

# Inferencia Bayesiana

## El problema de las urnas

*Se cuenta con 11 urnas etiquetadas según  $u = 0, 1, \dots, 10$ , que contienen diez bolas cada una. La urna  $u$  contiene  $u$  bolas azules y  $10 - u$  bolas blancas. Fede elige una urna  $u$  al azar y extrae con reposición  $N$  bolas, obteniendo  $n_A$  azules y  $N - n_A$  blancas. Nico, el amigo de Fede, observa atentamente. Si después de  $N = 10$  extracciones resulta  $n_A = 3$ , ¿cuál es la probabilidad de que la urna que Fede está usando sea la  $u$ ?*

La teoría de las probabilidades permite predecir una distribución sobre posibles valores de un resultado dado cierto conocimiento (o estado) del universo: **probabilidad hacia adelante**

La teoría de las probabilidades permite predecir una distribución sobre posibles valores de un resultado dado cierto conocimiento (o estado) del universo: **probabilidad hacia adelante**

Por el contrario, muchas veces estamos interesados en realizar inferencias sobre el estado del universo a partir de observaciones: **probabilidad inversa.**

La teoría de las probabilidades permite predecir una distribución sobre posibles valores de un resultado dado cierto conocimiento (o estado) del universo: **probabilidad hacia adelante**

Por el contrario, muchas veces estamos interesados en realizar inferencias sobre el estado del universo a partir de observaciones: **probabilidad inversa**.

$$p(\mathcal{H} \mid E) = \frac{p(E \mid \mathcal{H})p(\mathcal{H})}{p(E)}$$

$$p(\mathcal{H} \mid E) \propto p(E \mid \mathcal{H})p(\mathcal{H})$$

Conociendo  $N$ , si conociéramos  $u$  podríamos calcular las probabilidades de los diferentes  $n_A$ : **probabilidad hacia adelante**.

Conociendo  $N$ , si conociéramos  $u$  podríamos calcular las probabilidades de los diferentes  $n_A$ : **probabilidad hacia adelante**.

Aquí observamos un  $n_A$  y queremos calcular las probabilidades de los posibles valores de  $u$ : **probabilidad inversa**.

Conociendo  $N$ , si conociéramos  $u$  podríamos calcular las probabilidades de los diferentes  $n_A$ : **probabilidad hacia adelante**.

Aquí observamos un  $n_A$  y queremos calcular las probabilidades de los posibles valores de  $u$ : **probabilidad inversa**.

$$p(u \mid n_A, N) = \frac{p(n_A \mid u, N)p(u)}{p(n_A \mid N)}$$

- ▶  $N$  es una cantidad fija
- ▶  $n_A$  es otra cantidad fija: lo que observamos al realizar el experimento
- ▶  $u$  es la cantidad desconocida



Probabilidad conjunta de las cantidades observables (datos) y cantidades no observables (parámetros):

Probabilidad conjunta de las cantidades observables (datos) y cantidades no observables (parámetros):

$$p(u, n_A \mid N) = p(n_A \mid u, N)p(u)$$

Probabilidad conjunta de las cantidades observables (datos) y cantidades no observables (parámetros):

$$p(u, n_A \mid N) = p(n_A \mid u, N)p(u)$$

Podemos escribir la probabilidad de  $u$  condicionada a  $n_A$ :

$$\begin{aligned} p(u \mid n_A, N) &= \frac{p(u, n_A \mid N)}{p(n_A \mid N)} \\ &= \frac{p(n_A \mid u, N)p(u)}{p(n_A \mid N)} \end{aligned}$$

Probabilidad conjunta de las cantidades observables (datos) y cantidades no observables (parámetros):

$$p(u, n_A \mid N) = p(n_A \mid u, N)p(u)$$

Podemos escribir la probabilidad de  $u$  condicionada a  $n_A$ :

$$\begin{aligned} p(u \mid n_A, N) &= \frac{p(u, n_A \mid N)}{p(n_A \mid N)} \\ &= \frac{p(n_A \mid u, N)p(u)}{p(n_A \mid N)} \end{aligned}$$

Es la probabilidad de cada valor de  $u$  luego de haber observado  $n_A = 3$  bolas azules

La probabilidad marginal de  $u$  es

La probabilidad marginal de  $u$  es

$$p(u) = \frac{1}{11}$$

Es la probabilidad inicial de haber tomado la urna  $u$

La probabilidad de  $n_A$  dado  $u$  (y  $N$ ) es:

La probabilidad de  $n_A$  dado  $u$  (y  $N$ ) es:

$$p(n_A \mid u, N) = \binom{N}{n_A} \left(\frac{u}{10}\right)^{n_A} \left(1 - \frac{u}{10}\right)^{N-n_A}$$



La probabilidad de  $n_A$  dado  $u$  (y  $N$ ) es:

$$p(n_A | u, N) = \binom{N}{n_A} \left(\frac{u}{10}\right)^{n_A} \left(1 - \frac{u}{10}\right)^{N-n_A}$$

Como  $n_A = 3$  es fijo (¡son los datos observados!),  $p(n_A | u, N)$  es una función de  $u$ . Indica qué tan compatibles son los datos observados con los distintos valores de  $u$

El denominador,  $p(n_A \mid N) = p(n_A)$ , es

El denominador,  $p(n_A \mid N) = p(n_A)$ , es

$$\begin{aligned} p(n_A \mid N) &= \sum_u p(u, n_A \mid N) \\ &= \sum_u p(n_A \mid u, N) p(u) \\ &= \frac{1}{11} \sum_u p(n_A \mid u, N) \end{aligned}$$

Finalmente, la probabilidad de interés  $p(u \mid n_A, N)$  es

Finalmente, la probabilidad de interés  $p(u \mid n_A, N)$  es

$$p(u \mid n_A, N) = \frac{p(n_A \mid u, N)p(u)}{p(n_A \mid N)}$$

Finalmente, la probabilidad de interés  $p(u \mid n_A, N)$  es

$$p(u \mid n_A, N) = \frac{p(n_A \mid u, N)p(u)}{p(n_A \mid N)}$$

$$p(u \mid n_A, N) = \binom{N}{n_A} \left(\frac{u}{10}\right)^{n_A} \left(1 - \frac{u}{10}\right)^{N-n_A} \frac{1}{11} \frac{1}{p(n_A \mid N)}$$

Finalmente, la probabilidad de interés  $p(u \mid n_A, N)$  es

$$p(u \mid n_A, N) = \frac{p(n_A \mid u, N)p(u)}{p(n_A \mid N)}$$

$$p(u \mid n_A, N) = \binom{N}{n_A} \left(\frac{u}{10}\right)^{n_A} \left(1 - \frac{u}{10}\right)^{N-n_A} \frac{1}{11} \frac{1}{p(n_A \mid N)}$$

- ▶  $N$  es una cantidad fija
- ▶  $n_A$  es 3, otra cantidad fija: lo que observamos al realizar el experimento
- ▶  $u$  es la cantidad desconocida

Finalmente, la probabilidad de interés  $p(u \mid n_A, N)$  es

$$p(u \mid n_A, N) = \frac{p(n_A \mid u, N)p(u)}{p(n_A \mid N)}$$

$$p(u \mid n_A, N) = \binom{N}{n_A} \left(\frac{u}{10}\right)^{n_A} \left(1 - \frac{u}{10}\right)^{N-n_A} \frac{1}{11} \frac{1}{p(n_A \mid N)}$$

- ▶  $N$  es una cantidad fija
- ▶  $n_A$  es 3, otra cantidad fija: lo que observamos al realizar el experimento
- ▶  $u$  es la cantidad desconocida

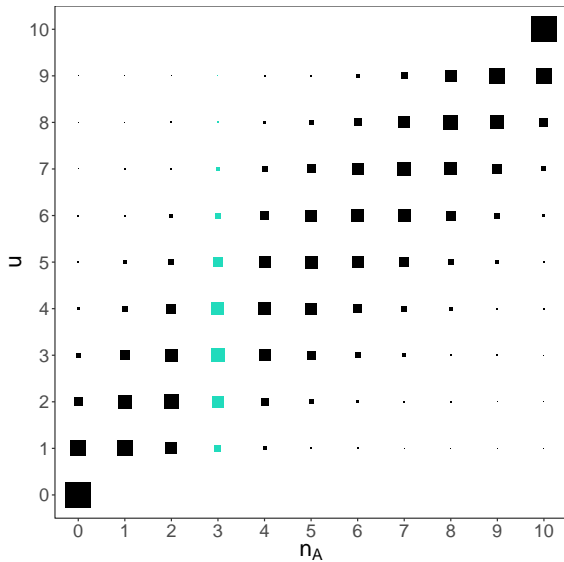
$p(u \mid n_A, N)$  es una función de  $u$ : es la credibilidad de los valores de  $u$  luego de observar los datos (es decir, condicionada a  $n_A = 3$ ).



