

Modelling Earthquake Waiting Times

Tutorial on Bayesian Data Analysis by Ankit Tewari

We're going to practise how to model the waiting times between serious earthquakes using a bayesian approach. The **United States Geological Survey** maintains a list of **significant earthquakes worldwide**. We will model the rate of earthquakes of magnitude 4.0+ in the state of California during 2015. An iid exponential model on the waiting time between significant earthquakes is appropriate if we assume:

- earthquake events are independent,
- the rate at which earthquakes occur does not change during the year, and
- the earthquake hazard rate does not change (i.e., the probability of an earthquake happening tomorrow is constant regardless of whether the previous earthquake was yesterday or 100 days ago).

Earthquake data:

The significant earthquakes of magnitude 4.0+ in the state of California during 2015 occurred on the following dates :

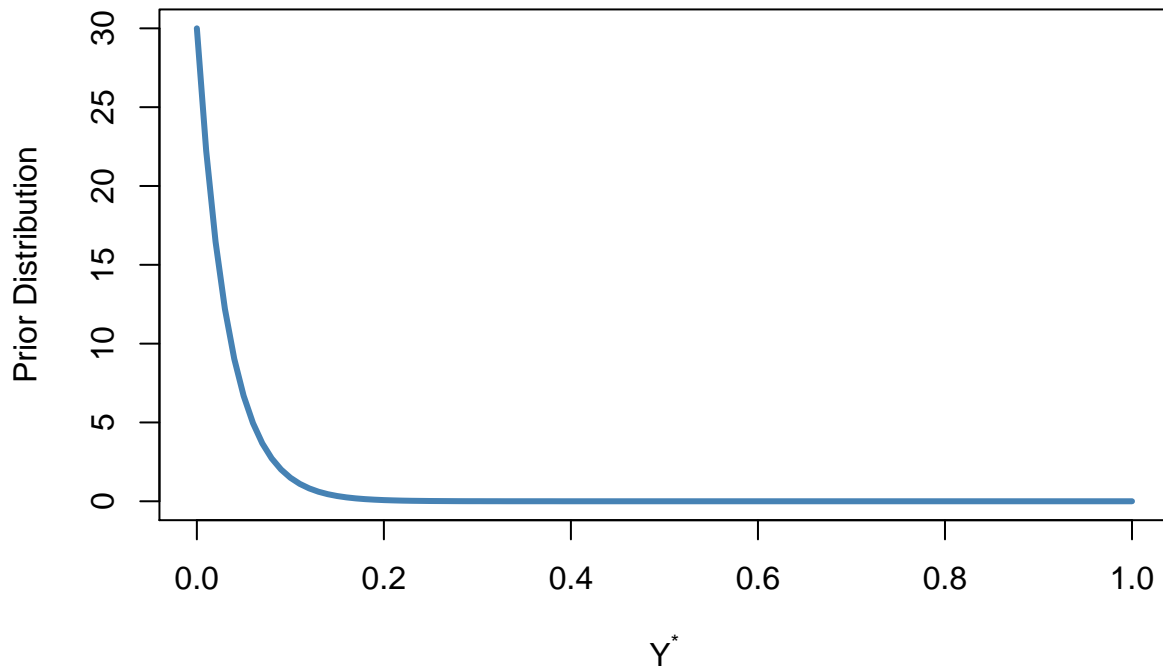
January 4, January 20, January 28, May 22, July 21, July 25, August 17, September 16, December 30. Let Y_i denote the waiting time in days between the i th earthquake and the following earthquake. Our model is $Y_i \sim i.i.d \text{ Exponential}(\lambda)$ where the expected waiting time between earthquakes is $E(Y) = 1/\lambda$ days.

Assume the conjugate prior $\lambda \sim \text{Gamma}(a, b)$. Suppose our prior expectation for λ is $1/30$, and we wish to use a prior effective sample size of one interval between earthquakes.

