Nanorepaso de probabilidad

probabilidad conjunta, condicional

$$p(\theta, y) = p(\theta|y)p(y)$$

Nanorepaso de probabilidad

probabilidad conjunta, condicional

$$p(\theta, y) = p(\theta|y)p(y)$$

simétrica

$$p(\theta, y) = p(y|\theta)p(\theta)$$

Nanorepaso de probabilidad

probabilidad conjunta, condicional

$$p(\theta, y) = p(\theta|y)p(y)$$

simétrica

$$p(\theta, y) = p(y|\theta)p(\theta)$$

marginalización

$$p(y) = \int p(y, \theta)d\theta$$

Nanorepaso de probabilidad

probabilidad conjunta, condicional

$$p(\theta, y) = p(\theta|y)p(y)$$

simétrica

$$p(\theta, y) = p(y|\theta)p(\theta)$$

marginalización

$$p(y) = \int p(y, \theta)d\theta$$

simétrica

$$p(\theta) = \int p(y, \theta)dy$$

Nanorepaso de probabilidad

probabilidad conjunta, condicional

$$p(\theta, y) = p(\theta|y)p(y)$$

simétrica

$$p(\theta, y) = p(y|\theta)p(\theta)$$

marginalización

$$p(y) = \int p(y, \theta)d\theta$$

simétrica

$$p(\theta) = \int p(y, \theta)dy$$

θ discreta

$$p(y) = \sum_{\theta} p(y, \theta)$$

Nanorepaso de probabilidad

probabilidad conjunta, condicional

$$p(\theta, y) = p(\theta|y)p(y)$$

simétrica

$$p(\theta, y) = p(y|\theta)p(\theta)$$

media, varianza

$$\mathrm{E}(y) = \int yp(y)dy$$

marginalización

$$p(y) = \int p(y, \theta)d\theta$$

simétrica

$$p(\theta) = \int p(y, \theta)dy$$

θ discreta

$$p(y) = \sum_{\theta} p(y, \theta)$$

Nanorepaso de probabilidad

probabilidad conjunta, condicional

$$p(\theta, y) = p(\theta|y)p(y)$$

simétrica

$$p(\theta, y) = p(y|\theta)p(\theta)$$

media, varianza

$$\mathrm{E}(y) = \int y p(y) dy$$

$$\mathrm{var}(y) = \int (y - \mathrm{E}(y))^2 p(y) dy$$

marginalización

$$p(y) = \int p(y, \theta) d\theta$$

simétrica

$$p(\theta) = \int p(y, \theta) dy$$

θ discreta

$$p(y) = \sum_{\theta} p(y, \theta)$$

Notación y nomenclatura

H hipótesis $\quad$ θ parámetros $\quad$ y, D datos

ocasionalmente vectores, y si hace falta, en **negrita**

Notación y nomenclatura

H hipótesis θ parámetros y, D datos

ocasionalmente vectores, y si hace falta, en **negrita**

Bayes

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\sum_\theta p(y|\theta)p(\theta)}$$

Notación y nomenclatura

H hipótesis θ parámetros y, D datos

ocasionalmente vectores, y si hace falta, en **negrita**

Bayes

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\sum_\theta p(y|\theta)p(\theta)}$$

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

Notación y nomenclatura

H hipótesis $\qquad \theta$ parámetros $\qquad y, D$ datos

ocasionalmente vectores, y si hace falta, en **negrita**

Bayes

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\sum_\theta p(y|\theta)p(\theta)}$$

$$p(\theta|y) \propto \underbrace{p(y|\theta)}_{\textit{likelihood}}p(\theta)$$

Notación y nomenclatura

H hipótesis θ parámetros y, D datos

ocasionalmente vectores, y si hace falta, en **negrita**

Bayes

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\sum_\theta p(y|\theta)p(\theta)}$$

$$p(\theta|y) \propto \underbrace{p(y|\theta)}_{\textit{likelihood}}\underbrace{p(\theta)}_{\textit{prior}}$$

Notación y nomenclatura

H hipótesis θ parámetros y, D datos

ocasionalmente vectores, y si hace falta, en **negrita**

Bayes

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\sum_\theta p(y|\theta)p(\theta)}$$

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

posterior *likelihood* *prior*

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

posterior　　　　*likelihood*　　　　*prior*

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

posterior *likelihood* *prior*

prior predictive

$$p(y) = \int p(y,\theta)d\theta = \int p(y|\theta)p(\theta)d\theta$$

$$\underbrace{p(\theta|y)}_{\textit{posterior}} \propto \underbrace{p(y|\theta)}_{\textit{likelihood}}\underbrace{p(\theta)}_{\textit{prior}}$$

prior predictive

$$p(y) = \int p(y, \theta) d\theta = \int p(y|\theta)p(\theta)d\theta$$

posterior predictive

$$\begin{aligned} p(\tilde{y}|y) &= \int p(\tilde{y}, \theta|y)d\theta \\ &= \int p(\tilde{y}|\theta, y)p(\theta|y)d\theta \\ &= \int p(\tilde{y}|\theta)p(\theta|y)d\theta \end{aligned}$$

$$p(\theta|y) \propto \underbrace{p(y|\theta)}_{\textit{likelihood}}\,\underbrace{p(\theta)}_{\textit{prior}}$$

posterior $\leftarrow p(\theta|y)$

prior predictive

$$p(y) = \int p(y,\theta)d\theta = \int p(y|\theta)p(\theta)d\theta$$

posterior predictive

$$p(\tilde{y}|y) = \int p(\tilde{y},\theta|y)d\theta$$

$$= \int p(\tilde{y}|\theta,y)p(\theta|y)d\theta$$

$$= \int p(\tilde{y}|\theta)p(\theta|y)d\theta$$

promedio con mi *posterior!*

¡Volvemos a la moneda!

$$D = 000000$$

$$D = 010010$$

¿Qué proceso produjo estas secuencias?

Inferencia Bayesiana

Inferencia Bayesiana

- Hipótesis H sobre los procesos que pueden haber generado los datos D

Inferencia Bayesiana

- Hipótesis *H* sobre los procesos que pueden haber generado los datos *D*

- Distribución de probabilidad sobre las hipótesis, dados los datos

Inferencia Bayesiana

- Hipótesis *H* sobre los procesos que pueden haber generado los datos *D*

- Distribución de probabilidad sobre las hipótesis, dados los datos

- *p(D|H)* probabilidad de que los datos *D* hayan sido generados por el proceso descrito por *H*

Inferencia Bayesiana

- Hipótesis *H* sobre los procesos que pueden haber generado los datos *D*
- Distribución de probabilidad sobre las hipótesis, dados los datos
- $p(D|H)$ probabilidad de que los datos *D* hayan sido generados por el proceso descrito por *H*
- Hipótesis mutuamente excluyentes: sólo un proceso generó *D*

Algunas hipótesis para la moneda

Procesos que pueden haber generado

$$D = 010010$$

Algunas hipótesis para la moneda

Procesos que pueden haber generado

$$D = 010010$$

- Moneda común: p(0) = 0.5
- Moneda cargada: p(0) = θ
- Modelo de *Markov*
- *Hidden Markov Model (HMM)*

Modelos probabilísticos gráficos (PGMs, Pearl)

Modelos probabilísticos gráficos (PGMs, Pearl)

- Los nodos son variables, las aristas representan dependencia

Modelos probabilísticos gráficos (PGMs, Pearl)

- Los nodos son variables, las aristas representan dependencia
- Aristas dirigidas representan influencia *causal*

Modelos probabilísticos gráficos (PGMs, Pearl)

- Los nodos son variables, las aristas representan dependencia
- Aristas dirigidas representan influencia *causal*
- Nodos *observables* y *latentes*

Modelos probabilísticos gráficos (PGMs, Pearl)

- Los nodos son variables, las aristas representan dependencia
- Aristas dirigidas representan influencia *causal*
- Nodos *observables* y *latentes*
- Notación de *placas*: variables repetidas

Modelos probabilísticos gráficos (PGMs, Pearl)

- Los nodos son variables, las aristas representan dependencia
- Aristas dirigidas representan influencia *causal*
- Nodos *observables* y *latentes*
- Notación de *placas*: variables repetidas
- Convenciones frecuentes:
  gris: observable, blanco: latente.
  circular: continuo, cuadrado: discreto.

Algunos Modelos Generativos

$$D = 010010$$

$$d_1\, d_2\, d_3\, d_4\, d_5\, d_6$$

Algunos Modelos Generativos

$$D = 010010$$

$$d_1 \, d_2 \, d_3 \, d_4 \, d_5 \, d_6$$

Moneda común

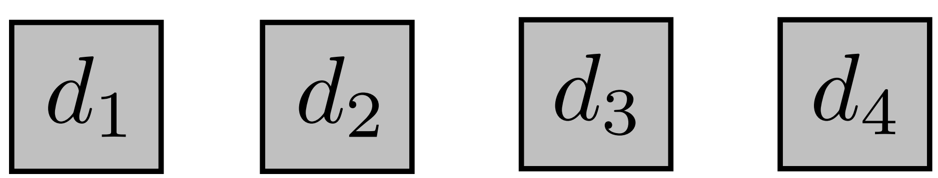
$$\boxed{d_1} \quad \boxed{d_2} \quad \boxed{d_3} \quad \boxed{d_4}$$

