Bayesian Notes Lulu

Romà Domènech Masana A20450272

March 2020

Contents

1 Hierarchical models

What is the difference between the prior distribution of θ being

$$\theta \sim \text{Beta}(\alpha, \beta) \tag{1}$$

and the parameter θ coming from a hyper distribution

$$\text{Beta}(\alpha, \beta) \tag{2}$$

? Is it the same thing?

I mean in the first case we are saying that

$$p(\theta) \sim \text{Beta}(\alpha, \beta) \tag{3}$$

and in the second case that

$$p(\theta|\alpha,\beta) \sim \text{Beta}(\alpha,\beta), \tag{4}$$

I don't understand the philosohpical difference between one and the other.
We could understand it in the following way: each farm has a pair α, β which gives a different prior $p(\theta)$ for each farm j. So this is a way of assuming a *local* prior $p(\theta)$ instead of the global, equal for all farms prior we used up to now.

2 Exam

Chapters 1-5

3 Conjugate distributions

What is normal-inverse-χ^2
Page 67

4 Prior distributions

Jeffrey's prior 2.8 page 53 is

$$p_J(\theta) = \sqrt{\det I(\theta)} \tag{5}$$

Page 53 formula 2.19 we have that if $\phi = h(\theta)$ for a parameter θ that has prior

$$p(\theta),$$

then

$$p(\phi) = p(\theta)\left|\frac{d\theta}{d\phi}\right| \tag{6}$$

5 multivariate models parameter estimation

The conujugate distribution for the univariate normal with unknown mean and variance is the normal-inverse-χ^2.

5.1 Drawing samples from the posterior bivariate distribution

Let's assume that we have two parameters $\mu,^2$ that we want to estimate. In particular, we want to know:

$$p(\mu,\sigma^2|y) \propto p(\mu,\sigma^2)p(y|\mu,\sigma^2). \tag{7}$$

But the difficulty here is, even if we know the posterior distribution in terms of the two **unknown** variables μ, σ^2, how do we sample observations from here? Being able to sample observations would be an easy way to compute the mean and standard deviation.

In 1D. From a pdf we can easily compute the CDF, and the, draw samples from the unif$[0,1]$ and then get the resulting sample from our 1D posterior distribution. But in 2D (or when we have 2 parameters to estimate) we can still compute a 2D CDF from the 2D PDF. But sampling observations from the unif$[0,1]$, leads to contour lines in the parameter space. Not points. However, there is a way around this, using marginal distributions:

We have that

$$p(\mu,\sigma^2|y) = p(\mu|\sigma^2,y)p(\sigma^2|y), \tag{8}$$

Then we can just compute first the posterior marginal for σ^2 and then, for each value of the σ^2, compute 1 value of the conditional posterior of μ. To compute a value of the marginal $p(\sigma^2|y)$ all we have to do is get a random value from the unif$[0,1]$ and then using the cdf of the marginal of σ^2, compute its sampled value.

6 Distributions

6.1 scaled inverse Chi squared

If a random variable $X \sim$ scale-inv $- \chi^2_{\nu,\tau^2}(x)$ then its pdf is

$$f_{X_{\nu,\tau^2}}(x) \propto e^{\frac{\nu\tau^2}{2x}} \frac{1}{x^{\nu/2+1}} \tag{9}$$

In particular in page 65 of *Bayesian data Analysis by Carlin et al.*

7 Lecture 26th March 2020

7.1 Ingorability

Chapter 8.

I is the **inclusing vector**.

She does something like the expectation maximization algorithm, but different from it:

$$p(y_{obs}, I|\theta, \phi) = \int p(y, I|\theta, \phi) dy_{missing}$$

Objective: $p(\theta, \phi)$. To get rid of ϕ we just integrage over ϕ like

$$p(\theta|x, y_{obs}, I) = \int p(\theta, \phi|x, y_{obs}, I) d\phi$$

One of the assumptions is that θ does not dpend on ϕ. If you had that θ depends on ϕ then lulu says *this is not reasonable.*

She wants to estimate the missing values y_{mis} by simulating θ, ϕ from the posterior distribution and then... I got lost. chapter 8

SHe says that $p(\theta|x, y_{obs}) = p(\theta|x, y_{obs}, I)$ if you truly believe ignorability. **multiple imputations**. Sif this happens you don't have to include the data process.

