

CURSO:

**REGRESION AVANZADA  
(CON ENFOQUE BAYESIANO)**

PROFESOR: LUIS E. NIETO BARAJAS

EMAIL: [lnieto@itam.mx](mailto:lnieto@itam.mx)

URL: <http://allman.rhon.itam.mx/~lnieto>

*Maestría en ciencia de datos*



## Regresión Avanzada

- **OBJETIVO:** El estudiante conocerá los principios básicos de la inferencia bayesiana y se familiarizará con el concepto de modelado estadístico en general. Conocerá algunas de las familias de modelos más comunes y será capaz de realizar un análisis estadístico bayesiano para estos modelos.
  
- **TEMARIO:**
  1. Introducción a la inferencia bayesiana.
  2. Introducción a MCMC y medidas de ajuste bayesianas
  3. Implementación en R (Winbugs, Openbugs y JAGS)
  4. Modelos lineales generalizados
  5. Modelos dinámicos
  6. Modelos jerárquicos o multinivel
  7. Modelos espaciales (optativo)
  
- **REFERENCIAS:**
  1. Bernardo, J. M. (1981). *Bioestadística: Una perspectiva Bayesiana*. Vicens Vives: Barcelona. (<http://www.uv.es/bernardo/Bioestadistica.pdf>)
  2. Gutiérrez-Peña, E. (1997). *Métodos computacionales en la inferencia Bayesiana*. Monografía IIMAS-UNAM Vol. 6, No. 15. (<http://www.dpye.iimas.unam.mx/eduardo/MCB/index.html>)
  3. Congdon, P. (2001). *Bayesian Statistical Modelling*. Wiley: Chichester.
  4. Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. (2002). *Bayesian Data Analysis*, 2a. edición. Chapman & Hall: Boca Raton.

5. Nieto-Barajas, L. E. & de Alba, E. (2014). Bayesian regression models. En *Predictive Modeling Applications in Actuarial Science*. E.W. Frees, R.A. Derrig & G. Meyers (eds.) Cambridge University Press, pp 334-366.
6. Banerjee, S., Carlin, B. P. & Gelfand, A. (2014). *Hierarchical Modeling and Analysis for Spatial Data*, 2a. edición. Chapman & Hall: Boca Raton.

➤ **PAQUETES ESTADÍSTICOS:** Durante el curso se manejarán varios paquetes estadísticos que nos servirán para entender mejor los conceptos y para realizar análisis Bayesianos.

1) R (<http://www.r-project.org/>)

Paquetes: R2WinBUGS, R2OpenBUGS, rjags

2) R Studio (<http://www.rstudio.com/>)

3) WinBUGS (<http://www.mrc-bsu.cam.ac.uk/bugs/>)

4) OpenBUGS (<http://www.openbugs.info/>)

5) JAGS (<http://sourceforge.net/projects/mcmc-jags/files/JAGS/>)

➤ **EVALUACIÓN:** El curso se evaluará de la siguiente manera:

- Tarea Examen - 40%
- Trabajo Final - 40%
- Exposición - 20%
- Tareas

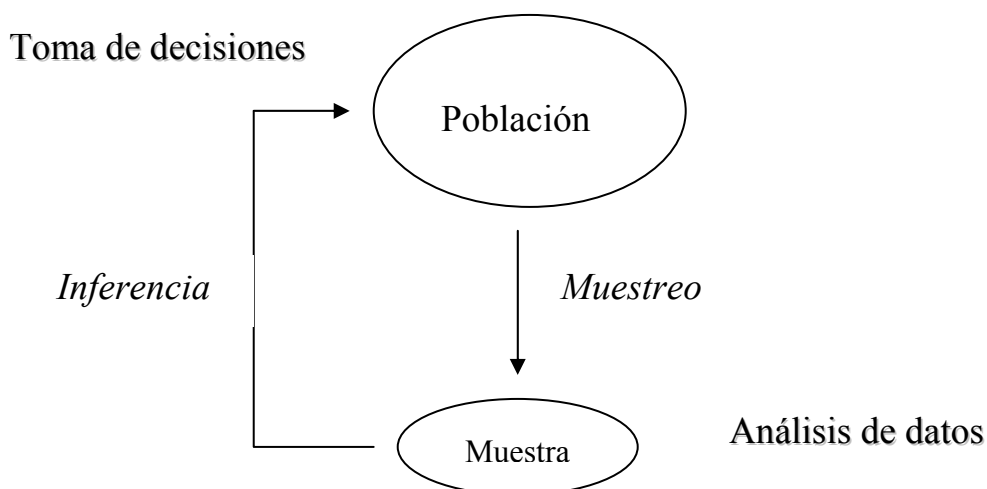
○ **NOTA:** Tanto el trabajo final como la exposición se realizarán en equipos de a lo más 3 integrantes. El objetivo del trabajo es enfrentar al alumno a un problema real en el que tendrá que mostrar su conocimiento aprendido modelando de manera adecuada un conjunto de datos, resolviendo objetivos particulares y tomando decisiones.

# 1. Introducción a la inferencia bayesiana

## 1.1 Fundamentos

- El OBJETIVO de la estadística, y en particular de la estadística Bayesiana, es proporcionar una metodología para analizar adecuadamente la información con la que se cuenta (*análisis de datos*) y decidir de manera razonable sobre la mejor forma de actuar (*teoría de decisión*).

