Estadística Bayesiana

Teoría y Conceptos Básicos

Eduardo Gutiérrez Peña

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas

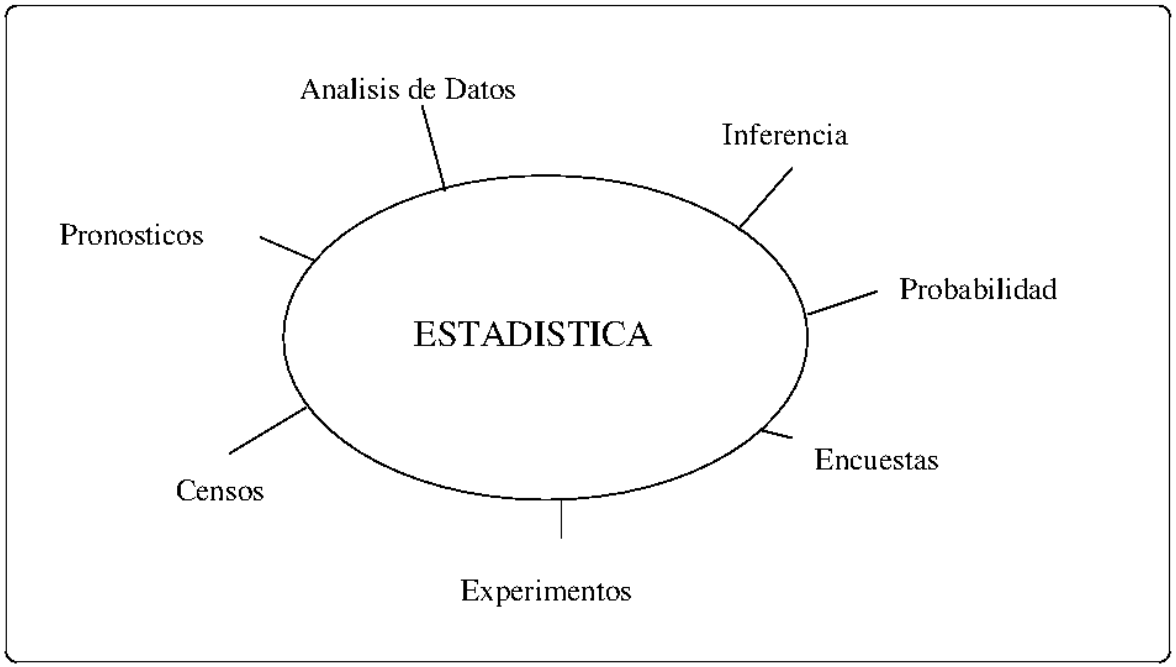
eduardo@sigma.iimas.unam.mx

XXXI Foro de Estadística - Universidad Autónoma Chapingo

Temario

Conceptos fundamentales

¿Qué es la Estadística?



Definiciones de Estadística

"Conjunto de técnicas para *describir* un fenómeno, a partir de un conjunto de datos que presentan *variabilidad*."

"Conjunto de métodos para alcanzar conclusiones acerca de una o varias características de interés de una *población* a partir de información parcial provista por una *muestra* de dicha población".

Ensayemos otra...

De manera muy general, puede decirse que la estadística es la disciplina que estudia los fenómenos *inciertos* (aleatorios), es decir, aquellos que no se pueden predecir con certeza.

El estudio se lleva a cabo a partir del posible conocimiento previo sobre el fenómeno y de *observaciones* que se realizan sobre el mismo.

Ensayemos otra...

De manera muy general, puede decirse que la estadística es la disciplina que estudia los fenómenos *inciertos* (aleatorios), es decir, aquellos que no se pueden predecir con certeza.

El estudio se lleva a cabo a partir del posible conocimiento previo sobre el fenómeno y de *observaciones* que se realizan sobre el mismo.

¿Variabilidad o incertidumbre?

¡Toda la Estadística es descriptiva!

*

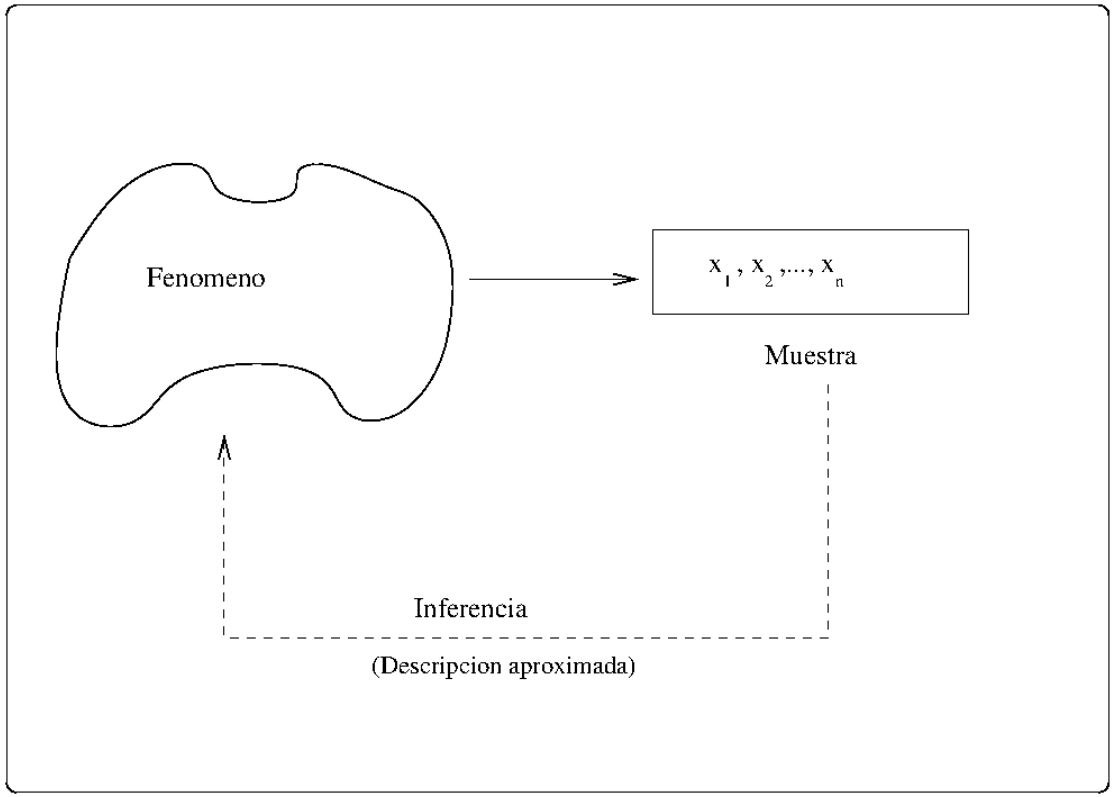
Caso A: se cuenta con todos los datos posibles del fenómeno bajo estudio (e.g. censos)

Descripción: Exacta $\longrightarrow$ Análisis Exploratorio de Datos

Caso B: se cuenta solamente con una parte de todos los datos posibles (e.g. encuestas)

Descripción: Aproximada $\longrightarrow$ Inferencia Estadística

En este último caso,



Pero... ¿cómo seleccionar la muestra?
¿cómo medir el grado de aproximación?

Solución:

Selección probabilística de la muestra (i.e. por sorteo)

$$x \longrightarrow X \longrightarrow \Pr[X = x]$$

Dato Variable (aleatoria) Modelo de probabilidad

Así,

$$\text{Describir el fenómeno} \iff \text{Describir el modelo}$$

Inferencia paramétrica y no paramétrica

- En ocasiones resulta conveniente suponer que

$$\Pr[X = x] = p(x|\theta) \quad \text{(si } X \text{ es discreta)}$$

donde $p(\cdot|\theta)$ tiene forma conocida pero el valor de θ es *desconocido*

Así,

$$\text{Describir el fenómeno} \iff \text{Caracterizar el valor de } \theta$$

- En otros casos, la propia forma funcional de $\Pr[X = x]$ se supone desconocida

A fin de cuentas... ¿qué es un modelo?

Métodos estadísticos tradicionales

Planteamientos más comunes de la Estadística clásica:

- Estimación puntual: $\hat{\theta}$
- Estimación por intervalo: $\theta \in (\underline{\theta}, \bar{\theta})$
- Prueba de hipótesis: $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$

Criterios: suficiencia, insesgamiento, varianza mínima, consistencia, eficiencia, confianza, significancia, potencia,...

¿Cómo y cuándo aplicar cada receta?

Veamos un ejemplo...

Problema: hacer inferencias sobre la proporción de individuos de una población determinada que sufren de cierta enfermedad.

Se selecciona una muestra *aleatoria* de individuos, de manera que cada individuo en la muestra sufra de la enfermedad con probabilidad θ independientemente de los otros individuos en la muestra (θ denota la proporción de individuos enfermos en la población).

La variable aleatoria X denota el número de individuos enfermos en la muestra.

El valor observado $X = x$ es usado para hacer inferencias acerca del parámetro (característica poblacional) θ.