Gráficamente...



Gráficamente...



Pasamos de una credibilidad *a priori* antes de observar los datos, a una *a posteriori* luego de observar  $n_A = 3$



## Intolerancia al gluten

*¿Pueden las personas alérgicas al gluten distinguir harina común de harina sin gluten en un ensayo ciego? En un experimento, de 35 sujetos, 12 identificaron correctamente la harina común y 23 se equivocaron o no supieron decir de qué harina se trataba.*

*Incluso si no hubiera alérgicos al gluten en el experimento, esperaríamos encontrar algunas identificaciones correctas... Basándonos en el número de identificaciones correctas, ¿cuántos de los sujetos son alérgicos al gluten y cuántos estaban adivinando?*

## Intolerancia al gluten

*¿Pueden las personas alérgicas al gluten distinguir harina común de harina sin gluten en un ensayo ciego? En un experimento, de 35 sujetos, 12 identificaron correctamente la harina común y 23 se equivocaron o no supieron decir de qué harina se trataba.*

*Incluso si no hubiera alérgicos al gluten en el experimento, esperaríamos encontrar algunas identificaciones correctas... Basándonos en el número de identificaciones correctas, ¿cuántos de los sujetos son alérgicos al gluten y cuántos estaban adivinando?*

Supongamos que una persona alérgica al gluten tiene una probabilidad de 0.90 de detectar la harina común mientras que una persona sin alergia detecta harina común con una probabilidad de 0.40 (y con una probabilidad de 0.6 se equivoca o no sabe decir).

Llamemos:

Llamemos:

- ▶  $N$  a la cantidad total de personas en el ensayo
- ▶  $N_a$  al número de personas alérgicas al gluten
- ▶  $\pi_a$  a la probabilidad de que un alérgico identifique correctamente
- ▶  $\pi_f$  a la probabilidad de que un no alérgico identifique correctamente
- ▶  $n_i$  al número de identificaciones correctas

Llamemos:

- ▶  $N$  a la cantidad total de personas en el ensayo
- ▶  $N_a$  al número de personas alérgicas al gluten
- ▶  $\pi_a$  a la probabilidad de que un alérgico identifique correctamente
- ▶  $\pi_f$  a la probabilidad de que un no alérgico identifique correctamente
- ▶  $n_i$  al número de identificaciones correctas

¿Cuáles son las cantidades conocidas? ¿Cuáles son las cantidades desconocidas? ¿Cómo es el modelo de probabilidad hacia adelante?  
¿Cómo es el problema inverso?



Conociendo  $N$ ,  $\pi_a$  y  $\pi_f$ , si conociéramos  $N_a$  podríamos calcular las probabilidades de los diferentes  $n_i$ : **probabilidad hacia adelante**

Conociendo  $N$ ,  $\pi_a$  y  $\pi_f$ , si conociéramos  $N_a$  podríamos calcular las probabilidades de los diferentes  $n_i$ : **probabilidad hacia adelante**

Aquí observamos  $n_i$  y queremos realizar inferencias sobre  $N_a$ : **probabilidad inversa**

Digamos que *a priori* cualquier número de  $N_a$  es igualmente probable o esperable:

Digamos que *a priori* cualquier número de  $N_a$  es igualmente probable o esperable:

$$p(N_a) = \frac{1}{36}$$

¿Cómo construimos la verosimilitud de los diferentes valores de  $N_a$   
 $p(n_i \mid N_a)$ ?

¿Cómo construimos la verosimilitud de los diferentes valores de  $N_a$   
 $p(n_i \mid N_a)$ ?

Pensemos de forma **generativa** (con el modelo de **probabilidad hacia adelante**). Imaginemos que conocemos  $N_a$  (además de  $N$ ,  $\pi_a$  y  $\pi_f$ ), ¿podríamos escribir un programa que simule diferentes valores de  $n_i$ ?

El número de identificaciones correctas  $n_i$  es la suma de las identificaciones correctas entre los  $N_a$  alérgicos ( $n_{ia}$ ) y los  $N - N_a$  no alérgicos ( $n_{if}$ ). ¿Cuántas identificaciones habrá en cada grupo?

El número de identificaciones correctas  $n_i$  es la suma de las identificaciones correctas entre los  $N_a$  alérgicos ( $n_{ia}$ ) y los  $N - N_a$  no alérgicos ( $n_{if}$ ). ¿Cuántas identificaciones habrá en cada grupo?

$$n_{ia} \sim Bi(N_a, \pi_a)$$

$$n_{if} \sim Bi(N - N_a, \pi_f)$$

$$n_i = n_{ia} + n_{if}$$



El número de identificaciones correctas  $n_i$  es la suma de las identificaciones correctas entre los  $N_a$  alérgicos ( $n_{ia}$ ) y los  $N - N_a$  no alérgicos ( $n_{if}$ ). ¿Cuántas identificaciones habrá en cada grupo?

$$n_{ia} \sim Bi(N_a, \pi_a)$$

$$n_{if} \sim Bi(N - N_a, \pi_f)$$

$$n_i = n_{ia} + n_{if}$$

```
N <- 35
pi_a <- 0.9
pi_f <- 0.4
N_a <- 10 # lo suponemos conocido para simular

n_ia <- rbinom(1, N_a, pi_a)
n_if <- rbinom(1, N-N_a, pi_f)

n_i <- n_ia + n_if
```

El número de identificaciones correctas  $n_i$  es la suma de las identificaciones correctas entre los  $N_a$  alérgicos ( $n_{ia}$ ) y los  $N - N_a$  no alérgicos ( $n_{if}$ ). ¿Cuántas identificaciones habrá en cada grupo?

$$n_{ia} \sim Bi(N_a, \pi_a)$$

$$n_{if} \sim Bi(N - N_a, \pi_f)$$

$$n_i = n_{ia} + n_{if}$$

```
N <- 35
pi_a <- 0.9
pi_f <- 0.4
N_a <- 10 # lo suponemos conocido para simular

n_ia <- rbinom(1, N_a, pi_a)
n_if <- rbinom(1, N-N_a, pi_f)

n_i <- n_ia + n_if
```

Sabríamos calcular las probabilidades de los diferentes valores de  $n_{ia}$  y  $n_{if}$ , ¿no?.

Recordemos que no conocemos  $N_a$ . En nuestro caso, la verosimilitud de cada valor de  $N_a$  es la probabilidad de observar  $n_i = 12$  para ese valor de  $N_a$ .