<u>when can we assume ignorability?</u>

1. missing at random, i.e. if
$$p(I|x, y\phi) = p(I|x, y_{obs}\phi)$$
this holds for instance if I is independent of hthe y themselves, or if example you ask people's income, that's why they shouldn¡t ask it.

2.

strongly ingorable if $p(I|..) doens'tdependon$y.
<u>Non-ignorable data</u>

1. Censored data, like you cannot know information of ill people or dead people.

2. patients choose something

7.2 Sample surveys

Simple random sampling. We collect info from them. I choose small n from the total population N. total number of different sets

$$\binom{N}{n}$$

We want to estimate

$$\bar{y} = \frac{n}{N}\bar{y}_{obs} + \frac{N-n}{n}\bar{y}_{mis}$$

. Inference on y_{mis} based on posterior predictive distribution.

If we have ignorability we can do

$$p(\theta|y_{obs}, I) = p(\theta|y_{obs})$$

. Sampling

$$\theta^l \sim p(\theta|y_obs)$$

.

If $N-n$ is large, we can approximate the pdf onf the missing values

$$p(\bar{y}_{mis}|\theta) \eqsim \mathcal{N}(\bar{y}_{mis}|\mu, \frac{\sigma^2}{N-n})$$

$$\mu = \mathbb{E}(y_i|\theta), \sigma^2 = \mathrm{Var}(y_i|\theta).$$

She does the CLT approximation when both n and $N-n$ are large for the pdf of the missing and non-missing observations.

8 Lecture 31.03.2020

8.1 Stratified sampling

Variance of y_{mis} is smaller if we stratify our sample.

The ratio N_j/N is very big. In general it means "I am going to separate the population in different stratum and then I am going to sample from each of the stratum". So the ratio of sample of each stratum whould keep the real data ratio. The population proportions N_j/N must be the same as the sample ratio n_j/n.

Models. No hierarchical structure.

$$y_{obs,j} = (y_{1j}, y_{2j}, y_{3j})$$

.

$$y_j \sim \mathrm{multin}(n, \theta_{1j}, \theta_{2j}, \theta_{3j}) \tag{10}$$

The prior should follow a conjugate **Dirichlet.**
Posterior wil be $p(\theta_j|y)$
To estimate the θ_j you just need the data y_j and I don't need the rest of the data.

Ojjective: $\bar{y}_1 - \bar{y}_2 \simeq \sum_{j=1}^{J} \frac{N_j}{N}(\theta_{1j} - \theta_{2j})$

Then we do this for each stratum. A best way of doing this is using a *hierarchical model.*

8.1.1 Hirearchical model

Assume

$$\alpha_{1j} = \frac{\theta_{1j}}{\theta_{1j}+\theta_{1j}} \text{ probability of preferrin Bush, given that} \tag{11}$$

is the probability of preferring Bush. We have 16 stratum. $\alpha_{2j} = 1 - \theta_{3j}$ probability of expressing preference.

This prior is informative.

8.1.2 Cluster sampling

Separate the N unites in the population innto K clusters. A sample of J clusters is drawn. And then sample n_j units from the N_j population within each sampled cluster $j = 1, ..., J$.

Propensity scores.

$$\begin{array}{c} N\text{weighting of an object} \\ y_j j\text{th weigtht} \end{array} \tag{12}$$

First we talk about **missing at random**.

