Tarea 3 - Inferencia II

Santiago Robatto, Sofia Terra

2025-09-01

Exercise 2.14 (Late bus). Li Qiang takes the 8:30am bus to work every morning. If the bus is late, Li Qiang will be late to work. To learn about the probability that her bus will be late (π), Li Qiang first surveys 20 other commuters: 3 think π is 0.15, 3 think π is 0.25, 8 think π is 0.5, 3 think π is 0.75, and 3 think π is 0.85.

1. Convert the information from the 20 surveyed commuters into a prior model for π.
2. Li Qiang wants to update that prior model with the data she collected: in 13 days, the 8:30am bus was late 3 times. Find the posterior model for π.
3. Compare and comment on the prior and posterior models. What did Li Qiang learn about the bus?

Respuestas:

1. Nuestra Realidad:

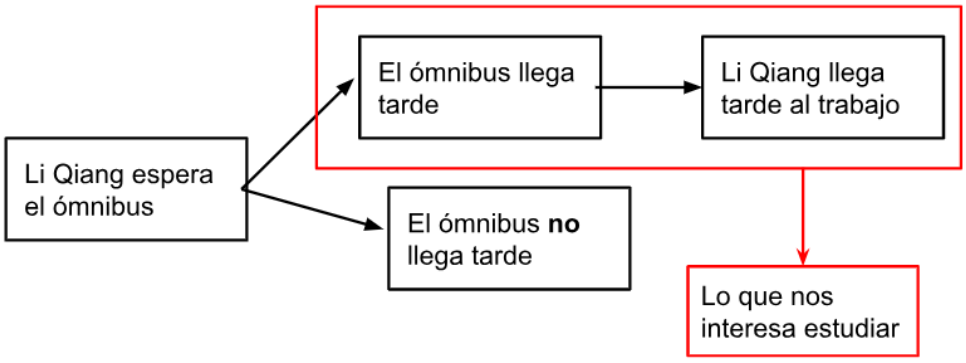


Figure 1: Diagrama que muestra la situacion actual

Datos que tenemos:

- π = probabilidad de que el ómnibus llegue tarde
- Si el ómnibus llega tarde, entonces Li Qiang llega tarde a trabajar.
- π es una variable aleatoria.

Nuestro análisis comienza con un **modelo a priori**, que nos dice qué valores puede tomar π, donde se le asigna una ponderación o peso a cada uno, **sumando todos en total 1**. Cada valor de π tiene una priori relativa que indica su plausibilidad. En nuestro caso, esto está determinado por la encuesta que realizó Li Qiang a 20 personas que se toman regularmente el ómnibus. Vemos que 8 personas creen que con un 50% de probabilidad el ómnibus va a llegar tarde, mientras que 3 creen que no llegará a tiempo con un 15% de probabilidad, las mismas que para el π=0.25, π=0.75 y π=0.85. Por lo tanto, nuestro modelo a priori queda determinado de la siguiente manera:

Table 1: Modelo Priori

π	0.15	0.25	0.5	0.75	0.85
$f(\pi)$	0.15	0.15	0.4	0.15	0.15

2. Ahora, Li Qiang quiere **actualizar** el modelo a priori con los datos que registró. Su observación fue que en 13 días, su ómnibus llegó tarde sólo 3 veces. Se busca obtener un modelo a posteriori. Para llegar a esto primero debemos determinar cómo se distribuyen nuestros datos. Utilizamos el **modelo Binomial**, ya que con este contamos el número de éxitos en n pruebas (siendo estas independientes entre sí). En nuestro caso, tendríamos en total 13 pruebas. Ella repite su "experimento" 13 veces, cada día se pregunta "¿llegó tarde el ómnibus?", si la respuesta es afirmativa, lo registra, siendo este un "éxito". Esto toma vida con una nueva variable aleatoria, Y. La dependencia condicional de Y sobre π se determina con dicha distribución Binomial a través de los parámetros n y π.

$$Y \mid \pi \sim \text{Bin}(n, \pi)$$
$$P(y \mid \pi) = \binom{n}{y} \pi^y (1-\pi)^{n-y}$$

