Regla de Bayes

Dr. Héctor Nájera
Dr. Curtis Huffman

03/03/2022

La sesión anterior

- La naturaleza dual de la probabilidad
  - Kolmogorov (teoría de la medida)
- Las distribuciones de probabilidad definen la medida de eventos
- Parámetros definen distribuciones

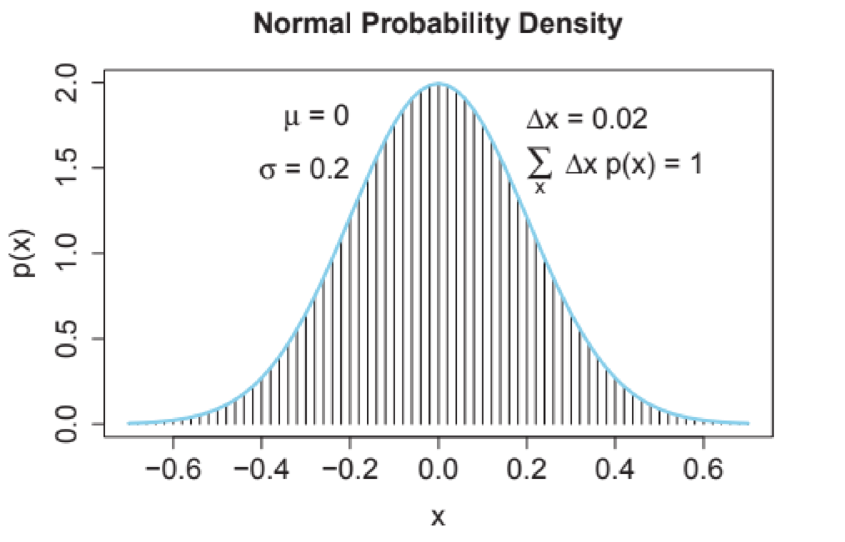


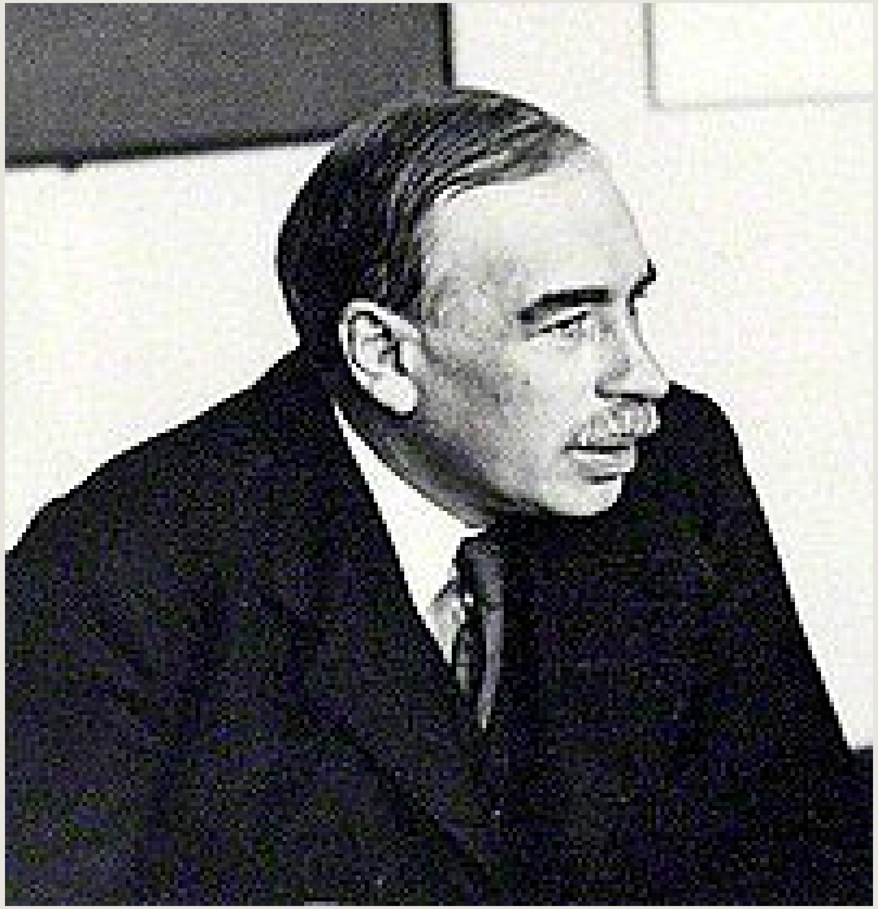
Figure 4.4: A normal probability density function, shown with a comb of narrow intervals. The integral is approximated by summing the width times height of each interval. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition*. Academic Press / Elsevier.

Probability Density Function

$$F(x) = P(a \le x \le b) = \int_a^b f(x)dx \ge 0$$

Probabilidad condicional

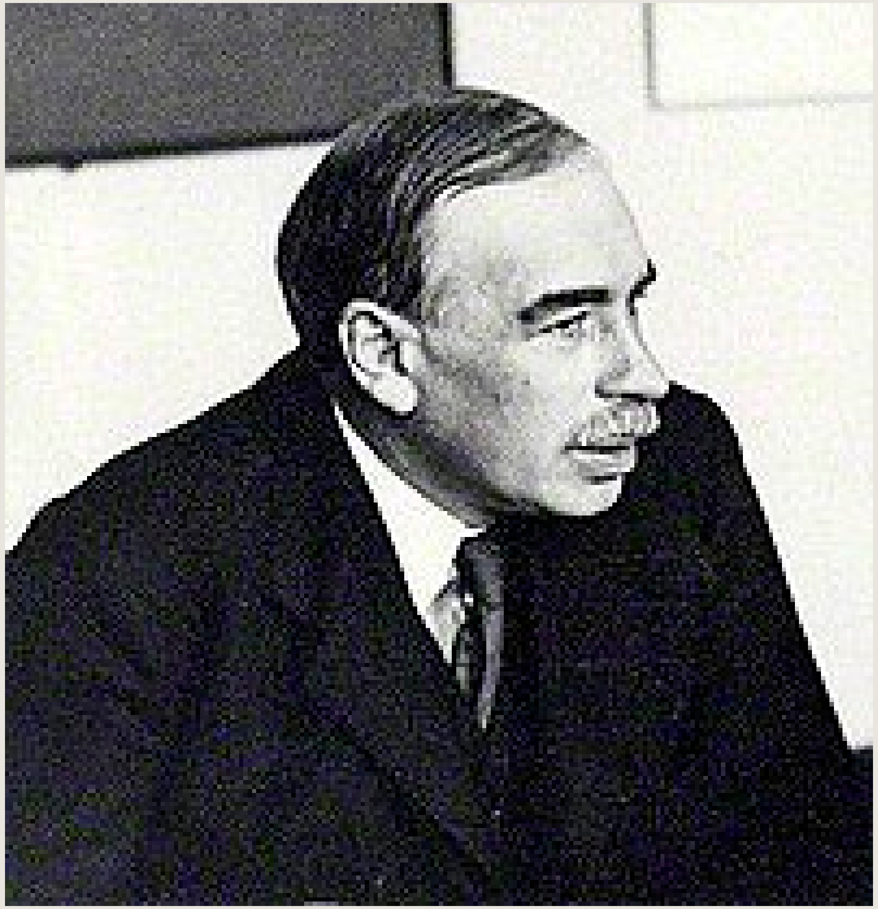
- All probabilities are conditional probabilities. Keynes, 1921
- Siempre usamos información para actualizar el espacio muestral o de referencia para estimar probabilidades.



John Maynard Keynes
1883-1942

Probabilidad condicional

- La probabilidad de que un libro escogido aleatoriamente de la estantería de su biblioteca esté forrado en cuero es condicional en los libros que de hecho están en su biblioteca y en cómo se haga esa elección “aleatoria”.



John Maynard Keynes
1883-1942

Esta clase

- Capítulo 5. Bayes´rule

(or Bayes rules)

Teorema de Bayes

- El capítulo 5 brinda las bases de cómo el teorema de Bayes puede utilizarse para hacer inferencia sobre parámetros vía construcción de distribuciones posteriores

Ideas centrales:

  - Definición formal de probabilidad condicional
  - Aplicación del teorema de Bayes a parámetros y datos
  - Influencia de los priors en la posterior
  - Por qué la inferencia bayesiana es difícil

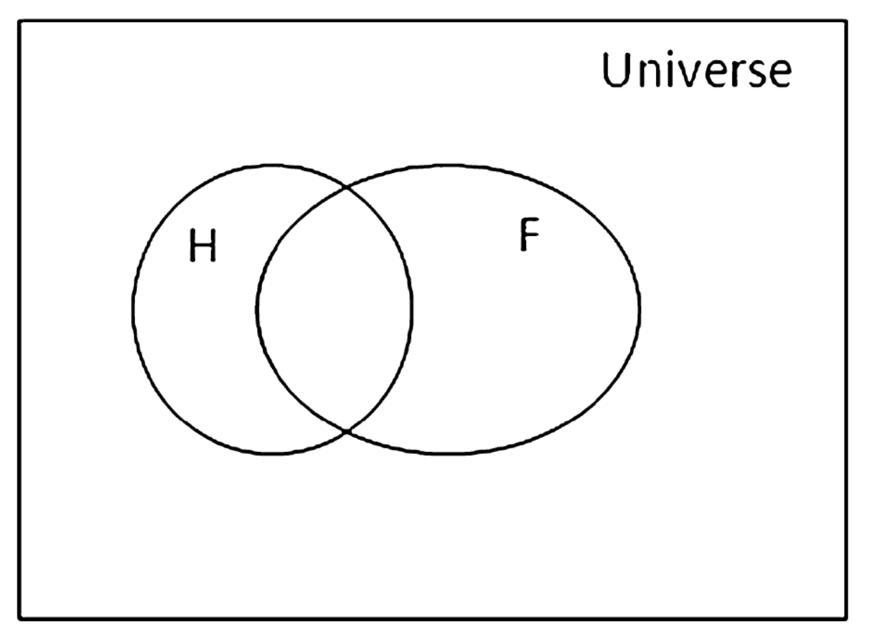
Teorema de Bayes

El objetivo típico consiste en determinar la probabilidad de una hipótesis (H) dados los datos (F; findings)

Invierte la probabilidad condicional

$$p(H|F) = \frac{p(F|H)\,p(H)}{p(F)}$$

La probabilidad de "antes" condicional en "después" ¿la probabilidad de que una baraja contenga cuatro ases dado que una de las manos repartidas tiene dos ases?

