecuentistas) no estaríamos considerando.

Por otro lado, veremos que los datos de una muestra particular van a modificar los valores de distribución a priori produciendo una nueva distribución de probabilidad que llamaremos *distribución a posteriori*. En este caso, cuando el tamaño de la muestra es grande la dependencia de la distribución a posteriori con respecto a la elección particular de la distribución a priori elegida es menor. Se dice entonces que los datos saturan la distribución a priori.

Por último, veremos que una de las principales ventajas de la aproximación bayesiana a la inferencia estadística es que su planteamiento se acomoda perfectamente al estudio estadístico de fenómenos que evolucionan en el tiempo. En esta aproximación es relativamente sencillo plantear la distribución de probabilidad para los valores de una muestra futura en función o condicionada a los valores que una muestra similar proporcionó en el pasado, lo que nos será de gran utilidad en los desarrollos que se verán en los próximos cursos.

De todas formas, hay que mencionar que estas dos aproximaciones a la inferencia estadística son en la mayoría de los casos complementarias y dependerá del problema concreto que una u otra proporcionen mejores resultados.

6.1.2. Distribuciones a priori y a posteriori

Para fijar conceptos, recordemos que el objetivo de partida de la inferencia estadística es realizar afirmaciones sobre los valores de un parámetro θ de la distribución poblacional a partir de los valores obtenidos en una muestra particular. Como hemos dicho, la idea principal de la aproximación bayesiana es considerar al parámetro θ como una v.a. que tiene asociada por tanto una distribución de probabilidad, $g(\theta)$. Esta distribución es la que hemos llamado *distribución a priori*. Por otro lado, sabemos de los temas anteriores que las muestras (o las v.a. que las representan) siguen otra distribución de probabilidad, la distribución muestral, $f(\vec{x};\theta)$, que por supuesto depende del valor del parámetro θ. La diferencia esencial es que en la aproximación frecuentista este valor era fijo aunque desconocido y ahora consideramos que es una v.a.

Desde este planteamiento general, y una vez tomada una muestra concreta, $\vec{x}_0$, podemos hacer uso del teorema de Bayes para calcular la distribución de probabilidad, $g(\theta|\vec{x}_0)$, de los valores del parámetro θ condicionada a los valores concretos que hemos obtenido en dicha muestra, es decir,

$$g(\theta|\vec{x}_0) = \frac{g(\theta)f(\vec{x}_0|\theta)}{\int_\Theta d\theta g(\theta)f(\vec{x}_0|\theta)}. \tag{6.1}$$

Aquí hay que puntualizar varias cosas. En (6.1) $\vec{x}_0$ no es una variable (aleatoria) sino que es el valor concreto que hemos obtenido en la muestra. Por tanto, $g(\theta|\vec{x}_0)$ no es una función de dos variables, θ y $\vec{x}_0$, sino que funcionalmente solo depende de la variable aleatoria θ (de hecho es su función de distribución). Segundo, la integración del denominador en (6.1), que en caso de ser θ una v.a. discreta se convertiría en una suma, se realiza sobre todos los posibles valores (continuos o discretos) del parámetro θ. Por tanto, es una constante, es decir, es un valor que solo depende de $\vec{x}_0$, que como hemos dicho es un valor dado. De este modo, (6.1) puede escribirse también como

$$g(\theta|\vec{x}_0) = C(\vec{x}_0)g(\theta)f(\vec{x}_0|\theta), \tag{6.2}$$

donde, $C(\vec{x}_0) = \left(\int_\Theta d\theta g(\theta)f(\vec{x}_0|\theta)\right)^{-1}$, es un factor de normalización, es decir, un factor que hace que se cumpla[2], $\int_\Theta d\theta g(\theta|\vec{x}_0) = 1$. La distribución de probabilidad $g(\theta|\vec{x}_0)$ obtenida a partir de la distribución a priori y de los datos tomados de una muestra es lo que se llama in la inferencia bayesiana *distribución a posteriori* del parámetro θ.

[2]Recordemos que toda distribución de probabilidad, $f(x)$, tiene que cumplir, $f(x) \geq 0$ y $\int dx f(x) = 1$.

Llegados a este punto el planteamiento inicial parece claro. A partir de la distribución muestral, evaluada para un valor concreto de la muestra, y junto con la distribución a priori calculamos la distribución a posteriori del parámetro θ, $g(\theta|\vec{x}_0)$. Como la distribución muestral la conocemos de forma teórica a partir de la distribución poblacional, la clave del cálculo de la distribución a posteriori está entonces en qué distribución a priori elegir ya que una vez tomada esa decisión el cálculo de la distribución a posteriori es directo, salvo obviamente por dificultades técnicas.

