# Estadística Bayesiana

Stefano Molina Martínez

Septiembre 2016

## Introducción

### Una intuición inicial

La estadística bayesiana hereda su nombre del teorema más famoso del estadístico Thomas Bayes, el cual fue planteado para describir la probabilidad de ocurrencia de un evento sujeta a condiciones relacionadas a él. El paradigma Bayesiano se basa en el Teorema de Bayes para llevar a cabo un aprendizaje adicional sobre los conocimientos que se tienen sobre un evento, la intuición es la que sigue.

*Definición.* Dados A y B dos eventos tales que $P(B) > 0$, la probabilidad condicional A dado B se define como

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

Suponer que se tiene **H**, una hipótesis que quiere ser probada y se posee evidencia para poner a prueba la veracidad de ella, la cual se denota como **E**, por la definición de probabilidad condicional, se tiene que

$$P(E \cap H) = P(E|H)P(H) = P(H|E)P(E)$$

Si la evidencia es verídica, es decir $P(E) > 0$, se obtiene la fórmula del Teorema de Bayes:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

## Inferencia Bayesiana

En la práctica se utilizan las variables $\theta$, una característica de la población en cuestión que sustituye a la hipótesis **H**, la cual es considerada variable aleatoria por ser desconocida y $\underline{x} = \{x_1, x_2, ..., x_n\}$, un conjunto de observaciones de la población, que sustituyen a la evidencia **E**, sobre la cual se busca hacer la inferencia.

En el sentido Bayesiano, los términos de esta expresión son conocidos como sigue:

$$f(\theta|\underline{x}) = \frac{f(\underline{x}|\theta)f(\theta)}{f(\underline{x})}$$

- $f(\theta)$ es llamada distribución *a priori*, el conocimiento muchas veces empírico que se tiene del evento antes de conocer la validez de la hipótesis.
- $f(\underline{x}|\theta)$ es la verosimilitud $f(\underline{x}|\theta) = L(\theta; \underline{x})$, asumiendo que $\underline{x}$ sea fijo dado $\theta$.
- $f(\underline{x})$ es la constante de propocionalidad.
- $f(\theta|\underline{x})$ se llama distribución posterior o *a posteriori*, es la probabilidad de la hipótesis dada la evidencia $\underline{x}$. También es conocido como *kernel* de la distribución, el cual se dice que es proporcional a la verosimilitud por la *a priori*:

$$f(\theta|\underline{x}) \propto L(\theta; \underline{x})f(\theta)$$

## Distribuciones *a priori*

Una distribución *a priori* para una variable $\theta$ posee todas las características de la distribución de una variable aleatoria con la excepción de que $\theta$ no es observable. Debido a lo anterior, la distribución de la variable $\theta$ está construida con los conocimientos o suposiciones que se tienen acerca de su verdadero valor. Las distribuciones *a priori* de una misma variable pueden variar según la cantidad de información que se posea sobre de ella o las fuentes que se consulten para obtenerla, por lo que cada una de éstas influirá en la forma y valores de la distribución posterior obtenida después del proceso de inferencia. Existen dos métodos populares para determinar las distribuciones *a priori*, uno es el de distribuciones conjugadas y el otro es el de distribuciones iniciales no informativas.

Usando distribuciones conjugadas se puede añadir información adicional a un modelo, el método consiste en basar la distribución *a priori* en la forma de la verosimilitud. Se dice que se tiene un par conjugado cuando la distribución *a priori* y la posterior pertecenen a la misma familia de distribuciones. Se llama familia de distribuciones al conjunto de distribuciones de probabilidad que poseen, entre otras cosas, la misma forma funcional y el mismo soporte.

Existen ocasiones en las que es poco conveniente usar las creencias o información *a priori* debido a un posible sesgo en éstas que podrían aportar a la distribución final o la forma distribucional no es conjugable, por lo que en situaciones como éstas se adoptan distribuciones *a priori* que aportan información mínima o nula y reducen el sesgo subjetivo, se conocen como distribuciones no informativas.

## Estimación

El problema de estimación estadística surge de la necesidad de aproximar el valor de uno o varios parámetros que definen alguna característica de una población, esta característica se puede representar como una variable aleatoria $X$ con densidad $f(x;\theta)$ conocida y un parámetro $\theta$ desconocido(Mood, 1963.). Para llevar a cabo la estimación, se cuenta con una muestra aleatoria de valores $\hat{x} = \{x_1, x_2, ..., x_n\}$, los cuales dependen del valor $\theta$. Existen dos tipos de estimación: puntual y por intervalos de confianza, llamados intervalos de credibilidad cuando se calculan por el método Bayesiano.

### Estimación puntual

La estimación puntual se lleva a cabo cuando se asigna el valor de una estadística en función de la muestra para representar al parámetro desconocido $\theta$. El valor asignado se denomina *estimador puntual*. En estadística matemática, el método que produce el mejor estimador insesgado es el de máxima verosimilitud, el cual basado en la muestra y su distribución, busca obtener el valor de $\theta$ que maximiza la verosimilitud $L(\theta; \underline{x})$.

En estadística Bayesiana, la estimación puntual se lleva a cabo penalizando la diferencia que existe entre el estimador $\hat{\theta}$ y el valor real $\theta$ que se supone aleatorio. La forma de llevar a cabo dicha penalización es a través de *funciones de pérdida* que se denotan $L(\hat{\theta}, \theta)$. El estimador puntual, de manera similar al método de máxima verosimilitud, se encuentra minimizando la pérdida esperada, en este caso, encontrando $\hat{\theta}$ que minimice bajo la distribución posterior.

$$\min_{\hat{\theta}} E_{\theta|\underline{x}}(L(\hat{\theta}, \theta)$$

Las funciones de pérdida más comunes involucran como resultado una caracterísitca muchas veces fácil de describir acerca de la distribución posterior: la función de pérdida cuadrática tiene por estimador $\hat{\theta} = E(\theta|\underline{x})$, para la función de valor absoluto, $\hat{\theta}$ es la mediana de la distribución y para la función de pérdida 0-1, el estimador es la moda de la distribución.