La función de densidad nos resume la **probabilidad condicional** de observar un cierto número de éxitos Y=y para cualquier π dada. Dado que sabemos que el ómnibus llegó tarde 3 veces, calcularemos la **verosimilitud** de nuestra función para cada valor de π. Obtenemos:

Table 2:

π	0.15	0.25	0.50	0.75	0.85
$L(\pi \mid y = 3)$	0.19	0.2516	0.035	≈ 0	≈ 0

Con esto ya podemos construir nuestro **modelo a posteriori**, el cual tiene la siguiente formula:

$$\text{posterior} = \frac{\text{prior} \cdot \text{likelihood}}{\text{normalizing constant}} \propto \text{prior} \cdot \text{likelihood}$$

Nuestra constante de normalización será la siguiente:

$$(0.1900 \cdot 0.15) + (0.2516 \cdot 0.15) + (0.0349 \cdot 0.4) + (0.0001 \cdot 0.15) + (0 \cdot 0.15) = 0.0802$$

Table 3: Modelo Posteriori

π	0.15	0.25	0.50	0.75	0.85
$f(\pi)$	0.15	0.15	0.40	0.15	0.15
$f(\pi \mid y = 3)$	0.3553	0.4706	0.1741	≈ 0	≈ 0

3. Gracias a la última tabla observamos que es mucho *más acertado decir que el ómnibus no llegará tarde*, en comparación con decir lo contrario. Nuestro modelo posteriori nos muestra que es casi imposible que con 75% o 85% de probabilidad el ómnibus llegue tarde, esto es un cambio extremadamente brusco en respecto a nuestro modelo a priori. Después de haber registrado que en 13 días llega tarde 3, las probabilidades de que llegue tarde bajan drásticamente, se actualizó el modelo. Lo que aprende Li Qiang es que puede confiar en que no llegará tarde al trabajo muy seguido.

Exercise 2.15(Cuckoo birds). Cuckoo birds are brood parasites, meaning that they lay their eggs in the nests of other birds (hosts), so that the host birds will raise the cuckoo bird hatchlings. Lisa is an ornithologist studying the success rate, π, of cuckoo bird hatchlings that survive at least one week. She is taking over the project from a previous researcher who speculated in their notes the following prior model for π.

π	0.6	0.65	0.7	0.75	Total
$f(\pi)$	0.3	0.4	0.2	0.1	1

Figure 2: Modelo priori ejercicio 2.15

1. If the previous researcher had been more sure that a hatchling would survive, how would the prior model be different?
2. If the previous researcher had been less sure that a hatchling would survive, how would the prior model be different?
3. Lisa collects some data. Among the 15 hatchlings she studied, 10 survived for at least one week. What is the posterior model for π?
4. Lisa needs to explain the posterior model for π in a research paper for ornithologists, and can't assume they understand Bayesian statistics. Briefly summarize the posterior model in context.

Exercise 2.19(Cuckoo birds redux). Repeat Exercise 2.15 utilizing simulation to approximate the posterior model of π

Respuestas:

1. El modelo a priori intenta asignar probabilidades a las tasas de supervivencia de los cuckoo bird hatchlings. De esta manera, podemos interpretar π como la probabilidad de que la tasa tome cierto valor. Por ejemplo el modelo asigna probabilidad 0.3 a una tasa de 60% de supervivencia (que podemos interpretar como que sobreviven 6 de 10 cuckoo birds hatchlings).

De esta manera, la seguridad de supervivencia se ve reflejada en la probabilidad de los valores de la tasa de supervivencia. Entonces si el investigador hubiese estado más seguro de que los hatchlings sobreviven, los valores de la tasa de supervivencia más cercanos a 1 serían más probables. Podemos hacer esto de dos maneras: Alterando el recorrido o sin alterarlo.

Si no alteramos el recorrido, el investigador hubiera estado más seguro de que los hatchlings sobrevirían si 0.7 y/o 0.75 tuviesen mayor probabilidad.