Recordemos que no conocemos  $N_a$ . En nuestro caso, la verosimilitud de cada valor de  $N_a$  es la probabilidad de observar  $n_i = 12$  para ese valor de  $N_a$ .

$$\begin{aligned} p(n_i = 12 \mid N_a) &= p(n_{ia} = 0 \mid N_a)p(n_{if} = 12 \mid N_a) \\ &\quad + p(n_{ia} = 1 \mid N_a)p(n_{if} = 11 \mid N_a) + \dots \end{aligned}$$

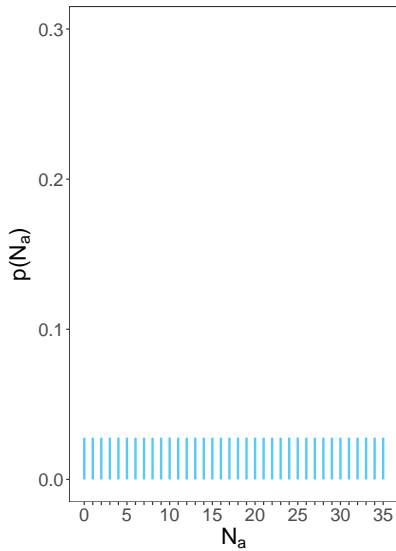
Recordemos que no conocemos  $N_a$ . En nuestro caso, la verosimilitud de cada valor de  $N_a$  es la probabilidad de observar  $n_i = 12$  para ese valor de  $N_a$ .

$$p(n_i = 12 \mid N_a) = p(n_{ia} = 0 \mid N_a)p(n_{if} = 12 \mid N_a) \\ + p(n_{ia} = 1 \mid N_a)p(n_{if} = 11 \mid N_a) + \dots$$

Queda como ejercicio calcular a mano  $p(n_i \mid N_a)$  o, mejor aún, escribir un programita que calcule  $p(n_i \mid N_a)$

Finalmente,

$$p(N_a \mid n_i) = \frac{p(n_i \mid N_a)p(N_a)}{p(n_i)}$$



## Vocabulario limitado

*Supongamos que existe un idioma con seis palabras:*

*$\{\text{perro, parra, farra, carro, corro, tarro}\}$*

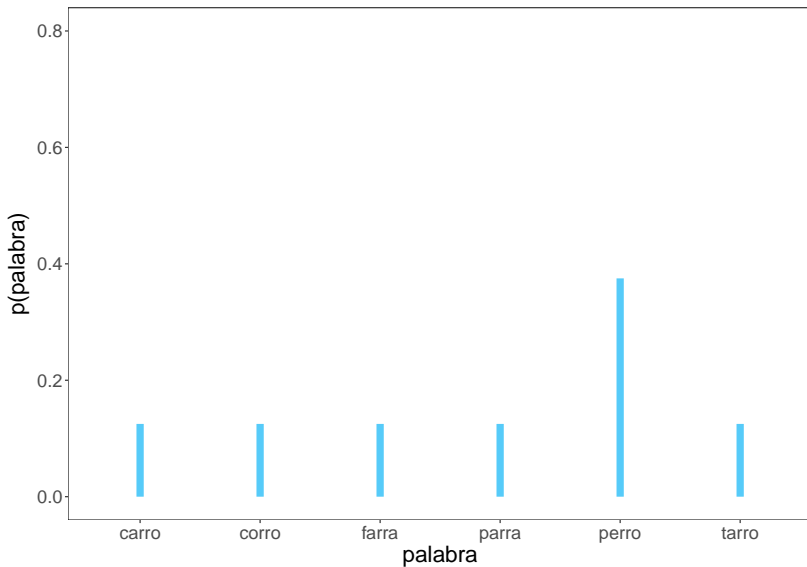
- ▶ Todas las palabras son igualmente probables, excepto por 'perro', que es  $\alpha = 3$  veces más probable que las otras.
- ▶ Cuando se tipean, un caracter se introduce erróneamente con probabilidad  $\pi = 0.1$ .
- ▶ Todas las letras tienen la misma probabilidad de producir un error de tipeo.
- ▶ Si una letra se tipeó mal, la probabilidad de cometer un error en otro caracter no cambia.
- ▶ Los errores son independientes a lo largo de una palabra.



- i. ¿Cuál es la probabilidad de escribir correctamente 'tarro'?
- ii. ¿Cuál es la probabilidad de tipear 'cerro' o 'curro' al querer escribir 'carro'?
- iii. Desarrollar un corrector gramatical para esta lengua: para las palabras tipeadas 'farra', 'birra' y 'locos', ¿cuál es la palabra que se quiso escribir?

- i. La probabilidad de escribir correctamente 'tarro' es  $(1 - \pi)^5$
- ii. La probabilidad de escribir correctamente 'cerro' o 'curro' al querer escribir 'carro' es  $\pi(1 - \pi)^4$
- iii. Allá vamos...

Estas son las probabilidades *a priori* de cada una de las palabras del vocabulario



Alguien escribe 'farra', ¿qué quiso escribir?

Alguien escribe 'farra', ¿qué quiso escribir?

¿Qué sería en este caso la verosimilitud?

Alguien escribe 'farra', ¿qué quiso escribir?

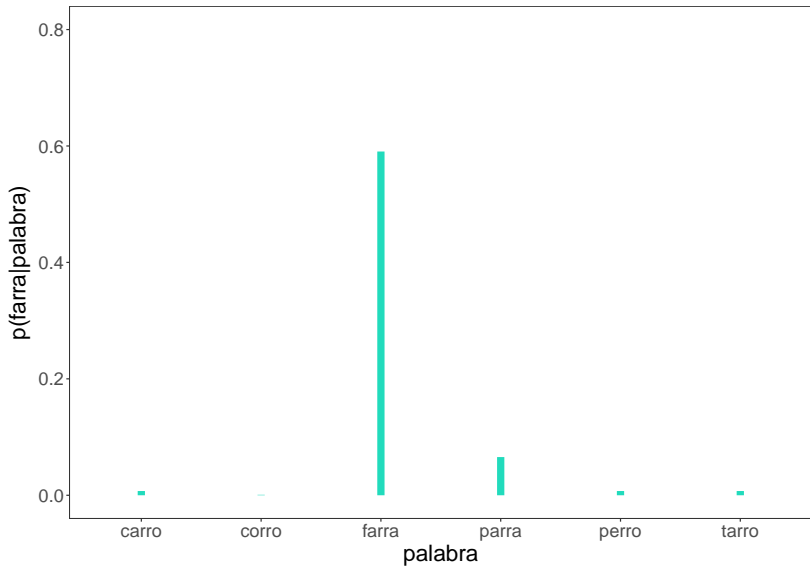
¿Qué sería en este caso la verosimilitud?

La verosimilitud de 'perro' es qué tan probable es escribir 'farra' cuando se quería escribir 'perro':  $p(\text{farra} \mid \text{perro}) = \pi^3(1 - \pi)^2$

Alguien escribe 'farra', ¿qué quiso escribir?

¿Qué sería en este caso la verosimilitud?

