Keypoints for Homework 3

Zhenglei Gao

November 5, 2003

1 Gelman 2.7

- For Y$\sim$ Bin(n,θ), then the distribution follow as:

$$\begin{aligned} P(Y=y|\theta) &\propto \theta^y(1-\theta)^{n-y} \\ &= (1-\theta)^n e^{ylog(\frac{\theta}{1-\theta})} \end{aligned}$$

so the natural parameter is: $\phi = log(\frac{\theta}{1-\theta})$ let $p(\phi) \propto 1$, then ,

$$\begin{aligned} p(\theta) &= p(\phi)|\frac{d\phi}{d\theta}| \\ &\propto \theta^{-1}(1-\theta)^{-1} \end{aligned}$$

- if y=n, $p(\theta|y) \propto \theta^{n-1}(1-\theta)^{-1}$,the integration of it goes to inf, which is improper, similar when y=0.

2 Lung Tumor Data

n patients, n_0 non-recurrent tumors and n_1 recurrent tumors; $g_i \sim Bernoulli, \Pr(g_i = 1) = \pi_0$ for non-recurrent, $\Pr((g_i = 1) = \pi_1$ for recurrent; $H_1 : \pi_0 = \pi_1 = \pi$, prior for $\pi \sim Unif[0,1]$; $H_2 : \pi_0 \neq \pi_1$, prior: independent uniform. Interest: association between the protein and the tumor type.

(a) The marginal density:

$$\begin{aligned} p(G|H_1) &= \int_0^1 p(G|\pi)p(\pi)d\pi \\ &= \int_0^1 \prod_{i=1}^n \pi^{g_i}(1-\pi)^{1-g_i}d\pi \\ &= \int_0^1 \pi^{\sum_{i=1}^n g_i}(1-\pi)^{n-\sum_{i=1}^n g_i}d\pi \\ &= \int_0^1 \pi^x(1-\pi)^{n-x}d\pi \\ &= \frac{\Gamma(x+1)\Gamma(n-x+1)}{\Gamma(n+2)} \end{aligned}$$

(b) The related marginal density:

$$\begin{aligned}
p(G|H_2) &= \int_0^1 p(G|\pi_0,\pi_1)p(\pi_0,\pi_1)d\pi \\
&= \int_0^1 \prod_{i=1}^{n_0} \pi_0^{g_i}(1-\pi_0)^{1-g_i} d\pi_0 \int_0^1 \prod_{i=1}^{n_1} \pi_1^{g_i}(1-\pi_1)^{1-g_i} d\pi_1 \\
&= \int_0^1 \pi_0^{x_0}(1-\pi_0)^{n_0-x_0} d\pi_0 \int_0^1 \pi_1^{x_1}(1-\pi_1)^{n_1-x_1} d\pi_1 \\
&= \frac{\Gamma(x_0+1)\Gamma(n_0-x_0+1)}{\Gamma(n_0+2)} \frac{\Gamma(x_1+1)\Gamma(n_1-x_1+1)}{\Gamma(n_1+2)}
\end{aligned}$$

(c) assign $\Pr(H_1) = 0.5$, the posterior probability:

$$\begin{aligned}
\Pr(H_1|G) &= \frac{p(G|H_1)p(H_1)}{p(G|H_1)p(H_1) + p(G|H_2)p(H_2)} \\
&= \frac{0.5p(G|H_1)}{0.5p(G|H_1) + 0.5p(G|H_2)} \\
&= \frac{p(G|H_1)}{p(G|H_1) + p(G|H_2)}
\end{aligned}$$

(d) $n_0 = 34, x_0 = 5; n_1 = 40, x_1 = 17$, then the Bayes factor:

$$\begin{aligned}
BF &= p(G|H_1)/p(G|H_2) \\
&= \frac{\Gamma(5+17+1)\Gamma(34+40-5-17+1)/\Gamma(34+40+2)}{\Gamma(5+1)\Gamma(34-5+1)/\Gamma(34+2)\times\Gamma(17+1)\Gamma(40-17+1)/\Gamma(40+2)} \\
&= 0.1294738
\end{aligned}$$

In this case, the posterior probability:

$$\begin{aligned}
\Pr(H_1|G) &= \frac{BF \times p(G|H_2)}{BF \times p(G|H_2) + p(G|H_2)} \\
&= \frac{0.13}{0.13+1} \\
&\approx 0.115
\end{aligned}$$

Surely, this is evidence for an association between the protein and the tumor type.