Por otro lado, podríamos alterar el recorrido. Por ejemplo, si sumamos 0.2 a cada tasa de supervivencia, tendríamos un modelo con tasas de supervivencias sumamente mayores:

Table 4:

π	0.8	0.85	0.9	0.95
$f(\pi)$	0.3	0.40	0.2	0.10

El razonamiento es análogo para la parte 2). Si el investigador hubiese pensado que la probabilidad de que un hatchling sobreviviese era menor, la distribución de las probabilidades de las tasas se "trasladaría" hacia la izquierda (el π tomaría valores más cercanos a cero). Para este caso, podríamos plantear un modelo a priori como el siguiente:

Aquí, las probabilidades para las tasas de supervivencia menores son más altas.

Table 5:

π	0.1	0.2	0.3	0.4
$f(\pi)$	0.4	0.4	0.1	0.1

3. Modelo posterior para π dado que recolectó datos los cuales muestran que sobrevivieron 10 de los 15 estudiados.

Tal como vimos en clase y en el libro en el ejemplo de ajedrez, podemos aplicar la Regla de Bayes para deducir el modelo posterior.

En nuestro caso, definimos la variable aleatoria Y, la cual sigue una distribución binomial.

Entonces, deduciremos el modelo posteriori sabiendo que y=10.

$$f(\pi \mid y = 10) = \frac{f(\pi) \cdot L(\pi \mid y = 10)}{f(y = 10)}, \quad \text{para } \pi \in \{0,6; 0,65; 0,7; 0,75\}$$

La distribución a priori ya la tenemos dada que viene dada por la letra.

Luego calcularemos la función de verosimilitud.

$$L(\pi \mid y) = f(y \mid \pi)$$
$$f(y \mid \pi) \sim \text{Bin}(15, \pi)$$

Podemos calcular las verosimilitudes para cada caso con la función de distribución binomial, usando n=15 dado que es el tamaño del modelo y π variando en función de cada verosimilitud. Realizamos las cuentas para y=10 dado que fue lo obtenido en el experimento.

Nuestra tabla con los valores de la verosimilitud queda de la siguiente manera:

Table 6:

π	0.6000	0.6500	0.7000	0.7500
$L(\pi \mid y = 10)$	0.1859	0.2123	0.2061	0.1651

Nuestra constante de normalización queda de la siguiente forma

$$(0.1859 \cdot 0.3) + (0.2123 \cdot 0.4) + (0.2061 \cdot 0.2) + (0.1651 \cdot 0.1) = 0.19842$$

Entonces tendremos:

$$f(\pi = 0.6 \mid y = 10) = \frac{0.1859 \cdot 0.3}{0.19842} = 0.281$$

$$f(\pi = 0.65 \mid y = 10) = \frac{0.2123 \cdot 0.4}{0.19842} = 0.428$$

$$f(\pi = 0.7 \mid y = 10) = \frac{0.2061 \cdot 0.2}{0.19842} = 0.208$$

$$f(\pi = 0.75 \mid y = 10) = \frac{0.1651 \cdot 0.1}{0.19842} = 0.083$$

La creencia inicial (a priori) ya sugería que el valor más plausible para la tasa de supervivencia era π=0.65 (con una probabilidad de 0.4). Los datos que se observaron (0.67) están muy alineados con esa creencia inicial. Por lo tanto, la evidencia actualizó nuestro modelo reforzando esa idea, y la probabilidad de que π=0.65 aumentó de 0.4 a 0.428 en el modelo a posteriori.

Ahora vamos a hacer la simulacion para una muestra de 10000 polluelos:

Primero creamos un vector que tenga los valores de π, el cual guardamos como data frame. Luego creamos otro vector que tenga tenga las prioris correspondientes a cada valor de π.

```
pi_vals <- data.frame(c(0.6, 0.65, 0.7, 0.75))
prior_probs <- c(0.3, 0.4, 0.2, 0.1)
```

Ahora definimos la simulacion para dichos valores de π, donde los valores de las prioris "ponderan" a sus correspondientes valores de π. Nuestra muestra es de tamaño 10000.

