Fundamentos de Inferencia Bayesiana

“La moneda”
Comparación de Modelos

¡Volvemos a la moneda!

$$D = 000000$$

$$D = 010010$$

¿Qué proceso produjo estas secuencias?

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

$$D = 010010$$

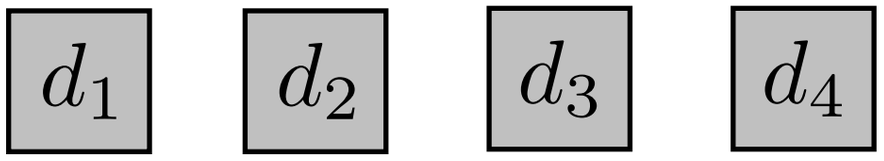
- Moneda común: p(0) = 0.5
- Moneda cargada: p(0) = θ
- Modelo de Markov
- Hidden Markov Model (HMM)

Algunos Modelos Generativos

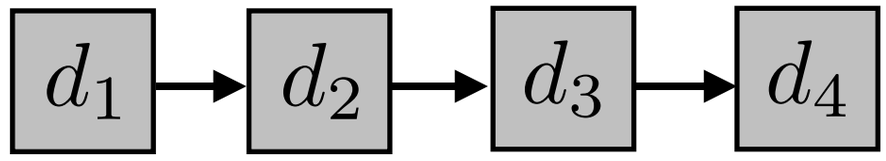
$$D = 010010$$

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

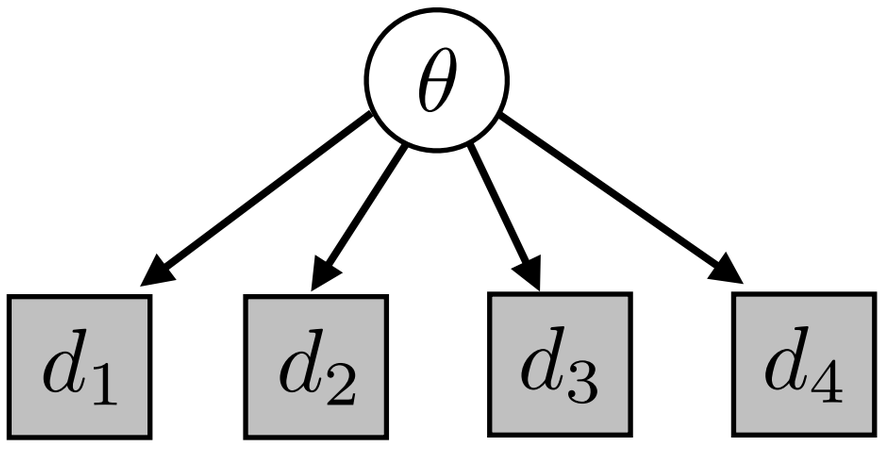
Moneda común



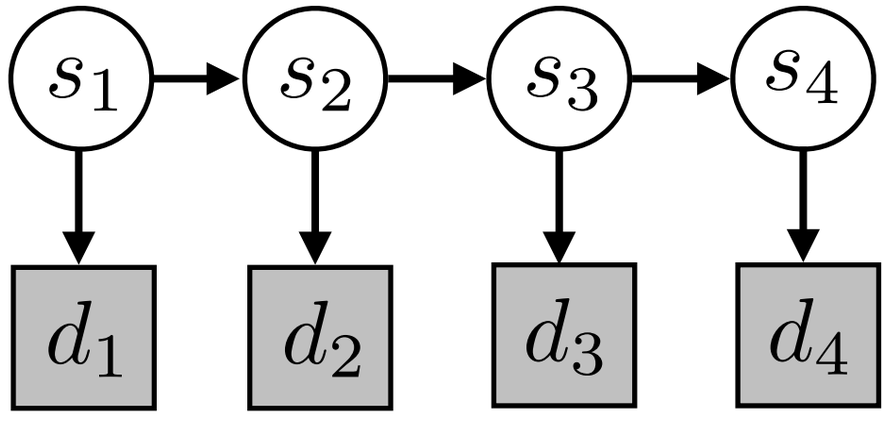
Modelo de *Markov*



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"