Algunos Modelos Generativos

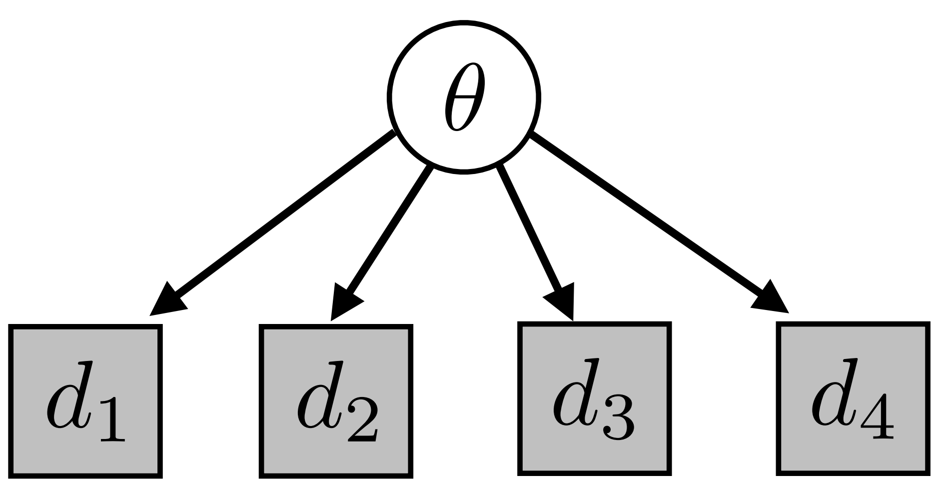
$$D = 010010$$

$$d_1\, d_2\, d_3\, d_4\, d_5\, d_6$$

Moneda común



Moneda cargada

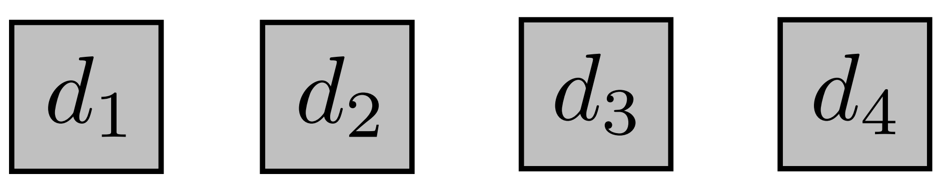


Algunos Modelos Generativos

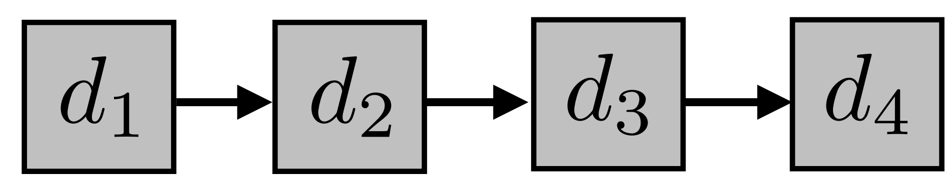
$$D = 010010$$

$d_1 \; d_2 \; d_3 \; d_4 \; d_5 \; d_6$

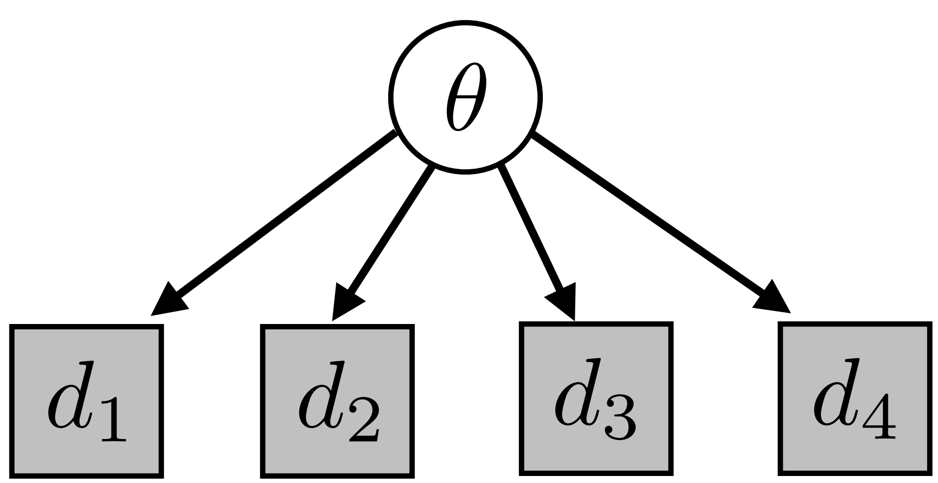
Moneda común



Modelo de *Markov*



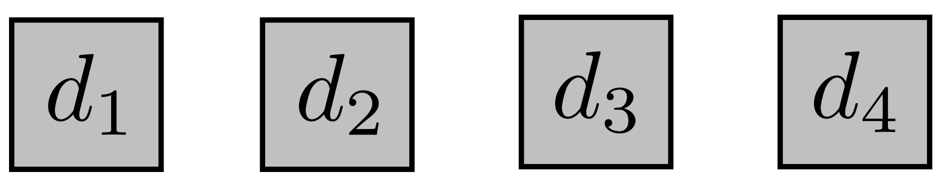
Moneda cargada



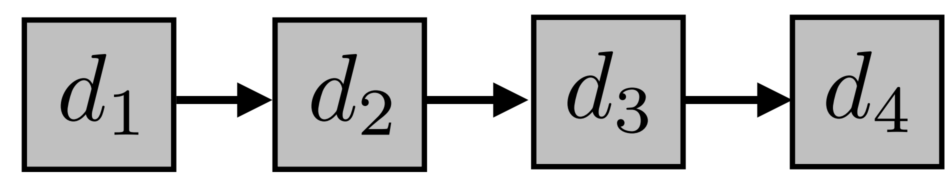
Algunos Modelos Generativos

$$D = 010010$$
$$d_1\, d_2\, d_3\, d_4\, d_5\, d_6$$

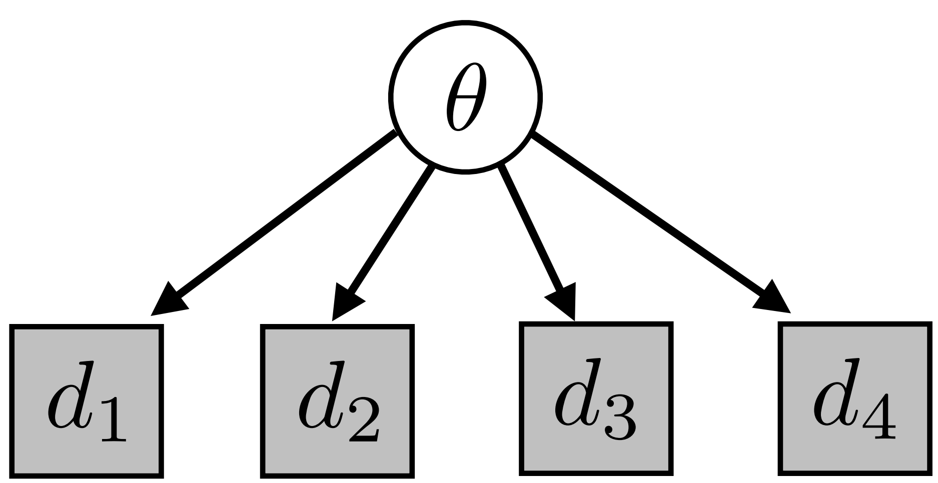
Moneda común



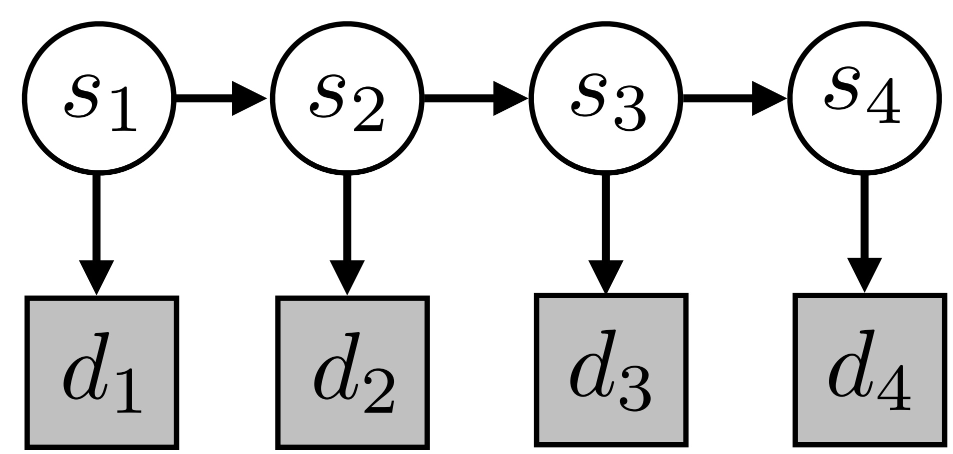
Modelo de *Markov*



Moneda cargada



Hidden Markov Model



Comparación de dos hipótesis simples

$$H_1\colon\ p(0) = 0.5 \quad \text{vs.} \quad H_2\colon\ p(0) = 1$$

Moneda común Moneda "dos caras"

Comparación de dos hipótesis simples

$$H_1\colon\ p(0) = 0.5 \quad \text{vs.} \quad H_2\colon\ p(0) = 1$$

Moneda común   Moneda “dos caras”

$$p(H|D) = \frac{p(D|H)p(H)}{p(D)}$$

Comparación de dos hipótesis simples

$$H_1\colon\ p(0) = 0.5 \quad \text{vs.} \quad H_2\colon\ p(0) = 1$$

Moneda común — Moneda “dos caras”

$$p(H|D) = \frac{p(D|H)p(H)}{p(D)}$$

Con dos hipótesis, comparamos las “chances” *(odds ratio)*

$$\frac{p(H_1|D)}{p(H_2|D)} = \frac{p(D|H_1)}{p(D|H_2)}\frac{p(H_1)}{p(H_2)}$$

$$\frac{p(H_1|D)}{p(H_2|D)} = \frac{p(D|H_1)}{p(D|H_2)} \frac{p(H_1)}{p(H_2)}$$

H_1: $p(0) = 0.5$

H_2: $p(0) = 1$

$$\frac{p(H_1|D)}{p(H_2|D)} = \frac{p(D|H_1)}{p(D|H_2)}\frac{p(H_1)}{p(H_2)}$$

H_1: $p(0) = 0.5$ $\qquad$ $p(H_1) = 999/1000$

H_2: $p(0) = 1$ $\qquad$ $p(H_2) = 1/1000$

$$\frac{p(H_1|D)}{p(H_2|D)} = \frac{p(D|H_1)}{p(D|H_2)} \frac{p(H_1)}{p(H_2)}$$

H_1: $p(0) = 0.5$

H_2: $p(0) = 1$

$p(H_1) = 999/1000$

$p(H_2) = 1/1000$

$$D = 010010$$

$p(D|H_1) = 1/2^6$

$p(D|H_2) = 0$

$$\frac{p(H_1|D)}{p(H_2|D)} = \frac{p(D|H_1)}{p(D|H_2)} \frac{p(H_1)}{p(H_2)}$$

H_1: $p(0) = 0.5$

H_2: $p(0) = 1$

$p(H_1) = 999/1000$

$p(H_2) = 1/1000$

$$D = 010010$$

$$p(D|H_1) = 1/2^6$$

$$p(D|H_2) = 0$$

$$\Downarrow$$

$$\frac{p(H_1|D)}{p(H_2|D)} = \infty$$

$$\frac{p(H_1|D)}{p(H_2|D)} = \frac{p(D|H_1)}{p(D|H_2)} \frac{p(H_1)}{p(H_2)}$$

H_1: $p(0) = 0.5$ $\qquad$ $p(H_1) = 999/1000$

H_2: $p(0) = 1$ $\qquad$ $p(H_2) = 1/1000$

$$D = 010010 \qquad D = 000000$$

$p(D|H_1) = 1/2^6$ $\qquad$ $p(D|H_1) = 1/2^6$

$p(D|H_2) = 0$ $\qquad$ $p(D|H_2) = 1$

$$\Downarrow$$

$$\frac{p(H_1|D)}{p(H_2|D)} = \infty$$

$$\frac{p(H_1|D)}{p(H_2|D)} = \frac{p(D|H_1)}{p(D|H_2)} \frac{p(H_1)}{p(H_2)}$$

H_1: $p(0) = 0.5$ $\qquad$ $p(H_1) = 999/1000$

H_2: $p(0) = 1$ $\qquad$ $p(H_2) = 1/1000$

$D = 010010$ $\qquad$ $D = 000000$

$$p(D|H_1) = 1/2^6 \qquad p(D|H_1) = 1/2^6$$

$$p(D|H_2) = 0 \qquad p(D|H_2) = 1$$

$$\Downarrow \qquad \Downarrow$$

$$\frac{p(H_1|D)}{p(H_2|D)} = \infty \qquad \frac{p(H_1|D)}{p(H_2|D)} \simeq 16$$

$$\frac{p(H_1|D)}{p(H_2|D)} = \frac{p(D|H_1)}{p(D|H_2)} \frac{p(H_1)}{p(H_2)}$$

H_1: $p(0) = 0.5$ $\qquad$ $p(H_1) = 999/1000$

H_2: $p(0) = 1$ $\qquad$ $p(H_2) = 1/1000$

$$D = 010010 \qquad D = 000000 \qquad D = 0000000000$$

$p(D\|H_1) = 1/2^6$	$p(D\|H_1) = 1/2^6$	$p(D\|H_1) = 1/2^{10}$
$p(D\|H_2) = 0$	$p(D\|H_2) = 1$	$p(D\|H_2) = 1$
$\Downarrow$	$\Downarrow$	
$\frac{p(H_1\|D)}{p(H_2\|D)} = \infty$	$\frac{p(H_1\|D)}{p(H_2\|D)} \simeq 16$	

$$\frac{p(H_1|D)}{p(H_2|D)} = \frac{p(D|H_1)}{p(D|H_2)} \frac{p(H_1)}{p(H_2)}$$

H_1: $p(0) = 0.5$

H_2: $p(0) = 1$

$p(H_1) = 999/1000$

$p(H_2) = 1/1000$

$$D = 010010$$

$$p(D|H_1) = 1/2^6$$

$$p(D|H_2) = 0$$

$$\Downarrow$$

$$\frac{p(H_1|D)}{p(H_2|D)} = \infty$$

$$D = 000000$$

$$p(D|H_1) = 1/2^6$$

$$p(D|H_2) = 1$$

$$\Downarrow$$

$$\frac{p(H_1|D)}{p(H_2|D)} \simeq 16$$

$$D = 0000000000$$

$$p(D|H_1) = 1/2^{10}$$

$$p(D|H_2) = 1$$

$$\Downarrow$$

$$\frac{p(H_1|D)}{p(H_2|D)} \simeq 1$$

$$\frac{p(H_1|D)}{p(H_2|D)} = \frac{p(D|H_1)}{p(D|H_2)} \frac{p(H_1)}{p(H_2)}$$

H_1: $p(0) = 0.5$ $\qquad$ $p(H_1) = 999/1000$

H_2: $p(0) = 1$ $\qquad$ $p(H_2) = 1/1000$

$$D = 010010 \qquad D = 000000 \qquad D = 0000000000$$

$p(D\|H_1) = 1/2^6$	$p(D\|H_1) = 1/2^6$	$p(D\|H_1) = 1/2^{10}$
$p(D\|H_2) = 0$	$p(D\|H_2) = 1$	$p(D\|H_2) = 1$
$\Downarrow$	$\Downarrow$	$\Downarrow$
$\frac{p(H_1\|D)}{p(H_2\|D)} = \infty$	$\frac{p(H_1\|D)}{p(H_2\|D)} \simeq 16$	$\frac{p(H_1\|D)}{p(H_2\|D)} \simeq 1$

Combina conocimiento previo con evidencia

Comparación de una hipótesis simple con una compleja

$$H_1\colon\ p(0) = 0.5 \quad \text{vs.} \quad H_2\colon\ p(0) = \theta$$

Comparación de una hipótesis simple con una compleja

$$H_1\colon\ p(0) = 0.5 \quad \text{vs.} \quad H_2\colon\ p(0) = \theta$$

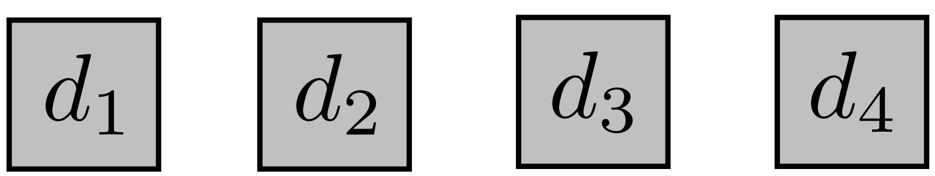
Moneda común

$\boxed{d_1}$ $\boxed{d_2}$ $\boxed{d_3}$ $\boxed{d_4}$

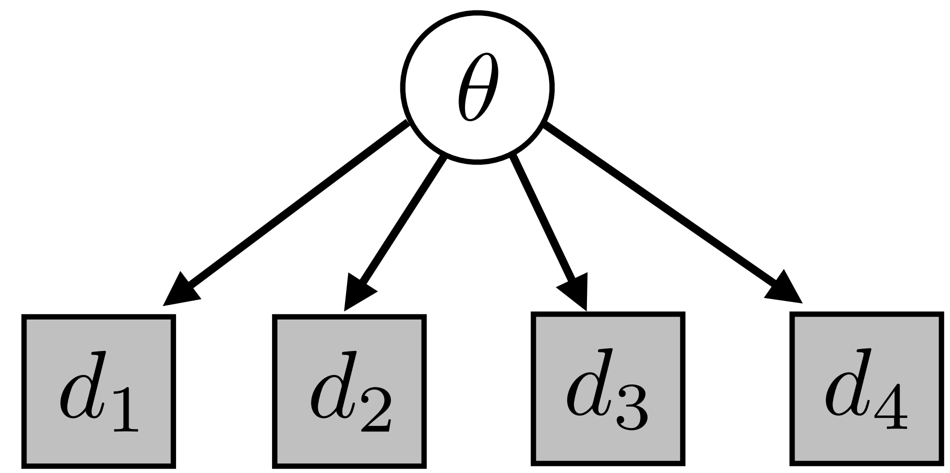
Comparación de una hipótesis simple con una compleja

