Université catholique de Louvain

Louvain School of Statistics

LSTAT2130 - Bayesian Statistics

Project - Group Q

Lionel Lamy - 1294-1700
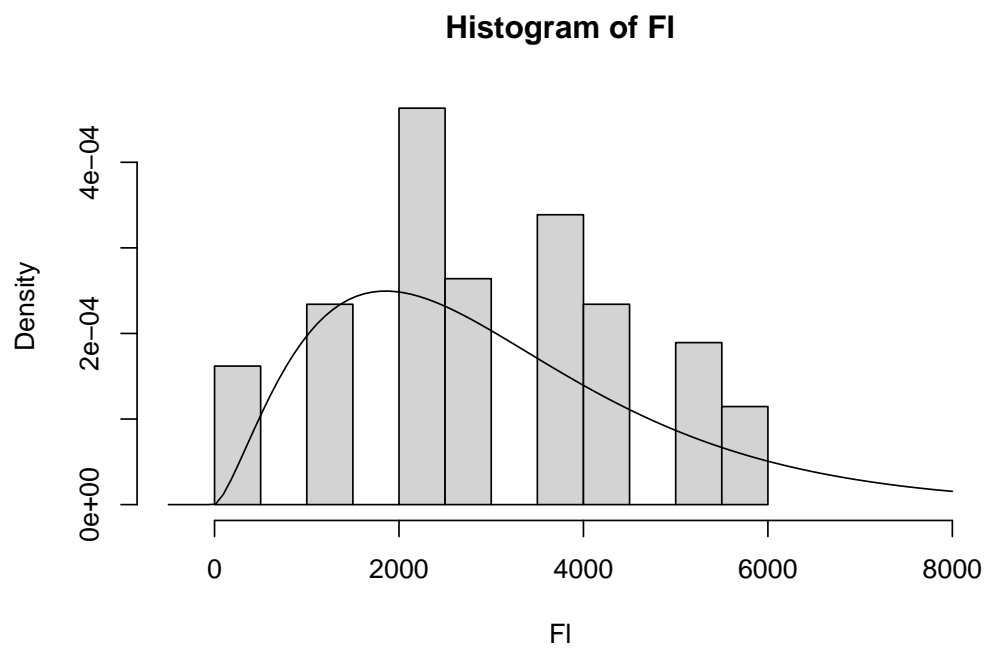
Adrien Kinart

Simon Lengendre

May 17, 2021

Contents

1 Introduction

```
##      shape
## 3089.944
```

```
##      shape
## 0.399687
```

Histogram of FI


```
##      shape
## 2914.882
```

```
##      shape
## 0.471615
```

Histogram of Wall
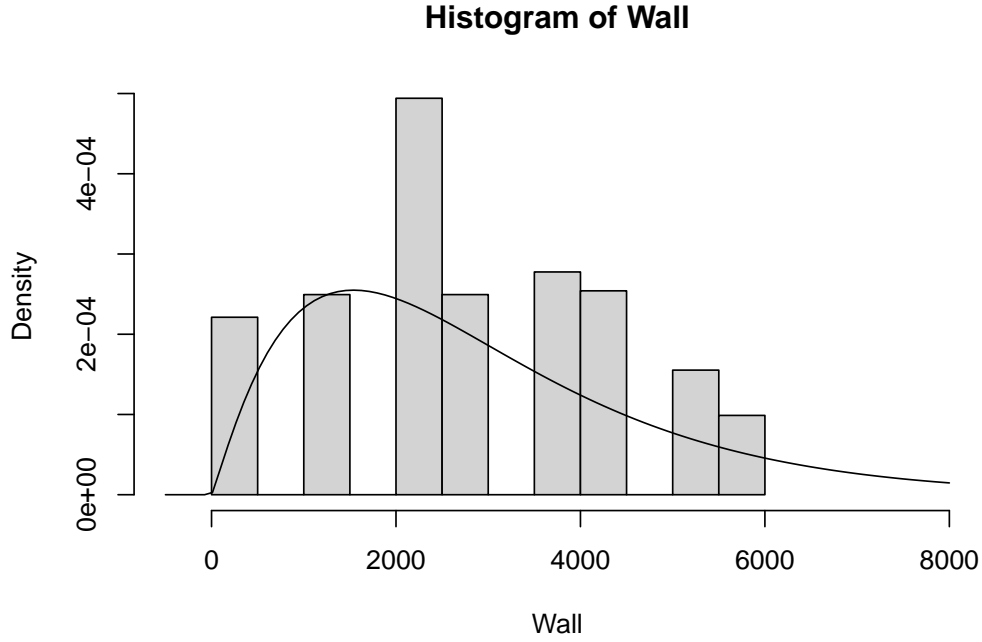

```
## [1] 0.05892352
```

```
## [1] 0.0001358382
```

