Example: a problem of male twins (Efron, 2003)

Pregnant with twins:
fraternal or identical?

Fraternal: 2/3 of all cases

Identical: 1/3 of all cases

What is the probability of identical twins IF both boys in sonogram?

Answer provided by Bayes theorem

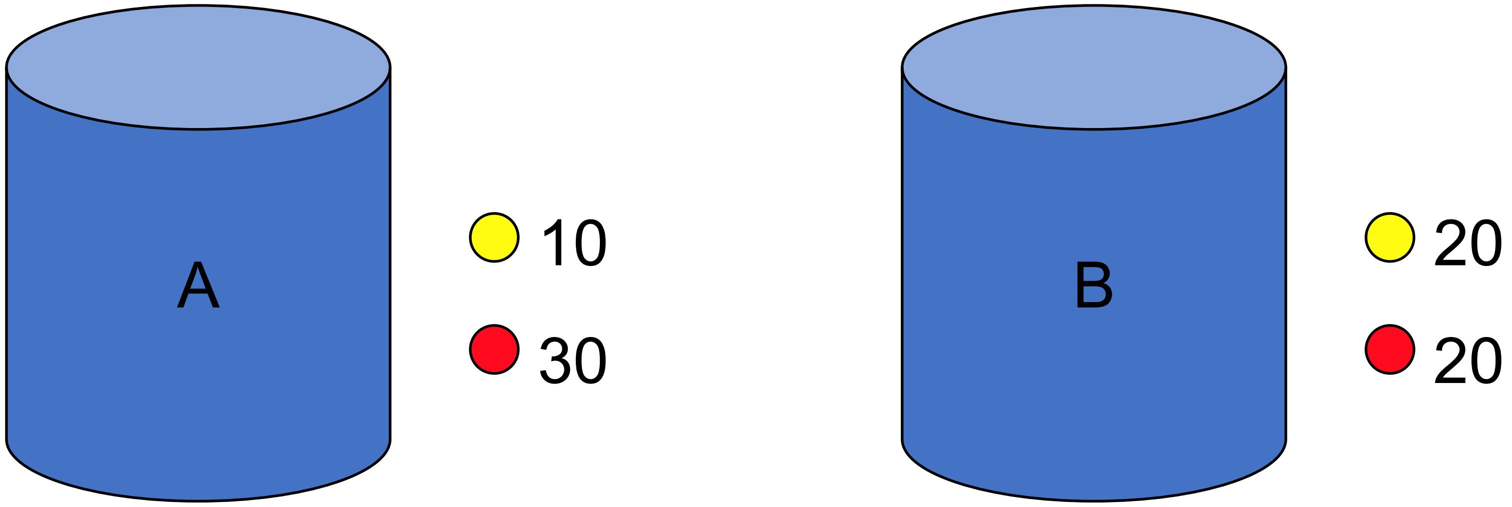
$$P(\text{Identical}|\text{Both boys}) = \frac{P(\text{Both boys}|\text{Identical})}{P(\text{Both boys})} P(\text{Identical})$$

$$
\begin{aligned}
P(\text{Identical}) &= 1/3 \\
P(\text{Fraternal}) &= 2/3 \\
P(\text{Both boys}|\text{Identical}) &= 1/2 \\
P(\text{Both boys}|\text{Fraternal}) &= 1/4
\end{aligned}
$$

$$
\begin{aligned}
P(\text{Both boys}) &= P(\text{Both boys}|\text{Identical})P(\text{Identical}) \\
&\quad + P(\text{Both boys}|\text{Fraternal})P(\text{Fraternal}) \\
&= (1/2)(1/3) + (1/4)(2/3) = 1/3
\end{aligned}
$$

$$
\begin{aligned}
P(\text{Identical}|\text{Both boys}) &= \frac{P(\text{Both boys}|\text{Identical})}{P(\text{Both boys})}P(\text{Identical}) \\
&= \frac{(1/2)}{(1/3)}(1/3) = 1/2
\end{aligned}
$$

A simple example with urns



Here we choose a ball as follows:

1. We choose the urn first
2. We draw a ball from that urn

What is the probability of drawing one red ball?

