Computational Statistics-Report

David Niederkofler, Erlend Lokna

2022-12-08

```
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
## 0.0506336 0.1095727
## 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
## 0.8426204 0.9140323
## attr(,"credMass")
## [1] 0.95
```

2. Spiders

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

2.1 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.

2.2 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
## 0.2409424 0.3409109
## attr(,"credMass")
## [1] 0.95
```

3. Hepatomegaly

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

3.1 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.

3.2 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
## 0.4565740 0.5669735
## 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( 5700067 , 312 )
## mean: 18269.44
## variance: 58.55591
## HPD intervall:
```

```
##    lower    upper
## 18254.30 18284.37
## 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
## 334.3112 338.3687
## 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
## 95.94962 98.13992
## 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
## 121.3487 123.8090
## 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
## 256.8031 260.3536
## 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:
```

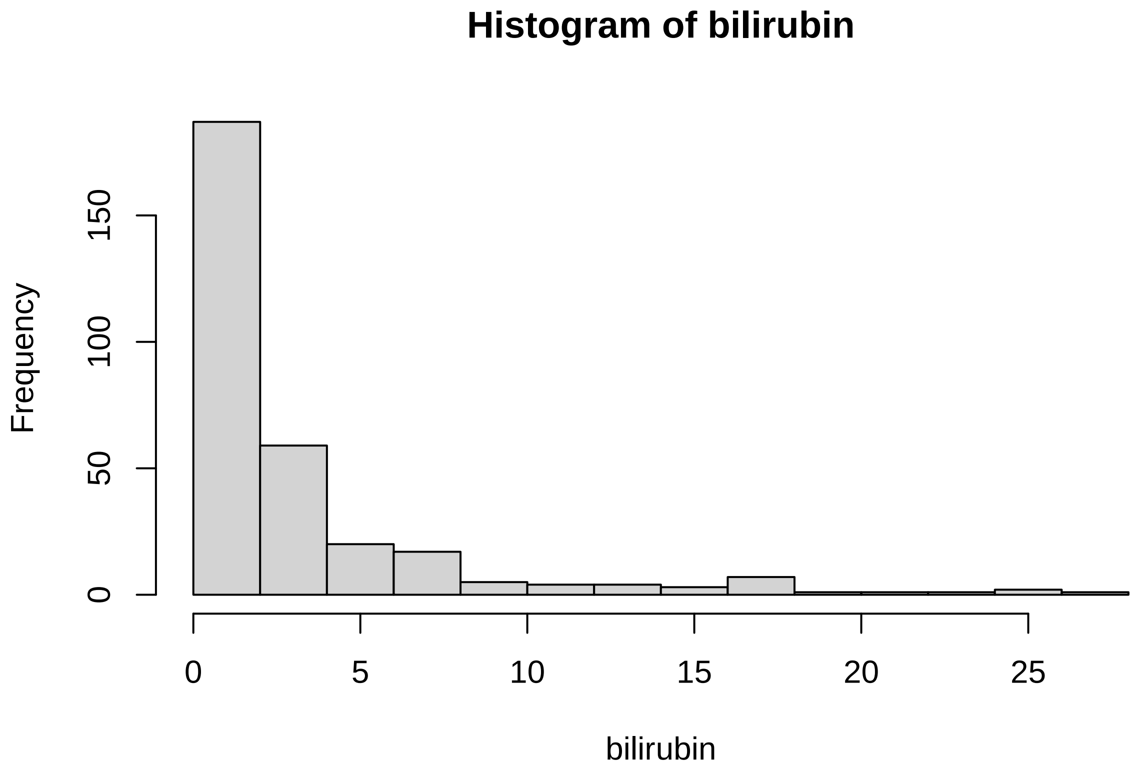
```
##    lower    upper
## 10.35539 11.08317
## attr(,"credMass")
## [1] 0.95
```

11. Bilirubin

The serum bilirubin of the patients is given in mg/dl.

11.1 Model Selection

We assume by inspecting the histogramm plot,



that the data follows a exponential distribution with parameter λ. Density is given by

$$f(x|\lambda) = \lambda e^{-\lambda x} \tag{3}$$

As a prior for λ we use, the jeffreys non-informative prior, namely: $h(\lambda) \propto \frac{1}{\lambda}$.