A good question then is that what is the value of a and b so that we can use them in our prior distribution? Well, as you might have guessed, we can use $a = 1$ and $b = 30$. It is so because in the exponential-gamma model, a is the prior effective sample size and The prior mean is $\frac{a}{b} = \frac{1}{30}$, and since we know the effective sample size $a = 1$, we have $b = 30$.

```
a = 1
b = 30
curve(dgamma(x, shape = a, rate = b), col="steelblue",
      lty=1, lwd=3, ylab = "Prior Distribution",
      xlab = expression(Y~"*"), main = "Modelling Earthquake Waiting Times")
```

Modelling Earthquake Waiting Times



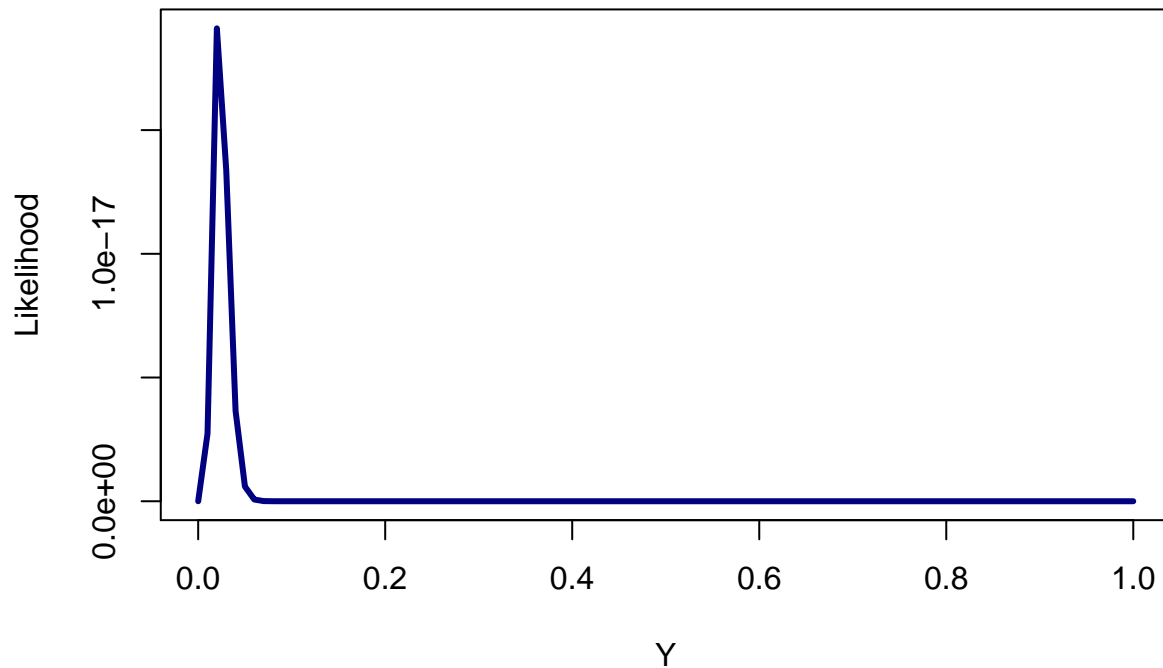
Which of the following is our data vector?

- $y = (16, 8, 114, 60, 4, 23, 30, 105)$
- $y = (3, 16, 8, 114, 60, 4, 23, 30, 105, 1)$
- $y = (3, 16, 8, 114, 60, 4, 23, 30, 105)$

Recall that we are modeling the waiting times between earthquakes in days. There are eight intervals between the first and last event. We are excluding four days of the year in which no events were observed. A more comprehensive model (e.g., censoring methods) would account for the fact that there were no major earthquakes Jan. 1 to Jan. 4 and Dec. 30 to Dec. 31. This is beyond the scope of this tutorial.

```
data <- c(16, 8, 114, 60, 4, 23, 30, 105)
plot( function(x) {return(x^length(data)*exp(-x*sum(data)))}, col="navyblue",
      lty=1, lwd=3, ylab = "Likelihood",
      xlab = expression(Y), main = "Modelling Earthquake Waiting Times")
```

