

# Fundamentos de Estadística Bayesiana

Juan Carlos Martínez-Ovando

ITAM

[Mini] Taller de Métodos Numéricos y Estadísticos en Cosmología 2017  
Cinvestav, CDMX  
6 de abril de 2017

# 1. Incertidumbre y aleatoriedad

# 1.1. Consideraciones

Incertidumbre  $\leftrightarrow$  Desconocimiento

## Datos

El proceso que genera los **datos**

$$\{x_1, \dots, x_n\},$$

es desconocido.

## Aleatoriedad intrínseca

Nuestro desconocimiento acerca de los **datos** lo manifestamos suponiendo **aleatoriedad intrínseca** empleando un modelo

$$\mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n).$$

# 1.1. Consideraciones

Incertidumbre  $\leftrightarrow$  Desconocimiento

## Datos

El proceso que genera los **datos**

$$\{x_1, \dots, x_n\},$$

es desconocido.

## Aleatoriedad intrínseca

Nuestro desconocimiento acerca de los **datos** lo manifestamos suponiendo **aleatoriedad intrínseca** empleando un modelo

$$\mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n).$$

# 1.1. Consideraciones

Incertidumbre  $\leftrightarrow$  Desconocimiento

## Datos

El proceso que genera los **datos**

$$\{x_1, \dots, x_n\},$$

es desconocido.

## Aleatoriedad intrínseca

Nuestro desconocimiento acerca de los **datos** lo manifestamos suponiendo **aleatoriedad intrínseca** empleando un modelo

$$\mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n).$$

## 1.2. Dependencia

Es un supuesto atribuible al modelo  $\mathbb{P}$  y **no** a los **datos**.

### Independencia

$$\mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n) = \prod_{j=1}^n \mathbb{P}(X_j \leq x_j).$$

### Intercambiabilidad

$$\mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n) = \mathbb{P}(X_{\sigma(1)} \leq x_1, \dots, X_{\sigma(n)} \leq x_n),$$

para toda permutación

$$\{\sigma(1), \dots, \sigma(n)\} \text{ de } \{1, \dots, n\}.$$

## 1.2. Dependencia

Es un supuesto atribuible al modelo  $\mathbb{P}$  y **no** a los **datos**.

### Independencia

$$\mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n) = \prod_{j=1}^n \mathbb{P}(X_j \leq x_j).$$

### Intercambiabilidad

$$\mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n) = \mathbb{P}(X_{\sigma(1)} \leq x_1, \dots, X_{\sigma(n)} \leq x_n),$$

para toda permutación

$$\{\sigma(1), \dots, \sigma(n)\} \text{ de } \{1, \dots, n\}.$$

## 1.2. Dependencia

Es un supuesto atribuible al modelo  $\mathbb{P}$  y **no** a los **datos**.

### Independencia

$$\mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n) = \prod_{j=1}^n \mathbb{P}(X_j \leq x_j).$$

### Intercambiabilidad

$$\mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n) = \mathbb{P}(X_{\sigma(1)} \leq x_1, \dots, X_{\sigma(n)} \leq x_n),$$

para toda permutación

$$\{\sigma(1), \dots, \sigma(n)\} \text{ de } \{1, \dots, n\}.$$



## 2. Subjetividad y modelación

## 2.1. Subjetividad

Intercambiabilidad *supone* que el orden en el cual los datos son recolectados/observados es indiferente.

### Teorema de representación

Atribuible a Bruno de Finetti, bajo intercambiabilidad

$$\mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n) = \int_{\Theta} \prod_{i=1}^n F_X(x_i|\theta) \Pi(d\theta)$$

- ▶  $F_X(x_i|\theta)$  es una distribución de probabilidades
- ▶  $\theta$  es un objeto estocástico **no observable** común a todos los datos (a.k.a. parámetro)
- ▶  $\Pi(\theta)$  es una medida de probabilidad

## 2.1. Subjetividad

Intercambiabilidad *supone* que el orden en el cual los datos son recolectados/observados es indiferente.