(a) **missing completely at random**

$$\begin{array}{rl} p(\theta|y_{obs,I}) & = p(\theta|y_{obs}) \\ & = p(\theta)p(y_{obs}|\theta) \\ & =\propto p(y_{obs}|\theta) \\ & = \prod_1^9 1\mathcal{N}(y_i|) \end{array} \tag{13}$$

$$I_i \sim (\pi), i \in [0, 1]$$

is unknown and independent of θ. Complicated case when π is a functio of θ. We assume :

$$\begin{array}{c} \pi = \frac{\theta}{\theta+1} \\ \theta > 0 \\ \downarrow \\ \theta = \frac{\pi}{1-\pi} \end{array} \tag{14}$$

So we get that

$$\begin{array}{c} p(\theta, \pi|y_{obs}, I) \propto ... \propto \mathcal{N}(\theta|\bar{y}_{obs}, 1/91) Bin(n = 91|N = 100, \pi) \\ \propto xp\,() \end{array} \tag{15}$$

where we assumed a non-informative prior $p(\theta, \pi) \propto 1$.

(b) **Censored data** y_i is missing iff y_i is greater than 200 ($y_i > 200$)

$$\begin{array}{rl} p(\theta, \pi|y_{obs}, I) & \propto p(\theta)p(y_{obs}, I|\theta) \\ & \propto p(y_{obs}, I|\theta) = \int p(y, I|\theta)dy_{miss} \\ & \propto \mathcal{N}(\theta|\bar{y}_{obs}, 1/91)[\Phi(\theta - 200)]^9 \end{array} \tag{16}$$

The missing values integration:

$$\begin{array}{r} \int p(y, I|\theta)dy_{miss} \propto \int \prod_1^9 1\mathcal{N}(y_{obs}|\theta, 1) \int \prod_1^9 1\mathcal{N}(y_{obs}|\theta, 1) \prod_1^9 \mathcal{N}(y_{miss}|\theta, 1) \\ \prod_1^9 1\mathcal{N}(y_{obs}|\theta, 1) \int \prod_1^9 \mathcal{N}(y_{miss}|\theta, 1) \\ \prod_1^9 1\mathcal{N}(y_{obs}|\theta, 1)[\Phi(\theta - 200)]^9 \end{array} \tag{17}$$

We get the same if we censor from ϕ upwards, We must take into account that

$$\phi \geq \max_j y_{obs,j}$$

Then we can see that we would get

$$p(\theta, \pi | y_{obs}, I) \quad \propto \mathcal{N}(\theta | \bar{y}_{obs}, 1/91)[\Phi(\theta - \phi)]^9 \tag{18}$$

Marginal posterior if we assume $p(N|\theta) \sim 1/N$

$$p(\theta | y_{obs}, I) \quad \propto \sum_9 1^\infty p(N|\theta) \tag{19}$$

(c) **truncated data with unknown truncation point:** They are all different missing data schemes

Important parts for Chapter 8 are sections: 8.2, 8.3 and 8.7.

9 Chapter 10

We will do Markov Chain simulation. Sample size calculation. Assume we know how to sample

$$\theta_i \sim p(\theta|y).$$

We want to get sample deviation.. We can estimate the posterior mean.
According the CLT.

$$\sqrt{N}(\bar{\theta} - \theta_0) \to \mathcal{N}(0, \sigma^2) \tag{20}$$

with standard deviation

$$std(\bar{\theta}) = \sqrt{\mathrm{Var}(\theta)} \approx \frac{\sigma}{\sqrt{N}}$$

$\tilde{\theta}_{m,N} - \theta_m$ **What si the asymptotic distribution of the sample median**?

$$\sqrt{N}(\tilde{\theta}_{m,N} - \theta_m) \to \mathcal{N}(0, \tfrac{1}{4f(0)^2}) \tag{21}$$

10 Exercises

10.1 Problem 5 (midterm exam) Inference about a normal population

We have the following sleeping hours of 20 students:

```
> y
 [1]  9.0  8.5  7.0  8.5  6.0 12.5  6.0  9.0  8.5
[10]  7.5  8.0  6.0  9.0  8.0  7.0 10.0  9.0  7.5
[19]  5.0  6.5
```