2 Question 1

Let $\theta_k := (\mu_k, \phi_k)$ be the set of parameters for a HNI with respect to region k.

2.1 (a) Theoretical probability

Let X be the monthly net income of 1123 Belgian households net income (HNI) older than 30 years. Regardless the 2 regions ($k = \{1, 2\}$ wrt Flanders and Wallonnia, respectively), is assumed it follows a Gamma distribution. It can be reparametrised in terms of its mean μ and dispersion parameter ϕ with the following trick:

$$\text{shape: } \kappa = \frac{1}{\phi}$$

$$\text{rate: } \lambda = \frac{1}{\phi\,\mu}$$

For both regions $k = \{1, 2\}$: This gives

$$f(x_k) = \frac{(\phi_k \mu_k)^{-1/\phi_k}}{\Gamma(\phi_k^{-1})}\, x_k^{1/\phi_k - 1} \exp\left(\frac{-x_k}{\phi_k \mu_k}\right)$$

Then, the probability to fall into a certain HNI interval I is:

$$P(x_k \in I_j) = \int_{I_j} \frac{(\phi_k \mu_k)^{-1/\phi_k}}{\Gamma(\phi_k^{-1})} x_k^{1/\phi_k - 1} \exp\left(\frac{-x_k}{\phi_k \mu_k}\right) \, dx$$

Using CDF writing, for a region k ,

$$p_j = \begin{cases} F(x_j, \kappa, \lambda) \text{ if j} =1 \\ F(x_{j+1}, \kappa, \lambda) - F(x_j, \kappa, \lambda) \text{ if } j \in \{2, ..9\} \\ 1 - F(x_j, \kappa, \lambda) \text{ if j} =10 \end{cases}$$

$$p_j = \begin{cases} \frac{1}{\Gamma(\kappa)} \gamma(\kappa, \lambda x_j) \text{ if j} =1 \\ \frac{1}{\Gamma(\kappa)} \left(\gamma(\kappa, \lambda x_{j+1}) - \gamma(\kappa, \lambda x_j) \right) \text{ if} j \in \{2, ..9\} \\ 1 - \frac{1}{\Gamma(\kappa)} \gamma(\kappa, \lambda x_j) \text{ if j} =10 \end{cases}$$

In terms of μ and ϕ

$$p_j = \begin{cases} \frac{1}{\Gamma(\phi^{-1})} \gamma(\phi^{-1}, (\mu\phi)^{-1} x_j) \text{ if j} =1 \\ \frac{1}{\Gamma(\phi^{-1})} \left(\gamma(\phi^{-1}, (\mu\phi)^{-1} x_{j+1}) - \gamma(\kappa, (\mu\phi)^{-1} x_j) \right) \text{ if} j \in \{2, ..9\} \\ 1 - \frac{1}{\Gamma(\phi^{-1})} \gamma(\phi^{-1}, (\mu\phi)^{-1} x_j) \text{ if j} =10 \end{cases}$$

