

COGNICIÓN BAYESIANA

Santiago Alonso-Díaz, PhD

Universidad Javeriana

POSTERIORS, PRIORS, & LIKELIHOODS

Prior (creencias)



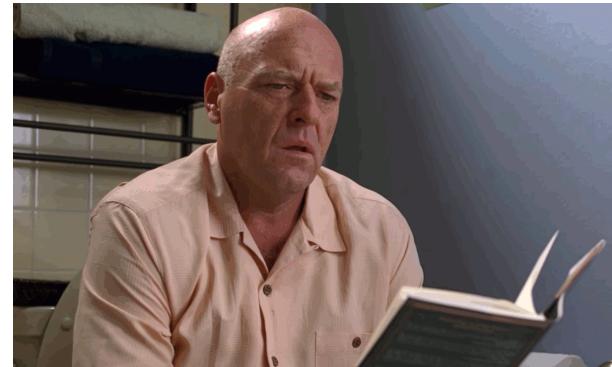
Likelihood (data)



Prior (creencias)



Likelihood (data)



¿EXTRATERRESTRES? ¿LIKELIHOOD? ¿PRIOR?



Bayes formalizó como combinar priors y likelihoods

$$Posterior = \frac{Prior \times Likelihood}{Marginal}$$

Veamos la misma formula con otros "nombres"

$$Belief_{t+1} = \frac{Belief_t \times Likelihood}{Marginal}$$

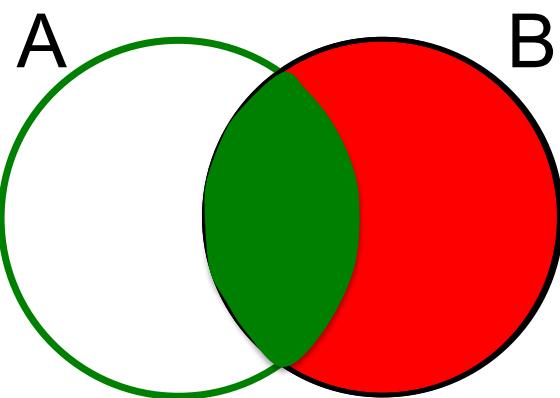
Veamos la misma formula expandida

$$p(Hypothesis|Data) = \frac{p(Hypothesis) \times p(Data|Hypothesis)}{p(Data)}$$

PRUEBA DEL TEOREMA

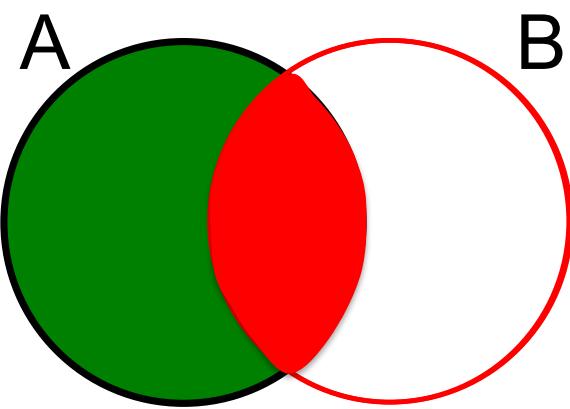
$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$

$$= \frac{\text{---} \text{ green } \text{ ---}}{\text{---} \text{ red } \text{ ---}}$$



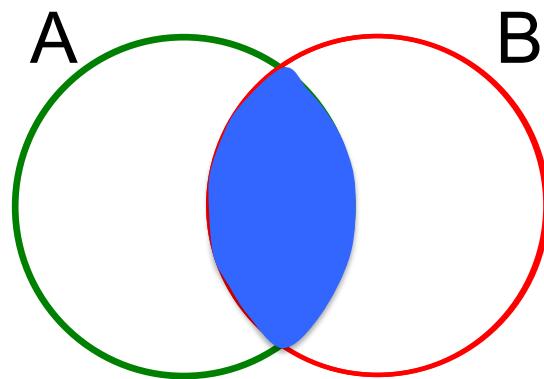
$$p(B|A) = \frac{p(A \cap B)}{p(A)}$$

$$= \frac{\text{Red Area}}{\text{Green Area}}$$



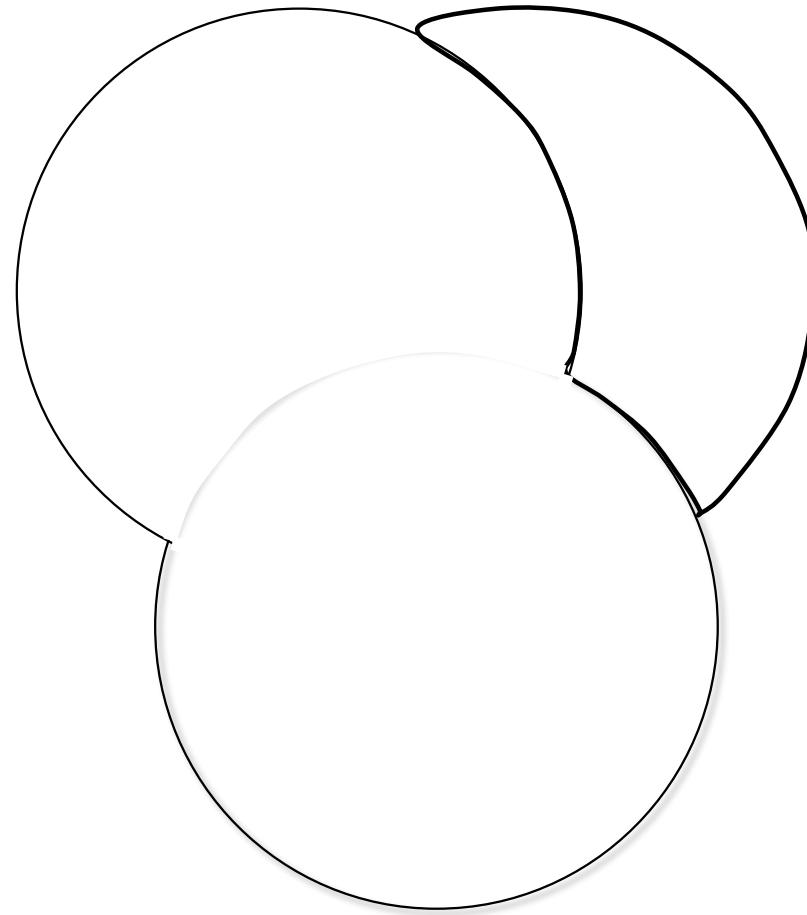
$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$