### Teorema de representación

Atribuible a Bruno de Finetti, bajo intercambiabilidad

$$\mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n) = \int_{\Theta} \prod_{i=1}^n F_X(x_i|\theta) \Pi(d\theta)$$

- ▶  $F_X(x_i|\theta)$  es una distribución de probabilidades
- ▶  $\theta$  es un objeto estocástico **no observable** común a todos los datos (a.k.a. parámetro)
- ▶  $\Pi(\theta)$  es una medida de probabilidad

## 2.2. Juicio inicial

### Teorema de representación

El Teorema de Representación es de **existencia** más **no de unicidad**

### Probabilidades subjetivas

- ▶ La aleatoriedad intrínseca no se interpreta como un límite de frecuencias
- ▶ Sino, aleatoriedad intrínseca es un juicio individual subjetivo (en función de nuestro conocimiento/desconocimiento)

### Distribución inicial

- ▶ La distribución  $\Pi(\theta)$  se interpreta como el juicio o creencia individual acerca del **modelo** antes de observar los **datos**.
- ▶ Aunque  $\Pi(\theta)$  esta definida sobre  $\Theta$ , en realidad representa una medida de probabilidad sobre la colección de modelos  $F(x|\theta)$ .

## 2.2. Juicio inicial

### Teorema de representación

El Teorema de Representación es de **existencia** más **no de unicidad**

### Probabilidades subjetivas

- ▶ La **aleatoriedad intrínseca** no se interpreta como un **límite de frecuencias**
- ▶ Sino, **aleatoriedad intrínseca** es un **juicio individual subjetivo** (en función de nuestro conocimiento/desconocimiento)

### Distribución inicial

- ▶ La distribución  $\Pi(\theta)$  se interpreta como el juicio o creencia individual acerca del **modelo** antes de observar los **datos**.
- ▶ Aunque  $\Pi(\theta)$  esta definida sobre  $\Theta$ , en realidad representa una medida de probabilidad sobre la colección de modelos  $F(x|\theta)$ .

## 2.2. Juicio inicial

### Teorema de representación

El Teorema de Representación es de **existencia** más **no de unicidad**

### Probabilidades subjetivas

- ▶ La **aleatoriedad intrínseca** no se interpreta como un **límite de frecuencias**
- ▶ Sino, **aleatoriedad intrínseca** es un **juicio individual subjetivo** (en función de nuestro conocimiento/desconocimiento)

### Distribución inicial

- ▶ La distribución  $\Pi(\theta)$  se interpreta como el juicio o creencia individual acerca del **modelo** antes de observar los **datos**.
- ▶ Aunque  $\Pi(\theta)$  esta definida sobre  $\Theta$ , en realidad representa una medida de probabilidad sobre la colección de modelos  $F(x|\theta)$ .

## 2.3. Ejemplo: Bernoulli-beta

### Variables observables

Supongamos que los **datos** son tales que

$$x_i = \begin{cases} 1 & \text{con base en la ocurrencia de un evento} \\ 0 & \text{no ocurrencia del evento} \end{cases}$$

### Desconocimiento/aleatoriedad

Desconocer el **mecanismo generador** de  $x_i$ s induce **aleatoriedad intrínseca**

$$f(x_i = j) = \begin{cases} \theta & \text{si } j = 1 \\ (1 - \theta) & \text{si } j = 0 \end{cases}$$

con  $0 \leq \theta \leq 1$ .

### Distribución Bernoulli

Lo anterior es **equivalente** a suponer que los datos son **generados** con una distribución Bernoulli, i.e.

$$x|\theta \sim F(x|\theta) = \text{Ber}(x_i|\theta).$$

## 2.3. Ejemplo: Bernoulli-beta

### Variables observables

Supongamos que los **datos** son tales que

$$x_i = \begin{cases} 1 & \text{con base en la ocurrencia de un evento} \\ 0 & \text{no ocurrencia del evento} \end{cases}$$

### Desconocimiento/aleatoriedad

Desconocer el **mecanismo generador** de  $x_i$ s induce **aleatoriedad intrínseca**

$$f(x_i = j) = \begin{cases} \theta & \text{si } j = 1 \\ (1 - \theta) & \text{si } j = 0 \end{cases}$$

con  $0 \leq \theta \leq 1$ .

### Distribución Bernoulli

Lo anterior es **equivalente** a suponer que los datos son **generados** con una distribución Bernoulli, i.e.

$$x|\theta \sim F(x|\theta) = \text{Ber}(x_i|\theta).$$



## 2.3. Ejemplo: Bernoulli-beta

### Variables observables

Supongamos que los **datos** son tales que

$$x_i = \begin{cases} 1 & \text{con base en la ocurrencia de un evento} \\ 0 & \text{no ocurrencia del evento} \end{cases}$$

### Desconocimiento/aleatoriedad

Desconocer el **mecanismo generador** de  $x_i$ s induce **aleatoriedad intrínseca**

$$f(x_i = j) = \begin{cases} \theta & \text{si } j = 1 \\ (1 - \theta) & \text{si } j = 0 \end{cases}$$

con  $0 \leq \theta \leq 1$ .

