MATHEMATICAL STATISTICS
2018-2019
PROJECT 3 : BAYESIAN ESTIMATION AND PREDICTION OF COUNTS OF PLANE ACCIDENTS

Recommended readings for this project.

1 Poisson process

`https://www.math.ust.hk/~maykwok/courses/ma246/04_05/04MA246L4B.pdf`

2 Definition of conjugate prior

`https://en.wikipedia.org/wiki/Conjugate_prior`

3 Introduction to Bayes using discrete priors

`https://cran.r-project.org/web/packages/LearnBayes/vignettes/DiscreteBayes.pdf`
`https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading11.pdf`

4 Introduction to posterior predictive distribution

`http://www2.stat.duke.edu/~rcs46/modern_bayes17/lecturesModernBayes17/lecture-1/01-intro-to-Bayes.pdf`
or
`https://www.youtube.com/watch?v=R9NQY2Hyl14`

Objective. The objective of this project is to estimate the mean number of events in a given period of time from a data set of events dates. We consider an application to a series of plane accidents. You will learn about conjugate priors, predictive probability and data-based simulations.
Work to do.

1) Using the first document, explain why the Poisson model is well adapted for modeling the number of events in a given period of time when events occur continuously at random in time with a small occurrence probability. We will assume in the sequel that the number $N(t, t+\Delta)$ of events occurring in the time interval $[t, t+\Delta]$ follows a Poisson distribution with mean denoted by θ_Δ., where Δ is some given time period.

2) Prove that if X follows a Poisson distribution with mean θ and if, in a Bayesian framework, we consider an a priori distribution for θ which is a Gamma distribution with parameters (α, β), then the posterior of θ given X is gamma distributed (specify its parameters). For the gamma distribution (see wikipedia) we will use the shape/scale parametrization where α is called the **shape** parameter and β the **scale** parameter.

3) Using the provided data base, estimate the mean number of accidents **per week, and per month** by the following three methods

   - maximum likelihood,
   - a bayesian estimator derived from a discrete prior taking 100 values between 0 and the maximum number of counts in the given time period
   - a bayesian estimator derived from a gamma prior as in question 2 (provide the formula) using some sensible value for a priori parameters. To help you for this choice : one can view the shape parameter as representing the overall count on the considered time period and the scale parameter as representing the inverse of the observation time duration. Bonus : you may try with a second choice.

4) Plot the following four density curves

   - the a priori Gamma density of θ that you have chosen
   - the a posteriori density of θ corresponding to your Gamma a priori
   - the a posteriori density of θ corresponding to the discrete uniform of the previous question
   - the theoretical asymptotic density of θ

5) Adopting a data-based simulation approach, you will generate $B = 500$ random samples of weekly and monthly events with the means being set to the MLE estimates of the previous question and the sample size corresponding to the number of weeks (resp : month) in your sample. Evaluate the bias and root mean square error of the maximum likelihood and the two bayesian estimators with this simulation and compare them.

6) Explain what are prior and posterior predictive distributions. Derive the posterior predictive distribution associated to the Gamma prior. For the weekly data, draw a sample from the prior predictive and one from the posterior predictive and find a graphical way of highlighting the difference between these two empirical distributions.