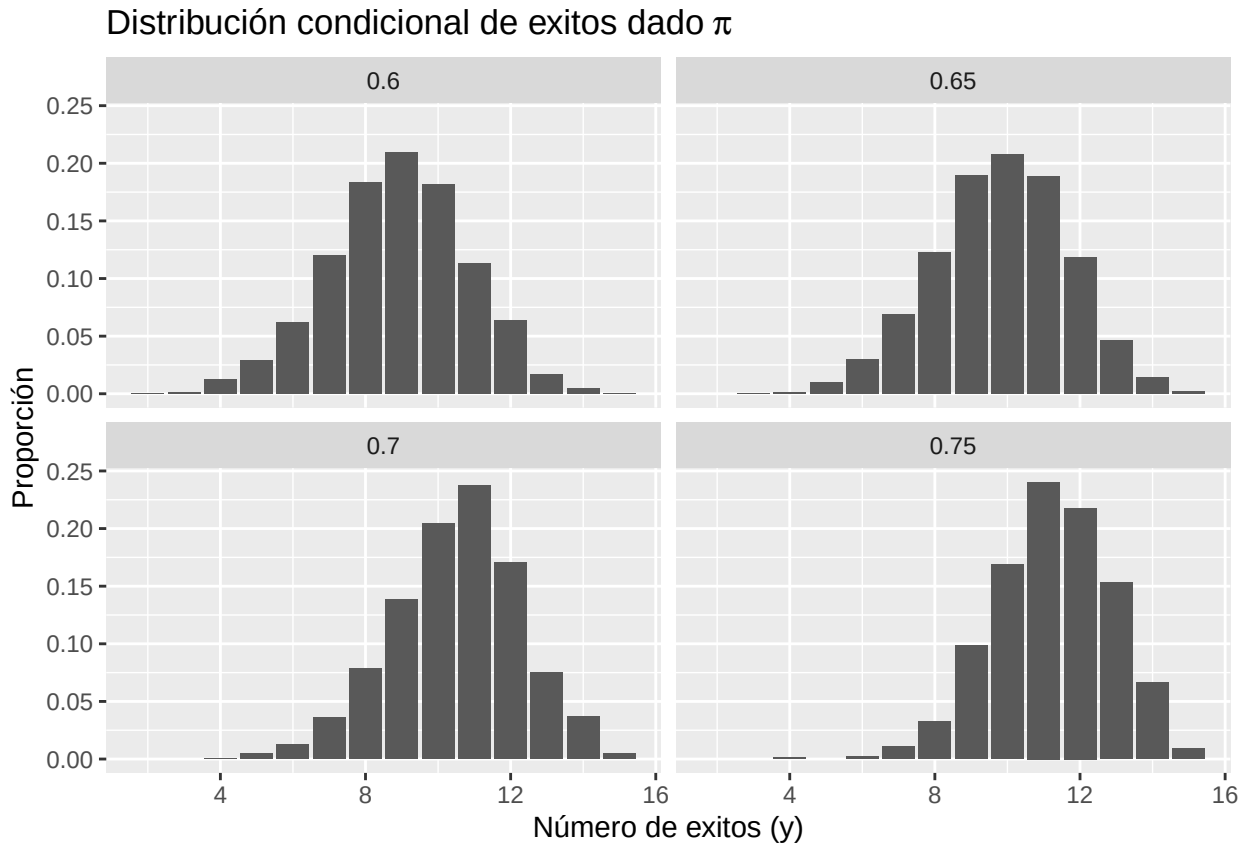
```
bird_sim <- dplyr::sample_n(pi_vals,
                            size = 10000,
                            weight = prior_probs,
                            replace = TRUE)
bird_sim <- as.data.frame(bird_sim)
```

Luego, usamos la distribucion binomial para modelar la cantidad de exitos.

```
colnames(bird_sim)[1] <- "pi"
bird_sim <- bird_sim %>%
  mutate(y = rbinom(10000, size = 15, prob = pi))
```

En este paso graficamos lo anterior.

```
ggplot(bird_sim, aes(x = y)) +
  stat_count(aes(y = after_stat(prop))) +
  facet_wrap(~ pi) +
  labs(title = bquote("Distribución condicional de exitos dado" ~pi),
       x = "Número de exitos (y)",
       y = "Proporción")
```



Ahora, filtramos las simulaciones anteriores para obtener las que tienen y = 10 exitos.

```
win_one <- bird_sim %>%
  dplyr::filter(y == 10)
```

Aquí obtenemos el resumen de la distribución a posteriori aproximada de π.

