bayesian_stats_ch3_poisson

inoue jin

2022-11-22

“3. **二項モデルとポアソンモデル**” より

■GSS の例: ポアソンモデル　総合的社会調査（GSS）より、40 歳の女性 155 人の学歴と子供の数に関するデータを収集。

学士号を持つかどうか $(Y = 1)$ で、女性の子供の数を比較する。

$Y_{i,1}$ を学士号を持たない n_1 人の女性の子供の数とし、$Y_{i,2}$ を学士号を持つ女性の子供の数とする。

サンプリングモデルは以下の通り

$$Y_{1,1}, \dots, Y_{n_1,1} \mid \theta_1 \sim \text{i.i.d. } Poisson(\theta_1),$$

$$Y_{1,2}, \dots, Y_{n_2,2} \mid \theta_2 \sim \text{i.i.d. } Poisson(\theta_2),$$

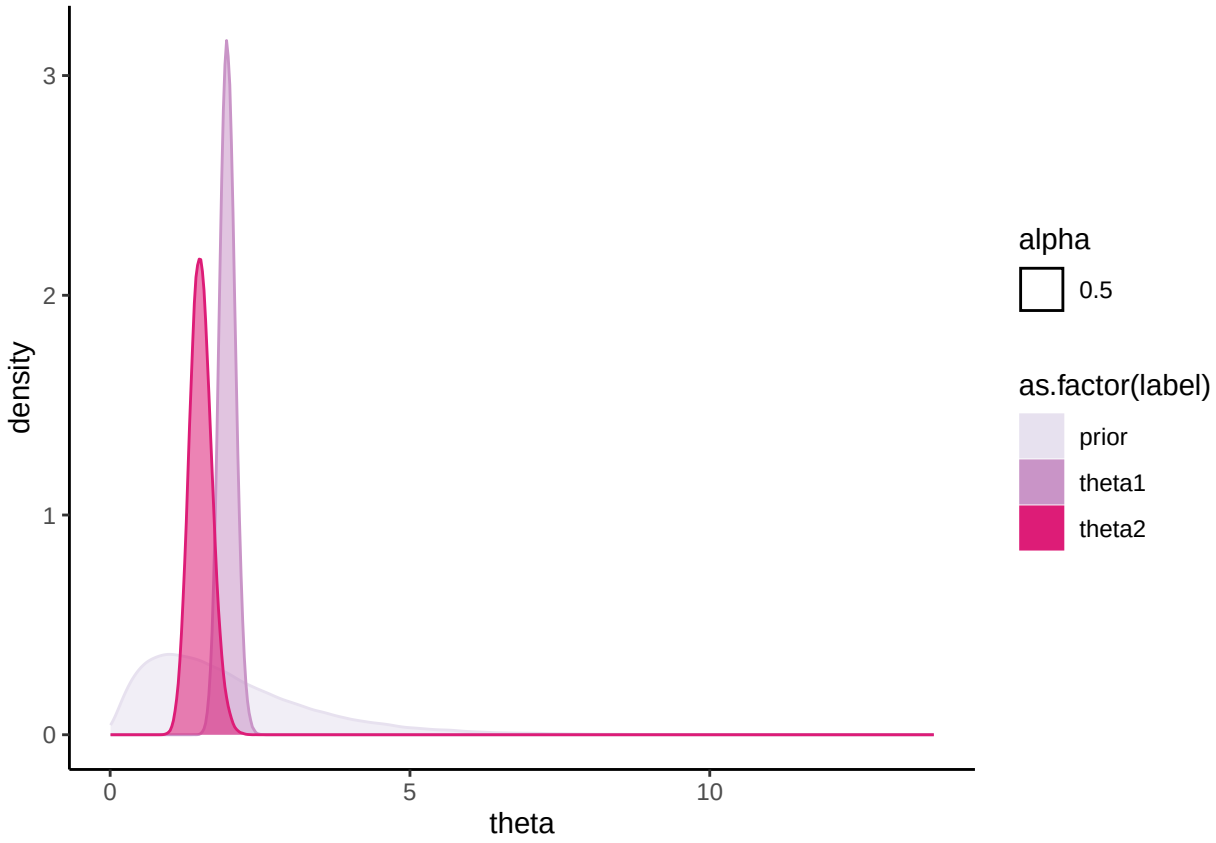
```
## [1] 1.955357
```

```
## [1] 1.704943 2.222679
```

