Bayesian statistics: project

ENSAE 2020/2021

Mastère Spécialisé - Data Science

Sarah LAUZERAL - Amale NOKRI

Abstract

The objective of this project is to estimate the mean number of events in a given period of time from a data set of plane accidents dates. We already analysed this data set with frequentist procedure in another class. Each line of this database corresponds to an airplane accident that took place between 01/01/1972 and 31/12/1975 and contains 3 columns: the exact date of the accident, the number of victims and the day of the accident from 1 to 1460 ($= 4 \times 365$) which respectively corresponds to 01/01/1972 and 31/12/1975.

More precisely, we will estimate the mean number of plane accidents per week and per month by different methods. On the one hand, we will compute the Bayesian estimator of the mean (by week & by month) derived from a non-informative prior and, on the other hand, we will compute the Bayesian estimator of the mean (by week & by month) derived from a gamma prior.

Contents

1 Motivation & methodology used

Let's explain why our problem can be modeled by a Poisson law. Events occur at random instants of time at an average rate of λ events per second. Let $N(t)$ be the number of event occurrences in $[0,t]$, $N(t)$ is a non decreasing, integer valued, continuous-time random process. Let's now assume that $[0,t]$ is divided into n subintervals of width $\Delta t = \frac{t}{n}$.

We can reasonably state the 2 following assumptions:

1. The probability of more than one event occurrences in a subinterval is negligible compared to the probability of observing one or zero events. That is, outcome in each subinterval is a Bernoulli trial.
2. Whether or not an event occurs in a subinterval is independent of the outcomes in other intervals. That is, these Bernoulli trials are independent.

These two assumptions together imply that the counting process $N(t)$ can be approximated by the binomial counting process that counts the number of successes in the n Bernoulli trials.

If the probability of an event occurrence in each subinterval is p, then the expected number of event occurrences in $[0,t]$ is np. Since events occur at the rate λ events per second, then

$$\lambda t = np \Leftrightarrow p = \frac{\lambda t}{n}$$

Let $n \to +\infty$, $p \to 0$ while $\lambda t = np$ remains fixed, the binomial distribution approaches a Poisson distribution with parameter λt.

$$\begin{aligned}
\mathbb{P}[N(t)=k] &= \binom{n}{k} p^k (1-p)^{n-k} \\
&= \binom{n}{k} \left(\frac{\lambda t}{n}\right)^k \left(1-\frac{\lambda t}{n}\right)^{n-k} \\
&= \frac{n!}{(n-k)!}\frac{1}{k!}\left(\frac{\lambda t}{n}\right)^k \left(1-\frac{\lambda t}{n}\right)^{n-k} \\
&\underset{n\to\infty}{\longrightarrow} n^k \frac{1}{k!}\left(\frac{\lambda t}{n}\right)^k exp\left[(n-k)ln\left(1-\frac{\lambda t}{n}\right)\right] \\
&\underset{n\to\infty}{\sim} \frac{(\lambda t)^k}{k!} exp\left[-\frac{\lambda t}{n}(n-k)\right] \\
&\underset{n\to\infty}{\longrightarrow} \frac{(\lambda t)^k}{k!} exp\left[-\lambda t\left(1-\frac{k}{n}\right)\right] \\
&\underset{n\to\infty}{\longrightarrow} \frac{(\lambda t)^k}{k!} e^{\lambda t}
\end{aligned}$$

The Poisson process $N(t)$ inherits properties of independent and stationary increments from the underlying binomial process. Hence, the probability mass function for the number of event occurrences in any interval of length t is given by the above formula.