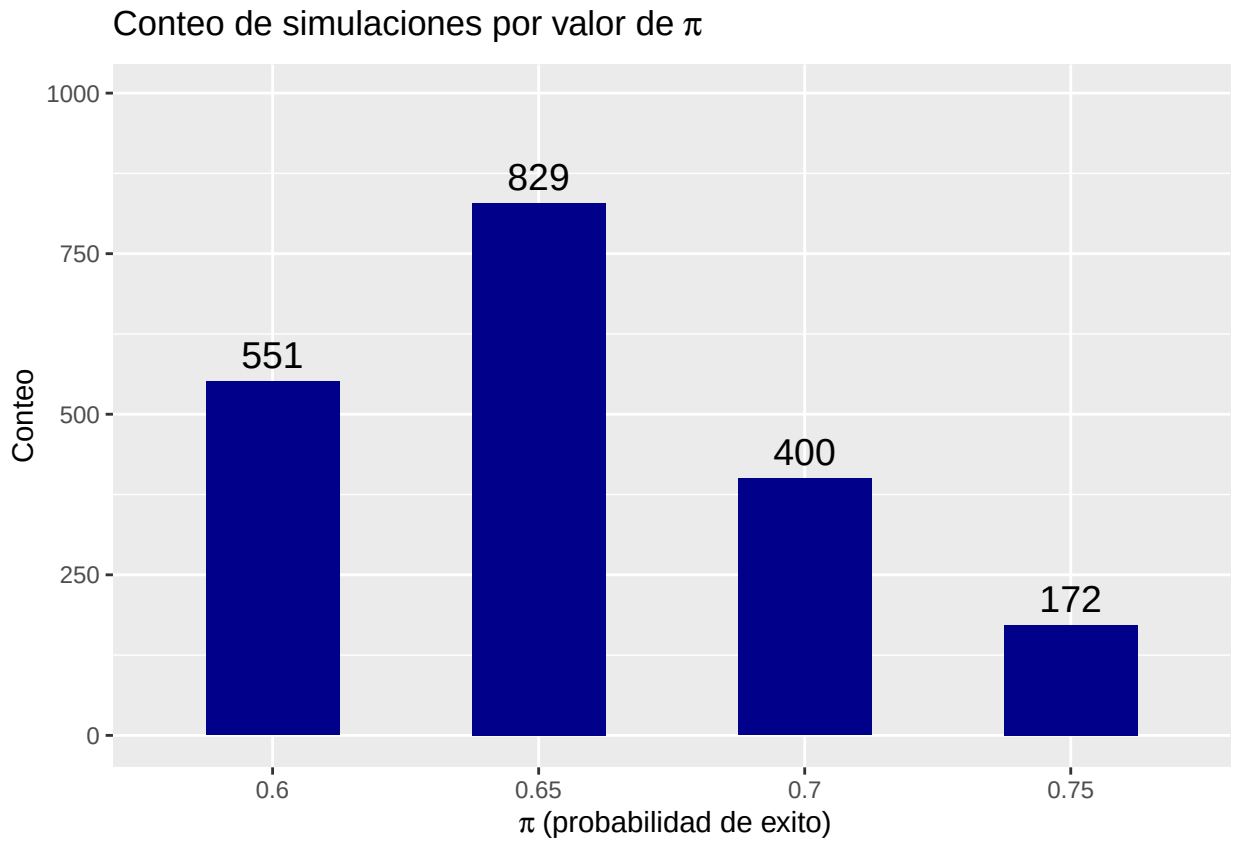
Table 7:

pi	n	percent
0.6	551	0.2822746
0.65	829	0.4246926
0.7	400	0.2049180
0.75	172	0.0881148
Total	1952	1.0000000

```
# Calcular la frecuencia (n) y el porcentaje asociado a cada valor de pi
posterior_df <- win_one %>%
  dplyr::count(pi) %>%
  dplyr::mutate(percent = n / sum(n))
```

```
# Gráfico de barras para los conteos simulados por cada valor de pi
ggplot(posterior_df, aes(x = as.factor(pi), y = n)) +
  geom_bar(stat = "identity", fill = "darkblue", width = 0.5) +
```