$$p(B|A) = \frac{p(A \cap B)}{p(A)}$$

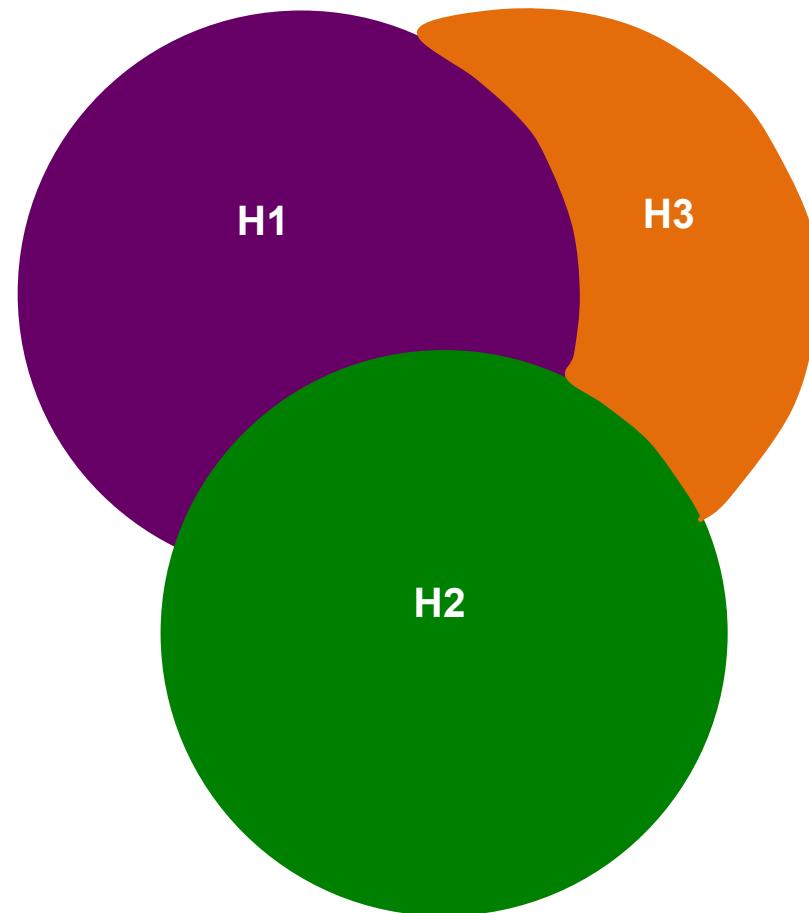


VISUALIZACIÓN DE POSTERIORS, PRIORS, Y LIKELIHOODS

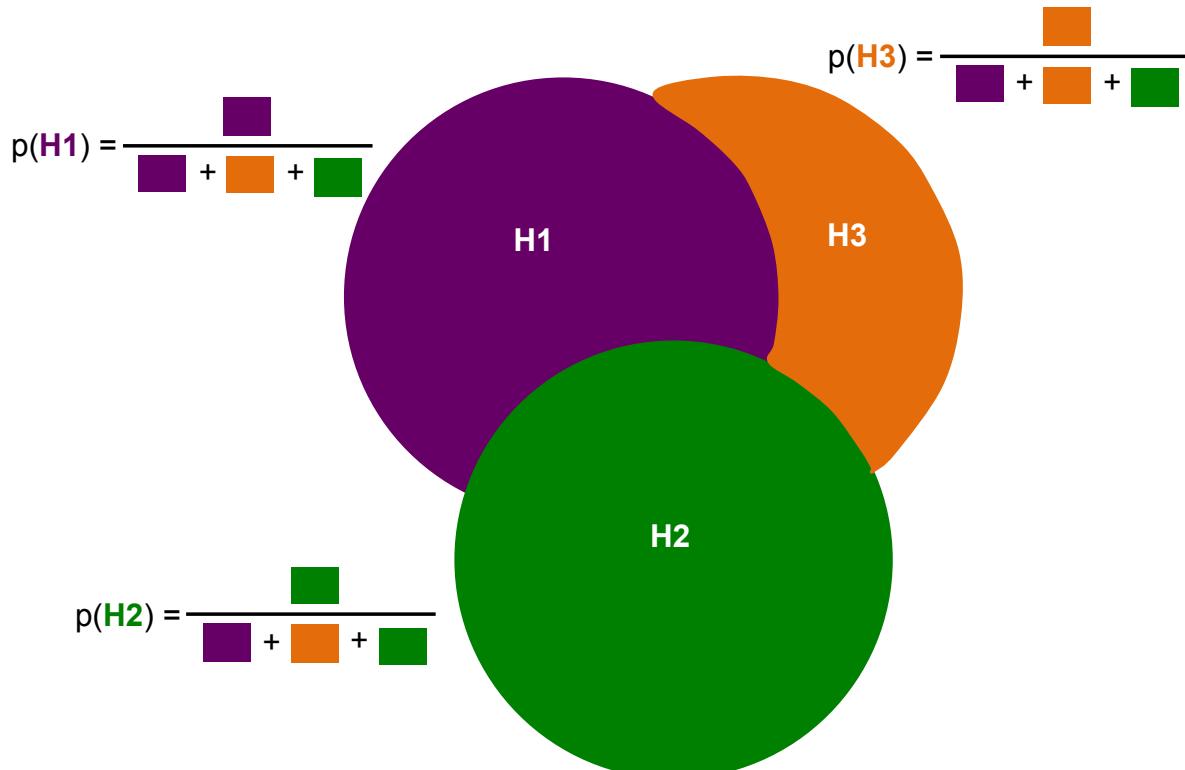
ESPACIO DE HIPÓTESIS (EVENTOS)



ESPACIO DE HIPÓTESIS (EVENTOS)

