Computational Statistics-Report

David Niederkofler, Erlend Lokna

2022-12-01

```r
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.

Bernoulli - Beta, Ascicles

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. Therefore we have the following initial information:

$$f(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}$$

$$L(\mathbf{x}|\theta) = \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i}$$

$$h(\theta) = Beta(a,b)$$

We proceed by calculating the posterior distribution for θ

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

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

1.1 Model selection

1.2 Results

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
## 0.8439134 0.9152898
## 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
## 0.2403824 0.3405487
## 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
## 0.4584070 0.5687331
## 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.