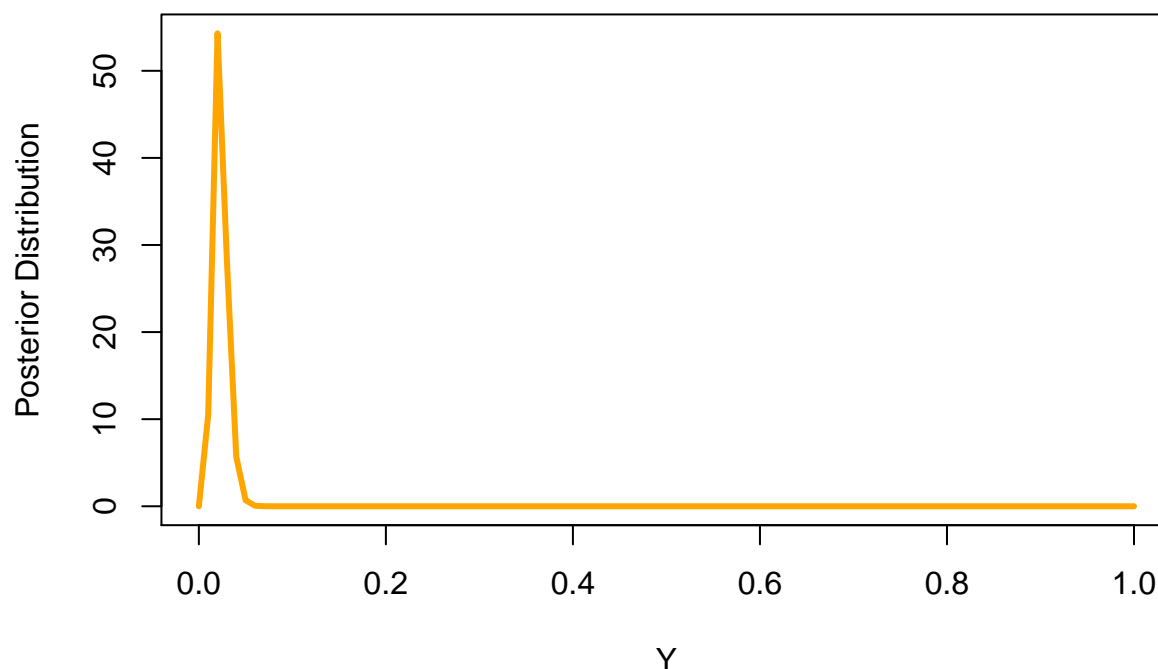
Modelling Earthquake Waiting Times



The posterior distribution is $\lambda|y \sim \text{Gamma}(\alpha, \beta)$ What is the value of α ? If you remember, we learnt that, This is $\alpha = a + n = 1 + 8$. Similarly, we learnt how to compute the value of β using the idea $\beta = b + \sum yi = 30 + 360$ which gives $\beta = 390$

```
curve( dgamma(x, shape = 9, rate = 390), col="orange",  
       lty=1, lwd=3, ylab = "Posterior Distribution",  
       xlab = expression(Y), main = "Modelling Earthquake Waiting Times")
```

Modelling Earthquake Waiting Times



Use R or Excel to calculate the upper end of the 95% equal-tailed credible interval for λ , the rate of major earthquakes in events per day. Round your answer to two decimal places. The full interval is (0.011, 0.040). Thus our posterior probability that $0.011 < \lambda < 0.040$ is 0.95.

