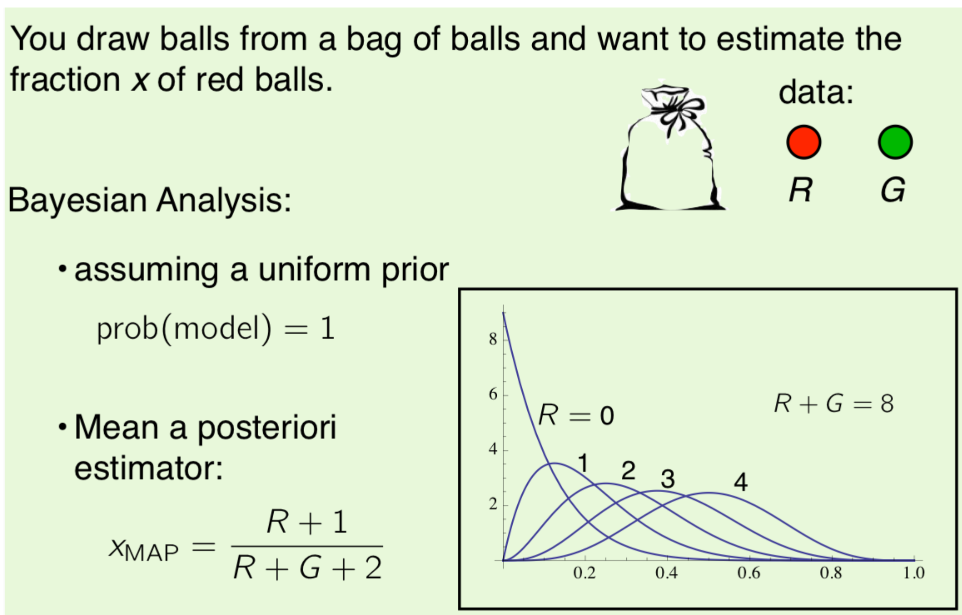
Solution to the question in the Bioinformatics class

Arthur Wang

October 17, 2019



1 Definition

B: the data collected. Number of red balls ($\#Red = R$) and green balls ($\#Green = G$).
A: the actual fraction of red balls in the bag ($x \in [0, 1]$).

QUESTION: What's the Bayesian estimation of x based on data **B**?

2 Lemma

Bayesian estimation of variable x

$$\tilde{x} = \int xP(x|B)dx \tag{1}$$

where $\tilde{x}$ is the Bayesian estimation and $P(x|B)$ is the probability of the hypothesis that the actual fraction of red balls in the bag equals to x under the condition **B**.

Bayesian formula

$$P(x|B) = \frac{P(x)P(B|x)}{\sum_x P(x)P(B|x)} \tag{2}$$

3 Solution

The calculation of $P(x|B)$

As mentioned in the slides, we have assumed the distribution of variable x as uniformed, which means $P(x) = constant = 1$. And then we can simplify equation 2 as:

$$P(x|B) = \frac{P(B|x)}{\sum_x P(B|x)} \tag{3}$$

x is actually a continuous rather than discrete ranging from 0 to 1. So we rewrite the sum in equation (3) as integral:

$$P(x|B) = \frac{P(B|x)}{\int_0^1 P(B|x)dx} \tag{4}$$

Drawing $R + G$ balls from the bag can be regarded as a sequence of $R + G$ independent experiments, in all of which the probability of a red ball is a constant x. The results follow binomial distribution $B(R + G, x)$, so we can derive:

$$P(B|x) = C_{R+G}^{R} x^R (1 - x)^G \tag{5}$$

where C_{R+G}^R is the combinatorial number. Then we can calculate the integral by parts and derive:

$$
\begin{aligned}
\int_0^1 P(B|x)dx &= \int_0^1 C_{R+G}^R x^R (1-x)^G dx && (6)\\
&= \frac{C_{R+G}^R}{R+1}\int_0^1 (1-x)^G dx^{R+1} && (7)\\
&= \frac{C_{R+G}^R}{R+1}[(x^{R+1}(1-x)^G)|_0^1 - \int_0^1 x^{R+1} d(1-x)^G] && (8)\\
&= \frac{C_{R+G}^R}{R+1}(x^{R+1}(1-x)^G)|_0^1 + \frac{C_{R+G}^R G}{R+1}\int_0^1 x^{R+1}(1-x)^{G-1}dx && (9)\\
&= 0 + \frac{C_{R+G}^R G}{R+1}\int_0^1 x^{R+1}(1-x)^{G-1}dx && (10)\\
&= \frac{C_{R+G}^R G(G-1)}{(R+1)(R+2)}\int_0^1 x^{R+2}(1-x)^{G-2}dx && (11)\\
&= \frac{C_{R+G}^R G(G-1)(G-2)}{(R+1)(R+2)(R+3)}\int_0^1 x^{R+3}(1-x)^{G-3}dx && (12)\\
&= ... && (13)\\
&= \frac{C_{R+G}^R G(G-1)...1}{(R+1)(R+2)...(R+G)}\int_0^1 x^{R+G}(1-x)^0 dx && (14)\\
&= C_{R+G}^R G!\frac{R!}{(R+G)!}\int_0^1 x^{R+G}(1-x)^0 dx && (15)\\
&= C_{R+G}^R \frac{G!R!}{(R+G)!}\frac{1}{R+G+1} && (16)\\
&= C_{R+G}^R \frac{G!R!}{(R+G+1)!} && (17)
\end{aligned}
$$

Combine equation (4), (5) and (17):

$$
\begin{aligned}
P(x|B) &= \frac{P(B|x)}{\int_0^1 P(B|x)dx} && (18)\\
&= \frac{C_{R+G}^R x^R (1-x)^G}{C_{R+G}^R \frac{G!R!}{(R+G+1)!}} && (19)\\
&= x^R(1-x)^G \frac{(R+G+1)!}{G!R!} && (20)
\end{aligned}
$$

The Bayesian estimation $\tilde{x}$

The Bayesian estimation $\tilde{x}$ can be calculated by equation (1), where $P(x|B)$ have been shown as equation (20):

$$\begin{aligned}\tilde{x} &= \int xP(x|B)dx && (21)\\ &= \int_0^1 x \cdot x^R(1-x)^G \frac{(R+G+1)!}{G!R!}dx && (22)\\ &= \frac{(R+G+1)!}{G!R!}\int_0^1 x^{R+1}(1-x)^G dx && (23)\end{aligned}$$

With integration by parts similar to equation (6) to (17) or just using the conclusion from equation (17), we can finally calculate:

$$\begin{aligned}\tilde{x} &= \int xP(x|B)dx && (24)\\ &= \frac{(R+G+1)!}{G!R!}\int_0^1 x^{R+1}(1-x)^G dx && (25)\\ &= \frac{(R+G+1)!}{G!R!}\frac{G!(R+1)!}{(R+G+2)!} && (26)\\ &= \frac{R+1}{R+G+2} && (27)\end{aligned}$$

4 Discussion

$\tilde{x}$ equals to $\frac{1}{2}$ when $R = G$, which is the same with the frequency $\frac{R}{R+G} = \frac{1}{2}$.
When $R = 0$, the frequency equals to 0 while $\tilde{x} = \frac{1}{G+2} \neq 0$.
In other words, Bayesian estimation think there's still a few red balls even if all balls we drew from the bag are green. But if we estimate x by frequency, there won't be any red balls if we found no red ones from the bag. There's still heated debate on these two theories.