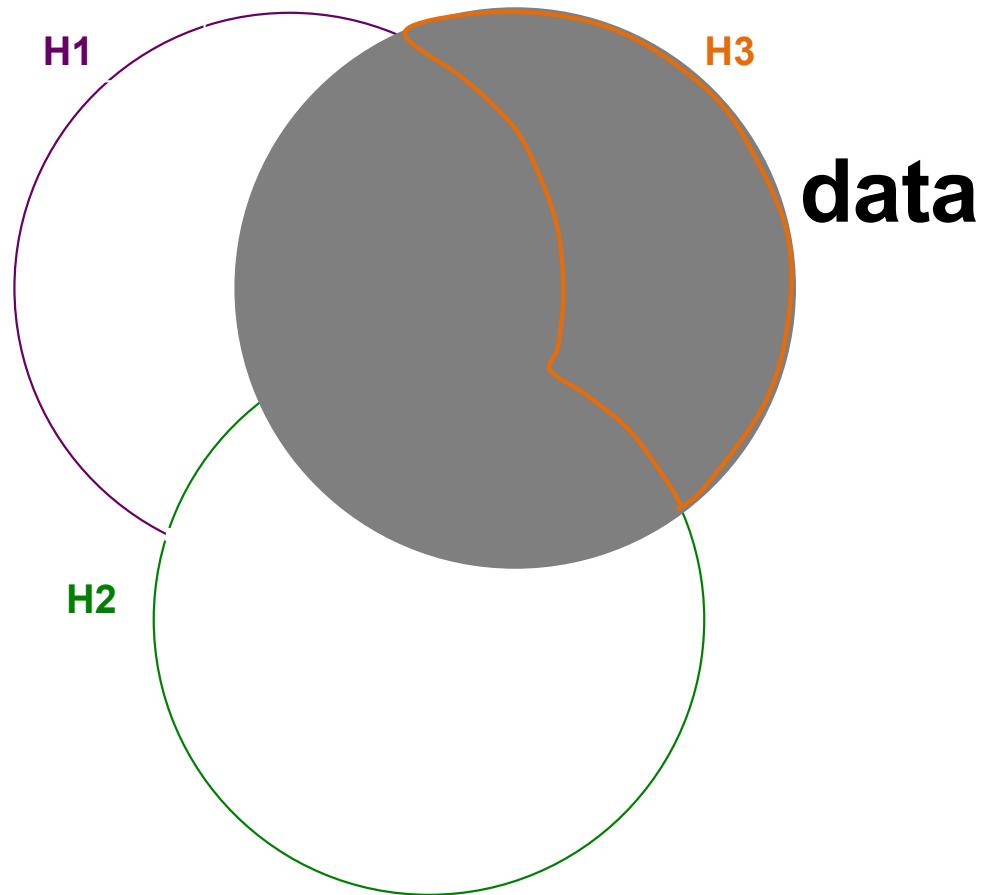


PRIORS

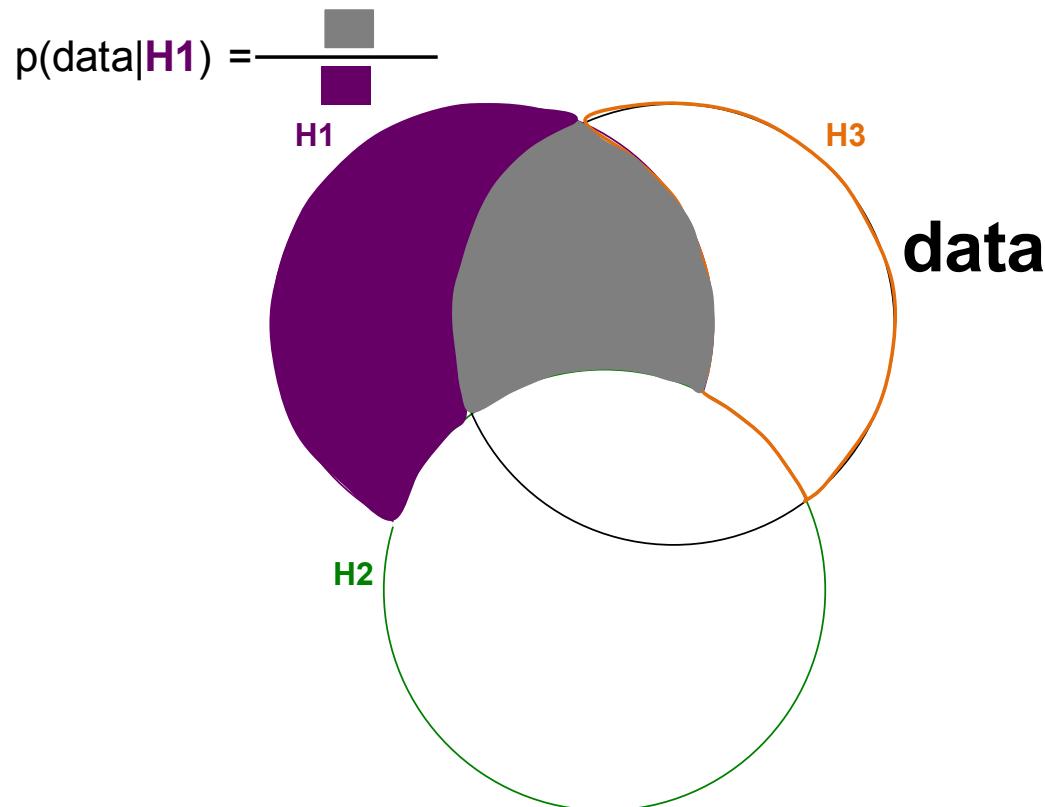


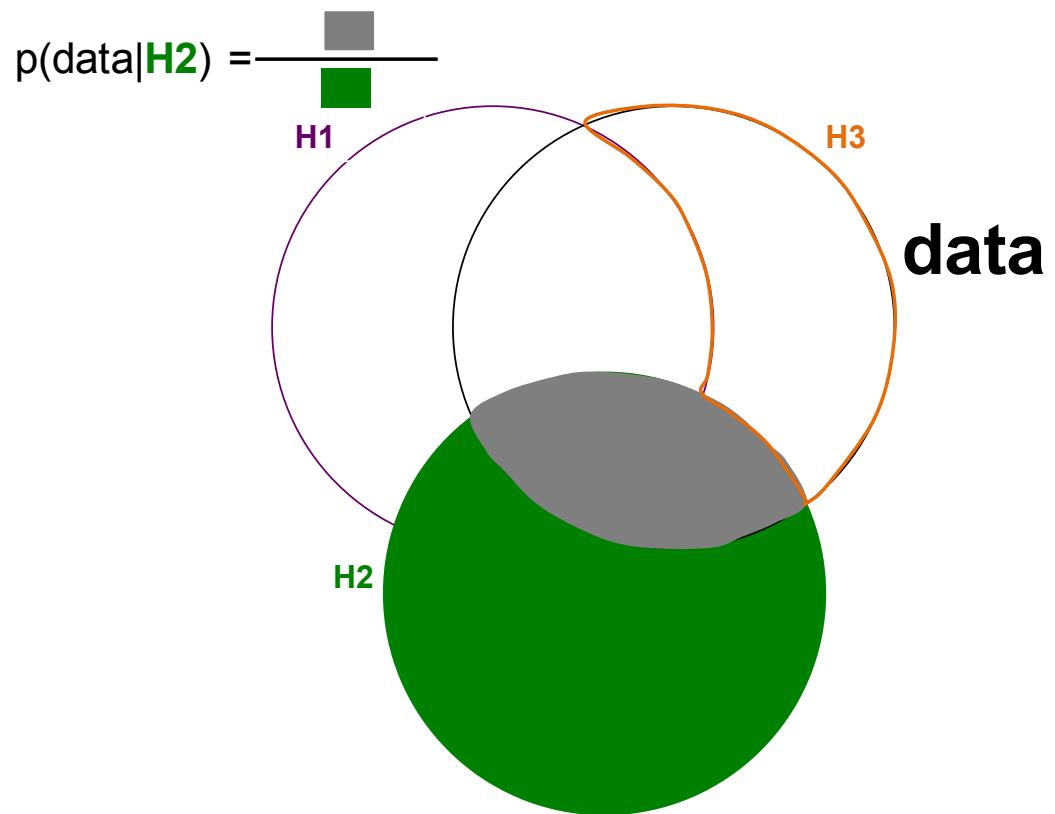
DATOS

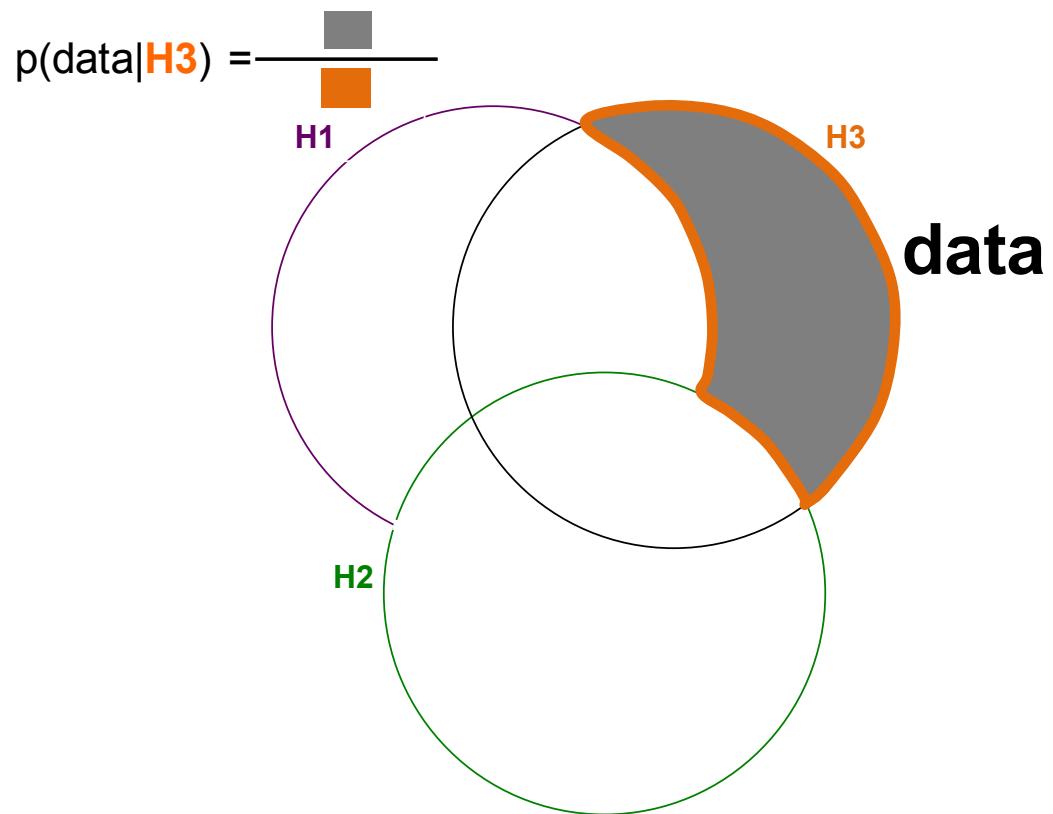
EL MARGINAL ES LA PROBABILIDAD DE LA DATA (GRIS DIVIDIDO TODA EL ÁREA DEL ESPACIO DE HIPÓTESIS)