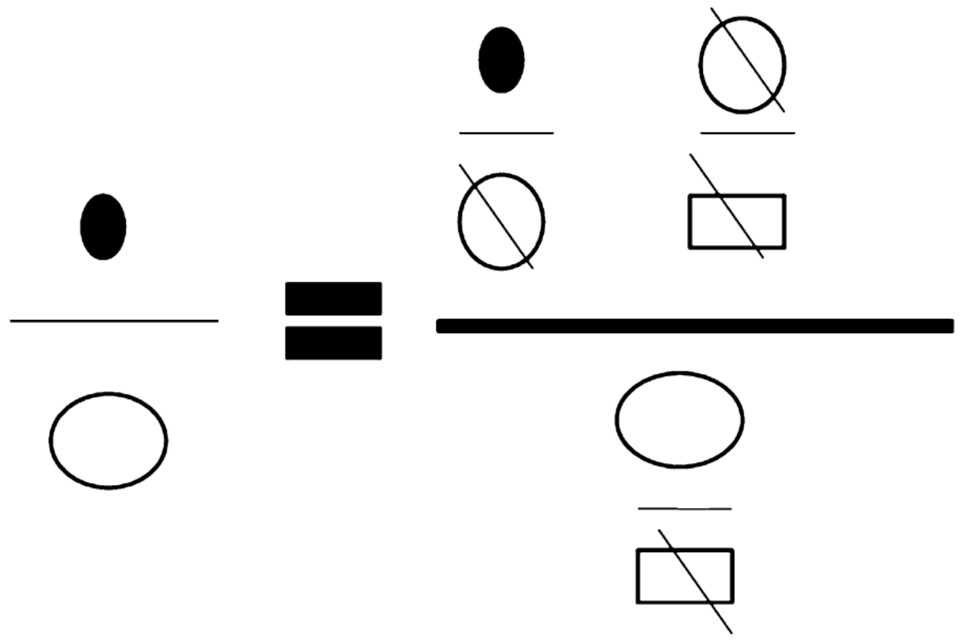


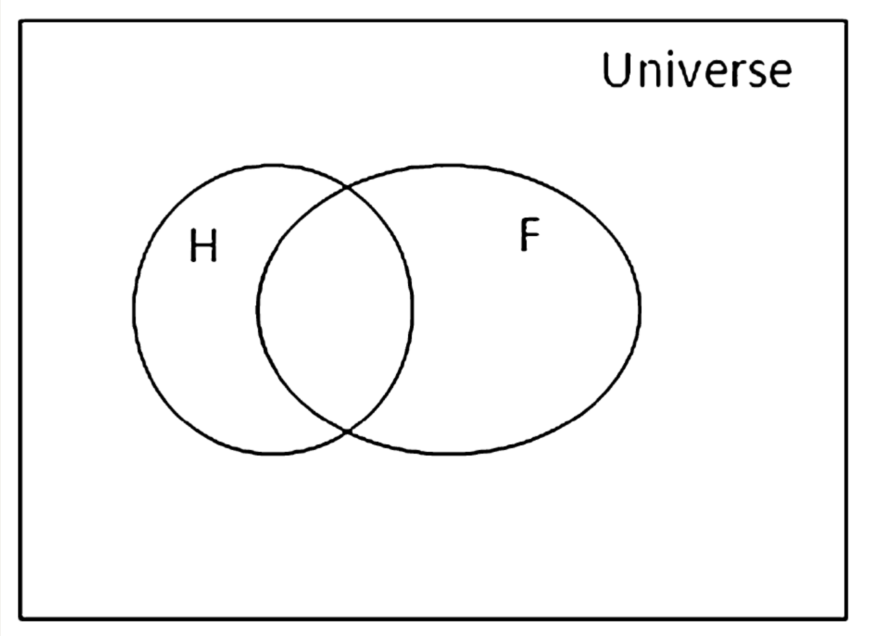
Trafimow, D. (2011).

Teorema de Bayes

El objetivo típico consiste en determinar la probabilidad de una hipótesis (H) dados los datos (F; findings)

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



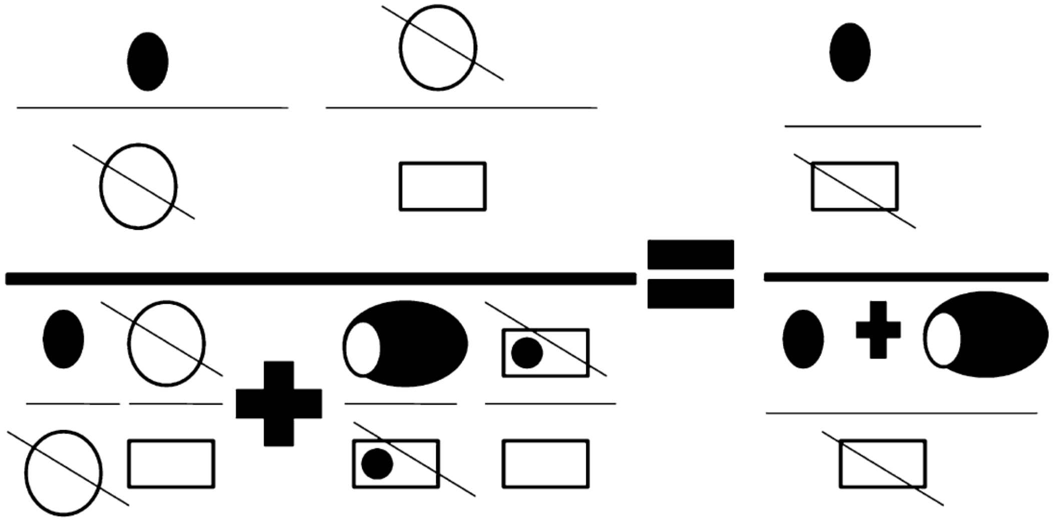


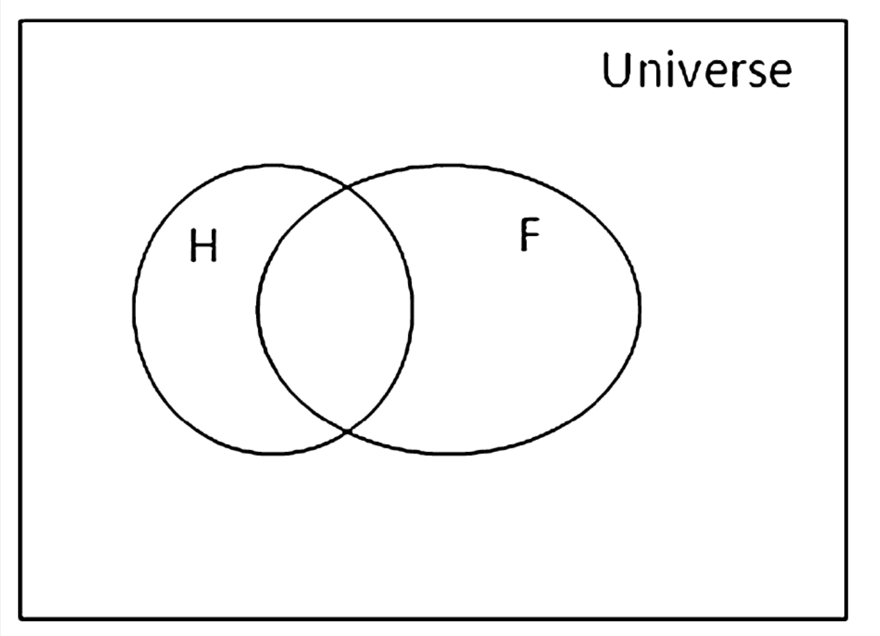
Trafimow, D. (2011).

Teorema de Bayes

El objetivo típico consiste en determinar la probabilidad de una hipótesis (H) dados los datos (F; findings)

$$\frac{p(F|H)p(H)}{p(F|H)p(H) + p(F|\sim H)p(\sim H)} = p(H|F)$$





Trafimow, D. (2011).

El caso de las dos monedas

Suponga que tenemos 2 monedas:

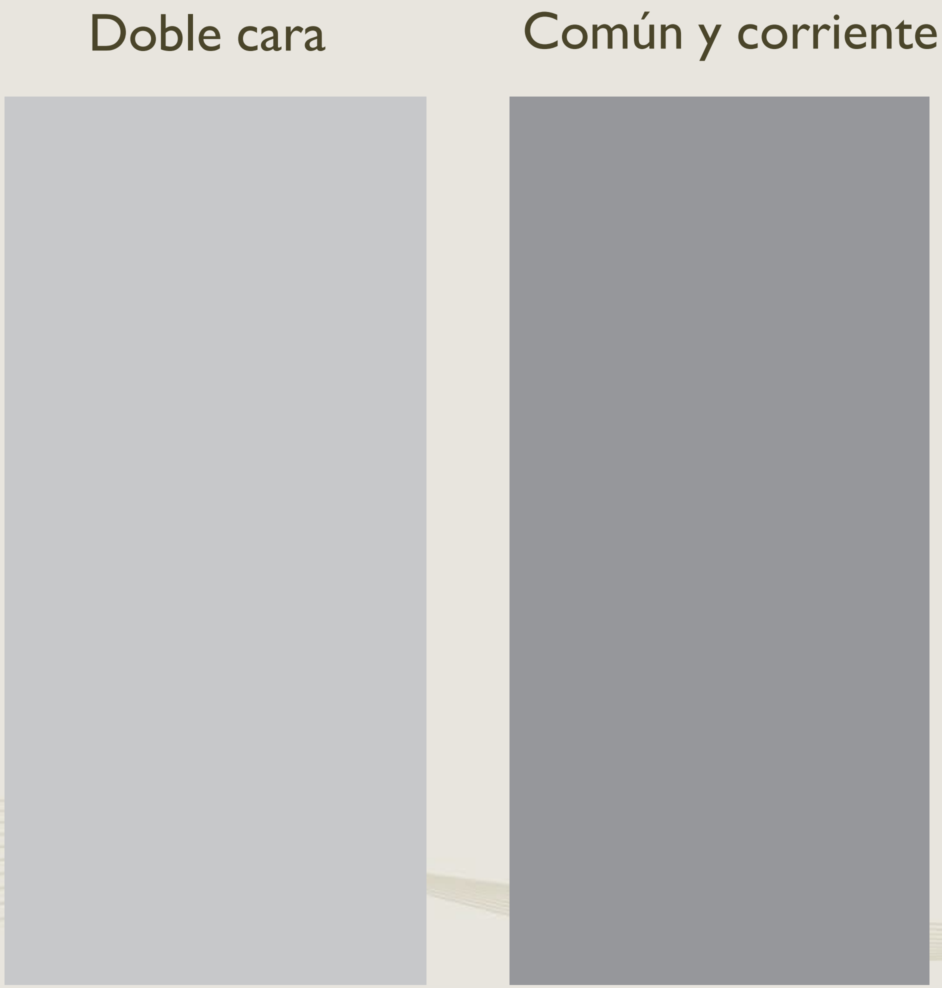
Común y corriente

Doble cara

¿cuál es la probabilidad de que hayamos escogido (***aleatoriamente***) la común y corriente?