11.2 Results

From the data we get the number of samples n and the sum of the samples s as

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

```
## [1] 1015.9
```

That means the posterior distribution for λ is $Gamma(n, s)$. Which turns out to be $Gamma(312, 1015.9)$. From that we get

```
## Posterior mean:  0.3071168
```

```
## Posterior mode:  0.3061325
```

```
## Centered 95% Confidence Interval: [ 0.2739805 , 0.3421174 ]
```

And the HPD confidence interval calculates to:

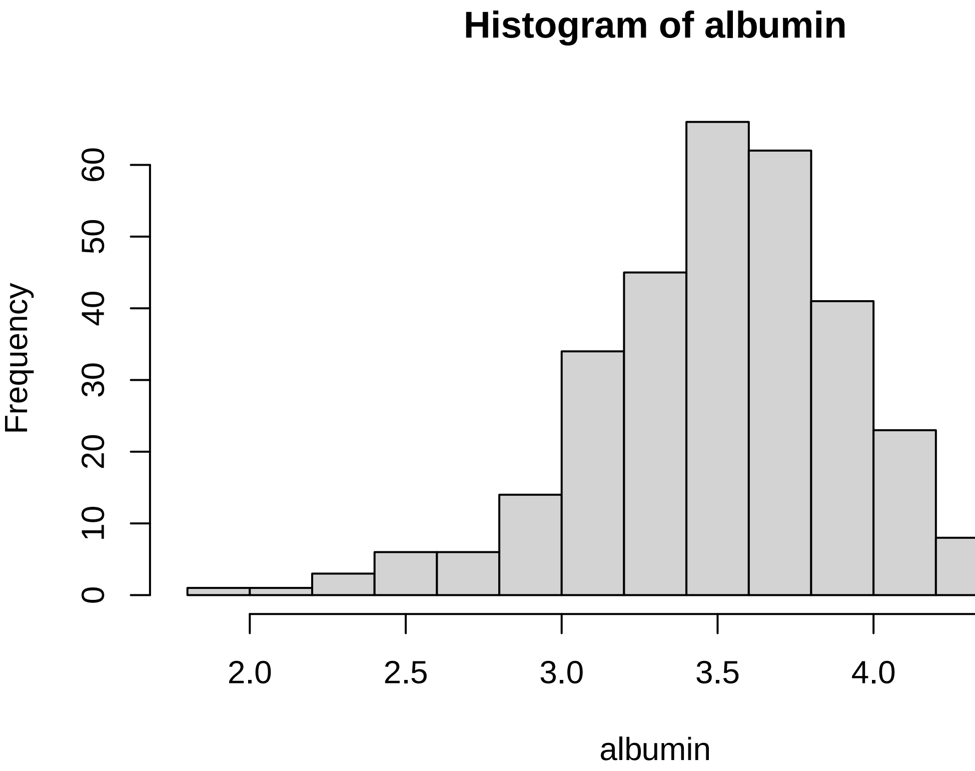
```
tst<-rgamma(1e5,312,1015.9)
hdi(tst)
```

```
##     lower     upper
## 0.2742004 0.3420808
## attr(,"credMass")
## [1] 0.95
```

12. Albumin

The Albumin is given in mg/dl.

12.1 Model selection



By the histogram plot of the data we see,

that albumin could be gamma distributed with shape and rate parameters a and b. We assume prior independence between a and b and use the marginal prior distributions $Gamma(0.001, 0.001)$ for both of them.

12.2 Results

Using OpenBUGS and MCMC methods, we get posterior information about the parameters a and b:

```
n<-length(albumin[!is.na(albumin)])
X<-albumin
data1<-list("X","n")
params<-c("a" , "b")
inits<-list(a=1,b=1)
fit1<-bugs(data=data1,inits=list(inits),parameters.to.save=params,"model_albu.txt",n.chains=1, n.iter=20
fit1$summary
```