- DIAGRAMA de la Estadística:



- Tipos de INFERENCIA:

	Clásica	Bayesiana
Paramétrica	√√√	√√
No paramétrica	√√	√

- La estadística esta basada en la TEORÍA DE PROBABILIDADES. Formalmente la probabilidad es una función que cumple con ciertas condiciones

(*axiomas de la probabilidad*), pero en general puede entenderse como una medida o cuantificación de la incertidumbre.

- Aunque la definición de función de probabilidad es una, existen varias interpretaciones de la probabilidad: clásica, frecuentista y subjetiva. La METODOLOGÍA BAYESIANA está basada en la interpretación subjetiva de la probabilidad y tiene como punto central el Teorema de Bayes.



Reverendo *Thomas Bayes* (1702-1761).

- El enfoque bayesiano realiza *inferencia estadística* en un contexto de teoría de decisión.
- La TEORÍA DE DECISIÓN propone un método de tomar decisiones basado en unos principios básicos sobre la *elección coherente* entre opciones alternativas.
- ELEMENTOS DE UN PROBLEMA DE DECISIÓN en ambiente de incertidumbre:  
Un problema de decisión se define por la cuarteta  $(D, E, C, \leq)$ , donde:
  - ❑  $D$  : Espacio de opciones.
  - ❑  $E$  : Espacio de eventos inciertos.
  - ❑  $C$  : Espacio de consecuencias.

- $\leq$  : Relación de preferencia entre las distintas opciones.
- CUANTIFICACIÓN de los sucesos inciertos y de las consecuencias.
  - La información que el decisor tiene sobre la posible ocurrencia de los eventos inciertos puede ser cuantificada a través de una *función de probabilidad* sobre el espacio  $E$ .
  - De la misma manera, es posible cuantificar las preferencias del decisor entre las distintas consecuencias a través de una *función de utilidad* de manera que  $c_{ij} \leq c_{i'j'} \Leftrightarrow u(c_{ij}) \leq u(c_{i'j'})$ .
- AXIOMAS DE COHERENCIA. Son una serie de principios que establecen las condiciones para tomar decisiones coherentemente y para aclarar las posibles ambigüedades en el proceso de toma de decisión.
- Teorema: Criterio de decisión Bayesiano.

*Considérese el problema de decisión definido por  $D = \{d_1, d_2, \dots, d_k\}$ , donde  $d_i = \{c_{ij} | E_j, j=1, \dots, m_i\}$ ,  $i=1, \dots, k$ . Sea  $P(E_{ij}|d_i)$  la probabilidad de que suceda  $E_{ij}$  si se elige la opción  $d_i$ , y sea  $u(c_{ij})$  la utilidad de la consecuencia  $c_{ij}$ . Entonces, la cuantificación de la opción  $d_i$  es su utilidad esperada, i.e.,*

$$\bar{u}(d_i) = \sum_{j=1}^{m_i} u(c_{ij})P(E_{ij}|d_i).$$

*La decisión óptima es aquella  $d^*$  tal que  $\bar{u}(d^*) = \max_i \bar{u}(d_i)$ .*

- RESUMIENDO: Si se aceptan los axiomas de coherencia, necesariamente se debe proceder de la siguiente manera:
  - 1) Asignar una utilidad  $u(c)$  para toda  $c$  en  $C$ .
  - 2) Asignar una probabilidad  $P(E)$  para toda  $E$  en  $E$ .
  - 3) Elegir la opción (óptima) que maximiza la utilidad esperada.

## 1.2 Proceso de aprendizaje y distribución predictiva

- La reacción natural de cualquiera que tenga que tomar una decisión cuyas consecuencias dependen de la ocurrencia de eventos inciertos  $E$ , es intentar reducir su incertidumbre obteniendo más información sobre  $E$ .
- LA IDEA es entonces recolectar información que *reduzca* la incertidumbre de los eventos inciertos, o equivalentemente, que mejore el conocimiento que se tiene sobre  $E$ .
- ¿De dónde obtengo información adicional?  
Encuestas, estudios previos, experimentos, etc.
- El problema central de la inferencia estadística es el de proporcionar una metodología que permita *asimilar* la información accesible con el objeto de mejorar nuestro conocimiento inicial.
- ¿Cómo utilizar  $Z$  para mejorar el conocimiento sobre  $E$ ?

$$P(E) \xrightarrow{! ?} P(E|Z)$$

Mediante el Teorema de Bayes.

- **TEOREMA DE BAYES:** Sean  $\{E_j, j \in J\}$  una partición finita de  $\Omega$  ( $E$ ), i.e.,

$E_j \cap E_k = \emptyset \quad \forall j \neq k$  y  $\bigcup_{j \in J} E_j = \Omega$ . Sea  $Z \neq \emptyset$  un evento. Entonces,

$$P(E_i|Z) = \frac{P(Z|E_i)P(E_i)}{\sum_{j \in J} P(Z|E_j)P(E_j)}, \quad i = 1, 2, \dots, k.$$

➤ Comentarios:

- 1) Una forma alternativa de escribir el Teorema de Bayes es:

$$P(E_i|Z) \propto P(Z|E_i)P(E_i)$$

$P(Z)$  es llamada constante de proporcionalidad.