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

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

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

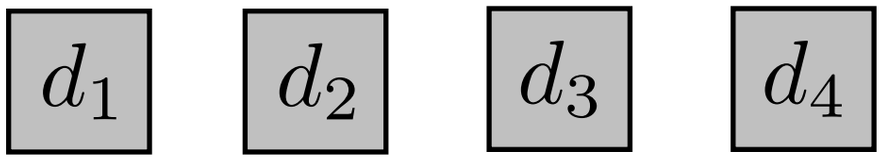
$D = 010010$	$D = 000000$	$D = 0000000000$
$p(D\|H_1) = 1/2^6$	$p(D\|H_1) = 1/2^6$	$p(D\|H_1) = 1/2^{10}$
$p(D\|H_2) = 0$	$p(D\|H_2) = 1$	$p(D\|H_2) = 1$
$\Downarrow$	$\Downarrow$	$\Downarrow$
$\frac{p(H_1\|D)}{p(H_2\|D)} = \infty$	$\frac{p(H_1\|D)}{p(H_2\|D)} \simeq 16$	$\frac{p(H_1\|D)}{p(H_2\|D)} \simeq 1$

Combina conocimiento previo con evidencia

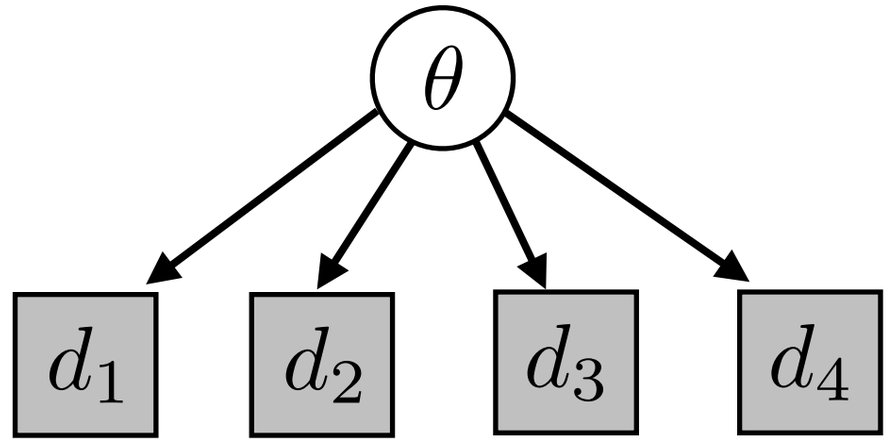
Comparación de una hipótesis simple con una compleja

