

# Teoría Bayesiana

Julián Jiménez-Cárdenas<sup>1</sup>

<sup>1</sup>Universidad Nacional de Colombia, Bogotá.

`juojimenezca@unal.edu.co`

## 1 Preliminares

- Espacio de Probabilidad
- Probabilidad Condicional e Independencia de Eventos
- Variables aleatorias
- Vectores Aleatorios

## 2 Función de Verosimilitud

- Estadística
- Estimación bayesiana

## 3 Cadenas de Markov Monte Carlo

- Introducción
- Algoritmo de Metropolis-Hastings

## 4 Referencias

# $\sigma$ -álgebra I

## Definición (Experimento Aleatorio)

*Un experimento se dice aleatorio si su resultado no se puede determinar de antemano.*

## Definición (Espacio de Muestra)

*El conjunto  $\Omega$  de todos los posibles resultados de un experimento aleatorio se llama espacio de muestra. Un elemento  $\omega \in \Omega$  se llama resultado o muestra.  $\Omega$  se dice discreto si es finito o contable.*

## $\sigma$ -álgebra II

### Definición ( $\sigma$ -álgebra)

Tome  $\Omega \neq \emptyset$ . Una colección  $\mathfrak{I}$  de subconjuntos de  $\Omega$  se llama  $\sigma$ -álgebra sobre  $\Omega$  si:

- ①  $\Omega \in \mathfrak{I}$ ,
- ② Si  $A \in \mathfrak{I}$ , entonces  $A^c \in \mathfrak{I}$  y,
- ③ Si  $A_1, A_2, \dots \in \mathfrak{I}$ , entonces  $\bigcup_{i=1}^{\infty} A_i \in \mathfrak{I}$ .

Los elementos de  $\mathfrak{I}$  se llaman eventos.

### Teorema

Si  $\Omega \neq \emptyset$  y  $\mathfrak{I}_1, \mathfrak{I}_2, \dots$  son  $\sigma$ -álgebras sobre  $\Omega$ , entonces  $\bigcap_{i=1}^{\infty} \mathfrak{I}_i$  es una  $\sigma$ -álgebra sobre  $\Omega$ .

# $\sigma$ -álgebra III

## Demostración.

Como  $\Omega \in \mathfrak{I}_j$ , para  $j = 1, 2, \dots$ ,  $\Omega \in \bigcap_{j=1}^{\infty} \mathfrak{I}_j$ . Si  $A \in \bigcap_{j=1}^{\infty} \mathfrak{I}_j$ ,  $A \in \mathfrak{I}_j$ , para  $j = 1, 2, \dots$ , de modo que  $A^c \in \mathfrak{I}_j$ , y  $A^c \in \bigcap_{j=1}^{\infty} \mathfrak{I}_j$ . Por último, si

$$A_1, A_2, \dots \in \bigcap_{j=1}^{\infty} \mathfrak{I}_j,$$

para todo  $j = 1, 2, \dots$ ,  $A_1, A_2, \dots \in \mathfrak{I}_j$ , de modo que

$$\bigcup_{i=1}^{\infty} A_i \in \mathfrak{I}_j \text{ y } \bigcup_{i=1}^{\infty} A_i \in \bigcap_{j=1}^{\infty} \mathfrak{I}_j.$$



## $\sigma$ -álgebra IV

### Definición ( $\sigma$ -álgebra generada)

Tome  $\Omega \neq \emptyset$  y  $\mathcal{A}$  como una colección de subconjuntos de  $\Omega$ . Si  $\mathcal{M} := \{\mathfrak{I} : \mathfrak{I} \text{ es una } \sigma\text{-álgebra sobre } \Omega \text{ que contiene a } \mathcal{A}\}$ ,

$$\sigma(\mathcal{A}) := \bigcap_{\mathfrak{I} \in \mathcal{M}} \mathfrak{I}$$

es la  $\sigma$ -álgebra más pequeña sobre  $\Omega$  que contiene a  $\mathcal{A}$ . Esta  $\sigma$ -álgebra se conoce como  $\sigma$ -álgebra generada por  $\mathcal{A}$ .

### Definición (Espacio de medida)

Tome  $\Omega \neq \emptyset$  y sea  $\mathfrak{I}$  una  $\sigma$ -álgebra sobre  $\Omega$ . La pareja  $(\Omega, \mathfrak{I})$  se llama espacio de medida.

## $\sigma$ -álgebra $\mathcal{V}$

$\emptyset$  es el evento imposible.  $\Omega$  es el evento seguro y  $\{\omega\}$ , con  $\omega \in \Omega$  es un evento simple. Decimos que el evento  $A$  ocurre después de llevar a cabo el experimento aleatorio si se obtiene un resultado en  $A$ , esto es,  $A$  ocurre si el resultado es algún  $\omega \in A$ .

- ① El evento  $A \cup B$  ocurre si y sólo si  $A$  ocurre,  $B$  pasa, o ambos ocurren.
- ② El evento  $A \cap B$  ocurre si y sólo si  $A$  y  $B$  ocurren a la vez.
- ③ El evento  $A^c$  ocurre si y sólo si  $A$  no ocurre.
- ④ El evento  $A - B$  ocurre si y sólo si  $A$  ocurre pero  $B$  no ocurre.

### Definición (Eventos mutuamente excluyentes)

*Dos eventos  $A$  y  $B$  se dicen mutuamente excluyentes si  $A \cap B = \emptyset$ .*

# Espacio de probabilidad I

## Definición (Frecuencia relativa)

*Para cada evento  $A$ , el número  $f_r(A) := \frac{n(A)}{n}$  se llama la frecuencia relativa de  $A$ , donde  $n(A)$  indica el número de veces que ocurre  $A$  en  $n$  repeticiones del experimento aleatorio.*

Cuando  $n \rightarrow \infty$ , se puede hablar de la probabilidad de que ocurra el evento  $A$ , normalizada de 0 a 1. La formalización de este concepto se encuentra en la idea del espacio de probabilidad.



# Espacio de probabilidad II

## Definición (Espacio de probabilidad)

Tome  $(\Omega, \mathfrak{I})$  como un espacio de medida. Una función real  $P$  sobre  $\mathfrak{I}$  que satisface las siguientes condiciones:

- 1  $P(A) \geq 0$  para todo  $A \in \mathfrak{I}$  (no negativa),
- 2  $P(\Omega) = 1$  (normalizada) y,
- 3 si  $A_1, A_2, \dots$  son eventos mutuamente excluyentes en  $\mathfrak{I}$ , esto es, si

$A_i \cap A_j = \emptyset$  para todo  $i \neq j$ , entonces

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i),$$

se llama medida de probabilidad sobre  $(\Omega, \mathfrak{I})$ . La tripleta  $(\Omega, \mathfrak{I}, P)$  se llama espacio de probabilidad.