$$p_j = \begin{cases} \frac{1}{\Gamma(\phi^{-1})} \int_0^{(\mu\phi)^{-1} x_j} t^{\phi^{-1}-1} e^{-t} dt \text{ if j} =1 \\ \frac{1}{\Gamma(\phi^{-1})} \int_{(\mu\phi)^{-1} x_j}^{(\mu\phi)^{-1} x_{j+1}} t^{\phi^{-1}-1} e^{-t} dt \text{ if} j \in \{2, ..9\} \\ \frac{1}{\Gamma(\phi^{-1})} \int_{(\mu\phi)^{-1} x_j}^{+\infty} t^{\phi^{-1}-1} e^{-t} dt \text{ if j} =10 \end{cases}$$

2.2   (b) Theoretical expression for the likelihood

On behalf of writing simplicity, the region index is removed. Since the frequency distribution in a given region is multinomial, i.e.

We have, writing $P := (p1, .., p_{10})$ and $X := (X_1, ... X_{10})$:

$$X|P \sim \text{Mul}(N, P) = \frac{x!}{x_1! \, ... \, x_{10}!} p_1^{x_1} \times ... \times p_{10}^{x_{10}} \text{ when } \sum_{j=1}^{10} p_j = 1$$

$$= 0 \text{ otherwise}$$

Up to a multiplicative constant, the likelihood can be written as follow:

$$L(\theta_k, D_k) = P(D_k | \mu_k, \phi_k) \propto \prod_{j=1}^{10} p_{k,j}^{x_{k,j}}$$

2.3 Taking approximation

p_j corresponds to the area in the jth interval. One can take the approximation mean the mean,e.g. $x_{Flanders,3} = (1500 + 1800)/2 = 1650$? On can approximate that with $f(x_i)\Delta$ where Δ is the unit of measurement.

$$p_j = P(x_j - \frac{\Delta_j}{2} < x_j < x_j + \frac{\Delta_j}{2}) \approx f(x_j)\Delta_j$$

$$\approx \frac{1}{\Gamma(\phi_k^{-1})}\left(\phi_k\mu\right)^{-1/\phi_k} x_j^{\frac{1}{\phi_k}-1} \exp\left(\frac{-x_j}{\phi_k\mu}\right)\Delta_j$$

This gives for the likelihood:

$$P(D|\mu,\phi) \propto \prod_{j=1}^{10} x_j^{\frac{1}{\phi_k}-1} \exp\left(\frac{-x_j}{\phi_k\mu}\right)\Delta_j$$

$$\propto \exp\left(\frac{-\sum x_j}{\phi_k\mu}\right) \prod_{j=1}^{10} x_j^{\frac{1}{\phi_k}-1}\Delta_j$$

2.4 Not taking approximation but the CDF differences

$$P(D|\kappa,\lambda) \propto \left(\frac{1}{\Gamma(\kappa)}\right)^{\sum_{i=1} x_j}\gamma(x_1,\kappa,\lambda)^{x_1}\left[\prod_{j=2}^{9}\left(\gamma(x_j,\kappa,\lambda)-\gamma(x_{j-1},\kappa,\lambda)\right)^{x_j}\right]\left(1-\gamma(x_{10},\kappa,\lambda)\right)^{x_{10}}$$

$$\propto \left(\frac{1}{\Gamma(\kappa)}\right)^{\sum_{i=1} x_j}\left(\int_0^{\lambda x_1} x^{\kappa-1}e^{-x}\mathrm{d}x\right)^{x_1}\left[\prod_{j=2}^{9}\left(\int_{\lambda x_{j-1}}^{\lambda x_j} x^{\kappa-1}e^{-x}\mathrm{d}x\right)^{x_j}\right]\left(\int_{\lambda x_{10}}^{+\infty} x^{\kappa-1}e^{-x}\mathrm{d}x\right)^{x_{10}}$$