- 2) A las  $P(E_j)$  se les llama probabilidades iniciales o a-priori y a las  $P(E_j|Z)$  se les llama probabilidades finales o a-posteriori. Además,  $P(Z|E_j)$  es llamada verosimilitud y  $P(Z)$  es llamada probabilidad marginal de la información adicional.

- Recordemos que todo esto de la cuantificación inicial y final de los eventos inciertos es para reducir la incertidumbre en un problema de decisión.

Supongamos que para un problema particular se cuenta con lo siguiente:

$P(E_{ij})$ : cuantificación inicial de los eventos inciertos

$u(c_{ij})$ : cuantificación de las consecuencias

$Z$ : información adicional sobre los eventos inciertos

**Teo. Bayes**

$$P(E) \rightleftarrows P(E|Z)$$



En este caso se tienen dos situaciones:

1) Situación inicial (a-priori):

$$P(E_{ij}), \quad u(c_{ij}), \quad \sum_j u(c_{ij})P(E_{ij}) \quad \longleftarrow \quad \text{Utilidad esperada inicial}$$

2) Situación final (a-posteriori):

$$P(E_{ij}|Z), \quad u(c_{ij}), \quad \sum_j u(c_{ij})P(E_{ij}|Z) \quad \longleftarrow \quad \text{Utilidad esperada final}$$

### Problema de Inferencia.

➤ PROBLEMA DE INFERENCIA. Sea  $\mathbf{F} = \{f(x|\theta), \theta \in \Theta\}$  una familia paramétrica indexada por el parámetro  $\theta \in \Theta$ . Sea  $X_1, \dots, X_n$  una m.a. de observaciones de  $f(x|\theta) \in \mathbf{F}$ . El problema de inferencia paramétrico consiste en aproximar el verdadero valor del parámetro  $\theta$ .

□ El problema de inferencia estadístico se puede ver como un problema de decisión con los siguientes elementos:

$\mathbf{D}$  = espacio de decisiones de acuerdo al problema específico

$\mathbf{E} = \Theta$  (espacio parametral)

$\mathbf{C} = \{(d, \theta) : d \in \mathbf{D}, \theta \in \Theta\}$

$\leq$  : Será representado por una función de utilidad o pérdida.

➤ La muestra proporciona información adicional sobre los eventos inciertos  $\theta \in \Theta$ . El problema consiste en cómo actualizar la información.

- Por lo visto con los axiomas de coherencia, el decisor es capaz de cuantificar su conocimiento acerca de los eventos inciertos mediante una función de probabilidades. Definamos,

$f(\theta)$  *la distribución inicial (ó a-priori)*. Cuantifica el conocimiento inicial sobre  $\theta$ .

$f(\mathbf{x}|\theta)$  *proceso generador de información muestral*. Proporciona información adicional acerca de  $\theta$ .

$f(\mathbf{x}|\theta)$  *la función de verosimilitud*. Contiene toda la información sobre  $\theta$  proporcionada por la muestra  $\mathbf{X} = (X_1, \dots, X_n)$ .

- Toda esta información acerca de  $\theta$  se combina para obtener un conocimiento final o a-posteriori después de haber observado la muestra. La forma de hacerlo es mediante el *Teorema de Bayes*:

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

donde  $f(\mathbf{x}) = \int_{\Theta} f(\mathbf{x}|\theta)f(\theta)d\theta$  ó  $\sum_{\theta} f(\mathbf{x}|\theta)f(\theta)$ .

Como  $f(\theta|\mathbf{x})$  es función de  $\theta$ , entonces podemos escribir

$$f(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)f(\theta)$$

Finalmente,

$f(\theta|\mathbf{x})$  *la distribución final (ó a-posteriori)*. Proporciona todo el conocimiento que se tiene sobre  $\theta$  (inicial y muestral).

- NOTA: Al tomar  $\theta$  el carácter de aleatorio, debido a que el conocimiento que tenemos sobre el verdadero valor  $\theta$  es incierto, entonces la función de

densidad que genera observaciones con información relevante para  $\theta$  es realmente una función de densidad condicional.

- Definición: Llamaremos una muestra aleatoria (m.a.) de tamaño  $n$  de una población  $f(x|\theta)$ , que depende de  $\theta$ , a un conjunto  $X_1, \dots, X_n$  de variables aleatorias condicionalmente independientes dado  $\theta$ , i.e.,

$$f(x_1, \dots, x_n | \theta) = f(x_1 | \theta) \cdots f(x_n | \theta).$$

En este caso, la función de verosimilitud es la función de densidad (condicional) conjunta de la m.a. vista como función del parámetro, i.e.,

$$f(\mathbf{x} | \theta) = \prod_{i=1}^n f(x_i | \theta).$$

- DISTRIBUCIÓN PREDICTIVA: La distribución predictiva es la función de densidad (marginal)  $f(x)$  que me permite determinar qué valores de la v.a.  $X$  resultan más probables.
- Lo que conocemos acerca de  $X$  está condicionado al valor del parámetro  $\theta$ , i.e.,  $f(x|\theta)$  (su función de densidad condicional). Como  $\theta$  es un valor desconocido,  $f(x|\theta)$  no puede utilizarse para describir el comportamiento de la v.a.  $X$ .
- Distribución predictiva inicial. Aunque el verdadero valor de  $\theta$  sea desconocido, siempre se dispone de cierta información sobre  $\theta$  (mediante su distribución inicial  $f(\theta)$ ). Esta información puede combinarse para poder dar información sobre los valores de  $X$ . La forma de hacerlo es:

$$f(x) = \int f(x|\theta)f(\theta)d\theta \quad \text{ó} \quad f(x) = \sum_{\theta} f(x|\theta)f(\theta)$$