## Notas I

Aplicando el teorema anterior de forma inductiva, para algunos eventos  $A_1, A_2, \dots, A_n \in \mathfrak{I}$ :

$$\begin{aligned} P(A_1 \cup A_2 \cup \dots \cup A_n) &= \sum_{i=1}^n P(A_i) - \sum_{i_1 < i_2} P(A_{i_1} \cap A_{i_2}) + \dots \\ &\quad + (-1)^{r+1} \sum_{i_1 < i_2 < \dots < i_r} P(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_r}) \\ &\quad + \dots + (-1)^{n+1} P(A_1 \cap A_2 \cap \dots \cap A_n). \end{aligned}$$

---

Tome  $(\Omega, \mathfrak{I}, P)$  como un espacio de probabilidad con  $\Omega$  finito o contable y  $\mathfrak{I} = \mathbb{P}(\Omega)$ . Tome  $\emptyset \neq A \in \mathfrak{I}$ . Es claro que

# Notas II

$$A = \bigcup_{\omega \in A} \{\omega\}, \text{ de modo que}$$

$$P(A) = \sum_{\omega \in A} P(\omega), \text{ donde } P(\omega) := P(\{\omega\}).$$

Así,  $P$  queda completamente definido por  $p_j := P(\omega_j)$ , donde  $\omega_j \in \Omega$ . El vector  $|\Omega|$ -dimensional  $p := (p_1, p_2, \dots)$  satisface las siguientes condiciones:

- $p_j \geq 0$  y
- $\sum_{j=1}^{\infty} p_j = 1$ .

Un vector que satisface las anteriores condiciones se llama **vector de probabilidad**.

# Introducción

Tome  $B$  como un evento cuya opción de ocurrir debe ser medida bajo la suposición de que otro evento  $A$  fue observado. Si el experimento se repite  $n$  veces bajo las mismas circunstancias, entonces la frecuencia relativa de  $B$  bajo la condición  $A$  se define como

$$f_r(B|A) := \frac{n(A \cap B)}{n(A)} = \frac{\frac{n(A \cap B)}{n}}{\frac{n(A)}{n}} = \frac{f_r(A \cap B)}{f_r(A)}, \text{ si } n(A) > 0.$$

Esto motiva la siguiente definición

# Probabilidad Condicional I

## Definición (Probabilidad condicional)

*Tome  $(\Omega, \mathfrak{I}, P)$  como un espacio de probabilidad. Si  $A, B \in \mathfrak{I}$ , con  $P(A) > 0$ , entonces la probabilidad del evento  $B$  bajo la condición  $A$  se define como sigue*

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

El siguiente teorema provee algunas propiedades de la probabilidad condicional.

# Probabilidad Condicional II

## Teorema (Medida de probabilidad condicional)

*Tome  $(\Omega, \mathfrak{I}, P)$  como un espacio de probabilidad y  $A \in \mathfrak{I}$ , con  $P(A) > 0$ . Entonces:*

- ①  *$P(\cdot|A)$  es una medida de probabilidad sobre  $\Omega$  centrada en  $A$ , esto es,  $P(A|A) = 1$ .*
- ② *Si  $A \cap B = \emptyset$ , entonces  $P(B|A) = 0$ .*
- ③  *$P(B \cap C|A) = P(B|A \cap C)P(C|A)$  si  $P(A \cap C) > 0$ .*
- ④ *Si  $A_1, A_2, \dots, A_n \in \mathfrak{I}$ , con  $P(A_1 \cap A_2 \cap \dots \cap A_{n-1}) > 0$ , entonces*

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \dots P(A_n|A_1 \cap A_2 \cap \dots \cap A_{n-1}).$$

# Teorema probabilidad total I

Los siguientes resultados son vitales para aplicaciones posteriores.

## Teorema (Teorema de probabilidad total)

*Tome  $A_1, A_2, \dots$  como una partición finita o contable de  $\Omega$ , esto es,  $A_i \cap A_j = \emptyset, \forall i \neq j$  y  $\bigcup_{i=1}^{\infty} A_i = \Omega$ , tal que  $P(A_i) > 0$ , para todo  $A_i \in \mathfrak{I}$ . Entonces, para todo  $B \in \mathfrak{I}$ :*

$$P(B) = \sum_i P(B|A_i)P(A_i).$$

# Teorema probabilidad total II

## Demostración.

Observe que

$$B = B \cap \Omega = B \cap \left( \bigcup_{i=1}^{\infty} A_i \right) = \bigcup_{i=1}^{\infty} B \cap A_i,$$

de modo que

$$P(B) = P\left(\bigcup_{i=1}^{\infty} B \cap A_i\right) = \sum_{i=1}^{\infty} P(B \cap A_i) = \sum_{i=1}^{\infty} P(B|A_i)P(A_i).$$





# Regla de Bayes I

Como corolario del teorema anterior, se obtiene un resultado conocido como **regla de Bayes**, que constituye la base para la **teoría Bayesiana**.

## Corolario (Regla de Bayes)

*Tome  $A_1, A_2, \dots$  como una partición finita o contable de  $\Omega$  con  $P(A_i) > 0$ , para todo  $i$ ; entonces, para todo  $B \in \mathfrak{I}$  con  $P(B) > 0$ :*

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_j P(B|A_j)P(A_j)}, \forall i.$$

# Regla de Bayes II

## Demostración.

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i)P(B|A_i)}{P(B)} = \frac{P(A_i)P(B|A_i)}{\sum_j P(B|A_j)P(A_j)}.$$

Con la partición  $A_1 = A, A_2 = A^c$  se obtiene la forma usual de la regla de Bayes. □

# Distribuciones a priori y a posteriori

## Definición (Distribuciones a priori y a posteriori)

*Tome  $A_1, A_2, \dots$  como una partición finita o contable de  $\Omega$ , con  $P(A_i) > 0$ , para todo  $i$ . Si  $P(B) > 0$ , con  $B \in \mathfrak{I}$ , entonces  $\{P(A_n)\}_n$  se llama *distribución a priori* (antes de que  $B$  ocurra), y  $\{P(A_n|B)\}_n$  se llama *distribución a posteriori* (después de que  $B$  ocurra).*

Algunas veces, la ocurrencia de un evento  $B$  no afecta la probabilidad de un evento  $A$ , es decir,

$$P(A|B) = P(A).$$

En este caso, se dice que el evento  $A$  es independiente del evento  $B$ . Esto motiva la siguiente definición.