$$H_1\!: \ p(0) = 0.5 \quad \text{vs.} \quad H_2\!: \ p(0) = \theta$$

Moneda común



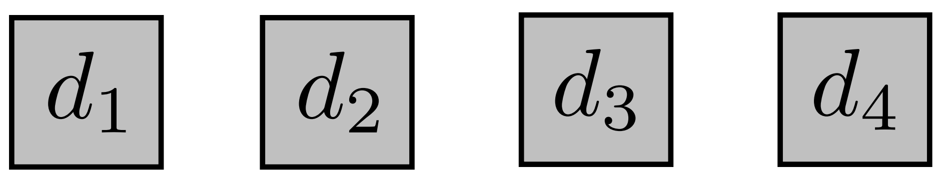
Moneda cargada



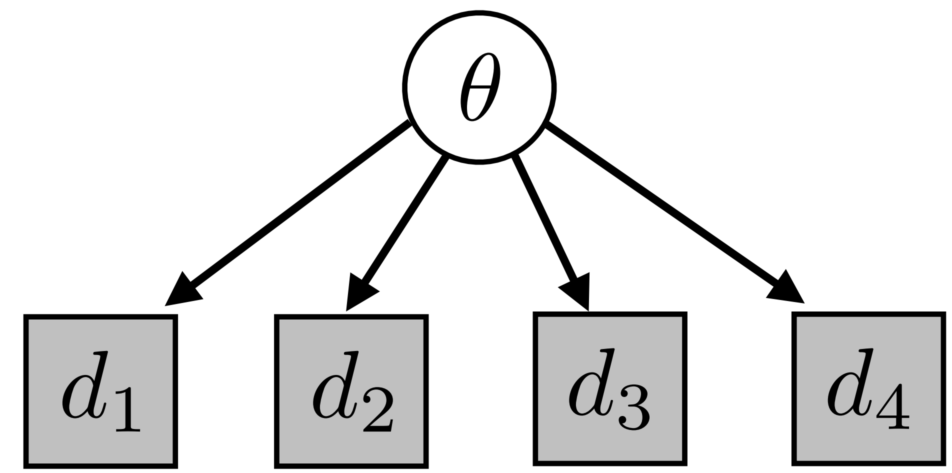
Comparación de una hipótesis simple con una compleja

$$H_1\colon\ p(0) = 0.5 \quad \text{vs.} \quad H_2\colon\ p(0) = \theta$$

Moneda común



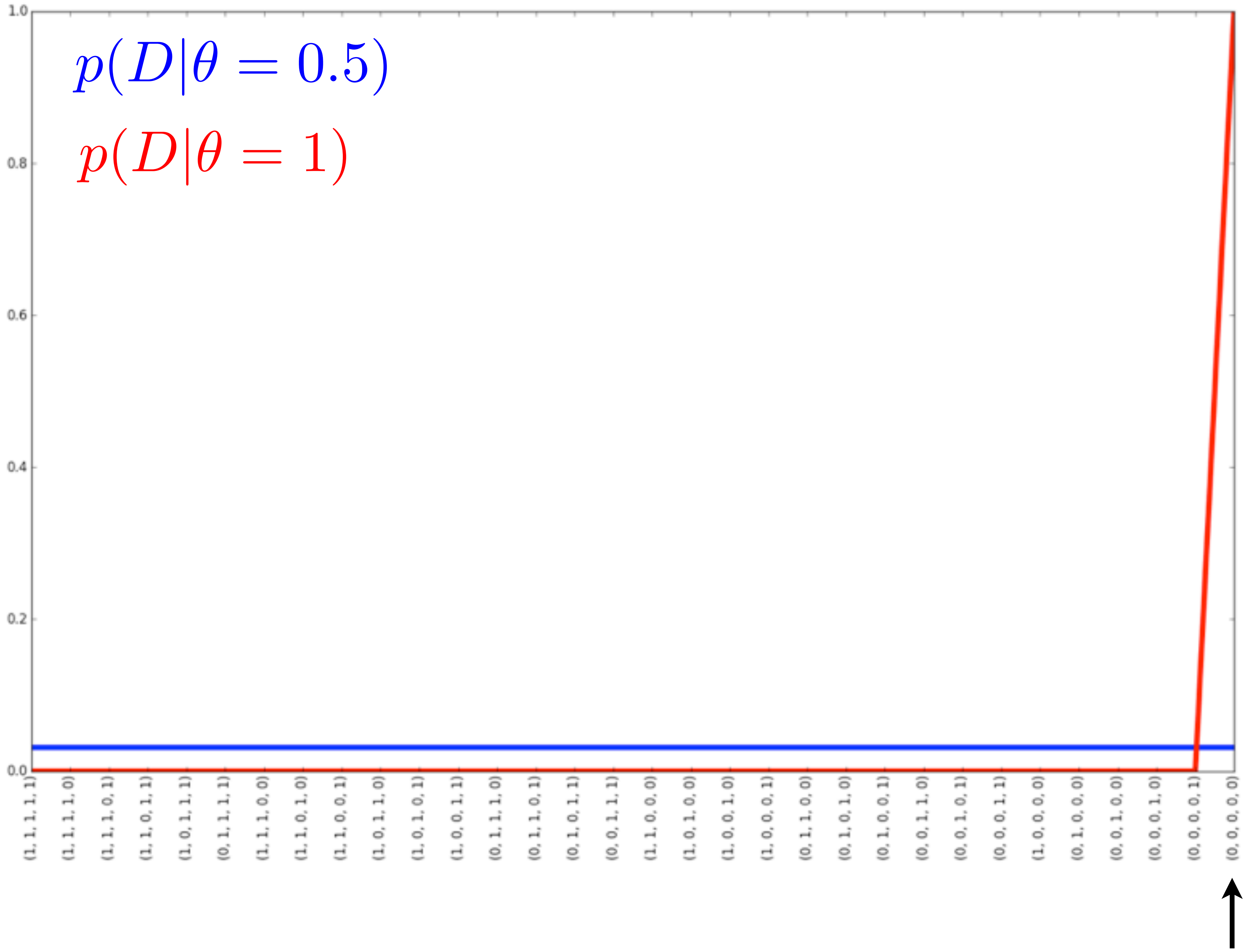
Moneda cargada



H_2 es más compleja:

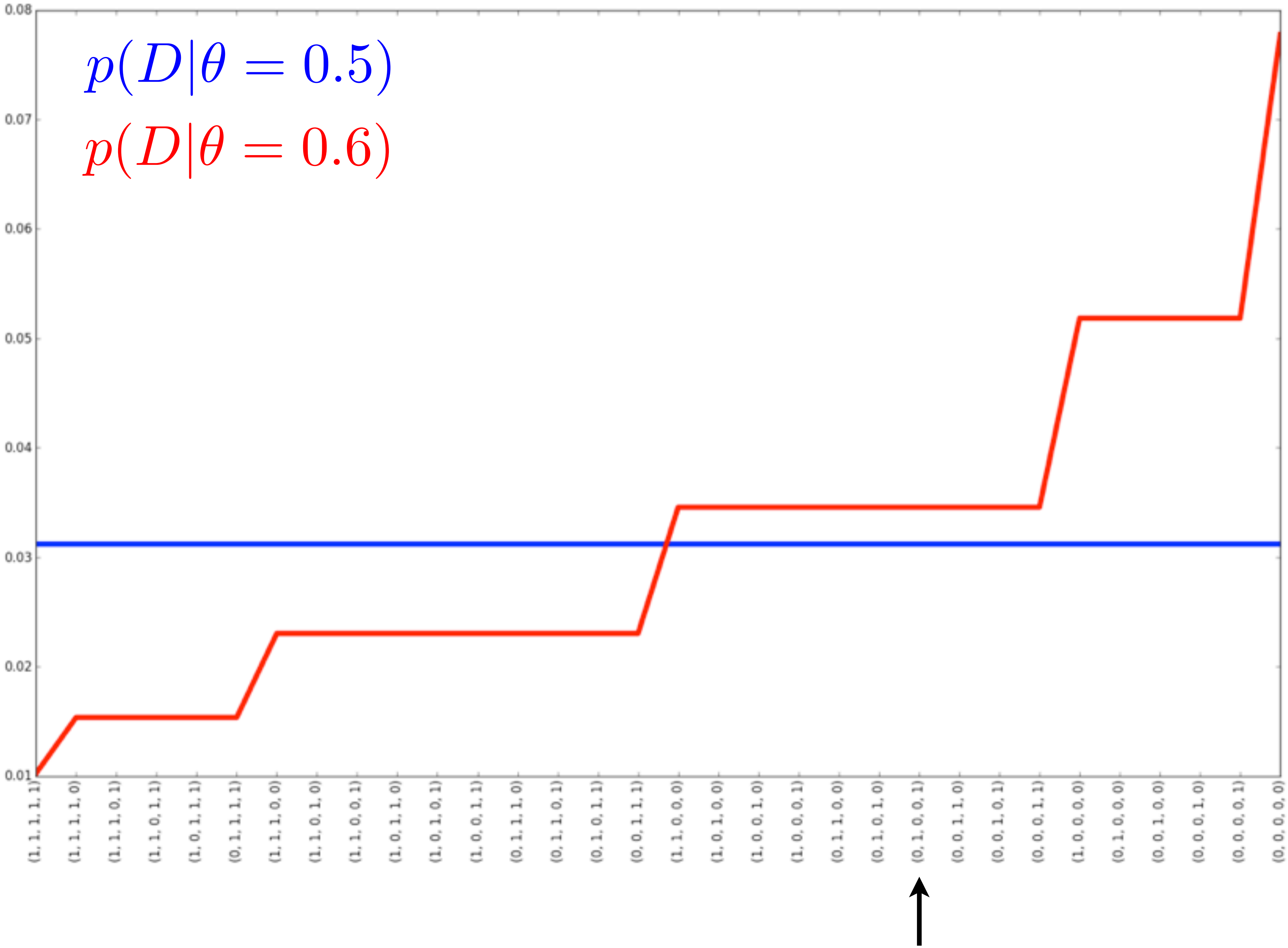
- *H_1* es un caso particular de *H_2*
- Para cualquier secuencia *D,* podemos elegir θ tal que *D* sea más probable bajo *H_2* que bajo *H_1*

$$D = 00000$$



$p(D|\theta = 0.5)$
$p(D|\theta = 1)$
1.0
0.8
0.6
0.4
0.2
0.0
(1, 1, 1, 1)
(1, 1, 1, 0)
(1, 1, 0, 1)
(1, 1, 0, 1)
(1, 0, 1, 1)
(0, 1, 1, 1)
(1, 1, 0, 0)
(1, 1, 0, 1)
(1, 1, 0, 1)
(1, 0, 1, 0)
(1, 0, 1, 1)
(1, 0, 0, 1)
(0, 1, 1, 0)
(0, 1, 0, 1)
(0, 1, 1, 1)
(0, 0, 1, 1)
(1, 1, 0, 0)
(1, 0, 0, 0)
(1, 0, 1, 0)
(1, 0, 0, 1)
(0, 1, 0, 0)
(0, 1, 1, 0)
(0, 1, 0, 1)
(0, 0, 1, 0)
(0, 0, 1, 1)
(0, 0, 0, 1)
(1, 0, 0, 0)
(0, 1, 0, 0)
(0, 0, 1, 0)
(0, 0, 0, 0)
(0, 0, 0, 1)
(0, 0, 0, 0)

$$D = 01101$$


$p(D|\theta = 0.5)$
$p(D|\theta = 0.6)$
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
(1, 1, 1, 1, 1)
(1, 1, 1, 1, 0)
(1, 1, 1, 0, 1)
(1, 1, 0, 1, 1)
(1, 0, 1, 1, 1)
(0, 1, 1, 1, 1)
(1, 1, 1, 0, 0)
(1, 1, 0, 1, 0)
(1, 1, 0, 0, 1)
(1, 0, 1, 1, 0)
(1, 0, 1, 0, 1)
(1, 0, 0, 1, 1)
(0, 1, 1, 1, 0)
(0, 1, 1, 0, 1)
(0, 1, 0, 1, 1)
(0, 0, 1, 1, 1)
(1, 1, 0, 0, 0)
(1, 0, 1, 0, 0)
(1, 0, 0, 1, 0)
(1, 0, 0, 0, 1)
(0, 1, 1, 0, 0)
(0, 1, 0, 1, 0)
(0, 1, 0, 0, 1)
(0, 0, 1, 1, 0)
(0, 0, 1, 0, 1)
(0, 0, 0, 1, 1)
(1, 0, 0, 0, 0)
(0, 1, 0, 0, 0)
(0, 0, 1, 0, 0)
(0, 0, 0, 1, 0)
(0, 0, 0, 0, 1)
(0, 0, 0, 0, 0)

Entonces..
¿cómo lidiamos con distintas complejidades?

Entonces..
¿cómo lidiamos con distintas complejidades?

- Estadística frecuentista: tests de hipótesis

Entonces..
¿cómo lidiamos con distintas complejidades?

- Estadística frecuentista: tests de hipótesis
- Teoría de información: longitud de descripción mínima

Entonces..
¿cómo lidiamos con distintas complejidades?

- Estadística frecuentista: tests de hipótesis
- Teoría de información: longitud de descripción mínima
- Inferencia Bayesiana: probabilidades

Entonces..
¿cómo lidiamos con distintas complejidades?

