VE414 Lecture 5

Jing Liu

UM-SJTU Joint Institute

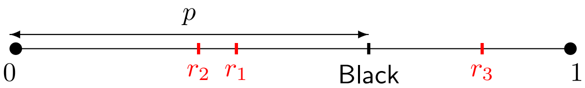
May 28, 2019

- Recall the structure of our Bayesian model so far is very simple, i.e.

$$
\begin{aligned}
Y &\sim f_Y \\
X \mid Y &\sim f_{X|Y} \\
Y \mid X = x &\sim f_{Y|X=x}
\end{aligned}
$$

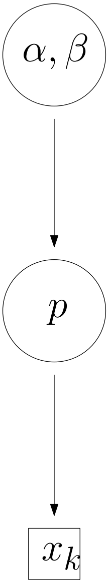
where $f_{Y|X=x} \propto f_{X=x|Y} \cdot f_Y$ according to Bayes theorem.

- For example, the uncertainty in Bayes' original problem



that is, our lack of knowledge on where the black ball, is modelled by,

$$
\begin{aligned}
P &\sim \mathrm{Beta}\left(\alpha, \beta\right) \\
X_k \mid P &\sim \mathrm{Binomial}\left(k, p\right) \\
P \mid X_k = x_k &\sim \mathrm{Beta}\left(\alpha + x_k, \beta + (k - x_k)\right)
\end{aligned}
$$


α, β
p
x_k

- Clearly not all uncertainty can be properly modelled by a simple structure.
- Imagine a study on the effectiveness of cardiac treatments in a hospital,

$$y_{71}/k_{71} = 10/14$$

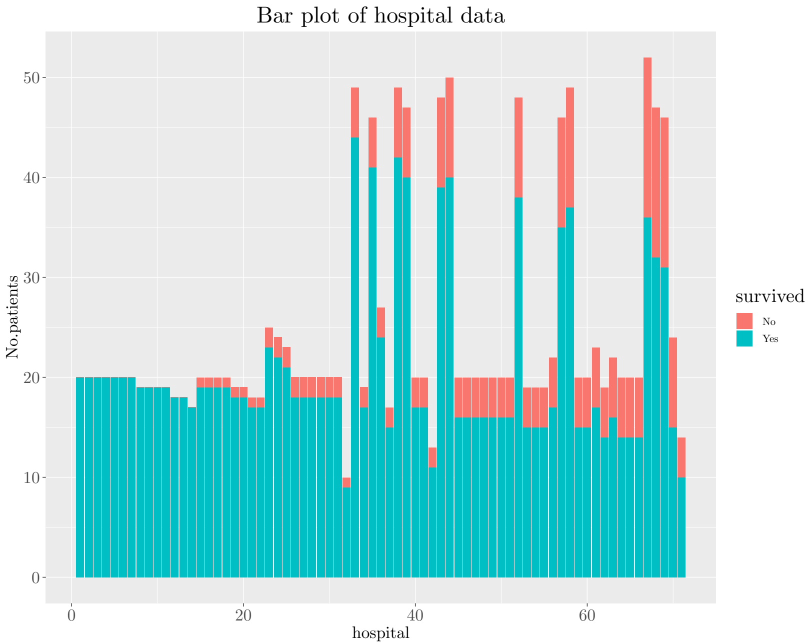
patients survived. Historically, similar data exist for other hospitals in the city

$$y_j/k_j \qquad \text{where } j = 1, 2, \ldots, 70.$$

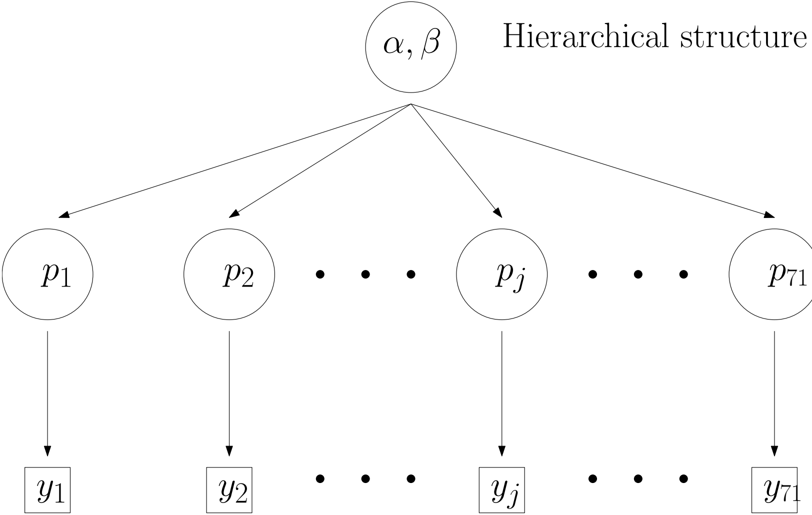
- It is clearly not ideal to collapse all data into a single value of y and k

$$y = \sum_{i=1}^{71} y_i; \qquad k = \sum_{i=1}^{71} k_i$$

since it is unlikely hospitals have the same equipments, let alone the doctors.
- It is more realistic to assume that the patents in hospital j having their own survival probability p_j, which might be related to each other in some sense.