$P(A) = P(B) = 1/2$ (probability of choosing either A or B)

$P(G|A) = 1/4$ (probability of drawing a yellow ball from A)

$P(R|A) = 3/4$ (probability of drawing a red ball from A)

$P(G|B) = 1/2$ (probability of drawing a yellow ball from B)

$P(R|B) = 1/2$ (probability of drawing a red ball from A)

and therefore

$$
\begin{aligned}
P(R) &= P(R|A) \cdot P(A) + P(R|B) \cdot P(B) \\
&= (3/4) \cdot (1/2) + (1/2) \cdot (1/2) = 5/8 = 0.625
\end{aligned}
$$

Now we play the reverse game with the help of a friend. The friend hides the names of the boxes and places them in front of us. We extract one ball from the box on the left and we use all our prior knowledge and Bayes' theorem to infer its name. Clearly, at the outset – i.e., *a priori*, given the information I that we initially have – the probability that the box on the left is actually box A is just $p_0(L = A|I) = 1/2$, because the two boxes A and B are equally probable. Now we start the game by extracting the first ball.

Extraction 1: the ball is red. Using the prior information we find that the probabilities of extraction of a red or a yellow ball from any of the two boxes (the *evidences*) are

$$p_0(R) = P(R|L = A, I)p_0(L = A, I) + P(R|L = B, I)p_0(L = B, I) = 5/8 = 0.625$$

$$p_0(Y) = P(Y|L = A, I)p_0(L = A, I) + P(Y|L = B, I)p_0(L = B, I) = 3/8 = 0.375$$

Therefore, using Bayes' theorem, the posterior probability for A is

$$p_1(L = A|R, I) = \frac{P(R|L = A, I)}{p_0(R)} p_0(L = A|I) = \frac{3/4}{5/8}(1/2) = 3/5 = 0.6$$

where $P(R|L = A, I) = P(R|A, I)$ is the *likelihood*.

After the first extraction the ball is reinserted in the box from which it was taken and we are ready for the next extraction.

Extraction 2: we extract a red ball, again. Then we use the old posterior probabilities $p_1(L = A, I)$ and $p_1(L = B, I)$ as the new prior values and we find

$$p_1(R) = P(R|L = A, I)p_1(L = A, I) + P(R|B, I)p_1(L = B, I) = 0.65$$

$$p_1(Y) = P(Y|L = A, I)p_1(L = A, I) + P(Y|B, I)p_1(L = B, I) = 0.35$$

and a repeated application of Bayes' theorem yields:

$$p_2(L = A|\{R, R\}, I) = \frac{P(R|L = A, I)}{p_1(R)} p_1(L = A|R, I) = 0.692308$$

How do we continue this procedure? We only have to generalize the previous steps to the generic n-th extraction.

Extraction n: in the generic case we consider the n-th extraction and we do not specify the colors of the extractions, we just indicate the whole series of extraction data with $\{D_k\}_{k=1,n}$. Using the posterior probabilities of step $n-1$

$$p_{n-1}(L = A|\{D_k\}_{k=1,n-1}, I) \quad \text{and} \quad p_{n-1}(L = B|\{D_k\}_{k=1,n-1}, I)$$

as the new prior probabilities we find the evidence at the n-th extraction

$$\begin{aligned} p_n(D_n|\{D_k\}_{k=1,n-1}, I) &= \\ &= P(D_n|L = A, I)p_{n-1}(L = A|\{D_k\}_{k=1,n-1}, I) \\ &\quad + P(D_n|L = B, I)p_{n-1}(L = B|\{D_k\}_{k=1,n-1}, I) \end{aligned}$$

and from Bayes' theorem we find the new posterior probabilities

$$p_n(L = A|\{D_k\}_{k=1,n}, I) = \frac{P(D_n|L = A, I)}{p_n(D_n|\{D_k\}_{k=1,n-1}, I)} p_{n-1}(L = A|\{D_k\}_{k=1,n-1}, I)$$