■事後分布　以下の事後分布の比較から、大まかな傾向として $\theta_1 > \theta_2$ となっていることが伺える。

```
# 予測分布のプロット
# それぞれの予測分布と共通の事前分布から乱数生成してプロットしてみる

N <- 100000

t1 <- tibble(theta = rgamma(N, a+sy1, b+n1),
             label = "theta1")
t2 <- tibble(theta = rgamma(N, a+sy2, b+n2),
             label = "theta2")
p <- tibble(theta = rgamma(N, a, b),
```

```
            label = "prior")

df <- bind_rows(t1, t2, p)


df %>%
  ggplot(aes(x = theta, color = as.factor(label))) +
  geom_density(aes(fill = as.factor(label), alpha = 0.5)) +
  scale_fill_brewer(palette= "PuRd")+
  scale_color_brewer(palette = "PuRd")+
  theme_classic()
```



■予測分布　予測分布の式を導出する

$$
\begin{aligned}
p(\tilde{y} \mid y_1, \dots, y_n) &= \int_0^\infty p(\tilde{y}, \theta \mid y_1, \dots, y_n) d\theta \\
&= \int_0^\infty p(\tilde{y} \mid \theta, y_1, \dots, y_n) p(\theta \mid y_1, \dots, y_n) d\theta \\
&= \int_0^\infty \frac{p(\tilde{y}, y_1, \dots, y_n \mid \theta) p(\theta)}{p(\theta, y_1, \dots, y_n)} p(\theta \mid y_1, \dots, y_n) d\theta \\
&= \int_0^\infty \frac{p(\tilde{y} \mid \theta) p(y_1 \mid \theta) \cdots p(y_n \mid \theta)}{p(y_1 \mid \theta) \cdots p(y_n \mid \theta)} p(\theta \mid y_1, \dots, y_n) d\theta \\
&= \int_0^\infty p(\tilde{y} \mid \theta) p(\theta \mid y_1, \dots, y_n) d\theta \\
&= \int_0^\infty dpois(\tilde{y}, \theta) dgamma(\theta, a + \sum_i y_i, b + n) d\theta \\
&= \int_0^\infty \left(\frac{\theta^{\tilde{y}} e^{-\theta}}{\tilde{y}!} \right) \left(\frac{(b+n)^{a+\sum_i y_i}}{\Gamma(a + \sum_i y_i)} \theta^{a+\sum_i y_i - 1} e^{-(b+n)\theta} \right) d\theta \\
&= \frac{\Gamma(a + \sum_i y_i + \tilde{y})}{\Gamma(\tilde{y}+1)\Gamma(a + \sum_i y_i)} \left(\frac{b+n}{b+n+1} \right)^{a+\sum_i y_i} \left(\frac{1}{b+n+1} \right)^{\tilde{y}}
\end{aligned}
$$

1 行目は周辺確率密度関数の定義より、2 行目は確率の公理 $P(F \cap G \mid H) = P(F \mid G \cap H) P(G \mid H)$ より従う。3 行目はベイズルール、4 行目はモデルの定義 $Y_1, \dots, Y_n \mid \theta$ i.i.d. $\sim Poisson(\theta)$ より θ を条件づけた後の独立性から従う。6,7 行目は定義より従う。8 行目はガンマ関数の関係 $1 = \int_0^\infty \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta}$ を利用して導ける。

なお、この予測分布は負の二項分布 $NegativeBinomial(a + \sum_i y_i, b + n)$ と一致する。

以下は、子供の数の事後予測分布を” 学士号なし” と” 学士号あり” の場合で分けて可視化したものである。平均出生率 θ の 2 つの事後分布の間の差に比べて、子供の数 $\tilde{Y}$ の 2 つの事後予測分布の間には大きな違いがない。

```
y0 <- 0:10
ngb1 <- tibble(p = dnbinom(y0,size = a+sy1, mu = (a+sy1)/(b+n1)),
               y = y0,
              label = "NoBachelor")

ngb2 <- tibble(p = dnbinom(y0, size = a+sy2,mu = (a+sy2)/(b+n2)),
               y = y0,
               label = "Bachelor")

pred <- bind_rows(ngb1, ngb2)

pred %>% glimpse()
```

```
## Rows: 22
## Columns: 3
```

```
## $ p     <dbl> 1.427473e-01, 2.766518e-01, 2.693071e-01, 1.755660e-01, 8.622930~
## $ y     <int> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, ~
## $ label <chr> "NoBachelor", "NoBachelor", "NoBachelor", "NoBachelor", "NoBache~
```

```
pred %>%
  ggplot(aes(x = y, y = p,color = label)) +
  geom_bar(width = 0.3, stat = "identity", position = "dodge",aes(fill = label, alpha = 0.9 ))+
  theme_pubr()
```