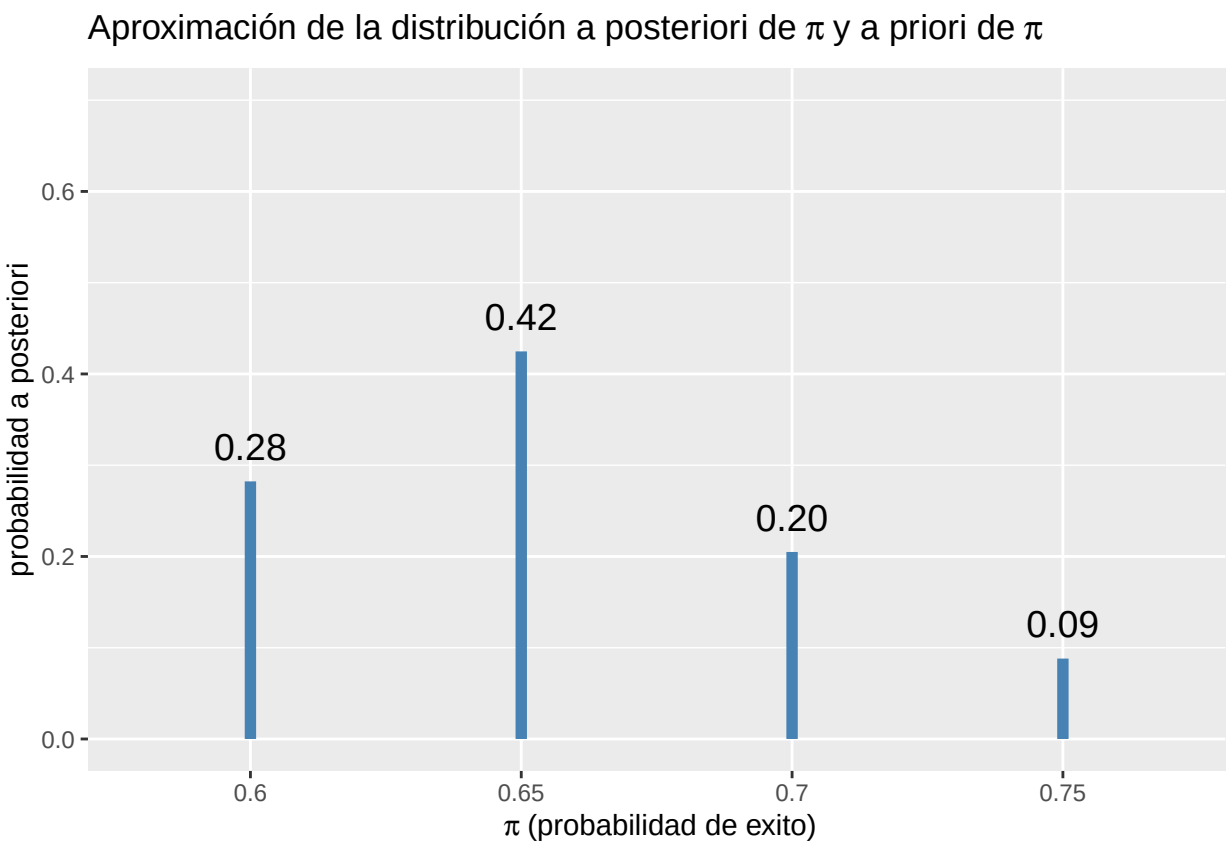
```
  geom_text(aes(label = n), vjust = -0.5, size = 5) +
  scale_x_discrete(labels = c("0.6", "0.65", "0.7", "0.75")) +
  labs(
    title = bquote("Conteo de simulaciones por valor de" ~pi),
    x = bquote( ~ pi ~ "(probabilidad de exito)"),
    y = "Conteo"
  ) +
  ylim(0, max(posterior_df$n) * 1.2)
```




```
# Gráfico de líneas
# representa la aproximación a la distribución a posteriori de $\pi$ dados los datos observados
grafico1<-ggplot(posterior_df, aes(x = as.factor(pi), y = percent)) +
  # Segmentos desde (x,0) hasta (x, percent)
  geom_segment(aes(x = as.factor(pi),
                   xend = as.factor(pi),
                   y = 0,
                   yend = percent),
               linewidth = 2, color = "steelblue") +
  geom_text(aes(label = scales::number(percent, accuracy = 0.01)),
            vjust = -0.8, size = 5) +
  labs(
    title = bquote("Aproximación de la distribución a posteriori de" ~ pi ~ "y a priori de" ~ pi),
    x = bquote( ~pi ~ "(probabilidad de exito)"),
    y = "probabilidad a posteriori"
  ) +
```

```
  ylim(0, 0.7)

grafico1
```