Ejemplo 6.1 *Ejs. 11.2 y 11.3 en Ref. [2]*

6.1.2.1. Tipos de distribuciones a priori

Según las características de la distribución a priori esta se puede llamar de distintas formas. Por ejemplo, decimos que una distribución a priori es *imprecisa* o no informativa si proporciona el mismo peso (probabilidad) a todos los posibles valores del parámetro θ. Este tipo de distribuciones a priori se utilizan principalmente cuando no se tiene información previa acerca del parámetro θ. Por ejemplo, si la única información que disponemos de un parámetro θ es que puede tomar cualquier valor dentro del intervalo $(0, 1)$, podemos considerar la distribución de densidad uniforme como la distribución a priori a utilizar.

En ocasiones puede ocurrir que la distribución a priori $g(\theta)$ no sea integrable, es decir, que no se cumpla, $\int_\Theta g(\theta)d\theta < \infty$. Decimos entonces que $g(\theta)$ es una distribución a priori *impropia*. En estos casos $g(\theta)$ no es, estrictamente hablando, una distribución de probabilidad ya que para que lo sea sabemos que tiene que cumplirse, $\int_\Theta g(\theta)d\theta = 1$. Sin embargo, en lo que estamos principalmente interesados en la aproximación bayesiana es en la probabilidad a posteriori, que es la que nos proporciona la distribución de probabilidad del parámetro θ una vez que el valor de la muestra viene dado por un valor concreto. Por tanto, de lo que debemos asegurarnos es de que esta distribución a posteriori sí sea efectivamente una distribución de probabilidad. Resulta a veces que aunque la distribución a priori sea impropia la distribución a posteriori resultante es integrable, y por ese motivo se utiliza a veces este tipo de distribuciones a priori.

Por último, mencionar un tipo importante de distribuciones a priori que aparecen con frecuencia en los problemas de inferencia bayesiana, las distribuciones a priori *conjugadas*. Se dice que una distribución a priori es conjugada con respecto a la distribución poblacional $f(x|\theta)$ si la distribución a posteriori obtenida a partir del teorema de Bayes pertenece a la misma familia de distribuciones que la distribución a priori. Es decir, si el efecto en la distribución a priori de los datos de la muestra se traduce principalmente en una modificación de los parámetros de dicha distribución y no en su relación funcional.

Ejemplo 6.2 *Demostrar que la distribución Gamma es conjugada con respecto a la distribución exponencial.*

6.1.2.2. Medidas de centralización y dispersión a posteriori

La idea esencial de la estadística bayesiana es que la distribución a posteriori contiene toda la información que disponemos del parámetro θ. Por un lado, contiene la información subjetiva de la distribución a priori y, por otro lado, la información proporcionada por la muestra concreta de los datos. Como distribución de probabilidad que es podemos describirla a partir de todas las medidas de una distribución que hemos visto en los temas anteriores. Todas estas medidas las llamaremos 'a posteriori' para indicar que provienen de la distribución a posteriori. Entre ellas, las principales medidas de centralización son:

- *Moda a posteriori*: es el valor θ_m para el que la distribución a posteriori es máxima y para el que se cumple por tanto, $\partial_\theta f(\theta|\vec{x}_0)_{|\theta=\theta_m} = 0$. Es por tanto el análogo en la estadística bayesiana al estimador de máxima verosimilitud en la estadística frecuentista. Por eso puede ser interesante utilizarlo como

medida de centralización de la distribución a posteriori. Sin embargo, como sabemos también puede tener ciertas desventajas como son el hecho de que pueda caer alejada del centro de la distribución en aquellas distribuciones con colas largas o que puedan existir dos o más valores modales.

- Otra medida de centralización de la distribución a posteriori es la *mediana a posteriori*, que será el valor θ_M que deja tanto a la derecha como a la izquierda de la distribución el 50 % de los datos, es decir, el valor que cumple,
$$\int_{-\infty}^{\theta_M} g(\theta|\vec{x}_0) = 0'5.$$
Es una buena medida de centralización cuyo principal inconveniente es que tiene que calcularse numéricamente en la mayoría de los casos de interés.

- *Media a posteriori*, o esperanza matemática del parámetro θ, dada por el valor
$$\bar{\theta} = \int_{\Theta} \theta g(\theta|\vec{x}_0) d\theta.$$
Aunque puede verse bastante afectada en algunas distribuciones por el efecto de la cola lateral es, por regla general, la medida de centralización que más se utiliza pues es la que minimiza la desviación cuadrática de los valores de la distribución. Es la principal medida de centralización que utilizaremos nosotros para describir la distribución a posteriori.