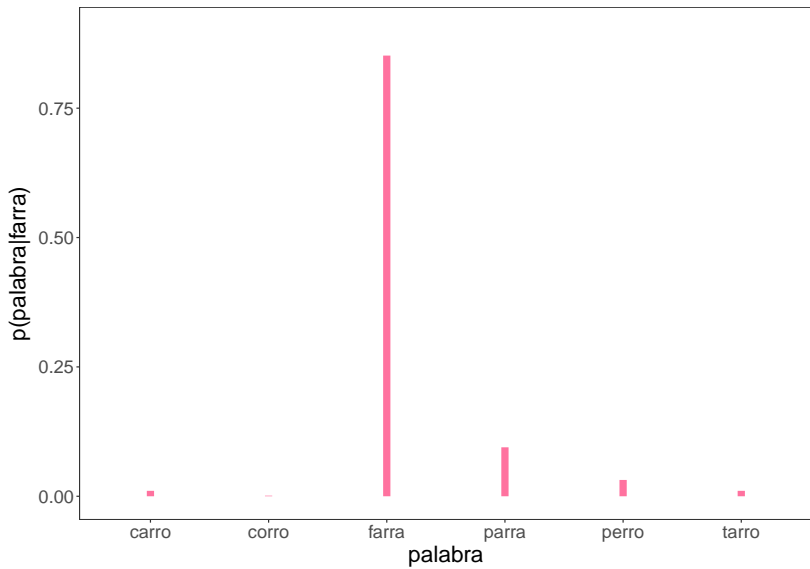
La verosimilitud de 'perro' es qué tan probable es escribir 'farra' cuando se quería escribir 'perro':  $p(\text{farra} \mid \text{perro}) = \pi^3(1 - \pi)^2$





Para obtener la probabilidad *a posteriori* de cada palabra, necesitamos combinar la información *a priori* con los datos (¿cuáles son los datos?). Aplicamos la Regla de Bayes:

$$p(\text{palabra} \mid \text{farra}) = \frac{p(\text{farra} \mid \text{palabra})p(\text{palabra})}{p(\text{farra})}$$



La inferencia bayesiana es la realocación de la credibilidad del conjunto de cantidades desconocidas (parámetros) de un modelo, una vez observado un conjunto de datos.

## Pequeño mundo

*Se desea estimar la proporción de agua que cubre el planeta Tierra. Para ello se arroja hacia arriba un “globo terráqueo antiestrés” y se registra la posición del dedo índice al volver a tomarlo.*

*Se arroja el globo 11 veces hacia arriba y se obtiene la siguiente secuencia:*

*TAAATT AATAA*

Llamemos:

Llamemos:

- ▶  $\pi$  a la proporción de agua en el planeta Tierra
- ▶  $N$  al número de tiradas
- ▶  $y$  al número de veces que salió agua

Llamemos:

- ▶  $\pi$  a la proporción de agua en el planeta Tierra
- ▶  $N$  al número de tiradas
- ▶  $y$  al número de veces que salió agua

$\pi$  es una cantidad continua entre 0 y 1. Esta vez no la discretizaremos.

*Prior*



## Prior

¿Cómo asignamos una credibilidad *a priori* para los valores de  $\pi_a$ ?

## Prior

¿Cómo asignamos una credibilidad *a priori* para los valores de  $\pi_a$ ?

Con una distribución de probabilidad.

## Prior

¿Cómo asignamos una credibilidad *a priori* para los valores de  $\pi_a$ ?

Con una distribución de probabilidad.

$$\pi \sim \text{Beta}(a, b)$$

## Prior

¿Cómo asignamos una credibilidad *a priori* para los valores de  $\pi_a$ ?

Con una distribución de probabilidad.

$$\pi \sim \text{Beta}(a, b)$$

$$p(\pi \mid a, b) = p(\pi) = \frac{\pi^{a-1}(1-\pi)^{b-1}}{B(a, b)}$$

## Prior

¿Cómo asignamos una credibilidad *a priori* para los valores de  $\pi_a$ ?

Con una distribución de probabilidad.

$$\pi \sim \text{Beta}(a, b)$$

$$p(\pi \mid a, b) = p(\pi) = \frac{\pi^{a-1}(1 - \pi)^{b-1}}{B(a, b)}$$

$$B(a, b) = \int_0^1 \pi^{a-1}(1 - \pi)^{b-1} d\pi = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)}$$

## Prior

¿Cómo asignamos una credibilidad *a priori* para los valores de  $\pi_a$ ?

Con una distribución de probabilidad.

$$\pi \sim \text{Beta}(a, b)$$

$$p(\pi \mid a, b) = p(\pi) = \frac{\pi^{a-1}(1-\pi)^{b-1}}{B(a, b)}$$

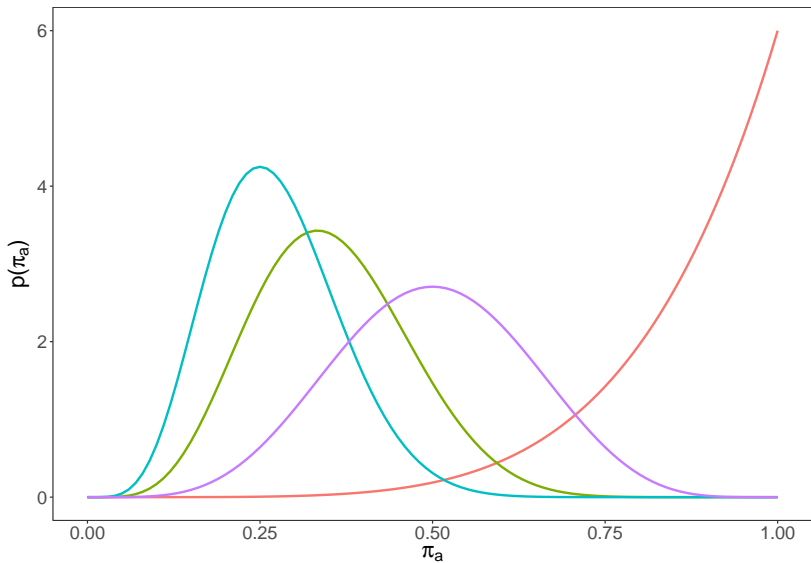
$$B(a, b) = \int_0^1 \pi^{a-1}(1-\pi)^{b-1} d\pi = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du$$

## La distribución beta

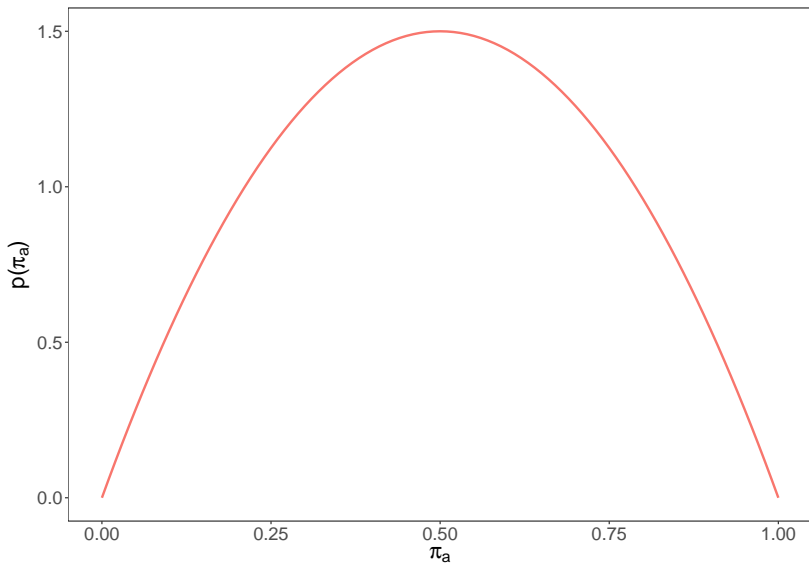


## La distribución beta





Una posible elección de valores para la distribución *a priori* es  $\text{Beta}(2, 2)$



## *Likelihood*

¿Cuál es la probabilidad de observar los datos que observamos para diferentes valores del parámetro?

## Likelihood

¿Cuál es la probabilidad de observar los datos que observamos para diferentes valores del parámetro?

$$Y \mid \pi, N \sim Bi(N, \pi)$$

$$p(y \mid \pi, N) = \binom{N}{y} \pi^y (1 - \pi)^{N-y} = p(y \mid \pi)$$

## Posterior

$$p(\pi \mid y) = \frac{p(y \mid \pi)p(\pi)}{p(y)}$$

## Posterior

$$p(\pi \mid y) = \frac{p(y \mid \pi)p(\pi)}{p(y)}$$

$$p(\pi \mid y) = \frac{\binom{N}{y} \pi^y (1 - \pi)^{N-y} \frac{\pi^{a-1} (1-\pi)^{b-1}}{B(a,b)}}{\int p(y \mid \pi) p(\pi) d\pi}$$

## Posterior

$$p(\pi \mid y) = \frac{p(y \mid \pi)p(\pi)}{p(y)}$$

$$p(\pi \mid y) = \frac{\binom{N}{y} \pi^y (1 - \pi)^{N-y} \frac{\pi^{a-1} (1-\pi)^{b-1}}{B(a,b)}}{\int p(y \mid \pi) p(\pi) d\pi}$$

La integral en el denominador suele ser un problema. Con dos parámetros es una integral doble, con tres parámetros, una triple, etc. Esta integral puede ser intratable (*intractable*) (no tener solución exacta, analítica, cerrada). No hay vaca vestida de uniforme que nos salve.