El caso de las dos monedas

¿Cuál es la probabilidad (a priori) de que hayamos escogido (***aleatoriamente***) la común y corriente?



Erickson, T. (2017)

El caso de las dos monedas

Tiro la moneda al aire sin saber cuál es y al revisar descubrimos que ha caído cara,
¿sabemos cuál moneda escogimos?,
¿cuál es la probabilidad (posterior) de que la moneda escogida sea la común y corriente?

Doble cara $\pi = 1$

Común y corriente $\pi = .5$

$$P(F1|H1) = \frac{P(F1).P(H1|F1)}{P(F1).P(H1|F1) + P(U1).P(H1|U1)}$$

$$= \frac{1/2.1/2}{1/2.1/2 + 1/2.1} = \frac{1}{3}$$

$$P(F1|H1) = 1/3$$

Erickson, T. (2017)

El caso de las dos monedas

Simulación de la regla de Bayes por el método de Monte Carlo

D10004 | f_x =D10002/C10002

	A	B	C	D
1	Coin	Flip	Head	Head|Fair
2	U	H	1	
3	U	H	1	
4	F	H	1	1
5	U	H	1	
6	F	H	1	1
7	U	H	1	
8	F	T		
9995	F	H	1	1
9996	F	T		
9997	F	T		
9998	U	H	1	
9999	U	H	1	
10000	F	T		
10001	F	H	1	1
10002	5019		7471	2490
10003				
10004	**P(F)=**	**0,5019**	**P(F1|H1)=**	**0,33329**

=D10002/C10002

A2=SI(ALEATORIO()<0.5,"F","U")

B2=SI(A2="U","H",SI(ALEATORIO()<0.5,"H","T"))

C2=SI(B2="H",1,"")

D2=SI(Y(A2="F",B2="H"),1,"")

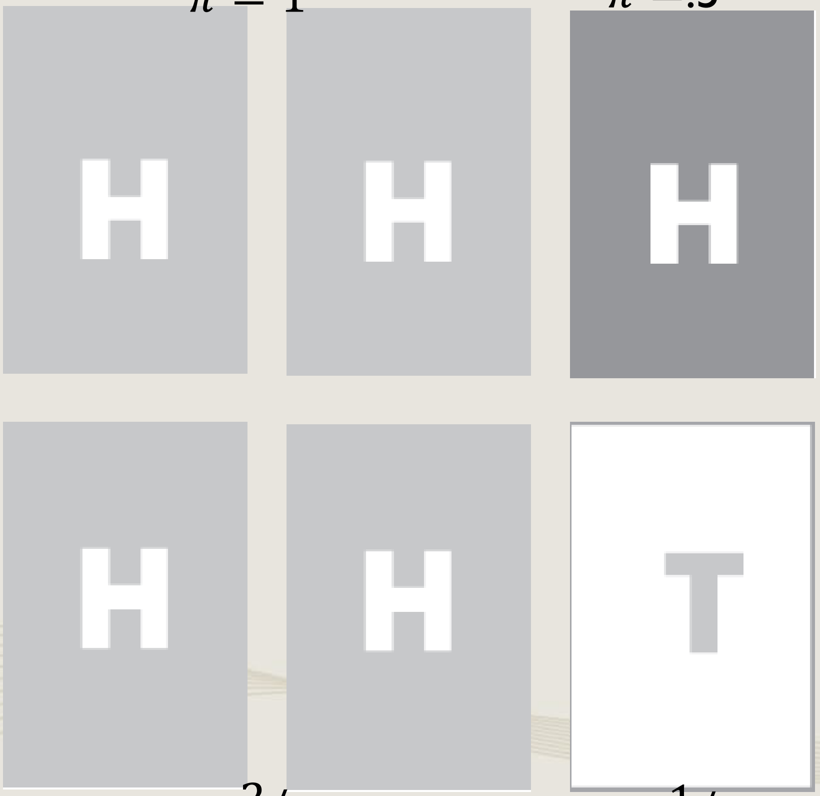
Gangur, M., & Svoboda, M. (2018).

Teaching Statistics, Volume: 40, Issue: 3, Pages: 83-87, First published: 06 April 2018, DOI: (10.1111/test.12158)

El caso de las dos monedas

Tiro la moneda al aire por segunda vez (sin revisar de qué tipo es) y al revisar descubrimos que ha caído cara nuevamente,
¿ahora sí ya sabemos cuál moneda escogimos?,
¿cuál es la probabilidad (posterior) de que la moneda escogida sea la común y corriente?

Doble cara $\pi = 1$ | Común y corriente $\pi = .5$



${}^{2}/_{3}$ | ${}^{1}/_{3}$

$$P(F2|H2) = \frac{P(F1).P(H2|F1)}{P(F1).P(H2|F1) + P(U1).P(H2|U1)}$$
$$= \frac{1/3.1/2}{1/3.1/2 + 2/3.1} = \frac{1}{5}$$

$$P(F2|H2) = 1/5$$

Erickson, T. (2017)

El caso de las dos monedas

Simulación de la regla de Bayes por el método de Monte Carlo

E10004 | =E10002/D10002

	A	B	C	D	E
1	Coin	First flip	Second flip	Head and Head	(Head and Head)|Fair
2	F	T			
3	U	H	H	1	
4	F	H	H	1	1
5	U	H	H	1	
6	F	H	T		
7	U	H	H	1	
8	U	H	H	1	
9995	F	T			
9996	F	T			
9997	U	H	H	1	
9998	U	H	H	1	
9999	F	H	H	1	1
10000	F	T			
10001	F	H	H	1	1
10002	5017			6239	1256
10003					
10004	**P(F)=**	**0,5017**		**P(F2|H2) =**	**0,20131**

=E10002/D10002

A2=SI(ALEATORIO()<0.5,"F","U")

B2=SI(A2="U","H",SI(ALEATORIO()<0.5,"H","T"))

C2=SI(B2="H",SI(A2="U","H",SI(ALEATORIO()>0.5,"H","T")),"")

D2=SI(Y(B2="H",C2="H"),1,"")

E2=SI(Y(A2="F",D2=1),1,"")

Gangur, M., & Svoboda, M. (2018).

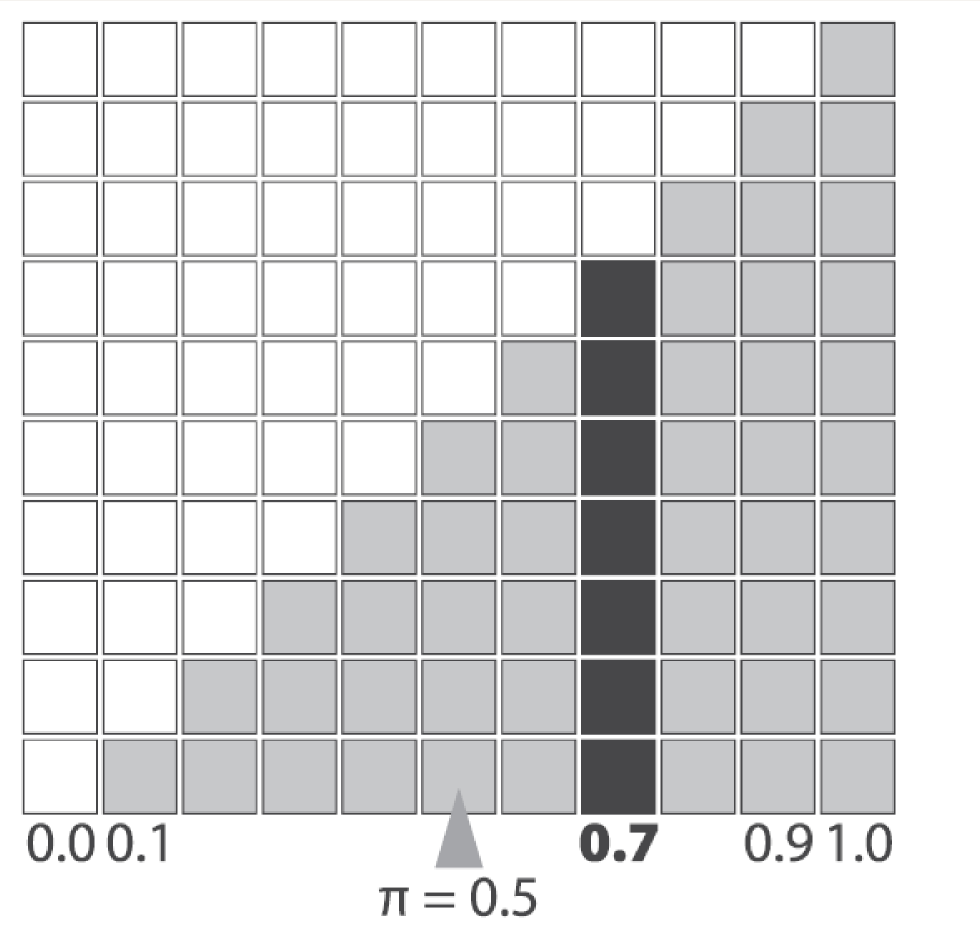
Teaching Statistics, Volume: 40, Issue: 3, Pages: 83-87, First published: 06 April 2018, DOI: (10.1111/test.12158)