Por otro lado, también podemos considerar las medidas de dispersión de la distribución a posteriori, entre las que destacamos la *varianza a posteriori*, y su raíz cuadrada, la desviación estándar a posteriori. La varianza a posteriori se define como
$$V = \int_{\Theta} (\theta - \bar{\theta})^2 g(\theta|\vec{x}_0) d\theta. \tag{6.3}$$
La desviación estándar a posteriori es la principal medida de dispersión que utilizaremos en las distribuciones a posteriori. También podemos calcular los percentiles de la distribución a posteriori de la forma habitual, como
$$k = 100 \times \int_{0}^{\theta_k} g(\theta|\vec{x}_0) d\theta,$$
y con ellos obtener el recorrido intercuartílico como medida de dispersión de la distribución a posteriori.

6.1.3. Estimación puntual, intervalos creíbles y test bayesianos

Para terminar esta breve introducción a la estadística bayesiana, no hay que olvidar que el objetivo principal de la inferencia estadística[3], y en este caso de la inferencia bayesiana, es poder hacer estimaciones y evaluar afirmaciones sobre el valor de los parámetros poblacionales a partir del valor de una o varias muestras. Como hemos visto en la inferencia frecuentista, las principales estimaciones que hemos estudiado son estimaciones puntuales o por intervalos y contrastes de hipótesis. En la estadística bayesiana la estimación puntual se obtiene mediante las medidas de centralización y dispersión a posteriori que hemos visto, o en general a partir de cualquier medida sobre la distribución a posteriori que se quiera evaluar[4]. Por otro lado, los análogos a los intervalos de confianza de la estadística frecuentista son los *intervalos creíbles* en la estadística bayesiana. Se construyen calculando los valores θ_r y θ_l que dejan a la derecha y a la izquierda de la distribución a posteriori, respectivamente, una probabilidad igual a $\alpha/2$. El intervalo $[\theta_l, \theta_r]$ determina entonces el intervalo para el que la probabilidad de encontrar el valor de θ es $1 - \alpha$.

[3]Hablamos siempre aquí de inferencia paramétrica

[4]Por ejemplo, la estimación puntual de los momentos de la distribución vendrá dada por el valor de los momentos de la distribución a posteriori.

Respecto de los test de hipótesis, el procedimiento para realizar los *test bayesianos* es distinto dependiendo de si el test es unilateral o bilateral. En los test unilaterales, del estilo de

$$H_0 : \theta \leq \theta_0 \quad , \quad H_1 : \theta > \theta_0, \tag{6.4}$$

al nivel de significación α, calculamos la probabilidad a posteriori de que la hipótesis nula sea verdadera, es decir, calculamos la probabilidad

$$P\left(H_0 : \theta \leq \theta_0 | \vec{x}_0\right) = \int_{-\infty}^{\theta_0} g(\theta|\vec{x}_0) d\theta, \tag{6.5}$$

y rechazamos la hipótesis nula si esta probabilidad es menor que el nivel de significación α.

En el caso de los test de hipótesis bilaterales para los que la hipótesis nula se corresponde con un valor concreto del parámetro, es decir, $H_0 : \theta = \theta_0$ frente a $H_1 : \theta \neq \theta_0$, no podemos seguir el procedimiento anterior ya que en tal caso, si la distribución a posteriori es una distribución continua, la función de probabilidad para cualquier valor concreto, en particular para $\theta = \theta_0$, es cero. En estos casos lo que haremos será calcular el intervalo creíble de la distribución a posteriori al nivel de credibilidad $(1 - \alpha) \times 100\,\%$, y rechazaremos la hipótesis nula si el valor θ_0 cae fuera de dicho intervalo. En caso contrario, aceptaremos (no rechazaremos) la hipótesis nula.

6.2. Inferencia bayesiana para la media

Comenzamos ahora a aplicar los métodos de la inferencia bayesiana a las principales distribuciones que hemos visto a lo largo del curso. En particular, calcularemos la estimación del valor p de la 'proporción de éxitos' en una distribución binomial, el valor de la media en una distribución de Poisson y el valor de la media en una distribución normal utilizando los métodos bayesianos. Haremos uso solo de dos distribuciones a priori: la distribución uniforme, que se utiliza cuando no disponemos de información a priori sobre la distribución del parámetro buscado, y de las distribuciones conjugadas, que como veremos nos simplifican enormemente los cálculos. Por supuesto, uno puede utilizar cualquier otra distribución a priori y realizar los cálculos numéricamente, pero en un curso introductorio como este las distribuciones uniformes y las distribuciones conjugadas nos facilitarán los cálculos y nos permiten por tanto centrarnos más en el propio desarrollo y en las interpretaciones.