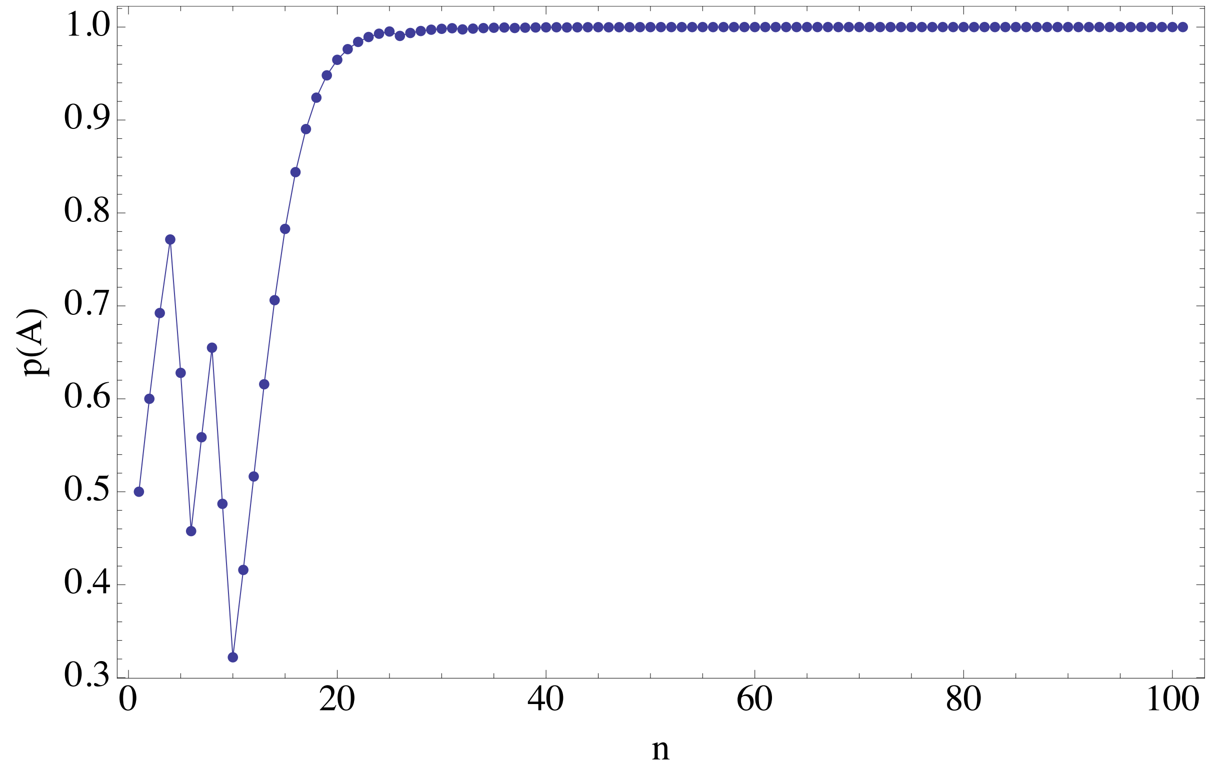
$$p_n(L = B|\{D_k\}_{k=1,n}, I) = \frac{P(D_n|L = B, I)}{p_n(D_n|\{D_k\}_{k=1,n-1}, I)} p_{n-1}(L = B|\{D_k\}_{k=1,n-1}, I)$$