Más de 2 hipótesis

- Suponga 11 diferentes hipótesis (igualmente probables) sobre la moneda misteriosa $\pi = \{0{,}0.1{,}0.2, \dots, 0.9, 1.0\}$
- Tiramos la moneda al aire y cae cara, ¿cuáles son las probabilidades posteriores?
- ¿Cuál es la probabilidad de $\pi = 0.7$?
- ¿Sigue siendo $1/11$ (algo así como 0.09)?

$$\mathrm{P}(\pi = 0.7|\mathrm{H1}) = 7/55$$

(algo así como 0.17)



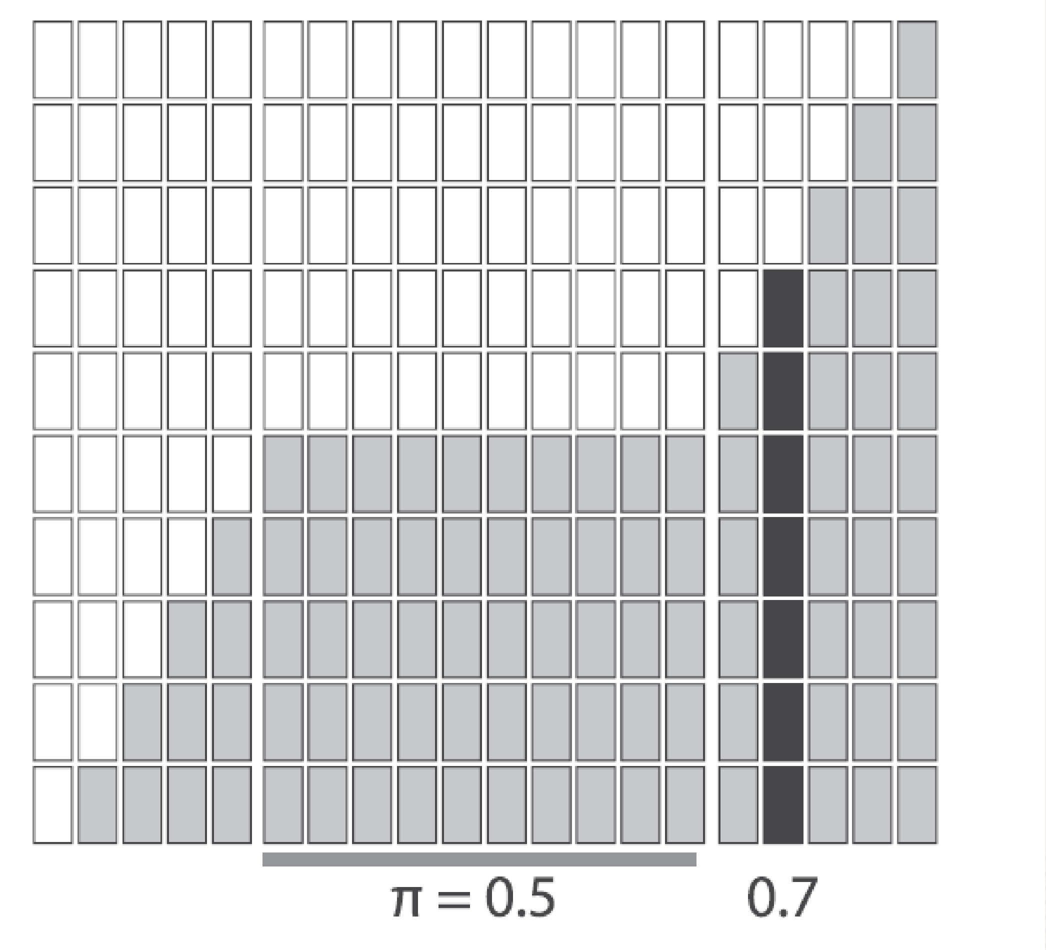
Erickson, T. (2017)

Más de 2 hipótesis

- ¿Cómo se vería un prior no-uniforme sobre las diferentes hipótesis usando el mismo diagrama de área donde podemos contar cajas?

$$P(\pi = 0.7 | H1) = 7/100$$

(pasa de 0.05 a 0.07)



Erickson, T. (2017)

Conclusiones. Influencia del tamaño de los datos sobre la posterior

- Hemos visto que la distribución posterior es un compromiso entre los a prioris y la función de verosimilitud
- Definimos dos p(θ)
- La posterior se aproxima a los datos a medida que la muestra se incrementa (con priors vagos es más claro)
- Entre más datos tengamos, mejor es la precisión

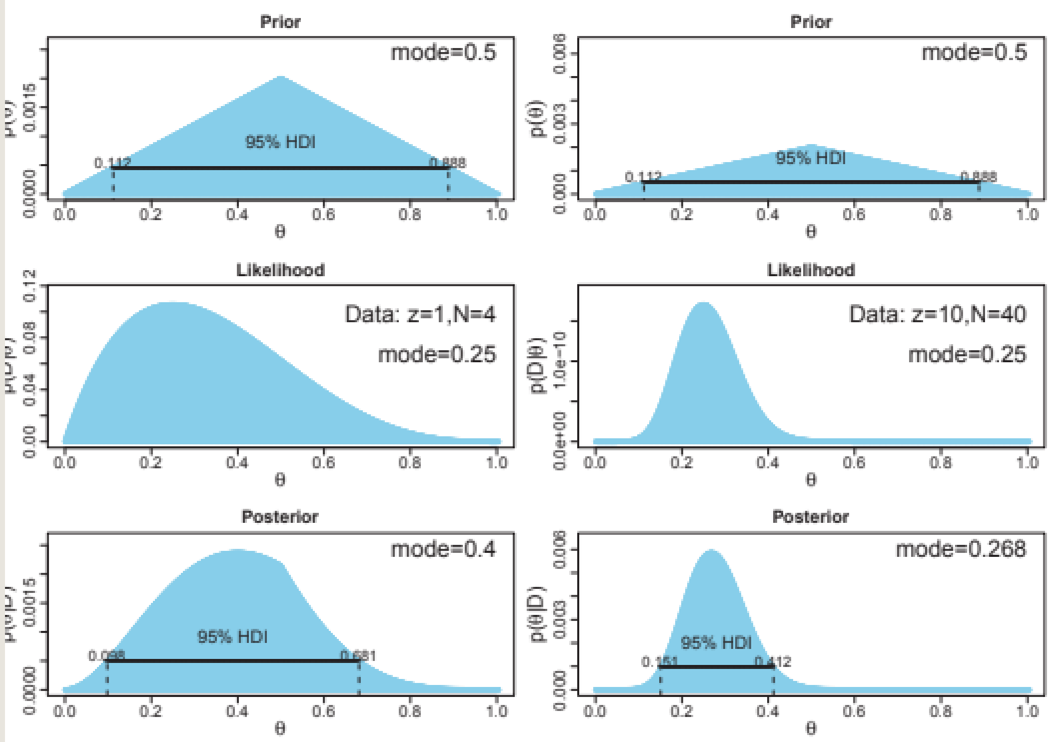


Figure 5.2: The two columns show different sample sizes with the same proportion of heads. The prior is the same in both columns but plotted on a different vertical scale. The influence of the prior is overwhelmed by larger samples, in that the peak of the posterior is closer to the peak of the likelihood function. Notice also that the posterior HDI is narrower for the larger sample. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Conclusiones. Influencia del a priori sobre la posterior

- En general, cuando la distribución de p(θ) es amplia comparada con la función de verosimilitud, p(θ) tiene poca influencia

- Con priors muy informativos (que están basados en muchos datos), se necesitarán muchos datos que lo contradigan para ver diferencias entre el prior y la distribución posterior.

- Con priors vagos, los datos fácilmente desplazan a la distribución a posteriori.

- "It can be a serious blunder not to use strong prior information when it is available" p. 114.

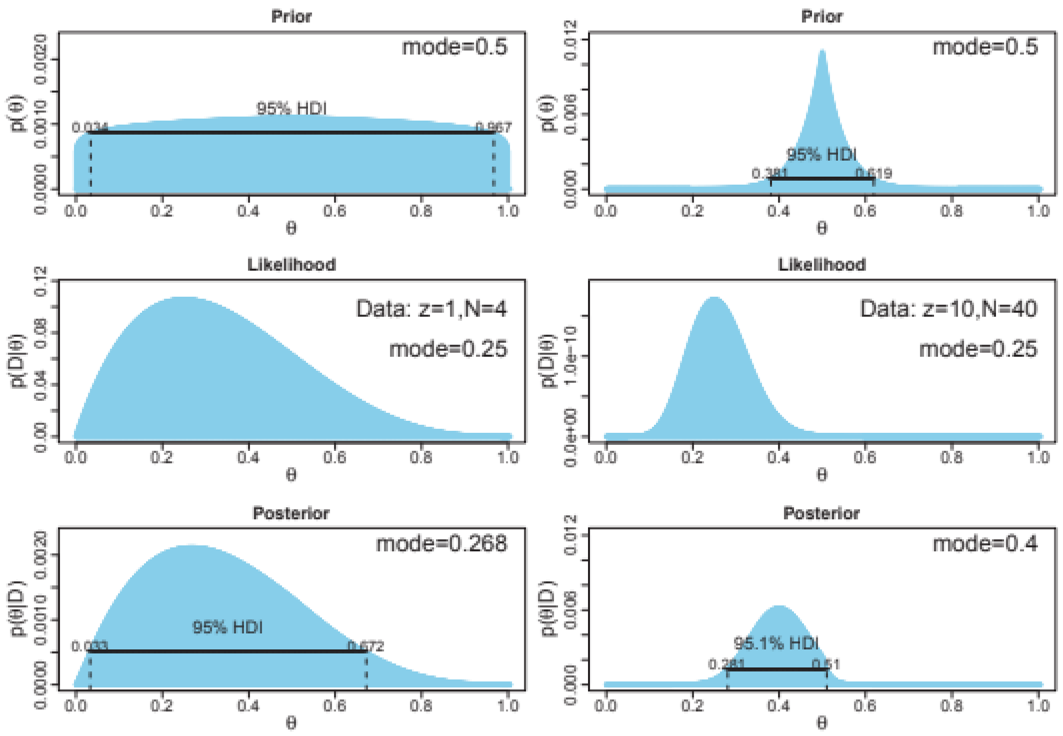


Figure 5.3: The left side is the same small sample as the left side of Figure 5.2 but with a flatter prior. The right side is the same larger sample as the right side of Figure 5.2 but with a sharper prior. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Aproximación de la posterior vía grids – lista de posibles explicaciones-