6.2.1. De una población binomial

6.2.1.1. Distribuciones a priori y a posteriori

Recordemos que un experimento binomial es aquel que corresponde a la realización de n pruebas independientes cuyos resultados individuales solo pueden ser 'éxito' o 'fracaso'. En tal caso, llamamos p a la proporción de éxitos del total de pruebas y asumimos que dicha proporción o probabilidad se mantiene constante a lo largo de las sucesivas pruebas. Como ya hemos visto en temas anteriores, la probabilidad de obtener x 'éxitos' en un total de n pruebas es

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x} \quad , \; x = 0, 1, \ldots, n. \tag{6.6}$$

En los problemas de inferencia frecuentista p era un parámetro fijo, aunque posiblemente desconocido. Ahora sin embargo es al revés, $x = x_0$ será un valor fijo dado por el resultado de la muestra y la variable aleatoria

será precisamente el valor de la proporción, p. Por tanto, la distribución (6.6) hay que mirarla ahora como una función del parámetro p, es decir,

$$f(x_0|p) \propto p^{x_0}(1-p)^{n-x_0} \quad , \; 0 \leq p \leq 1, \tag{6.7}$$

donde hemos considerado explícitamente solo los términos que dependen de p (ya que x_0 y n son dos valores fijos). Como hemos visto en las secciones precedentes, para construir la probabilidad a posteriori necesitamos elegir una distribución a priori para el parámetro p, que llamaremos $g(p)$. Una vez elegida, la distribución a posteriori $g(p|x_0)$ viene dada por el teorema de Bayes

$$g(p|x_0) = \frac{g(p)f(x_0|p)}{\int_0^1 g(p)f(x_0|p)dp} = C_1\, g(p)\, f(x_0|p), \tag{6.8}$$

donde C_1 es una constante de normalización que garantiza que la distribución de probabilidad a posteriori cumple la condición de normalización,

$$\int_0^1 g(p|x_0)dp = 1. \tag{6.9}$$

Si no tenemos ninguna información previa sobre el valor de p o si queremos ser tan objetivos como nos sea posible podemos considerar como distribución a priori la distribución uniforme, que asigna el mismo valor de la probabilidad a todos los valores del intervalo $[0, 1]$, es decir,

$$g(p) = \begin{cases} 1 \quad \text{si} \quad 0 \leq p \leq 1 \\ \\ 0 \quad \text{en otro caso} \end{cases} \tag{6.10}$$

En tal caso, la probabilidad a posteriori coincide con la distribución de probabilidad del parámetro p dada por (6.7). Esta distribución, como función del parámetro p, es un caso particular de una distribución que se conoce con el nombre de *distribución beta*, $B(x; a, b)$, que depende de dos parámetros, a y b, y cuya densidad de probabilidad es,

$$f(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} \quad , 0 \leq x \leq 1, \tag{6.11}$$

donde

$$\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt, \tag{6.12}$$

es una integral que se puede calcular por métodos numéricos[5]. El factor delante de las potencias de x en (6.11) es el que asegura que $f(x; a, b)$ es una distribución de probabilidad,

$$\int_0^1 f(x; a, b)dx = 1 \quad \forall a, b \tag{6.13}$$

Por tanto, siempre que nos encontremos con una distribución de probabilidad cuya dependencia funcional en x sea $x^a(1-x)^b$ sabemos que su constante de normalización será, $\Gamma(a+b)/\Gamma(a)\Gamma(b)$. Esto nos va a ser muy útil para evitar cálculos integrales tediosos y para comprobar que la distribución beta es precisamente la distribución a priori conjugada para las observaciones binomiales. En efecto, supongamos que elegimos como distribución a priori para la proporción de éxitos en una serie de ensayos binomiales, p, la distribución beta, $B(p; a, b)$. En tal caso, la distribución a posteriori resulta

$$g(p|x_0) = C_2\, p^{a-1}(1-p)^{b-1}p^{x_0}(1-p)^{n-x_0} = C_2\, p^{a'-1}(1-p)^{b'-1} = B(a', b'), \tag{6.14}$$

[5]Si, $a = n$, es un número natural se puede calcular también de forma explícita y resulta, $\Gamma(n) = (n-1)!$.

donde C_2 es una constante de normalización y,

$$a' = a + x_0 \quad , \quad b' = b + n - x_0. \tag{6.15}$$

Es decir, el efecto de la muestra en la distribución a priori beta es un desplazamiento de los parámetros a y b de la distribución pero no su relación funcional. Por tanto no hay que calcular ninguna integral ni hacer ningún cálculo complejo, tan solo las sumas y restas que aparecen en (6.15). Esto es lo que hace que las distribuciones a priori conjugadas sean tan útiles y prácticas[6]. Por otro lado, de lo expuesto después de (6.13) sabemos que la constante de normalización C_2 en (6.14) es

$$\frac{\Gamma(a'+b')}{\Gamma(a')\Gamma(b')} = \frac{\Gamma(n+a+b)}{\Gamma(x_0+a)\Gamma(n-x_0+b)}. \tag{6.16}$$