2 Description of the Bayesian techniques used

Let's assume that the model can be written as follow : $X_{1:n}|\theta \underset{i.i.d}{\sim} \mathcal{P}(\theta)$

$$\begin{aligned}\mathbb{P}[x_{1:n}|\theta] &= \prod_{i=1}^{n} \mathbb{P}[X_i = x_i|\theta] \\ &= \prod_{i=1}^{n} e^{-\theta} \frac{\theta^{x_i}}{x_i!} \\ &= e^{-n\theta} \frac{\theta^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}\end{aligned}$$

2.1 Non-informative prior

Uniform prior: $\mathbb{P}(\theta) \sim \mathcal{U}[0, c]$ with $c = max$(number of plane accidents per week or month)
We believe a priori that $\mathbb{P}(\theta)$ is equally likely to take 100 values between 0 and c.
By Bayes' theorem, the posterior distribution can be written as

$$\mathbb{P}[\theta|x_{1:n}] = \frac{\mathbb{P}[x_{1:n}|\theta] \times \mathbb{P}(\theta)}{\mathbb{P}(x_{1:n})} \propto \mathbb{P}[x_{1:n}|\theta] \times \mathbb{P}(\theta) \propto \mathbb{P}[x_{1:n}|\theta]$$

Since $\mathbb{P}(\theta) \propto constant$.
So the posterior distribution of this model is:

$$\mathbb{P}[\theta|x_{1:n}] \propto e^{-n\theta} \frac{\theta^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}$$

Hence we can deduce the estimator of the mean which is equal to $\widehat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i$

2.2 Informative prior

Gamma prior: $\forall \theta \in \mathbb{R}_+^*, \mathbb{P}(\theta) = \theta^{\alpha-1} \frac{e^{-\frac{\theta}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}}$
We know that the posterior distribution can be computed as follow thanks to the Bayes formula:

$$\mathbb{P}[\theta|x_{1:n}] = \frac{\mathbb{P}[x_{1:n}, \theta]}{\mathbb{P}(x_{1:n})} = \frac{\mathbb{P}[x_{1:n}|\theta] \times \mathbb{P}(\theta)}{\int_0^{+\infty} \mathbb{P}[x_{1:n}|\theta]\mathbb{P}(\theta)d\theta}$$

Let's first compute the joint distribution (numerator of the posterior):

$$\begin{aligned}\mathbb{P}[x_{1:n}, \theta] &= e^{-n\theta} \frac{\theta^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!} \times \theta^{\alpha-1} \frac{e^{-\frac{\theta}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}} \\ &= \theta^{\sum_{i=1}^{n} x_i + \alpha - 1} \frac{e^{-\theta\left(n+\frac{1}{\beta}\right)}}{\prod_{i=1}^{n} x_i!\Gamma(\alpha)\beta^{\alpha}}\end{aligned}$$

Now we will compute the marginal distribution:

$$\begin{aligned}\mathbb{P}(x_{1:n}) &= \int_0^{+\infty} \theta^{\sum_{i=1}^{n} x_i + \alpha - 1} \frac{e^{-\theta\left(n+\frac{1}{\beta}\right)}}{\prod_{i=1}^{n} x_i!\Gamma(\alpha)\beta^{\alpha}} d\theta \\ &= \frac{1}{\prod_{i=1}^{n} x_i!\Gamma(\alpha)\beta^{\alpha}} \int_0^{+\infty} \theta^{\sum_{i=1}^{n} x_i + \alpha - 1} \times e^{-\theta\left(n+\frac{1}{\beta}\right)} d\theta \\ &= \frac{1}{\prod_{i=1}^{n} x_i!\Gamma(\alpha)\beta^{\alpha}} \Gamma\left(\sum_{i=1}^{n} x_i + \alpha\right) \left(\frac{1}{n+\frac{1}{\beta}}\right)^{\sum_{i=1}^{n} x_i + \alpha} \int_0^{+\infty} \frac{\theta^{\sum_{i=1}^{n} x_i + \alpha - 1} \times e^{-\theta\left(n+\frac{1}{\beta}\right)}}{\Gamma\left(\sum_{i=1}^{n} x_i + \alpha\right) \left(\frac{1}{n+\frac{1}{\beta}}\right)^{\sum_{i=1}^{n} x_i + \alpha}} d\theta\end{aligned}$$

We notice that:

$$\frac{\theta^{\sum_{i=1}^n x_i+\alpha-1}\times e^{-\theta\left(n+\frac{1}{\beta}\right)}}{\Gamma\left(\sum_{i=1}^n x_i+\alpha\right)\left(\frac{1}{n+\frac{1}{\beta}}\right)^{\sum_{i=1}^n x_i+\alpha}} \sim \Gamma\left(\sum_{i=1}^n x_i+\alpha, \frac{1}{n+\frac{1}{\beta}}\right)$$