If we write $[0;x_1],[x_1;x_2],...,[x_9;x_{10}],[x_{10},+\infty]$ as $I_1,I_2,...,I_9$ and I_{10}. The notation can be lightened. With respect to the region k, this gives:

$$P(D_k|\kappa_k,\lambda_k) \propto \left(\frac{1}{\Gamma(\kappa)}\right)^{\sum_{i=1} x_j} \prod_{j=1}^{10}\left(\int_{\lambda_k I_j} x^{\kappa_k-1}e^{-x}\mathrm{d}x\right)^{x_{k,j}}$$

Writing in terms of μ_k and ϕ_k:

$$P(D_k|\mu_k,\phi_k) \propto \prod_{j=1}^{10}\left(\frac{1}{\Gamma(\phi^{-1})}\int_{\frac{I_j}{\phi_k\mu_k}} x^{(\phi_k^{-1}-1)}e^{-x}\mathrm{d}x\right)^{x_{k,j}}$$

2.5 NOT taking approximation but with PDF definitions:

3 Question 2 : Priors

- Statement 1: we are at 95% convinced that the mean net monthly household income in a given region is in the interval (2400, 3600). If one assumes a normal distribution for the mean μ_k (à justifier), then it is possible to get a prior of the distriubtion for both regions:

For both regions $\mu_0 = 3000$. Then, to get the standard deviation:

$$3000 - t_{(n_k-1,1-\alpha/2)} \frac{s_k}{\sqrt{n_k}} = 2400$$

$$\rightarrow \hat{\sigma}_0 = \frac{s_k}{\sqrt{n_k}} = \frac{600}{t_{(n_k-1,1-\alpha/2)}}$$

$\hat{\sigma}_{Fl} = 306$

So we have

$$\mu \sim N(\mu_0 = 3000, \sigma_0 = 306)$$

$$\pi(\mu) \propto \sigma_0^{-1/2} \exp\left(-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right)$$

- dispersion parameter:

If one considers that the parameter can be in any point within the interval (0.0, 10.0) with the same probability, then one could say that it follows a uniform distribution between those to interval

$$\phi_k \sim U(a = 0, b = 10) \propto 1_{0,10}$$

Using the entropy theorem, the conjugate prior would be the product of the two last quantity:

$$\text{prior:} P(\mu_k, \phi_k) = P(\mu_k) \; P(\phi_k)$$

$$\propto \sigma_0^{-1/2} \exp\left(-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right) 1_{0;10}$$

4 Question 3

4.1 Question 3a: posterior

4.1.1 With approximation

$\sigma^2 = \phi\mu^2$

$$P(\mu, \phi|D) \propto \exp\left(\frac{-\sum x_j}{\phi_k \mu}\right) \prod_{j=1}^{10} x_j^{\frac{1}{\phi_k}-1} \Delta_j \sigma^{-1/2} \exp\left(-\frac{1}{2\sigma^2}(\bar{x}_k - \mu)^2\right) 1_{0,10}$$

$$\propto \exp\left(\frac{-\sum x_j}{\phi_k \mu}\right) \prod_{j=1}^{10} x_j^{\frac{1}{\phi_k}-1} \Delta_j \left(\frac{1}{\sqrt{\phi}\mu}\right) \exp\left(-\frac{1}{2\phi\mu^2}(\bar{x}_k - \mu)^2\right) 1_{0,10}$$

4.1.2 Without approximation

$$P(\mu_k, \phi_k | D) \propto \left(\prod_{j=1}^{10} \left(\frac{1}{\Gamma(\phi^{-1})} \int_{\frac{I_j}{\phi_k \mu_k}} x^{(\phi_k^{-1}-1)} e^{-x} \mathrm{d}x \right)^{x_{k,j}} \right) \sigma_0^{-1/2} \exp\left(- \frac{1}{2\sigma_0^2} (\mu - \mu_0)^2 \right) 1_{0;10}$$