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

Moneda común



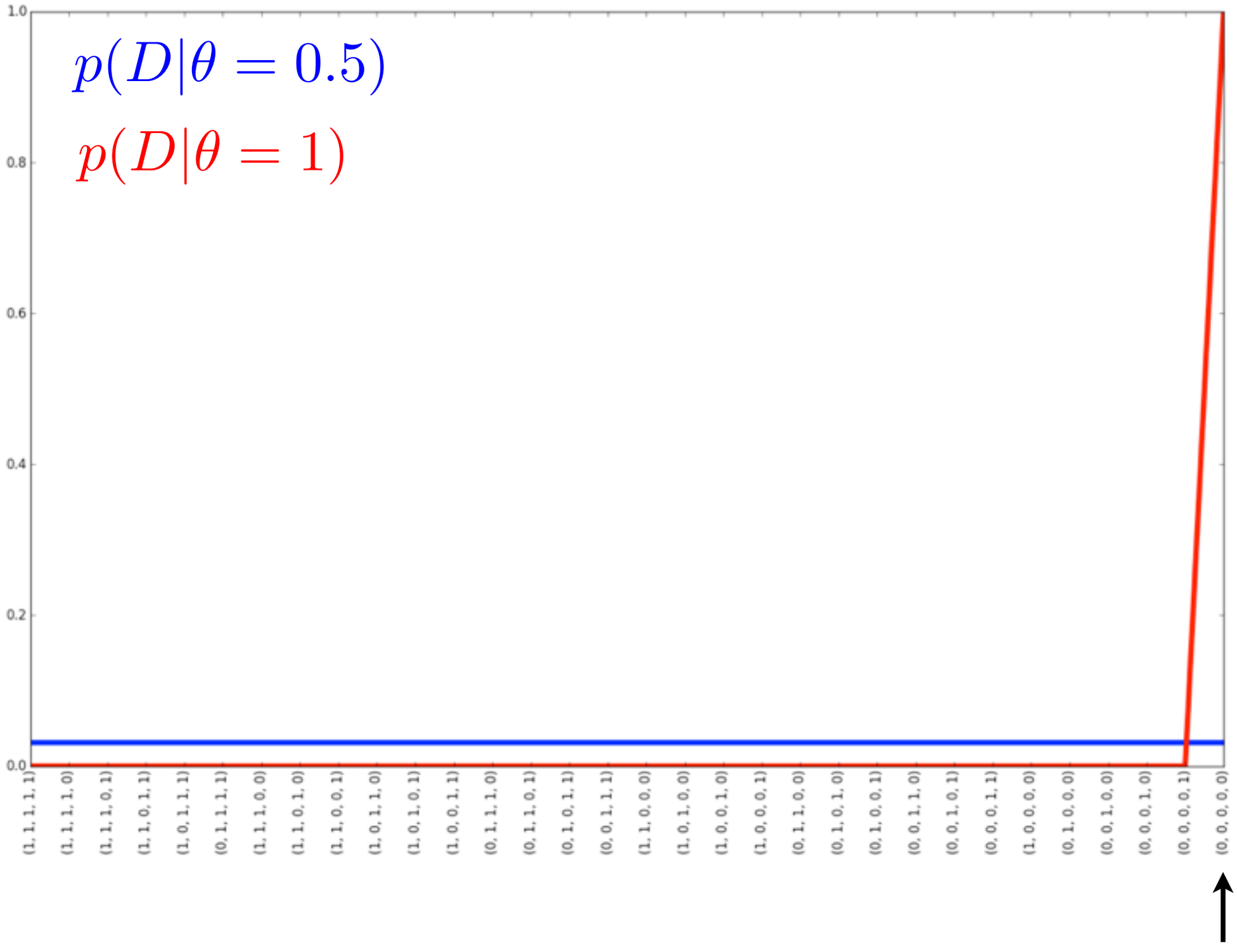
Moneda cargada



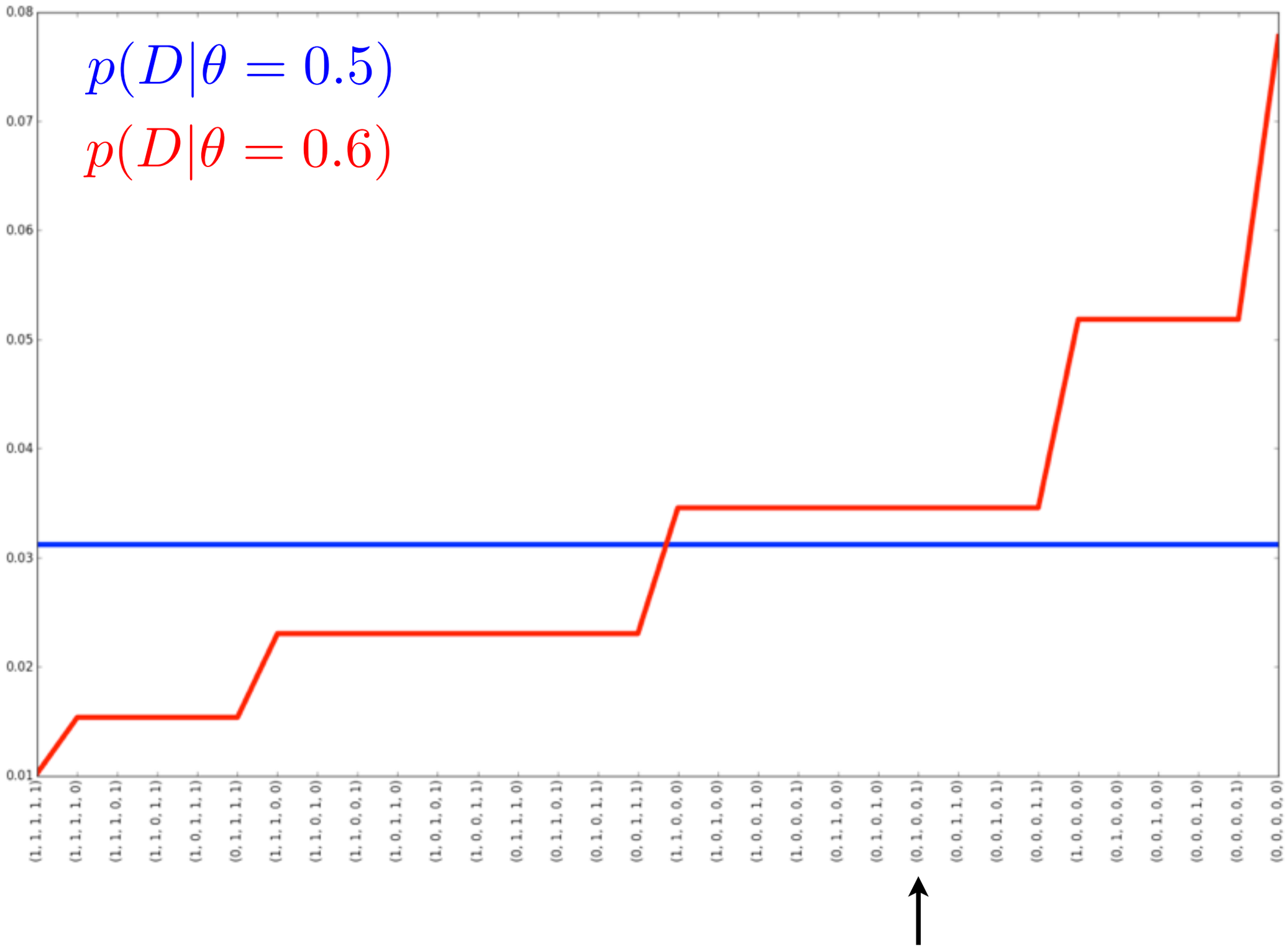
H_2 es más compleja:

- *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, 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)

$$D = 01001$$


$p(D|\theta = 0.5)$
$p(D|\theta = 0.6)$
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?

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

H_1: $p(0) = 0.5$

H_2: $p(0) = \theta$

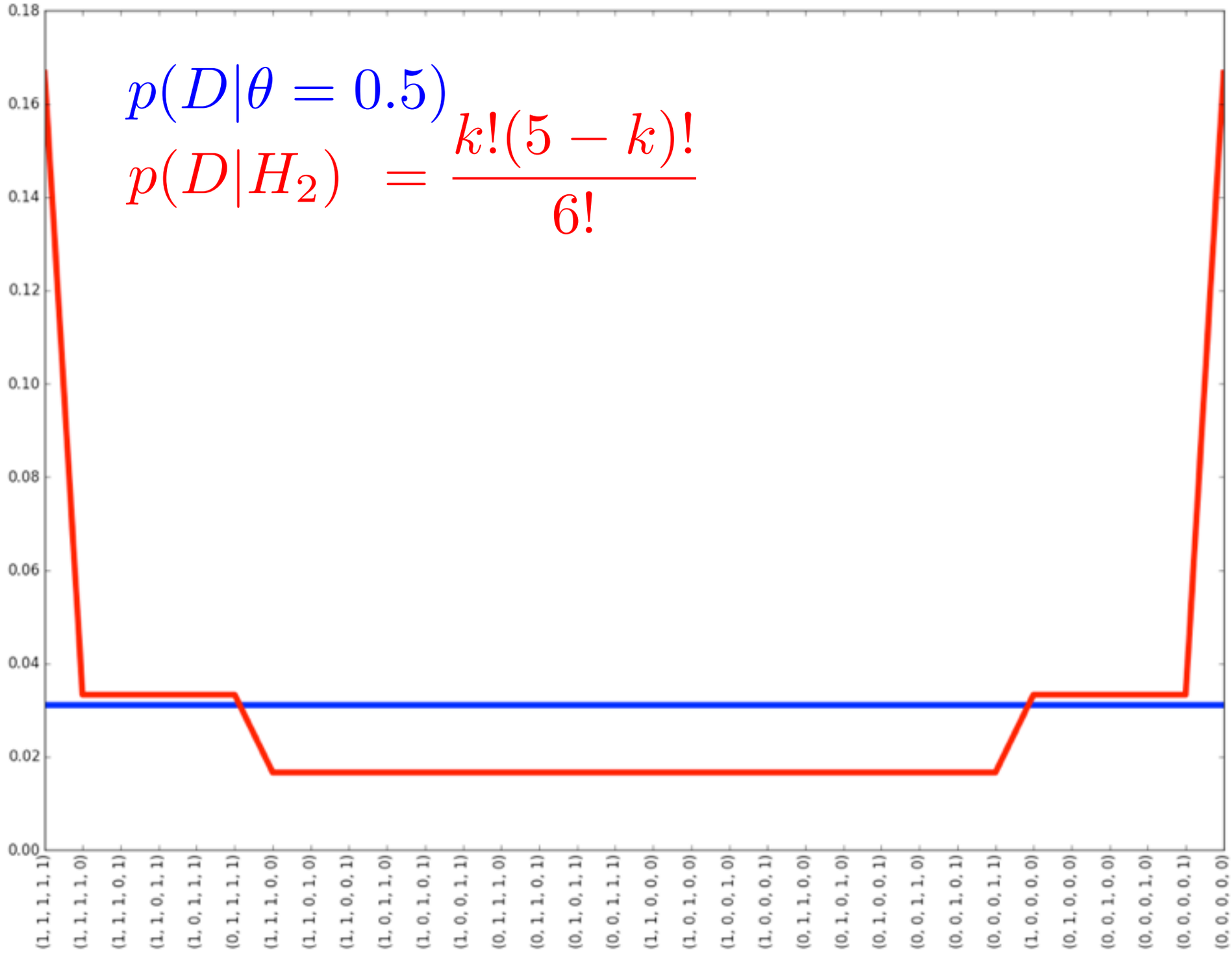
$$\underline{p(H_1|D)} \;=\; \underline{p(D|H_1)}\;\underline{p(H_1)}$$
$$p(H_2|D) \qquad p(D|H_2)\,p(H_2)$$