4. Lisa, la cual es una investigadora de ciencias, construye un modelo sobre la supervivencia de polluelos de aves cuckoo, la cual pone sus huevos en nidos de otras especies. Su objetivo es ver cuántos de esos huevos sobreviven al menos una semana. El modelo que construye justamente es una simplificación de la realidad, donde se basa en suposiciones, datos y predicciones para llegar a conclusiones. Para empezar, ella toma la información ya recabada por un colega y la establece como su "base", es lo que inicialmente sabe. A esto se le llama el modelo "a priori" en la Estadística Bayesiana. Sin embargo, el misma puede estar desactualizado o no ser el más adecuado al contexto actual, por lo tanto, Lisa decide realizar un pequeño estudio, donde observa si los polluelos sobreviven o no. Si uno sobrevive, lo registra como un "éxito". Ella repite este procedimiento 15 veces, de las cuales 10 registra "éxitos". Luego, utiliza esta información para actualizar su modelo a uno nuevo, el cual llamamos "posteriori". Con esto, podrá hacer predicciones sobre si un polluelo futuro sobrevivirá o no. En resumen, el modelo "a posteriori" lo que hace es combinar la creencia inicial de Lisa con los datos que ella recolectó ("la realidad") para producir nueva información acerca de la supervivencia de los polluelos.

En este caso, dado que la evidencia es congruente con lo planteado por el modelo a priori, se reafirman las creencias iniciales.

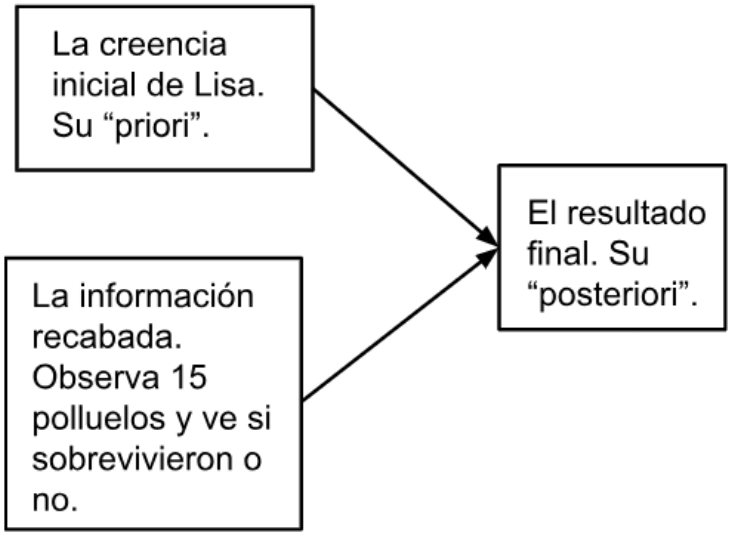


Figure 3: Resumen de como funciona el modelo bayesiano.