- Estadística frecuentista: tests de hipótesis
- Teoría de información: longitud de descripción mínima
- Inferencia Bayesiana: probabilidades

$$H_1\colon\ p(0) = 0.5$$

$$H_2\colon\ p(0) = \theta$$

Entonces..
¿cómo lidiamos con distintas complejidades?

- Estadística frecuentista: tests de hipótesis
- Teoría de información: longitud de descripción mínima
- Inferencia Bayesiana: probabilidades

$$H_1\colon\ p(0) = 0.5$$
$$H_2\colon\ p(0) = \theta$$

$$\frac{p(H_1|D)}{p(H_2|D)} = \frac{p(D|H_1)}{p(D|H_2)}\frac{p(H_1)}{p(H_2)}$$

Entonces..
¿cómo lidiamos con distintas complejidades?

- Estadística frecuentista: tests de hipótesis
- Teoría de información: longitud de descripción mínima
- Inferencia Bayesiana: probabilidades

$$H_1\colon\ p(0) = 0.5$$
$$H_2\colon\ p(0) = \theta$$

$$\frac{p(H_1|D)}{p(H_2|D)} = \frac{p(D|H_1)}{p(D|H_2)}\frac{p(H_1)}{p(H_2)}$$

$$p(D|H_1) = 1/2^N$$

Entonces..
¿cómo lidiamos con distintas complejidades?

- Estadística frecuentista: tests de hipótesis
- Teoría de información: longitud de descripción mínima
- Inferencia Bayesiana: probabilidades

$$H_1\!: \; p(0) = 0.5$$
$$H_2\!: \; p(0) = \theta$$

$$\frac{p(H_1|D)}{p(H_2|D)} = \frac{p(D|H_1)}{p(D|H_2)}\frac{p(H_1)}{p(H_2)}$$

$$p(D|H_1) = 1/2^N$$

$$p(D|H_2) = \int_0^1 p(D|\theta)p(\theta|H_2)d\theta$$

promediamos sobre θ

Entonces..
¿cómo lidiamos con distintas complejidades?

- Estadística frecuentista: tests de hipótesis
- Teoría de información: longitud de descripción mínima
- Inferencia Bayesiana: probabilidades

$$H_1\!: \; p(0) = 0.5$$
$$H_2\!: \; p(0) = \theta$$

$$\frac{p(H_1|D)}{p(H_2|D)} = \frac{p(D|H_1)}{p(D|H_2)} \frac{p(H_1)}{p(H_2)}$$

$$p(D|H_1) = 1/2^N$$

1 (uniforme)

$$p(D|H_2) = \int_0^1 p(D|\theta)p(\theta|H_2)d\theta$$

promediamos sobre θ

Entonces..
¿cómo lidiamos con distintas complejidades?

- Estadística frecuentista: tests de hipótesis
- Teoría de información: longitud de descripción mínima
- Inferencia Bayesiana: probabilidades

$$H_1\colon\ p(0) = 0.5$$
$$H_2\colon\ p(0) = \theta$$

$$\frac{p(H_1|D)}{p(H_2|D)} = \frac{p(D|H_1)}{p(D|H_2)}\frac{p(H_1)}{p(H_2)}$$

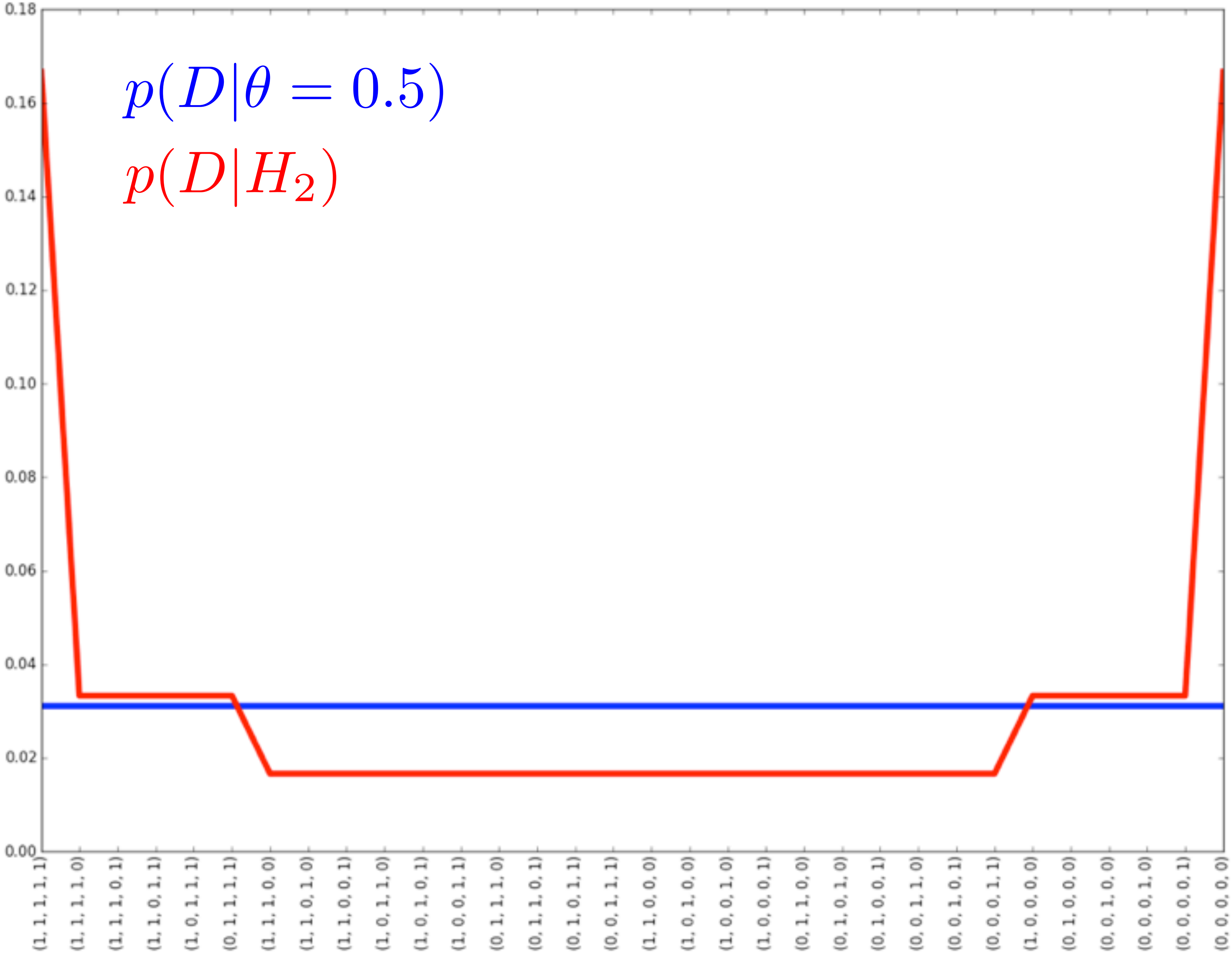
$$p(D|H_1) = 1/2^N$$

1 (uniforme)

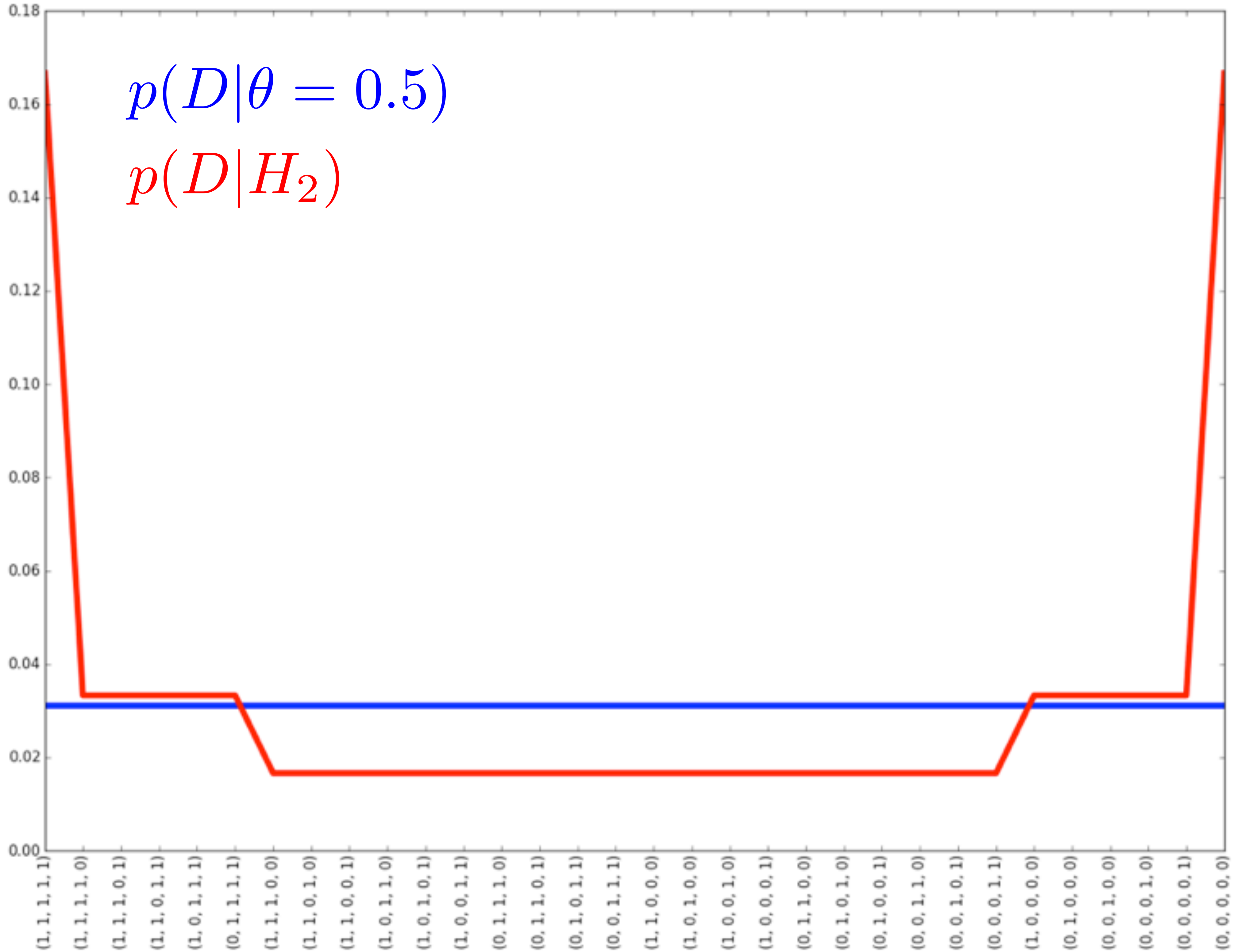
$$p(D|H_2) = \int_0^1 p(D|\theta)p(\theta|H_2)d\theta$$

promediamos sobre θ

$$\theta^k(1-\theta)^{N-k}$$

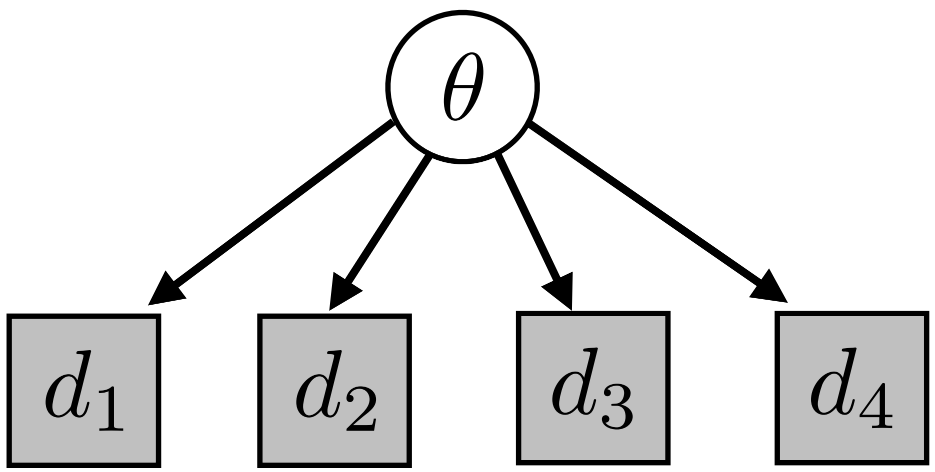


$p(D|\theta = 0.5)$
$p(D|H_2)$
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
(1, 1, 1, 1, 1)
(1, 1, 1, 1, 0)
(1, 1, 1, 0, 1)
(1, 1, 0, 1, 1)
(1, 0, 1, 1, 1)
(0, 1, 1, 1, 1)
(1, 1, 1, 0, 0)
(1, 1, 0, 1, 0)
(1, 1, 0, 0, 1)
(1, 0, 1, 1, 0)
(1, 0, 1, 0, 1)
(1, 0, 0, 1, 1)
(0, 1, 1, 1, 0)
(0, 1, 1, 0, 1)
(0, 1, 0, 1, 1)
(0, 0, 1, 1, 1)
(1, 1, 0, 0, 0)
(1, 0, 1, 0, 0)
(1, 0, 0, 1, 0)
(1, 0, 0, 0, 1)
(0, 1, 1, 0, 0)
(0, 1, 0, 1, 0)
(0, 1, 0, 0, 1)
(0, 0, 1, 1, 0)
(0, 0, 1, 0, 1)
(0, 0, 0, 1, 1)
(1, 0, 0, 0, 0)
(0, 1, 0, 0, 0)
(0, 0, 1, 0, 0)
(0, 0, 0, 1, 0)
(0, 0, 0, 0, 1)
(0, 0, 0, 0, 0)



Automáticamente, la complejidad resulta penalizada
(Occam's Razor)

Comparación de *infinitas* hipótesis

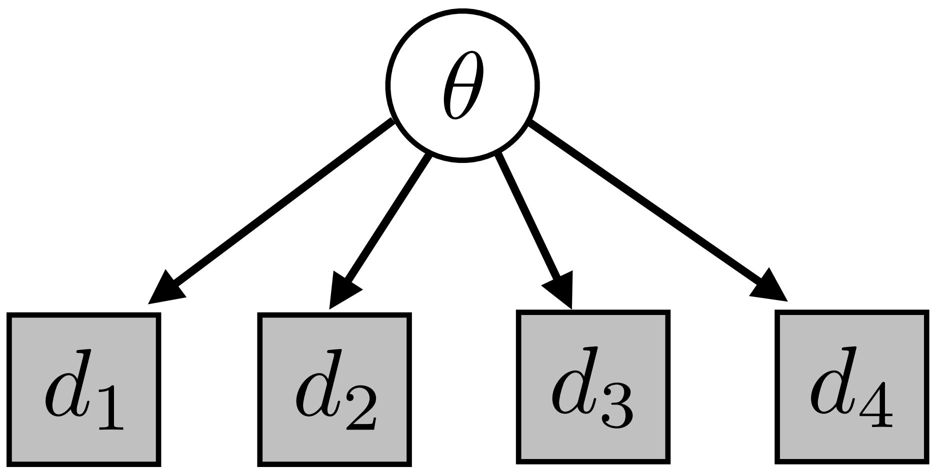


$$p(0) = \theta$$

nos preguntamos por el valor de θ

Cada valor de θ es una hipótesis *H*

Comparación de *infinitas* hipótesis



$$p(0) = \theta$$

nos preguntamos
por el valor de θ

Cada valor de θ es una hipótesis *H*

Volvemos al “experimento” inicial...