Figure 1: Sample data

Now we have that using the noninformative prior

$$p(\mu, \log) \propto 1,$$

or equivalently,

$$p(\mu, \sigma^2) = p(\mu, \log \sigma) \begin{vmatrix} 1 & 0 \\ 0 & \frac{1}{2\sigma^2} \end{vmatrix} \propto \tfrac{1}{\sigma^2}. \tag{22}$$

where $h_1(\mu, \sigma^2) = \mu, h_2(\mu, \sigma^2) = 1/2 \log(\sigma^2)$. This will be the prior we will use. We will use the data distribution:

$$y \sim \mathcal{N}(\mu, \sigma^2).$$

Finally, to be able to sample from the posterior bivariate distribution $p(\mu, \sigma^2|y)$, we will use the decomposition

$$p(\mu, \sigma^2|y) = p(\mu|\sigma^2, y)p(\sigma^2|y), \tag{23}$$

which will allow us to first sample a value of σ^2 using the marginal distribution (and 1 dimensional) of σ^2, and then use this sampled value together with the data to get the corresponding sampled value of μ using the conditional distribution of μ given σ^2 and the data. So what we will do is try to find the distributions $p(\mu|\sigma^2, y)$ and $p(\sigma^2|y)$, that satisfy Equation 22. Let's get started, we have that:

$$\begin{aligned} p(\mu, \sigma^2|y) &\propto p(\mu, \sigma^2)p(y|\mu, \sigma^2) \\ &\propto \frac{1}{\sigma^2}\frac{1}{\sigma^n}e^{\frac{-1}{2\sigma^2}\sum_1^n (y_i-\mu)^2} \\ &\propto \frac{1}{\sigma/\sqrt{n}}e^{\frac{-1}{2\sigma^2}(\mu-\bar{y})^2}\frac{1}{\sigma^{n+1}}e^{\frac{-1}{\sigma^2}(n-1)S_n^2} \\ &= p(\mu|\sigma^2, y)p(\sigma^2|y), \end{aligned} \tag{24}$$

where

$$\begin{aligned} p(\mu|\sigma^2, y) &= \frac{1}{\sigma/\sqrt{n}}e^{\frac{-1}{2\sigma^2}(\mu-\bar{y})^2} \sim \mathcal{N}(\bar{y}, \sigma^2/n), \\ p(\sigma^2|y) &= \frac{1}{\sigma^{n+1}}e^{\frac{-1}{\sigma^2}(n-1)S_n^2} \sim \text{scaled-inv}\chi^2(\nu = n-1, \tau^2 = S_n^2). \end{aligned} \tag{25}$$

So now we have all the ingredients to do the whole problem. Let's compute a sample from the posterior first:

(a) We draw first 1000 samples of σ^2 from the scaled inverse χ^2_{n-1,S_n^2} that we just deduced:

We get the sampling distribution that can be found at Figure 2.

