Kelvin Niu

Homework 4

There are *n* fishermen who go fishing in a pond with fisherman *i* fishing for t_i hours. Let y_i be the number of bites for fisherman *i* and θ be the rate per hour of bites in the pond. We are interested in testing the following three hypotheses:
H_0: $\theta \in (0, 1)$
H_1: $\theta \in [1, 5)$
H_2: $\theta \in [5, \infty)$

Bayesian approach

A Bayesian approach to assessing the hypotheses based on collected data would be to look at the posterior probabilities of each hypothesis. That is, if we let

$$M = \begin{cases} 2, if\ H_2\ is\ true \\ 1, if\ H_1\ is\ true \\ 0, if\ H_0\ is\ true \end{cases}$$

then we are interested in the posterior probabilities P(M = 2 | data), P(M = 1 | data), and P(M = 0 | data). It would be reasonable select the hypothesis with the greatest posterior probability.

Assuming a Gamma(a, b) prior (truncated to the range of each hypothesis) and solving for the posterior probabilities, we get

$$P(M = 2 \mid data) = \frac{1}{1 + \left[\frac{F'(1)}{F(1)} + \frac{F'(5) - F'(1)}{F(5) - F(1)}\right] * \frac{1 - F(5)}{1 - F'(5)}}$$

$$P(M = 1 \mid data) = \frac{1}{1 + \left[\frac{1 - F'(5)}{1 - F(5)} + \frac{F'(1)}{F(1)}\right] * \frac{F(5) - F(1)}{F'(5) - F'(1)}}$$

$$P(M = 0 \mid data) = \frac{1}{1 + \left[\frac{1 - F'(5)}{1 - F(5)} + \frac{F'(5) - F'(1)}{F(5) - F(1)}\right] * \frac{F(1)}{F'(1)}}$$

where $F(c)$ is the c.d.f. of $Gamma(a, b)$ and $F'(c)$ is the c.d.f. of $Gamma(a + \sum y_i, b + \sum t_i)$.

[Derivation is attached in appendix]

Simulation of approach

To simulate this approach, I picked different values of theta and for each value I simulated 1000 trials of this approach. I then recorded the percentage of times that the approach correctly guessed the hypothesis that was true. For each trial, I set a Gamma(2, rate = 2/3) prior (since in limited my fishing experience, the hourly rate of bites is probably more likely to be between [1, 5) or 5+ than (0, 1)) and the combined hours of fishing as 30.

Table: Assessment of Bayesian approach under a Gamma(2, 2/3) prior and 30 combined hours of fishing for different values of θ

True θ	Percent correct
0.5	100.0%
0.9	92.5%
1.1	38.9%
3	100.0%
4.9	32.0%
5.1	85.6%
7	100.0%

This approach is very accurate when the true value of θ is in the middle of a hypothesis's range. However, when the value of θ is close to the border between two hypotheses, the performance becomes poorer.

<u>Adding more narrow hypotheses (more bins)</u>
To observe the effect of narrower hypotheses, I tested the following group of hypotheses using the method above:
H_0: $\theta \in (0, 1)$
H_1: $\theta \in [1, 2)$
H_2: $\theta \in [2, 3)$
H_3: $\theta \in [3, 4)$
H_4: $\theta \in [4, 5)$
H_5: $\theta \in [5, \infty)$

Again, I simulated 1000 trials, using a Gamma(2, rate = 2/3) prior and 30 combined hours of fishing using the same theta values as before. The results are below, comparing the original set of 3 hypotheses with the new narrower set of 6 hypotheses.

Table: Assessment of Bayesian approach for two sets of hypotheses

True θ	Percent correct original	Percent correct narrower
0.5	100.0%	100.0%
0.9	92.5%	76.2%
1.1	38.9%	67.5%
3.1	100.0%	66.0%
4.9	32.0%	75.1%
5.1	85.6%	39.9%
7.1	100.0%	100.0%

Generally, the Bayesian method of posterior probabilities performs poorer when there are more hypotheses and narrower bins. This makes sense because as we try to narrow down the true value of theta, we lose some accuracy. However, in a few cases where the value of theta was considered on the border of an original hypothesis and became "less on the border" (relative to the size of the interval) of a new narrower hypothesis, the performance actually improved.

<u>Testing two ponds against each other</u>

Let θ_A be the rate of pond A and θ_B be the rate of pond B. To test whether fishing is better in one pond than another, we are interested in testing the following hypotheses:

H_0: $\theta_A = \theta_B$

H_1: $\theta_A \neq \theta_B$

Again, we can look at the posterior probabilities of each hypothesis and select the one with the highest posterior probability.

That is, if we let

$$M = \begin{cases} 1, if\ H_1\ is\ true \\ 0, if\ H_0\ is\ true \end{cases}$$

then we are interested in the posterior probabilities P(M = 1 | data) and P(M = 0 | data). In calculating the posterior probabilities, we would need to assume a single prior under the null

H_0: $\theta \sim$ Gamma(a, b)

and two priors under the alternative

H_1: $\theta_A \sim$ Gamma(c, d) and $\theta_B \sim$ Gamma(e, f)

