Specifying Bayesian Models

Javier Esteban Aragoneses, Mauricio Marcos Fajgenbaun, Danyu Zhang, Daniel Alonso

February 25th, 2021

Introduction

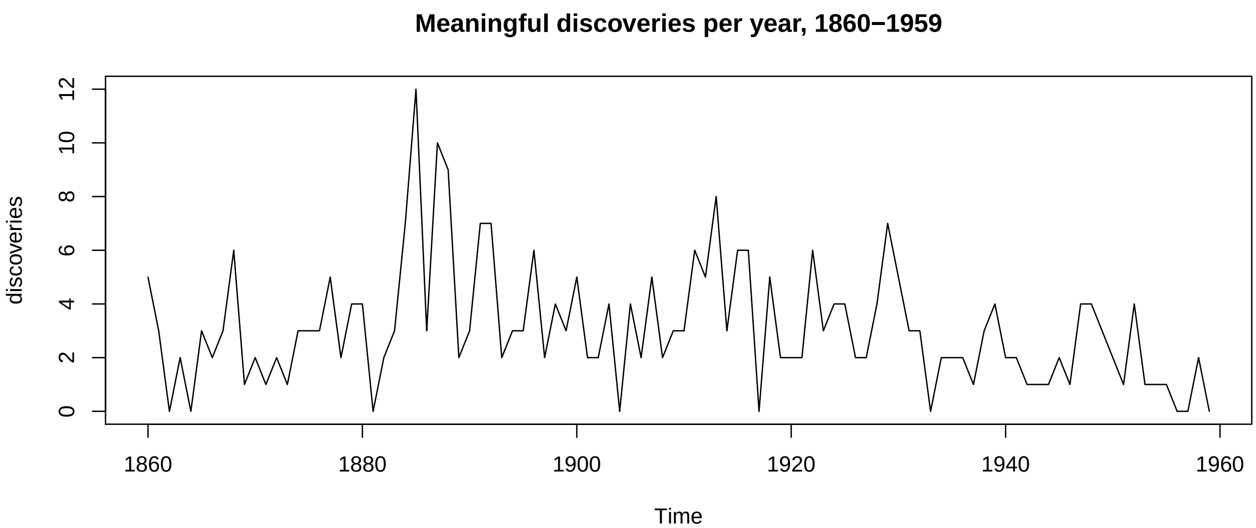
In this research, we want to provide a conditional estimation of the number of scientific discoveries per year. Our dataset consist of a table with the number of discoveries per year, from 1860 to 1959.

Discoveries dataset

The discoveries dataset is a 100-point time series which contains a list of significant discoveries per year. These discoveries represent *great* inventions or important scientific discoveries.

Observations start in the year 1860 and end in 1959. The dataset originally comes from a publication in *The World Almanac and Book of Facts*, 1975 edition, pages 315–318.

We can visualize the time series as follows:



Model

As our data is expressed as a ratio (a number per year), we will make use of a model based in a Poisson distribution. In other words, we consider this problem as a Poisson distribution problem with rate parameter λ, being our sample space: $y = 0, 1, 2 \ldots$. This distribution is a distribution for count variables: we can consider every year as a discrete count of discoveries. For this random variable:

$$E[Y|\theta] = \theta,$$

$$Var[Y|\theta] = \theta$$

If we consider the 100 y_i random variables as independent and identically distributed with mean θ we can say that the joint pdf may be expressed:

$$\begin{aligned} Pr(Y_1 = y_1, \ldots, Y_{100} = y_{100}|\theta) &= \prod_{i=1}^{100} p(y_i|\theta) \\ &= \prod_{i=1}^{100} \frac{1}{y_i!}\theta^{y_i} e^{-100*\theta} \\ &= c(y_1, \ldots, y_{100})\theta^{\sum y_i} e^{-100*\theta} \end{aligned}$$

Prior Selection

Let´s recall that a class of prior densities is conjugate for a sampling model, if the posterior distribution is also in the class. In this case, as we are working with a Poisson sampling model, our posterior distribution for θ has a particular form that makes us think that the best is to use a conjugate prior such as a gamma distribution. The pdf of the gamma distribution is as follows:

$$f(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$$

The parameters alfa and beta in the gamma distribution are shape and inverse-scale parameters, respectively. The mean of a gamma distribution is $E[\theta] = \frac{\alpha}{\beta}$ and the variance is $Var[\theta] = \frac{\alpha}{\beta^2}$. It is interesting to point out that this distribution is very flexible: small changes in alfa o beta produce a big variation in the shape or scale of the density.

Basing ourselves in the fact that there has been approximately 100 scientific discoveries during the 17th century (in average, 1 discovery every year), we will say that for the time period that we are studying there may had been 3 discoveries a year. This is due to the fact that we assume that during the next century there had been a big increase in the rate of discoveries, as a consequence of technology achievements.

Thus, we will use a gamma distribution as a prior, with parameters (3,1) to calculate the posterior predictive distribution of a future observation.

For calculating the posterior distribution of θ, we can say that the posterior expectation for θ is a convex combination of the prior expectation and the sample average, as follows:

$$E[\theta|y_1, \ldots, y_{100}] = \frac{a + \sum y_i}{b + 100} = E[\tilde{Y}|y_1, \ldots, y_{100}]$$

Results

Posterior mean

The posterior mean is a weighted average of the prior mean and the data estimates.

The expected value for future values given our sample is as follows:

$E[\tilde{Y}|y_1, \ldots, y_{100}] \approx 3.1$

Posterior 95% quantile-based confidence interval

According to the credible interval, the probability that the mean of discoveries is between *2.77* and *3.45* is about *0.95*.