LIKELIHOODS

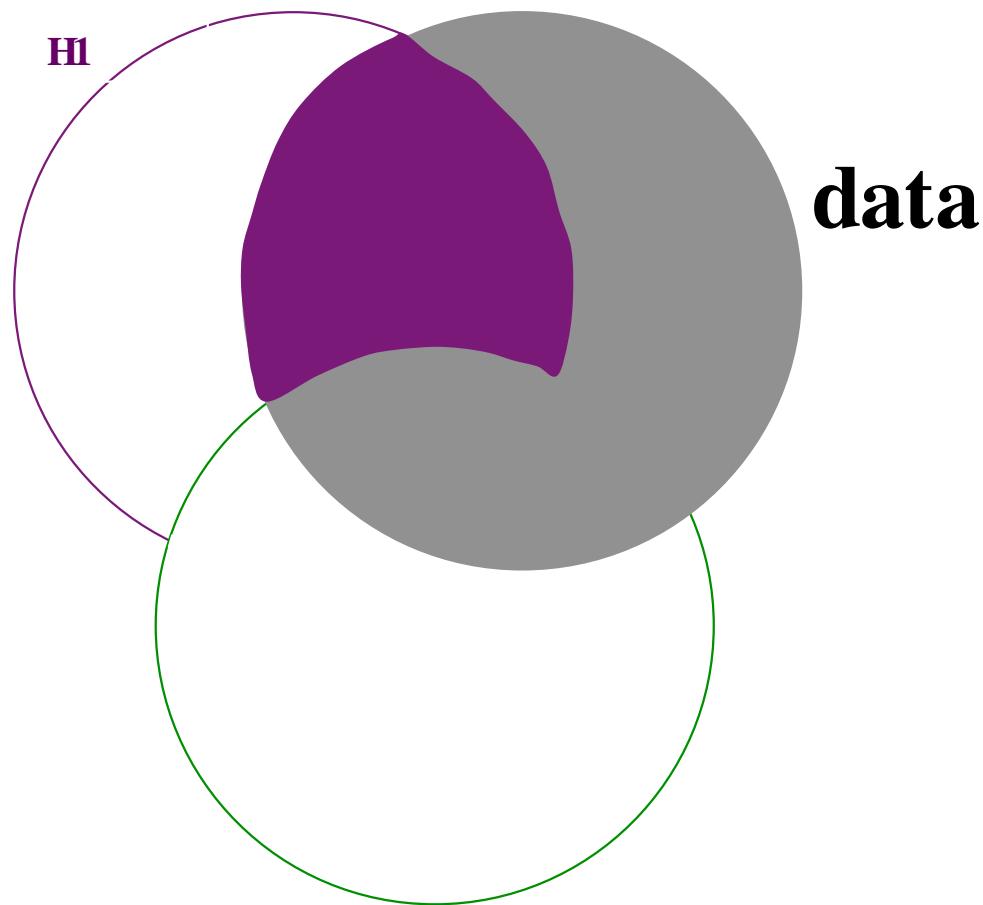






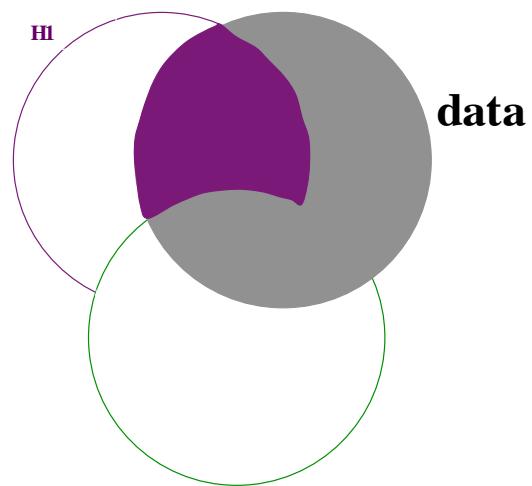
POSTERIOR

Posterior = $p(\text{H1}|\text{data})$

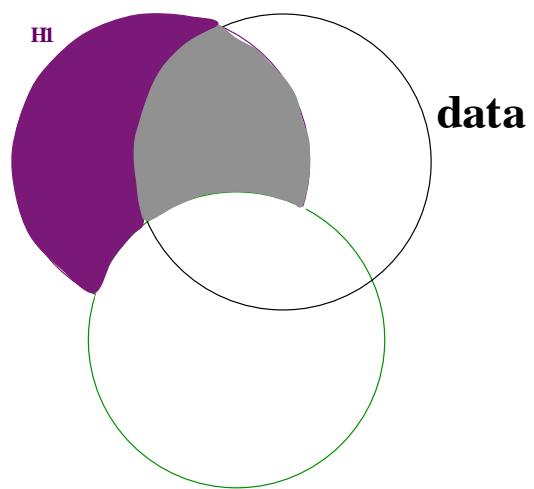


POSTERIOR VS. LIKELIHOOD

Posterior = $p(\text{H1}|\text{data})$



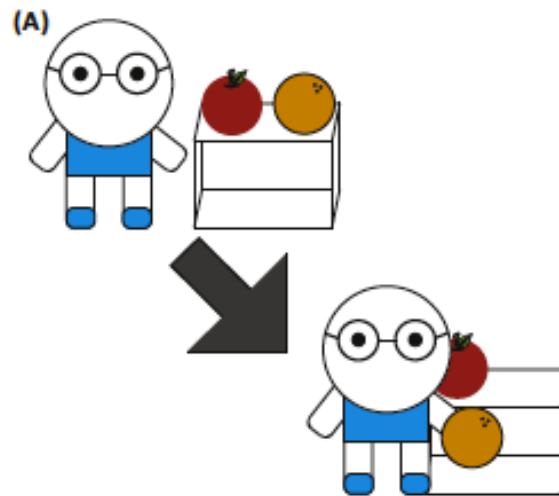
Likelihood = $p(\text{data}|\text{H1})$



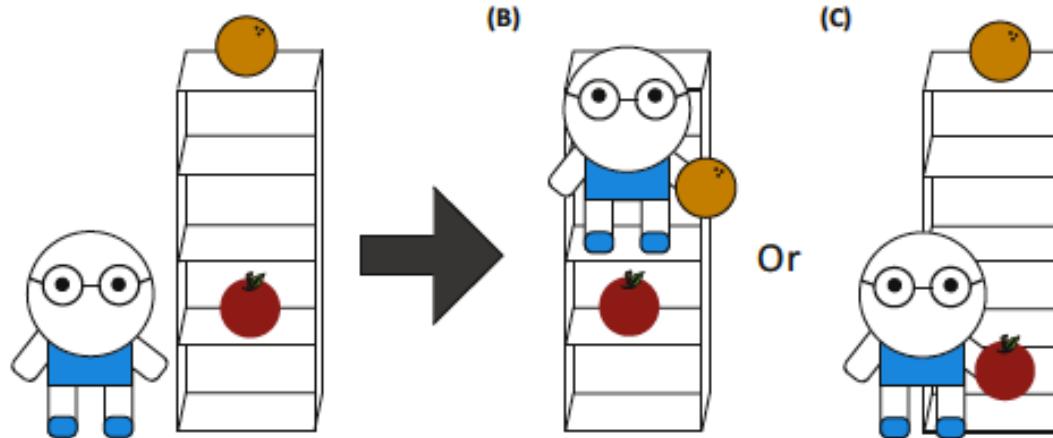
COGNICIÓN BAYESIANA:

COMBINAR DATOS Y CREENCIAS CON EL TEOREMA

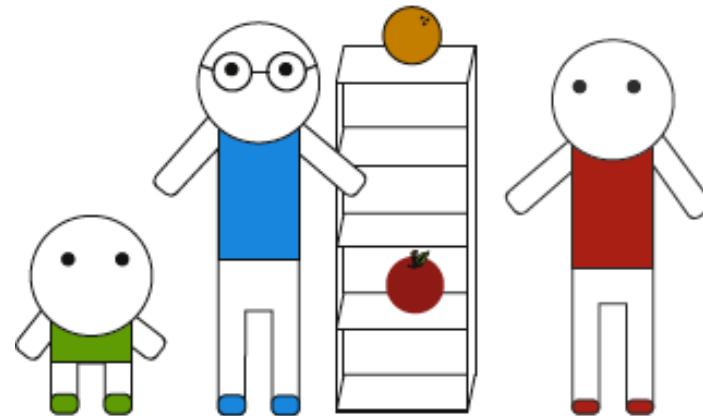
¿Qué prefiere **azulito**?



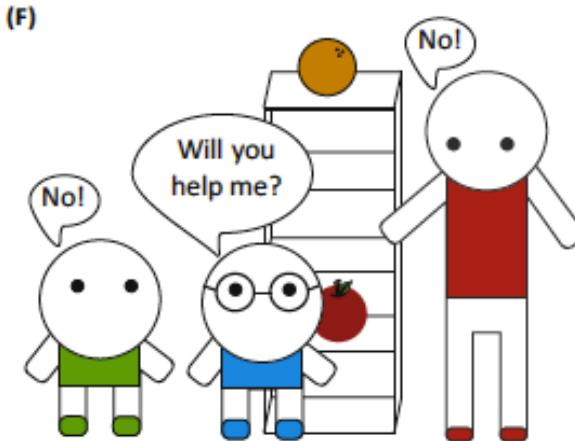
¿Qué prefiere **azulito**?