# Eventos Independientes I

## Definición (Eventos independientes)

*Dos eventos  $A$  y  $B$  se dicen independientes si y sólo si*

$$P(A \cap B) = P(A)P(B).$$

*Si esta condición no se tiene, se dice que los eventos son dependientes.*

# Variables aleatorias I

## Definición (Variable aleatoria)

Tome  $(\Omega, \mathfrak{I}, P)$  como un espacio de probabilidad. Una variable aleatoria es un mapa  $X : \Omega \rightarrow \mathbb{R}$  tal que, para todo  $A \in \mathbb{B}$ ,  $X^{-1}(A) \in \mathfrak{I}$ , donde  $\mathbb{B}$  es la  $\sigma$ -álgebra de Borel sobre  $\mathbb{R}$  ( $\sigma$ -álgebra más pequeña que contiene todos los intervalos de la forma  $(-\infty, a]$ ).

El conjunto de posibles valores de  $X$  es

$\mathbb{S} := \{x \in \mathbb{R} : \exists \omega \in \Omega \text{ tal que } X(\omega) = x\}$ , conocido como **soporte de la variable aleatoria**  $X$ .

# Variables aleatorias II

Si  $X$  es una variable aleatoria definida sobre un espacio de probabilidad  $(\Omega, \mathfrak{I}, P)$ , se introduce la notación

$$\{X \in B\} := \{\omega \in \Omega : X(\omega) \in B\}, \text{ con } B \in \mathbb{B}.$$

## Definición (Variable aleatoria discreta)

*Una variable aleatoria  $X$  se dice discreta cuando el soporte  $\mathbb{S}$  de  $X$  es un subconjunto finito o contable de  $\mathbb{R}$ . Para  $x \in \mathbb{S}$ , la función  $f(x) = P(X = x)$  se llama función de densidad de probabilidad (pdf para abreviar).*

# Variables aleatorias III

## Definición (Variable aleatoria continua)

*Una variable aleatoria  $X$  se dice continua si el soporte  $\mathbb{S}$  de  $X$  es la unión de uno o más intervalos y si existe una función no negativa y real  $f(x)$  tal que  $P(X \leq x) = \int_{-\infty}^x f(t)dt$ . La función  $f(x)$  se llama función de densidad de probabilidad (pdf).*

Algunas propiedades de la *pdf* discreta son las siguientes:

- 1  $f(x) \geq 0, \forall x \in \mathbb{S}$  y  $f(x) = 0, \forall x \notin \mathbb{S}$ .
- 2  $\sum_{x \in \mathbb{S}} f(x) = 1$ .

## Variables aleatorias IV

$$\textcircled{3} \quad P(X \in B) = \sum_{x \in B} f(x).$$

Análogamente, para la *pdf* continua:

$$\textcircled{1} \quad f(x) \geq 0, \forall x \in \mathbb{S} \text{ y } f(x) = 0, \forall x \notin \mathbb{S}.$$

$$\textcircled{2} \quad \int_{\mathbb{S}} f(x) dx = 1.$$

$$\textcircled{3} \quad P(X \in B) = \int_B f(x) dx.$$

### Definición (Función de distribución acumulativa)

*La función de distribución acumulativa (CDF, para abreviar) de una variable aleatoria se define como la función  $F(x) = P(X \leq x)$ .*

El siguiente teorema resume algunas propiedades importantes de una *CDF*.



# Variables aleatorias V

## Teorema

Si  $X$  es una variable aleatoria, con CDF  $F(x)$ , entonces:

- 1  $\lim_{x \rightarrow -\infty} F(x) = 0$  y  $\lim_{x \rightarrow \infty} F(x) = 1$ .
- 2  $F(x)$  es no decreciente; esto es,  $F(x) \leq F(y)$ , siempre que  $x \leq y$ .
- 3  $F(x)$  es continua por derecha.
- 4  $P(a < X \leq b) = F(b) - F(a)$ .

# Variables aleatorias VI

## Teorema

Si  $X$  es una variable aleatoria discreta con CDF  $F(x)$  y soporte  $\mathbb{S} = \{x_0, x_1, \dots\}$ , con  $x_0 < x_1 < \dots$ , entonces, para  $x_k \in \mathbb{S}$ ,

$$f(x_k) = F(x_k) - F(x_{k-1}).$$

## Teorema

Para una variable aleatoria continua,  $f(x) = \frac{dF}{dx}$ ,  $\forall x \in \mathbb{R}$ .

# Vectores Aleatorios I

## Definición (Vector aleatorio)

*Un vector aleatorio  $\vec{X} = (X_1, X_2, \dots, X_k)$  es un vector  $k$ -dimensional, donde  $X_1, \dots, X_k$  son variables aleatorias. Un vector aleatorio se dice discreto cuando cada una de las variables aleatorias que lo conforman son discretas, y continuo cuando son continuas.*

## Definición (Variable aleatoria bivariada)

*Un vector aleatorio bidimensional  $\vec{X} = (X_1, X_2)$  se llama variable aleatoria bivariada.*

## Vectores Aleatorios II

De modo similar al caso de las variables aleatorias, los vectores aleatorios tienen *pdf*, un soporte y una *CDF*. El soporte de un vector aleatorio  $k$ –dimensional es el conjunto de valores que puede tomar, denotado por  $\mathbb{S}_{\vec{X}} \subseteq \mathbb{R}^k$ .

### Definición (Función de densidad de probabilidad adjunta discreta)

*Tome  $\vec{X}$  como un vector aleatorio discreto  $k$ –dimensional. La pdf adjunta de  $\vec{X}$  se define como*

$$f(\vec{x}) := f(x_1, x_2, \dots, x_k) = P(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k)$$

*para  $\vec{x} = (x_1, x_2, \dots, x_k) \in \mathbb{S}_{\vec{X}}$ .*

# Vectores Aleatorios III

La *pdf* conjunta discreta tiene las siguientes propiedades:

- 1  $0 \leq f(x_1, x_2, \dots, x_k) \leq 1, \forall \vec{x} \in \mathbb{S}_{\vec{X}}.$
- 2  $\sum_{\vec{x} \in \mathbb{S}_{\vec{X}}} f(\vec{x}) = 1.$
- 3 Para cualquier subconjunto  
 $B \subseteq \mathbb{S}_{\vec{X}}, P(\vec{X} \in B) = \sum_{\{\vec{x} \in \mathbb{S}_{\vec{X}}: \vec{x} \in B\}} f(\vec{x}).$

## Vectores Aleatorios IV

### Definición (Función de densidad de probabilidad adjunta continua)

Tome  $\vec{X}$  como un vector aleatorio  $k$ -dimensional continuo. La pdf continua de  $\vec{X}$  se define como cualquier función no negativa  $f(\vec{x})$  que satisfaga las siguientes propiedades:

- 1  $f(x_1, \dots, x_k) > 0, \forall \vec{x} \in \mathbb{S}_{\vec{X}}$ .
- 2  $\int_{\mathbb{S}_{\vec{X}}} f(x_1, \dots, x_k) dx_1 \cdots dx_k = 1$ .
- 3 Para cualquier subconjunto  $B \subset \mathbb{S}_{\vec{X}}$ ,  $P(\vec{X} \in B) = \int_B f(x_1, \dots, x_k) dx_1 \cdots dx_k$ .