Recordando que:

$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

Recordando que:

$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

Resulta

$$p(\pi \mid y) \propto p(y \mid \pi)p(\pi)$$



Recordando que:

$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

Resulta

$$p(\pi \mid y) \propto p(y \mid \pi)p(\pi)$$

$$p(\pi \mid y) \propto \binom{N}{y} \pi^y (1 - \pi)^{N-y} \frac{1}{B(a, b)} \pi^{a-1} (1 - \pi)^{b-1}$$

Recordando que:

$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

Resulta

$$p(\pi \mid y) \propto p(y \mid \pi)p(\pi)$$

$$p(\pi \mid y) \propto \binom{N}{y} \pi^y (1 - \pi)^{N-y} \frac{1}{B(a, b)} \pi^{a-1} (1 - \pi)^{b-1}$$

$$p(\pi \mid y) \propto \binom{N}{y} \frac{1}{B(a, b)} \pi^{(y+a)-1} (1 - \pi)^{(N-y+b)-1}$$

Recordando que:

$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

Resulta

$$p(\pi \mid y) \propto p(y \mid \pi)p(\pi)$$

$$p(\pi \mid y) \propto \binom{N}{y} \pi^y (1 - \pi)^{N-y} \frac{1}{B(a, b)} \pi^{a-1} (1 - \pi)^{b-1}$$

$$p(\pi \mid y) \propto \binom{N}{y} \frac{1}{B(a, b)} \pi^{(y+a)-1} (1 - \pi)^{(N-y+b)-1}$$

$$p(\pi \mid y) = KC \pi^{(y+a)-1} (1 - \pi)^{(N-y+b)-1}$$

Recordando que:

$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

Resulta

$$p(\pi \mid y) \propto p(y \mid \pi)p(\pi)$$

$$p(\pi \mid y) \propto \binom{N}{y} \pi^y (1 - \pi)^{N-y} \frac{1}{B(a, b)} \pi^{a-1} (1 - \pi)^{b-1}$$

$$p(\pi \mid y) \propto \binom{N}{y} \frac{1}{B(a, b)} \pi^{(y+a)-1} (1 - \pi)^{(N-y+b)-1}$$

$$p(\pi \mid y) = KC \pi^{(y+a)-1} (1 - \pi)^{(N-y+b)-1}$$

$$p(\pi \mid y) = K^* \pi^{(y+a)-1} (1 - \pi)^{(N-y+b)-1}$$

Para que  $\int_0^1 p(\pi | y) d\pi = 1$ , debe ser

Para que  $\int_0^1 p(\pi \mid y) d\pi = 1$ , debe ser

$$K^* = \frac{1}{B(y+a, N-y+b)} = \frac{\Gamma[(y+a) + (N-y+b)]}{\Gamma(y+a)\Gamma(N-y+b)}$$

Para que  $\int_0^1 p(\pi | y) d\pi = 1$ , debe ser

$$K^* = \frac{1}{B(y+a, N-y+b)} = \frac{\Gamma[(y+a) + (N-y+b)]}{\Gamma(y+a)\Gamma(N-y+b)}$$

Por lo tanto, resulta que la distribución *a posteriori* es Beta de parámetros  $y+a$  y  $N-y+b$

$$p(\pi | y) = \frac{\pi^{(y+a)-1} (1-\pi)^{(N-y+b)-1}}{B(y+a, N-y+b)}$$

$$\pi | y \sim \text{Beta}(y+a, N-y+b)$$

¿Qué hicimos?



## ¿Qué hicimos?

Nos las arreglamos para encontrar la solución exacta al problema de inferir el parámetro de una distribución binomial a partir del número de éxitos observados.

## ¿Qué hicimos?

Nos las arreglamos para encontrar la solución exacta al problema de inferir el parámetro de una distribución binomial a partir del número de éxitos observados.

El *prior* y el *posterior* tienen la misma forma distribucional. Esto ocurre por la elección del *prior* y el *likelihood*.

Una distribución  $\mathcal{F}$  se dice conjugada de una verosimilitud  $\mathcal{L}$  si cuando la distribución *a priori* es  $\mathcal{F}$ , la distribución *a posteriori* también es  $\mathcal{F}$

## Pequeño mundo

*Se desea estimar la proporción de agua que cubre el planeta Tierra. Para ello se arroja hacia arriba un “globo terráqueo antiestrés” y se registra la posición del dedo índice al volver a tomarlo.*

*Se arroja el globo 11 veces hacia arriba y se obtiene la siguiente secuencia:*

*TAAATTAAATAA*

## Pequeño mundo

*Se desea estimar la proporción de agua que cubre el planeta Tierra. Para ello se arroja hacia arriba un “globo terráqueo antiestrés” y se registra la posición del dedo índice al volver a tomarlo.*

*Se arroja el globo 11 veces hacia arriba y se obtiene la siguiente secuencia:*

*TAAATTAAATAA*

$$Y \mid \pi \sim \text{Binomial}(N, \pi)$$

$$\pi \sim \text{Beta}(a, b)$$

con  $N = 11$ ,  $a = 2$  y  $b = 2$ .

Al observar  $y = 7$  resulta

$$\pi \mid y \sim \text{Beta}(a + y, b + N - y)$$

$$p(\pi \mid y) = \text{Beta}(2 + 7, 2 + 4)$$



## Más ejemplos

Queremos estimar la probabilidad  $\pi$  de que salga cara al arrojar una moneda.



## Más ejemplos

Queremos estimar la probabilidad  $\pi$  de que salga cara al arrojar una moneda.

Credibilidad *a priori*:  $\text{Beta}(2, 2)$

## Más ejemplos

Queremos estimar la probabilidad  $\pi$  de que salga cara al arrojar una moneda.

Credibilidad *a priori*:  $\text{Beta}(2, 2)$

¿Cómo cambia nuestra creencia si...

1. ...realizamos 6 tiradas y observamos 4 caras?
2. ...realizamos 60 tiradas y observamos 40 caras?
3. ...realizamos 2 tiradas y observamos 2 caras?
4. ...realizamos 40 tiradas y observamos 40 caras?
5. ...realizamos 4 tiradas y obtenemos 3 caras y luego realizamos 2 tiradas más y observamos 1 caras?