¿ $p(0)$ en la próxima tirada? ¿3/6=0.5?

¿ $p(0)$ en la próxima tirada? ¿3/6=0.5?

¿Y ahora? ¡¿1?!

¿ $p(0)$ en la próxima tirada? ¿3/6=0.5?

¿Y ahora? ¡¿1?!

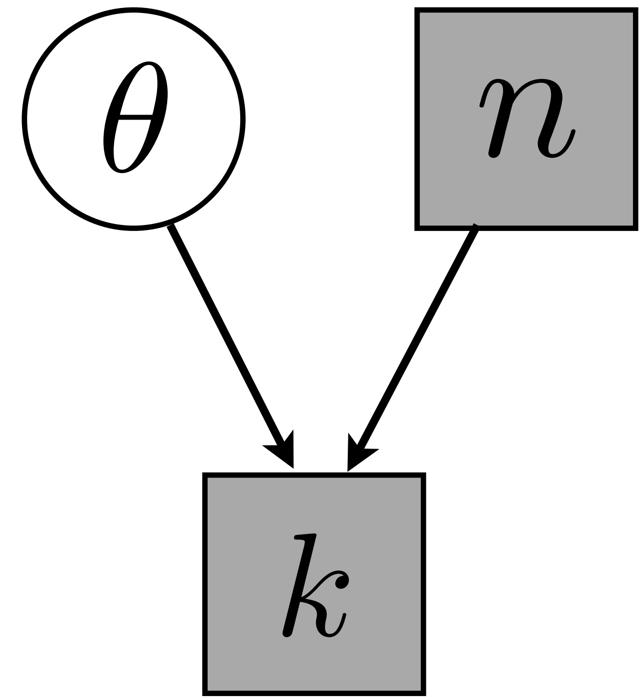
Inferencia Bayesiana, incorporamos conocimientos previos.
Primer paso: modelo.

Modelo

likelihood:  $p(01011|\theta) = \theta^2(1-\theta)^3$

Modelo

likelihood:  $p(01011|\theta) = \theta^2(1-\theta)^3$

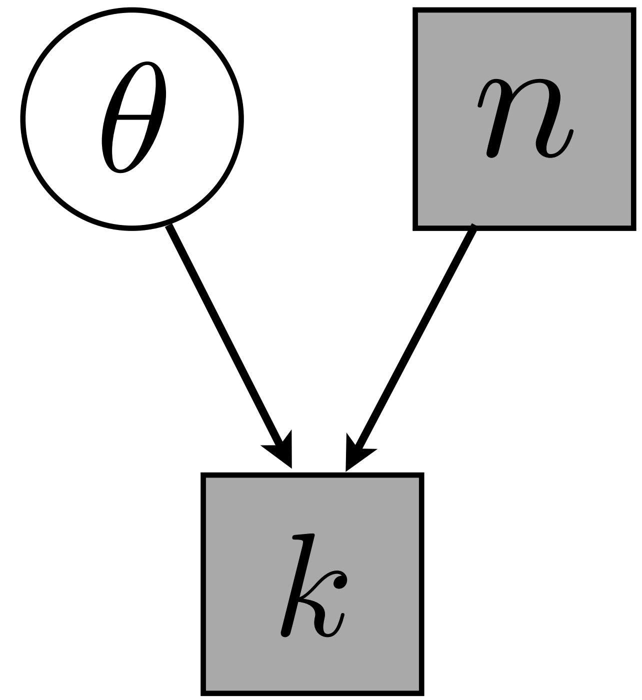


Modelo

likelihood:

$$p(01011|\theta) = \theta^2(1-\theta)^3$$

$$p(k|\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

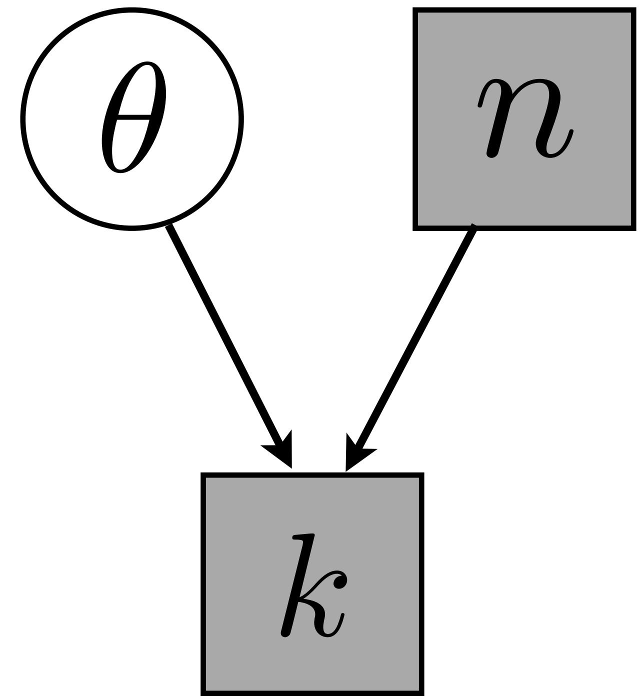


Modelo

likelihood:  $p(01011|\theta) = \theta^2(1-\theta)^3$

número de *caras*

$$p(k|\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$



Modelo

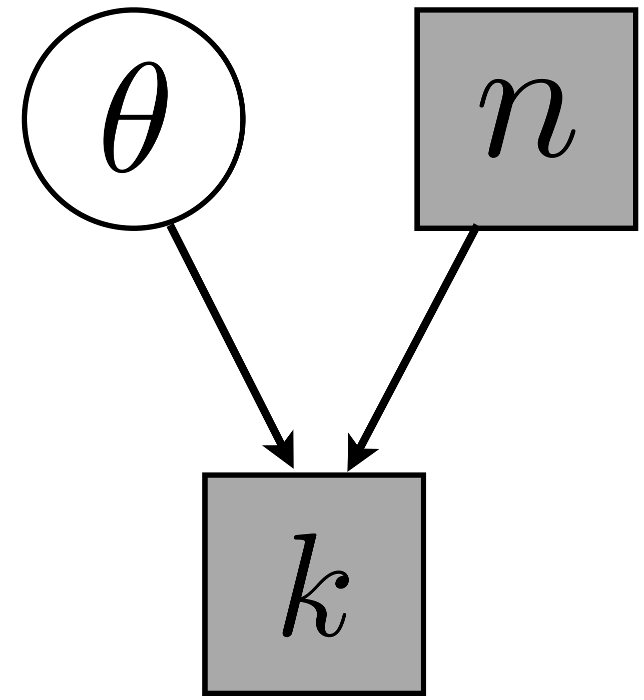
likelihood:

$$p(01011|\theta) = \theta^2(1-\theta)^3$$

número de *caras*

$$p(k|\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

número de *tiradas*



Modelo

likelihood:

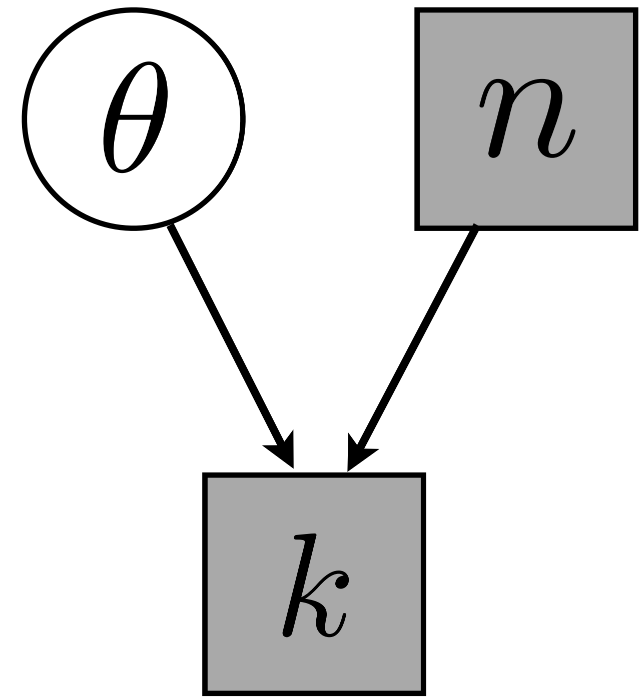
$$p(01011|\theta) = \theta^2(1-\theta)^3$$

número de *caras*

$$p(k|\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

número de *tiradas*

$$k \sim \text{Binomial}(\theta, n)$$



Modelo

likelihood:

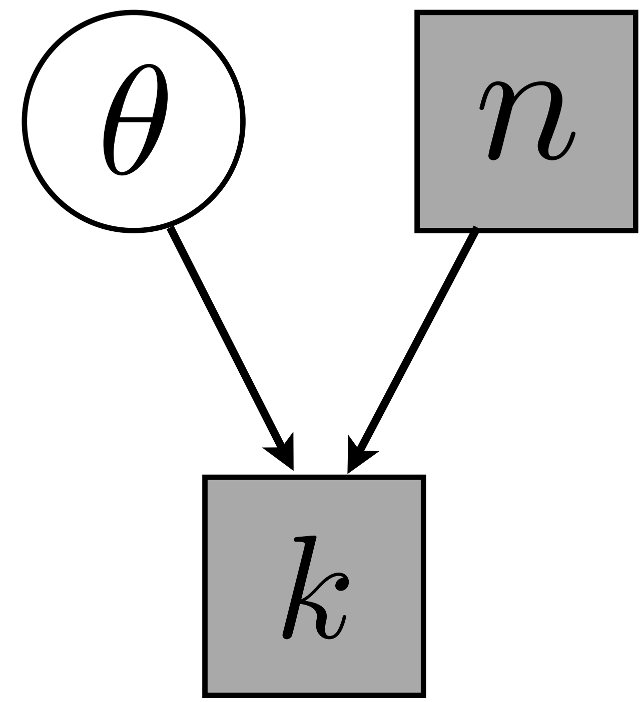
$$p(01011|\theta) = \theta^2(1-\theta)^3$$

número de *caras*

$$p(k|\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

número de *tiradas*

$$k \sim \mathrm{Binomial}(\theta, n)$$



prior:

¿Cómo lo elegimos?

Podemos pensar en experiencia previa *ficticia*

Eligiendo el *prior*...

Si pienso que tiré la moneda y obtuve 1000 *caras* y 1000 *cecas*..

likelihood: $$p(D|\theta) \propto \theta^k (1-\theta)^{n-k}$$

Eligiendo el *prior*...

Si pienso que tiré la moneda y obtuve 1000 *caras* y 1000 *cecas*..

likelihood: $$p(D|\theta) \propto \theta^k (1-\theta)^{n-k}$$

prior: $$p(\theta) \propto \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

Eligiendo el *prior*...

Si pienso que tiré la moneda y obtuve 1000 *caras* y 1000 *cecas*..

likelihood: $$p(D|\theta) \propto \theta^k (1-\theta)^{n-k}$$

prior: $$p(\theta) \propto \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

α número ficticio de *caras*

β número ficticio de *cecas*

Eligiendo el *prior*...

