Statistical Inference
R: Assessment 2

István Z. Kiss and Francesco Di Lauro

1. The deadline for this assessment is 14:00 Thursday, Week 11, 29th of April 2021.

2. Please submit online and use your corresponding Canvas site for submission, i.e. if you are taking this course at Level 7, please submit via the Canvas site corresponding to Level 7 (same for Level 6).

3. Your work must be submitted as ONE SINGLE PDF file. This may contain R code, figures produced in R to support your answers, handwritten calculations that you typed up or photographed and inserted it into your word or other editor.

4. You can use a text editor of your choice.

5. Each figure must be annotated with title, axes labels and legends (when multiple graphs are shown on the same plot).

1 Faulty fibres

The number of flaws in 1-km lengths of fibres can be assumed to be independent and follow a Poisson distribution with mean $\lambda(> 0)$ (this is the same as the parameter of the distribution). The prior distribution of λ is $\Gamma(2, 12)$ (here the shape is $\alpha = 2$ and the rate is $\beta = 12$). When 10 randomly selected 1-km lengths are inspected, 8 flaws are found in total.

1. Show that the posterior distribution of λ is $\Gamma(10, 22)$. You should do this calculation by hand and write it down or type it. Using R, plot the posterior.

2. Compute the maximum likelihood estimate of λ, as it is done within the frequentist approach.

3. Using R, or otherwise find the mean and the maximum a posteriori (MAP) (you can use a formula for the mean posterior and MAP can be read out from your plot, an approximation is fine). Compare these to the maximum likelihood estimate. Are there big differences? What was the mean of the prior and how does this affect the posterior?

4. Using R, find the posterior probability that λ is less than 0.646.

5. Using R, find the high density probability credible region at 90%, and produce a plot that shows the posterior and two vertical lines at both ends of the credible region that you have found. Approximate values are acceptable.

6. **At Level 7 only:** With the posterior in mind, write down the formula for the probability that a newly-selected 1-km long fibre contains no errors.

Hint: The likelihood is based on assuming independent samples so it is simply the product of the probabilities of observing each data point on its own. When writing down the likelihood you may want to treat the number of faults in each 1-km fibre as i.i.d. random variables X_i with $i = 1, 2, \dots, 10$. This means that you have 10 observations $x_1, x_2, \dots, x_{10}$ (these are now realisations and no longer random variables). Make sure that you use the Γ function in R with shape and rate. Sometimes the Γ distribution is used with shape and scale but here we use it as shape and rate, as in the lecture notes. Read the help documentation about the Γ distribution to familiarise yourself with this.

2 Model Selection

The number of daily visitors of a recently opened caffe is modelled as a random variable. The owner would like to get a better understanding of the number of customers the caffe may get. A small data set is available where each entry is the number of customers during the first five days of the caffe being open (you can find this in the file *datapoints.csv*. Looking at the dataset, you are not sure if the number of customers is best captured by a Poisson or a Geometric distribution, that is $\mathbb{P}(X = k) = p(1-p)^k$. You then decide to perform model selection, that is work out which distribution fits the data best. This is important for the owner since the model will be used to plan staff numbers and shifts.

1. For the Poisson case, show that $\Gamma(\alpha, \beta)$ (α is the shape, and β the rate) is a conjugate prior.

2. For the Geometric distribution, show that instead $Beta(\alpha, \beta)$ is a conjugate prior.

3. The prior for the Poisson distribution chosen is $\Gamma(5, 0.5)$, whereas for the Geometric is $Beta(5, 10)$. Plot the priors and the posterior distributions for both the models.

4. Using R, find the maximum a posteriori (MAP) and the conditional mean for each model. An approximate value is acceptable.

5. Using R, compute the Bayes factor of the two models, assuming that you assign them equal probability $P(M_{Poisson}) = P(M_{Geometric}) = \frac{1}{2}$. Which model is favoured by the Bayes factor?

6. **At Level 7 only:** For some reason, you now believe that a $\Gamma(12, 0.1)$ is more reasonable as a prior for the Poisson model. Repeat the last point. Which distribution would you choose now?

Hint: To compute the Bayes factor, you need to integrate a posterior distribution that depends on one parameter. To do so, you can use the built-in routine called *integrate* in R. This function takes as inputs a function (that you need to define), and two numbers that will be the limits of integration. Since for the Poisson distribution, the rate goes from 0 to ∞, in principle, it is a good idea to put a cut-off (you might want to check visually where to place it by inspecting the posterior distributions). Remember that for Bayes factor calculations, priors need to be normalised.