### Distribución Bernoulli

Lo anterior es **equivalente** a suponer que los datos son **generados** con una distribución Bernoulli, i.e.

$$x|\theta \sim F(x|\theta) = \text{Ber}(x_i|\theta).$$

## 2.3. Ejemplo: Bernoulli-beta

### Configuración

Cada valor específico de  $\theta$  en  $(0, 1)$  define un modelo de probabilidad particular, por ejemplo:

M1  $Ber(x|0,1)$

M2  $Ber(x|0,33)$

M3  $Ber(x|0,999)$

M4  $Ber(x|0,001)$

M5 etc.

### Juicio inicial

Definido como la **creencia** acerca de la **plausibilidad** de cada configuración para describir los **datos** que estan por observarse.

**Posibles alternativas:**

- ▶  $\Pi_1(\theta) = \sum_{k=1}^K \alpha_k \delta_{\theta_k}(\theta)$  para un conjunto de  $\theta_k$ s en  $(0, 1)$
- ▶  $\Pi_2(\theta) = U(\theta|0, 1)$
- ▶  $\Pi_3(\theta) = Be(\theta|a, b)$
- ▶ entre muchas otras,

## 2.3. Ejemplo: Bernoulli-beta

### Configuración

Cada valor específico de  $\theta$  en  $(0, 1)$  define un modelo de probabilidad particular, por ejemplo:

M1  $Ber(x|0,1)$

M2  $Ber(x|0,33)$

M3  $Ber(x|0,999)$

M4  $Ber(x|0,001)$

M5 etc.

### Juicio inicial

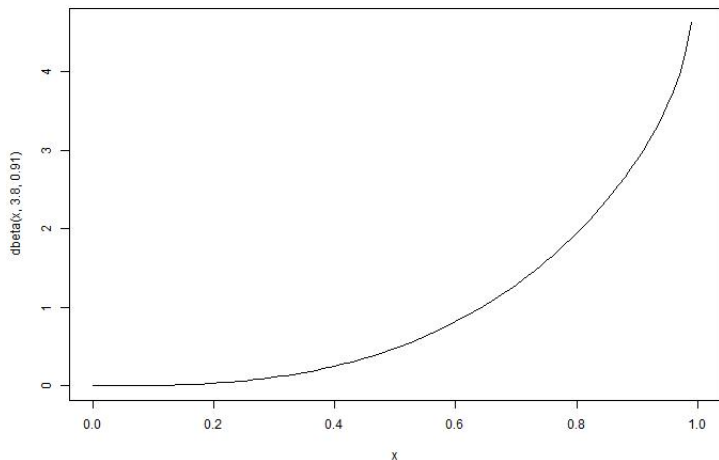
Definido como la **creencia** acerca de la **plausibilidad** de cada configuración para describir los **datos** que estan por observarse.

**Posibles alternativas:**

- ▶  $\Pi_1(\theta) = \sum_{k=1}^K \alpha_k \delta_{\theta_k}(\theta)$  para un conjunto de  $\theta_k$ s en  $(0, 1)$
- ▶  $\Pi_2(\theta) = U(\theta|0, 1)$
- ▶  $\Pi_3(\theta) = Be(\theta|a, b)$
- ▶ entre muchas otras,

## 2.3. Ejemplo: Bernoulli-beta

Figura: Una posible distribución  $\Pi(\theta)$



## 2.3. Ejemplo: Bernoulli-beta

### Reflexión

Recordemos:

- ▶ Los **datos**  $x_i$ s son **observables**
- ▶ La variable (**parámetro**)  $\theta$  **no es observable**.
- ▶  $\Pi(\theta)$  **no es única**, sino **subjetiva**

### Modelación

Así, el modelo subjetivo es una medida de probabilidad conjunta de **observables** y **no observables**,

$$\begin{aligned}\mathbb{P}(x_1, \dots, x_n, \theta) &= F(x_1, \dots, x_n | \theta) \times \Pi(\theta) \\ &= \prod_{i=1}^n F(x_i | \theta) \times \Pi(\theta)\end{aligned}$$

## 2.3. Ejemplo: Bernoulli-beta

### Reflexión

Recordemos:

- ▶ Los **datos**  $x_i$ s son **observables**
- ▶ La variable (**parámetro**)  $\theta$  **no es observable**.
- ▶  $\Pi(\theta)$  **no es única**, sino **subjetiva**

### Modelación

Así, el modelo subjetivo es una medida de probabilidad conjunta de **observables** y **no observables**,

$$\begin{aligned}\mathbb{P}(x_1, \dots, x_n, \theta) &= F(x_1, \dots, x_n | \theta) \times \Pi(\theta) \\ &= \prod_{i=1}^n F(x_i | \theta) \times \Pi(\theta)\end{aligned}$$

### 3. Aprendizaje y predicción

## 3.1. Aprendizaje

### Aprendizaje bayesiano/inferencia

*El aprendizaje bayesiano, con base en un conjunto de datos,  $x_i$ s consiste en la actualización de  $\Pi(\theta)$  condicional en los datos, i.e.*

$$\Pi(\theta|x_1, \dots, x_n) = \frac{F(x_1, \dots, x_n|\theta)\Pi(\theta)}{\mathbb{P}(x_1, \dots, x_n)}$$

donde

$$\mathbb{P}(x_1, \dots, x_n) = \int_{\Theta} F(x_1, \dots, x_n|\theta)\Pi(d\theta)$$

Las ecuaciones anteriores se refieren al **Teorema de Bayes** de probabilidades inversas.

### Comentario

- ▶ El denominador en  $\Pi(\theta|x_1, \dots, x_n)$ , i.e.  $\mathbb{P}(x_1, \dots, x_n)$  se conoce como la constante de normalización de la distribución actualizada para  $\theta$
- ▶ Generalmente, la constante de normalización no puede obtenerse de manera analítica cerrada, pero puede aproximarse empleando **métodos numéricos**.



## 3.1. Aprendizaje

### Aprendizaje bayesiano/inferencia

*El aprendizaje bayesiano, con base en un conjunto de datos,  $x_i$ s consiste en la actualización de  $\Pi(\theta)$  condicional en los datos, i.e.*

$$\Pi(\theta|x_1, \dots, x_n) = \frac{F(x_1, \dots, x_n|\theta)\Pi(\theta)}{\mathbb{P}(x_1, \dots, x_n)}$$

donde

$$\mathbb{P}(x_1, \dots, x_n) = \int_{\Theta} F(x_1, \dots, x_n|\theta)\Pi(d\theta)$$

Las ecuaciones anteriores se refieren al **Teorema de Bayes** de probabilidades inversas.

### Comentario

- ▶ El denominador en  $\Pi(\theta|x_1, \dots, x_n)$ , i.e.  $\mathbb{P}(x_1, \dots, x_n)$  se conoce como la constante de normalización de la distribución actualizada para  $\theta$
- ▶ Generalmente, la constante de normalización no puede obtenerse de manera analítica cerrada, pero puede aproximarse empleando **métodos numéricos**.

## 3.2. Ejemplo: Bernoulli-beta

### Especificación

Continuemos con el modelo Bernoulli-beta del ejemplo 2.3.

$$\begin{aligned} f(x_1, \dots, x_n | \theta) &= \prod_{i=1}^n \text{Ber}(x_i | \theta) \\ &\propto \theta^{\#\{x_i : x_i = 1\}} (1 - \theta)^{\#\{x_i : x_i = 0\}} \end{aligned}$$

y

$$\begin{aligned} \pi(\theta) &= \text{Be}(\theta | a, b) \\ &\propto \theta^{a-1} (1 - \theta)^{b-1} \mathbb{I}_{(0,1)}(\theta). \end{aligned}$$

### Aprendizaje bayesiano

Así, la distribución actualizada para  $\theta$  es,

$$\pi(\theta | x_1, \dots, x_n) \propto \theta^{\#\{x_i : x_i = 1\} + a - 1} (1 - \theta)^{\#\{x_i : x_i = 0\} + b - 1} \mathbb{I}_{(0,1)}(\theta).$$

## 3.2. Ejemplo: Bernoulli-beta

### Especificación

Continuemos con el modelo Bernoulli-beta del ejemplo 2.3.