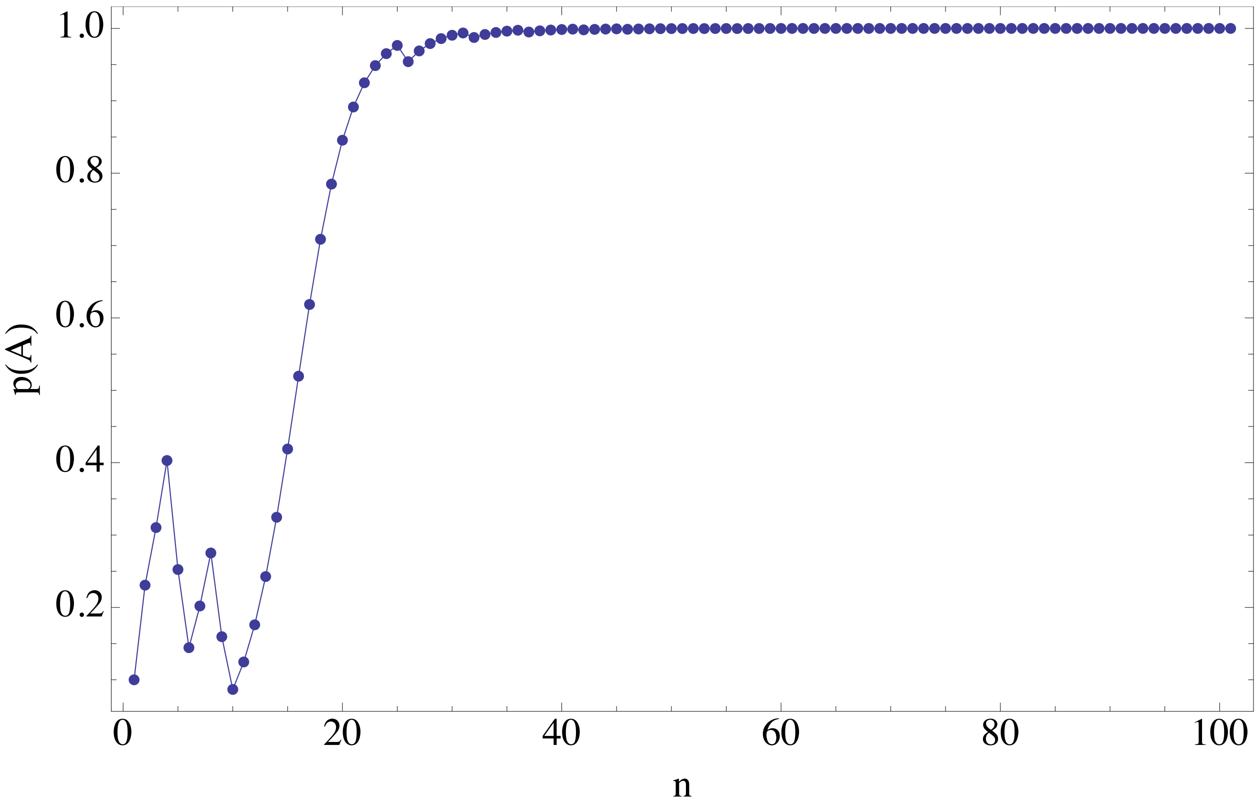
100 extractions ...

R, R, R, Y, Y, R, R, Y, Y, R, R, R, R, R, R, R, R, R, R, R, R, R, R, R, Y,
R, R, R, R, R, Y, R, R, R, R, Y, R, R, R, R, Y, R, R, R, Y, R, R, R, R, R,
R, Y, R, R, Y, R, R, R, R, R, R, Y, R, R, R, R, Y, R, R, Y, R, Y, R, R, Y,
Y, R, R, Y, R, R, R, Y, R, R, Y, R, R, R, R, R, R, R, R, R, R, Y, Y, R, R

posterior probabilities

$$p_n(L = A|\{D_k\}_{k=1,n}, I)$$





Posterior probability $p_n(L = A|\{D_k\}_{k=1,n}, I)$ vs. n in the simulated experiment described in the text with a much lower initial prior ($p_0(L = A|I) = 0.1$).

A new game

Now there is just one box with N balls and an unknown number N_R of red balls.

Problem: use repeated extractions to estimate the number of red balls.