Si pienso que tiré la moneda y obtuve 1000 *caras* y 1000 *cecas*..

likelihood: $p(D|\theta) \propto \theta^k (1-\theta)^{n-k}$

prior: $p(\theta) \propto \theta^{\alpha-1} (1-\theta)^{\beta-1}$

α número ficticio de *caras*

β número ficticio de *cecas*

Distribución $\mathrm{Beta}(\alpha, \beta)$, *conjugada* de la *Binomial*

Distribuciones conjugadas

$$p(\theta|D) \propto p(D|\theta)p(\theta)$$

$$p(\theta|D) \propto \theta^k(1-\theta)^{n-k}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Distribuciones conjugadas

$$p(\theta|D) \propto p(D|\theta)p(\theta)$$

$$p(\theta|D) \propto \underbrace{\theta^k(1-\theta)^{n-k}}_{\textit{likelihood}}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Distribuciones conjugadas

$$p(\theta|D) \propto p(D|\theta)p(\theta)$$

$$p(\theta|D) \propto \underbrace{\theta^k(1-\theta)^{n-k}}_{\textit{likelihood}}\underbrace{\theta^{\alpha-1}(1-\theta)^{\beta-1}}_{\textit{prior}}$$

Distribuciones conjugadas

$$p(\theta|D) \propto p(D|\theta)p(\theta)$$

$$p(\theta|D) \propto \underbrace{\theta^{k}(1-\theta)^{n-k}}_{\textit{likelihood}}\underbrace{\theta^{\alpha-1}(1-\theta)^{\beta-1}}_{\textit{prior}}$$

$$= \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}$$

Distribuciones conjugadas

$$p(\theta|D) \propto p(D|\theta)p(\theta)$$

$$p(\theta|D) \propto \underbrace{\theta^k(1-\theta)^{n-k}}_{\textit{likelihood}}\underbrace{\theta^{\alpha-1}(1-\theta)^{\beta-1}}_{\textit{prior}}$$

$$= \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}$$

$$\mathrm{Beta}(k+\alpha, n-k+\beta)$$

Distribuciones conjugadas

$$p(\theta|D) \propto p(D|\theta)p(\theta)$$

$$p(\theta|D) \propto \underbrace{\theta^k(1-\theta)^{n-k}}_{\textit{likelihood}}\underbrace{\theta^{\alpha-1}(1-\theta)^{\beta-1}}_{\textit{prior}}$$

$$= \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}$$

$$\mathrm{Beta}(k+\alpha, n-k+\beta)$$

real + ficticio

Distribuciones conjugadas

$$p(\theta|D) \propto p(D|\theta)p(\theta)$$

$$p(\theta|D) \propto \underbrace{\theta^k(1-\theta)^{n-k}}_{\textit{likelihood}}\underbrace{\theta^{\alpha-1}(1-\theta)^{\beta-1}}_{\textit{prior}}$$

$$= \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}$$

$$\mathrm{Beta}(k+\alpha, n-k+\beta)$$

real + ficticio

$$\mathrm{Beta}(.) \propto \mathrm{Binomial}(.)\mathrm{Beta}(.)$$

Distribuciones conjugadas

$$p(\theta|D) \propto p(D|\theta)p(\theta)$$

$$p(\theta|D) \propto \underbrace{\theta^k(1-\theta)^{n-k}}_{\textit{likelihood}}\underbrace{\theta^{\alpha-1}(1-\theta)^{\beta-1}}_{\textit{prior}}$$