Kelvin Niu

Derivation of posterior probabilities

$$M = \begin{cases} 2, & \text{if } H_2 \text{ is true} \\ 1, & \text{if } H_1 \text{ is true} \\ 0, & \text{if } H_0 \text{ is true} \end{cases}$$

let us set an uninformed prior on M : $P(M=2) = P(M=1) = P(M=0) = 1/3$

The posterior probability for hypothesis k is

$$P(M=k \mid \text{data}) = \frac{P(M=k)L(\text{data} \mid M=k)}{P(M=2)L(\text{data} \mid M=2) + P(M=1)L(\text{data} \mid M=1) + P(M=0)L(\text{data} \mid M=0)}$$

$$= \frac{1}{\frac{P(M=2)L(\text{data} \mid M=2)}{P(M=k)L(\text{data} \mid M=k)} + \frac{P(M=1)L(\text{data} \mid M=1)}{P(M=k)L(\text{data} \mid M=k)} + \frac{P(M=0)L(\text{data} \mid M=0)}{P(M=k)L(\text{data} \mid M=k)}}$$

$$= \frac{1}{\frac{L(\text{data} \mid M=2)}{L(\text{data} \mid M=k)} + \frac{L(\text{data} \mid M=1)}{L(\text{data} \mid M=k)} + \frac{L(\text{data} \mid M=0)}{L(\text{data} \mid M=k)}} \qquad (1)$$

The marginal likelihood under hypothesis k is

$$P(\text{data} \mid M=k) = \int_R L(y_{1:n} \mid \theta)\, \pi(\theta)\, d\theta$$

(2), where R is the relevant region for hypothesis k

$y_i \sim \text{Poisson}(t_i \theta)$

$$L(y_{1:n} \mid \theta) = \prod_{i=1}^{n} \frac{(t_i\theta)^{y_i} e^{-t_i\theta}}{y_i!} = \theta^{\Sigma y_i} e^{-\Sigma t_i \theta} \prod_{i=1}^{n} \frac{t_i^{y_i}}{y_i!}$$

For each hypothesis, let us set a truncated gamma prior on θ

$H_0: \theta \sim \text{Gamma}_{(0,1)}(a,b)$

$H_1: \theta \sim \text{Gamma}_{[1,5)}(a,b)$

$H_2: \theta \sim \text{Gamma}_{[5,\infty)}(a,b)$

Thus, the marginal likelihood from equation (2) becomes

$$P(\text{data} \mid M=k) = \int_R \left[\prod_{i=1}^{n} \frac{t_i^{y_i}}{y_i!}\right] \theta^{\Sigma y_i} e^{-\Sigma t_i \theta} \frac{c(a,b)\, \theta^{a-1} e^{-b\theta}}{F(R;a,b)}$$

constant for Gamma (a,b)

area of region R for Gamma (a,b)

$$= \frac{c(a,b)}{F(R;a,b)} \prod_{i=1}^{n} \frac{t_i^{y_i}}{y_i!} \int_R \theta^{a+\Sigma y_i - 1} e^{-(b+\Sigma t_i)\theta} \frac{c(\tilde{a},\tilde{b})/F(R;\tilde{a},\tilde{b})}{c(\tilde{a},\tilde{b})/F(R;\tilde{a},\tilde{b})}$$

$\tilde{a} = a + \Sigma y_i$

$\tilde{b} = b + \Sigma t_i$

$$= \frac{c(a,b)\, F(R;\tilde{a},\tilde{b})}{c(\tilde{a},\tilde{b})\, F(R;a,b)} \prod_{i=1}^{n} \frac{t_i^{y_i}}{y_i!}$$

$$= C \cdot \frac{F(R;\tilde{a},\tilde{b})}{F(R;a,b)} \qquad (3)$$

constant $C = \frac{c(a,b)}{c(\tilde{a},\tilde{b})} \prod_{i=1}^{n} \frac{t_i^{y_i}}{y_i!}$

Plugging in result (3) into equation (1), we get the posterior probabilities

$$P(M=0 \mid \text{data}) = \frac{1}{1 + \left[\frac{F([5,\infty);\tilde{a},\tilde{b})}{F([5,\infty);a,b)} + \frac{F([1,5);\tilde{a},\tilde{b})}{F([1,5);a,b)}\right] \cdot \frac{F((0,1);a,b)}{F((0,1);\tilde{a},\tilde{b})}}$$

$$= \boxed{\frac{1}{1 + \left[\frac{1-F'(5)}{1-F(5)} + \frac{F'(5)-F'(1)}{F(5)-F(1)}\right] \cdot \frac{F(1)}{F'(1)}}}$$

where $F(c)$ is the cdf of Gamma (a,b) and $F'(c)$ is the cdf of Gamma $(\tilde{a},\tilde{b})$

$$P(M=1 \mid data) = \frac{1}{1 + \left[\frac{F([5,\infty); \tilde{a}, \tilde{b})}{F([5,\infty); a, b)} + \frac{F((0,1); \tilde{a}, \tilde{b})}{F((0,1); a, b)} \right] \cdot \frac{F([1,5); a, b)}{F([1,5); \tilde{a}, \tilde{b})}}$$

$$= \boxed{\frac{1}{1 + \left[\frac{1 - F'(5)}{1 - F(5)} + \frac{F'(1)}{F(1)} \right] \cdot \frac{F(5) - F(1)}{F'(5) - F'(1)}}}$$

$$P(M=2 \mid data) = \frac{1}{1 + \left[\frac{F((0,1); \tilde{a}, \tilde{b})}{F((0,1); a, b)} + \frac{F([1,5); \tilde{a}, \tilde{b})}{F([1,5); a, b)} \right] \cdot \frac{F([5,\infty); a, b)}{F([5,\infty); \tilde{a}, \tilde{b})}}$$

$$= \boxed{\frac{1}{1 + \left[\frac{F'(1)}{F(1)} + \frac{F'(5) - F'(1)}{F(5) - F(1)} \right] \cdot \frac{1 - F(5)}{1 - F'(5)}}}$$