$$\begin{aligned} f(x_1, \dots, x_n | \theta) &= \prod_{i=1}^n \text{Ber}(x_i | \theta) \\ &\propto \theta^{\#\{x_i : x_i=1\}} (1 - \theta)^{\#\{x_i : x_i=0\}} \end{aligned}$$

y

$$\begin{aligned} \pi(\theta) &= \text{Be}(\theta | a, b) \\ &\propto \theta^{a-1} (1 - \theta)^{b-1} \mathbb{I}_{(0,1)}(\theta). \end{aligned}$$

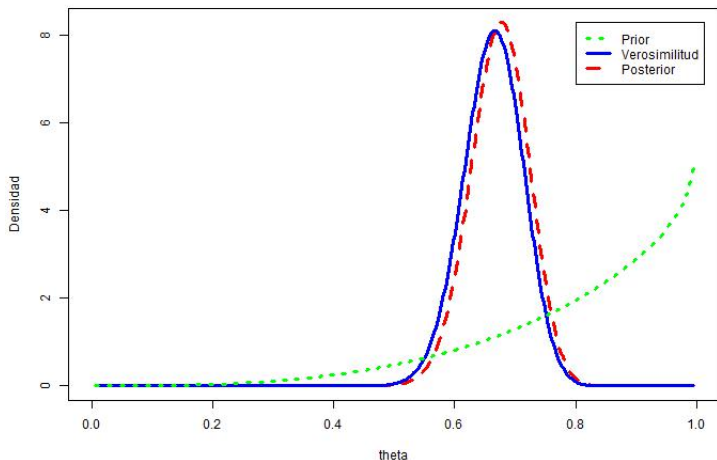
### Aprendizaje bayesiano

Así, la distribución actualizada para  $\theta$  es,

$$\pi(\theta | x_1, \dots, x_n) \propto \theta^{\#\{x_i : x_i=1\} + a - 1} (1 - \theta)^{\#\{x_i : x_i=0\} + b - 1} \mathbb{I}_{(0,1)}(\theta).$$

## 3.2. Ejemplo: Bernoulli-beta

Figura: Distribución  $\Pi(\theta)$  actualizada con base en los **datos**



## 3.3. Predicción

### Reflexión

*Predicción es el objetivo último cuando se define un modelo estocástico.*

**Se anticipan posibles resultados de eventos observables con base en información pasada..**

### Distribución predictiva

La **previsión** de posibles esenarios no observados aun se produce con

$$\mathbb{P}(x^f | x_1, \dots, x_n) = \int_{\Theta} \underbrace{F(x^f | \theta)}_{\text{Intrínseca}} \underbrace{\Pi(d\theta | x_1, \dots, x_n)}_{\text{Epistémica}}.$$

Cuando la constante de normalización de  $\Pi(\theta | x_1, \dots, x_n)$  no se puede calcular de manera analítica cerrada, la distribución predictiva puede aproximarse mediante el método de Monte Carlos, donde

$$\hat{\mathbb{P}}(x^f | x_1, \dots, x_n) = \frac{1}{M} \sum_{m=1}^M F(x^f | \theta^{(m)}).$$

## 3.3. Predicción

### Reflexión

*Predicción es el objetivo último cuando se define un modelo estocástico.*

**Se anticipan posibles resultados de eventos observables con base en información pasada..**

### Distribución predictiva

La **previsión** de posibles escenarios no observados aun se produce con

$$\mathbb{P}(x^f | x_1, \dots, x_n) = \int_{\Theta} \underbrace{F(x^f | \theta)}_{\text{Intrínseca}} \underbrace{\Pi(d\theta | x_1, \dots, x_n)}_{\text{Epistémica}}.$$

Cuando la constante de normalización de  $\Pi(\theta | x_1, \dots, x_n)$  no se puede calcular de manera analítica cerrada, la distribución predictiva puede aproximarse mediante el método de Monte Carlos, donde

$$\hat{\mathbb{P}}(x^f | x_1, \dots, x_n) = \frac{1}{M} \sum_{m=1}^M F(x^f | \theta^{(m)}).$$

## 3.4. Resultados importantes

### Aspectos importantes

*Conforme el conjunto de **datos** es más grande, la incertidumbre epistémica se reduce.*

*En el caso límite, la **incertidumbre epistémica** se desvanece, i.e.*

$$\lim_{n \rightarrow \infty} \pi(\theta | x_1, \dots, x_n) = \delta_{\{\theta^*\}}(\theta), \quad (1)$$