- Supongamos que se cuenta con información adicional (información muestral)  $X_1, X_2, \dots, X_n$  de la densidad  $f(x|\theta)$ , por lo tanto es posible tener un conocimiento final sobre  $\theta$  mediante su distribución final  $f(\theta|\underline{x})$ .

- Distribución predictiva final. Supongamos que se quiere obtener información sobre los posibles valores que puede tomar una nueva v.a.  $X_F$  de la misma población  $f(x|\theta)$ . Si  $X_F$  es independiente de la muestra  $X_1, X_2, \dots, X_n$ , entonces

$$f(x_F|\underline{x}) = \int f(x_F|\theta)f(\theta|\underline{x})d\theta \quad \text{ó} \quad f(x_F|\underline{x}) = \sum_{\theta} f(x_F|\theta)f(\theta|\underline{x})$$

- EJEMPLO 6: *Lanzar una moneda*. Se tiene un experimento aleatorio que consiste en lanzar una moneda. Sea  $X$  la v.a. que toma el valor de 1 si la moneda cae sol y 0 si cae águila, i.e.,  $X \sim \text{Ber}(\theta)$ . En realidad se tiene que  $X|\theta \sim \text{Ber}(\theta)$ , donde  $\theta$  es la probabilidad de que la moneda caiga sol.

$$f(x|\theta) = \theta^x (1 - \theta)^{1-x} I_{\{0,1\}}(x).$$

El conocimiento inicial que se tiene acerca de la moneda es que puede ser una moneda deshonesto (dos soles).

$$P(\text{honesta}) = 0.95 \text{ y } P(\text{deshonesta}) = 0.05$$

¿Cómo cuantificar este conocimiento sobre  $\theta$ ?

$$\left. \begin{array}{l} \text{moneda honesta} \Leftrightarrow \theta = 1/2 \\ \text{moneda deshonesto} \Leftrightarrow \theta = 1 \end{array} \right\} \theta \in \{1/2, 1\}$$

por lo tanto,

$$P(\theta = 1/2) = 0.95 \quad \text{y} \quad P(\theta = 1) = 0.05$$

es decir,

$$f(\theta) = \begin{cases} 0.95, & \text{si } \theta = 1/2 \\ 0.05, & \text{si } \theta = 1 \end{cases}$$

Supongamos que al lanzar la moneda una sola vez se obtuvo un sol, i.e,  $X_1=1$ . Entonces la verosimilitud es

$$P(X_1 = 1|\theta) = \theta^1(1 - \theta)^0 = \theta.$$

Combinando la información inicial con la verosimilitud obtenemos,

$$\begin{aligned} P(X_1 = 1) &= P(X_1 = 1|\theta = 1/2)P(\theta = 1/2) + P(X_1 = 1|\theta = 1)P(\theta = 1) \\ &= (0.5)(0.95) + (1)(0.05) = 0.525 \end{aligned}$$

$$P(\theta = 1/2|X_1 = 1) = \frac{P(X_1 = 1|\theta = 1/2)P(\theta = 1/2)}{P(X_1 = 1)} = \frac{(0.5)(0.95)}{0.525} = 0.9047$$

$$P(\theta = 1|X_1 = 1) = \frac{P(X_1 = 1|\theta = 1)P(\theta = 1)}{P(X_1 = 1)} = \frac{(1)(0.05)}{0.525} = 0.0953$$

es decir,

$$f(\theta|x_1 = 1) = \begin{cases} 0.9047, & \text{si } \theta = 1/2 \\ 0.0953, & \text{si } \theta = 1 \end{cases}$$

La distribución predictiva inicial es

$$\begin{aligned} P(X = 1) &= P(X = 1|\theta = 1/2)P(\theta = 1/2) + P(X = 1|\theta = 1)P(\theta = 1) \\ &= (0.5)(0.95) + (1)(0.05) = 0.525 \end{aligned}$$

$$\begin{aligned} P(X = 0) &= P(X = 0|\theta = 1/2)P(\theta = 1/2) + P(X = 0|\theta = 1)P(\theta = 1) \\ &= (0.5)(0.95) + (0)(0.05) = 0.475 \end{aligned}$$

es decir,

$$f(x) = \begin{cases} 0.525, & \text{si } x = 1 \\ 0.475, & \text{si } x = 0 \end{cases}$$