Azulito pide ayuda ¿Pereza?



¿Quién es más malo? ¿verde? ¿rojo?



Inferencia sobre estados latentes de **azulito** (e.g. preferencias) con el teorema de Bayes

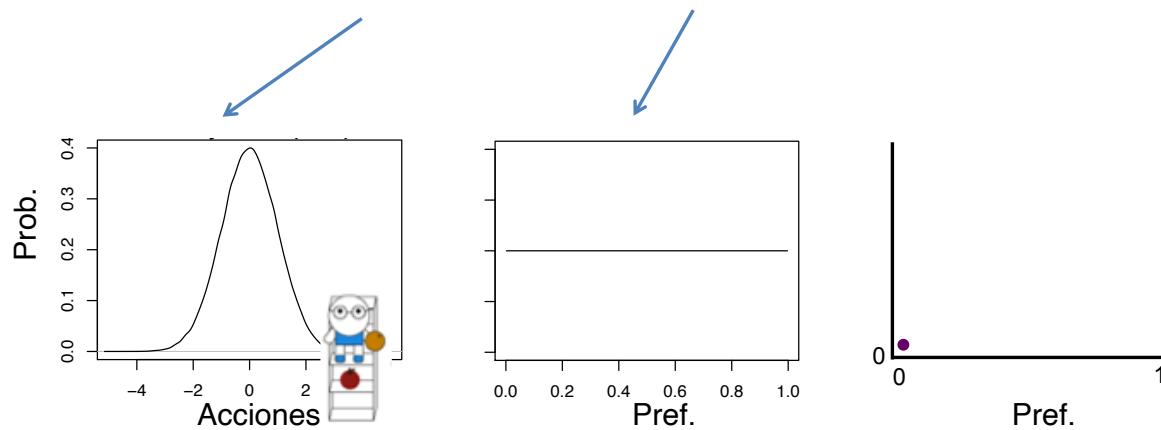
$$p(\text{Preferencia}|\text{Accion}) \propto p(\text{Accion}|\text{Preferencia})p(\text{Preferencia})$$

Por ejemplo, likelihood normal y prior uniforme

Hipótesis 1. Pref. = 0

$$0.01 \quad 0.2$$

$$p(\text{Acciones}|\text{Pref}) \cdots p(\text{Pref}) \propto p(\text{Pref}|\text{Acciones})$$



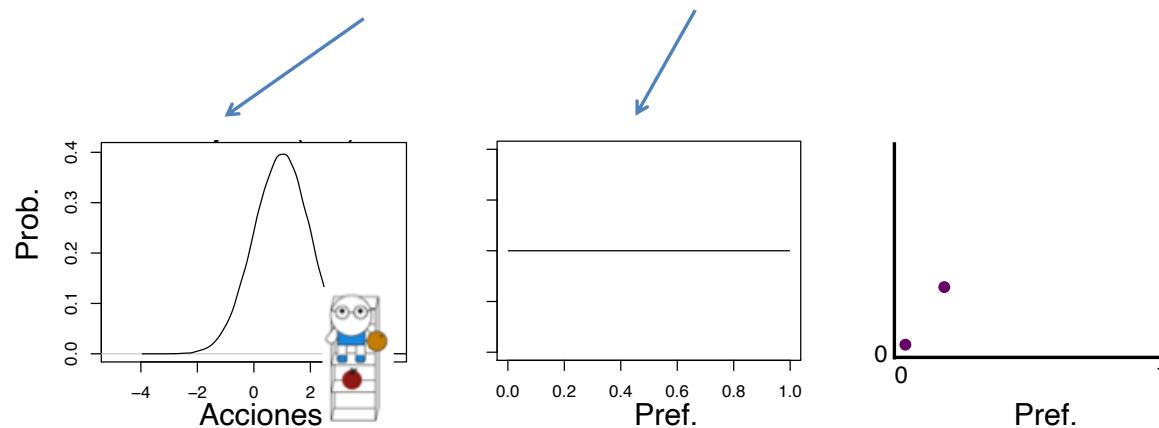
Hipótesis 2. Pref. = 0.2

0.1

$$p(\text{Acciones}|\text{Pref}) \cdots p(\text{Pref}) = p(\text{Pref}|\text{Acciones})$$

0.2

0.02



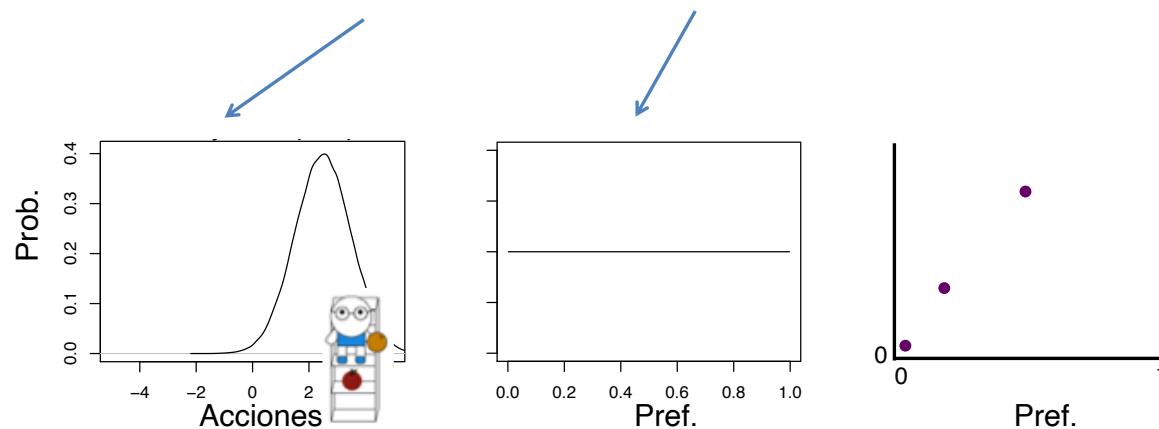
Hipótesis 3. Pref. = 0.5

0.4

$$p(\text{Acciones}|\text{Pref}) \cdots p(\text{Pref}) = p(\text{Pref}|\text{Acciones})$$

0.2

0.08



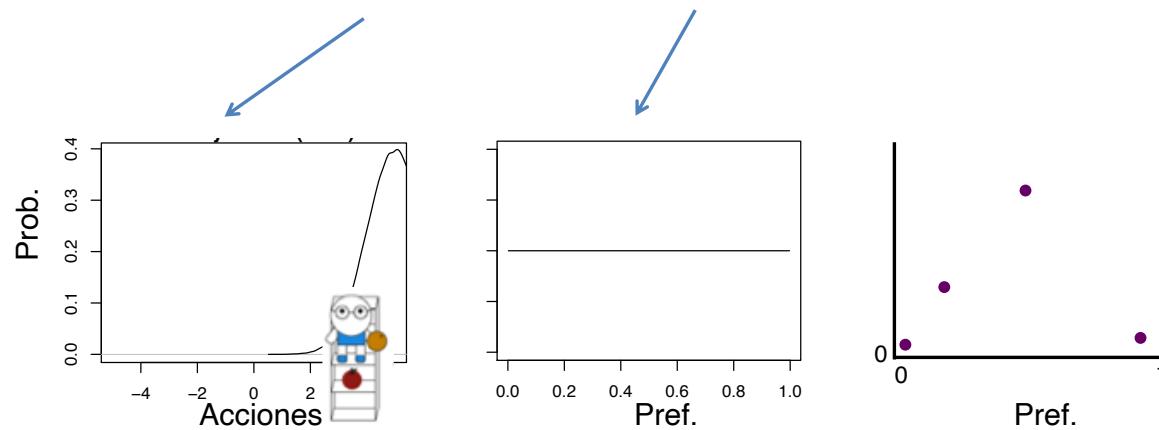
Hipótesis 4. Pref. = 0.8

0.05

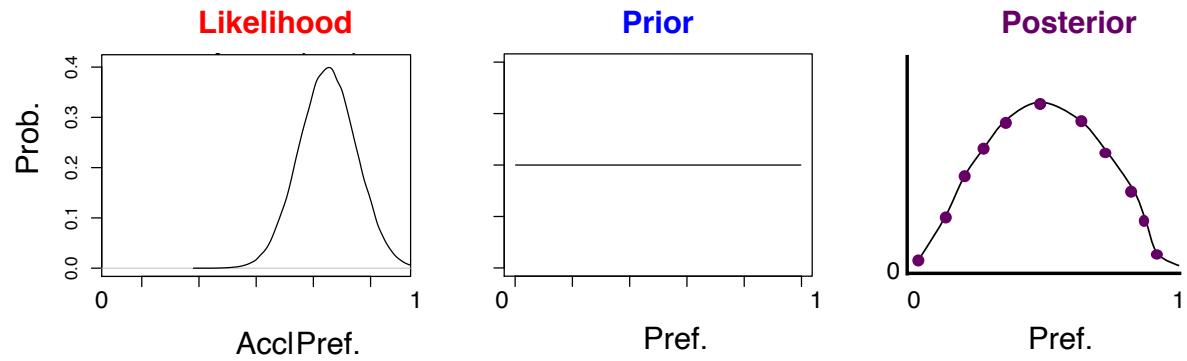
$p(\text{Acciones}|\text{Pref})$... $p(\text{Pref}) = p(\text{Pref}|\text{Acciones})$