1.  $\pi \mid y \sim \text{Beta}(2 + 4, 2 + 2)$
2.  $\pi \mid y \sim \text{Beta}(2 + 40, 2 + 20)$
3.  $\pi \mid y \sim \text{Beta}(2 + 2, 2 + 0)$
4.  $\pi \mid y \sim \text{Beta}(2 + 40, 2 + 0)$
5.  $\pi \mid y \sim \text{Beta}((2 + 3) + 1, (2 + 1) + 1)$

4 caras en 6 tiradas



40 caras en 60 tiradas



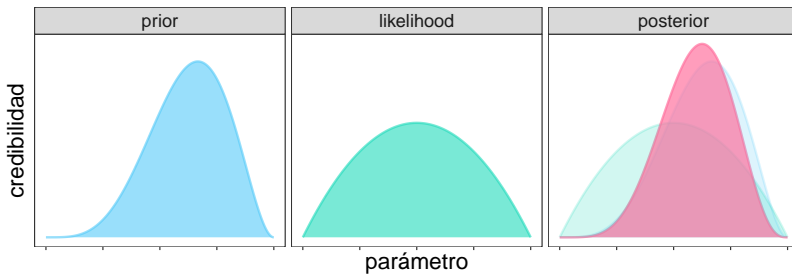
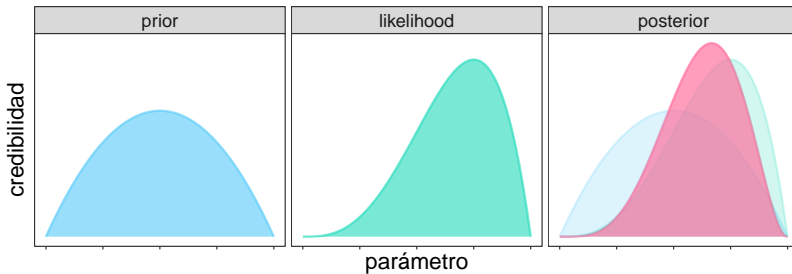
2 caras en 2 tiradas



40 caras en 40 tiradas



3 caras en 4 tiradas, luego 1 cara en 2 tiradas





# Características generales

La inferencia bayesiana presenta ciertas características que se repiten independientemente de las distribuciones elegidas.

## Compromiso

Vamos a formalizar lo que observamos en el ejemplo para el modelo Beta–Binomial. Para esto será útil el siguiente resultado:

Si  $X \sim \text{Beta}(a, b)$

$$\mathbb{E}(X) = \frac{a}{a+b}$$

La distribución *a priori* es  $\text{Beta}(a, b)$  y la distribución *a posteriori* es  $\text{Beta}(y + a, N - y + b)$ .

La distribución *a priori* es  $\text{Beta}(a, b)$  y la distribución *a posteriori* es  $\text{Beta}(y + a, N - y + b)$ . La media del *posterior* es:

$$\begin{aligned}\mathbb{E}[p(\pi \mid y)] &= \frac{y + a}{a + b + N} \\&= \frac{y}{a + b + N} + \frac{a}{a + b + N} \\&= \frac{N}{a + b + N} \frac{y}{N} + \frac{a + b}{a + b + N} \frac{a}{a + b} \\&= \frac{N}{a + b + N} \frac{y}{N} + \frac{a + b}{a + b + N} \mathbb{E}[p(\pi)]\end{aligned}$$

La distribución *a posteriori* representa un balance (promedio ponderado o *combinación convexa*) entre la proporción observada y la proporción esperada *a priori*. Hay un *shrinkage* hacia la media del *prior*.

## Secuencialidad

Si primero observamos  $y_1$  en  $N_1$  y luego observamos  $y_2$  en  $N_2$ ...  
Con el primer conjunto de datos pasamos del *prior* al *posterior* y luego esa distribución se convierte en el nuevo *prior*:

## Secuencialidad

Si primero observamos  $y_1$  en  $N_1$  y luego observamos  $y_2$  en  $N_2$ ...  
Con el primer conjunto de datos pasamos del *prior* al *posterior* y luego esa distribución se convierte en el nuevo *prior*:

$$\text{Beta}(a, b) \rightarrow \text{Beta}(y_1 + a, N_1 - y_1 + b)$$

## Secuencialidad

Si primero observamos  $y_1$  en  $N_1$  y luego observamos  $y_2$  en  $N_2$ ...  
Con el primer conjunto de datos pasamos del *prior* al *posterior* y luego esa distribución se convierte en el nuevo *prior*:

$$\text{Beta}(a, b) \rightarrow \text{Beta}(y_1 + a, N_1 - y_1 + b)$$

$$\text{Beta}(y_1 + a, N_1 - y_1 + b) \rightarrow \text{Beta}(y_2 + y_1 + a, N_2 - y_2 + N_1 - y_1 + b)$$



## Secuencialidad

Si primero observamos  $y_1$  en  $N_1$  y luego observamos  $y_2$  en  $N_2$ ...  
Con el primer conjunto de datos pasamos del *prior* al *posterior* y luego esa distribución se convierte en el nuevo *prior*:

$$\text{Beta}(a, b) \rightarrow \text{Beta}(y_1 + a, N_1 - y_1 + b)$$

$$\text{Beta}(y_1 + a, N_1 - y_1 + b) \rightarrow \text{Beta}(y_2 + y_1 + a, N_2 - y_2 + N_1 - y_1 + b)$$

$$\text{Beta}(a, b) \rightarrow \text{Beta}((y_1 + y_2) + a, (N_1 + N_2) - (y_1 + y_2) + b)$$

## Secuencialidad

Si primero observamos  $y_1$  en  $N_1$  y luego observamos  $y_2$  en  $N_2$ ...  
Con el primer conjunto de datos pasamos del *prior* al *posterior* y luego esa distribución se convierte en el nuevo *prior*:

$$\text{Beta}(a, b) \rightarrow \text{Beta}(y_1 + a, N_1 - y_1 + b)$$

$$\text{Beta}(y_1 + a, N_1 - y_1 + b) \rightarrow \text{Beta}(y_2 + y_1 + a, N_2 - y_2 + N_1 - y_1 + b)$$

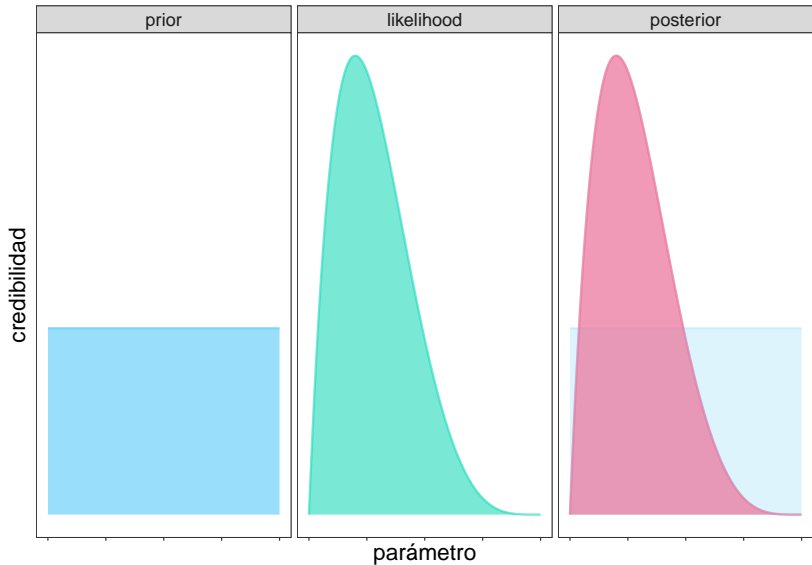
$$\text{Beta}(a, b) \rightarrow \text{Beta}((y_1 + y_2) + a, (N_1 + N_2) - (y_1 + y_2) + b)$$

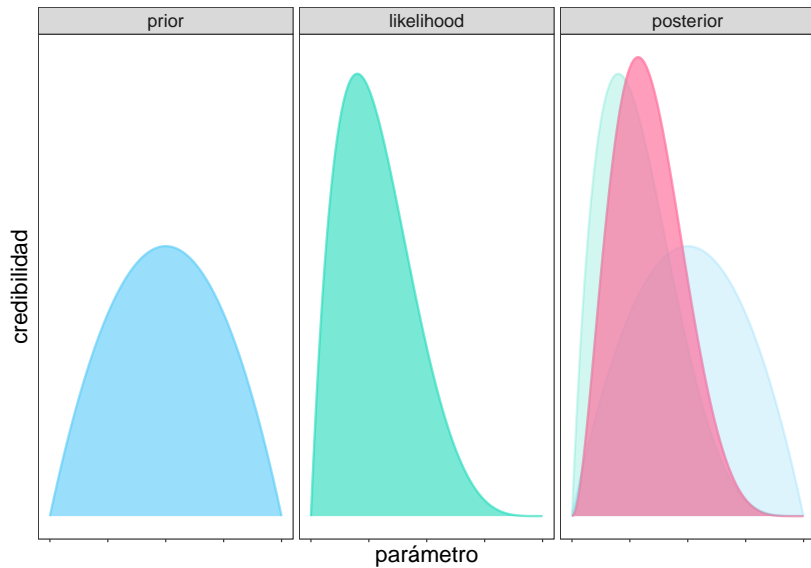
Es idéntico a observar  $y_1 + y_2$  en  $N_1 + N_2$