Llegados a este punto, hay que destacar un par de cosas. Por un lado, la distribución beta representa en realidad a muchos y distintos tipos de distribuciones ya que según el valor de a y b su relación funcional puede ser bastante diferente. Por otro lado, existe una manera de incorporar información previa del parámetro p dentro de la distribución a priori, aun usando la misma distribución beta. Sabemos que la media y la desviación estándar de la distribución beta vienen dados por los valores,

$$E[B(x;a,b)] = \frac{a}{a+b}, \tag{6.17}$$

y

$$\sigma[B(x;a,b)] = \sqrt{\frac{ab}{(a+b)^2(a+b+1)}}, \tag{6.18}$$

respectivamente. Por tanto, si pensamos que la distribución a priori del parámetro p debería tener un valor medio p_0 y una desviación estándar σ_0 podemos elegir como distribución a priori una distribución beta con los parámetros a y b que resuelvan las ecuaciones

$$p_0 = \frac{a}{a+b} \quad , \quad \sigma_0 = \sqrt{\frac{p_0(1-p_0)}{(a+b+1)}}. \tag{6.19}$$

Por último, destacar que cuando el número de datos es muy elevado el efecto de la distribución a priori puede llegar a ser muy pequeño. Se dice entonces que los datos *saturan* la distribución a priori. Esto puede verse fácilmente a partir de los valores dada en (6.15). Si n y x_0 son muy grandes el efecto de los valores a y b de la distribución a priori es menor ya que resulta, $a' \approx x_0$ y $b' \approx n - x_0$.

6.2.1.2. Estimación del parámetro p

Como dijimos en la primera sección, para realizar estimaciones sobre el valor del parámetro p utilizaremos la distribución a posteriori calculada en el apartado anterior. Por ejemplo, para hacer **estimación puntual** del parámetro p utilizaremos la media a posteriori. Si hemos utilizado la función beta como distribución a priori resulta que la media a posteriori es,

$$\hat{p} = \frac{a'}{a'+b'}, \tag{6.20}$$

[6]Distribuciones conjugadas a priori solo existen cuando la distribución de la observación, $f(x_0|\theta)$, es un miembro de la familia *exponencial*, es decir, cuando puede escribirse como

$$f(x_0|\theta) = a(\theta)b(x_0)e^{c(\theta)\times d(x_0)},$$

y en estos casos resulta de especial utilidad debido a la propiedad que se acaba de ver.

y para su desviación estándar utilizaremos igualmente la desviación estándar a posteriori que en el caso que nos ocupa resulta ser,

$$\hat{\sigma}_p = \sqrt{\frac{a'b'}{(a'+b')^2(a'+b'+1)}}. \tag{6.21}$$

Para calcular el **intervalo creíble** para el parámetro p tendríamos que calcular los percentiles $\theta_{k=97'5\,\%}$ y $\theta_{k=2'5\,\%}$ de la distribución beta a posteriori utilizando cualquier aplicación de cálculo numérico. Para hacer los cálculos 'a mano' podemos aproximar la distribución a posteriori $B(p; a', b')$ por una distribución normal de media $\hat{p}$ y desviación estándar $\hat{\sigma}_p$ dados por (6.20) y (6.21), respectivamente. En ese caso, el intervalo creíble resultará ser

$$I[p] = [\hat{p} \pm z_{\alpha/2}\hat{\sigma}_p], \tag{6.22}$$

teniendo en cuenta que la aproximación de la distribución beta por la distribución normal funciona relativamente bien para valores, $a' \geq 10$ y $b' \geq 10$.

Respecto de los **test de hipótesis**, como dijimos en la primera parte del tema, cuando el test de hipótesis sea bilateral calcularemos el intervalo creíble y rechazaremos la hipótesis nula, $H_0 : p = p_0$, si el valor de p_0 cae fuera del intervalo creíble. En el caso de los test de hipótesis unilaterales tenemos que calcular la probabilidad de la hipótesis nula $H_0 : p \leq p_0$ como

$$P(p \leq p_0) = \int_0^{p_0} B(p; a', b')dp, \tag{6.23}$$

y ver si esta probabilidad es menor que el nivel de significación α, en cuyo caso rechazaremos la hipótesis nula H_0. Como hemos dicho, en el caso de que los valores de a' y b' cumplan, $a' \geq 10$ y $b' \geq 10$, la distribución beta se puede aproximar por la distribución normal, $N(\hat{p}, \hat{\sigma}_p)$, y en tal caso los valores de $P(p \leq p_0)$ se pueden calcular a partir de las tablas de la distribución normal tipificando la variable p.

6.2.2. De una distribución de Poisson

Recordemos que la distribución de Poisson se utiliza para contar el número de ocurrencias de eventos relativamente extraños que suceden aleatoriamente a lo largo del tiempo a un promedio constante, λ, teniendo en cuenta que los sucesos deben ocurrir cada uno en un instante distinto (no pueden ocurrir dos eventos a la vez).