Bar plot of hospital data
No.patients
0
10
20
30
40
50
hospital
0
20
40
60
survived
No
Yes

- If we assume p_j's follow a common distribution, then we have the following



- It is reasonable to assume the following

$$P_j \sim \mathrm{Beta}\,(\alpha, \beta)$$
$$Y_j \mid P_j \sim \mathrm{Binomial}\,(k_j, p_j)$$

Q: How to specify α and β, and how to obtain the posterior of P_{71}?

- We have discussed collapsing historical data, $j = 1, 2, \ldots 70$, with current data $j = 71$ is not ideal, that is the following estimation is not ideal

$$P \sim \mathrm{Beta}\,(1, 1)$$
$$Y \mid P \sim \mathrm{Binomial}\,(k = 1739, p)$$
$$P \mid X_{1739} = 1472 \sim \mathrm{Beta}\,(1 + 1472, 1 + (1739 - 11472))$$

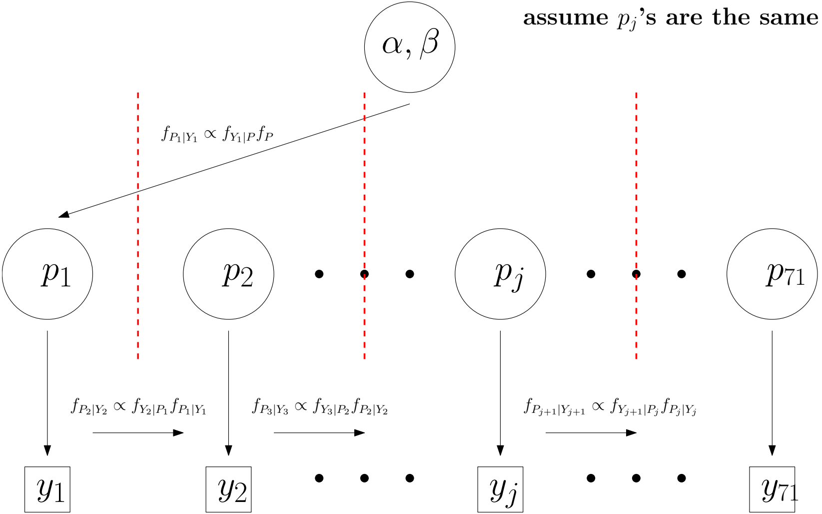
which treat it as Bayes' original problem, and should not be taken seriously.

- Using historical data sequentially as before to obtain a prior is just as bad.

$$\text{uniform prior} \xrightarrow{y_1} \text{posterior as prior} \xrightarrow{y_2} \cdots \xrightarrow{y_{71}} \text{posterior}$$

- It is just as bad since it is equivalent to collapsing the data, in which we

assume p_j's are the same



Q: How can we incorporate the information in the historical data into our prior?

- Perhaps one rather simple and nature way in this case is to combine

  frequentist and Bayesian approaches

- The historical data, $j = 1, 2, \ldots, 70$, can be treated as observed proportions

$$P \sim \mathrm{Beta}\,(\alpha, \beta)$$

- That is, instead of treating p_j's as unobserved and having the lay of y_j's, let

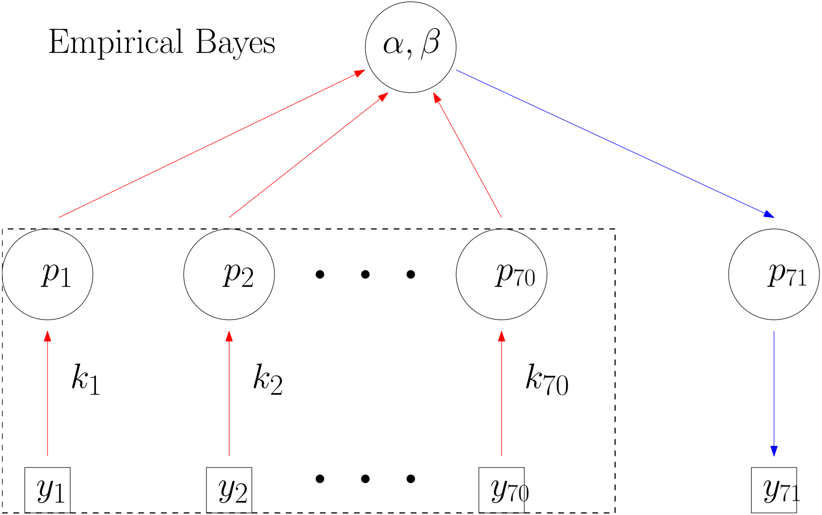
$$p_j = y_j / k_j \qquad \text{for} \quad j = 1, 2, \ldots 70$$

- However, we keep the current data on hospital 71 as it is

$$y_{71} = 10, \qquad k_{71} = 14$$

  and still in search for the posterior of the random variable P_{71} given the data.