4.2 3.b

the log likelihood is, up to an additive constant:

$$l(\mu, \phi) \propto \sum_{j=1}^{10} x_j \ln \left(\frac{1}{\Gamma(\phi^{-1})} \int_{\frac{I_j}{\phi_k \mu_k}} x^{(\phi_k^{-1}-1)} e^{-x} \mathrm{d}x \right)$$

so the log-posterior is:

$$h(\mu, \phi) \propto l(\mu, \phi) + \ln(\pi(\mu, \phi))$$

$$h(\mu, \phi) = C^t + \sum_j x_j \ln \left(F(\mu, \phi, x_{j+1}) - F(\mu, \phi, x_j) \right) - \frac{1}{2\sigma_0^2} (\mu - \mu_0)^2$$

```
##      shape
## -1755.997
```

```
##      shape
## -956.3546
```

5 4.

5.1 Estimation

Start from the log-posterior $h(\mu_k, \phi_k)$. One assumes the distribution is unimodale then, if it is normally distributed, then the mode=mean, which means that the mean can be found by optimizing the log-likelihood with respect to its parameters. (aussi assumer μ et ϕ indépendant??), then blablabla (expliquez avec hessienne).
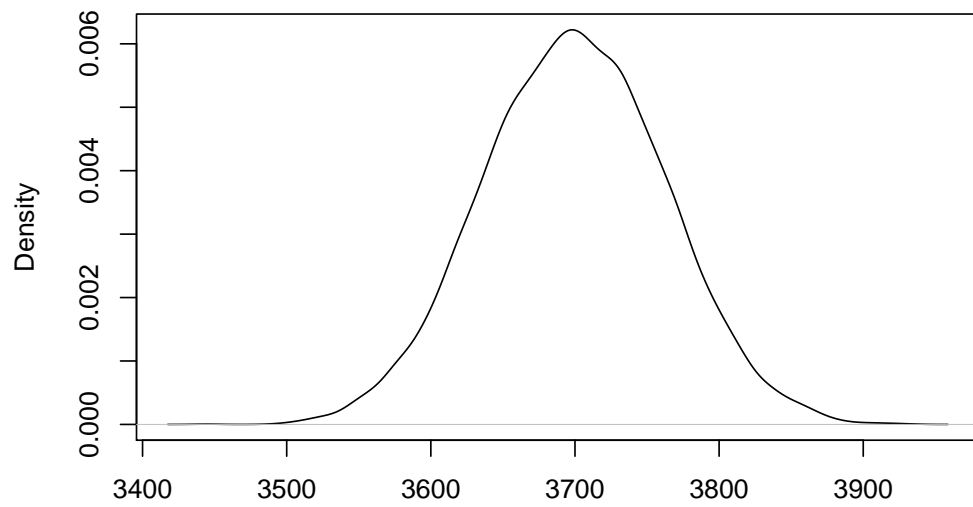
Based on http://www.sumsar.net/blog/2013/11/easy-laplace-approximation/#:~:text=Laplace%20App

5.1.1 Flanders
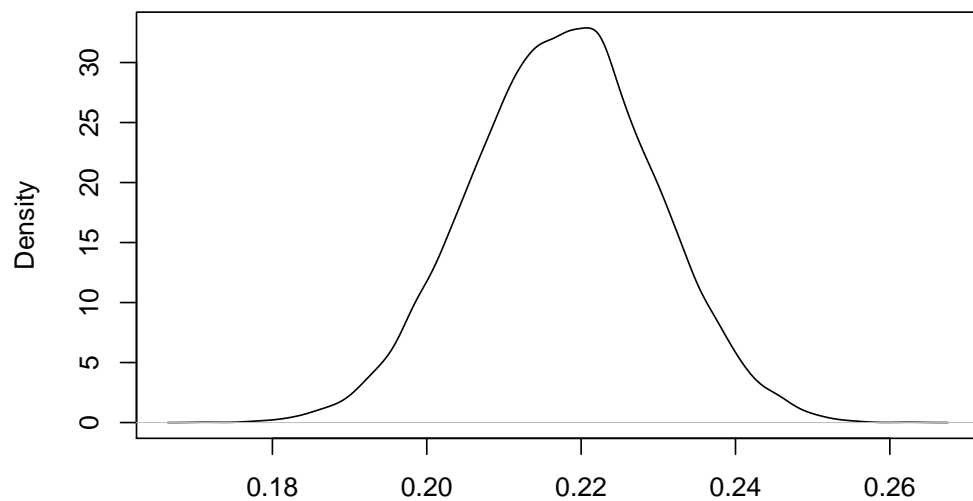
```
##       mu     phi
## 3698.91    0.22
```

```
##           mu  phi
## mu  3952.206 0.04
## phi    0.040 0.00
```