Now the likelihood function is given by the set of probabilities

$$P(Y|N_R, I) = (1 - N_R/N); \quad P(R|N_R, I) = N_R/N$$

and we use Bayes' theorem for many competing and mutually exclusive hypotheses

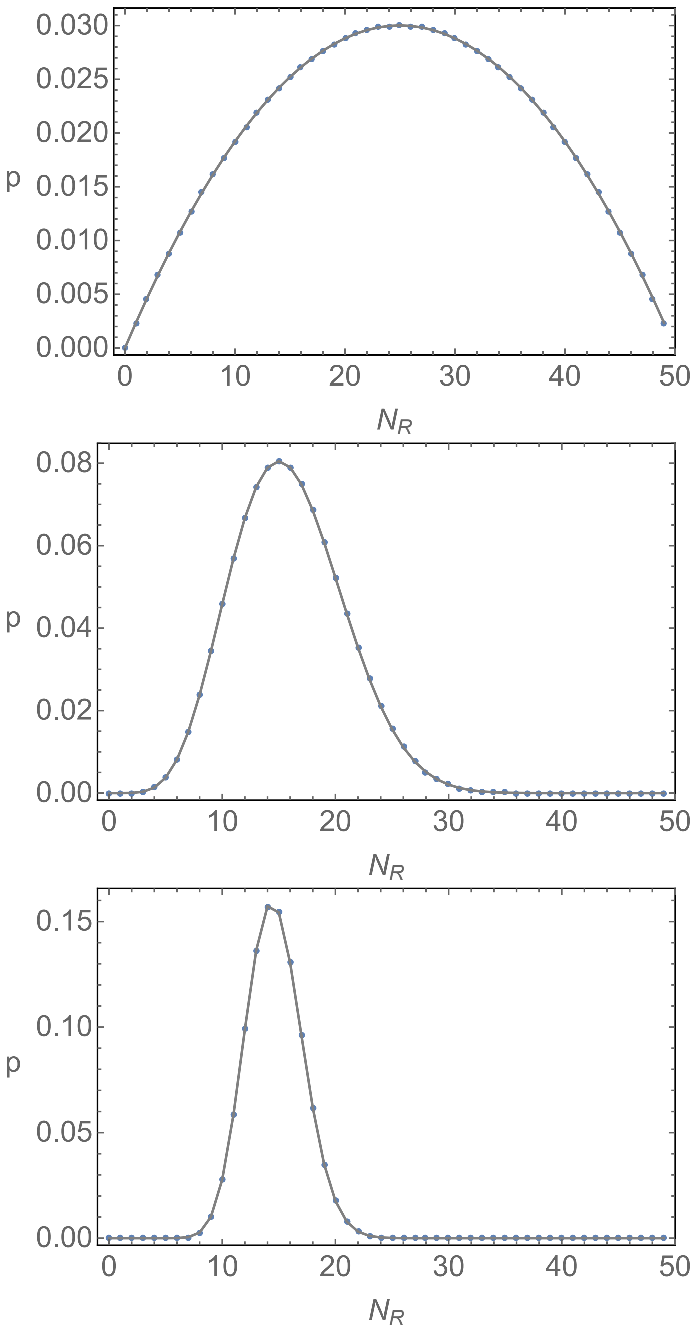
$$P(A_{N_R}|D, I) = \frac{P(D|A_{N_R}, I)}{\sum_n P(D|A_n, I)P(A_n)} P(A_{N_R}|I)$$