# Vectores Aleatorios V

## Definición (Función de distribución acumulativa adjunta)

Tome  $\vec{X} = (X_1, X_2, \dots, X_k)$  como un vector aleatorio  $k$ -dimensional. La CDF adjunta de  $\vec{X}$  se define como

$$F(x_1, x_2, \dots, x_k) = P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_k \leq x_k)$$

$$\forall (x_1, \dots, x_k) \in \mathbb{R}^k.$$

# Vectores Aleatorios VI

## Definición (Función de densidad de probabilidad marginal)

Tome  $\vec{X} = (X_1, \dots, X_k)$  como un vector aleatorio  $k$ —dimensional. La función de densidad de probabilidad marginal de la variable aleatoria  $X_i$  es, para los casos discreto y continuo:

$$f_i(x_i) = \underbrace{\sum_{x_1 \in \mathbb{S}_X} \cdots \sum_{x_k \in \mathbb{S}_{X_k}}}_{\text{quitando la suma sobre } x_i} f(x_1, \dots, x_k)$$

$$f_i(x_i) = \underbrace{\int_{x_1 \in \mathbb{S}_{X_1}} \cdots \int_{x_k \in \mathbb{S}_{X_k}}}_{\text{quitando la integral sobre } x_i} f(x_1, \dots, x_k) \prod_{n \neq i} dx_n$$



# Vectores Aleatorios VII

## Definición (Función de densidad de probabilidad condicional)

*Tome  $\vec{X}(X_1, \dots, X_k)$  como un vector aleatorio  $k$ —dimensional. Para un valor fijo de  $x_i$ , donde  $f_i(x_i) > 0$ , la función de densidad de probabilidad condicional para  $\vec{Y}|X_i$ ; donde  $\vec{Y}$  es un vector aleatorio  $(k - 1)$ —dimensional con todas las variables aleatorias de  $\vec{X}$ , a excepción de  $X_i$ ; es*

$$f(\vec{y}|x_i) = \frac{f(x_1, \dots, x_k)}{f_i(x_i)},$$

*donde  $\vec{y} \in \mathbb{S}_{\vec{Y}}$ .*

# Vectores Aleatorios VIII

## Definición (Colección independiente de variables aleatorias)

*Una colección de variables aleatorias  $\{X_1, X_2, \dots, X_k\}$  se dice independiente cuando*

$$F(x_1, x_2, \dots, x_k) = \prod_{i=1}^k F_i(x_i), \forall \vec{x} \in \mathbb{R}^k,$$

*donde  $F_i(x_i)$  es la CDF marginal de la variable aleatoria  $X_i$  (determinada a partir de la pdf marginal:  $F_i(x_i) := P(X_i \leq x_i)$ ).*

# Vectores Aleatorios IX

## Definición (Colección independiente de variables aleatorias)

*También se puede definir la independencia entre variables aleatorias usando las pdf, en el sentido de que la misma colección de variables aleatorias se dice independiente cuando*

$$f(x_1, x_2, \dots, x_k) = \prod_{i=1}^k f_i(x_i), \forall \vec{x} \in \mathbb{S}_{\vec{X}},$$

*donde  $f_i(x_i)$  es la pdf marginal asociada a la variable aleatoria  $X_i$ .*

# Vectores Aleatorios $\mathbf{X}$

**Definición (Colección de variables aleatorias independientes idénticamente distribuidas)**

*Una colección de variables aleatorias  $\{X_1, X_2, \dots, X_k\}$  se dice independiente e idénticamente distribuidas (iid, para abreviar) si y sólo si  $X_1, X_2, \dots, X_k$  son variables aleatorias independientes y la pdf de cada variable aleatoria es idéntica.*

## Estimación paramétrica

- 1 Un modelo probabilístico  $f(x, \theta)$ , especificado con los valores de los parámetros desconocidos.
- 2 Un conjunto de posibles valores de  $\theta$  bajo consideración, llamado el espacio de parámetros, denotado por  $\Theta$ .
- 3 Una muestra aleatoria de  $n$  observaciones del modelo probabilístico.
- 4 Un conjunto de estimadores puntuales para los valores de los parámetros desconocidos, basados en la información contenida en la muestra aleatoria.
- 5 Las propiedades específicas de los estimadores que permiten evaluar la precisión y eficiencia del estimador.

# Muestra y Estadística I

## Definición (Muestra)

*Una colección de variables aleatorias  $X_1, \dots, X_k$  se llama muestra de tamaño  $n$ . Una muestra de  $n$  variables aleatorias independientes  $X_1, \dots, X_n$  se llama muestra aleatoria.*

## Definición (Estadística y estimador)

*Dada una muestra  $X_1, X_2, \dots, X_n$ , una estadística  $T = T(X_1, \dots, X_n)$  es una función de la muestra que no depende de ningún otro parámetro desconocido. Un estimador es una estadística que se usa para determinar una cantidad desconocida, y el estimado es el valor observado del estimador (evaluando la función en la muestra).*

## Muestra y Estadística II

### Definición (Distribución muestral)

*Para una muestra  $X_1, \dots, X_n$  y una estadística  $T = T(X_1, \dots, X_n)$ , la distribución muestral de la estadística  $T$  es la distribución de probabilidad asociada a la variable aleatoria  $T$ . La pdf de la distribución muestral se denota como  $f_T(t; \theta)$ .*

## Muestra y Estadística III

### Definición (Valor esperado)

*Tome  $X$  como una variable aleatoria con pdf  $f(x)$  en  $\mathbb{S}_X$ . El valor esperado de la variable aleatoria  $X$ , denotado por  $E(X)$ , se define como*

$$E(X) = \sum_{x \in \mathbb{S}_X} xf(x)$$

*cuando  $X$  es una variable aleatoria discreta, y como*

$$E(X) = \int_{x \in \mathbb{S}_X} xf(x)dx$$

*cuando  $X$  es una variable aleatoria continua.*



## Muestra y Estadística IV

### Definición (Estimador imparcial)

*Una estadística  $T$  se dice estimador imparcial de un parámetro  $\theta$  cuando  $E(T) = \theta$ ,  $\forall \theta \in \Theta$ . Una estadística se conoce como estimador parcial de  $\theta$  cuando  $E(T) \neq \theta$ , y la parcialidad de una estadística  $T$  para estimar un parámetro  $\theta$  se define como  $\text{Bias}(T; \theta) = E(T) - \theta$ .*