¿Cuál es la probababilidad de no hospitalización de la población vacunada?

$$P(\theta|D) = p(\theta|Datos + Modelo) = p(\theta|verosimilitud)$$

Verosimilitud (Datos y modelo):

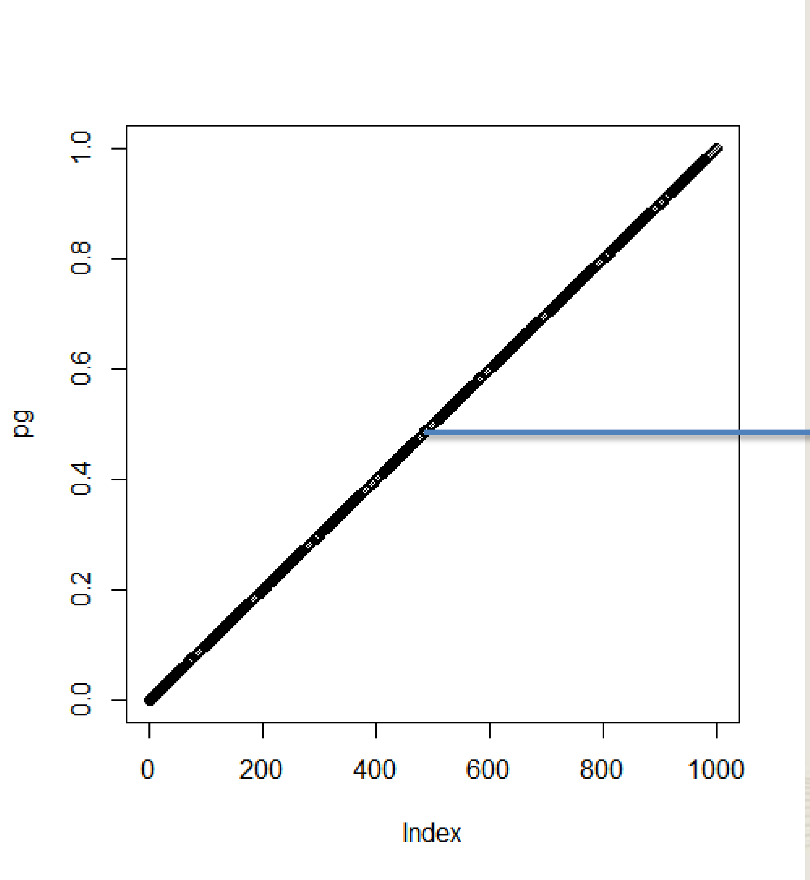
$$NHV \sim binomial(N, p)$$
$$P \sim uniform(0,1)$$

Supuestos sobre la formas posibles en la que los datos ocurrieron

Formas de observar Hospital y no en el hospital (NHV) dada una probabilidad – condicional en cierta explicación-

Aproximación de la posterior vía grids – lista de posibles explicaciones-

- ¿Sirve la vacuna?



Verosimilitud (Datos y modelo):

$$\text{NHV} \sim binomial(N, p)$$
$$P \sim uniform(0,1)$$

p=.5

50% de efectividad

Porque necesito estimar la probabilidad relativa de cada posibilidad. Para cada valor de p:

Pr(NHV|p)Pr(p)

Planteo un espacio de solución

- Las posibles probabilidades del parámetro
- Van de 0 a 1 en intervalos de .001

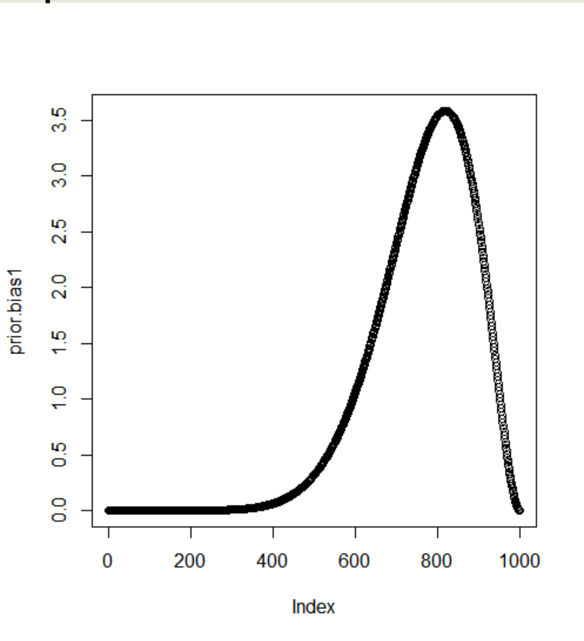
Aproximación vía grids –lista de posibles explicaciones-

- ¿Sirve la vacuna? –Pocos datos-

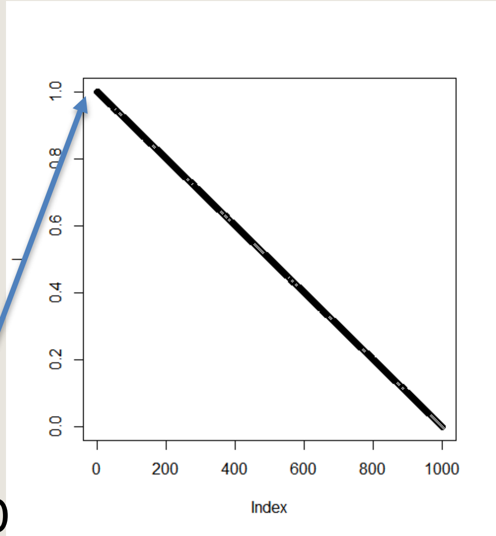
Creemos que la vacuna sirve –para la mayoría de casos de la lista pensamos que habrá éxito-

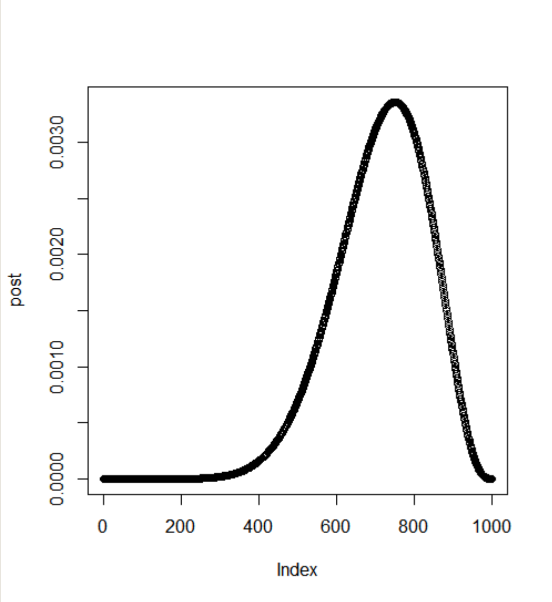
Los datos nos dicen que alguien que se vacunó terminó en el hospital.
Los parámetros de la lista menos probables tienen más peso

La posterior nos dice que ese caso no importa



p_1=.000





prior.bias1<-dbeta(pg,10,3)

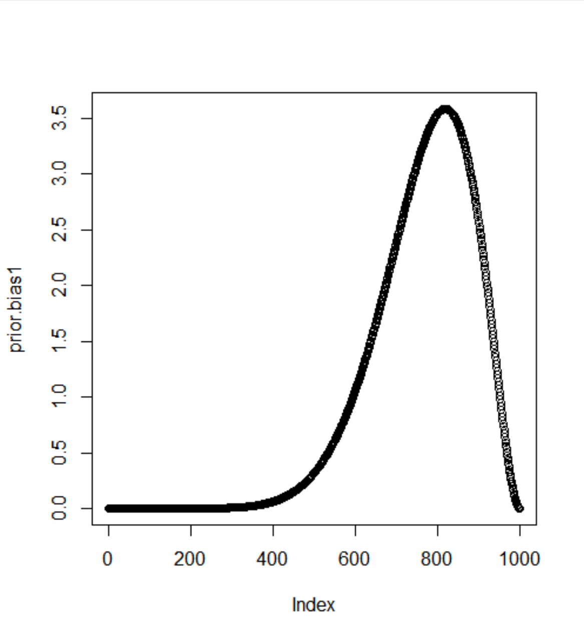
l<-dbinom(0,1,prob=pg)

post<-l* prior.bias1
post<-post/sum(post)
plot(post)

Segundo ejemplo: Más datos

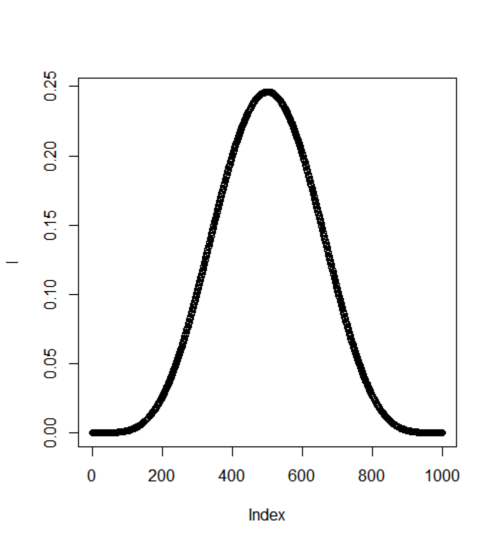
- ¿Srive la vacuna? –Pocos datos-

Creemos que sirve



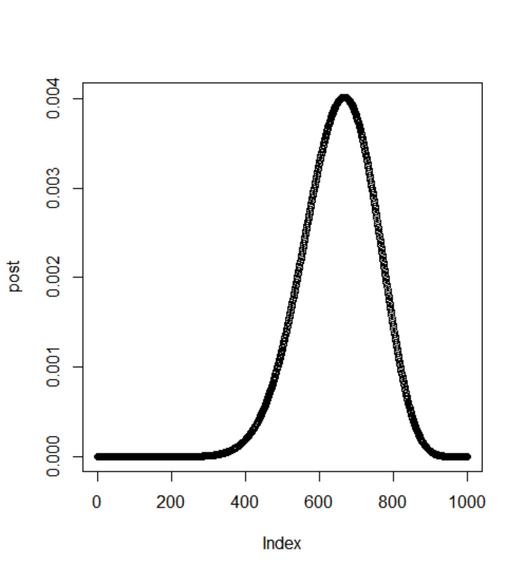
prior.bias1<-dbeta(pg,10,3)

