Computational Statistics-Report

David Niederkofler, Erlend Lokna

2022-12-06

```
mydata<-read.table("Report2_Dataset.txt", header=FALSE)
```

Statistical Analysis of Covariates

It is important to mention the use of notation before we proceed. We will in this section use the notation $\mathbf{x} = \{x_1, x_2, \cdots, x_n\}$ as our covariate vector.

Ascicles

1.1 Model selection

Since the Ascicles - covariate has a 0-1 outcome we can assume that it is Bernoulli distributed with parameter θ. A natural conjugate prior for the Bernoulli distribution is the Beta distribution. The posterior beta distribution for the parameter is given by

$$Beta(\theta|a + \sum_{i=1}^{n} x_i, b + n - \sum_{i+1}^{n} x_i)$$

1.2 Results

The following results where found using the posterior beta distribution with a=1 and b=1 (Uniform distributed) for the ascicles data:

```
## Posterior mean:  0.08227848
```

```
## Posterior mode:  0.07961783
```

```
## Centered 95% Confidence Interval: [ 0.05235453 , 0.1119428 ]
```

With the following HPD interval:

```
##      lower      upper
## 0.05074464 0.10984480
## attr(,"credMass")
## [1] 0.95
```

1. Sex

The sex of the patients is encoded in a binary variable, where 0 means *male* and 1 means *female*.

1.1 Model selection

We assume a Bernoulli model $Ber(\theta)$ for the sex of the patient conditional on one parameter θ, the probability of the patient to be female. The density function is given by

$$f(x|\theta) = \theta^x (1 - \theta)^{1-x}, \tag{1}$$

where $x \in \{0,1\}$. As a prior distribution for θ we use the natural conjugate family of the Bernoulli distribution, namely the Beta distribution, $Beta(a,b)$, with two shape parameters $a=b=2$ to give more weight to the middle of the interval $[0,1]$, knowing how females and males are represented in the general population. The density is given by

$$h(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}, \tag{2}$$

for $\theta \in [0,1]$.

1.2 Results

From the given dataset we get the sample size n and the sum of the observations s:

```
n<-length(mydata$V6[!is.na(mydata$V6)])
s<-sum(mydata$V6)
n
```

```
## [1] 312
```

```
s
```

```
## [1] 276
```

Therefore the posterior distribution is $Beta(2+s, 2+n-s)$, which turns out to be $Beta(278, 38)$. From that we get

```
## Posterior mean:  0.8797468
```

```
## Posterior mode:  0.8821656
```

```
## Centered 95% Confidence Interval: [ 0.8417454 , 0.9132003 ]
```

And the HPD confidence Interval calculates to:

```
tst<-rbeta(1e5,278,38)
hdi(tst)
```

```
##     lower     upper
## 0.8433744 0.9145678
## attr(,"credMass")
## [1] 0.95
```

2. Spiders

The presence of spiders is encoded in a Binary variable, where 1 means spiders are present.

1.1 Model selection

We assume a Bernoulli model $Ber(\theta)$ for the presence of spiders in patients conditional on one parameter θ, the probability of the presence of spiders in the patient. The density function is given as stated earlier. As a prior distribution for θ we use the natural conjugate family of the Bernoulli distribution, namely the Beta distribution, $Beta(a,b)$, with two shape parameters $a=b=1$, because we have no prior information. The density is given as above.

1.2 Results

From the given dataset we get the sample size n and the sum of the observations s:

```
n<-length(mydata$V9[!is.na(mydata$V9)])
s<-sum(mydata$V9)
n
```

```
## [1] 312
```

```
s
```

```
## [1] 90
```

Therefore the posterior distribution is $Beta(1+s, 1+n-s)$, which turns out to be $Beta(91, 223)$. From that we get

```
## Posterior mean:  0.2911392
```

```
## Posterior mode:  0.2898089
```

```
## Centered 95% Confidence Interval: [ 0.2410228 , 0.341131 ]
```

And the HPD confidence interval calculates to:

```
tst<-rbeta(1e5,91,223)
hdi(tst)
```

```
##     lower     upper
## 0.2392598 0.3393446
## attr(,"credMass")
## [1] 0.95
```

3. Hepatomegaly

The presence of hepatomegaly is encoded in a Binary variable, where 1 means hepatomegaly is present.

