Homework #01

Statistical Methods in Data Science II & Lab

2021/2022

deadline April 22th, 2022

```
#{r, include=FALSE, include=FALSE, warning=FALSE} ##opts_chunk$set(out.lines = 23) #
```


Barboni Alessio, 2027647

*Your answers for each data analysis question should discuss the problem, data, model, method, conclusions.

Fully Bayesian conjugate analysis of Rome car accidents

Consider the car accident in Rome (year 2016) contained in the `data.frame` named `roma`. Select your data using the following code

```
mydata <- subset(roma,subset=sign_up_number==104)
str(mydata)
```

```
## 'data.frame':    19 obs. of  5 variables:
##  $ week          : int  2 3 4 5 6 7 8 9 10 11 ...
##  $ weekday       : chr  "Saturday" "Saturday" "Saturday" "Saturday" ...
##  $ hour          : int  9 9 9 9 9 9 9 9 9 9 ...
##  $ car_accidents : int  3 2 4 1 4 8 4 8 3 5 ...
##  $ sign_up_number: int  104 104 104 104 104 104 104 104 104 104 ...
```

The column `car_accidents` contains the number of car accidents $Y_i = y_i$ occurred in a specific weekday during a specific hour of the day in some of the weeks of 2016. Using the observed outcomes of the number of car accidents do a fully Bayesian analysis using as a statistical model a conditionally i.i.d. Poisson distribution with unknown parameter. Take into account that it is known that the average number of hourly car accidents occurring in Rome during the day is **3.22**. In particular do the following:

1. describe your observed data
2. justify as best you can your choices for the ingredients of your Bayesian model especially for the choices you make for the prior distribution
3. report your main inferential findings using your posterior distribution of the unknown parameter in terms of

   a) possible alternative point estimates with comments on how similar they are and, in case, why

b) posterior uncertainty

c) interval estimates justifying your (possibly best) choices

d) suitable comments on the differences between the prior and the posterior

e) (optional) Provide a formal definition of the posterior predictive distribution of $Y_{next}|y_1, ..., y_n$ and try to compare the posterior predictive distribution for a future observable with the actually observed data

Bulb lifetime

You work for Light Bulbs International. You have developed an innovative bulb, and you are interested in characterizing it statistically. You test 20 innovative bulbs to determine their lifetimes, and you observe the following data (in hours), which have been sorted from smallest to largest.

1, 13, 27, 43, 73, 75, 154, 196, 220, 297,
344, 610, 734, 783, 796, 845, 859, 992, 1066, 1471

Based on your experience with light bulbs, you believe that their lifetimes Y_i can be modeled using an exponential distribution conditionally on θ where $\psi = 1/\theta$ is the average bulb lifetime.

1. Write the main ingredients of the Bayesian model.
2. Choose a conjugate prior distribution $\pi(\theta)$ with mean equal to 0.003 and standard deviation 0.00173.
3. Argue why with this choice you are providing only a vague prior opinion on the average lifetime of the bulb.
4. Show that this setup fits into the framework of the conjugate Bayesian analysis
5. Based on the information gathered on the 20 bulbs, what can you say about the main characteristics of the lifetime of yor innovative bulb? Argue that we have learnt some relevant information about the θ parameter and this can be converted into relevant information about $1/\theta$
6. However, your boss would be interested in the probability that the average bulb lifetime $1/\theta$ exceeds 550 hours. What can you say about that after observing the data? Provide her with a meaningful Bayesian answer.