Más datos nos dicen que cayeron 5 de 10 no terminaron en el hospital



l<-dbinom(5,10,prob=pg)

La posterior nos dice que no es 50/50

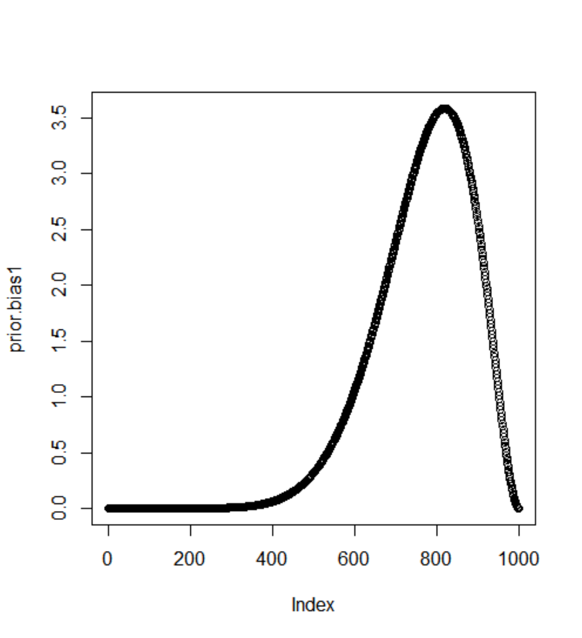


post<-l* prior.bias1
post<-post/sum(post)
plot(post)

Segundo ejemplo

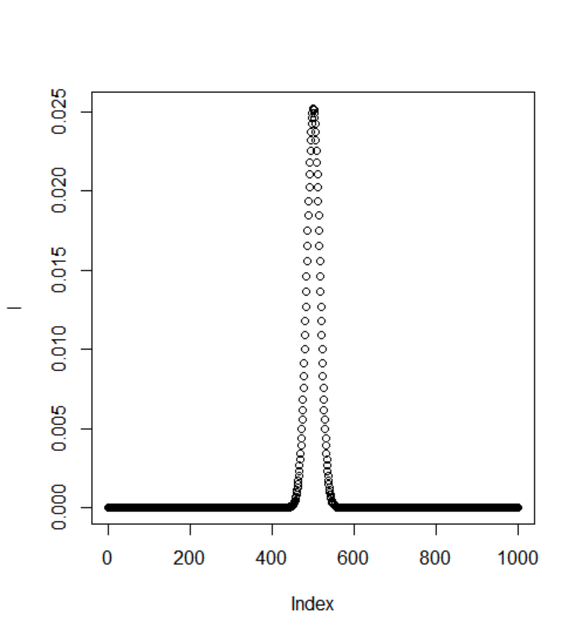
- ¿Sirve la vacuna? –Pocos datos-

Creemos que sirve



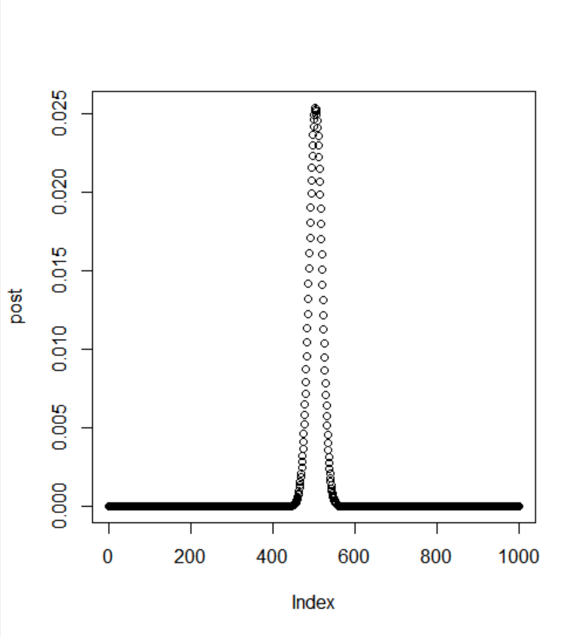
prior.bias1<-dbeta(pg,10,3)

Más datos nos dicen que de 10,000 vacunados, la mitad termina en el hospital hospital



l<-dbinom(5000,10000,prob=pg)

La posterior refleja los datos –verosimilitud-

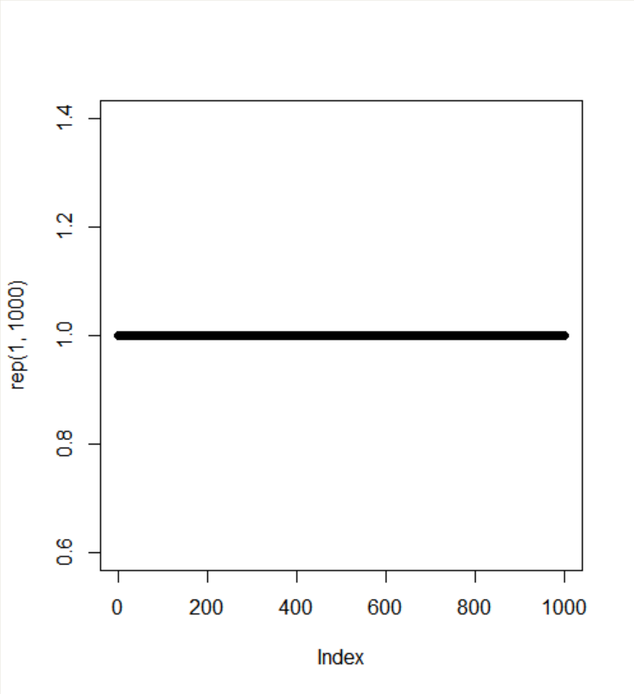


post<-l* prior.bias1
post<-post/sum(post)
plot(post)

Segundo ejemplo: B

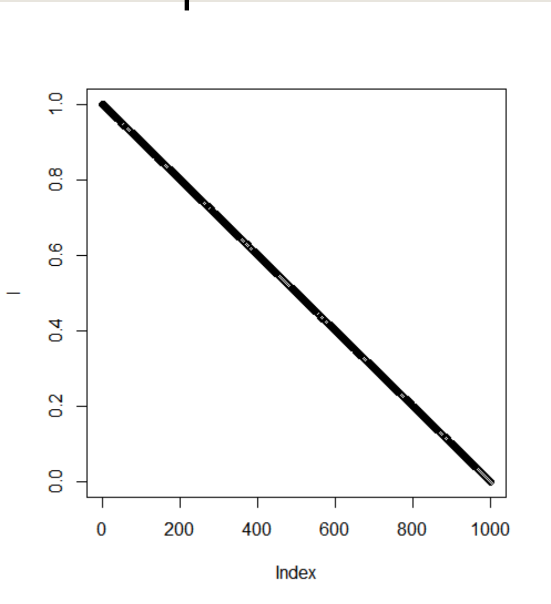
- ¿Sirve la vacuna? –Pocos datos-

Prior Uniforme: NO SABEMOS NADA



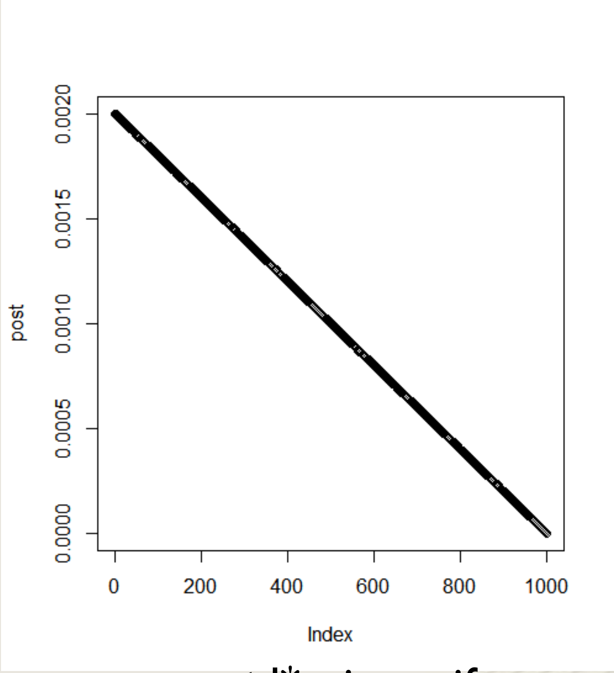
prior.uniform<-rep(1,1000)

Los datos nos dicen que el primer vacunado terminó en el hospital



l<-dbinom(0,1,prob=pg)

La posterior nos dice que la vacuna no sirve

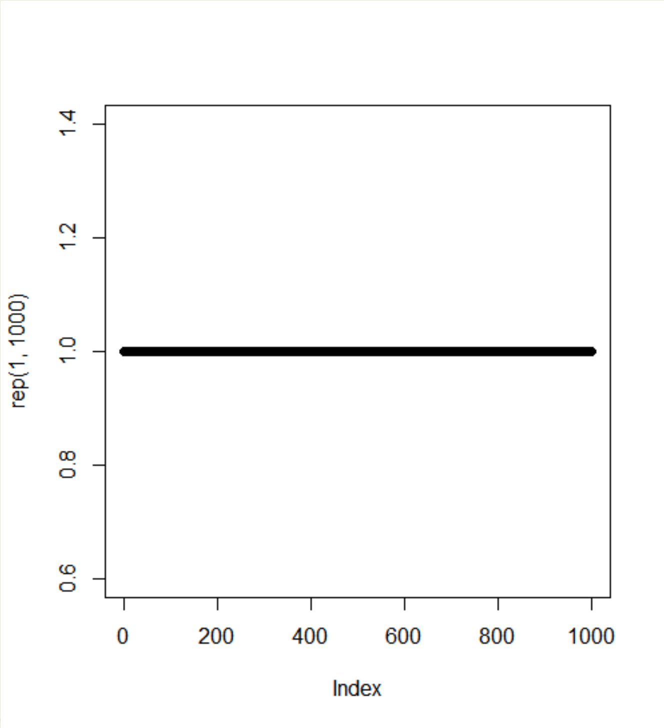


post<-l*prior.uniform
post<-post/sum(post)
plot(post)