Hence we deduce that:

$$\int_0^{+\infty} \frac{\theta^{\sum_{i=1}^n x_i+\alpha-1}\times e^{-\theta\left(n+\frac{1}{\beta}\right)}}{\Gamma\left(\sum_{i=1}^n x_i+\alpha\right)\left(\frac{1}{n+\frac{1}{\beta}}\right)^{\sum_{i=1}^n x_i+\alpha}} d\theta = 1$$

Thus we have:

$$\mathbb{P}(x_{1:n}) = \frac{\Gamma\left(\sum_{i=1}^n x_i+\alpha\right)}{\prod_{i=1}^n x_i!\Gamma(\alpha)\beta^\alpha}\left(\frac{1}{n+\frac{1}{\beta}}\right)^{\sum_{i=1}^n x_i+\alpha}$$

To conclude, the posterior distribution is:

$$\begin{aligned}\mathbb{P}[\theta|x_{1:n}] &= \theta^{\sum_{i=1}^n x_i+\alpha-1}\frac{e^{-\theta\left(n+\frac{1}{\beta}\right)}}{\prod_{i=1}^n x_i!\Gamma(\alpha)\beta^\alpha}\frac{\prod_{i=1}^n x_i!\Gamma(\alpha)\beta^\alpha}{\Gamma\left(\sum_{i=1}^n x_i+\alpha\right)}\left(n+\frac{1}{\beta}\right)^{\sum_{i=1}^n x_i+\alpha} \\ &= \theta^{\sum_{i=1}^n x_i+\alpha-1}\frac{e^{-\theta\left(n+\frac{1}{\beta}\right)}}{\Gamma\left(\sum_{i=1}^n x_i+\alpha\right)\left(\frac{1}{n+\frac{1}{\beta}}\right)^{\sum_{i=1}^n x_i+\alpha}} \\ &\sim \Gamma\left(\sum_{i=1}^n x_i+\alpha, \frac{1}{n+\frac{1}{\beta}}\right)\end{aligned}$$

Therefore the estimator of the mean is equal to $\widehat{\theta} = \frac{\sum_{i=1}^n x_i+\alpha}{n+\frac{1}{\beta}}$

3 Explanation and interpretation of the results

Let's now compute our estimators per month and per week in order to compare them.

3.1 Weekly estimators

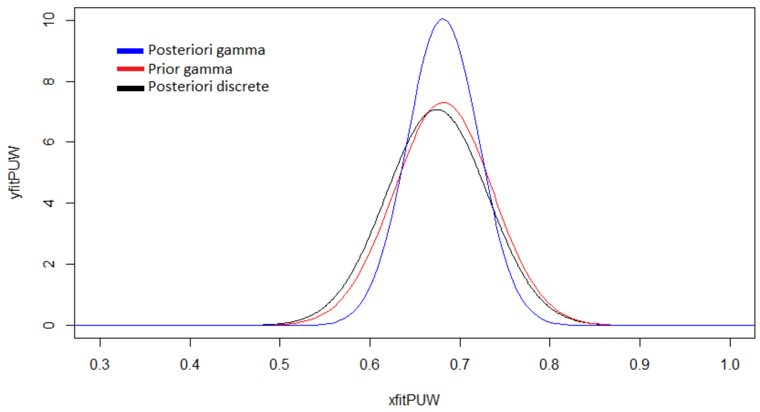

Figure 1: Weekly distribution

The posteriori gamma distribution is more accurate than the gamma prior and the uniform prior. The gamma posteriori is continuous so the probability is going to be more accurate than the uniform posteriori which, in this case, as we can see, is close to the gamma posteriori because of the set of θ (100 values between 0 and 4).

3.1.1 Non-informative prior

There are 212 weeks in our data set and the maximum number of accident per week is $c = 4$.
Using the prior probabilities previously defined, the mean number of accidents per week is estimated to $\widehat{\theta} = \frac{1}{212}\sum_{i=1}^{212} x_i \approx 0.6745$ accidents per week.