Las inferencias pueden tomar la forma de

- un *estimador puntual*: $\hat{\theta} = 0.1$
- un *intervalo de confianza*: $\theta \in (0.08, 0.12)$ con 95 % de confianza
- una *prueba de hipótesis*: rechazar $H_0 : \theta < 0.07$ con $\alpha = 0.05$
- un *pronóstico*: predecir cuántos individuos sufrirán de la enfermedad el año próximo
- una *decisión*: aplicar un nuevo tratamiento a los individuos que padecen la enfermedad

Estas inferencias se realizan especificando un modelo probabilístico, $p(x|\theta)$, que determina las probabilidades de los posibles valores de X para un valor dado de θ, e.g.

$$X \sim \text{Bin}(\theta, n),$$

de manera que el problema de inferencia estadística se reduce a hacer inferencias sobre θ con base en el valor observado $X = x$.

Principio de Máxima Verosimilitud: valores de θ que asignan una probabilidad alta al valor observado x son más "verosímiles" que aquellos valores de θ que asignan a x una probabilidad pequeña.

Si todo esto suena muy bien... ¿Para qué otro enfoque?

Si todo esto suena muy bien... ¿Para qué otro enfoque?

Notemos lo siguiente:

El parámetro θ es desconocido, pero se considera *constante*, no aleatorio.

De ahí que en la terminología clásica se hable de "verosimilitud", "confianza", "nivel de significancia", etc., y no de **probabilidad**.

Sin embargo, es común que la gente interprete intuitivamente a un intervalo de confianza del 95 % para θ, digamos $(0.08, 0.12)$, como si $\Pr(0.08 < \theta < 0.12) = 0.95$.

De manera similar, no es raro que la gente interprete el nivel de significancia descriptivo (p-value) como la probabilidad de que la hipótesis nula sea verdadera.

El enfoque bayesiano

Idea: diseñar una Teoría Estadística, basada en una pequeña serie de principios básicos, que nos permita *estructurar* la solución a cualquier problema de inferencia.

La vía: la Teoría de la Decisión

¿Para qué una Teoría Estadística?

- Para darle a la Estadística una estructura coherente
- Porque con otros enfoques pueden presentarse casos en los que: (i) no hay una solución razonable; (ii) se presentan paradojas.

Teorema de Bayes. Dados dos eventos A y B tales que $\Pr(B) > 0$,

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

Si $\{A_i : i = 1, 2, \ldots, M\}$ es un conjunto exhaustivo de eventos mutuamente excluyentes, entonces

$$\Pr(A_i|B) = \frac{\Pr(B|A_i)\ \Pr(A_i)}{\sum_{j=1}^{M} \Pr(B|A_j)\ \Pr(A_j)}.$$

Ejemplo. Pruebas de diagnóstico.

- Se desarrolla una nueva prueba para detectar el VIH con una *sensitividad* de 95 % y una *especificidad* del 98 %.
- En una población con una prevalencia de VIH de 1/1000, ¿cuál es la probabilidad de que una persona cuya prueba resulta positiva realmente tenga el VIH?

Sean

$A =$ "la persona tiene VIH" y $A^c =$ "la persona *no* tiene VIH"

$B =$ "la prueba resulta positiva"

- Sensitividad de 95 % significa que $\Pr(B|A) = 0.95$
- Especificidad de 98 % significa que $\Pr(B^c|A^c) = 0.98$

Queremos calcular $\Pr(A|B)$. El Teorema de Bayes nos dice que

$$\Pr(A|B) = \frac{\Pr(B|A)\ \Pr(A)}{\Pr(B|A)\ \Pr(A) + \Pr(B|A^c)\ \Pr(A^c)}.$$

Es decir,

$$\Pr(A|B) = \frac{0.95 \times 0.001}{(0.95 \times 0.001) + (0.02 \times 0.999)} = 0.045$$

¡Más del 95 % de las personas cuya prueba resulta positiva en realidad no tienen el VIH!

Discusión

- Nuestra intuición no es suficientemente buena al procesar evidencia probabilística.
- El punto crucial es *¿de qué manera el resultado de la prueba debe cambiar mis juicios sobre el evento de que la persona tenga VIH?*
- La prevalencia de VIH puede pensarse como la probabilidad *a priori* que describe nuestros juicios sobre el evento de que la persona tenga VIH *antes de conocer el resultado de la prueba*: $\Pr(A) = 0.001$.
- Al observar un resultado positivo, nuestros juicios cambian y la probabilidad del evento se modifica: $\Pr(A|B) = 0.045$. Ésta es la probabilidad *a posteriori* que describe nuestros juicios sobre la ocurrencia de A *después de conocer el resultado de la prueba*.

Reformulación del Ejemplo

Sea θ un parámetro que toma el valor 1 si la persona tiene el VIH y el valor 0 si no lo tiene.

Sea X una variable aleatoria que toma el valor 1 si la prueba resulta positiva y el valor 0 en caso contrario.

Sabemos que

$$\Pr(X = 1|\theta = 1) = 0.95 \qquad \Pr(X = 0|\theta = 1) = 0.05$$

$$\Pr(X = 1|\theta = 0) = 0.02 \qquad \Pr(X = 0|\theta = 0) = 0.98$$

y

$$\Pr(\theta = 1) = 0.001 \qquad \Pr(\theta = 0) = 0.999$$

Entonces

$$\Pr(\theta = 1|X = 1) = 0.045 \qquad \Pr(\theta = 0|X = 1) = 0.955$$

Si la prueba resulta positiva (es decir, si $X = 1$):

- El estimador de máxima verosimilitud es $\hat{\theta} = 1$ debido a que

$$\Pr(X = 1|\theta = 1) = 0.95 \text{ y } \Pr(X = 1|\theta = 0) = 0.02$$

- Al probar las hipótesis $H_0 : \theta = 0$ vs $H_1 : \theta = 1$, usando la estadística de prueba X, el p-valor es 0.02.

  Esto se debe a que

$$\Pr(X = 0|\theta = 0) = 0.98 \text{ y } \Pr(X = 1|\theta = 0) = 0.02$$

  lo que llevaría a rechazar H_0 si se utilizara un nivel de significancia del 0.05.

En cualquier caso, la conclusión es que la persona tiene VIH.

Por otro lado, como se mencionó anteriormente, el Teorema de Bayes nos dice que

$$\Pr(\theta = 1|X = 1) = \frac{\Pr(X = 1|\theta = 1)\ \Pr(\theta = 1)}{\Pr(X = 1)}$$

Es decir,

$$\Pr(\theta = 1|X = 1) = \frac{0.95 \times 0.001}{0.021} = 0.045$$

Por lo tanto, *en contra* de la conclusión obtenida usando los métodos tradicionales, el análisis desde el punto de vista bayesiano indica que es mucho más probable que la persona *no* tenga VIH a pesar de que la prueba resultó *positiva*.

¿A qué se debe esta discrepancia?

Discusión

- El uso del Teorema de Bayes en pruebas de diagnóstico es bastante común y no causa controversias.

- Mucho más controversial es el uso del Teorema de Bayes en análisis estadísticos generales, en los que los parámetros son las cantidades desconocidas de interés y por lo tanto se requiere especificar probabilidades sobre sus valores.

Diferencias

- *Inferencia estadística tradicional*:

  ¿Qué nos dicen los datos X acerca del parámetro θ?

  (Ignora toda evidencia externa)

- *Inferencia bayesiana*:

  ¿Cómo cambian nuestros juicios originales acerca del valor de la cantidad desconocida θ a la luz de los datos X?

  (Puede tomar en cuenta cualquier evidencia externa)

En general tenemos:

(1) Datos, X; y

(2) Cantidades desconocidas, θ, cuyo valor nos interesa.

Las cantidades desconocidas descritas por θ pueden ser: parámetros del modelo, observaciones faltantes, mediciones que no podemos observar directamente o con suficiente precisión, etc.

Como estadísticos, postulamos un modelo de probabilidad

$$p(x|\theta)$$

Desde el punto de vista bayesiano, además,

- θ debe tener una *distribución de probabilidad*, $p(\theta)$, que refleje nuestra incertidumbre *inicial* acerca de su valor.
- X es conocido, así que debemos condicionar en su valor observado, x.

Por lo tanto, nuestro conocimiento acerca del valor de θ queda descrito a través de su *distribución final*

$$p(\theta|x)$$

El Teorema de Bayes nos dice cómo encontrarla:

$$p(\theta|x) = \frac{p(\theta)\, p(x|\theta)}{\int p(\theta)\, p(x|\theta)\, \mathrm{d}\theta}$$

*

El Teorema de Bayes es la clave del *proceso de aprendizaje*.

Interpretación subjetiva de la probabilidad

¿Cómo debe interpretarse la probabilidad?

Existen por lo menos tres interpretaciones:

- *Clásica*: basada en ciertas *simetrías* o en propiedades físicas de objetos tales como dados, cartas de una baraja, bolas dentro de una urna, etc.
- *Frecuentista*: basada en el límite de frecuencias relativas de eventos repetibles *bajo condiciones similares*.
- *Subjetiva*: refleja juicios personales acerca de eventos únicos.

Un ejemplo...

¿Cuál es la probabilidad que *tú* asignarías en este momento al evento

$$A = \text{“El PRI ganará las elecciones presidenciales en el 2018”?}$$

- ¿Quiere decir esto que podemos reportar cualquier número que queramos?

  ¡No! Las probabilidades que asignemos deben ser coherentes, i.e., deben obedecer las leyes de la probabilidad. Además, deben reflejar honestamente nuestro estado de conocimiento.

  Para ser tomadas en serio, las probabilidades que asignemos deben tener relación con la realidad. Usualmente estas probabilidades son asignadas por expertos y/o con base en información (muestral) previa.