### Definición (Estimador asintóticamente imparcial)

*Una estadística  $T_n = T(X_1, \dots, X_n)$  se conoce como estimador asintóticamente imparcial de un parámetro  $\theta$  cuando*

$$\lim_{n \rightarrow \infty} \text{Bias}(T_n; \theta) = 0.$$

## Muestra y Estadística V

Algunas variables útiles para determinar la precisión y exactitud del estimador son las siguientes:

$$SE(T) := \sqrt{E((T - E(T))^2)} := \sqrt{Var(T)}.$$

$$MSE(T; \theta) = E((T - \theta)^2).$$

Una estadística que contiene toda la información relevante acerca de  $\theta$  en una muestra se conoce como estadística suficiente.

## Muestra y Estadística VI

### Definición (Estadística suficiente)

*Tome  $X_1, \dots, X_n$  como una muestra de variables aleatorias iid con pdf común  $f(x; \theta)$ , para  $\theta \in \Theta \subseteq \mathbb{R}^d$ . Un vector de estadísticas  $\vec{S}(\vec{X}) := (S_1(\vec{X}), \dots, S_k(\vec{X}))$  se dice que es una estadística suficiente  $k$ -dimensional para un parámetro  $\theta$  si y sólo si la distribución condicional de  $\vec{X}$  dado  $S = s$  no depende de  $\theta$ , para ningún valor de  $s$ .*

## Muestra y Estadística VII

### Definición (Función de verosimilitud)

Para una muestra  $X_1, \dots, X_n$ , la función de verosimilitud  $L(\theta|\vec{X})$  es la pdf adjunta de  $\vec{X} = (X_1, \dots, X_n)$ , es decir,

$$L(\theta|\vec{X}) = f(x_1, \dots, x_n; \theta)$$

La función logarítmica de verosimilitud  $\ell(\theta)$  se define como el logaritmo de la función de verosimilitud.

Cuando  $X_1, \dots, X_n$  es una muestra de variables aleatorias *iid*, se puede escribir la función de verosimilitud como

$$L(\theta) = \prod_{i=1}^n f(x_i; \theta).$$

## Muestra y Estadística VIII

### Teorema (Teorema de factorización de Neyman-Fisher)

*Tome  $X_1, \dots, X_n$  como una muestra de variables aleatorias iid con pdf  $f(x; \theta)$ , y espacio de parámetros  $\Theta$ . Una estadística  $S(\vec{X})$  es suficiente para  $\theta$  si y sólo si  $L(\theta)$  se puede factorizar como*

$$L(\theta) = g(S(\vec{x}); \theta)h(\vec{x}),$$

*donde  $g(S(\vec{x}); \theta)$  no depende de  $\vec{x} = (x_1, \dots, x_n)$ , excepto a través de  $S(\vec{x})$ , y  $h(\vec{x})$  no depende de  $\theta$ .*

## Ley de verosimilitud

Tome  $X_1, \dots, X_n$  como una muestra de variables aleatorias *iid* con *pdf* común  $f(x; \vec{\theta})$  y espacio de parámetros  $\Theta$ . Para  $\vec{\theta} \in \Theta$ , mientras mayor sea el valor de  $L(\vec{\theta})$ , el modelo probabilístico con parámetro  $\vec{\theta}$  se ajusta más a los datos observados. Entonces, el grado con el cual la información de la muestra da soporte a un parámetro  $\vec{\theta}_0 \in \Theta$ , en comparación con otro parámetro  $\vec{\theta}_1 \in \Theta$  es igual a la razón entre sus verosimilitudes

$$\Lambda(\vec{\theta}_0, \vec{\theta}_1) = \frac{L(\vec{\theta}_0)}{L(\vec{\theta}_1)}.$$

## Definición (Función de Score)

*Tome  $X_1, \dots, X_n$  como una muestra de variables aleatorias con función de verosimilitud  $L(\vec{\theta})$ , para  $\vec{\theta} \in \Theta$ . Si la función de verosimilitud logarítmica  $\ell(\vec{\theta})$  es diferenciable, la función de Score se define como*

$$Sc(\vec{\theta}) = \nabla_{\vec{\theta}} \ell(\theta),$$

*de tal modo que una condición necesaria para que  $\vec{\theta} \in \Theta$  sea un máximo es que  $Sc(\theta) = \vec{0}$ .*

# Estimación bayesiana I

En la estimación paramétrica puntual bayesiana, el parámetro  $\theta$  se trata como una variable aleatoria, con su propia *pdf*  $\pi(\theta; \lambda)$ .

Las inferencias de  $\theta$  en la aproximación bayesiana están basadas en la distribución de  $\theta$  dados los valores observados de una muestra aleatoria  $\vec{x} = (x_1, \dots, x_n)$ , llamada distribución posterior, y denotada por  $f(\theta|\vec{x})$ .

Usando el teorema de Bayes y el teorema de la probabilidad total, en el caso de que  $\theta$  es una variable aleatoria continua,

$$f(\theta|\vec{x}) = \frac{f(\vec{x}, \theta; \lambda)}{f_{\vec{x}}(\vec{x})} = \frac{f(\vec{x}|\theta)\pi(\theta; \lambda)}{\int_{\mathbb{S}_{\theta}} f(\vec{x}|\theta)\pi(\theta; \lambda)d\theta}.$$

De modo similar, cuando  $\theta$  es una variable aleatoria discreta,



## Estimación bayesiana II

$$f(\theta|\vec{x}) = \frac{f(\vec{x}|\theta)\pi(\theta;\lambda)}{\sum_{\theta \in \mathbb{S}_\theta} f(\vec{x}|\theta)\pi(\theta;\lambda)}.$$

La distribución posterior combina la información disponible de  $\theta$  en la distribución previa y la función de verosimilitud para producir una distribución actualizada que contiene toda la información disponible de  $\theta$ .

El siguiente teorema indica que la distribución posterior depende de la muestra  $\vec{x}$  sólo bajo una estadística suficiente para  $\theta$ .