0.2

0.01



Hipótesis TODAS. CR = [0,1]



DOS FORMAS PARA OBTENER LA POSTERIOR

ANÁLISIS MATEMÁTICO

SIMULACIONES (E.G. MCMC)

FORMA 1: ANÁLISIS MATEMÁTICO.

CASO: LIKELIHOOD & PRIOR NORMAL

Descripción de un problema estadístico simple:

1. Una sola observación X para justificar una hipótesis llamada H .
2. ¿Qué tanto creemos en H dada la observación X ?

Solución en lenguaje natural:

Creencia en H es proporcional a la probabilidad de que X ocurra si H es verdad. Todo ponderado por mi creencia previa en H

En lenguaje de Bayes:

Posterior \propto *Likelihood* \times *Prior*

$$p(H|X) \propto p(X|H) \times p(H)$$

En lenguaje de distribuciones de probabilidad:

$$p(H|X) \propto Normal(X|\mu = H, \sigma^2) \times Normal(H|\mu = H_0, \sigma_0^2)$$

Nota: conocemos H_0 , σ_0^2 , y σ^2 .

Ahora podemos empezar la solución analítica

Recordemos la formula de la densidad normal de una variable x con promedio μ y desv. estandar σ

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Tenemos que multiplicar dos normales:

$$p(H|X) \propto Normal(X|\mu = H, \sigma^2) \times Normal(H|\mu = H_0, \sigma_0^2)$$

Reemplazemos las normales por sus formulas:

$$p(H|X) \propto \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-H)^2}{2\sigma^2}} \times \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{\frac{-(H-H_0)^2}{2\sigma_0^2}}$$

Luego de algo de algebra ...

$$p(H|X) = \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{\frac{-(\mu-\mu_1)^2}{2\sigma_1^2}}$$

Donde

$$\sigma_1^2 = \left(\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2} \right)^{-1}$$

$$\mu_1 = \frac{\mu_0\sigma_0^{-2} + X\sigma^{-2}}{\sigma^{-2} + \sigma_0^{-2}}$$

Nota: cuando X son n observaciones independientes, σ^2 se divide por n y X es \bar{X}

En lenguaje natural:

Nuestra creencia en H tiene una probabilidad gaussiana.

$$p(H|X) \sim Normal(\mu_1, \sigma_1^2)$$

Su promedio μ_1 depende de μ_0 (promedio del prior) y \bar{X} (promedio de la data), ponderados y escalados por la precisión de sus distribuciones (σ_0^{-2} y σ^{-2}).

$$\frac{\mu_0 \sigma_0^{-2} + \bar{X} \frac{\sigma^{-2}}{n}}{\frac{\sigma^{-2}}{n} + \sigma_0^{-2}}$$

Su varianza σ_1^2 depende de las desviaciones estandar del likelihood y el prior

$$\left(\frac{1}{\frac{\sigma^2}{n}} + \frac{1}{\sigma_0^2} \right)^{-1}$$

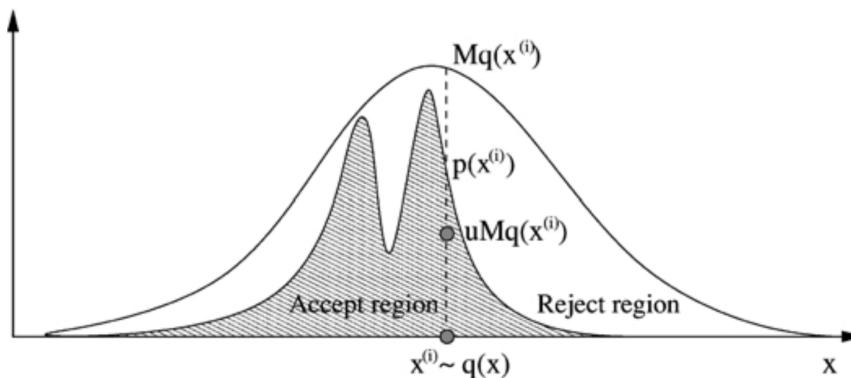
FORMA 2: SIMULACIONES (MCMC)

Rejection sampling:

Tomar muestras de $p(x)$ (gris) indirectamente con una distribución más simple $Mq(x)$ más alta que $p(x)$ (blanca). N son iteraciones (no muestras).

```
Set i = 1
Repeat until i = N
    1. Sample  $x^{(i)} \sim q(x)$  and  $u \sim \mathcal{U}_{(0,1)}$ .
    2. If  $u < \frac{p(x^{(i)})}{Mq(x^{(i)})}$  then accept  $x^{(i)}$  and increment the counter i by
        1. Otherwise, reject.
```

Figure 1. Rejection sampling algorithm. Here, $u \sim \mathcal{U}_{(0,1)}$ denotes the operation of sampling a uniform random variable on the interval $(0, 1)$.



Fuente: Andrieu, et al, (2003)

Algoritmo en lenguaje natural: solo aceptar propuestas bien probables en $p(x)$ relativo a $Mq(x)$.