1.1 Model selection

We assume a Bernoulli model $Ber(\theta)$ for the presence of hepatomegaly in the patient, conditional on one parameter θ, the probability of the presence of hepatomegaly in the patient. The density function is given as stated earlier. As a prior distribution for θ we use the natural conjugate family of the Bernoulli distribution, namely the Beta distribution, $Beta(a, b)$, with two shape parameters $a = b = 1$, because we have no prior information. The density is given as above.

1.2 Results

From the given dataset we get the sample size n and the sum of the observations s:

```
n<-length(mydata$V8[!is.na(mydata$V8)])
s<-sum(mydata$V8)
n
```

```
## [1] 312
```

```
s
```

```
## [1] 160
```

Therefore the posterior distribution is $Beta(1+s, 1+n-s)$, which turns out to be $Beta(161, 153)$. From that we get

```
## Posterior mean:  0.5126582
```

```
## Posterior mode:  0.5127389
```

```
## Centered 95% Confidence Interval: [ 0.4575015 , 0.5678225 ]
```

And the HPD confidence interval calculates to:

```
tst<-rbeta(1e5,161,153)
hdi(tst)
```

```
##      lower     upper
## 0.4572249 0.5666820
## attr(,"credMass")
## [1] 0.95
```

4. Histologic stage

The Histologic stage of the disease is a number in $\{1, 2, 3, 4\}$, where the stage increases with severeness. We will give here the frequencies of the stages in the dataset.

```
##   1   2   3   4
##  16  67 120 109
```

We see that, most patients have been diagnosed in the last to stages of the disease.

5. Age

The age of the patient in days.

5.1 Model selection

The data seems to follow a poisson distribution $Poi(\lambda)$. Using the non informative Jeffreys prior, we can derive that the posterior for the parameter λ is Gamma distributed.

$$\theta | x \sim Gamma(\alpha = \frac{1}{2} + \sum_{i=1}^{n} x_i, \beta = n)$$

5.2 Results

$$\theta | x \sim Gamma(\frac{1}{2} + s, n)$$

```
## posterior distribution: Gamma( 5700066 , 312 )
## mean: 18269.44
## variance: 58.55591
## HPD intervall:
```

```
##    lower    upper
## 18254.75 18284.78
## attr(,"credMass")
## [1] 0.95
```

6. Cholesterol

6.1 Model selection

We assume that the data is sampled from a poisson, $Poi(\lambda)$, distribution, and we use the non informative Jeffreys prior for the rate parameter in the bayesian analysis.

$$\mathbf{x} \sim Poi(\lambda)$$

$$h(\lambda) \propto \lambda^{-\frac{1}{2}}$$

6.2 Results

```
## posterior distribution: Gamma( 104941.5 , 312 )
## mean: 336.351
## variance: 1.078048
## HPD intervall:
```

```
##    lower    upper
## 334.3068 338.3873
## attr(,"credMass")
## [1] 0.95
```

7. Urine

7.1 Model selection

$$\mathbf{x} \sim Poi(\lambda)$$
$$h(\lambda) \propto \lambda^{-\frac{1}{2}}$$

7.2 Results

```
## posterior distribution: Gamma( 30271.5 , 312 )
## mean: 97.02404
## variance: 0.3109745
## HPD intervall:
```

```
##    lower    upper
## 95.93102 98.11615
## attr(,"credMass")
## [1] 0.95
```