(e) We use Fisher's exact test to determine whether there is a differencr in π_0 and π_1, which assume a hypergeometric distribution for protein expression. The null hypothesis assumes equality. This test gives a p-value of 0.01125, so we reject $\pi_0 = \pi_1$

(f) Assuming $H_2, \pi_0 \sim U[0,1]$, the joint poterior distribution for π_0 and π_1:

$$p(\pi_0,\pi_1|G) \propto p(G|\pi_0,\pi_1)p(\pi_0,\pi_1)$$

so, the kernel for posterior distribution of π_0 is:

$$\begin{aligned}
p(\pi_0|G) &\propto \pi_0^{x_0}(1-\pi_0)^{n_0-x_0} \\
&\sim Beta(x_0+1, n_0-x_0+1) \\
&= Beta(6,30)
\end{aligned}$$

so, the poterior mean of π_0 is:$(x_0+1)/(n_0+2) = 6/36 = 1/6$
Similarily, we can get posterior distribution of π_1:

$$\begin{aligned} p(\pi_1|G) &\propto Beta(x_1+1, n_1-x_1+1) \\ &= Beta(18, 24) \end{aligned}$$

and the posterior mean $E(\pi_1|G) = (x_1+1)/(n_1+2) = 18/42 = 3/7$

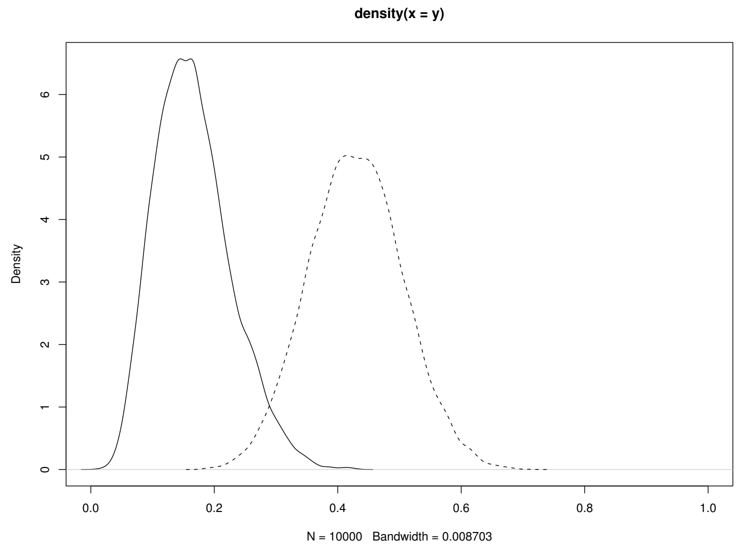

Figure 1: posterior distributions for pi_0 and pi_1(the dashed line)

(g) For a randomly selected patient, denote that A_1=the incidence of a recurrent aggressive tumor, A_2=the incidence of non-recurrent tumor, B_1= the present of the protein, B_2=the absent of the protein. Then, if $\Pr(A_1) = 15\%$,

$$\begin{aligned} \theta &= \Pr(A_1|B_1) \\ &= \frac{\Pr(B_1|A_1)\Pr(A_1)}{\Pr(B_1|A_1)\Pr(A_1) + \Pr(B_1|A_2)\Pr(A_2)} \\ &= \frac{\pi_1 \times 15\%}{\pi_1 \times 15\% + \pi_0 \times 85\%} \\ &= \frac{3\pi_1}{3\pi_1 + 17\pi_0} \end{aligned}$$

(h) With the data values above,the likelihood function for π_0, π_1, is:

$$L(\pi_0, \pi_1) = p(G|\pi_0, \pi_1) = \pi_0^{x_0}(1-\pi_0)^{n_0-x_0} d\pi_0 \pi_1^{x_1}(1-\pi_1)^{n_1-x_1}$$

So, $\pi_0 \sim Beta(x_0+1, n_0-x_0+1) = Beta(6, 30), \pi_1 \sim Beta(x_1+1, n_1-x_1+1) = (18, 24)$the MLE of π_0, π_1,respectively, is:

$$\begin{aligned} \hat{\pi}_0 &= x_0/n_0 = 5/34 \\ \hat{\pi}_1 &= x_1/n_1 = 17/40 \end{aligned}$$

the MLE of θ:

$$\begin{aligned}\hat{\theta} &= \frac{3\hat{\pi}_1}{3\hat{\pi}_1 + 17\hat{\pi}_0} \\ &= \frac{3 \times 17/40}{3 \times 17/40 + 17 \times 5/34} \\ &= 51/151\end{aligned}$$