La distribución predictiva final es

$$P(X_F = 1 | x_1 = 1) = P(X_F = 1 | \theta = 1/2)P(\theta = 1/2 | x_1 = 1) + P(X_F = 1 | \theta = 1)P(\theta = 1 | x_1 = 1) \\ = (0.5)(0.9047) + (1)(0.0953) = 0.54755$$

$$P(X_F = 0 | x_1 = 1) = P(X_F = 0 | \theta = 1/2)P(\theta = 1/2 | x_1 = 1) + P(X_F = 0 | \theta = 1)P(\theta = 1 | x_1 = 1) \\ = (0.5)(0.9047) + (0)(0.0953) = 0.45235$$

es decir,

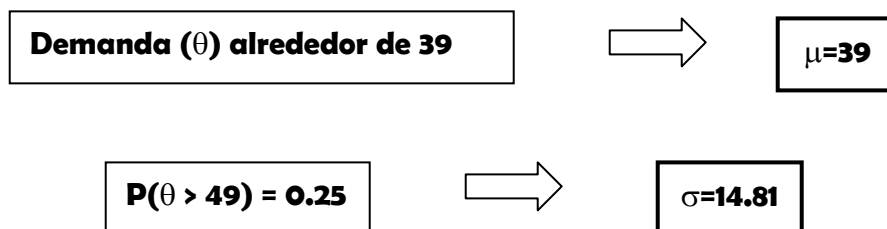
$$f(x_F | x_1 = 1) = \begin{cases} 0.548, & \text{si } x_F = 1 \\ 0.452, & \text{si } x_F = 0 \end{cases}.$$

- EJEMPLO 7: *Proyectos de inversión*. Las utilidades de un determinado proyecto pueden determinarse a partir de la demanda ( $\theta$ ) que tendrá el producto terminal. La información inicial que se tiene sobre la demanda es que se encuentra alrededor de \$39 millones de pesos y que el porcentaje de veces que excede los \$49 millones de pesos es de 25%.

De acuerdo con la información proporcionada, se puede concluir que una distribución normal modela “adecuadamente” el comportamiento inicial, entonces

$$\theta \sim N(\mu, \sigma^2),$$

donde  $\mu = E(\theta) = \text{media}$  y  $\sigma^2 = \text{Var}(\theta) = \text{varianza}$ . Además



¿Cómo?

$$P(\theta > 49) = P\left(Z > \frac{49 - 39}{\sigma}\right) = 0.25 \Rightarrow Z_{0.25} = \frac{49 - 39}{\sigma},$$

$$\text{como } Z_{0.25} = 0.675 \text{ (valor de tablas)} \Rightarrow \sigma = \frac{10}{0.675}$$

Por lo tanto,  $\theta \sim N(39, 219.47)$ .

Para adquirir información adicional sobre la demanda, se considerarán 3 proyectos similares cuyas utilidades dependen de la misma demanda. Supongamos que la utilidad es una variable aleatoria con distribución Normal centrada en  $\theta$  y con una desviación estándar de  $\sigma=2$ .

$$X|\theta \sim N(\theta, 4) \quad \text{y} \quad \theta \sim N(39, 219.47)$$

Se puede demostrar que la distribución predictiva inicial toma la forma

$$X \sim N(39, 223.47)$$

¿Qué se puede derivar de esta distribución predictiva?

$$P(X > 60) = P\left(Z > \frac{60 - 39}{\sqrt{223.47}}\right) = P(Z > 1.4047) = 0.0808,$$

lo cual indica que es muy poco probable tener una utilidad mayor a 60.

Suponga que las utilidades de los 3 proyectos son:  $x_1=40.62$ ,  $x_2=41.8$ ,  $x_3=40.44$ .

Se puede demostrar que si

$$X|\theta \sim N(\theta, \sigma^2) \quad \text{y} \quad \theta \sim N(\theta_0, \sigma_0^2) \Rightarrow \theta|X \sim N(\theta_1, \sigma_1^2)$$

$$\text{donde, } \theta_1 = \frac{\frac{n}{\sigma^2} \bar{x} + \frac{1}{\sigma_0^2} \theta_0}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}} \quad \text{y} \quad \sigma_1^2 = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}.$$

Por lo tanto,

$$\bar{x}=40.9533, \theta_0 = 39, \sigma^2 = 4, \sigma_0^2 = 219.47, n=3$$

$$\theta_1 = 40.9415, \sigma_1^2 = 1.3252 \therefore \theta|\underline{x} \sim N(40.9415, 1.3252)$$

### 1.3 Distribuciones iniciales informativas, no informativas y conjugadas

- Existen diversas clasificaciones de las distribuciones iniciales. En términos de la cantidad de información que proporcionan se clasifican en *informativas* y *no informativas*.
- **DISTRIBUCIONES INICIALES INFORMATIVAS:** Son aquellas distribuciones iniciales que proporcionan información relevante e importante sobre la ocurrencia de los eventos inciertos  $\theta$ .
- **DISTRIBUCIONES INICIALES NO INFORMATIVAS:** Son aquellas distribuciones iniciales que no proporcionan información relevante o importante sobre la ocurrencia de los eventos inciertos  $\theta$ .
- Existen varios criterios para definir u obtener una distribución inicial no informativa:
  - 1) *Principio de la razón insuficiente:* Bayes (1763) y Laplace (1814, 1952). De acuerdo con este principio, en ausencia de evidencia en contra, todas las posibilidades deberían tenerla misma probabilidad inicial.
    - En particular, si  $\theta$  puede tomar un número finito de valores, digamos  $m$ , la distribución inicial no informativa, de acuerdo con este principio es:



$$f(\theta) = \frac{1}{m} I_{\{\theta_1, \theta_2, \dots, \theta_m\}}(\theta)$$

- ¿Qué pasa cuando el número de valores (m) que puede tomar  $\theta$  tiende a infinito?

$$f(\theta) \propto \text{cte.}$$

En este caso se dice que  $f(\theta)$  es una distribución inicial impropia, porque no cumple con todas las propiedades para ser una distribución inicial propia.