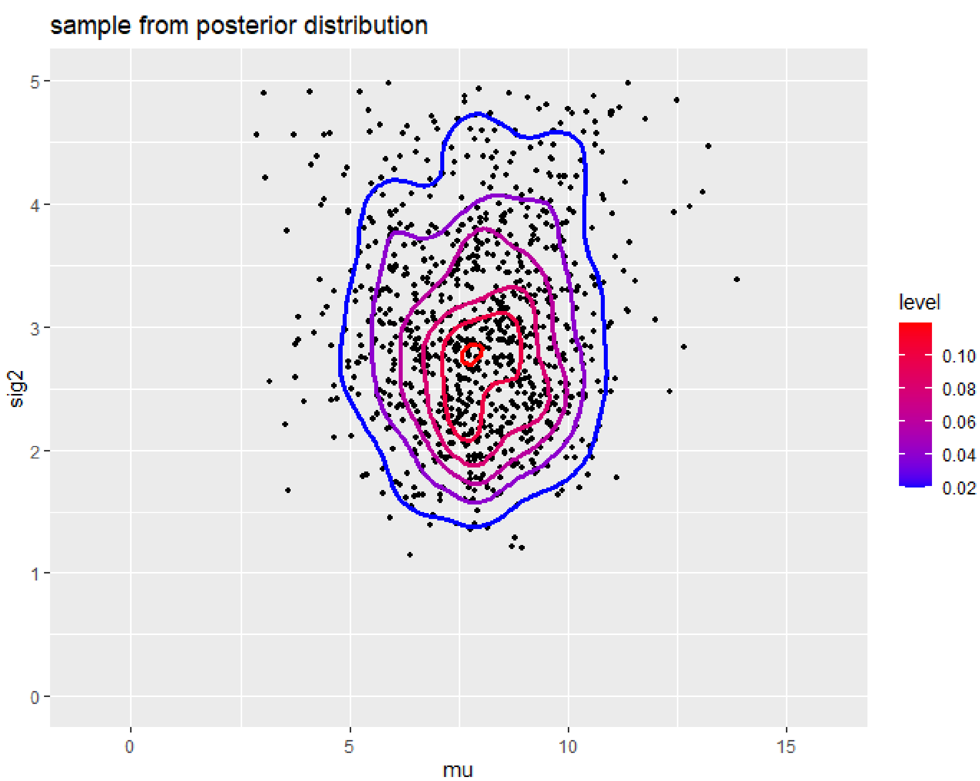

Figure 2: Posterior sample from the two parameters. Each contour level is 0.02 units apart.

(b) Compute 95% confidence intervals for μ and for σ. We can use the sample percentiles from the marginal sample. That is, in order to compute the sample conficence interval of μ, we take these 1000 values and forget about what the corresponding values of σ^2 are. Thus we get using R the results displayed on Figure 3. This gives us the 90% confidence intervals:

$$\begin{aligned} \Pr(\mu \in [5.03, 10.58]) &= 0.9 \\ \Pr(\sigma \in [1.34, 2.31]) &= 0.9 \end{aligned} \tag{26}$$

```
> quantile(mu,c(0.05,0.95))
      5%      95%
 5.02851 10.58366
> quantile(sqrt(sig),c(0.05,0.95))
      5%      95%
1.342180 2.310436
```

Figure 3: Posterior quantiles of μ and σ.

(c) Now we are asked to estimate the mean and variance of

$$p_{0.75} = \mu + 0.674\sigma. \tag{27}$$

The way I would do it is using the sampled join parameters to compute

$$\begin{aligned}
\mathbb{E}p_{0.75} &= \mathbb{E}\mu + 0.674\mathbb{E}\sigma \\
&= 7.92 + 0.674 * 1.77 \\
&\simeq 9.11 \\
\operatorname{Var} p_{0.75} &= \operatorname{Var}\mu + (0.674)^2 \operatorname{Var}\sigma - 2 * 0.674 \operatorname{Cov}(\mu,\sigma) \\
&= 3.17 + (0.674)^2 * 0.094 - 2 * 0.674 * (-0.004) \\
&= 3.24 \\
&\downarrow \\
\text{Std } p_{0.75} &\simeq 1.8
\end{aligned} \tag{28}$$

So the upper quartile $p_{0.75} \simeq 9.11 \pm 1.8$. This doesn't seem to contradict the initial 20 data samples.

```
> mean(mu)
[1] 7.915616
> mean(sqrt(sig))
[1] 1.765822
> var(mu)
[1] 3.173814
> var(sqrt(sig))
[1] 0.09430578
> mean(mu)+0.674*mean(sqrt(sig))
[1] 9.10578
> cov(mu,sqrt(sig))
[1] -0.003928852
```

Figure 4: Posterior mean and variance of μ, σ.

10.2 Exercise 6