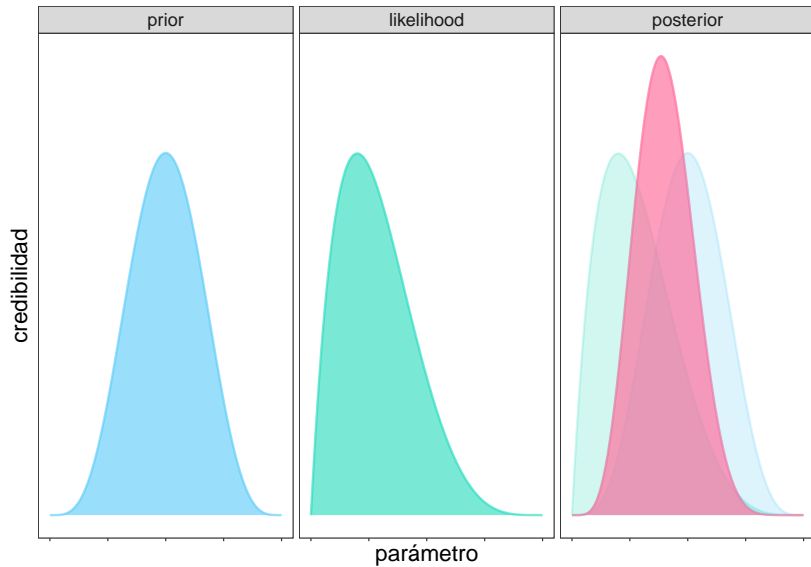
La Regla de Bayes permite combinar dos fuentes de información: la información *a priori* (lo que sabemos hasta el momento), y la nueva información (representada por la verosimilitud). La distribución *a posteriori* representa un compromiso entre la verosimilitud de los datos y la credibilidad *a priori*.