with a uniform initial prior

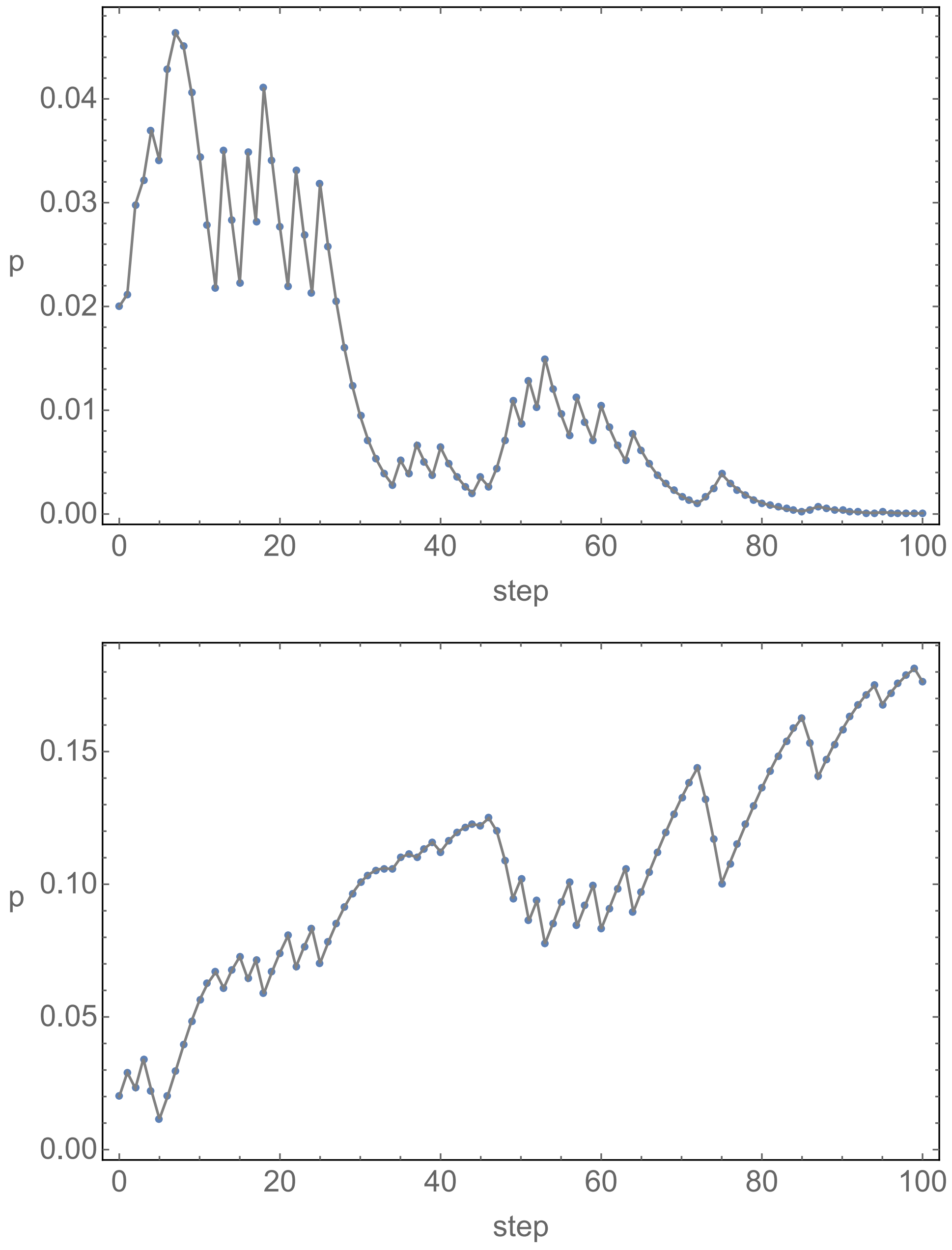
$$P(A_{N_R}|I) = 1/N$$

Experiment: 100 extractions (with N = 50), what is the estimate of N_R?

Y, R, Y, R, R, Y, Y, Y, Y, Y, Y, Y, R, Y, Y, R, Y, R, Y, Y, Y, R, Y, Y, R,
Y, Y, Y, Y, Y, Y, Y, Y, Y, R, Y, R, Y, Y, R, Y, Y, Y, Y, R, Y, R, R, R, Y,
R, Y, R, Y, Y, Y, R, Y, Y, R, Y, Y, Y, R, Y, Y, Y, Y, Y, Y, Y, Y, R, R, R,
Y, Y, Y, Y, Y, Y, Y, Y, Y, Y, R, R, Y, Y, Y, Y, Y, Y, Y, R, Y, Y, Y, Y, R.