- So we pretty much remove a layer from the structure, and use it separately



- Using the 70 proportions,

$$p_j = y_j / k_j \qquad \text{for} \quad j = 1, 2, \ldots 70$$

we could estimate α and β using an frequentist approach, e.g.

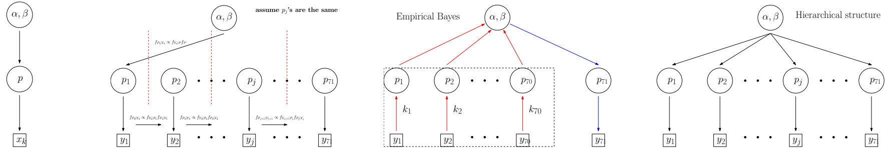
$$\tilde{\alpha} = \bar{p}\left(\frac{\bar{p}\,(1-\bar{p})}{s_p^2} - 1\right); \qquad \tilde{\beta} = (1-\bar{p})\left(\frac{\bar{p}\,(1-\bar{p})}{s_p^2} - 1\right)$$

- Here the method of moments (MoM) is used since the MLE is not analytic.
- The prior is set to be $P_{71} \sim \mathrm{Beta}\left(\tilde{\alpha}, \tilde{\beta}\right)$, from which we obtain

$$P_{71} \mid Y_{71} = 10 \sim \mathrm{Beta}\left(\tilde{\alpha} + 10, \tilde{\beta} + 4\right)$$

which is known as the pseudo-posterior, using Bayes theorem as usual.
- This half frequentist half Bayesian method is known as empirical Bayes.

- To derive a full Bayesian analysis for the hierarchical structure, we have to understand the difference in how data provide information



- In a simple structure, the data x_k, provide the information directly on the distribution of p, however, in a hierarchical structure, the historical data,

$$\mathbf{y}_h = \begin{bmatrix} y_1 & y_2 & \cdots & y_{70} \end{bmatrix}^{\mathrm{T}}$$

  provide no information on p_{71} directly; it is only through α and β, that $\mathbf{y}_h$ is useful to improve our knowledge on p_{71} in a way similar to empirical Bayes.

- Since $\mathbf{y}_h$ improves our understanding on α and β, we have to treat α and β as random as well. The distribution $f_{\alpha,\beta}$ is known as hyperprior.

- Therefore the joint posterior distribution

$$f_{\{\mathbf{P},\alpha,\beta\}|\mathbf{Y}=\mathbf{y}}$$

where $\mathbf{y} = \begin{bmatrix}\mathbf{y}_h & y_{71}\end{bmatrix}^{\mathrm{T}}$ denotes all the data, and $\mathbf{P}$ the random vector

$$\mathbf{P} = \begin{bmatrix} P_1 \\ \vdots \\ P_{71} \end{bmatrix}$$

would capture our most up to date understanding of the system according to the Bayesian hierarchical model, in which we have to specify a likelihood

$$\mathcal{L}(\mathbf{P}, \mathbf{y}) = f_{\mathbf{y}|\{\mathbf{P},\alpha,\beta\}}$$

and a joint prior in terms of a conditional prior and a hyperprior

$$f_{\mathbf{P},\alpha,\beta} = f_{\mathbf{P}|\{\alpha,\beta\}} \cdot f_{\alpha,\beta}$$

- Recall we have been using a single-parameter version of Bayes theorem so far

$$f_{Y|X} \propto f_{X|Y} f_Y$$

and the continuous version is based on the definition conditional probability

$$\begin{aligned}
F_{Y|X}(y \mid x) &= \lim_{\varepsilon \to 0^+} \Pr\left(Y \le y \mid X \in (x, x+\varepsilon]\right) \\
&= \lim_{\varepsilon \to 0^+} \frac{\Pr\left(Y \le y, X \in (x, x+\varepsilon]\right)}{\Pr\left(x < X \le x+\varepsilon\right)} \\
&= \lim_{\varepsilon \to 0^+} \frac{F_{X,Y}(x+\varepsilon, y) - F_{X,Y}(x, y)}{F_X(x+\varepsilon) - F_X(x)} = \frac{\partial F_{X,Y}(x,y)/\partial x}{f_X(x)}
\end{aligned}$$

$$\begin{aligned}
\implies f_{Y|X}(y \mid x) &= \frac{\partial}{\partial y} \frac{\partial F_{X,Y}(x,y)/\partial x}{f_X(x)} \\
&= \frac{1}{f_X(x)} \frac{\partial^2 F_{X,Y}(x,y)}{\partial y \partial x} \\
&= \frac{f_{X,Y}(x,y)}{f_X(x)} \propto f_{X|Y}(x \mid y) \cdot f_Y(y)
\end{aligned}$$