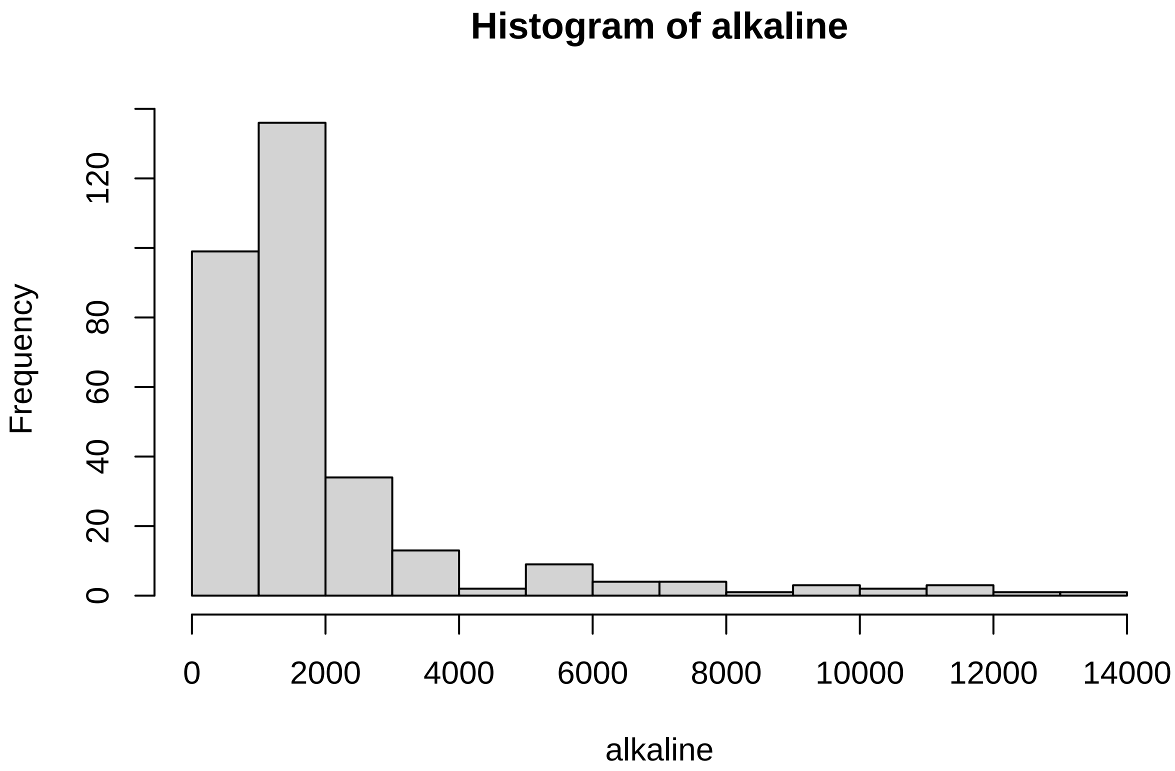
```
##                 mean       sd   2.5%    25%    50%    75%    97.5%
## a          66.32709 4.995273  57.66  62.76  66.30  69.38  77.1405
## b          18.84450 1.424988  16.37  17.83  18.84  19.71  21.9400
## deviance 365.10028 1.880193 363.30 363.70 364.50 365.80 370.1000
```

13. Alkaline

The data contains the units of alkaline phosphatase per liter of the patients.

13.1 Model selection

Since the units of alkaline per liter are integers, we assume that it is a counting process. Therefore we want to assume, that the data is poisson distributed conditional on one parameter λ. The histogram plot justifies our assumptions:



The density function of a single obervation is given as

$$f(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!} \tag{4}$$

As a prior for λ we use the natural conjugate prior of the poisson distribution which is the gamma distribution. To not give a lot of prior information, we will use $Gamma(0.001, 0.001)$.

13.2 Results

From our data we get the sample size n and the sum s over the sample:

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

```
## [1] 618588.6
```

The posterior distribution for λ is given by $Gamma(s + 0.001, n + 0.001)$ which in our case results to $Gamma(618588.601, 312.001)$. This yields to:

```
## Posterior mean:  1982.649
```

```
## Posterior mode:  1982.646
```

```
## Centered 95% Confidence Interval: [ 1977.712 , 1987.593 ]
```

And the HPD confidence interval calculates to:

```
tst<-rgamma(1e5,618588.601,312.001)
hdi(tst)
```

```
##    lower    upper
## 1977.708 1987.573
## attr(,"credMass")
## [1] 0.95
```