## Estimación bayesiana III

### Teorema

*Si  $X_1, \dots, X_n$  es una muestra de variables independientes iid con pdf común  $f(x|\theta)$ ,  $S$  es una estadística suficiente para  $\theta$ , y  $\pi(\theta; \lambda)$  una distribución previa para  $\theta$ , entonces la distribución posterior de  $\theta$  dado  $\vec{X}$  depende de la muestra sólo a través de una estadística suficiente  $S$ .*

# Teorema de Bayes

El teorema de Bayes estipula que

$$P(\theta) = \frac{L(\theta)\pi(\theta)}{\int_{\theta \in \mathbb{S}_\theta} L(\theta)\pi(\theta)d\theta} = \frac{L(\theta)\pi(\theta)}{Z}, \quad (1)$$

donde  $P(\theta) := f(\theta|\vec{x})$  es la distribución posterior,  $L(\theta) = f(\vec{x}|\theta)$  es la función de verosimilitud,  $\pi(\theta)$  es la distribución previa y la constante  $Z$  se conoce como evidencia.

## Aproximación del valor esperado I

Considere ahora una función  $f(\theta)$  del parámetro o parámetros del modelo que va a  $\mathbb{R}$ . El valor esperado de esta función sobre todo el soporte de  $\theta$  es

$$\underbrace{E_P[f(\theta)]}_{\text{valor esperado respecto a } P} = \int_{\theta \in \mathbb{S}_\theta} f(\theta) P(\theta) d\theta. \quad (2)$$

Considerando una partición  $\mathcal{P} = \{\theta_1 < \theta_2 < \dots < \theta_{n+1}\}$  (con  $\mathbb{S}_\theta \subseteq \mathbb{R}$ ). Definiendo  $\Delta\theta_i = \theta_{i+1} - \theta_i$  como el desplazamiento entre los elementos de la partición y  $\bar{\theta}_i = \frac{\theta_{i+1} + \theta_i}{2}$  como el punto medio

## Aproximación del valor esperado II

entre  $\theta_{i+1}$  y  $\theta_i$ , el valor esperado de  $f(\theta)$  se puede aproximar de la siguiente forma:

$$E_P[f(\theta)] \approx \sum_{i=1}^n f(\bar{\theta}_i) P(\bar{\theta}_i) \Delta\theta_i. \quad (3)$$

La generalización a más dimensiones es directa: se descompone el soporte  $\mathbb{S}_\theta \subseteq \mathbb{R}^N$  en  $n$  cuboides  $N$ —dimensionales. La contribución de cada uno de estos cuboides es proporcional al producto del peso  $f(\bar{\theta}_i)P(\bar{\theta}_i)$  (donde  $\bar{\theta}_i$  es el centro geométrico del  $i$ —ésimo cuboide) y al volumen

$$\Delta\theta_i = \prod_{j=1}^N \Delta\theta_{i,j},$$

## Aproximación del valor esperado III

donde  $\Delta\theta_{i,j}$  es el ancho del  $i$ -ésimo cuboide en la  $j$ -ésima dimensión.  
 Escribiendo el valor esperado de la siguiente forma

$$\begin{aligned}
 E_P[f(\theta)] &= \int_{\theta \in \mathbb{S}_\theta} f(\theta) P(\theta) d\theta = \frac{\int_{\theta \in \mathbb{S}_\theta} f(\theta) P(\theta) d\theta}{\underbrace{\int_{\theta \in \mathbb{S}_\theta} P(\theta) d\theta}_1} \\
 &= \frac{\int_{\theta \in \mathbb{S}_\theta} f(\theta) ZP(\theta) d\theta}{\int_{\theta \in \mathbb{S}_\theta} ZP(\theta) d\theta}
 \end{aligned}$$

y tomando  $\tilde{P}(\theta) = ZP(\theta)$ , se puede aproximar la evidencia a través del mismo procedimiento, dado que  $Z = \int_{\theta \in \mathbb{S}_\theta} \tilde{P}(\theta) d\theta$ :

## Aproximación del valor esperado IV

$$E_P[f(\theta)] = \frac{\int_{\theta \in \mathbb{S}_\theta} f(\theta) \tilde{P}(\theta) d\theta}{\int_{\theta \in \mathbb{S}_\theta} \tilde{P}(\theta) d\theta} \approx \frac{\sum_{i=1}^n f(\bar{\theta}_i) \tilde{P}(\bar{\theta}_i) \Delta\theta_i}{\sum_{i=1}^n \tilde{P}(\bar{\theta}_i) \Delta\theta_i}.$$

Algunas desventajas de este procedimiento son:

- ❶ El crecimiento exponencial de la cantidad de cubos necesarios para cubrir el soporte cuando aumenta su dimensión.
- ❷ Los pesos arbitrarios seleccionados al momento de crear la rejilla para aproximar la integral.

# Media pesada muestral I

La **media pesada muestral** de un conjunto de  $\{f_1, \dots, f_n\}$  observaciones con peso  $\{\omega_1, \dots, \omega_n\}$  de la siguiente forma:

$$f_{mean} = \frac{\sum_{i=1}^n \omega_i f_i}{\sum_{i=1}^n \omega_i}. \quad (4)$$

Si se toma  $f_i := f(\bar{\theta}_i)$  y  $\omega_i := \tilde{P}(\bar{\theta}_i) \Delta \theta_i$ , se observa que el valor esperado se puede escribir de manera aproximada como una media pesada muestral.

El **tamaño de muestra efectivo**  $n_{eff}$  es una herramienta para calcular de manera eficiente la media pesada muestral, basada en el hecho de que no todas las muestras dan la misma información.



## Media pesada muestral II

De manera formal, se define  $n_{eff}$  del siguiente modo[3]:

$$n_{eff} = \frac{(\sum_{i=1}^n \omega_i)^2}{\sum_{i=1}^n \omega_i^2}. \quad (5)$$

Intuitivamente, el mejor caso es cuando todos los pesos son iguales ( $\omega_i = \omega$ ), donde

$$n_{eff}^{best} = \frac{(n\omega)^2}{n\omega^2} = n,$$

y el peor caso es cuando todo el peso está concentrado en una única muestra, ( $\omega_j = \omega$ , para algún  $j$  y  $\omega_i = 0$  en otro caso):

$$n_{eff}^{worst} = \frac{\omega^2}{\omega^2} = 1.$$

# Distribución propuesta I

Se debe procurar entonces que los pesos tiendan a ser una constante. En teoría, si se conoce la distribución posterior lo suficientemente bien, para  $n$  lo suficientemente grande, se podría ajustar  $\Delta\theta_i$  para que los pesos  $\omega_i = \tilde{P}(\overline{\theta_i})\Delta\theta_i$  sean uniformes a cierto nivel de precisión. Esta uniformidad ocurre cuando

$$\Delta\theta_i \propto \frac{1}{\tilde{P}(\overline{\theta_i})}, \text{ para todo } i.$$