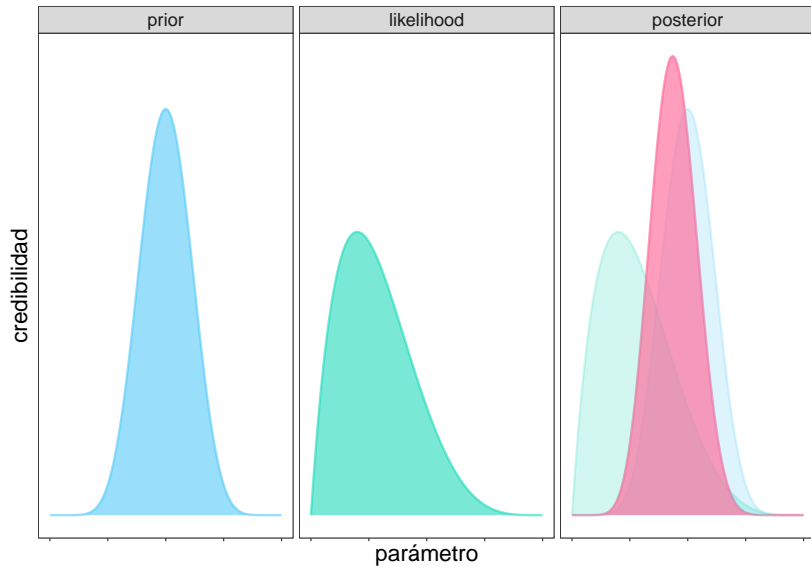
Más ejemplos

## Más ejemplos

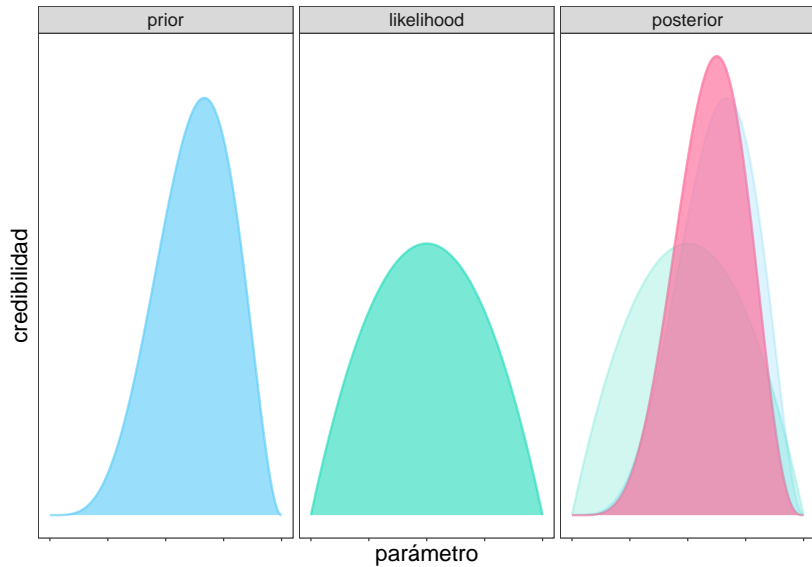


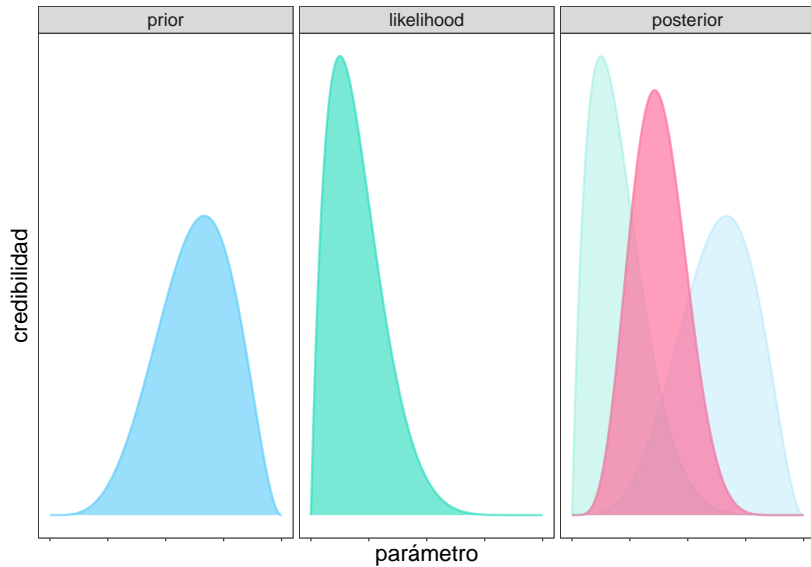


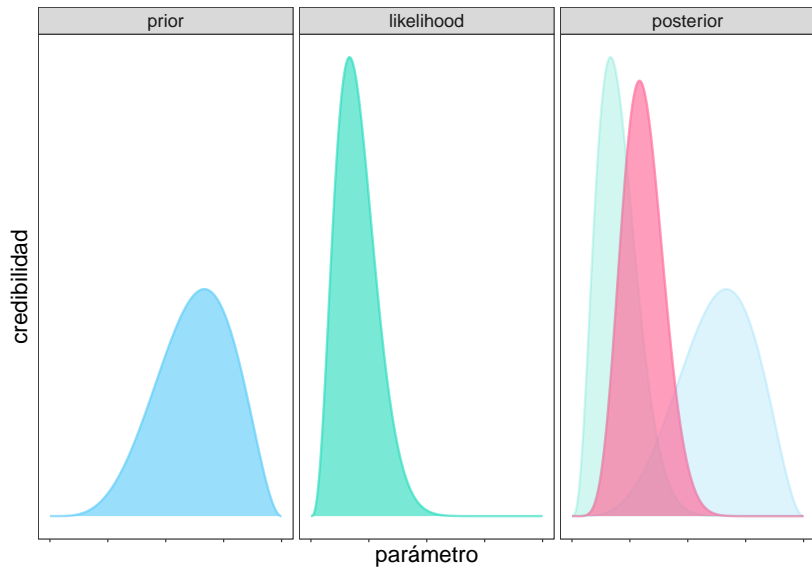


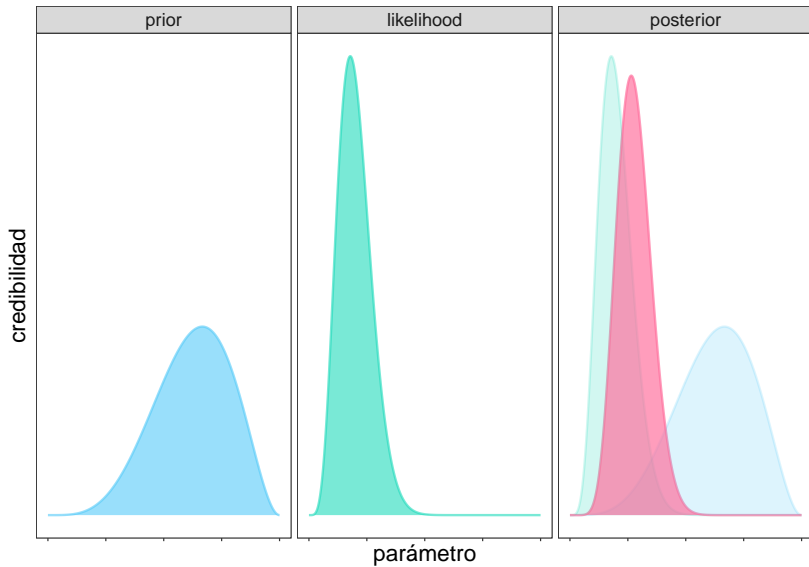


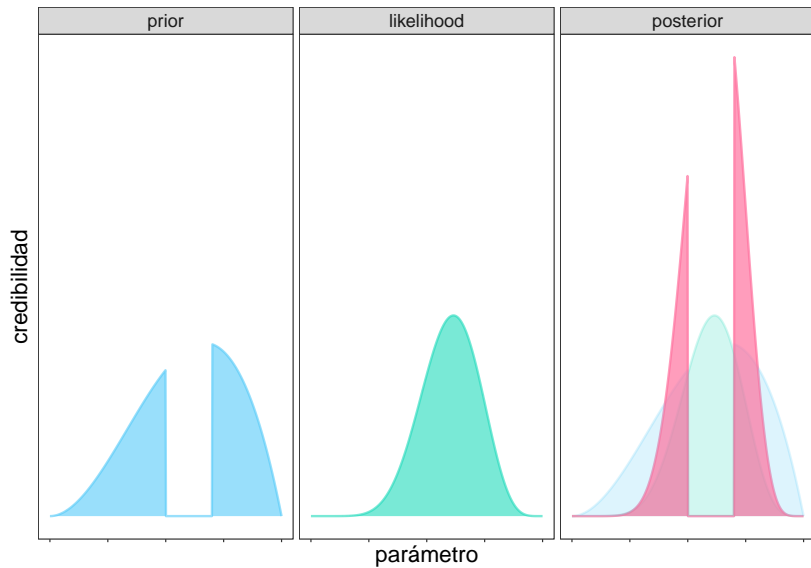


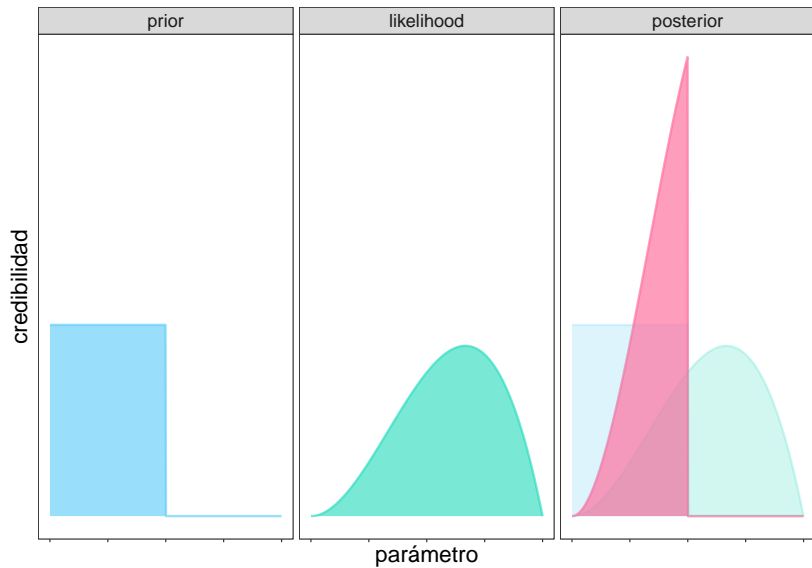












# Predicciones

Distribución predictiva *a posteriori* (también distribución posterior predictiva) (en inglés *posterior predictive distribution*): queremos predecir un valor futuro de la variable de interés,  $\tilde{y}$ . Más aún, interesa la distribución de  $\tilde{y}$  *a posteriori*, es decir, luego de observar los datos  $y$ :  $\tilde{y} \mid y$

$$p(\tilde{y} \mid y) = \int p(\tilde{y} \mid \pi) p(\pi \mid y) d\pi$$

- ▶  $\tilde{y}$  tiene una distribución de probabilidad
- ▶ Si  $\pi$  fuera fijo, la distribución de  $\tilde{y}$  viene dada por  $p(\tilde{y} \mid \pi)$  (la verosimilitud, aunque ahora es función de  $\tilde{y}$ )
- ▶ Pero ahora hay incertidumbre en  $\pi$  (tiene una distribución *a posteriori*), por lo tanto se hace una ponderación para los distintos valores de  $\pi$  ( $\pi$  varía en la integral anterior)
- ▶ Combinamos lo que no sabemos porque es aleatorio *per se*, con aquello que desconocemos (aunque podemos reducir nuestra incertidumbre recolectando más información)



Para el caso binomial que venimos estudiando, consideramos una realización más (tirar el globo terráqueo y agarrarlo). ¿Cuál es la probabilidad de obtener  $A$  (agua)?

$$\begin{aligned} p(\tilde{y} = 1 \mid y) &= \int_0^1 \pi p(\pi \mid x) d\pi = \mathbb{E}[p(\pi \mid x)] \\ &= \frac{y + a}{y + a + N - y + b} = \frac{y + a}{N + a + b} \end{aligned}$$

```
muestras_pi <- rbeta(2000,a+y,b+N-y) # muestras del posterior  
x_new <- rbinom(2000,1,muestras_pi) # predicciones para cada valor de pi
```

Consideremos un caso particular:

*En una bolsa hay bolitas negras y blancas, queremos saber cuál es la probabilidad de sacar una bolita negra. A priori no sabemos nada. Sacamos (con reposición) tres veces una bolita. Las tres veces sale negra. ¿Cuál es la probabilidad de que la próxima bolita sea negra?*

$$p(\tilde{y} = 1 \mid y) = \frac{y + 1}{N + 2}$$

Los parámetros tienen una distribución de probabilidad. Incorporar la incertidumbre en el valor de  $\pi$  nos permite no entusiasmarnos tanto con los datos, hacer predicciones más conservadoras con pocos datos, regularizar.