3.1.2 Informative prior

There are 212 weeks in our data set so $\beta = \frac{1}{212}$.
Using the prior probabilities previously defined, the mean number of accidents per week is estimated to $\widehat{\theta} = \frac{142+142}{212+212} \approx 0.6698$ accidents per week.

3.2 Monthly estimators

For this part we will calibrate our gamma prior such that $\alpha = 142$ which corresponds to the total number of plane accidents in our data set and β will be equal to the inverse of the time duration considered.

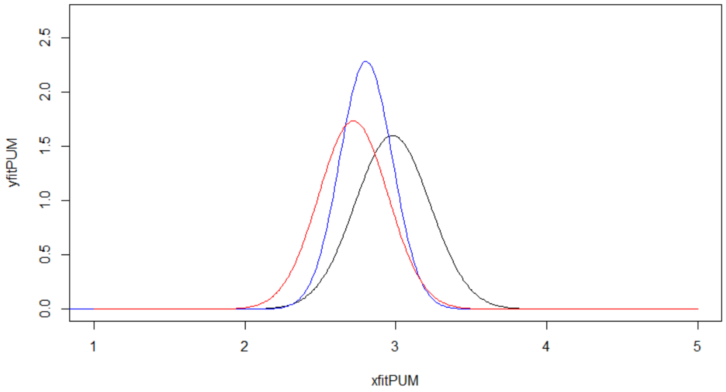

Figure 2: Monthly distributions

In this case, the difference between the 3 densities is more noticeable than per week, because of the set of values (100 values between 0 and 7). It makes the uniform be less accurate.

3.2.1 Non-informative prior

There are 48 months in our data set and the maximum number of accident per month is $c = 7$.
Using the prior probabilities previously defined, the mean number of accidents per month is estimated to $\widehat{\theta} = \frac{1}{48}\sum_{i=1}^{48} x_i \approx 2.9792$ accidents per month.

3.2.2 Informative prior

There are 48 weeks in our data set so $\beta = \frac{1}{48}$.
Using the prior probabilities previously defined, the mean number of accidents per month is estimated to $\widehat{\theta} = \frac{142+142}{48+48} \approx 2.9583$ accidents per month.

3.3 Data-based simulation approach

Considering now $B = 500$ random samples using the bootstrap method such that: $\widehat{\theta}^{(b)} = T(X_1^{(b)}, ..., X_n^{(b)})$. We want to evaluate the bias and root mean square error (RMSE) of both Bayesian estimators using this simulation.

Bias is a measure of accuracy of an estimator. It measures the difference between the expected value of the parameter and the actual parameter such that

$$\widehat{Bias} = \frac{1}{B}\sum_{b=1}^{B} \widehat{t}^{(b)} - \widehat{t} = \overline{\widehat{t'}} - \widehat{t}$$

However, to conclude on the accuracy of an estimator, it is necessary to also take into account its variance, that's why we will take a look at the Root Mean Square Error (RMSE) of our estimators.

$$\widehat{RMSE} = \sqrt{\frac{1}{B}\sum_{b=1}^{B} \left(\widehat{t}^{(b)} - \widehat{t}\right)^2}$$

We obtain the following results for the weekly estimators:

	Uniform prior	Gamma prior
`Bias`	-0.001377367	-0.0006886792
`RMSE`	0.05656052	0.02828026

Table 1: Weekly estimators comparison

and the following results for the monthly estimators:

	Uniform prior	Gamma prior
`Bias`	0.005791667	0.002895833
`RMSE`	0.2500226	0.1250113

Table 2: Monthly estimators comparison

4 Conclusion

The goal of this project was to estimate the mean number of plane accidents in a given time period by several methods : a uniform prior and a gamma prior. To compare these methods, we estimated the bias and the root mean square error of each estimator To compare their accuracy.

Per week estimators both have a negative bias, meaning that they overestimate the true θ on average whereas per month estimators both have a positive bias, meaning that they underestimate the true θ. Furthermore, the Bayesian estimator derived from a Gamma prior has the lowest bias in both time period. It also has the lowest RMSE and is therefore the most accurate estimator.

Furthermore, one can notice that the bias and the RMSE of the monthly estimator are larger than the per week estimator since the sample size is larger for the second one.