8 SGOT

8.1 Model selection

$$\mathbf{x} \sim Poi(\lambda)$$
$$h(\lambda) \propto \lambda^{-\frac{1}{2}}$$

8.2 Results

```
poisson_jeffrey(mydata$V16)
```

```
## posterior distribution: Gamma( 38238.08 , 312 )
## mean: 122.5579
## variance: 0.3928139
## HPD intervall:
```

```
##    lower    upper
## 121.3463 123.8056
## attr(,"credMass")
## [1] 0.95
```

9. Plateles

9.1 Model selection

$$\mathbf{x} \sim Poi(\lambda)$$
$$h(\lambda) \propto \lambda^{-\frac{1}{2}}$$

9.2 Results

```
poisson_jeffrey(mydata$V18)
```

```
## posterior distribution: Gamma( 80676.5 , 312 )
## mean: 258.5785
## variance: 0.8287773
## HPD intervall:

##    lower    upper
## 256.8205 260.3943
## attr(,"credMass")
## [1] 0.95
```

10. Prothrombin

10.1 Model selection

$$\mathbf{x} \sim Poi(\lambda)$$
$$h(\lambda) \propto \lambda^{-\frac{1}{2}}$$

10.2 Results

```
poisson_jeffrey(mydata$V19)
```

```
## posterior distribution: Gamma( 3346.9 , 312 )
## mean: 10.72724
## variance: 0.03438219
## HPD intervall:

##    lower    upper
## 10.36086 11.08776
## attr(,"credMass")
## [1] 0.95
```

Appendix

Bernoulli/Beta

A natural conjugate prior for the Bernoulli distribution is the Beta distribution.

$$f(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}$$
$$L(\mathbf{x}|\theta) = \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i}$$
$$h(\theta) = Beta(a,b)$$

We proceed by calculating the posterior distribution for θ

$$h(\theta|\mathbf{x}) \propto L(\mathbf{x}|\theta)h(\theta) = \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i}\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}I(0<\theta<1)$$

$$\propto Beta(\theta|a + \sum_{i=1}^{n} x_i, b + n - \sum_{i+1}^{n} x_i)$$

Poisson/Gamma

If our data $X_1, \cdot, X_n$ are iid Poisson(λ) distributed then a gamma(α, β) prior on λ is a conjugate prior. The Likelyhood function is:

$$L(\lambda|\mathbf{x}) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{\sum_{i=1}^{n} x_i}}{x_i!} = \frac{e^{-\lambda}\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}$$

Our gamma prior has the expression:

$$h(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$$

Using bayes rule we find the following posterior:

$$h(\lambda|\mathbf{x}) \propto h(\lambda)L(\mathbf{x}|\lambda) \propto \lambda^{\sum_{i=1}^{n} x_i+\alpha-1}e^{-(n+\beta)\lambda}$$

$$\propto gamma(\sum_{i=1}^{n} x_i + \alpha, n + \beta)$$

Poisson/Jeffreys prior

The density distribution for poisson is equal to

$$f(n|\lambda) = e^{-\lambda}\frac{\lambda^n}{n!}$$

The jeffreys prior $h(\theta$ is a non informative prior distrubution for a parameter space and its proportionality is expressed as

$$h(\theta) \propto \sqrt{detI(\theta)}$$

$$I(\theta) = -E[\frac{\partial^2}{\partial\theta^2}logf(x|\theta)] = \frac{1}{\theta}$$

And the following jeffreys prior is thus

$$h(\theta) \propto \theta^{-\frac{1}{2}}I_{\theta>0}$$

The posterior is calculated as follows

$$h(\theta|x) \propto f(\mathbf{x}|\theta)h(\theta) \propto e^{-n\theta}\theta^{-\frac{1}{2}+\sum_{i=1}^{n} x_i}$$

which is in fact a gamma distribution

$$\theta|x \sim Gamma(\alpha = \frac{1}{2} + \sum_{i=1}^{n} x_i, \beta = n)$$