Discrete posterior distribution for the problem of one box with an unknown number of red balls. Upper panel: the posterior distribution after 2 draws; middle panel: the posterior distribution after 20 draws; bottom panel: the posterior distribution after 80 draws. As the number of draws increases, our experiment provides more and more information, and the width of the posterior distribution decreases. This decrease of the width of the posterior distribution means that our confidence in the result is correspondingly larger.



Posterior probability associated with $N_R = 24$ (upper panel) and $N_R = 14$ (lower panel). As more and more data are collected, the posterior probability of $N_R = 24$ drops to zero, while that of $N_R = 14$ has a steady upward trend.

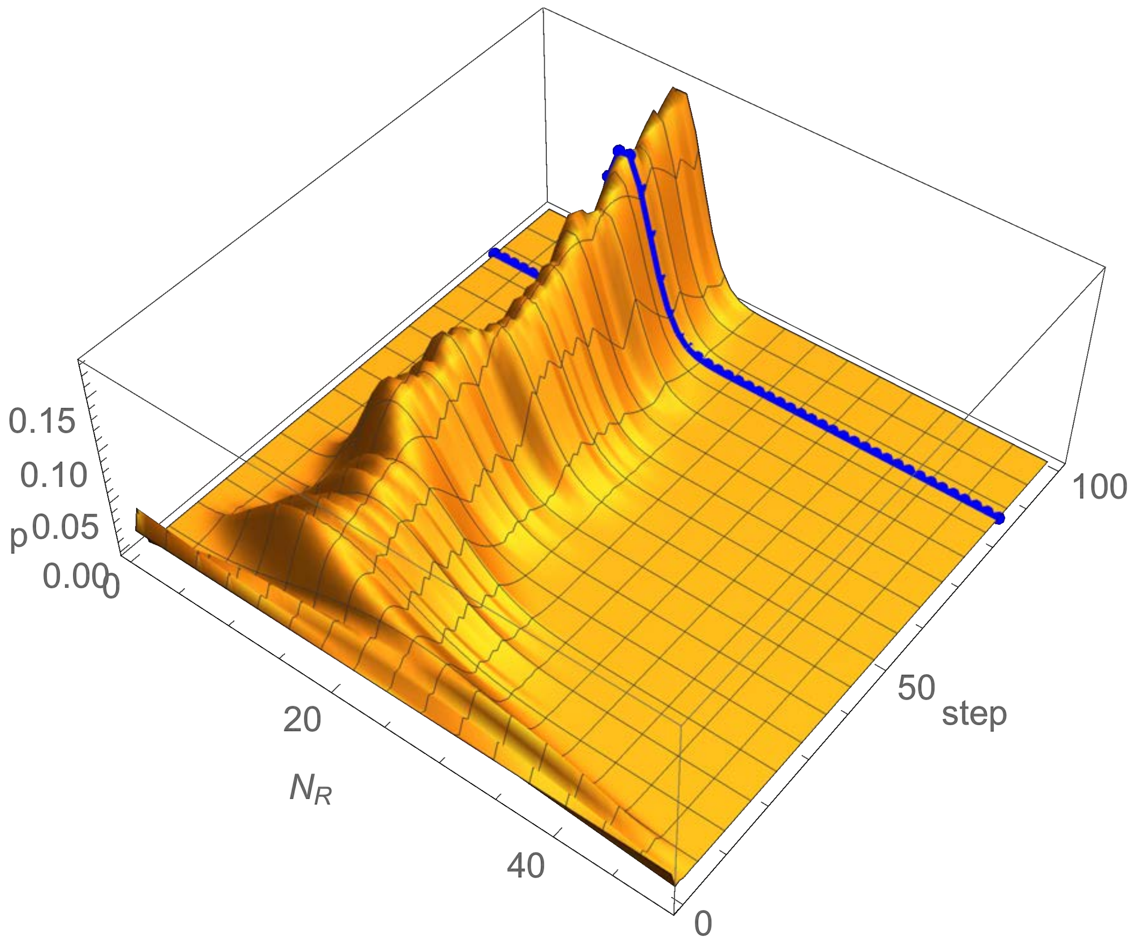


Figure 3.6: Combined representation that shows the evolution of the probability distribution as more and more data are collected. The blue line corresponds to the posterior probability distribution shown in the bottom panel of figure 3.4.