*donde*

*$\theta^*$  es el "verdadero" valor de  $\theta$ .*

### Conciliación de opiniones

*Se ha demostrado que aun en el caso donde dos individuos asignan distribuciones iniciales distintas, digamos  $\Pi_1$  y  $\Pi_2$ , el proceso de aprendizaje de ambos converge con el tamaño de los **datos**, i.e.*

$$\lim_{n \rightarrow \infty} \pi_1(\theta | x_1, \dots, x_n) = \lim_{n \rightarrow \infty} \pi_2(\theta | x_1, \dots, x_n) \quad (2)$$

*y ambos convergen al "verdadero" valor de  $\theta^*$ .*

## 3.4. Resultados importantes

### Aspectos importantes

*Conforme el conjunto de **datos** es más grande, la incertidumbre epistémica se reduce.*

*En el caso límite, la **incertidumbre epistémica** se desvanece, i.e.*

$$\lim_{n \rightarrow \infty} \pi(\theta | x_1, \dots, x_n) = \delta_{\{\theta^*\}}(\theta), \quad (1)$$

*donde*

*$\theta^*$  es el "verdadero" valor de  $\theta$ .*

### Conciliación de opiniones

*Se ha demostrado que aun en el caso donde dos individuos asignan distribuciones iniciales distintas, digamos  $\Pi_1$  y  $\Pi_2$ , el proceso de aprendizaje de ambos converge con el tamaño de los **datos**, i.e.*

$$\lim_{n \rightarrow \infty} \pi_1(\theta | x_1, \dots, x_n) = \lim_{n \rightarrow \infty} \pi_2(\theta | x_1, \dots, x_n) \quad (2)$$

*y ambos convergen al "verdadero" valor de  $\theta^*$ .*



## 3.5. Inferencia

### Inferencia bayesiana

Siendo que toda la información contenida en los datos se resume en

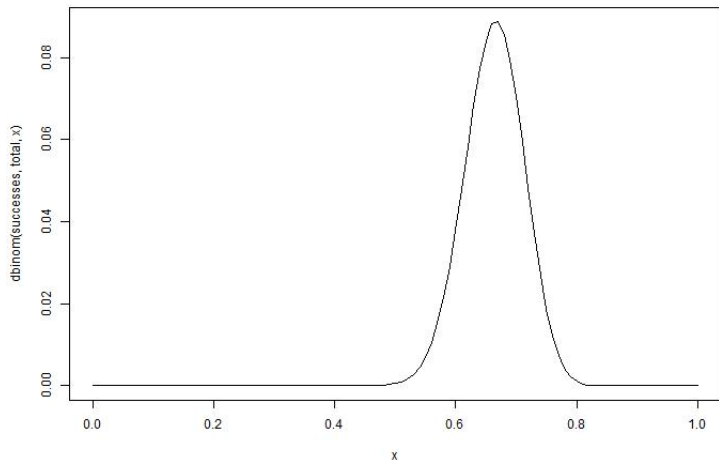
$$\Pi(\theta|x_1, \dots, x_n),$$

cualquier problema inferencial puede derivarse de esta distribución, e.g.

- ▶ **Estimación puntual.**- Media, moda o mediana de la distribución.
- ▶ **Estimación por intervalos.**- Regiones inter-cuantílicas de la distribución.
- ▶ **Pruebas de hipótesis.**- Con probabilidades cuantificadas respecto a la distribución.

## 3.5. Inferencia

Figura: Distribución final de  $\theta$  en el modelo Bernoulli-beta



## 4. Herramientas computacionales

## 4.1. Variables latentes

### Inferencia bayesiana

Siendo que toda la información contenida en los datos se resume en

$$\Pi(\theta|x_1, \dots, x_n),$$

cualquier problema inferencial puede derivarse de esta distribución, e.g.

- ▶ **Estimación puntual.**- Media, moda o mediana de la distribución.
- ▶ **Estimación por intervalos.**- Regiones inter-cuantílicas de la distribución.
- ▶ **Pruebas de hipótesis.**- Con probabilidades cuantificadas respecto a la distribución.

## 4.2. Aproximaciones numéricas

### Inferencia bayesiana

Siendo que toda la información contenida en los datos se resume en

$$\Pi(\theta|x_1, \dots, x_n),$$

cualquier problema inferencial puede derivarse de esta distribución, e.g.

- ▶ **Estimación puntual.**- Media, moda o mediana de la distribución.
- ▶ **Estimación por intervalos.**- Regiones inter-cuantílicas de la distribución.
- ▶ **Pruebas de hipótesis.**- Con probabilidades cuantificadas respecto a la distribución.