Ejemplo:

Preguntas de opción múltiple

Al hacer inferencias sobre un parámetro θ, generalmente se cuenta con algún tipo de información (juicios, creencias) acerca de su valor, incluso antes de observar los datos.

Consideremos las siguientes tres situaciones:

- Una mujer afirma que puede detectar, con un sólo sorbo de una taza de café, si la leche fue agregada antes o después del café. La mujer detecta correctamente el orden en diez tazas.
- Un experto en música afirma que puede distinguir entre una página de una obra de Hayden y una de Mozart. El experto clasifica correctamente diez páginas.
- Un amigo ebrio afirma que puede predecir el resultado del lanzamiento de una moneda honesta. El amigo predice correctamente el resultado de diez lanzamientos.

En cada uno de los tres casos, el modelo es $X \sim \text{Bin}(\theta, 10)$ y se observa $x = 10$, de manera que se rechaza la hipótesis $H_0 : \theta \leq 0.5$ en favor de $H_1 : \theta > 0.5$.

Por lo tanto, en términos de los datos observados, nos veríamos obligados a hacer las mismas inferencias en los tres casos.

Sin embargo, dada nuestra información inicial, *muy probablemente permaneceríamos escépticos acerca de la capacidad del amigo ebrio, ligeramente impresionados por la bebedora de café y sólo un poco sorprendidos por el experto en música.*

El ejemplo anterior muestra que las inferencias deben basarse tanto en los *datos* como en la *información inicial*, incluso si ésta es de naturaleza subjetiva.

La teoría bayesiana proporciona el mecanismo para combinar estas dos fuentes de información de una manera natural.

Como consecuencia, y a diferencia de los métodos clásicos, no es necesario desarrollar criterios *ad hoc* (por ejemplo, insesgamiento, potencia) para juzgar si un procedimiento determinado es bueno en algún sentido.

Distintas distribuciones iniciales pueden dar lugar a inferencias distintas.
¿Es esto una ventaja o una desventaja del enfoque bayesiano?

El precio adicional que hay que pagar es la especificación de una distribución de probabilidad sobre θ que describa la información que se tiene sobre su valor.

Cabe mencionar que los procedimientos clásicos también se basan (implícitamente) en apreciaciones subjetivas (¿Por qué un modelo normal?, ¿Por qué $\alpha = 0.05$?)

El proceso de aprendizaje

Los cuatro pasos a seguir dentro del enfoque bayesiano:

1. Especificación de un modelo muestral, $p(x|\theta)$
2. Especificación de una distribución inicial, $p(\theta)$
3. Cálculo de la distribución final, $p(\theta|x)$, vía el Teorema de Bayes
4. Resumen de la información contenida en $p(\theta|x)$ para hacer inferencias sobre las cantidades de interés (parámetros, observaciones futuras, etc.)

Modelo muestral (*Verosimilitud*)

El problema de elegir un modelo para describir el proceso que generó los datos es esencialmente el mismo que desde el punto de vista clásico.

El modelo elegido dependerá del problema en turno y del propósito del análisis.

En ocasiones, la forma en la que se obtuvieron los datos puede sugerir modelos apropiados *como punto de partida* (e.g., muestreo binomial, conteos Poisson).

Con frecuencia, el modelo refleja una hipótesis cuya plausibilidad es verificada posteriormente en el contexto de los datos (e.g., Y y X se relacionan linealmente entre sí).

Modelo muestral (*Verosimilitud*)

El problema de elegir un modelo para describir el proceso que generó los datos es esencialmente el mismo que desde el punto de vista clásico.

El modelo elegido dependerá del problema en turno y del propósito del análisis.

En ocasiones, la forma en la que se obtuvieron los datos puede sugerir modelos apropiados *como punto de partida* (e.g., muestreo binomial, conteos Poisson).

Con frecuencia, el modelo refleja una hipótesis cuya plausibilidad es verificada posteriormente en el contexto de los datos (e.g., Y y X se relacionan linealmente entre sí).

"Todos los modelos son incorrectos, pero algunos modelos son más útiles que otros." (*George E.P. Box*)

Distribución inicial

Este es un aspecto fundamental del enfoque bayesiano.

El análisis es subjetivo dado que depende del conocimiento que el investigador tiene antes de observar los datos (y que describe a través de *su* distribución inicial).

Sin embargo, si la distribución inicial es razonable, su efecto sobre las inferencias disminuye conforme se tienen más datos.

En ocasiones tenemos una idea vaga de la forma que debería tener la distribución inicial. Tal vez incluso somos capaces de asignar valores, por ejemplo, a su media y su varianza, pero no podemos ser más precisos.

En estos casos es común usar una distribución inicial consistente con nuestra información pero cuya forma sea conveniente, e.g. tal que dé lugar a análisis más sencillos. ($\rightarrow$ Familias conjugadas)

En otros casos puede considerarse que no se tiene información inicial sobre el valor del parámetro (o, por algún motivo, no es deseable incluir nuestra información inicial en el análisis).

En estas situaciones nos gustaría poder utilizar una distribución inicial que refleje nuestra ignorancia acerca del valor del parámetro.

En términos generales siempre es posible encontrar este tipo de distribuciones iniciales *no-informativas*.

Sin embargo, excepto en modelos relativamente simples, esta labor es complicada y no está exenta de problemas.

Distribución final

En términos de variables aleatorias, el Teorema de Bayes toma la forma

$$p(\theta|x) = \frac{p(\theta)p(x|\theta)}{\int p(\tilde{\theta})p(x|\tilde{\theta})d\tilde{\theta}} \ .$$

El denominador, $p(x) = \int p(\tilde{\theta})p(x|\tilde{\theta})d\tilde{\theta}$, no depende de θ, por lo que es común escribir

$$p(\theta|x) \propto p(\theta)p(x|\theta).$$

* En la práctica, el cálculo de la distribución final puede ser un asunto complicado, especialmente si la dimensión del parámetro no es pequeña.

* Sin embargo, para ciertas combinaciones de distribuciones iniciales y verosimilitudes es posible simplificar el análisis. ($\rightarrow$ Familias conjugadas)

* En otros casos se requieren aproximaciones analíticas y/o técnicas computacionales relativamente sofisticadas. ($\rightarrow$ **¡Sesión de mañana!**)

Inferencia

El enfoque bayesiano proporciona inferencias más completas en el sentido de que toda la información disponible sobre el valor de θ queda representada a través de la distribución final.

Es decir, desde el punto de vista bayesiano, el problema de inferencia se reduce a encontrar $p(\theta|x)$: la distribución final *es* la inferencia.

La única receta de la Inferencia Bayesiana. . .

. . .consiste en encontrar la distribución condicional de todas aquellas cantidades de interés cuyo valor desconocemos dado el valor conocido de las variables observadas.

Por supuesto, en la práctica generalmente es deseable resumir este tipo de inferencias en la forma de una estimación puntual, una estimación por intervalo, una prueba de hipótesis, etc.

Ejemplo: **eliminación de parámetros de ruido**.

Robustez

- En Estadística, independientemente del enfoque que se utilice, es importante entender hasta qué punto el modelo usado es robusto antes posibles violaciones a los supuestos.
- Lo anterior también es cierto dentro del enfoque bayesiano en lo que se refiere a la especificación de la distribución inicial.
- En ocasiones el modelo es tal que las inferencias *no* se modifican sustancialmente ante cambios moderados en la distribución final. Esto ocurre, por ejemplo, cuando el tamaño de la muestra es suficientemente grande.
- En otros casos, sin embargo, puede ocurrir que incluso cambios aparentemente insignificantes en la distribución inicial produzcan inferencias completamente distintas.

Algunos autores sugieren que, en la práctica, es conveniente comparar los resultados de los análisis derivados de por lo menos tres distribuciones iniciales distintas:

- Una distribución inicial no-informativa
- Una distribución inicial (tentativa) que refleje los aspectos más importantes nuestra información inicial
- Una distribución inicial (tal vez artificialmente) más informativa

La idea es que, si las inferencias no son muy distintas en cada uno de estos casos, el análisis (dados los datos observados) será relativamente robusto en lo que se refiere a la elección de la distribución inicial. No será necesario entonces preocuparse demasiado por especificar una distribución inicial con mucha precisión.

En caso contrario, es importante hacer el esfuerzo necesario para especificar una distribución que refleje genuinamente nuestra información inicial.

Un ejemplo simple de inferencia bayesiana (distribución Binomial)

- Datos: x éxitos en n ensayos independientes, cada uno con probabilidad de éxito θ.

  Por ejemplo, θ puede representar la tasa de respuesta ante cierta dosis de una sustancia tóxica, y x el número de individuos, de un total de n expuestos, que presentan efectos adversos.

- Función de verosimilitud:

$$p(x|\theta) = \text{Bin}(x|\theta; n) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \propto \theta^x(1-\theta)^{n-x}$$

- Distribución inicial:

$$p(\theta) = \text{Beta}(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \propto \theta^{a-1}(1-\theta)^{b-1}$$

- Distribución final:

$$\begin{aligned} p(\theta|x) &\propto p(\theta)\, p(x|\theta) \\ &\propto \theta^{x+a-1}(1-\theta)^{n-x+b-1} \\ &\propto \text{Beta}(\theta|x+a, n-x+b) \end{aligned}$$

Notemos que tanto la distribución inicial como la final son Beta.