- The same approach can be used to derive a multi-parameter version, e.g.

$$F_{\{Y,Z\}|X} = \lim_{\varepsilon \to 0^+} \Pr\left(Y \le y, Z \le z \mid X \in (x, x+\varepsilon]\right) = \frac{\partial F_{X,Y,Z}/\partial x}{f_X}$$

$$\implies f_{\{Y,Z\}|X} = \frac{\partial^2}{\partial y \partial z} \frac{\partial F_{X,Y,Z}/\partial x}{f_X} = \frac{f_{X,Y,Z}}{f_X} = \frac{f_{X|\{Y,Z\}} \cdot f_{Y,Z}}{f_X}$$

$$\propto f_{X|\{Y,Z\}} \cdot f_{Y,Z} \propto f_{X|\{Y,Z\}} \cdot f_{Y|Z} \cdot f_Z$$

- In this case, with the assumption of independence, the joint posterior is

$$\begin{aligned}
f_{\{\mathbf{P},\alpha,\beta\}|\mathbf{Y}} &\propto f_{\mathbf{Y}|\{\mathbf{P},\alpha,\beta\}} \cdot {\color{red} f_{\mathbf{P}|\{\alpha,\beta\}}} \cdot {\color{blue} f_{\alpha,\beta}} \\
&= \prod_{j=1}^{71} \binom{k_j}{y_j} p_j^{y_j} (1-p_j)^{k_j - y_j} {\color{red} \prod_{j=1}^{71} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p_j^{\alpha-1} (1-p_j)^{\beta-1}} {\color{blue} f_{\alpha,\beta}} \\
&= f_{\alpha,\beta} \prod_{j=1}^{71} \binom{k_j}{y_j} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p_j^{y_j+\alpha-1} (1-p_j)^{k_j - y_j + \beta - 1}
\end{aligned}$$

- The conditional posterior of $\mathbf{P}$ given α and β is

$$f_{\mathbf{P}|\{\alpha,\beta,\mathbf{Y}\}} = \frac{f_{\{\mathbf{P},\alpha,\beta\}|\mathbf{Y}}}{f_{\{\alpha,\beta\}|\mathbf{Y}}} \propto f_{\{\mathbf{P},\alpha,\beta\}|\mathbf{Y}}$$
$$\propto \prod_{j=1}^{71} p_j^{y_j+\alpha-1}(1-p_j)^{k_j-y_j+\beta-1}$$

$$= \prod_{j=1}^{71} \underbrace{\frac{\Gamma(\alpha+\beta+k_j)}{\Gamma(\alpha+y_j)\Gamma(\beta+k_j-y_j)}}_{\text{normalisation constant } c} p_j^{y_j+\alpha-1}(1-p_j)^{k_j-y_j+\beta-1}$$

where the normalisation constant c is due to the fact that components of $\mathbf{P}$ have independent posterior densities of the form

$$p_j^{\alpha_j^*-1}(1-p_j)^{\beta_j^*-1}$$

thus c is found by finding the normalisation constant of a beta distribution.

- The marginal posterior of α and β can be obtained since the joint posterior

$$f_{\{\mathbf{P},\alpha,\beta\}|\mathbf{Y}} \propto f_{\alpha,\beta} \prod_{j=1}^{71} \binom{k_j}{y_j} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p_j^{y_j+\alpha-1}(1-p_j)^{k_j-y_j+\beta-1}$$

and the conditional posterior of $\mathbf{P}$ given α and β

$$f_{\mathbf{P}|\{\alpha,\beta,\mathbf{Y}\}} = \prod_{j=1}^{71} \frac{\Gamma(\alpha+\beta+k_j)}{\Gamma(\alpha+y_j)\Gamma(\beta+k_j-y_j)} p_j^{y_j+\alpha-1}(1-p_j)^{k_j-y_j+\beta-1}$$

have been worked out, thus the marginal posterior of α and β is given by

$$\begin{aligned} f_{\{\alpha,\beta\}|\mathbf{Y}} &= \frac{f_{\{\mathbf{P},\alpha,\beta\}|\mathbf{Y}}}{f_{\mathbf{P}|\{\alpha,\beta,\mathbf{Y}\}}} \\ &\propto f_{\alpha,\beta} \prod_{j=1}^{71} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+y_j)\Gamma(\beta+k_j-y_j)}{\Gamma(\alpha+\beta+k_j)} \end{aligned}$$