Segundo ejemplo: B

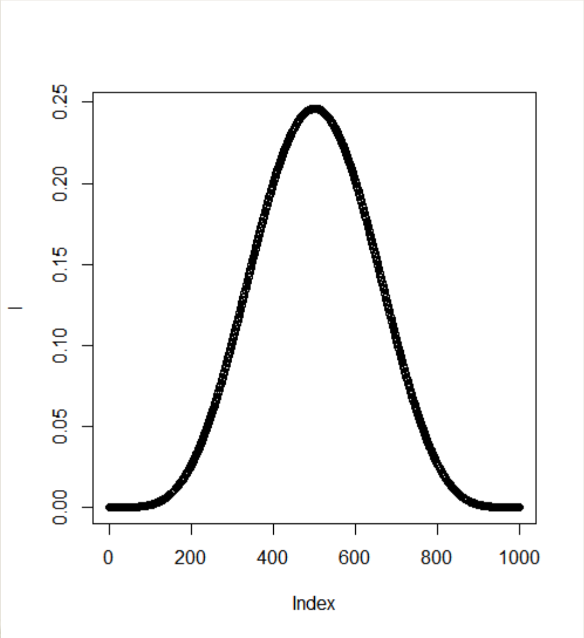
- ¿Sirve la vacuna? –Pocos datos-

Prior Uniforme: NO SABEMOS NADA



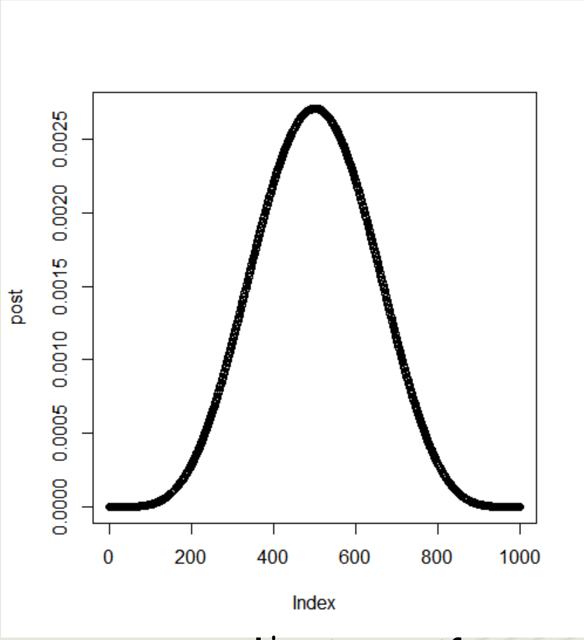
prior.bias1<-dbeta(pg,10,3)

Los datos nos dicen que 5 de 10 terminan en el hospital



l<-dbinom(5,10,prob=pg)

Priors uniformes dejan que dominen totalmente los datos –Es deseable?-



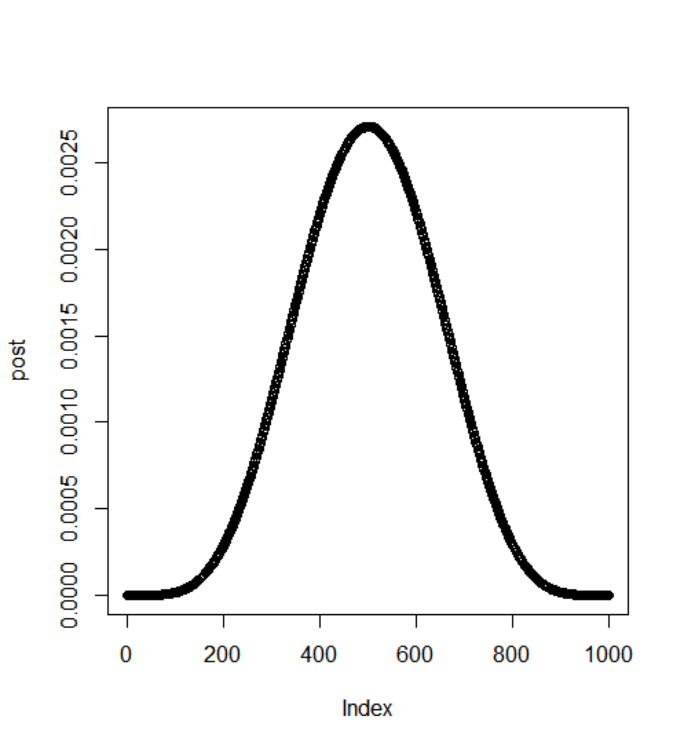
post<-l*prior.uniform
post<-post/sum(post)
plot(post)

La inferencia Bayesiana puede ser difícil

Esta posterior no se puede "producir" para la gran mayoría de los problemas que nos interesa:

- Múltiples parámetros –Altas dimensiones-
- La posterior de un parámetro emerge de la realización simultánea del resto de parámetros

$$V \sim binomial(N, p)$$
$$P \sim uniform(0,1)$$



Podemos crear múltiples grids para casos donde cada parámetro depende de la realización de otros parámetros

Ejemplo

Modelo científico: Estatura influye el Peso

Modelo estadístico: $y_i = \alpha + \beta X_i$

$$y_i \sim Normal(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta X_i$$

Distribución de los datos –Verosimilitud- y Modelo lineal (deterministica)

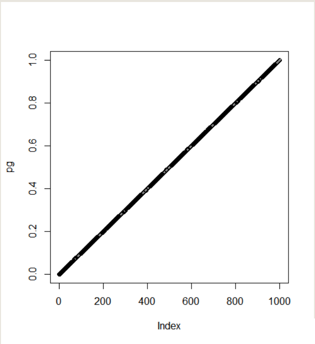
$$\alpha \sim Normal(0,1)$$
$$\beta \sim Normal(0,1)$$
$$\sigma \sim HalfNormal(0,1)$$

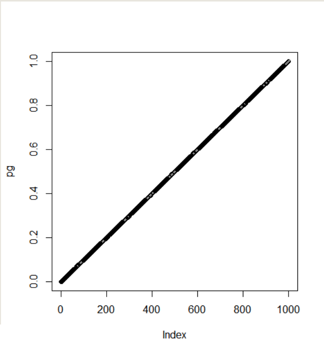
Priors

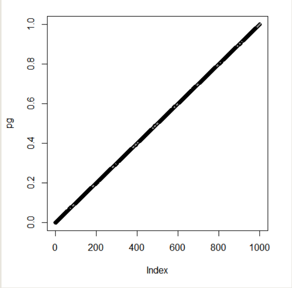
$$\Pr(\alpha, \beta, \sigma | y, x)$$

Posterior conjunta: Los tres parámetros al mismo tiempo

Para 100 valores para los parámetros son 1000000 de cálculos







$$\begin{aligned}\Pr(\alpha, \beta, \sigma | W, H) \propto\ & \mathrm{Normal}(W | \mu, \sigma) \\ & \times \mathrm{Normal}(\alpha | 60,10) \\ & \times \mathrm{LogNormal}(\beta | 0,1) \\ & \times \mathrm{Uniform}(\sigma | 0,10)\end{aligned}$$

La inferencia Bayesiana puede ser difícil

- A medida que aumentan las hipótesis, la construcción de la posterior se complica
- A medida que aumentan los parámetros y las hipótesis, la construcción de la posterior se imposibilita
- ¿Cómo se crean esas distribuciones a partir del uso del teorema de Bayes?

$$p(\theta|D) = p(D|\theta)\, p(\theta) \,/\, p(D)$$

La inferencia Bayesiana puede ser difícil

$$p(\theta|D) \; = \; p(D|\,\theta)\, p(\theta) \,/\, p(D)$$

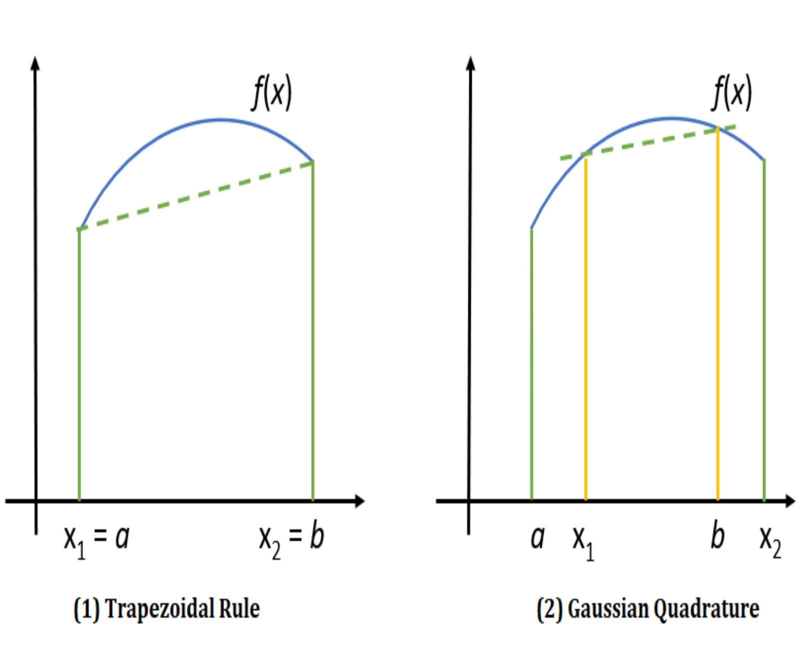
Posterior = verosimilitud * a priori / evidencia

En el caso de parámetros continuos la ecuación puede ser imposible de resolver analíticamente

Si temenos un parámetro con 1000 valores y tenemos DOS parámetros tenemos:

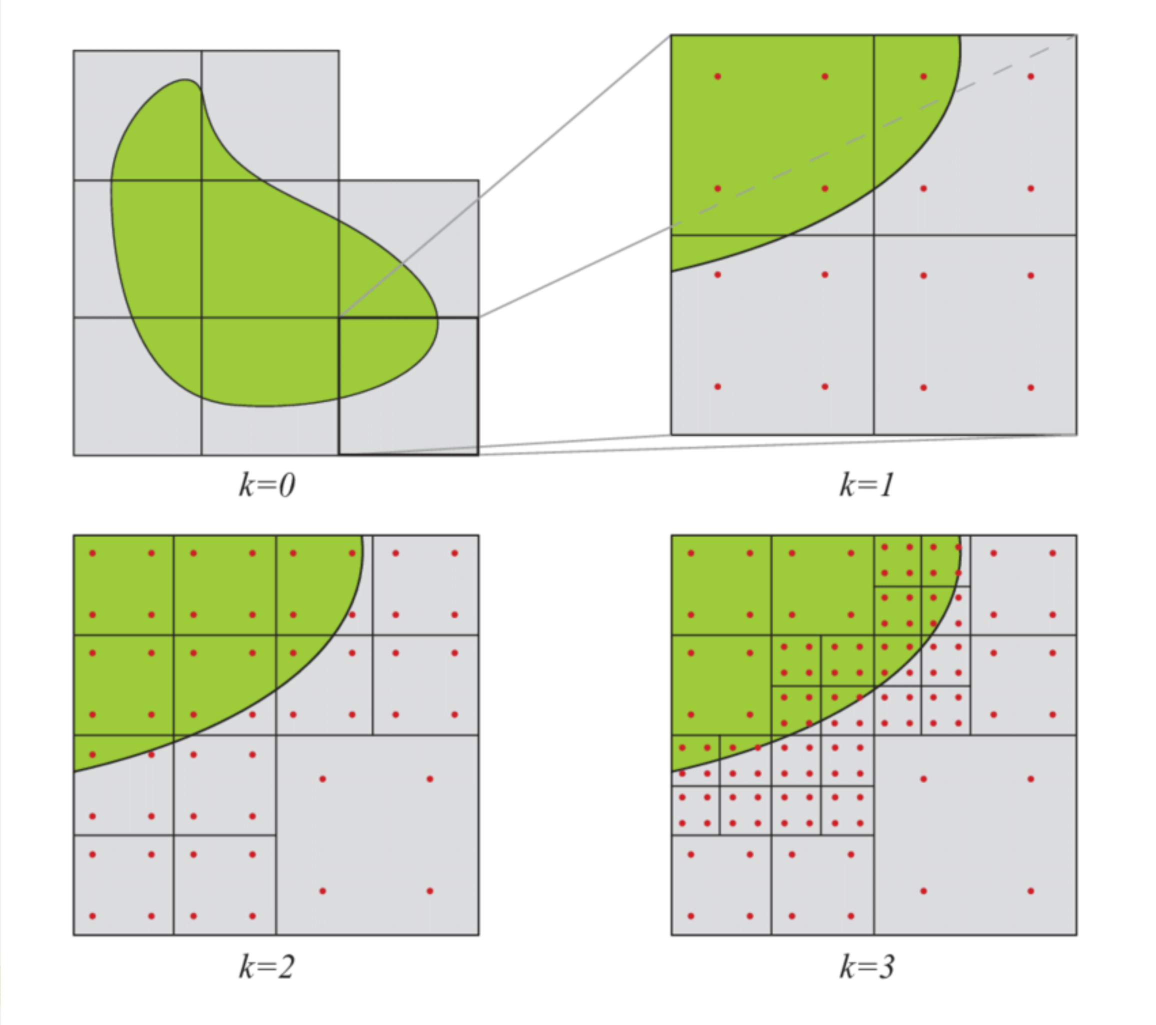
1,000^2 combinaciones de parámetros

Soluciones (?)



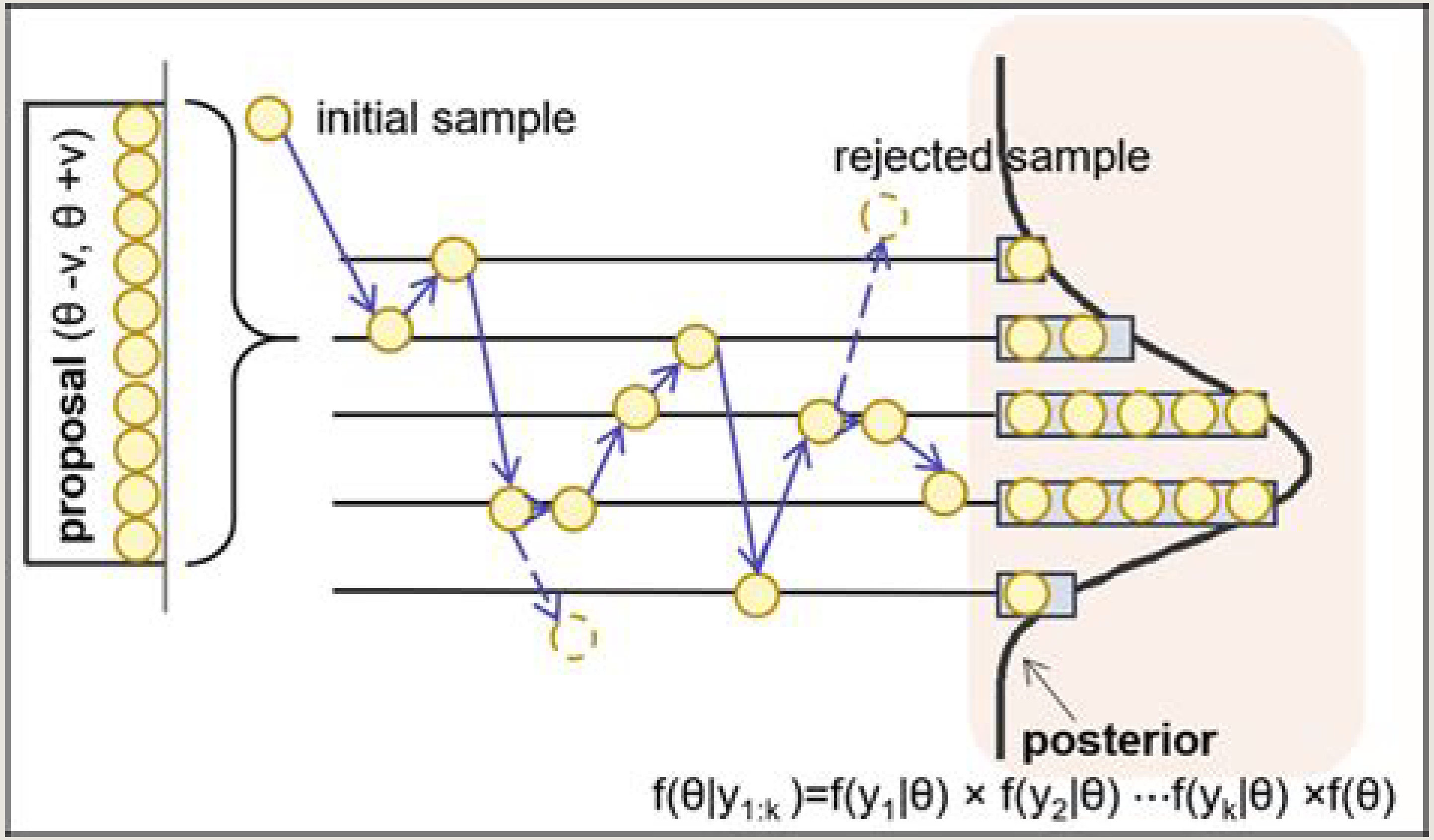
Aproximación vía solución analítica

- Soluciono la integral para algunos segmentos
- ¿Cuáles segementos elegir? Con prueba y error.



Métodos estocásticos

¿Muestreo?



Referencias

Andrew Gelman (2011), "Induction and Deduction in Bayesian Data Analysis", Special Topic: Statistical Science and Philosophy of Science RMM Vol. 2, 2011, 67–78

Bernoulli, J. (1713). *Ars conjectandi, opus posthumum: accedit tractatus de seriebus infinitis, et epistola Gallice scripta de ludo pilæ reticularis*. Impensis Thurnisiorum Fratrum.

Brooks, S. P. (2003). Bayesian computation: a statistical revolution. Philosophical Transactions of the Royal Society of London. Series A: Mathematical, *Physical and Engineering Sciences*, 361(1813), 2681-2697.

Erickson, T. (2017). Beginning Bayes. *Teaching Statistics*, 39(1), 30-35.

Gangur, M., & Svoboda, M. (2018). Simulation of Bayes' rule by means of Monte Carlo method. *Teaching Statistics*, 40(3), 83-87.

Keynes, J. M. A. (1921). *Treatise on Probability*, London: Macmillan.

Kruschke, J. (2014). Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan.

Neyman, J. (1992 [1934]). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. En *Breakthroughs in Statistics* (pp. 123-150). Springer, New York, NY.

Rincón, L. (2007). *Curso intermedio de probabilidad*. UNAM, Facultad de Ciencias.

Rosenkrantz, R. D. (1989). Where do we stand on maximum entropy?(1978). In *ET Jaynes: Papers on probability, statistics and statistical physics* (pp. 210-314). Springer, Dordrecht.

Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in the twentieth century*. Macmillan.

Shafer, G. (1996). The significance of Jacob Bernoulli's Ars Conjectandi for the philosophy of probability today. *Journal of Econometrics*, 75(1), 15-32.

Trafimow, D. (2011). Using pictures to enhance students' understanding of Bayes' theorem. *Teaching Statistics*, 33(3), 83-84.

Anexo de R

- prior.uniform<-rep(1,1000)
- prior.bias1<-dbeta(pg,10,3)
- prior.bias2<-dbeta(pg,3,10)
- prior.bias3<-dbeta(pg,100,100)

- pg<-seq(0,1,len=1000) #Espacio de las posibles soluciones
- l<-dbinom(0,1,prob=pg) #Likelihood –Datos y modelo-

- post<-l*prior.uniform
- post<-post/sum(post)
- plot(post)

CONTACTO

Dr. Héctor Nájera y Dr. Curtis Huffman
Investigadores

Programa Universitario de Estudios del Desarrollo (PUED)
Universidad Nacional Autónoma de México (UNAM)
Antigua Unidad de Posgrado (costado sur de la Torre II de Humanidades), planta baja.
Campus Central, Ciudad Universitaria, Ciudad de México, México.
Tel. (+52) 55 5623 0222, Ext. 82613 y 82616

Tel. (+52) 55 5622 0889
Email: hecatalan@hotmail.com, chuffman@unam.mx

PONTE PUMA
PONTE EL CUBREBOCAS

UNAM La Universidad de la Nación