En este caso se dice que la familia de distribuciones Beta es *conjugada* para el modelo Binomial.

Supongamos que, dada la información inicial disponible, se determina que $E(\theta) = 0.40$ y que $\Pr(\theta > 0.54) = 0.10$

Esto implica que $a = 9.2$ y $b = 13.8$

Interpretación: esta información inicial es equivalente a la de una muestra de tamaño $a + b = 23$ en la que se obtuvieron $a = 9.2$ éxitos.

Para la distribución Beta(a, b) se sabe que la media está dada por $m = a/(a + b)$ y la varianza por $s^2 = m(1 - m)/(a + b + 1)$

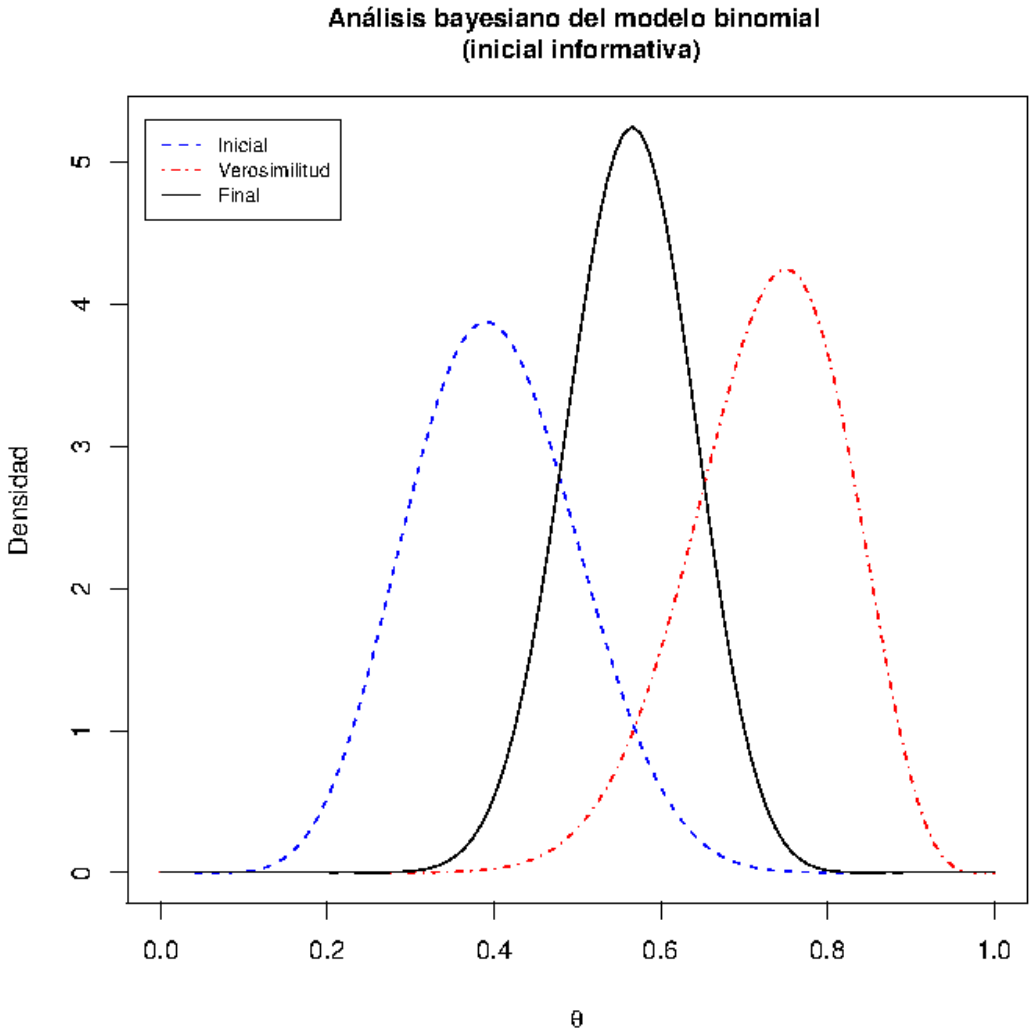
Entonces, *a priori*, la media de θ es $m = 0.40$ y la desviación estándar es $s = 0.1$

Supongamos ahora que, al realizar un experimento con $n = 20$ individuos expuestos, observamos $x = 15$ individuos afectados.

Desglose de la información	*Inicial*	*Datos*	*Final*
Éxitos	9.2	15	24.2
Fracasos	13.8	5	18.8
Total	23	20	43

La media y la desviación estándar de la distribución final de θ están dadas por $E(\theta|x) = 0.563$ y $\text{sd}(\theta|x) = 0.075$, respectivamente.

Notemos que $\Pr(\theta > 0.54|x) = 0.62$


Análisis bayesiano del modelo binomial
(inicial informativa)
Inicial
Verosimilitud
Final
Densidad
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0
θ

Caso no informativo

Supongamos que no se tiene o no se desea utilizar la información inicial.

Esto se puede especificar a través de una distribución inicial uniforme, lo que implica que $a = b = 1$.

En este caso, con $x = 15$ individuos afectados de un total de $n = 20$ individuos expuestos, tenemos:

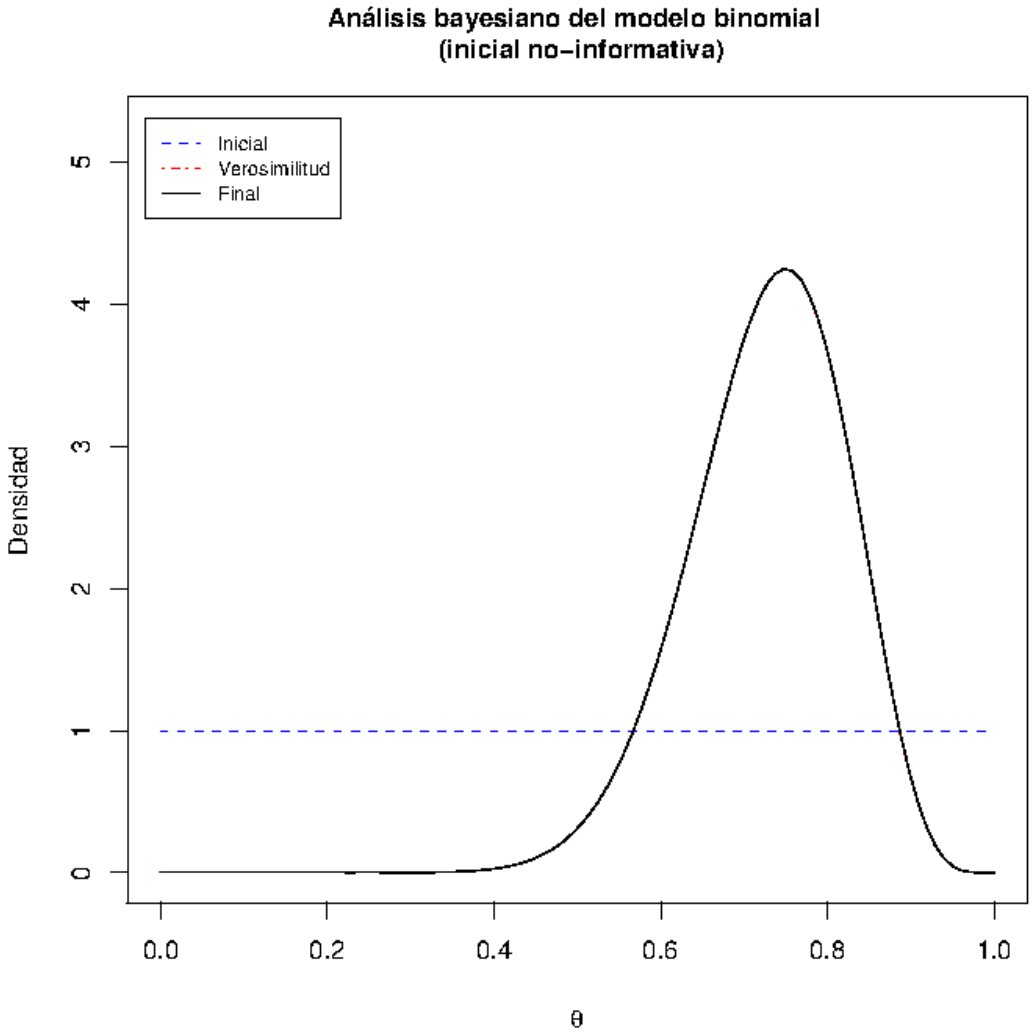
Desglose de la información	*Inicial*	*Datos*	*Final*
Éxitos	1	15	16
Fracasos	1	5	6
Total	2	20	22

La media y la desviación estándar de la distribución final de θ están dadas por $E(\theta|x) = 0.727$ y $\text{sd}(\theta|x) = 0.093$, respectivamente.

Por otro lado, la *moda* de la distribución final es igual a 0.75, valor que coincide con el *estimador de máxima verosimilitud* para θ en este caso.

Cabe hacer notar que en este caso $\Pr(\theta > 0.54|x) = 0.97$

Supongamos ahora que estamos interesados en probar la hipótesis $H_0 : \theta \leq 0.40$. Entonces, la probabilidad $\Pr(\theta \leq 0.40|x) = 0.0008$ puede usarse para determinar que los datos *no* apoyan esta hipótesis nula.