Como sabemos, la probabilidad de que sucedan x sucesos en el intervalo de tiempo considerado viene dada por la distribución

$$f(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}\ , x = 0, 1, 2, \dots\ . \tag{6.24}$$

Como función de λ esta densidad de probabilidad puede escribirse como

$$f(x|\lambda) \propto e^{-\lambda}\lambda^x\ , \lambda > 0, \tag{6.25}$$

que tiene la forma de una distribución gamma, $\Gamma(a, b)$, cuya densidad de probabilidad viene dada por la función

$$f(x; a, b) = \frac{b^a x^{a-1} e^{-bx}}{\Gamma(a)}, \tag{6.26}$$

donde $\Gamma(a)$ está definida en (6.12). La función (6.25) es un caso particular de la función gamma (6.26), con $a = x + 1$ y $b = 1$.

Por otro lado, supongamos que tomamos una muestra de tamaño n, $x_1, \ldots, x_n$. En tal caso, la densidad de probabilidad de la muestra para un valor λ de la distribución de Poisson resulta

$$f(x_1, \ldots, x_n|\lambda) = \prod_{i=1}^{n} f(x_i|\lambda) \propto \lambda^{\sum x_i} e^{-n\lambda}, \tag{6.27}$$

que también tiene la forma de una distribución $\Gamma(a, b)$ con, $a = \sum_i x_i + 1$ y $b = n$. Resulta entonces que la distribución gamma es la distribución a priori conjugada de la distribución de Poisson de manera que si utilizamos la función, $g(\lambda) = \Gamma(\lambda; a, b)$, como distribución a priori la distribución a posteriori resulta,

$$g(\lambda, \vec{x}) \propto g(\lambda) f(\vec{x}|\lambda) \propto \lambda^{a-1} e^{-b\lambda} \lambda^{n\bar{x}} e^{-n\lambda} = \lambda^{a'} e^{-b'\lambda}, \tag{6.28}$$

que tiene la forma de una distribución $\Gamma(a', b')$, con los parámetros a' y b' desplazados respecto de los valores a y b de la distribución a priori[7],

$$a' = a + \sum_i x_i \quad , \quad b' = b + n. \tag{6.29}$$

Sabemos además que la media y la varianza de la distribución $\Gamma(r', v')$ vienen dadas por,

$$E(\lambda|\vec{x}) = \frac{a'}{b'} \quad , \quad V(\lambda|\vec{x}) = \frac{a'}{(b')^2}, \tag{6.30}$$

que serán los estimadores a posteriori de la media y la varianza. Como sucedía en el caso de la distribución binomial, una forma especialmente interesante de elegir la distribución a priori es entonces tomar una distribución gamma, $\Gamma(a, b)$, con los parámetros iniciales a y b dados por,

$$a = \frac{\lambda_0^2}{\sigma_\lambda^2} \quad , \quad b = \frac{\lambda_0}{\sigma_\lambda^2}, \tag{6.31}$$

que son los valores asociados a una distribución gamma con media λ_0 y varianza σ_λ^2.

Ejemplo 6.3 *Ej. 11.7 en Ref. [2]*

Ejemplo 6.4 *Ej. 17 en Ref. [1]*

Los valores dados en (6.30) son precisamente los que utilizaremos como estimadores a posteriori de la media y la varianza de la distribución de Poisson. Para construir los intervalos creíbles utilizaremos también la aproximación de la función gamma $\Gamma(a, b)$ a la distribución normal, $N(\frac{a}{b}, \frac{\sqrt{a}}{b})$, que puede considerarse una buena aproximación para, $a > 30$. En ese caso, el intervalo creíble bayesiano para las observaciones de un proceso de Poisson puede escribirse como,

$$I[\lambda] = [\hat{\lambda} \pm z_{\alpha/2} \hat{\sigma}_\lambda]. \tag{6.32}$$

Para los test de hipótesis seguiremos los mismos pasos que vimos con la distribución binomial. Cuando el test de hipótesis sea bilateral calcularemos el intervalo creíble y rechazaremos la hipótesis nula, $H_0 : \lambda = \lambda_0$, si el valor de λ_0 cae fuera del intervalo creíble. Para los test de hipótesis unilaterales tendríamos que calcular la probabilidad de la hipótesis nula $H_0 : \lambda \leq \lambda_0$ como

$$P(\lambda \leq \lambda_0) = \int_0^{\lambda_0} \Gamma(\lambda; r', v') d\lambda, \tag{6.33}$$

y ver si esta probabilidad es menor que el nivel de significación α, en cuyo caso rechazaremos la hipótesis nula H_0. Para no tener que calcular numéricamente la integral, utilizaremos en la práctica (cuando $a > 30$) la aproximación a la normal que se ha comentado anteriormente.

[7] La distribución uniforme puede verse como un caso límite de la distribución gamma, $\Gamma(a, b)$, con $a = 1$ y $b = 0$. Por tanto, haber elegido la distribución uniforme como distribución a priori equivale a tomar estos dos valores en (6.29).