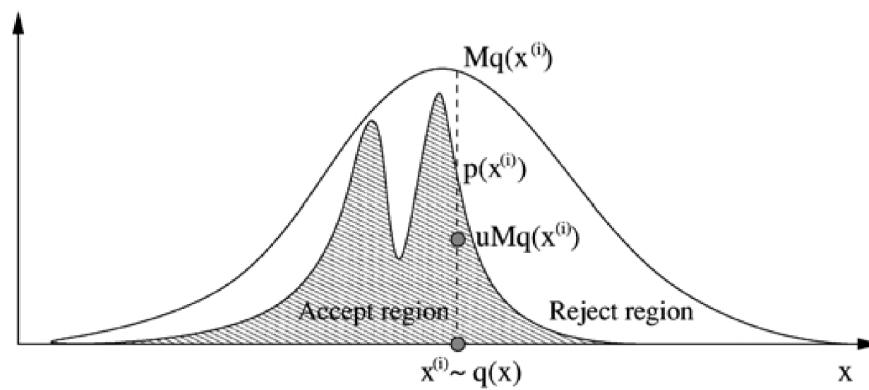


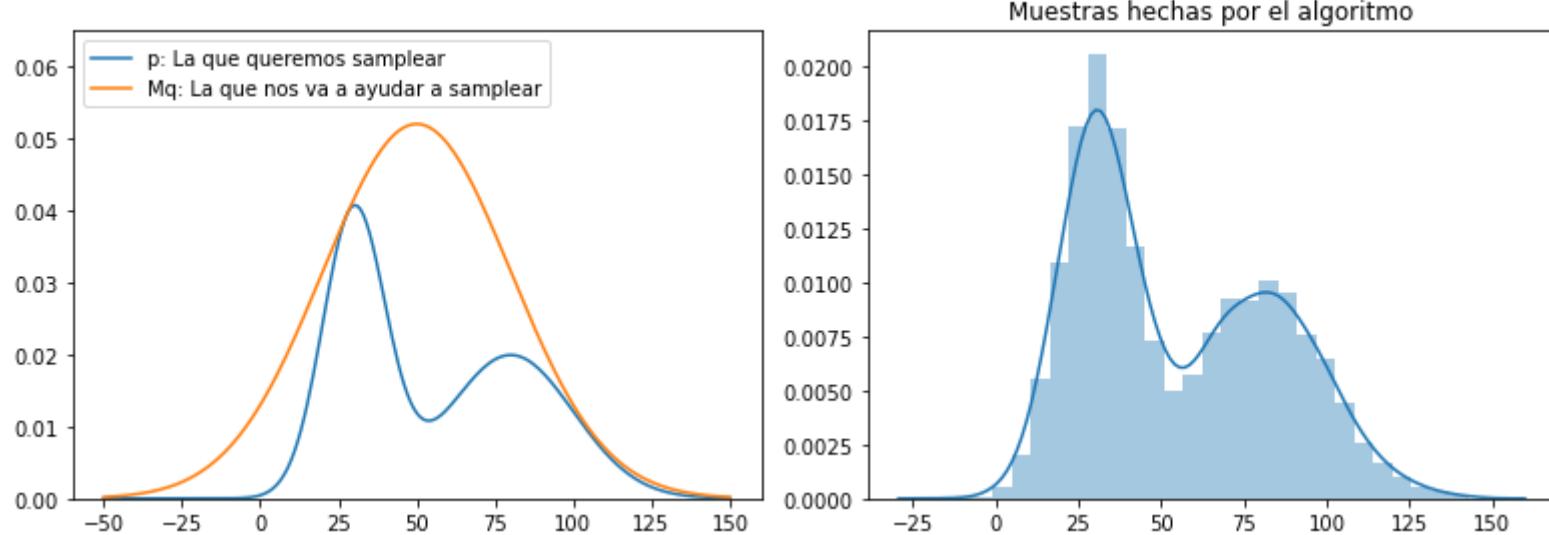
Figure 2. Rejection sampling: Sample a candidate $x^{(i)}$ and a uniform variable u . Accept the candidate sample if $uMq(x^{(i)}) < p(x^{(i)})$, otherwise reject it.

```
In [2]: #Tomado de Agustinus Kristiadi's blog (https://wiseodd.github.io/techblog/2015/10/21/rejection-sampling/)  
  
def p(x):  
    #Distribución que nos interesa tomar muestras  
    #Sin estandarizar: la integral -inf,inf no da 1.  
  
    return st.norm.pdf(x, loc=30, scale=10) + st.norm.pdf(x, loc=80, scale=20)  
  
def q(x):  
    #Distribución más simple que nos va a ayudar a tomar muestras  
    return st.norm.pdf(x, loc=info_q[0], scale=info_q[1])  
  
  
def rejection_sampling(iter=1000):  
    #El algoritmo para tomar muestras  
    samples = []  
    for i in range(iter):  
        x_proposed = np.random.normal(info_q[0], info_q[1])  
        u = np.random.uniform(0, M*q(x_proposed))  
  
        if u <= p(x_proposed):  
            samples.append(x_proposed)  
  
    return np.array(samples)
```

```
In [3]: # graficas
info_q = [50,30]
x = np.arange(-50, 151)
M = max(p(x) / q(x)) #Asegura que Mq > p

fig, ax = plt.subplots(1,2, figsize=(11,4))
ax[0].plot(x, p(x), label = 'p: La que queremos samplear')
ax[0].plot(x, M*q(x), label = 'Mq: La que nos va a ayudar a samplear')
ax[0].set_ylim([0,0.065])
ax[0].legend(loc='upper left')

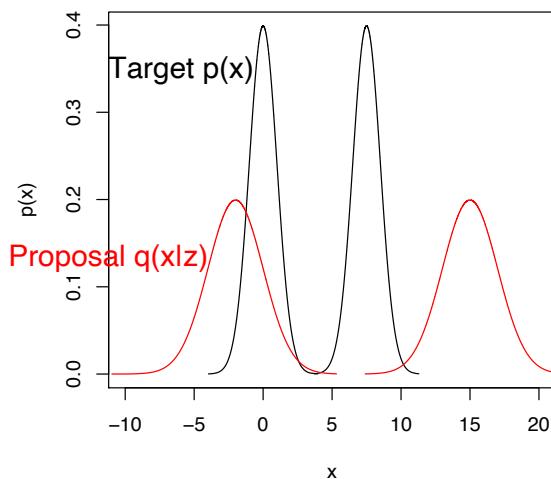
s = rejection_sampling(iter=10000)
sns.distplot(s, ax = ax[1])
plt.title('Muestras hechas por el algoritmo')
plt.tight_layout()
plt.show()
```



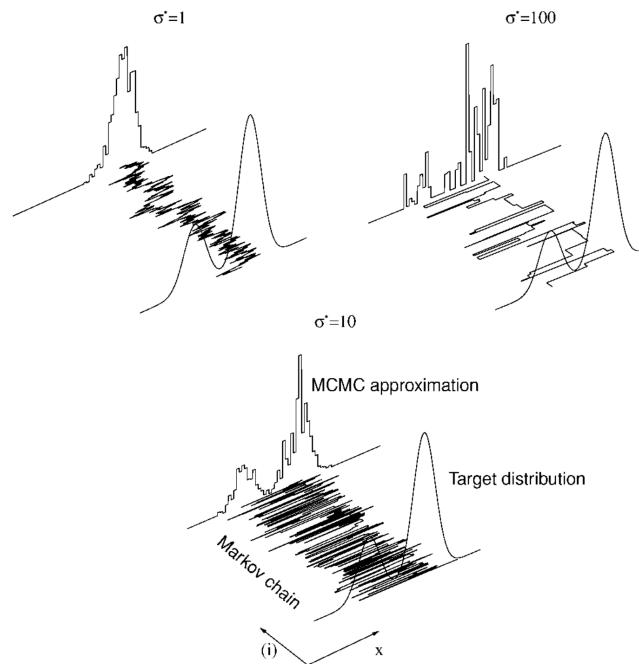
Sampleo MCMC con Metropolis Hastings:

Tomar muestras de $p(x)$ via una distribución más simple $q(x)$ (roja) que se puede mover. Los movimientos son markovianos: solo dependen de la posición actual.

```
1. Initialise  $x^{(0)}$ .  
2. For  $i = 0$  to  $N - 1$   
    – Sample  $u \sim \mathcal{U}_{[0,1]}$ .  
    – Sample  $x^* \sim q(x^*|x^{(i)})$ .  
    – If  $u < \mathcal{A}(x^{(i)}, x^*) = \min\left\{1, \frac{p(x^*)q(x^{(i)}|x^*)}{p(x^{(i)})q(x^*|x^{(i)})}\right\}$   
         $x^{(i+1)} = x^*$   
    else  
         $x^{(i+1)} = x^{(i)}$ 
```



Metropolis Hastings puede generar muestras de distribuciones arbitrarias. Acá vemos ejemplos de diferentes propuestas $q(x)$ con diferente movilidad (desviaciones estandar)



Fuente: Andrieu, et al, (2003)

Algunas alternativas de algoritmos para samplear:

- PyMC (Python)
- Edward (Python)
- Stan (Python, R, Julia, Stata, Matlab)
- JAGS (Python, R)