Análisis bayesiano del modelo binomial
(inicial no−informativa)
Inicial
Verosimilitud
Final
Densidad
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0
θ

Distribución predictiva

Hasta el momento sólo hemos discutido el problema de hacer inferencias acerca del valor desconocido de un parámetro.

En muchas situaciones, sin embargo, el propósito de formular un modelo estadístico es hacer *predicciones* sobre el valor de una o más observaciones futuras.

Este problema se resuelve de manera más elegante desde el punto de vista bayesiano que desde el punto de vista clásico.

Al hacer inferencias predictivas sobre el valor de una observación futura con base en un modelo, deben tomarse en cuenta dos fuentes de incertidumbre:

- Incertidumbre sobre el valor del parámetro (sobre el cual se pueden hacer inferencias con base en la distribución final).

- Incertidumbre por el hecho de que cualquier observación futura es aleatoria en sí misma (aún si conociéramos el verdadero valor del parámetro, no podríamos predecir con certeza el valor de una observación futura).

Dentro del enfoque clásico de la Estadística, es común ajustar el modelo con base en los datos (obteniendo un estimador puntual $\hat{\theta}$), y entonces hacer predicciones con base en el modelo $p(x|\hat{\theta})$ como si éste fuera el modelo verdadero.

De esta manera, se ignora completamente la primera fuente de incertidumbre, lo que produce predicciones que aparentan ser más precisas de lo que realmente son.

En contraste, el enfoque bayesiano toma en cuenta las dos fuentes de incertidumbre de manera natural.

Distribución predictiva

Supongamos que tenemos una muestra observada $\boldsymbol{x} = (x_1, \ldots, x_n)'$ de $p(x|\theta)$ y que se desea hacer inferencias acerca del valor futuro de $Y = X_{n+1}$.

Dada una distribución inicial $p(\theta)$, el Teorema de Bayes produce la distribución final $p(\theta|\boldsymbol{x})$.

Siguiendo la "*única receta de la inferencia bayesiana*", debemos entonces encontrar la distribución condicional de Y dado el valor observado de $\boldsymbol{x}$.

Dicha distribución está dada por

$$\begin{aligned} p(y|\boldsymbol{x}) &= \int p(y|\theta, \boldsymbol{x}) p(\theta|\boldsymbol{x})\, d\theta \\ &= \int p(y|\theta) p(\theta|\boldsymbol{x})\, d\theta \\ &= E_{p(\theta|\boldsymbol{x})}[p(y|\theta)] \end{aligned}$$

y se conoce como la *distribución predictiva (final)*.

Continuación del ejemplo (distribución Binomial)

- Supongamos que estamos considerando detener el estudio si por lo menos 25 de 40 nuevos individuos tratados presentan efectos adversos.

Con base en la información disponible, ¿Cuál es la probabilidad de que detengamos el estudio?

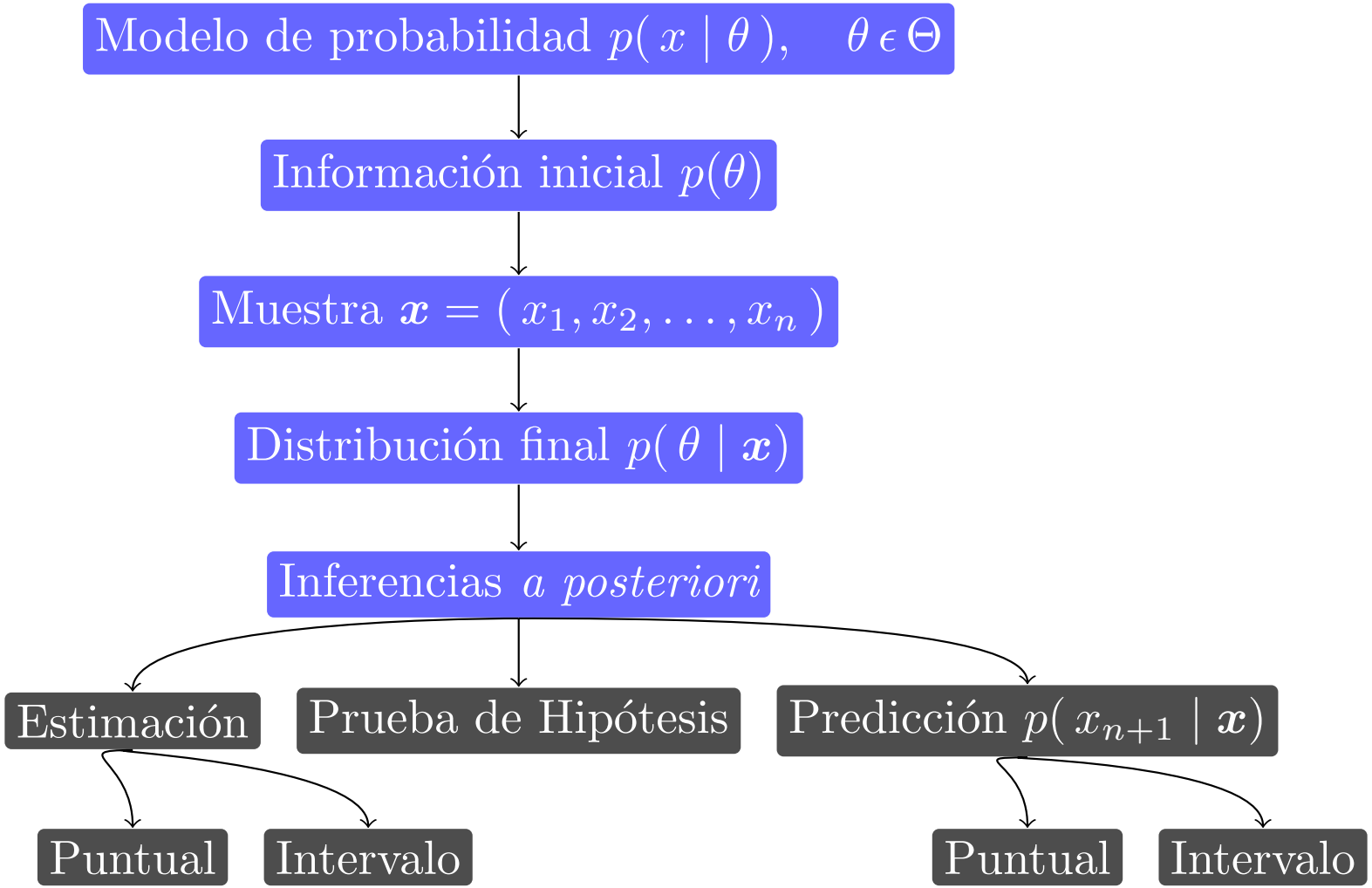
- Estamos considerando observar n^* ensayos adicionales y nos interesa *predecir* el número de 'éxitos', X^*, en esos n^* ensayos.

La distribución predictiva (final) es *Binomial-Beta:*

$$p(x^*|x) = \binom{n^*}{x^*} \frac{\Gamma(n+a+b)\Gamma(x^*+x+a)\Gamma(n^*-x^*+n-x+b)}{\Gamma(x+a)\Gamma(n-x+b)\Gamma(n^*+n+a+b)}.$$

- Esta distribución tiene media $E(X^*|x) = 22.5$ y desviación estándar $\text{sd}(X^*|x) = 4.3$. Además, es tal que $\Pr(X^* \geq 25|x) = 0.33$.

Recapitulando...



Análisis secuencial

Hemos visto que el Teorema de Bayes proporciona el mecanismo para actualizar nuestro estado de información, llevándonos de la distribución inicial a la distribución final.

Esta distribución final se convierte entonces en la *nueva* distribución inicial antes de observar nuevos datos.

Dado $p(\theta)$, supongamos que observamos $X_1 = x_1$ de la densidad $p(x|\theta)$. Por el Teorema de Bayes,

$$p(\theta|x_1) \propto p(\theta)\, p(x_1|\theta).$$

Ésta es nuestra nueva distribución inicial antes de observar $X_2 = x_2$ de la densidad $p(x|\theta)$, condicionalmente independiente de X_1.

Aplicando de nuevo el Teorema de Bayes, obtenemos

$$\begin{aligned} p(\theta|x_1, x_2) &\propto p(\theta|x_1)p(x_2|\theta, x_1) \\ &\propto \{\, p(\theta)\, p(x_1|\theta)\,\}\, p(x_2|\theta) \\ &= p(\theta)\, p(x_1, x_2|\theta). \end{aligned}$$

Éste es el mismo resultado que hubiésemos obtenido de haber actualizado de un solo golpe la distribución inicial $p(\theta)$ con base en la muestra completa $\{x_1, x_2\}$.

Este argumento puede extenderse, por inducción, a cualquier número de observaciones.

Los procedimientos clásicos de análisis secuencial no necesariamente son coherentes en este sentido.

El concepto de intercambiabilidad

Definición. Las variables aleatorias $X_1, \ldots, X_n$ son *(finitamente) intercambiables* bajo una medida de probabilidad P si la distribución inducida por P satisface

$$p(x_1, \ldots, x_n) = p(x_{\pi(1)}, \ldots, x_{\pi(n)})$$

para toda permutación π definida sobre el conjunto $\{1, 2, \ldots, n\}$.