laplace approximation for mu in Flanders

N = 10000   Bandwidth = 8.965

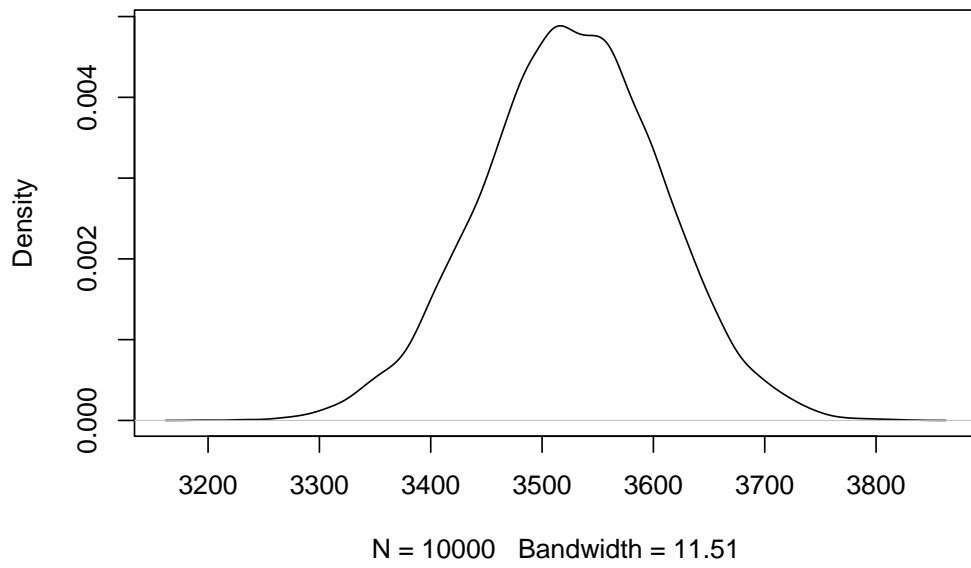**laplace approximation for phi in Flanders**
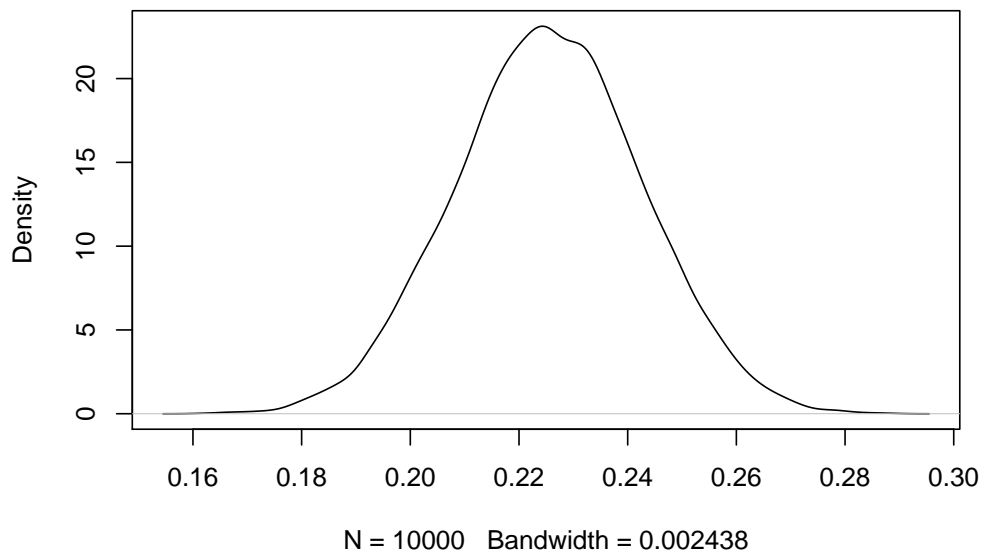