## Distribución propuesta II

Cuando  $n \rightarrow \infty$ , el espaciamiento  $\Delta\theta$  cambia como función de  $\theta$ . Esto motiva la definición de la densidad de puntos  $Q(\theta)$ , conocida como **distribución propuesta**, basada en la resolución variable  $\Delta\theta(\theta)$  en la rejilla infinita como función de  $\theta$ :

$$Q(\theta) \propto \frac{1}{\Delta\theta(\theta)}.$$

Usando  $Q(\theta)$ , se puede reescribir el valor esperado como

$$E_P[f(\theta)] = \frac{\int_{\theta \in \mathbb{S}_\theta} f(\theta) \tilde{P}(\theta) d\theta}{\int_{\theta \in \mathbb{S}_\theta} \tilde{P}(\theta) d\theta} = \frac{\int_{\theta \in \mathbb{S}_\theta} f(\theta) \frac{\tilde{P}(\theta)}{Q(\theta)} Q(\theta) d\theta}{\int_{\theta \in \mathbb{S}_\theta} \frac{\tilde{P}(\theta)}{Q(\theta)} Q(\theta) d\theta}$$

## Distribución propuesta III

$$E_P[f(\theta)] = \frac{E_Q[f(\theta)\tilde{P}(\theta)/Q(\theta)]}{E_Q[\tilde{P}(\theta)/Q(\theta)]}.$$

En palabras, la rejilla de  $n$  elementos en el límite de infinita resolución se manifiesta en una nueva distribución  $Q(\theta)$ , con la cual se puede escribir el valor esperado  $E_P[f(\theta)]$  en términos de los valores esperados  $E_Q[f(\theta)\tilde{P}(\theta)/Q(\theta)]$  y  $E_Q[\tilde{P}(\theta)/Q(\theta)]$ . Estos valores esperados generando una muestra aleatoria de  $n$  elementos a partir de  $Q(\theta)$ .

## Distribución propuesta IV

Esto motiva a escoger  $Q(\theta)$  de manera que se puedan generar muestras de manera fácil y directa. Generando  $n$  muestras  $\{\theta_1, \dots, \theta_n\}$  de esta distribución, con pesos asociados  $q_i$  y definiendo

$$f(\theta_i) = f_i, \quad \tilde{\omega}_i := \tilde{\omega}(\theta_i) = \tilde{P}(\theta_i)/Q(\theta_i),$$

el valor esperado se puede aproximar como

$$E_P[f(\theta)] = \frac{E_Q[f(\theta)\tilde{P}(\theta)/Q(\theta)]}{E_Q[\tilde{P}(\theta)/Q(\theta)]} \approx \frac{\sum_{i=1}^n f_i \tilde{\omega}_i q_i}{\sum_{i=1}^n \tilde{\omega}_i q_i}.$$

Si además se toma  $Q(\theta)$  de modo que las muestras sean *iid*, los correspondientes pesos  $q_i$  se reducen a  $1/n$ , de manera que

## Distribución propuesta V

$$E_P[f(\theta)] \approx \frac{n^{-1} \sum_{i=1}^n f_i \tilde{\omega}_i}{n^{-1} \sum_{i=1}^n \tilde{\omega}_i}.$$

El denominador de la última expresión es nuevamente una aproximación directa de la evidencia,

$$Z = \int_{\theta \in \mathbb{S}_\theta} \tilde{P}(\theta) d\theta \approx n^{-1} \sum_{i=1}^n \tilde{\omega}_i.$$

# Cadenas de Markov I

Los pasos a seguir para calcular el valor esperado son los siguientes.

- 1 Se debe generar  $n$  muestras *iid*  $\{\theta_1, \dots, \theta_n\}$  a partir de  $Q(\theta)$ .
- 2 Se calculan sus correspondientes pesos  $\tilde{w}_i = \tilde{P}(\theta_i)/Q(\theta_i)$ .
- 3 Se estima  $E_P[f(\theta)]$  aproximando  $E_Q[f(\theta)\tilde{P}(\theta)/Q(\theta)]$  y  $E_Q[\tilde{P}(\theta)/Q(\theta)]$  a través de los pesos de las muestras.

Los métodos *MCMC* buscan generar muestras de tal modo que los pesos asociados  $\{\tilde{w}_1, \dots, \tilde{w}_n\}$  sean constantes.  $Q(\theta)$  juega un papel fundamental para lograr este cometido, y para ilustrar, considere los siguientes casos.

## Cadenas de Markov II

- Tome  $Q(\theta) = Q^{unif}(\theta)$ , definida sobre un cuboide de volumen  $V$ , de la siguiente manera

$$Q^{unif}(\theta) = \begin{cases} 1/V, & \text{si } \theta \text{ está dentro del cuboide o} \\ 0 & \text{en otro caso.} \end{cases}$$

Los pesos en este caso serán proporcionales a la distribución posterior:

$$\tilde{\omega}_i^{unif} = \frac{\tilde{P}(\theta_i)}{Q^{unif}(\theta_i)} = V\tilde{P}(\theta_i) \propto P(\theta_i).$$



## Cadenas de Markov III

- Tome  $Q(\theta) = Q^{prior}(\theta) = \pi(\theta)$  como la distribución previa de  $\theta$ . Los pesos en este caso se pueden calcular mediante la función de verosimilitud.

$$\tilde{\omega}_i^{prior} = \frac{\tilde{P}(\theta_i)}{Q^{prior}(\theta_i)} = \frac{ZP(\theta_i)}{\pi(\theta_i)} = \frac{L(\theta_i)\pi(\theta_i)}{\pi(\theta_i)} = L(\theta_i).$$

- Tome  $Q(\theta) = Q^{post}(\theta) = P(\theta)$  como la distribución posterior de  $\theta$ , de modo que los pesos serán constantes e iguales a la evidencia  $Z$ :

$$\tilde{\omega}_i^{post} = \frac{\tilde{P}(\theta_i)}{Q^{post}(\theta_i)} = \frac{ZP(\theta_i)}{P(\theta_i)} = Z.$$

## Cadenas de Markov IV

Siguiendo la idea del último caso, si uno desea que sus pesos sean constantes, se debe procurar que  $Q(\theta)$  sea lo más cercana posible a  $P(\theta)$ . Los modelos *MCMC* buscan generar muestras con pesos proporcionales a la distribución posterior, para obtener un estimado óptimo del valor esperado.