6.2.3. De una distribución normal

Supongamos ahora que tenemos una muestra de tamaño n, $x_1, \ldots, x_n$, de una población que sigue una distribución normal de media μ desconocida y varianza σ^2 conocida. La probabilidad conjunta de la muestra, dado el valor μ de la media, puede manipularse (ver Refs. [1, 2]) para escribirse como

$$f(x_1, \ldots, x_n|\mu) \propto e^{-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2}. \tag{6.34}$$

Si elegimos una distribución a priori uniforme, $g(\mu) = 1$, la probabilidad a posteriori que resulta es precisamente la densidad de probabilidad (6.34), es decir,

$$g(\mu|x_1, \ldots, x_n) \propto e^{-\frac{n}{2\sigma^2}(\mu-\bar{x})^2}. \tag{6.35}$$

Si en cambio elegimos como distribución a priori una densidad de probabilidad normal[8] de media m y varianza s^2,

$$g(\mu) \propto e^{-\frac{1}{2s^2}(\mu-m)^2}, \tag{6.36}$$

resulta una probabilidad a posteriori que también tiene la forma de una distribución normal de media m' y desviación estándar s', es decir,

$$f(\mu|x_1, \ldots, x_n) \propto e^{-\frac{1}{2(s')^2}(\mu-m')^2}, \tag{6.37}$$

con (ver, Ref. [2])

$$m' = \frac{\sigma^2 m + n s^2 \bar{x}}{\sigma^2 + n s^2} \quad , \quad (s')^2 = \frac{\sigma^2 s^2}{(\sigma^2 + n s^2)}. \tag{6.38}$$

El efecto de la muestra sobre la distribución a priori normal es desplazar la media m al valor m' y modificar la desviación s al valor s'. Estas expresiones pueden escribirse de forma más simplificada definiendo las *precisiones* β de las distribuciones muestral y a priori como

$$\beta_{\text{muestral}} = \frac{n}{\sigma^2} \quad , \quad \beta_{\text{priori}} = \frac{1}{s^2}. \tag{6.39}$$

En tal caso resulta que la precisión de la distribución a posteriori es aditiva, es decir, es la suma de las precisión de la distribución a priori más la precisión de la distribución de la muestra, esto es,

$$\beta_{\text{posteriori}} = \beta_{\text{muestral}} + \beta_{\text{priori}}, \tag{6.40}$$

y la media de la distribución a posteriori resulta ser la media ponderada de las medias de las distribuciones a priori y de la muestra,

$$m' = \frac{\beta_{\text{priori}}}{\beta_{\text{posteriori}}} m + \frac{\beta_{\text{muestral}}}{\beta_{\text{posteriori}}} \bar{x}, \tag{6.41}$$

donde los pesos de cada media resultan ser las proporciones de sus respectivas precisiones en la precisión total o a posteriori. Si el tamaño de la muestra n es muy grande, entonces, teniendo en cuenta los valores dados en (6.39), se cumple que, $\beta_{\text{muestral}} \gg \beta_{\text{priori}}$, y por tanto,

$$\beta_{\text{posteriori}} \approx \beta_{\text{muestral}} \quad , \quad m' \approx \bar{x}, \tag{6.42}$$

es decir, como decíamos anteriormente, la muestra satura en estos casos la distribución a priori.

Como la distribución a posteriori es una normal con media m' y desviación típica s', los intervalos creíbles para el valor de μ al $(1-\alpha)\cdot 100$ de credibilidad vendrán dados por

$$I[\mu] = [m' \pm z_{\alpha/2} s'], \tag{6.43}$$

[8] La distribución normal es la familia conjugada para las observaciones de distribuciones normales con varianza conocida.

que serán también los intervalos que utilizaremos para los test de hipótesis bilaterales. En el caso de que la varianza σ^2 sea desconocida utilizaremos para el cálculo de s' la cuasivarianza muestral,

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2, \tag{6.44}$$

en vez de la varianza σ^2, pero entonces el intervalo de credibilidad (al nivel $(1-\alpha)\cdot 100$ de credibilidad) vendrá dado por

$$I[\mu] = [m' \pm t_{n-1,\alpha/2}s'], \tag{6.45}$$

donde $t_{n-1,\alpha/2}$ es el valor de la distribución t de Student de $n-1$ grados de libertad que deja a la derecha una probabilidad $\alpha/2$.

Por último, para los test de hipótesis unilaterales, utilizaremos los valores de la distribución normal para calcular la probabilidad de la hipótesis nula, $P(\mu \leq \mu_0)$, tipificando la variable μ, y la rechazaremos solo en el caso en que dicha probabilidad está por debajo del nivel de significación elegido, α.