**Estimación por intervalos**

En estadística matemática, un intervalo de confianza al $(1-\alpha)$ es aquel en el que en el rango de valores que lo definen, podrá ser observado el valor real del parámetro en un $(1-\alpha) * 100\,\%$ de los casos. El análisis de intervalo de confianza depende del cálculo de variables pivotales que permiten estimar el rango de los intervalos dado cierto $\alpha$ de confianza.

En estadística Bayesiana, un intervalo de credibilidad se define de la misma manera que en estadística matemática, pero su cálculo depende únicamente de la distribución posterior obtenida por el método Bayesiano. Un intervalo de credibilidad se define como sigue:

$$P(a < \theta < b|\underline{x}) = 1 - \alpha$$

sustituyendo la igualdad por un $\geq$ en el caso discreto.

Los intervalos de credibilidad más comunes son los intervalos simétricos y los de máxima credibilidad.

**Intervalo simétrico** es aquel que deja la misma masa de probabilidad en cada cola de la distribución.

$$P(a < \theta < b|\underline{x}) = 1 - \alpha$$
$$P(\theta \leq a) = P(\theta \geq b) = \frac{\alpha}{2}$$

**Intervalo de máxima credibilidad** es aquel que minimiza la longitud del intervalo sujeto al valor de $\alpha$.

$$
\begin{aligned}
\min_{a,b} \quad & b - a \\
\text{s.a.} \quad & P(a < \theta < b) = 1 - \alpha \\
& a, b \in \mathbb{R}
\end{aligned}
$$

## Distribuciones predictivas

A pesar de que la idea fundamental de la estadística Bayesiana radica en encontrar el parámetro $\theta$ para la distribución posterior $f(\theta|\underline{x})$, muchas veces el verdadero problema proviene de encontrar cómo sería una observación futura, $x_{n+1}|\underline{x}$: su distribución, esperanza y varianza. Para esto se utilizan las distribuciones predictivas.

Las distribuciones predictivas se basan en el mismo fundamento de la estadística Bayesiana: utilizan la distribución de el parámetro $\theta$ para determinar la distribución de una $x|\theta$. Las distribuciones predictivas pueden ser a priori o posteriores.

Las distribuciones **predictivas a priori** ocupan el conocimiento que se tiene acerca del parámetro $\theta$ así como la distribución conocida de las variables $\underline{x}|\theta$, la cual se asume para la observación $x_{n+1}$ para determinar la distribución de la observación siguiente sin tomar en cuenta los valores observados para $\underline{x}$. De esta manera se está eliminando la influencia del parámetro $\theta$ el cual se mantiene desconocido.

$$\begin{aligned} f(\underline{x}) &= \int\limits_{\theta} f(\underline{x}, \theta) d\theta \\ &= \int\limits_{\theta} f(\underline{x}|\theta) f(\theta) d\theta \end{aligned}$$

Por otro lado, las distribuciones **predictivas posteriores** determinan la distribución para la observación siguiente basadas en el método Bayesiano previamente explicado: con base en la distribución posterior $f(\theta|\underline{x})$ y la distribución conocida para $x_i|\theta$, se busca la distribución de la observación siguiente dadas las observaciones pasadas. En éstas distribuciones, la influencia de la variable $\theta$ se elimina después de conocer su distribución posterior $f(\theta|\underline{x})$.

$$\begin{aligned} f(x_{n+1}|\underline{x}) &= \int\limits_{\theta} f(x_{n+1}, \theta|\underline{x}) d\theta \\ &= \int\limits_{\theta} f(x_{n+1}|\underline{x}) f(\theta|\underline{x}) d\theta \end{aligned}$$

En éstas distribuciones es fácil probar que existe una forma analítica para determinar el valor esperando de la siguiente observación en caso de no ser necesario el conocimiento de la distribución, esto es posible dado el conocimiento de la Ley de Esperanzas Iteradas y el Teorema de Green.

$$
\begin{aligned}
E[x_{n+1}] &= \int\limits_{x_{n+1}} x_{n+1} f(x_{n+1}) dx_{n+1} \\
&= \int\limits_{x_{n+1}} x_{n+1} [\int\limits_{\theta} f(x_{n+1}|\theta) f(\theta) d\theta] dx_{n+1} \\
&= \int\limits_{x_{n+1}} \int\limits_{\theta} x_{n+1} f(x_{n+1}|\theta) f(\theta) d\theta dx_{n+1} \\
&= \int\limits_{\theta} \left[ \int\limits_{x_{n+1}} x_{n+1} f(x_{n+1}|\theta) dx_{n+1} \right] f(\theta) d\theta \\
&= \int\limits_{\theta} E[x_{n+1}|\theta] f(\theta) d\theta
\end{aligned}
$$

De la misma manera, se puede probar para el valor esperado de la predictiva posterior, la cual se define

$$
\begin{aligned}
E[x_{n+1}|\underline{x}] &= \int\limits_{x_{n+1}} x_{n+1} f(x_{n+1}|\underline{x}) dx_{n+1} \\
&= \int\limits_{\theta} E[x_{n+1}|\theta] f(\theta|\underline{x}) d\theta
\end{aligned}
$$