Los modelos *MCMC* logran esto creando una cadena de valores de parámetros correlacionados  $\{\theta_1 \rightarrow \dots \rightarrow \theta_n\}$  al cabo de  $n$  iteraciones de tal modo que el número  $m(\theta)$  de iteraciones hechas en cada región particular  $\delta_\theta$ , centrada en  $\theta$  es proporcional a la densidad posterior  $P(\theta)$ . En otras palabras, la densidad de muestras generadas por el modelo *MCMC*

$$\rho(\theta) := \frac{m(\theta)}{n}$$

## Cadenas de Markov V

en la posición  $\theta$  integrada sobre  $\delta_\theta$  es aproximadamente

$$\int_{\theta \in \delta_\theta} P(\theta) d\theta \approx \int_{\theta \in \delta_\theta} \rho(\theta) d\theta \approx n^{-1} \sum_{j=1}^n \mathbb{I}[\theta_j \in \delta_\theta],$$

donde  $\mathbb{I}[\cdot]$  es la función indicadora, equivalente a 1 si la condición a la que está siendo evaluada es verdadera, y cero si es falsa.

Cuando  $n \rightarrow \infty$ , se garantiza que  $\rho(\theta) \rightarrow P(\theta)$  en cualquier punto  $\theta$  [2]. Con una aproximación razonable de  $\rho(\theta)$ , se pueden usar las muestras  $\{\theta_1 \rightarrow \dots \rightarrow \theta_n\}$  generadas por  $\rho(\theta)$  para estimar la evidencia

$$Z = \int_{\theta \in \mathbb{S}_\theta} \frac{\tilde{P}(\theta)}{\rho(\theta)} \rho(\theta) d\theta = E_\rho[\tilde{P}(\theta)/\rho(\theta)] \approx n^{-1} \sum_{i=1}^n \frac{\tilde{P}(\theta_i)}{\rho(\theta_i)}.$$

## Cadenas de Markov VI

Además, como el modelo *MCMC* produce una serie de  $n$  muestras de la distribución posterior, el valor esperado de  $f(\theta)$  se reduce a

$$E_P[f(\theta)] \approx \frac{n^{-1} \sum_{i=1}^n f_i \tilde{\omega}_i}{n^{-1} \sum_{i=1}^n \tilde{\omega}_i} = \frac{n^{-1} \sum_{i=1}^n f_i}{n^{-1} \sum_{i=1}^n 1} = n^{-1} \sum_{i=1}^n f_i,$$

que es la expresión del promedio aritmético de los valores  $f_i = f(\theta_i)$ .

# Metropolis-Hastings I

Se desea generar muestras  $\theta_i \rightarrow \theta_{i+1}$  de modo que la distribución de las muestras finales  $\rho(\theta)$  sea estacionaria cuando  $n \rightarrow \infty$  (que converja) y sea igual a  $P(\theta)$ . La primera condición se puede satisfacer usando el **balance detallado**, que refiere a la idea de que la probabilidad sea conservada cuando uno se mueve de una posición a otra (es decir, el proceso es reversible). Formalmente, esto implica que

$$M(\theta_{i+1}|\theta_i)M(\theta_i) = M(\theta_{i+1}, \theta_i) = M(\theta_i|\theta_{i+1})M(\theta_{i+1}),$$

donde  $M(\theta_{i+1}|\theta_i)$  es la probabilidad de moverse de  $\theta_i$  a  $\theta_{i+1}$  y  $M(\theta_i|\theta_{i+1})$  es la probabilidad de moverse de  $\theta_{i+1}$  a  $\theta_i$ . Reescribiendo esta última igualdad:

## Metropolis-Hastings II

$$\frac{M(\theta_{i+1}|\theta_i)}{M(\theta_i|\theta_{i+1})} = \frac{M(\theta_{i+1})}{M(\theta_i)} = \frac{P(\theta_{i+1})}{P(\theta_i)}. \quad (6)$$

Se propone una nueva posición  $\theta_i \rightarrow \theta'_{i+1}$  usando la distribución propuesta  $Q(\theta'_{i+1}|\theta_i)$ . Se decide si aceptar ( $\theta_{i+1} = \theta'_{i+1}$ ) o rechazar ( $\theta_{i+1} = \theta_i$ ) esta nueva posición con una **probabilidad de transición**  $T(\theta'_{i+1}|\theta_i)$ . Combinando ambas distribuciones, se obtiene la probabilidad de moverse a una nueva posición

$$M(\theta_{i+1}|\theta_i) = Q(\theta_{i+1}|\theta_i) T(\theta_{i+1}|\theta_i).$$

## Metropolis-Hastings III

Para encontrar  $T$ , se hace uso de la condición de balance detallado (6).

$$\frac{T(\theta_{i+1}|\theta_i)}{T(\theta_i|\theta_{i+1})} = \frac{P(\theta_{i+1})}{P(\theta_i)} \frac{Q(\theta_i|\theta_{i+1})}{Q(\theta_{i+1}|\theta_i)}.$$

El criterio de Metropolis [4]:

$$T(\theta_{i+1}|\theta_i) = \min \left[ 1, \frac{P(\theta_{i+1})}{P(\theta_i)} \frac{Q(\theta_i|\theta_{i+1})}{Q(\theta_{i+1}|\theta_i)} \right]$$

satisface esta condición.

# Algoritmo de Metropolis-Hastings

El algoritmo para generar la muestra es el siguiente:

- 1 Se propone una nueva posición  $\theta_i \rightarrow \theta'_{i+1}$ , generando una muestra de la distribución propuesta  $Q(\theta'_{i+1}|\theta_i)$ .
- 2 Se calcula la probabilidad de transición  $T(\theta'_{i+1}|\theta_i)$ .
- 3 Se genera un número aleatorio  $u_{i+1}$  a partir de una distribución uniforme entre 0 y 1.
- 4 Si  $u_{i+1} \leq T(\theta'_{i+1}|\theta_i)$ , se acepta el movimiento y se toma  $\theta_{i+1} = \theta'_i$ . Si  $u_{i+1} > T(\theta'_{i+1}|\theta_i)$ , se rechaza el movimiento y se toma  $\theta_{i+1} = \theta_i$ .
- 5 Se incrementa  $i = i + 1$  y se repite el proceso.



## Referencias I



L. Blanco, V. Arunachalam, and S. Dharmaraja.

*Introduction to Probability and Stochastic Processes with Applications.*

Wiley, 2012.



Steve Brooks, Andrew Gelman, and Galin L. Jones.

*Handbook of Markov chain Monte Carlo.*

CRC Press, 2011.



Leslie Kish.

*Survey sampling.*

Wiley, 1995.

## Referencias II



Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller.  
Equation of state calculations by fast computing machines.  
*The Journal of Chemical Physics*, 21(6):1087–1092, 1953.



R. Rossi.  
*Mathematical statistics: an introduction to likelihood based inference*.  
John Wiley Sons, Inc., 2018.



Joshua S. Speagle.  
A conceptual introduction to markov chain monte carlo methods, 2019.