2) *Distribución inicial invariante*: Jeffreys (1946) propuso una distribución inicial no informativa invariante ante reparametrizaciones, es decir, si  $\pi_\theta(\theta)$  es la distribución inicial no informativa para  $\theta$  entonces,  $\pi_\varphi(\varphi) = \pi_\theta(\theta(\varphi)) |J_\theta(\varphi)|$  es la distribución inicial no informativa de  $\varphi = \varphi(\theta)$ . Esta distribución es generalmente impropia.

- La regla de Jeffreys consiste en lo siguiente: Sea  $\mathbf{F} = \{f(x|\theta) : \theta \in \Theta\}$ ,  $\Theta \subset \mathbb{R}^d$  un modelo paramétrico para la variable aleatoria  $X$ . La distribución inicial no informativa de Jeffreys para el parámetro  $\theta$  con respecto al modelo  $\mathbf{F}$  es

$$\pi(\theta) \propto |\det\{I(\theta)\}|^{1/2}, \quad \theta \in \Theta,$$

donde  $I(\theta) = -E_{X|\theta} \left\{ \frac{\partial^2 \log f(X|\theta)}{\partial \theta \partial \theta'} \right\}$  es la matriz de información de Fisher

- EJEMPLO 9: Sea  $X$  una v.a. con distribución condicional dado  $\theta$ ,  $\text{Ber}(\theta)$ , i.e.,  $f(x|\theta) = \theta^x (1-\theta)^{1-x} I_{\{0,1\}}(x)$ ,  $\theta \in (0,1)$ .

$$\log f(x|\theta) = x \log(\theta) + (1-x) \log(1-\theta) + \log I_{\{0,1\}}(x)$$

$$\frac{\partial}{\partial \theta} \log f(x|\theta) = \frac{x}{\theta} - \frac{1-x}{1-\theta}$$

$$\frac{\partial^2}{\partial \theta^2} \log f(x|\theta) = -\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2}$$

$$I(\theta) = -E_{X|\theta} \left\{ -\frac{X}{\theta^2} - \frac{1-X}{(1-\theta)^2} \right\} = \frac{E(X|\theta)}{\theta^2} + \frac{1-E(X|\theta)}{(1-\theta)^2} = \dots = \frac{1}{\theta(1-\theta)}$$

$$\pi(\theta) \propto \left\{ \frac{1}{\theta(1-\theta)} \right\}^{1/2} = \theta^{-1/2} (1-\theta)^{-1/2} I_{(0,1)}(\theta)$$

$$\therefore \pi(\theta) = \text{Beta}(\theta|1/2, 1/2).$$

3) *Criterio de referencia*: Bernardo (1986) propuso una nueva metodología para obtener distribuciones iniciales mínimo informativas o de referencia, basándose en la idea de que los datos contienen toda la información relevante en un problema de inferencia.

- La distribución inicial de referencia es aquella distribución inicial que maximiza la distancia esperada que hay entre la distribución inicial y la final cuando se tiene un tamaño de muestra infinito.
- Ejemplos de distribuciones iniciales de referencia se encuentran en el formulario.
- **DISTRIBUCIONES CONJUGADAS**: Las distribuciones conjugadas surgen de la búsqueda de cuantificar el conocimiento inicial de tal forma que la distribución final sea fácil de obtener de “manera analítica”. Debido a los avances tecnológicos, esta justificación no es válida en la actualidad.
  - Definición: Familia conjugada. *Se dice que una familia de distribuciones de  $\theta$  es conjugada con respecto a un determinado*

*modelo probabilístico  $f(\mathbf{x}|\theta)$  si para cualquier distribución inicial perteneciente a tal familia, se obtiene una distribución final que también pertenece a ella.*

- EJEMPLO 10: Sea  $X_1, X_2, \dots, X_n$  una m.a. de  $\text{Ber}(\theta)$ . Sea  $\theta \sim \text{Beta}(a, b)$  la distribución inicial de  $\theta$ . Entonces,

$$f(\mathbf{x}|\theta) = \theta^{\sum x_i} (1 - \theta)^{n - \sum x_i} \prod_{i=1}^n I_{\{0,1\}}(x_i)$$

$$f(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1} I_{(0,1)}(\theta)$$

$$\Rightarrow f(\theta|\mathbf{x}) \propto \theta^{a+\sum x_i-1} (1-\theta)^{b+n-\sum x_i-1} I_{(0,1)}(\theta)$$

$$\therefore f(\theta|\mathbf{x}) = \frac{\Gamma(a_1+b_1)}{\Gamma(a_1)\Gamma(b_1)} \theta^{a_1-1} (1-\theta)^{b_1-1} I_{(0,1)}(\theta),$$

donde  $a_1 = a + \sum x_i$  y  $b_1 = b + n - \sum x_i$ . Es decir,  $\theta|\mathbf{x} \sim \text{Beta}(a_1, b_1)$ .