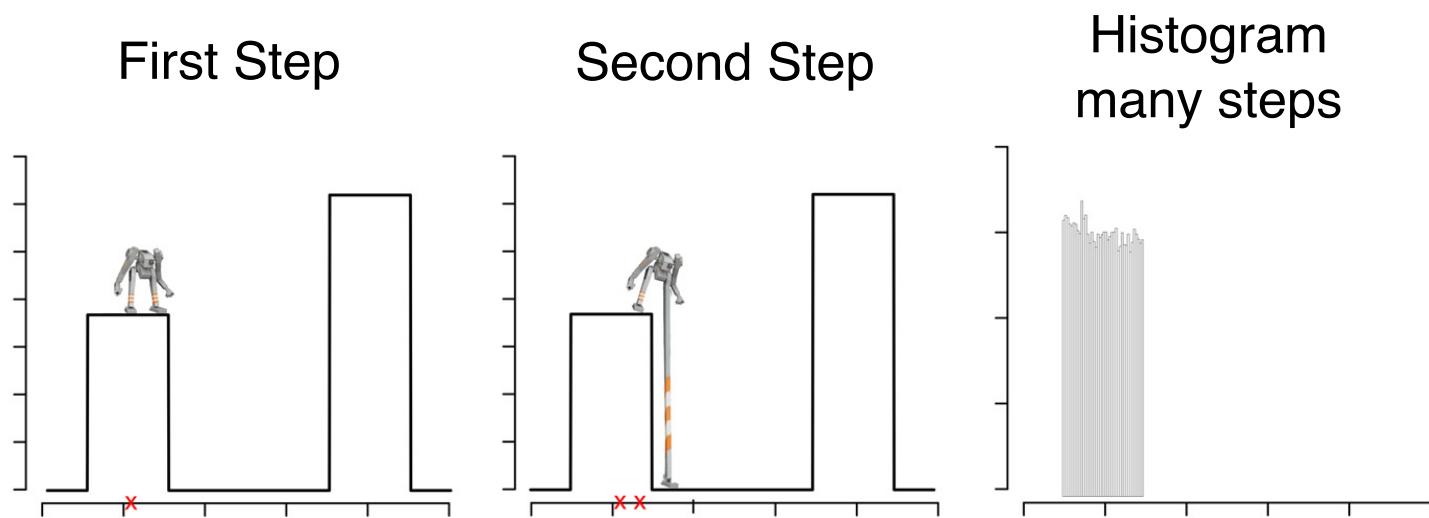
COGNICIÓN BAYESIANA: SAMPLEADORES

Un sampleador Bayesiano via MCMC puede explicar sesgos clásicos en toma de decisiones

Property	Ideal Bayesian reasoning	Bayesian sampler
Strictly follows laws of probability	Yes	Only asymptotically, otherwise can show systematic biases
Represents all hypotheses simultaneously in the brain	Yes	No, only represents one or a few at a time
Is easy to generate examples	Yes, though requires a sampling mechanism	Yes, just selects a sample
Can find all likely hypotheses, even if surprising	Yes	No, leading to unpacking effects and conjunction fallacies
Can evaluate relative probabilities of far-apart or incommensurable hypotheses	Yes	No, leading to forms of base-rate neglect and other forms of the conjunction fallacy
Is more effective in 'small' worlds compared to large context-rich worlds	Yes, better with small, well-defined worlds	No, better when context guides sampler to relevant hypotheses, rather than aggregating over broad hypothesis spaces
Produces stochastic and autocorrelated behavior	No	Yes

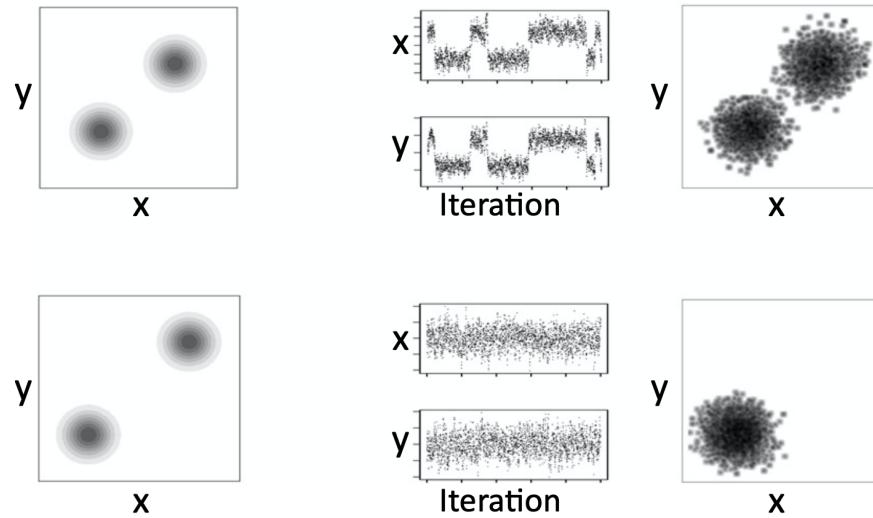
Fuente: Sanborn & Chater (2016)

El algoritmo puede no explorar todo el espacio de hipótesis



Fuente: Sanborn & Chater (2016)

El algoritmo puede no explorar todo el espacio de hipótesis



Fuente: Sanborn & Chater (2016)

Veamos un caso que puede, en principio, explicarse por características del sampleador

Mitad de ustedes cierren los ojos. No los abran hasta que les diga.

La otra mitad hagan la siguiente multiplicación en menos de 3 segundos y escriban el resultado en un papel (NO HABLEN).

$$8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$$

Abran los ojos.

Los que acaban de abrir los ojos, hagan esta multiplicación, en menos de 3 segundos y escriban el resultado en un papel:

$$1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7 \times 8$$

La diferencia se da por anchoring effects.

Es importante. Por ejemplo, en economía anchoring effects
"*...challenge the central premise of welfare economics that choices reveal true preferences...*" (Ariely, et al, 2003)(pp. 102)

TABLE I
AVERAGE STATED WILLINGNESS-TO-PAY SORTED BY QUINTILE OF THE SAMPLE'S
SOCIAL SECURITY NUMBER DISTRIBUTION

Quintile of SS# distribution	Cordless trackball	Cordless keyboard	Average wine	Rare wine	Design book	Belgian chocolates
1	\$ 8.64	\$16.09	\$ 8.64	\$11.73	\$12.82	\$ 9.55
2	\$11.82	\$26.82	\$14.45	\$22.45	\$16.18	\$10.64
3	\$13.45	\$29.27	\$12.55	\$18.09	\$15.82	\$12.45
4	\$21.18	\$34.55	\$15.45	\$24.55	\$19.27	\$13.27
5	\$26.18	\$55.64	\$27.91	\$37.55	\$30.00	\$20.64
Correlations	.415	.516	0.328	.328	0.319	.419
	$p = .0015$	$p < .0001$	$p = .014$	$p = .0153$	$p = .0172$	$p = .0013$

The last row indicates the correlations between Social Security numbers and WTP (and their significance levels).

¿Cómo explicarlo? Es difícil, pero aca va una hipótesis Bayesiana:
 Sampleador sesgado que empieza con el anchor. Solo después de
 varias iteraciones cae el sesgo.

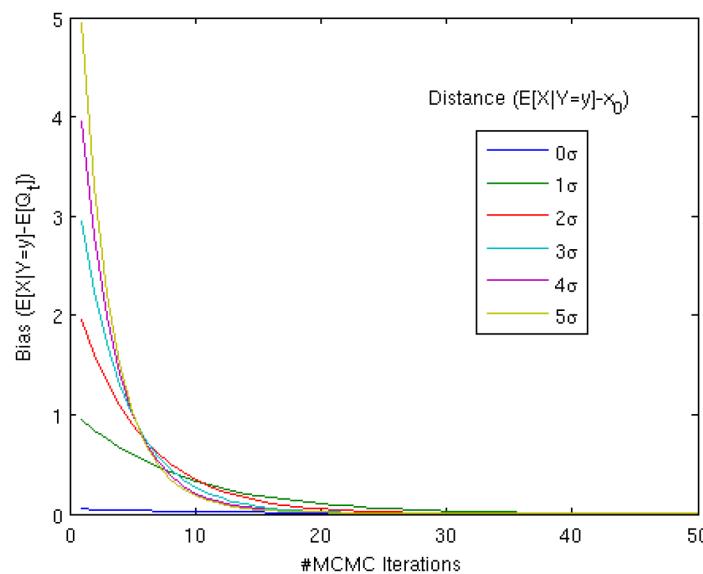


Figure 1: Bias of the mean of the approximation Q_t , i.e. $|\mathbb{E}[\tilde{X}_t] - \mathbb{E}[X|Y = y]|$ where $X_t \sim Q_t$, as a function of the number of iterations t of our Metropolis-Hastings algorithm. The five lines show this relationship for different posterior distributions whose means are located $1\sigma_p, \dots, 5\sigma_p$ away from the prior mean (σ_p is the standard deviation of the prior). As the plot shows, the bias decays geometrically with the number of iterations in all five cases.

CONCLUSIONES

¿Por qué cognición bayesiana?

Es un mundo ambiguo/incierto. El cerebro tiene que inferir la probabilidad de diferentes hipótesis