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

$$p(D|H_2) = \int_0^1 \overbrace{p(D|\theta)}^{\theta^k(1-\theta)^{N-k}} \overbrace{p(\theta|H_2)}^{1\text{(uniforme)}} d\theta$$

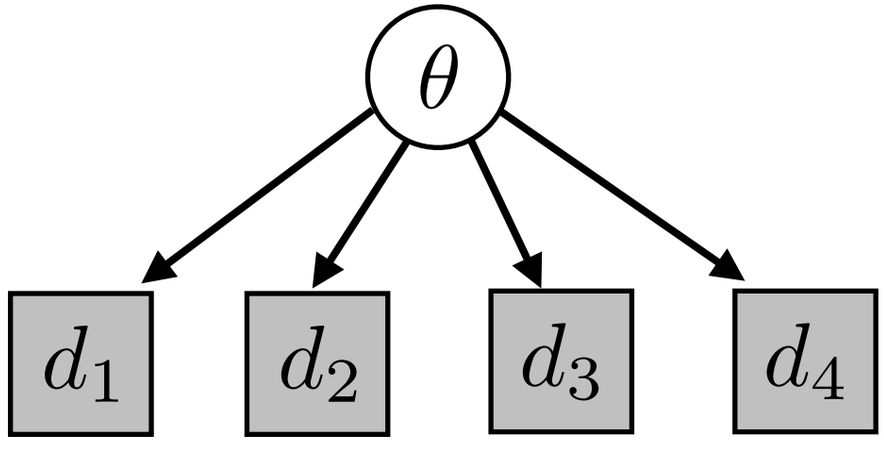
promediamos sobre θ

$$= \int_0^1 \theta^k (1-\theta)^{N-k} d\theta = \mathrm{B}(k+1, n-k+1) = \frac{k!(N-k)!}{(N+1)!}$$



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*

Volvemos al “experimento” inicial...

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

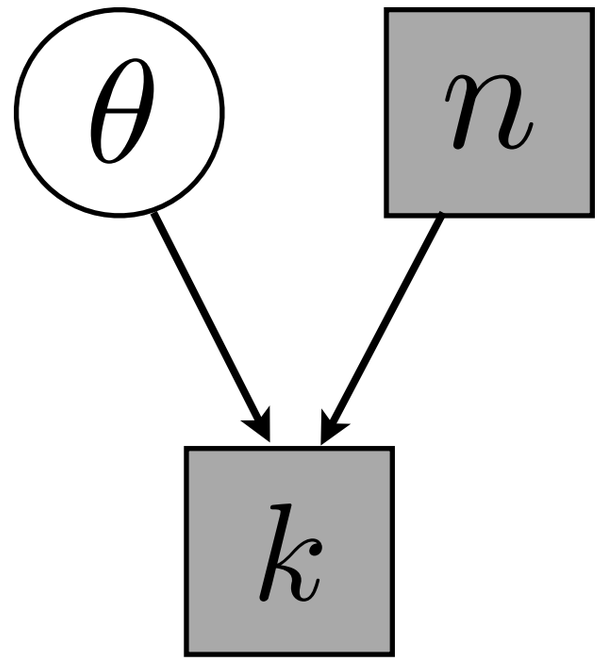
¿Y ahora? ¡¿1?!