- En otras palabras, las "etiquetas" que identifican a cada una de las variables no proporcionan información alguna.
- Si las variables aleatorias $X_1, \ldots, X_n$ son independientes e idénticamente distribuidas entonces son intercambiables.
- Sin embargo, $X_1, \ldots, X_n$ pueden ser intercambiables sin ser independientes.

Definición. La sucesión infinita de variables aleatorias $X_1, X_2, \ldots$ es *(infinitamente) intercambiable* si toda subsucesión finita es intercambiable en el sentido de la definición anterior.

- El concepto de intercambiabilidad es fundamental en la construcción de los *modelos jerárquicos* que discutiremos en la sesión de mañana.
- El siguiente teorema, que presentaremos en su forma más simple, permite integrar –en un paradigma unificado– los conceptos estadísticos frecuentistas asociados a modelos paramétricos con el concepto de probabilidad como grado de creencia (interpretación subjetiva).
- El resultado proporciona una justificación del enfoque Bayesiano.
- Otra justificación la proporciona la Teoría de la Decisión, que discutiremos más adelante.

Teorema de Representación (Bruno de Finetti)

Si $X_1, X_2, \ldots$ es una sucesión infinita de variables aleatorias definidas sobre $\{0, 1\}$, intercambiables con respecto a la medida de probabilidad P, entonces existe una distribución Q tal que la distribución conjunta $p(x_1, \ldots, x_n)$ tiene la forma

$$p(x_1, \ldots, x_n) = \int_0^1 \left\{ \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} \right\} dQ(\theta),$$

donde $Q(\theta) = \lim_{n\to\infty} \Pr(Y_n/n \leq \theta)$, con $Y_n = X_1 + \cdots + X_n$, y $\theta = \lim_{n\to\infty} Y_n/n$ (c. s.).

El Teorema de Representación tiene un significado muy profundo desde el punto de vista de la modelación subjetiva.

El resutaldo nos dice que el modelo predictivo para una sucesión intercambiable de variables aleatorias binarias puede ser descrito en términos de una situación en la que:

(i) condicional en el valor de una variable aleatoria, θ, las variables aleatorias X_i se consideran independientes con distribución Bernoulli;

(ii) a θ se le asigna una distribución de probabilidad Q.

Por la Ley de los Grandes Números, $\theta = \lim_{n\to\infty} Y_n/n$ (c. s.), de manera que Q puede interpretarse como una descripción de los juicios acerca del límite de la frecuencia relativa de los "éxitos" en una sucesión de ensayos Bernoulli.

Corolario. Si $X_1, X_2, \ldots$ es una sucesión infinita de variables aleatorias definidas sobre $\{0, 1\}$ e intercambiables con respecto a la medida de probabilidad P, entonces

$$p(x_{m+1}, \ldots, x_n | x_1, \ldots, x_m) = \int_0^1 \left\{ \prod_{i=m+1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} \right\} dQ(\theta | x_1, \ldots, x_m)$$

donde $1 \leq m < n$,

$$dQ(\theta | x_1, \ldots, x_m) = \frac{\left\{ \prod_{i=1}^{m} \theta^{x_i} (1-\theta)^{1-x_i} \right\} dQ(\theta)}{\int_0^1 \left\{ \prod_{i=1}^{m} \theta^{x_i} (1-\theta)^{1-x_i} \right\} dQ(\theta)} \qquad (1)$$

y $Q(\theta) = \lim_{n \to \infty} \Pr(Y_n/n \leq \theta)$.

- La expresión (1) no es más que una versión del Teorema de Bayes.
- Notemos que la forma de la representación no cambia.
- En la terminología usual, la *distribución inicial* $Q(\theta)$ ha sido actualizada a través del T. de Bayes, obteniéndose la *distribución final* $Q(\theta | x_1, \ldots, x_m)$.

Teoría de la decisión

- Nos hallamos frente a un problema de decisión cuando debemos elegir entre dos o más formas de actuar.

  La mayor parte de nuestras decisiones cotidianas son triviales (e.g. elegir una película para el fin de semana).

  En otras ocasiones, las consecuencias de nuestras decisiones pueden ser muy importantes y deben ser consideradas cuidadosamente antes de llegar a una conclusión (e.g. elegir una carrera).

- Nuestro interés aquí *no* es describir cómo la gente toma decisiones, sino cómo debería tomarlas si quiere ser *coherente*.

- Cualquier problema de inferencia estadística puede en principio ser visto como un problema de decisión.

  Una teoría de la decisión racional nos permitiría producir una teoría estadística que no presente contradicciones internas.

Elementos de un problema de decisión en ambiente de incertidumbre

- $\mathcal{A} = \{a_1, \ldots, a_k\}$: Conjunto de acciones potenciales

  Debe definirse de manera que sea *exhaustivo* (i.e. que agote todas las posibilidades que en principio parezcan razonables) y *excluyente* (i.e. que la elección de uno de los elementos de $\mathcal{A}$ excluya la elección de cualquier otro).

- $\mathcal{E} = \{E_1, \ldots, E_m\}$: Conjunto de eventos

  Contiene todos los eventos relevantes al problema de decisión.

- $\mathcal{C} = \mathcal{A} \times \mathcal{E}$: Conjunto de consecuencias posibles

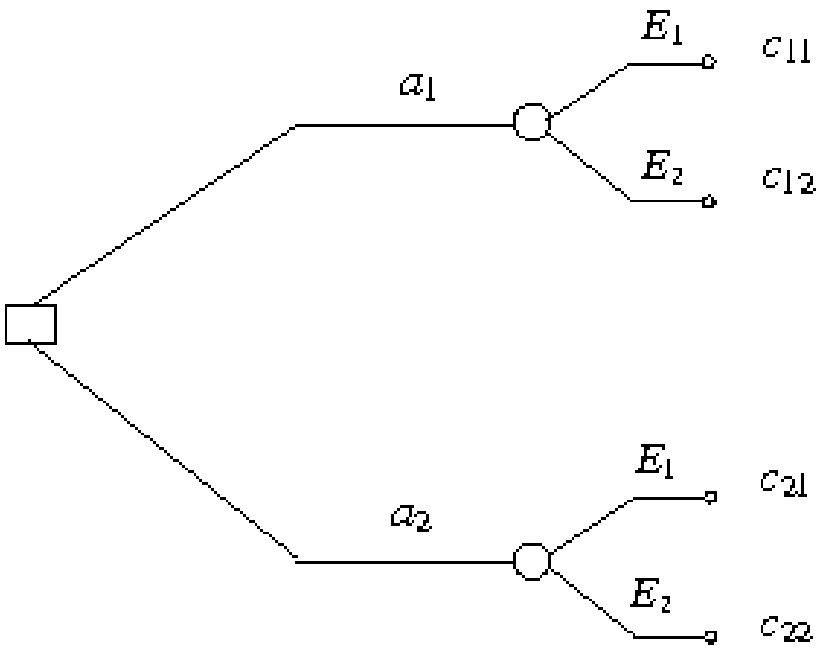
  Describe las consecuencias de elegir una acción $a \in \mathcal{A}$ cuando ocurre un evento $E \in \mathcal{E}$. Por lo tanto podemos escribir $c = (a, E)$.

- $\preceq$: Relación de preferencia entre las distintas consecuencias

  Se define de manera que $c_1 \preceq c_2$ si c_1 no es preferible a c_2 $(c_1, c_2 \in \mathcal{C})$.

Es posible representar la estructura del problema de decisión mediante un *árbol de decisión*.

Por ejemplo, en el caso más simple:



Tanto el conjunto de acciones como el de eventos relevantes puede contener un número infinito de elementos.

En general, el conjunto de eventos relevantes puede ser distinto para cada una de las acciones potenciales a_i, en cuyo caso lo denotamos por $\mathcal{E}_i = \{E_{i1}, \cdots, E_{im_i}\}$

La idea es producir un criterio para elegir la *mejor* acción, tomando en cuenta nuestras preferencias sobre las posibles consecuencias así como nuestra incertidumbre sobre los eventos en $\mathcal{E}$.

La teoría bayesiana se basa en *Axiomas de Coherencia* que describen intuitivamente lo que debe entenderse por comportamiento racional. Por ejemplo:

- *Comparabilidad*: para cada para de consecuencias c_1 y c_2 en $\mathcal{C}$, una y sólo una de las siguiente condiciones puede ser cierta: $c_1 \prec c_2$, $c_1 \succ c_2$ ó $c_1 \sim c_2$
- *Transitividad*: si $c_1 \preceq c_2$ y $c_2 \preceq c_3$ entonces $c_1 \preceq c_3$

Probabilidad y utilidad

De los axiomas se deriva lo siguiente:

- La información que el decisor tiene sobre la verosimilitud de los distintos eventos relevantes al problema de decisión debe ser cuantificada a través de una *medida de probabilidad.*
- De la misma manera, las preferencias del decisor entre las distintas consecuencias debe de cuantificarse a través de una *función de utilidad.*

  A cada una de las consecuencias c se le asigna un número $u(c)$ que mide la utilidad que c tiene para el decisor, de manera tal que

$$c_i \preceq c_j \text{ si y sólo si } u(c_i) \leq u(c_j).$$

Solución bayesiana

Maximización de la utilidad esperada

El resultado fundamental de la teoría bayesiana de decisiones en ambiente de incertidumbre establece que debe elegirse aquella acción a_{i_*} tal que

$$\bar{u}(a_{i_*}) = \max_i \bar{u}(a_i)$$

donde

$$\bar{u}(a_i) = \sum_{j=1}^{m} u(a_i, E_j) \cdot \Pr(E_j) \qquad (i = 1, \ldots, k)$$

denota la *utilidad esperada* de la acción a_i.