- Más ejemplos de familias conjugadas se encuentran en el formulario.

## 1.4 Problemas de inferencia paramétrica

- Los problemas típicos de inferencia son: estimación puntual, estimación por intervalos y prueba o contraste de hipótesis.
- ESTIMACIÓN PUNTUAL. El problema de estimación puntual visto como problema de decisión se describe de la siguiente manera:
  - $\mathbf{D} = \mathbf{E} = \Theta$ .

- $v(\tilde{\theta}, \theta)$  la pérdida de estimar mediante  $\tilde{\theta}$  el verdadero valor del parámetro de interés  $\theta$ . Considérense tres funciones de pérdida:

1) *Función de pérdida cuadrática:*

$$v(\tilde{\theta}, \theta) = a(\tilde{\theta} - \theta)^2, \text{ donde } a > 0$$

En este caso, la decisión óptima que minimiza la pérdida esperada es

$$\tilde{\theta} = E(\theta).$$

**La mejor estimación de  $\theta$  con pérdida *cuadrática* es la *media* de la distribución de  $\theta$  al momento de producirse la estimación.**

2) *Función de pérdida absoluta:*

$$v(\tilde{\theta}, \theta) = a|\tilde{\theta} - \theta|, \text{ donde } a > 0$$

En este caso, la decisión óptima que minimiza la pérdida esperada es

$$\tilde{\theta} = \text{Med}(\theta).$$

**La mejor estimación de  $\theta$  con pérdida *absoluta* es la *mediana* de la distribución de  $\theta$  al momento de producirse la estimación.**

3) *Función de pérdida vecindad:*

$$v(\tilde{\theta}, \theta) = 1 - I_{B_\varepsilon(\tilde{\theta})}(\theta),$$

donde  $B_\varepsilon(\tilde{\theta})$  denota una vecindad (bola) de radio  $\varepsilon$  con centro en  $\tilde{\theta}$ .

En este caso, la decisión óptima que minimiza la pérdida esperada cuando  $\varepsilon \rightarrow 0$  es

$$\tilde{\theta} = \text{Moda}(\theta).$$

**La mejor estimación de  $\theta$  con pérdida *vecindad* es la moda de la distribución de  $\theta$  al momento de producirse la estimación.**

- EJEMPLO 11: Sean  $X_1, X_2, \dots, X_n$  una m.a. de una población  $\text{Ber}(\theta)$ . Supongamos que la información inicial que se tiene se puede describir mediante una distribución Beta, i.e.,  $\theta \sim \text{Beta}(a, b)$ . Como demostramos en el ejemplo pasado, la distribución final para  $\theta$  es también una distribución Beta, i.e.,

$$\theta|\underline{x} \sim \text{Beta}\left(a + \sum_{i=1}^n X_i, b + n - \sum_{i=1}^n X_i\right).$$

La idea es estimar puntualmente a  $\theta$ ,

- 1) Si se usa una función de pérdida cuadrática:

$$\tilde{\theta} = E(\theta|\underline{x}) = \frac{a + \sum x_i}{a + b + n},$$

- 2) Si se usa una función de pérdida vecindad:

$$\tilde{\theta} = \text{Moda}(\theta|\underline{x}) = \frac{a + \sum x_i - 1}{a + b + n - 2}.$$

- ESTIMACIÓN POR INTERVALO. El problema de estimación por intervalo visto como problema de decisión se describe de la siguiente manera:

- $\mathbf{D} = \{D : D \subset \Theta\},$

donde,  $D$  es un *intervalo de probabilidad* al  $(1-\alpha)$  si  $\int_D f(\theta) d\theta = 1 - \alpha$ .

*Nota:* para un  $\alpha \in (0, 1)$  fijo no existe un único intervalo de probabilidad.

- $\mathbf{E} = \Theta.$

- $v(D, \theta) = \|D\| - I_D(\theta)$  la pérdida de estimar mediante  $D$  el verdadero valor del parámetro de interés  $\theta$ .

Esta función de pérdida refleja la idea intuitiva que para un  $\alpha$  dado es preferible reportar un intervalo de probabilidad  $D^*$  cuyo tamaño sea mínimo. Por lo tanto,