Modelo

likelihood:

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

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



prior:

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

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

posterior:

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

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

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,

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

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

posterior — *likelihood* — *prior*

prior predictive

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

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!*

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?)

$$k = n = 6$$

$$\alpha = 1, \beta = 1$$
$$p(0|6,6) = 7/8 = 0.875$$

$$\alpha = 100, \beta = 100$$
$$p(0|6,6) = 106/206 \simeq 0.51$$

MODELO

Creencias

prior

Mecanismo

likelihood

variables *latentes*

DATOS

Explicación

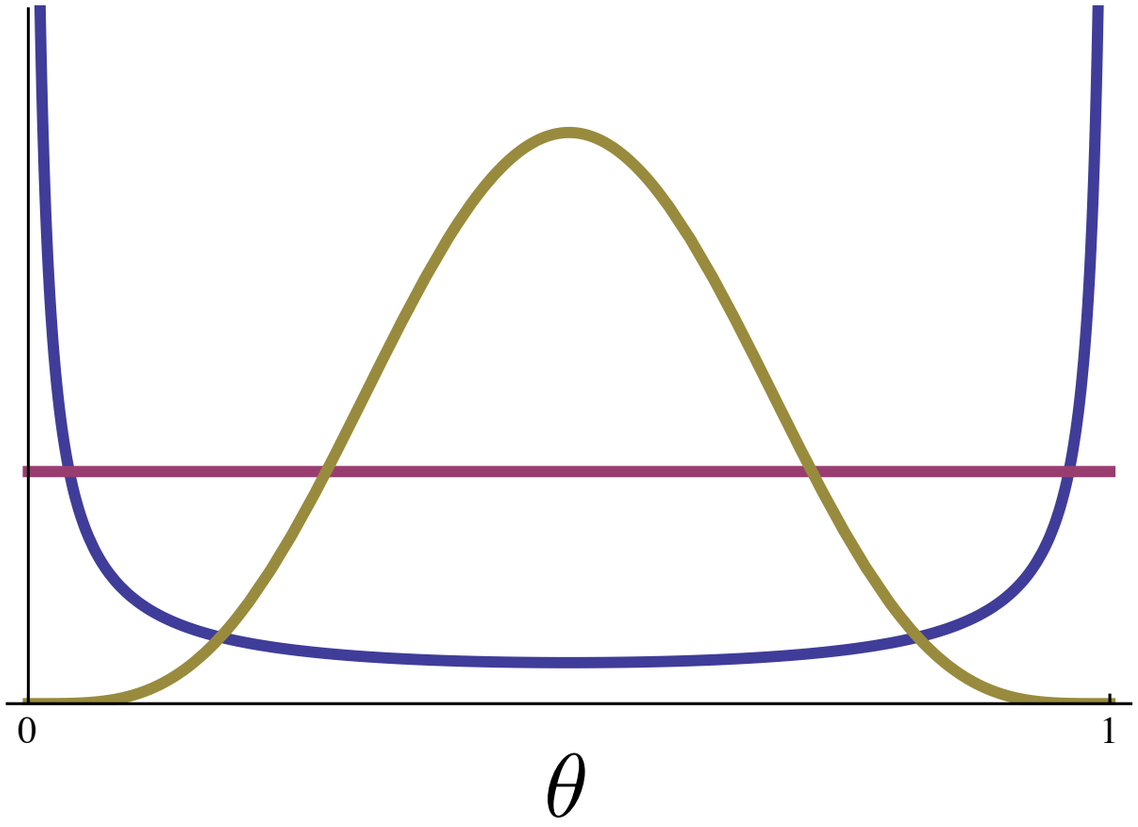
posterior

Predicción

posterior predictive

variables *observables*

Vuelta al *prior…* Codificando nuestra ignorancia



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

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

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

Dependencia fuerte sobre el *prior*

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

pero...

- no vimos tantas (200) tiradas de moneda
  -conocimiento previo más *fuerte* que la experiencia real

- ni fueron perfectamente balanceadas (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

-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!

Práctica

Wagenmakers & Lee: 3.1, 3.2, 3.3