Equivalentemente, la mejor acción es la que *minimiza la pérdida esperada*

Otros criterios

Se han propuesto otras formas de resolver problemas de decisión en ambiente de incertidumbre. Aquí describiremos dos de ellas.

Notemos que si el conjunto de eventos relevantes es el mismo para cada una de las acciones, entonces el problema de decisión puede representarse de manera conveniente mediante una tabla como la siguiente:

$\Pr(E)$	$\Pr(E_1)$	$\Pr(E_2)$	$\ldots$	$\Pr(E_m)$
$u(a, E)$	E_1	E_2	$\ldots$	E_m
a_1	$u(a_1, E_1)$	$u(a_1, E_2)$	$\ldots$	$u(a_1, E_m)$
a_2	$u(a_2, E_1)$	$u(a_2, E_2)$	$\ldots$	$u(a_2, E_m)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
a_k	$u(a_k, E_1)$	$u(a_k, E_2)$	$\ldots$	$u(a_k, E_m)$

Criterio maximin (Criterio minimax en caso de funciones de pérdida)

Sea

$$u_m(a_i) = \min_j u(a_i, E_j) \qquad (i = 1, \ldots, k).$$

El criterio maximin consiste entonces en elegir aquella acción a_{i_*} tal que

$$u_m(a_{i_*}) = \max_j u_m(a_j).$$

Criterio condicional (Criterio de la consecuencia más probable)

Sea E_{j*} tal que $\Pr(E_{j_*}) = \max_j \Pr(E_j)$ y definamos

$$u_p(a_i) = u(a_i, E_{j_*}) \qquad (i = 1, \ldots, k).$$

El criterio de la consecuencia más probable consiste en elegir la acción a_{i_*} tal que

$$u_p(a_{i_*}) = \max_i u_p(a_i).$$

Ejemplo. Al prepararse para el examen final, un estudiante debe decidir entre repasar con mucho detalle una de las dos partes de su curso, o repasar con menos detalle las dos partes.

El estudiante juzga que lo más probable es que el examen contenga más preguntas de la segunda parte. Analizaremos este problema de acuerdo a los distintos criterios de decisión mencionados antes.

- Espacio de acciones: $\mathcal{A} = \{a_1, a_2, a_3\}$

$$a_1 = \text{“}\textit{Repasar con detalle la primera parte}\text{”}$$
$$a_2 = \text{“}\textit{Repasar con detalle la segunda parte}\text{”}$$
$$a_3 = \text{“}\textit{Repasar todo el curso con menos detalle}\text{”}$$

- Conjunto de eventos relevantes al problema: $\mathcal{E} = \{E_1, E_2, E_3\}$

$$E_1 = \text{“}\textit{El examen contiene más preguntas de la primera parte}\text{”}$$
$$E_2 = \text{“}\textit{El examen contiene más preguntas de la segunda parte}\text{”}$$
$$E_3 = \text{“}\textit{El examen está equilibrado}\text{”}$$

Una tabla de utilidades razonable sería entonces del tipo

$\Pr(E)$	p	q	$1-p-q$
$u(a,E)$	E_1	E_2	E_3
a_1	0.9	0.2	0.5
a_2	0.2	0.9	0.5
a_3	0.6	0.6	0.7

Por hipótesis $\Pr(E_2) > \Pr(E_1)$ y $\Pr(E_2) > \Pr(E_3)$.

- *Criterio maximin*:

En este caso $u_m(a_1) = 0.2$, $u_m(a_2) = 0.2$ y $u_m(a_3) = 0.6$, por lo que el criterio maximin recomienda elegir a_3.

- *Criterio condicional*:

Dado que $q > p$ y $q > 1 - p - q$, tenemos que $u_p(a_1) = 0.2$, $u_p(a_2) = 0.9$ y $u_p(a_3) = 0.6$. Por lo tanto, el criterio condicional recomienda elegir a_2.

- *Criterio de la utilidad esperada máxima*:

$$\begin{aligned}
\bar{u}(a_1) &= 0.9\,p + 0.2q + 0.5(1-p-q) \\
&= 0.5 + 0.4p - 0.3q \\
\bar{u}(a_2) &= 0.2p + 0.9q + 0.5(1-p-q) \\
&= 0.5 - 0.3p + 0.4q \\
\bar{u}(a_3) &= 0.6p + 0.6q + 0.7(1-p-q) \\
&= 0.7 - 0.1p - 0.1q
\end{aligned}$$

Por ejemplo,

$$Si \quad p = 0.33 \quad y \quad q = 0.50 \ \textit{entonces la mejor acción es } a_3.$$

mientras que

$$Si \quad p = 0.33 \quad y \quad q = 0.60 \ \textit{entonces la mejor acción es } a_2.$$

Problemas de decisión estadísticos

En el contexto de la Estadística, los elementos de un problema de decisión en ambiente de incertidumbre son los siguientes:

- El espacio de *acciones potenciales* disponibles: $\mathcal{A}$
- El espacio parametral, que contiene los posibles *estados de la naturaleza*: Θ
- El espacio de las *consecuencias*: $\mathcal{C} = \mathcal{A} \times \Theta$

Recordemos que, para poder resolver un problema de decisión, es necesario *cuantificar* tanto la incertidumbre sobre Θ como las consecuencias en $\mathcal{C}$.

La única forma racional de cuantificar la incertidumbre es a través de una *medida de probabilidad*, $p(\theta)$, y las consecuencias deben cuantificarse por medio de una *función de utilidad*, $u(a, \theta)$.

En la literatura estadística es más común trabajar, de manera equivalente, en términos de una *función de pérdida* $L(a, \theta)$.

Dicha función de pérdida puede definirse, a partir de una función de utilidad, como

$$L(a, \theta) = B(\theta) - A\, u(a, \theta)$$

donde $A > 0$ y $B(\theta)$ es una función de θ cuyo valor esperado existe.

En este caso, el resultado fundamental de la teoría es que debe elegirse aquella acción que *minimice la pérdida esperada*

$$L^*(a) = \int_\Theta L(a, \theta)\, p(\theta)\, d\theta.$$

Por otra parte, en problemas de inferencia estadística por lo regular se cuenta con información adicional en la forma de una muestra $X_1, \ldots, X_n \sim p(x|\theta)$.

¿Cómo incorporar esta información?

El Teorema de Bayes nos permite combinar las dos fuentes de información (la inicial y la muestral) y de esta manera producir la distribución final $p(\theta|\boldsymbol{x})$.

En este caso, la solución bayesiana al problema de decisión consiste en elegir aquella acción que *minimice la pérdida esperada final*

$$L_x^*(a) = \int_\Theta L(a, \theta)\, p(\theta|\boldsymbol{x})\, d\theta.$$

Procesos de inferencia como problemas de decisión

Sea

$$\mathcal{F} = \{p(x|\theta) : \theta \in \Theta\}$$

una familia paramétrica de distribuciones de probabilidad.

Problema: hacer inferencias sobre el valor de θ.

- Estimación puntual: en este caso $\mathcal{A} = \Theta$, $\mathcal{E} = \sigma(\Theta)$,

  $p(\theta)$ es una distribución de probabilidad sobre Θ, y

  $L(\hat{\theta}, \theta)$ es una función de pérdida.

Como ejemplo, supongamos que usamos la función de pérdida cuadrática: $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$.

Entonces

$$\begin{aligned} L_x^*(\hat{\theta}) = & \int_\Theta L(\hat{\theta}, \theta)\, p(\theta|\boldsymbol{x})\, d\theta \\ = & E_{\theta|x}[\,(\hat{\theta} - \theta)^2\,]. \end{aligned}$$

Notemos que

$$\begin{aligned} E_{\theta|x}[\,(\hat{\theta} - \theta)^2\,] = & E_{\theta|x}[\,(\hat{\theta} - E_{\theta|x}[\theta] + E_{\theta|x}[\theta] - \theta)^2\,] \\ = & E_{\theta|x}[\,(\hat{\theta} - E_{\theta|x}[\theta])^2] + E_{\theta|x}[(\theta - E_{\theta|x}[\theta])^2\,] \\ = & E_{\theta|x}[\,(\hat{\theta} - E_{\theta|x}[\theta])^2\,] + \mathrm{Var}_{\theta|x}[\theta], \end{aligned}$$

de manera que $E_{\theta|x}[\,(\hat{\theta} - \theta)^2\,]$ es mínimo cuando $\hat{\theta} = E_{\theta|x}[\theta]$.

Por lo tanto, la acción óptima (el estimador bayesiano) es

$$\hat{\theta}_* = E_{\theta|x}[\theta].$$

- Prueba de hipótesis: supongamos que deseamos contrastar

$$H_o : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta = \theta_1$$

En este caso

$$\mathcal{A} = \{a_0, a_1\}$$

con

$$a_0 = \quad \text{“Actuar como si } H_0 \text{ fuera cierta”}$$

$$a_1 = \quad \text{“Actuar como si } H_1 \text{ fuera cierta”}$$

y

$$\mathcal{E} = \{\theta_0, \theta_1\}.$$

Como ejemplo, supongamos la siguiente función de pérdida:

$L(a,\theta)$	θ_0	θ_1
a_0	0	k_0
a_1	k_1	0

donde $k_0 > 0$ y $k_1 > 0$.