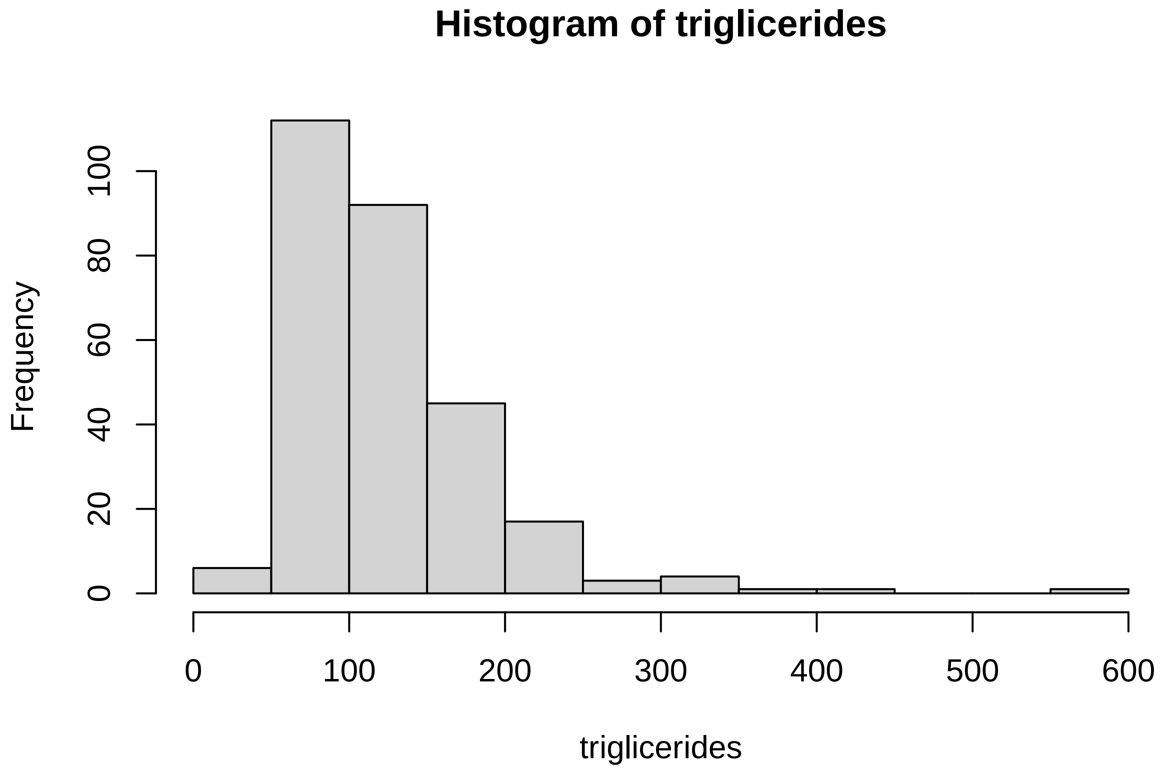
14. Triglicerides

Triglicerides of the patients in mg/dl.

14.1 Model selection

By the histogram plot of the data we see,

```
## Warning: NAs durch Umwandlung erzeugt
```



that triglicerides could be gamma distributed with shape and rate parameters a and b. We assume prior independence between a and b and use the marginal prior distributions $Gamma(0.001, 0.001)$ for both of them.

14.2 Results

By OpenBUGS and MCMC methods we get posterior information about the parameters a and b:

```
n<-length(triglicerides[!is.na(triglicerides)])
X<-triglicerides
```

```
data1<-list("X","n")
params<-c("a" , "b")
inits<-list(a=1,b=1)
fit1<-bugs(data=data1,inits=list(inits),parameters.to.save=params,"model_albu.txt",n.chains=1, n.iter=2
fit1$summary
```

```
##                  mean          sd         2.5%       25%        50%       75%
## a        4.742428e+00 0.410163943 3.988975e+00 4.457e+00 4.7310e+00 5.010e+00
## b        3.775649e-02 0.003443284 3.136975e-02 3.535e-02 3.7645e-02 4.001e-02
## deviance 2.732715e+03 1.974568784 2.731000e+03 2.731e+03 2.7320e+03 2.734e+03
##               97.5%
## a           5.58705
## b           0.04484
## deviance 2738.00000
```

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

$$h(\theta) \propto \sqrt{det I(\theta)}$$

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

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

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