The interval in terms of $1/\lambda$, the expected number of days between events is (24.7, 94.8). Note that although $E(1/\lambda) \neq 1/E(\lambda)$, we can take the reciprocals of quantiles since $P(X < q) = P(\frac{1}{q} < \frac{1}{X})$. Just remember that the lower end of the interval becomes the upper end and vice versa.

```
qgamma(p=0.975, shape=9, rate=390)
```

```
## [1] 0.04041843
```

The posterior predictive density for a new waiting time y^* in days is:

$$f(y^*|y) = \int f(y^*|\lambda) \cdot f(\lambda|y) d\lambda$$

or,

$$f(y^*|y) = \frac{\beta^\alpha \Gamma(\alpha + 1)}{(\beta + y^*)^{\alpha+1} \Gamma(\alpha)} I_{\{y^* \geq 0\}}$$

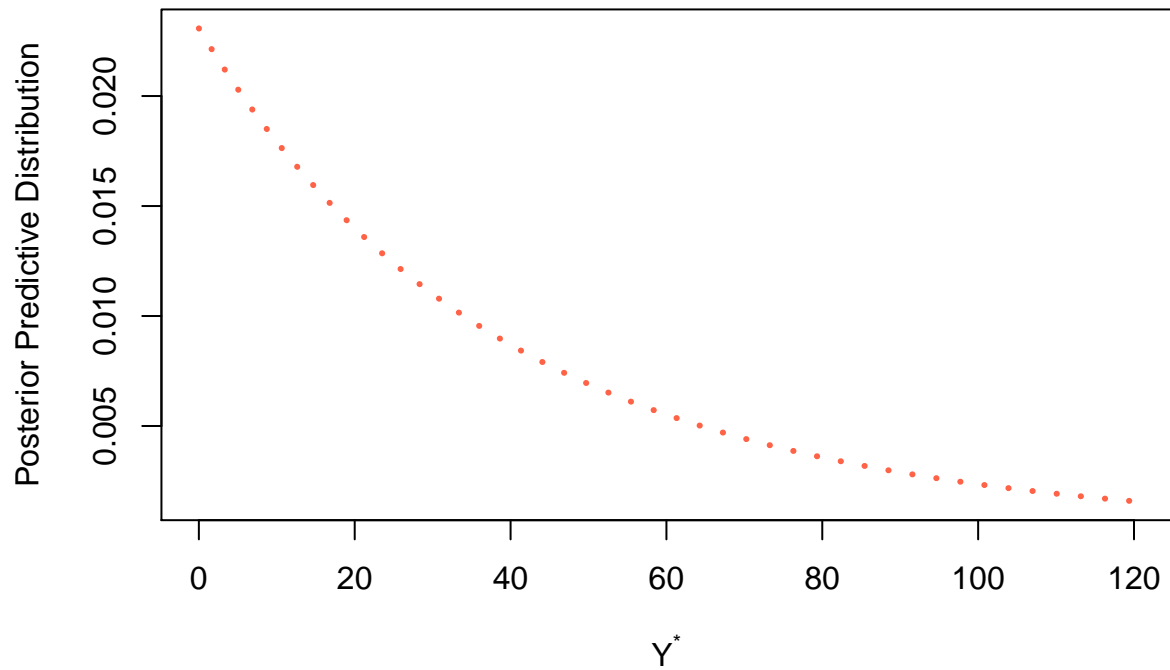
Now, as we know that $\Gamma(\alpha + 1) = \alpha$, we can deduce the above formula as,

$$f(y^*|y) = \frac{\beta^\alpha \alpha}{(\beta + y^*)^{\alpha+1}} I_{\{y^* \geq 0\}}$$

where $f(\lambda|y)$ is the $Gamma(\alpha, \beta)$ posterior found earlier. Using R to evaluate this posterior predictive probability distribution function, we can arrive at the following plot-

```
alpha = a + length(data)
beta = b + sum(data)
curve( (((390^9)*9)/(390 + x)^10), xlim = c(0, 120), col="tomato",
      lty=3, lwd=3, ylab = "Posterior Predictive Distribution",
      xlab = expression(Y~"*"), main = "Predicting Future Earthquake Waiting Times")
```

Predicting Future Earthquake Waiting Times



Given the data, this is the distribution of the waiting time (in days) between significant earthquakes. It turns out that the first significant 4.0+ magnitude earthquake in California in 2016 occurred on January 6.