(i) to compute the approximate poterior mean, just sample from the poterior distributions of $\pi_0 and \pi_1$. Use R, we can get poterior mean, median, and a 90% posterior credible interval for θ,respectively:

$$\begin{aligned}E(\theta|x) &= 0.3306 \\ median(\theta|x) &= 0.3177 \\ CI_{90\%}(\theta|x) &= (0.1983658, 0.5050169)\end{aligned}$$

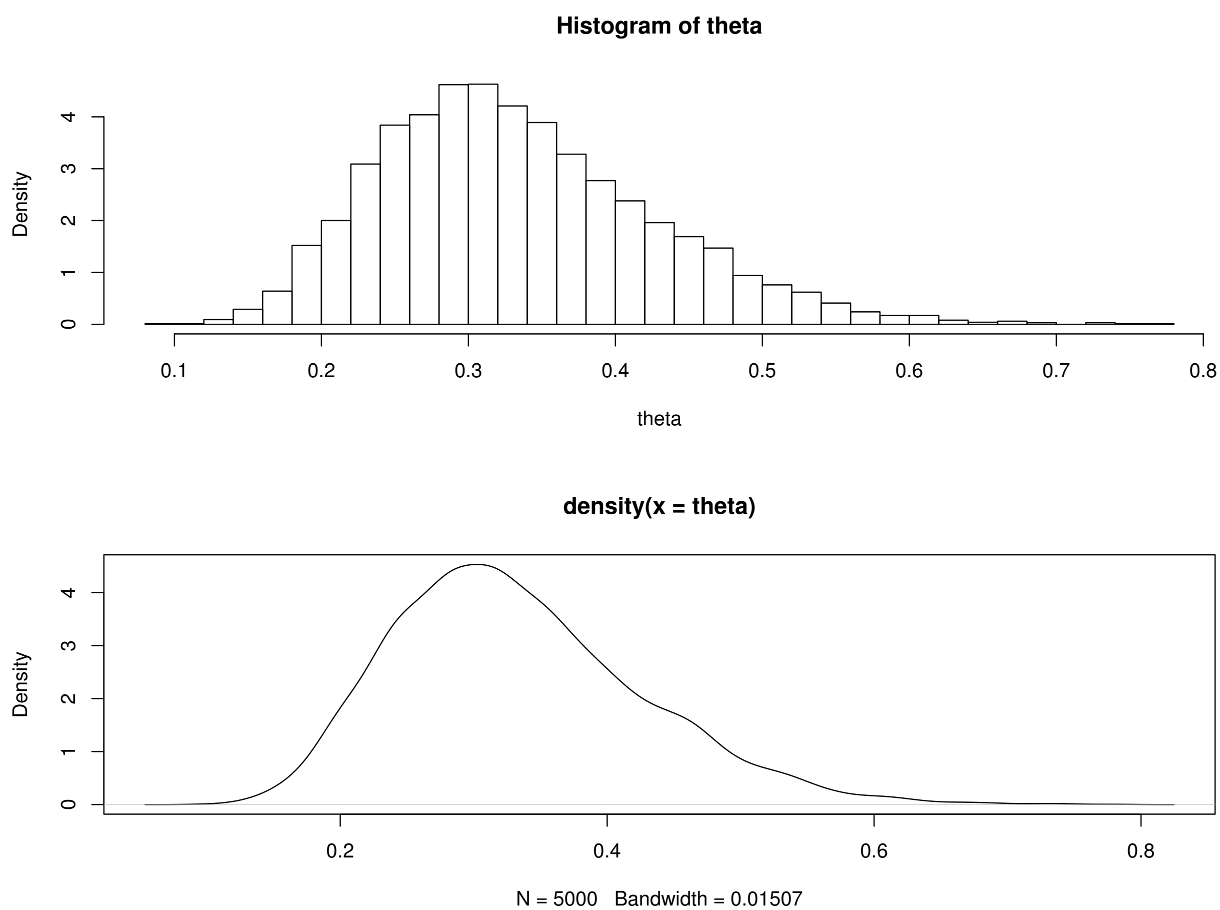

Figure 2: Approximate Poterior Density for θ