**La mejor estimación por intervalo de  $\theta$  es el intervalo  $D^*$  cuya longitud es mínima.**

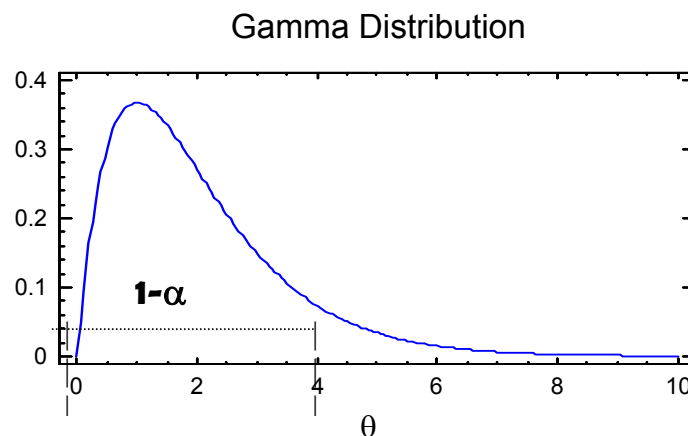
- El intervalo  $D^*$  de longitud mínima satisface la propiedad de ser un *intervalo de máxima densidad*, es decir

$$\text{si } \theta_1 \in D^* \text{ y } \theta_2 \notin D^* \Rightarrow f(\theta_1) \geq f(\theta_2)$$

- ¿Cómo se obtiene el intervalo de mínima longitud (máxima densidad)?

Los pasos a seguir son:

- Localizar el punto más alto de la función de densidad (posterior) de  $\theta$ .
- A partir de ese punto trazar líneas rectas horizontales en forma descendiente hasta que se acumule  $(1-\alpha)$  de probabilidad.



- CONTRASTE DE HIPÓTESIS. El problema de contraste de hipótesis es un problema de decisión sencillo y consiste en elegir entre dos modelos o hipótesis alternativas  $H_0$  y  $H_1$ . En este caso,
- $D = E = \{H_0, H_1\}$
  - $v(d, \theta)$  la función de pérdida que toma la forma,

$v(d, \theta)$	$H_0$	$H_1$
$H_0$	$v_{00}$	$v_{01}$
$H_1$	$v_{10}$	$v_{11}$

donde,  $v_{00}$  y  $v_{11}$  son la pérdida de tomar una decisión correcta (generalmente  $v_{00} = v_{11} = 0$ ),

$v_{10}$  es la pérdida de rechazar  $H_0$  (aceptar  $H_1$ ) cuando  $H_0$  es cierta y

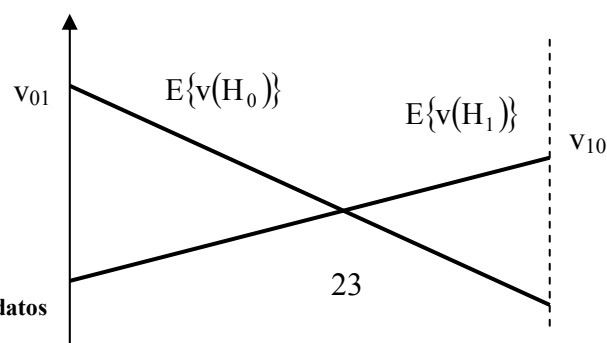
$v_{01}$  es la pérdida de no rechazar  $H_0$  (aceptar  $H_0$ ) cuando  $H_0$  es falsa.

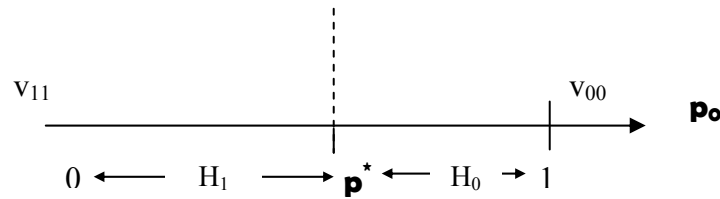
Sea  $p_0 = P(H_0)$  = probabilidad asociada a la hipótesis  $H_0$  al momento de tomar la decisión (inicial o final). Entonces, la pérdida esperada para cada hipótesis es:

$$E\{v(H_0)\} = v_{00}p_0 + v_{01}(1 - p_0) = v_{01} - (v_{01} - v_{00})p_0$$

$$E\{v(H_1)\} = v_{10}p_0 + v_{11}(1 - p_0) = v_{11} - (v_{11} - v_{10})p_0$$

cuya representación gráfica es de la forma:





donde,  $p^* = \frac{V_{01} - V_{11}}{V_{10} - V_{11} + V_{01} - V_{00}}.$

Finalmente, la solución óptima es aquella que minimiza la pérdida esperada:

$$\text{si } E\{v(H_0)\} < E\{v(H_1)\} \Leftrightarrow \frac{p_0}{1-p_0} > \frac{V_{01} - V_{11}}{V_{10} - V_{00}} \Leftrightarrow p_0 > p^* \Rightarrow H_0$$

**$H_0$  si  $p_0$  es suficientemente grande comparada con  $1-p_0$ .**

$$\text{si } E\{v(H_0)\} > E\{v(H_1)\} \Leftrightarrow \frac{p_0}{1-p_0} < \frac{V_{01} - V_{11}}{V_{10} - V_{00}} \Leftrightarrow p_0 < p^* \Rightarrow H_1$$

**$H_1$  si  $p_0$  es suficientemente pequeña comparada con  $1-p_0$ .**

$$\text{si } p_0 = p^* \Rightarrow H_0 \text{ ó } H_1$$

**Indiferente entre  $H_0$  y  $H_1$  si  $p_0$  no es ni suficientemente grande ni suficientemente pequeña comparada con  $1-p_0$ .**