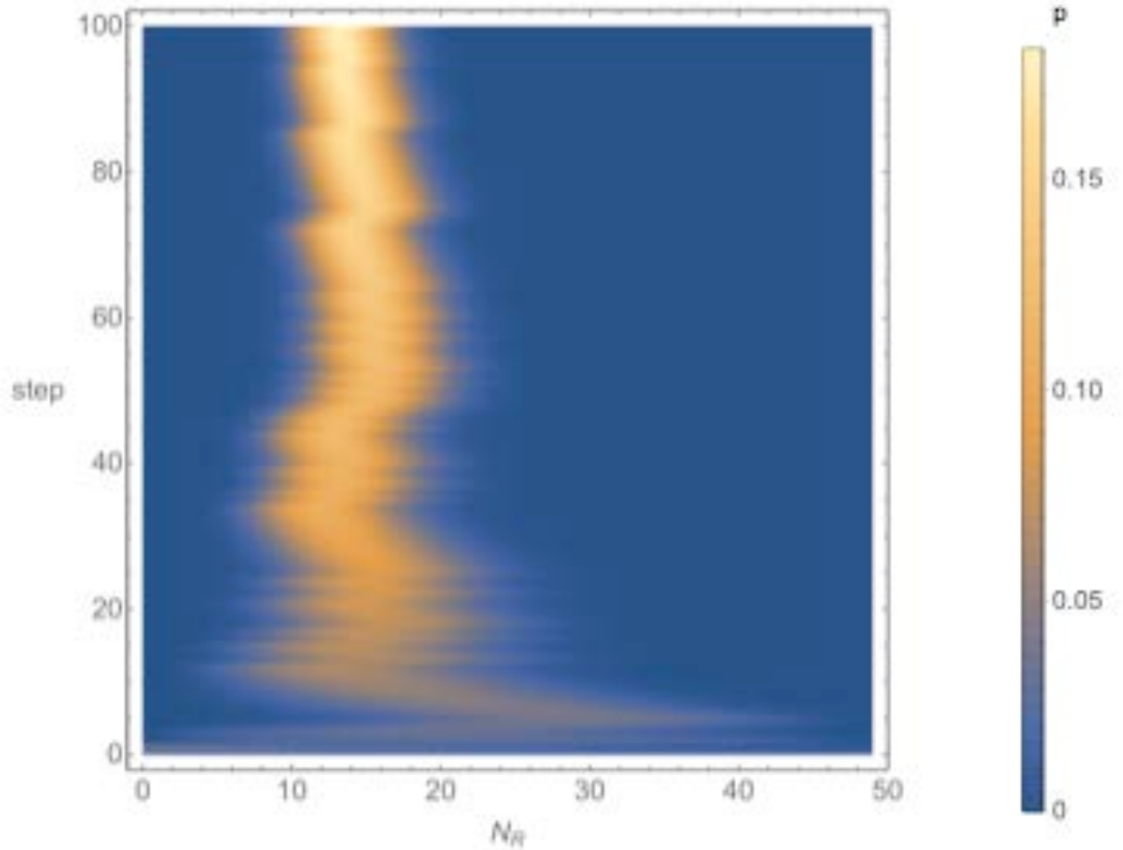


Figure 3.7: Another representation – a *density plot* – equivalent to the 3D plot shown in figure 3.6. While probability corresponds to a height in figure 3.6, here it corresponds to a colour shown in the colour scale on the right.