6.3. Distribuciones predictivas bayesianas

Una de las ventajas de la estadística bayesiana es la relativa facilidad para desarrollar un método que nos proporcione la probabilidad de una observación futura dada la condición de que las n observaciones pasadas vienen dadas por unos ciertos valores, es decir,

$$f(x_{n+1}|x_1,\ldots,x_n). \tag{6.46}$$

Esto es lo que se denomina *distribución predictiva* y es fundamental a la hora de poder predecir o estimar resultados futuros en fenómenos aleatorios. El proceso es el siguiente. Como sabemos, las $n+1$ variables aleatorias que representan la muestra, $x_1,\ldots,x_n,x_{n+1}$, siguen una distribución $f(x_i|\mu), \forall i = 1,\ldots,n+1$. En particular,

$$f(x_{n+1}|\mu), \tag{6.47}$$

que por ejemplo puede ser cualquiera de las distribuciones que hemos visto en este tema (binomial, Poisson, normal, ...). Para concretar, asumiremos que es una distribución normal de media μ desconocida y varianza σ^2 conocida. Desde la perspectiva de la estadística frecuentista la idea más coherente sería estimar el parámetro μ a partir de la muestra $x_1,\ldots,x_n$, y con ese valor, llamémosle $\hat{\mu}$ (que en el caso que nos ocupa podría ser la media muestral, $\bar{x}$), calcular la distribución, $f(x_{n+1}|\hat{\mu})$. Sin embargo, haciéndolo así estaríamos dándole demasiada importancia al valor $\hat{\mu}$ ya que estaríamos asumiendo que es el único valor del parámetro μ que contribuye al comportamiento futuro de los datos, sin tener en cuenta que dicho valor es tan solo una medida (más o menos) representativa del valor muestral de μ pero con una incertidumbre asociada a dicho valor dada por la dispersión de los datos muestrales, que puede ser en algunos casos muy grande.

En la aproximación bayesiana, el proceso es diferente. Por un lado, hemos visto que el resultado de combinar la distribución muestral con una distribución a priori resulta, mediante el teorema de Bayes, en una distribución a posteriori, $g(\mu|x_1,\ldots,x_n)$, que nos proporciona el peso relativo que tienen los distintos valores del parámetro μ. La idea entonces es no tener en cuenta solo un valor concreto de μ para predecir el comportamiento futuro de los datos sino calcular la suma ponderada de la contribución de todos estos valores de μ a la probabilidad de una muestra futura, es decir, calcular $f(x_{n+1}|\mu)$ para cada valor de μ y sumar todos estos valores de forma ponderada de manera que los que estén asociados a un valor muy probable de μ tengan más peso que aquellos términos que estén asociados a valores poco probables de μ (que aun así también están tenidas en cuenta). El peso relativo de cada valor de μ lo va a dar en la aproximación bayesiana la

probabilidad a posteriori calculada a partir de la muestra (y de la distribución a posteriori). Esto se expresa matemáticamente como,

$$f(x_{n+1}|x_1,\ldots,x_n) = \int f(x_{n+1}|\mu)\, g(\mu|x_1,\ldots,x_n)d\mu, \tag{6.48}$$

que es el valor promedio de la función $f(x_{n+1}|\mu)$ en la distribución de valores de μ dada por la distribución a posteriori calculada a partir de la muestra. En el caso en el que las observaciones se hagan sobre una población normal hemos visto que tanto la distribución $f(x_{n+1}|\mu)$ como la distribución a posteriori[9] $g(\mu|x_1,\ldots,x_n)$ son distribuciones normales, la primera con media μ y desviación típica σ, y la segunda con media m_n y desviación típica s_n dadas por (6.38). En ese caso resulta,

$$f(x_{n+1}|x_1,\ldots,x_n) \propto \int e^{-\frac{1}{2\sigma^2}(y_{n+1}-\mu)^2} e^{-\frac{1}{2s_n^2}(\mu-m_n)^2} d\mu. \tag{6.49}$$

Desarrollando los cuadrados de los exponentes, reorganizando términos y realizando la integral sobre μ (que es la integral de una distribución normal), se obtiene [1]

$$f(x_{n+1}|x_1,\ldots,x_n) \propto e^{-\frac{1}{2(\sigma_n^2+s_n^2)}(x_{n+1}-m_n)^2}, \tag{6.50}$$

es decir, la distribución de posibles valores de una observación futura, x_{n+1}, es una distribución normal centrada en el valor de la media muestral, m_n, y cuya varianza es la suma de las varianzas poblacional y muestral, $\sigma_n^2 + s_n^2$.

Referencias

[1] W. M. Bolstad. *Introduction to Bayesian Statistics.* John Willey & Sons, Hoboken, New Jersey, 2007.

[2] M. R. Spiegel, J. Schiller, and R. A. Srinivasan. *Probabilidad y estadística.* Schaum. Mc Graw Hill, 2013.

[9] Asumiendo una distribución a priori uniforme o normal, como se ha visto en el apartado anterior.