En este caso

$$L_x^*(a_0) = L(a_0,\theta_0)\cdot p(\theta_0|\boldsymbol{x}) + L(a_0,\theta_1)\cdot p(\theta_1|\boldsymbol{x}) = k_0\, p(\theta_1|\boldsymbol{x})$$

$$L_x^*(a_1) = L(a_1,\theta_0)\cdot p(\theta_0|\boldsymbol{x}) + L(a_0,\theta_1)\cdot p(\theta_1|\boldsymbol{x}) = k_1\, p(\theta_0|\boldsymbol{x})$$

Debe rechazarse H_0 si y sólo si

$$L_x^*(a_0) > L_x^*(a_1).$$

Es decir, si y sólo si

$$\frac{p(\theta_1|\boldsymbol{x})}{p(\theta_0|\boldsymbol{x})} > \frac{k_1}{k_0}.$$

Equivalentemente, si y sólo si

$$\frac{p(\boldsymbol{x}|\theta_0)}{p(\boldsymbol{x}|\theta_1)} < \frac{k_0\, p(\theta_1)}{k_1\, p(\theta_0)}.$$

En particular, si $k_0 = k_1$ entonces H_0 se rechaza si y sólo si

$$p(\theta_1|\boldsymbol{x}) > p(\theta_0|\boldsymbol{x}).$$

El problema de reportar inferencias como problema de decisión

- Aún si no se tiene en mente un problema de decisión específico, nuestra descripción de la incertidumbre presente en una situación dada puede ser usada por otros (e.g., reportes meteorológicos).
- En otros casos, el reporte de inferencias puede ser un fin en sí mismo (no sólo un medio), independiente de cualquier problema de decisión "práctico".
- En esta situación, el espacio de acciones potenciales es el espacio de *todas las distribuciones de probabilidad que podrían representar nuestro estado de información* al momento de tomar la decisión.
- El papel del estadístico sería análogo al de un estudiante que se enfrenta a una pregunta de opción múltiple y al que se le pide responderla con una distribución de probabilidad sobre las posibles respuestas.

Aproximaciones asintóticas

Aproximación normal asintótica

Bajo ciertas condiciones de regularidad, y para tamaños de muestra grandes,

$$p(\theta|\boldsymbol{x}) \approx N(\theta|\hat{\theta}, V(\hat{\theta})),$$

donde $\hat{\theta}$ denota al estimador de máxima verosimilitud para θ y $V(\hat{\theta})$ es la inversa de la matriz de información de Fisher evaluada en $\hat{\theta}$.

En este caso, prácticamente cualquier resumen inferencial de interés, (e.g. distribuciones marginales o momentos de funciones lineales de θ) puede aproximarse fácilmente.

Sin embargo, en aplicaciones específicas no siempre es fácil determinar si la aproximación normal es adecuada para el tamaño de muestra dado.

Es conveniente trabajar en términos de una parametrización $\phi = \phi(\theta)$ tal que la distribución final de ϕ sea más parecida a una distribución normal.

Ejemplo: Distribución Binomial.

- Verosimilitud: $p(x|\theta) = \text{Bin}(x|\theta; n) \propto \theta^x(1-\theta)^{n-x}$
- EMV: $\hat{\theta} = x/n$
- Información de Fisher: $I(\theta) = n\,\theta^{-1}(1-\theta)^{-1}$
- Distribución final: $p(\theta|x) = \text{Beta}(\theta|x+a, n-x+b)$
- Aproximación normal: $p(\theta|x) \approx N(\theta|\hat{\theta}, \hat{\theta}(1-\hat{\theta})/n)$

Ejercicio: Supongan que $n = 10$, $x = 1$, $a = 1$ y $b = 1$. Calculen y comparen gráficamente la aproximación con la verdadera densidad final de θ.

Ahora consideren la reparametrización $\phi = \log\{\theta/(1-\theta)\}$, encuentren la distribución final de ϕ y calculen la correspondiente aproximación asintótica.

Comparen gráficamente esta aproximación con la verdadera densidad final de ϕ. ¿Cuál aproximación es mejor?

Aproximación de Laplace

Supongamos que se desea calcular una integral de la forma

$$I = \int q(\boldsymbol{\theta}) \, \exp\{-n \, h(\boldsymbol{\theta})\} \, d\boldsymbol{\theta}$$

donde $q : \mathbb{R}^d \to \mathbb{R}$ y $h : \mathbb{R}^d \to \mathbb{R}$ son funciones suaves de $\boldsymbol{\theta}$.

Supongamos también que $h(\cdot)$ tiene un mínimo en $\hat{\boldsymbol{\theta}}$.

El método de Laplace aproxima I a través de

$$\hat{I} = (2\,\pi/n)^{d/2} \, |\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})|^{1/2} \, q(\hat{\boldsymbol{\theta}}) \, \exp\{-n \, h(\hat{\boldsymbol{\theta}})\},$$

donde

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \left\{ \frac{\partial^2 h(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T \, \partial \boldsymbol{\theta}} \right\}^{-1}.$$

Proposición. Conforme $n \to \infty$,

$$\hat{I} = I\,\{1 + O(n^{-1})\}.$$

Ejemplo: Supongamos que se desea calcular $E(g(\boldsymbol{\theta})|\boldsymbol{x})$.

Sean $q(\boldsymbol{\theta}) = g(\boldsymbol{\theta})$ y $h(\boldsymbol{\theta}) = -\frac{1}{n}\,\log p(\boldsymbol{\theta}|\boldsymbol{x})$; es decir, $p(\boldsymbol{\theta}|\boldsymbol{x}) = \exp\{-nh(\boldsymbol{\theta})\}$.

Entonces

$$E(g(\boldsymbol{\theta})|\boldsymbol{x}) \approx g(\hat{\boldsymbol{\theta}})\;(2\,\pi/n)^{d/2}\,|\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})|^{1/2}\,p(\hat{\boldsymbol{\theta}}|\boldsymbol{x})$$

La aproximación de Laplace es particularmente útil para aproximar densidades marginales.

Sea $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, $\boldsymbol{\theta}_1 \in \mathbb{R}^{d_1}$ y $\boldsymbol{\theta}_2 \in \mathbb{R}^{d-d_1}$. Supongamos que la distribución de $\boldsymbol{\theta}$ se puede escribir como

$$p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \propto q(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \exp\{-h(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\}$$

y que nos interesa calcular la densidad marginal de $\boldsymbol{\theta}_1$, *i.e.*

$$p(\boldsymbol{\theta}_1) \propto \int q(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \exp\{-h(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\}\, d\boldsymbol{\theta}_2.$$

Para cada valor de $\boldsymbol{\theta}_1$, definamos $q_{\theta_1}(\boldsymbol{\theta}_2) = q(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ y $h_{\theta_1}(\boldsymbol{\theta}_2) = h(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$.

Finalmente, supongamos que $h_{\theta_1}(\cdot)$ tiene un mínimo en $\hat{\boldsymbol{\theta}}_2 = \hat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1)$.

Entonces

$$\hat{p}(\boldsymbol{\theta}_1) \propto |\boldsymbol{\Sigma}(\boldsymbol{\theta}_1)|^{1/2}\, p(\boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1)),$$

donde $\boldsymbol{\Sigma}(\boldsymbol{\theta}_1) = \boldsymbol{\Sigma}_{\theta_1}(\hat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1))$, con

$$\boldsymbol{\Sigma}_{\theta_1}(\boldsymbol{\theta}_2) = \left\{ \frac{\partial^2 h_{\theta_1}(\boldsymbol{\theta}_2)}{\partial \boldsymbol{\theta}_2^T\, \partial \boldsymbol{\theta}_2} \right\}^{-1}.$$

Ejemplo: Distribución logística bivariada

$$p(\theta_1, \theta_2) = \frac{2e^{-\theta_1}e^{-\theta_2}}{(1 + e^{-\theta_1} + e^{-\theta_2})^3}, \quad (\theta_1, \theta_2) \in \mathbb{R}^2.$$

La densidad marginal de θ_1 es

$$p(\theta_1) = \frac{e^{-\theta_1}}{(1 + e^{-\theta_1})^2}, \quad \theta_1 \in \mathbb{R}.$$

Ejercicio: Calculen la aproximación de Laplace para la densidad marginal de θ_1, tomando $q(\theta_1, \theta_2) \equiv 1$ y $h(\theta_1, \theta_2) = -\log p(\theta_1, \theta_2)$.

Fin de la primera parte

A manera de conclusión...

Los siguientes tres aspectos fundamentales caracterizan al enfoque bayesiano:

Información inicial: cada problema es único y tiene su propio contexto, del cual se deriva la información inicial sobre el parámetro (o cualquier otra característica) de interés.

Probabilidad subjetiva: se reconoce explícitamente que toda asignación de probabilidades es subjetiva (i.e., dependen del estado de información del individuo que las asigna). No pretende ser un enfoque "objetivo".

Coherencia interna: al considerar a θ como aleatorio, los métodos bayesianos de inferencia se desarrollan de manera natural a partir de la teoría de la probabilidad y por lo tanto no presentan contradicciones internas.

Muchas gracias por su atención

Muchas gracias por su atención

¡La sesión de mañana será mucho más interesante!