$$= \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}$$

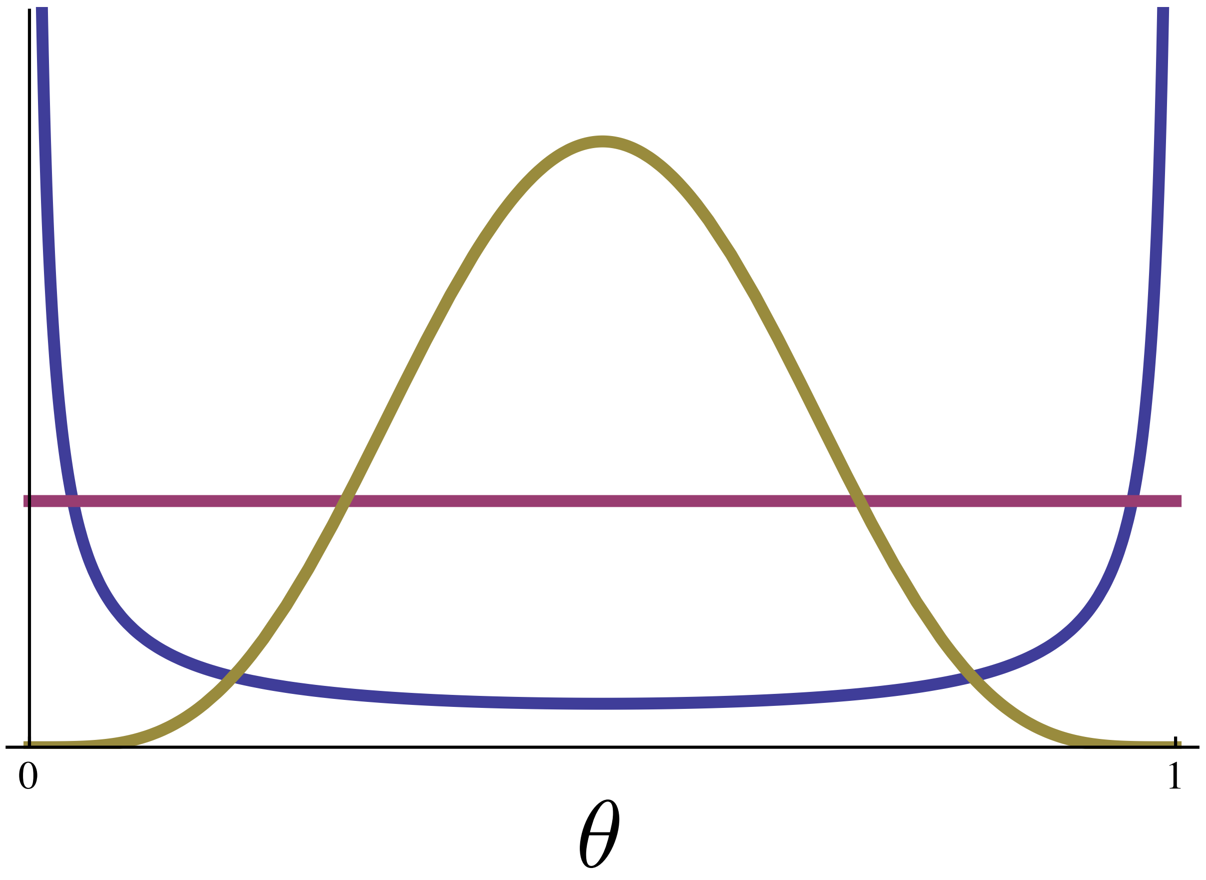
$$\mathrm{Beta}(k+\alpha, n-k+\beta)$$

real + ficticio

$$\mathrm{Beta}(.) \propto \mathrm{Binomial}(.)\mathrm{Beta}(.)$$

Propiedad útil para el cómputo y la interpretación

Distribución $\mathrm{Beta}(\alpha, \beta)$



$$\mathrm{Beta}(\alpha, \alpha)$$

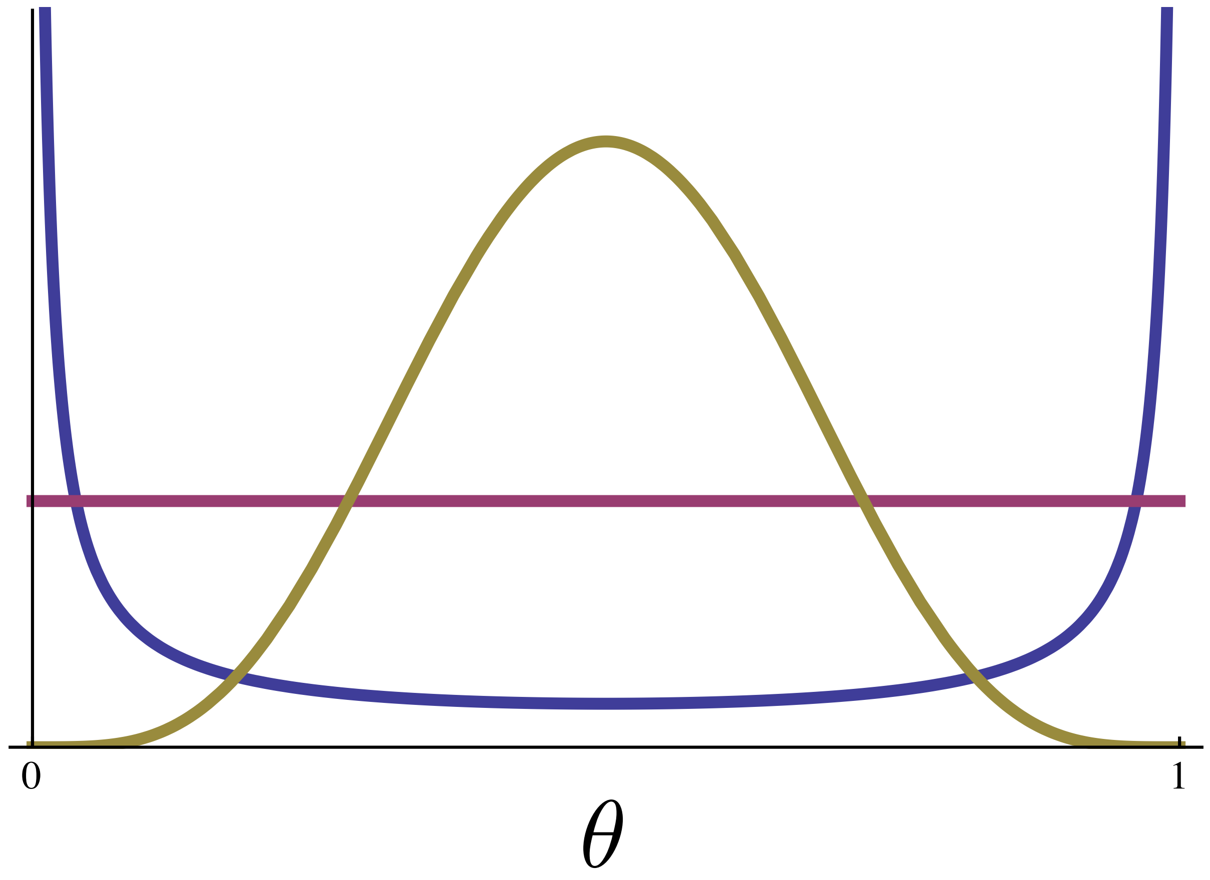
α

0.01

1

5

Distribución $\mathrm{Beta}(\alpha, \beta)$



$\mathrm{Beta}(\alpha, \alpha)$

α

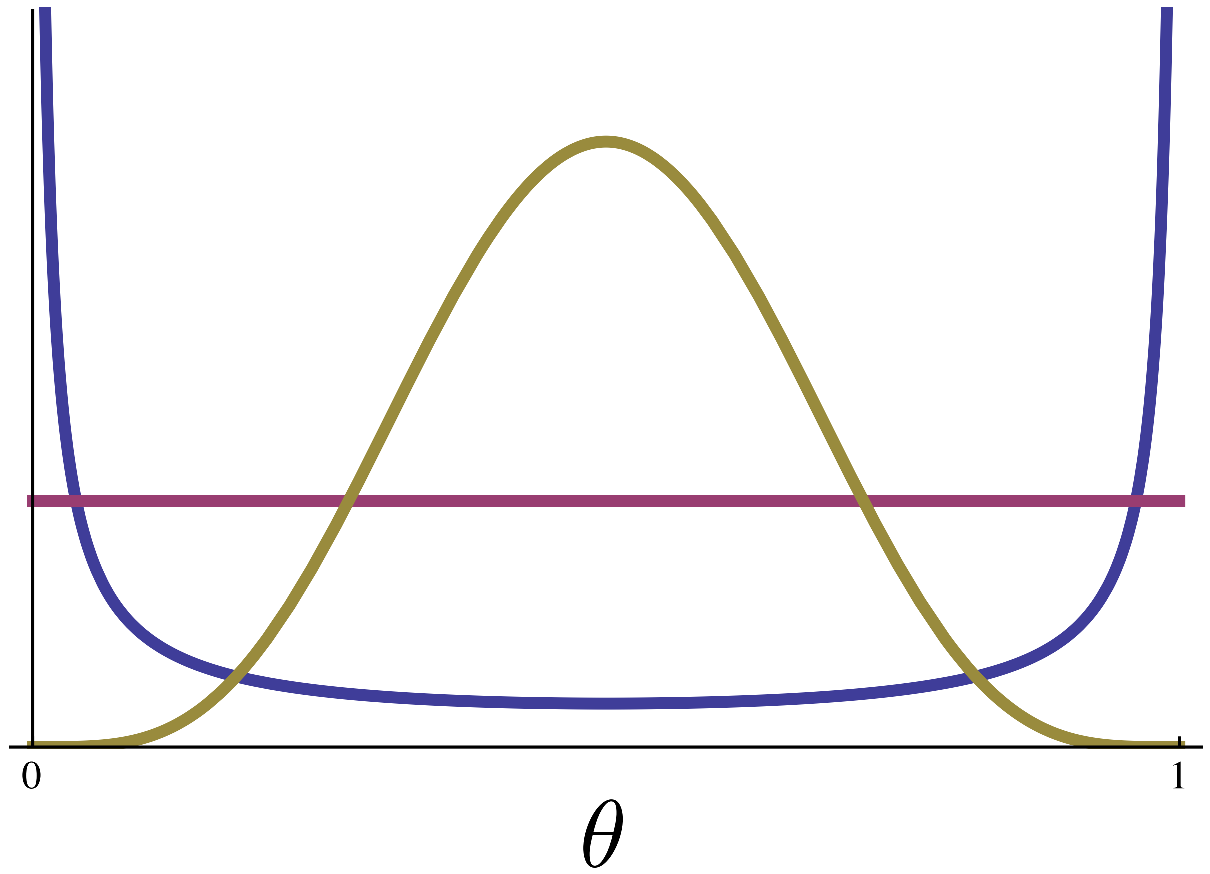
0.01

1

5

$\alpha, \beta = 1$ Distribución uniforme

Distribución $\mathrm{Beta}(\alpha, \beta)$



$$\mathrm{Beta}(\alpha, \alpha)$$

α

0.01

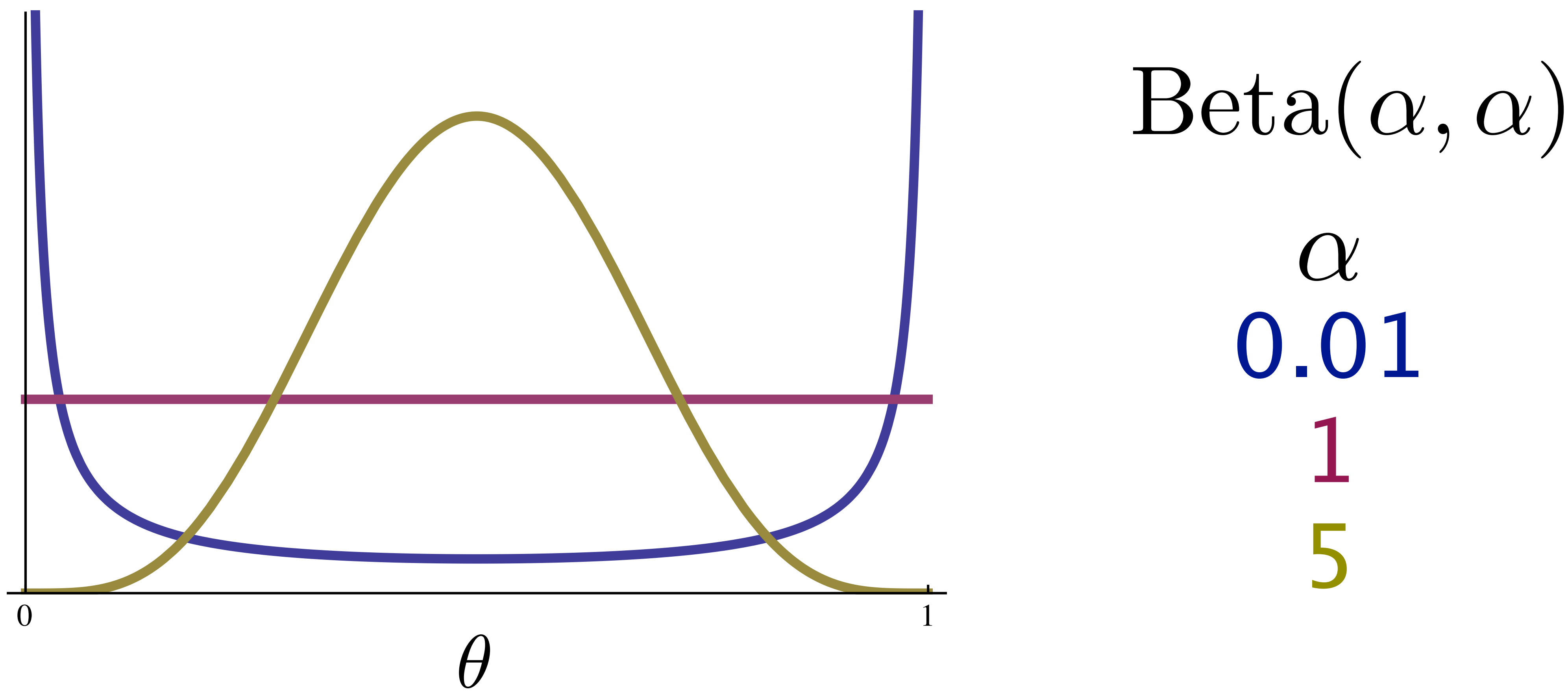
1

5

$\alpha, \beta = 1$ Distribución uniforme

$\alpha, \beta = \frac{1}{2}$ *Jeffrey's prior:* del principio de invariancia frente a transformaciones de variables

Distribución $\mathrm{Beta}(\alpha, \beta)$



$\alpha, \beta = 1$ Distribución uniforme

$\alpha, \beta = \frac{1}{2}$ *Jeffrey's prior:* del principio de invariancia frente a transformaciones de variables

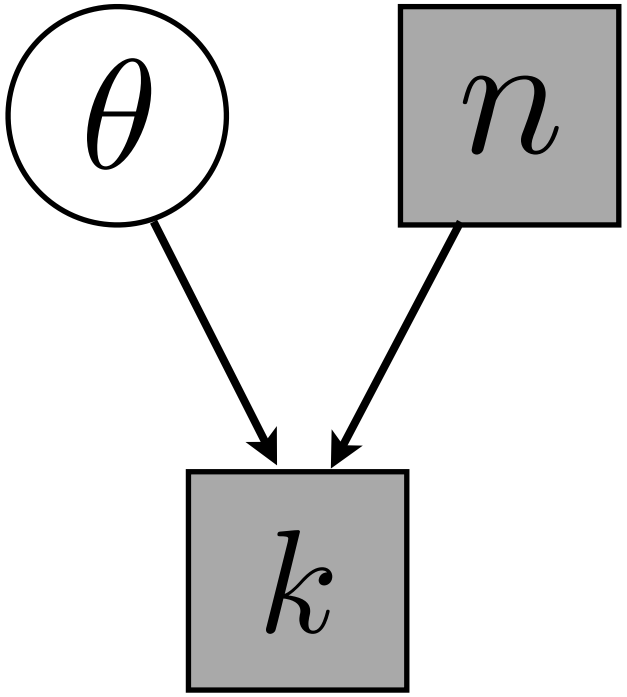
$\alpha, \beta = 0$ *Haldane,* impropia: pueden dar *posteriors propias*

Modelo

likelihood:

$$p(k|\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

$$k \sim \text{Binomial}(\theta, n)$$



prior:

$$\theta \sim \text{Uniform}(0,1) = \text{Beta}(1,1)$$

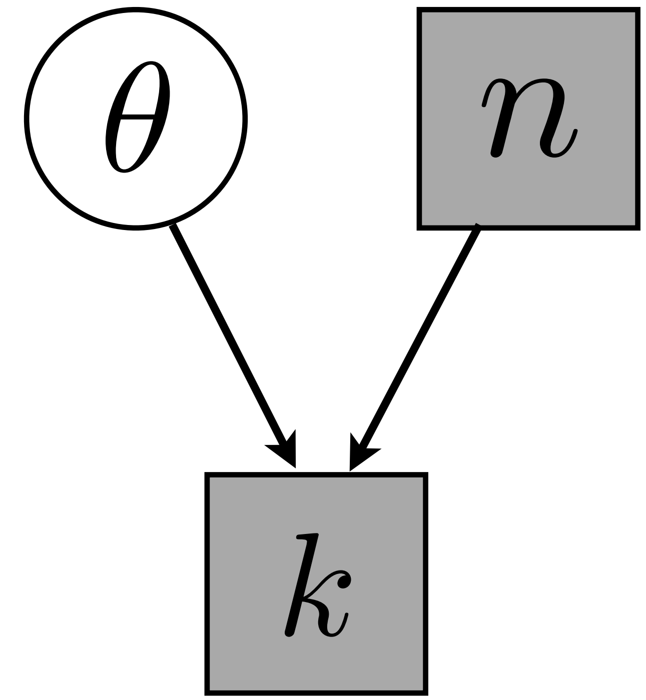
$$\theta \sim \text{Beta}(100,100)$$

Modelo

likelihood:

$$p(k|\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

$$k \sim \mathrm{Binomial}(\theta, n)$$



prior:

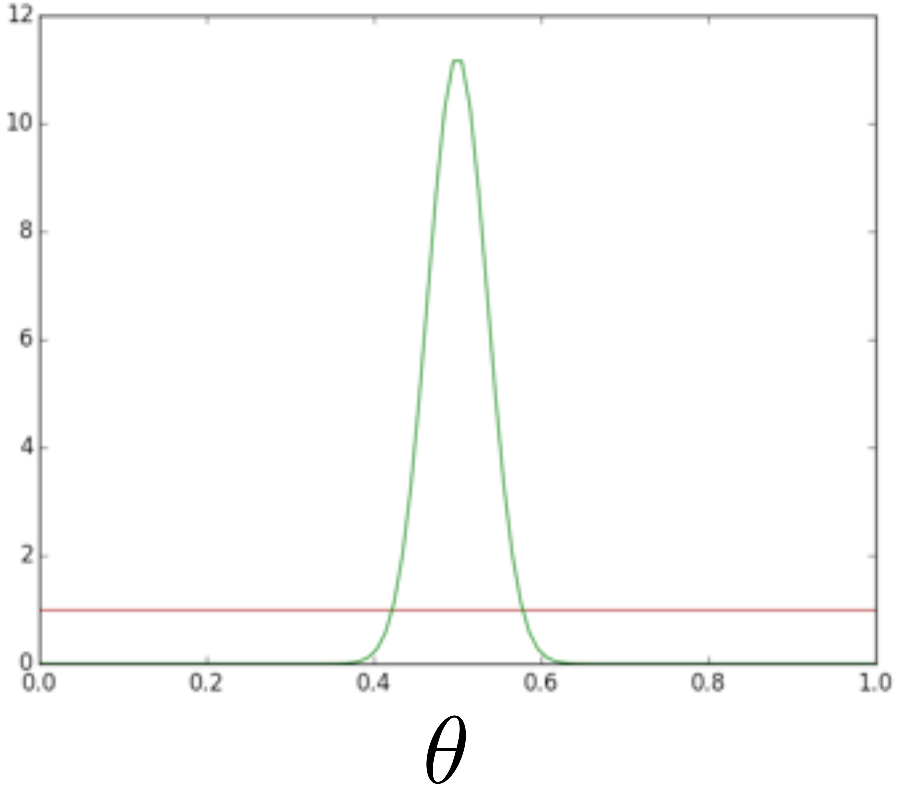
$$\theta \sim \mathrm{Uniform}(0,1) = \mathrm{Beta}(1,1)$$

$$\theta \sim \mathrm{Beta}(100,100)$$

posterior:

$$p(\theta|D) = \mathrm{Beta}(k+1, n-k+1)$$

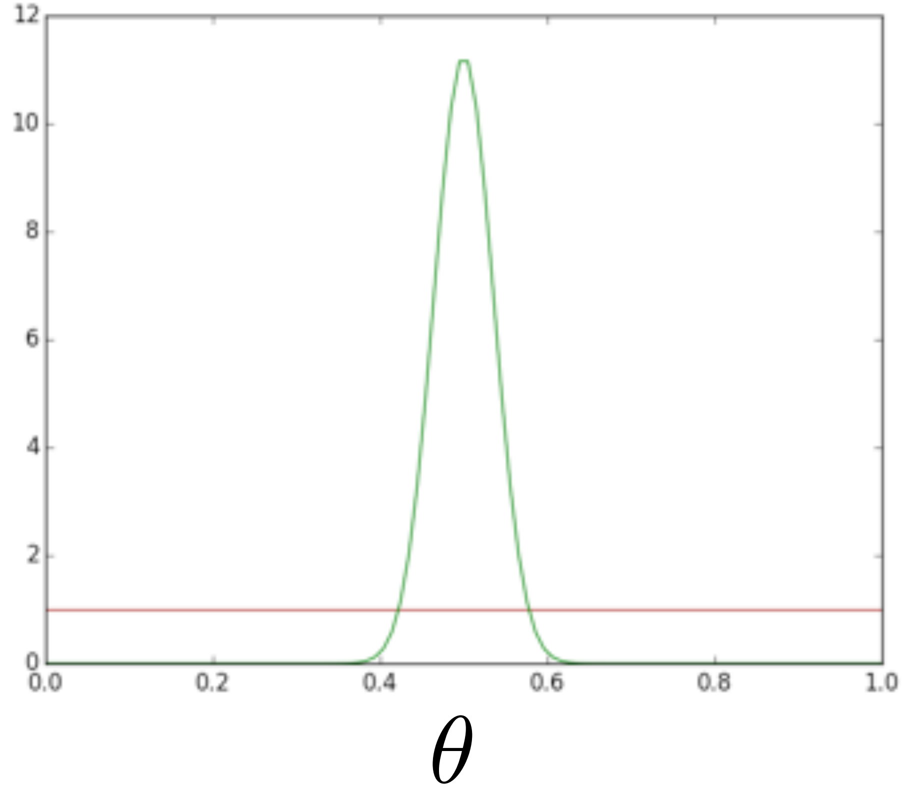
$$p(\theta|D) = \mathrm{Beta}(k+100, n-k+100)$$