N = 10000   Bandwidth = 0.001686

5.1.2   Wallonia

```
##        mu       phi
## 3525.57     0.23
```

```
##             mu   phi
## mu   6639.722 0.038
## phi     0.038 0.000
```
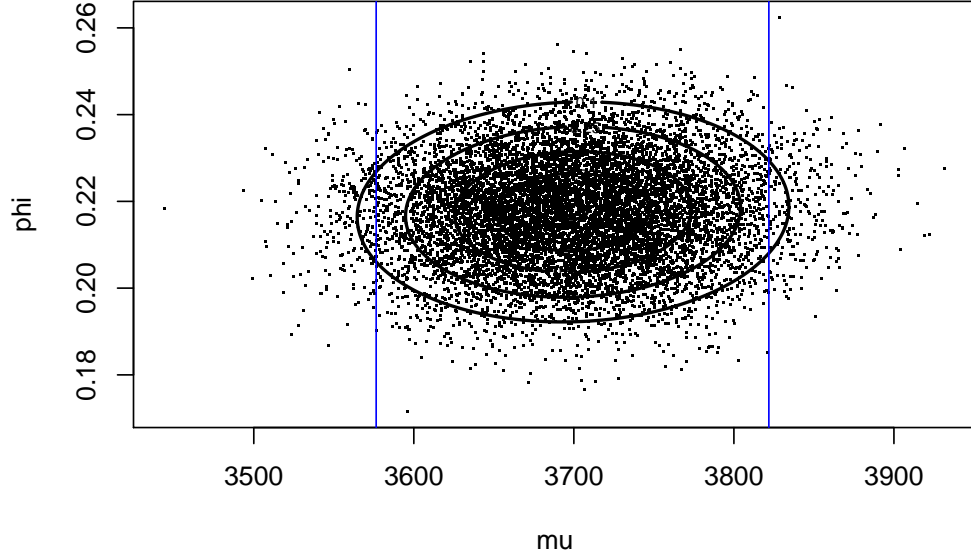
8

laplace approximation for mu in Wallonia

Density

0.004

0.002

0.000

3200 3300 3400 3500 3600 3700 3800

N = 10000   Bandwidth = 11.51

laplace approximation for phi in Wallonia

Density

20

15

10

5

0

0.16   0.18   0.20   0.22   0.24   0.26   0.28   0.30

N = 10000   Bandwidth = 0.002438

5.2   Credible intervals Laplce approximation

5.2.1   Flanders

The credible region is given here below:

We needs to get the credible interval for μ. This means that we needs the marginal posterior distribution:

$$P(\mu|D) \propto \int p(\mu, \phi|D)\mathrm{d}\phi$$

It can be shown that the marginal (univariate) distribution of the bivariate Gaussian distribution $N(\mu_\theta, \Sigma)$ with $\mu_\theta = (E(\mu), E(\phi))$ and

$$\Sigma = \begin{pmatrix} \sigma_\mu & \sigma_{\mu,\phi} \\ \sigma_{\phi,\mu} & \sigma_\phi \end{pmatrix}$$

also follows a normal distribution. See here. Hence, for μ, one has:

$$\mu|D \sim N(\mu, \sigma_\mu)$$

We needs that 95% of the marginal posterior to fall into the interval for Flanders:

```
##        5%     97.5%
## 3593.351 3823.121
```

A   Appendix

A.1   Figures

A.2   Code

> **Note**
>
> For reproducibility purposes, the complete R project containing the source code and the results is available on github.com.