LSTAT2130 - Introduction to Bayesian Statistics
Project 2020-21

A survey was conducted in 2018 in Belgium among a series of households randomly selected from the National Register. Among the many questions asked, one concerned net disposable income. We limited our attention here to 1128 respondents aged 30 years or older. They were asked to select one of 10 possible net income ranges for their household. The following frequency table summarises the information collected, broken down by the two main regions of the country:

	Net monthly household income (in euros)										
Region	< 1200	$[1200, 1500)$	[1500,1800)	[1800,2300)	[2300,2700)	[2700,3300)	[3300,4000)	[4000,4900)	[4900,6000)	6000+	Total
Flanders	25	69	65	106	80	106	136	94	76	46	803
Wallonia	17	36	47	58	47	53	59	54	33	21	425

Assume that the distribution of monthly net income X is adequately described by a Gamma. As a reminder, $X \sim \mathcal{G}(\kappa, \lambda)$ if its density $f(x)$ is such that

$$f(x) = \frac{\lambda^\kappa}{\Gamma(\kappa)} x^{\kappa-1} \mathrm{e}^{-\lambda x} \quad \text{if} \quad x \geq 0.$$

The mean and variance of X are given by

$$\mathrm{E}(X) = \mu = \frac{\kappa}{\lambda} \quad ; \quad \mathrm{V}(X) = \frac{\kappa}{\lambda^2} = \frac{1}{\kappa}\mu^2 = \phi\mu^2 \ .$$

In order to facilitate your answers to the questions below, I suggest to reparameterize the gamma density using the mean μ and the dispersion parameter ϕ. In practice, that means that obtaining with `R` the density $f(x; \mu, \phi)$ or the cdf $F(x; \mu, \phi)$ of the Gamma for given μ and ϕ values would require typing

```
dgamma(x,1/phi,1/(phi*mu)) ; pgamma(x,1/phi,1/(phi*mu))
```

repectively. In particular, it implies that to obtain the probability that $2300 \leq X < 2700$ when $\mu = 2500$ and $\phi = 1.0$, one can just type

```
mu = 2500 ; phi = 1.0
kappa = 1/phi ; lambda = 1/(phi*mu)
pgamma(2700,kappa,lambda)-pgamma(2300,kappa,lambda)
```

These differences of cdf's will play a crucial role to obtain the likelihood.

Questions

1. Let $\theta_k = (\mu_k, \phi_k)$ be the set of parameters for a household net income distribution in Region k ($k = 1, 2$):

   (a) Using the notations defined on Page 1, write the theoretical expression for the probability that a randomly selected household in the population of Region k has an income in the jth interval or semi-interval ($j = 1, \ldots, 10$) mentioned in the frequency table.

   (b) Based on your previous answer and remembering that the likelihood is (up to a multiplicative constant) the probability of the observed data given the assumed probability model for the data generating mechanism, write the theoretical expression for the likelihood $L(\theta_k; \mathcal{D}_k)$ associated to the data $\mathcal{D}_k$ collected in Region k.
   [Hint: the frequency distribution in a given region is multinomial]

2. Propose priors for μ_k and ϕ_k translating the following prior information:

   - Statement 1: *we are at 95% convinced that the mean net monthly household income in a given region is in the interval* $(2400, 3600)$.
   - Statement 2: *we are certain that* ϕ *lies in the interval* $(0.0, 10.0)$.

3. Starting from your previous answers:

   (a) Obtain a theoretical expression for the joint posterior of the income distribution parameters θ_1 in Flanders.

   (b) Write a R function `lpost(theta,freq)` computing the logarithm of the joint posterior for a given vector of parameters in `theta` and an observed vector of frequencies in `freq`.
   [Numerical advice: direcly compute your expressions on the log-scale]

4. Write R code providing the Laplace approximation to the joint posterior of θ_1. Based on that approximation, provide an approximate 95% credible interval for the mean net monthly household income in Flanders.

5. Using pure R code,

   (a) Implement a componentwise Metropolis algorithm where the components of θ_1 are updated one at a time.

   (b) Visualize and confirm the convergence of your chains for μ_1 & ϕ_1 (possibly with the additional help of the `coda` package).

   (c) Compute a credible interval for μ_1 and compare it with the one obtained with the Laplace approximation.

6. Redo the same exercise using JAGS and compare your results with those obtained in Question 5.

7. Repeat the same analysis with JAGS on the data from Wallonia.

8. Starting from the chains generated with JAGS for the mean income in Flanders (μ_1) and Wallonia (μ_2), obtain a 95% credible interval for the mean income difference, $\mu_1 - \mu_2$. What can you conclude from this ?

INSTRUCTIONS

- By **Tuesday 25th May 2021 at 13:00**, each group of 3 students must transmit their results to Hortense DOMS by uploading the following 2 documents to the MoodleUCL platform:

  1. A single PDF file containing the report detailing the answers to all the preceding questions (using the same structure and numbering as within the questionnaire). The software code must be in appendix and referred clearly in the main text.

     The file should be named using your family names in a row as in the following example:

     Smith-Jones-Brown_Report.pdf

  2. A single text file containing the commented software code (R and JAGS only) enabling to reproduce the claimed results. Its subdivision must follow the same structure as in the questionnaire.

     The file should be named using your family names in a row as in the following example:

     Smith-Jones-Brown_Rcode.R

  There is no second chance for this report for a later exam session, see below.

- Each group of 3 students must work independently !! Any detected fraud will lead to a severe penalty.
- Any change to the agreed group composition will lead to a zero score for your project.
- Each member of a group must work on all aspects of the project (no "specialization").

YOUR FINAL MARK FOR THIS COURSE:

Your final mark (E: max 20 points) for LSTAT2130 will be obtained by rounding to the closet integer the sum of the results at:

- the written exam (W: max 15 points) in June or August-September ;
- the project (P: max 5 points) (written report in May, no second chance for a later session):