prior

$$\theta \sim \mathrm{Uniform}(0, 1) = \mathrm{Beta}(1, 1)$$

$$\theta \sim \mathrm{Beta}(100, 100)$$



prior

$$\theta \sim \mathrm{Uniform}(0, 1) = \mathrm{Beta}(1, 1)$$

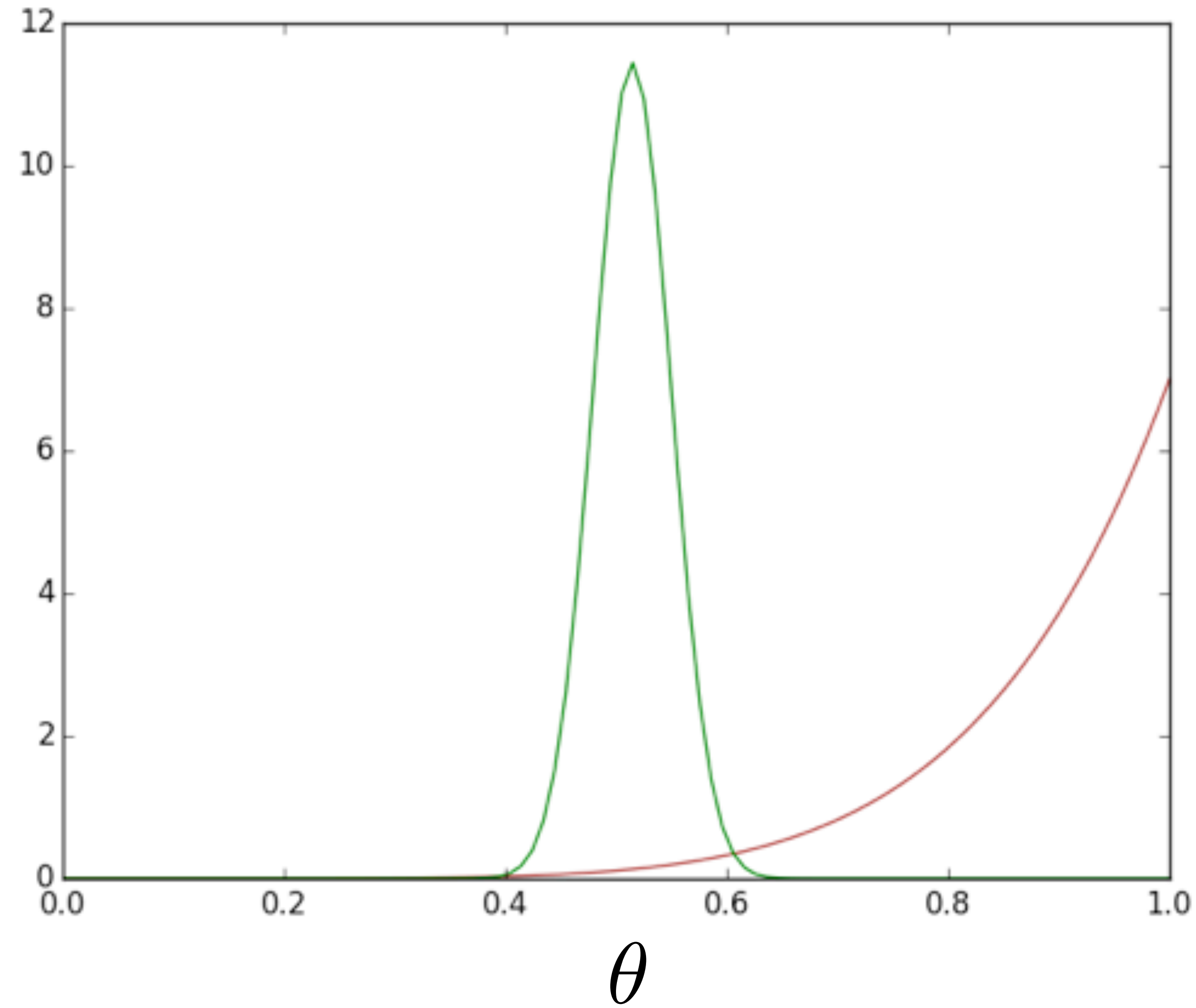
$$\theta \sim \mathrm{Beta}(100, 100)$$

posterior

D: k=6, n=6

$$p(\theta|D) = \mathrm{Beta}(k + 1, n - k + 1)$$

$$p(\theta|D) = \mathrm{Beta}(k + 100, n - k + 100)$$



Media *a posteriori*

$$p(\theta|n,k) = \mathrm{Beta}(k+\alpha, n-k+\beta)$$

Media *a posteriori*

$$p(\theta|n,k) = \mathrm{Beta}(k+\alpha, n-k+\beta)$$

$$\mathrm{Beta}(\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Media *a posteriori*

$$p(\theta|n,k) = \mathrm{Beta}(k+\alpha, n-k+\beta)$$

$$\mathrm{Beta}(\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$E[\theta|n,k] = \int_0^1 \theta p(\theta|n,k)d\theta = \frac{k+\alpha}{n+\alpha+\beta}$$

Media *a posteriori*

$$p(\theta|n,k) = \mathrm{Beta}(k+\alpha, n-k+\beta)$$

$$\mathrm{Beta}(\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$E[\theta|n,k] = \int_0^1 \theta p(\theta|n,k)d\theta = \frac{k+\alpha}{n+\alpha+\beta}$$

Varianza *a posteriori*

$$\mathrm{var}(\theta|n,k) = \frac{\mathrm{E}(\theta|n,k)(1-\mathrm{E}(\theta|n,k))}{\alpha+\beta+n+1}$$

Media *a posteriori*

$$p(\theta|n,k) = \mathrm{Beta}(k+\alpha, n-k+\beta)$$

$$\mathrm{Beta}(\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$E[\theta|n,k] = \int_0^1 \theta p(\theta|n,k)d\theta = \frac{k+\alpha}{n+\alpha+\beta}$$

Varianza *a posteriori*

$$\mathrm{var}(\theta|n,k) = \frac{\mathrm{E}(\theta|n,k)(1-\mathrm{E}(\theta|n,k))}{\alpha+\beta+n+1}$$

Cuando k y $n-k$ crecen con α y β fijos, Media *a posteriori*

$$p(\theta|n,k) = \mathrm{Beta}(k+\alpha, n-k+\beta)$$

$$\mathrm{Beta}(\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$E[\theta|n,k] = \int_0^1 \theta p(\theta|n,k)d\theta = \frac{k+\alpha}{n+\alpha+\beta}$$

Varianza *a posteriori*

$$\mathrm{var}(\theta|n,k) = \frac{\mathrm{E}(\theta|n,k)(1-\mathrm{E}(\theta|n,k))}{\alpha+\beta+n+1}$$

Cuando k y $n-k$ crecen con α y β fijos,

$$\mathrm{E}(\theta|k,n) \approx k/n$$

Media *a posteriori*

$$p(\theta|n,k) = \mathrm{Beta}(k+\alpha, n-k+\beta)$$

$$\mathrm{Beta}(\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$E[\theta|n,k] = \int_0^1 \theta p(\theta|n,k)d\theta = \frac{k+\alpha}{n+\alpha+\beta}$$

Varianza *a posteriori*

$$\mathrm{var}(\theta|n,k) = \frac{\mathrm{E}(\theta|n,k)(1-\mathrm{E}(\theta|n,k))}{\alpha+\beta+n+1}$$

Cuando k y $n-k$ crecen con α y β fijos,

$$\mathrm{E}(\theta|k,n) \approx k/n \qquad \mathrm{var}(\theta|k,n) \approx \frac{1}{n}\frac{k}{n}\left(1-\frac{k}{n}\right) \to 0$$

Posterior predictiva

$$p(\theta|n,k) = \mathrm{Beta}(k+\alpha, n-k+\beta)$$

$$p(0|n,k) = \int_0^1 p(0|\theta)p(\theta|n,k)d\theta = \frac{k+\alpha}{n+\alpha+\beta}$$

Posterior predictiva

$$p(\theta|n, k) = \mathrm{Beta}(k + \alpha, n - k + \beta)$$

$$p(0|n, k) = \int_0^1 p(0|\theta)p(\theta|n, k)d\theta = \frac{k + \alpha}{n + \alpha + \beta}$$

(¿Por qué vale esto?)

Posterior predictiva

$$p(\theta|n,k) = \mathrm{Beta}(k+\alpha, n-k+\beta)$$

$$p(0|n,k) = \int_0^1 p(0|\theta)p(\theta|n,k)d\theta = \frac{k+\alpha}{n+\alpha+\beta}$$

(¿Por qué vale esto?) $\mathbb{P}$

Posterior predictiva

$$p(\theta|n,k) = \mathrm{Beta}(k+\alpha, n-k+\beta)$$

$$p(0|n,k) = \int_0^1 p(0|\theta)p(\theta|n,k)d\theta = \frac{k+\alpha}{n+\alpha+\beta}$$

(¿Por qué vale esto?) $\mathbb{P}$

$$k = n = 6$$

Posterior predictiva

$$p(\theta|n,k) = \mathrm{Beta}(k+\alpha, n-k+\beta)$$

$$p(0|n,k) = \int_0^1 p(0|\theta)p(\theta|n,k)d\theta = \frac{k+\alpha}{n+\alpha+\beta}$$

(¿Por qué vale esto?) $\mathbb{P}$

$$k = n = 6$$

$$\alpha = 1, \beta = 1$$

$$p(0|6,6) = 7/8 = 0.875$$

Posterior predictiva

$$p(\theta|n,k) = \mathrm{Beta}(k+\alpha, n-k+\beta)$$

$$p(0|n,k) = \int_0^1 p(0|\theta)p(\theta|n,k)d\theta = \frac{k+\alpha}{n+\alpha+\beta}$$

(¿Por qué vale esto?) $\mathbb{P}$

$$k = n = 6$$

$$\alpha = 1, \beta = 1$$

$$p(0|6,6) = 7/8 = 0.875$$

$$\alpha = 100, \beta = 100$$

$$p(0|6,6) = 106/206 \simeq 0.51$$

Dependencia fuerte sobre el *prior*

de dónde viene: experiencia previa (asumimos que las monedas son parecidas, vimos otras monedas)

Dependencia fuerte sobre el *prior*

de dónde viene: experiencia previa (asumimos que las monedas son parecidas, vimos otras monedas)

pero...

Dependencia fuerte sobre el *prior*

de dónde viene: experiencia previa (asumimos que las monedas son parecidas, vimos otras monedas)

pero...

- no vimos 200 tiradas de moneda

Dependencia fuerte sobre el *prior*

de dónde viene: experiencia previa (asumimos que las monedas son parecidas, vimos otras monedas)

pero...

- no vimos 200 tiradas de moneda
  - -conocimiento previo más *fuerte* que la experiencia

Dependencia fuerte sobre el *prior*

de dónde viene: experiencia previa (asumimos que las monedas son parecidas, vimos otras monedas)

pero...

- no vimos 200 tiradas de moneda
  - -conocimiento previo más *fuerte* que la experiencia
- ni fueron 100 y 100

Dependencia fuerte sobre el *prior*

de dónde viene: experiencia previa (asumimos que las monedas son parecidas, vimos otras monedas)

pero...

- no vimos 200 tiradas de moneda
  - -conocimiento previo más *fuerte* que la experiencia
- ni fueron 100 y 100
  - -conocimiento previo más *suave* que la experiencia

Dependencia fuerte sobre el *prior*

de dónde viene: experiencia previa (asumimos que las monedas son parecidas, vimos otras monedas)

pero...

- no vimos 200 tiradas de moneda
  - conocimiento previo más *fuerte* que la experiencia
- ni fueron 100 y 100
  - conocimiento previo más *suave* que la experiencia
- no es lo mismo ver 200 de una moneda que 20 de 10 distintas

Dependencia fuerte sobre el *prior*

de dónde viene: experiencia previa (asumimos que las monedas son parecidas, vimos otras monedas)

pero...

- no vimos 200 tiradas de moneda
  - -conocimiento previo más *fuerte* que la experiencia
- ni fueron 100 y 100
  - -conocimiento previo más *suave* que la experiencia
- no es lo mismo ver 200 de una moneda que 20 de 10 distintas
  - -conocimiento previo más *estructurado* que la experiencia

Dependencia fuerte sobre el *prior*

de dónde viene: experiencia previa (asumimos que las monedas son parecidas, vimos otras monedas)

pero...

- no vimos 200 tiradas de moneda
  - -conocimiento previo más *fuerte* que la experiencia
- ni fueron 100 y 100
  - -conocimiento previo más *suave* que la experiencia
- no es lo mismo ver 200 de una moneda que 20 de 10 distintas
  - -conocimiento previo más *estructurado* que la experiencia

Teoría: monedas manufacturadas por un proceso estandarizado

Teoría: monedas manufacturadas por un proceso estandarizado

Teoría: monedas manufacturadas por un proceso estandarizado

-Justifica generalizar de otras monedas

Teoría: monedas manufacturadas por un proceso estandarizado

-Justifica generalizar de otras monedas

-Justifica *priors* más fuertes y suaves

Teoría: monedas manufacturadas por un proceso estandarizado

-Justifica generalizar de otras monedas

-Justifica *priors* más fuertes y suaves

-Explica por qué 10 tiradas de 20 monedas es mejor que 200 de una sola

Teoría: monedas manufacturadas por un proceso estandarizado

-Justifica generalizar de otras monedas

-Justifica *priors* más fuertes y suaves

-Explica por qué 10 tiradas de 20 monedas es mejor que 200 de una sola

Limitaciones:

Teoría: monedas manufacturadas por un proceso estandarizado

-Justifica generalizar de otras monedas

-Justifica *priors* más fuertes y suaves

-Explica por qué 10 tiradas de 20 monedas es mejor que 200 de una sola

Limitaciones:

-¿Podemos representar cualquier tipo de conocimiento como un número de observaciones ficticias?

Teoría: monedas manufacturadas por un proceso estandarizado

-Justifica generalizar de otras monedas

-Justifica *priors* más fuertes y suaves

-Explica por qué 10 tiradas de 20 monedas es mejor que 200 de una sola

Limitaciones:

-¿Podemos representar cualquier tipo de conocimiento como un número de observaciones ficticias?

-Si tiramos 25 veces la moneda y sale 25 veces cara.. raro

Teoría: monedas manufacturadas por un proceso estandarizado

-Justifica generalizar de otras monedas

-Justifica *priors* más fuertes y suaves

-Explica por qué 10 tiradas de 20 monedas es mejor que 200 de una sola

Limitaciones:

-¿Podemos representar cualquier tipo de conocimiento como un número de observaciones ficticias?

-Si tiramos 25 veces la moneda y sale 25 veces cara.. raro

-Pero con el prior de 100 y 100 que usamos obtenemos:

$$p(0|25,25) = 125/225 \simeq 0.56$$ ¡no tan raro!

Teoría: monedas manufacturadas por un proceso estandarizado

-Justifica generalizar de otras monedas

-Justifica *priors* más fuertes y suaves

-Explica por qué 10 tiradas de 20 monedas es mejor que 200 de una sola

Limitaciones:

-¿Podemos representar cualquier tipo de conocimiento como un número de observaciones ficticias?

-Si tiramos 25 veces la moneda y sale 25 veces cara.. raro

-Pero con el prior de 100 y 100 que usamos obtenemos:

$$p(0|25,25) = 125/225 \simeq 0.56$$ ¡no tan raro!

...¡Modelos jerárquicos!