Homework 1: Likelihood Methods in Biology

Laura Jimenez

February 3, 2016

1. *Your sample is n independent measurements from a Poisson distribution:*

$$P[X = k] = \frac{e^{-\lambda}\lambda^k}{k!}. \tag{1}$$

(a) *If the data are the values $X_1, X_2, ..., X_n$, solve analytically (using calculus) for the maximum likelihood estimator for λ.*

Assuming that the observations are independent and that they come from a Poisson distribution given by ecuation (1), after some algebraic simplifications, the likelihood function has the following expression:

$$\begin{aligned} \mathcal{L}(\lambda) &\propto \prod_{i=1}^{n} P[X_i = x_i] = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i} \prod_{i=1}^{n} \frac{1}{x_i!} \\ &\propto e^{-n\lambda}\lambda^{n\bar{x}}, \end{aligned} \tag{2}$$

where $\bar{x}$ is the average of the measurements and the factor $\prod_{i=1}^{n} \frac{1}{x_i!}$ doesn't depend on the parameter of interest, so it is a constant.
After taking the natural logarithm of the expression given in (2), we get the log-likelihood function for λ,

$$\ell(\lambda) = -n\lambda + n\bar{x}\log\lambda. \tag{3}$$

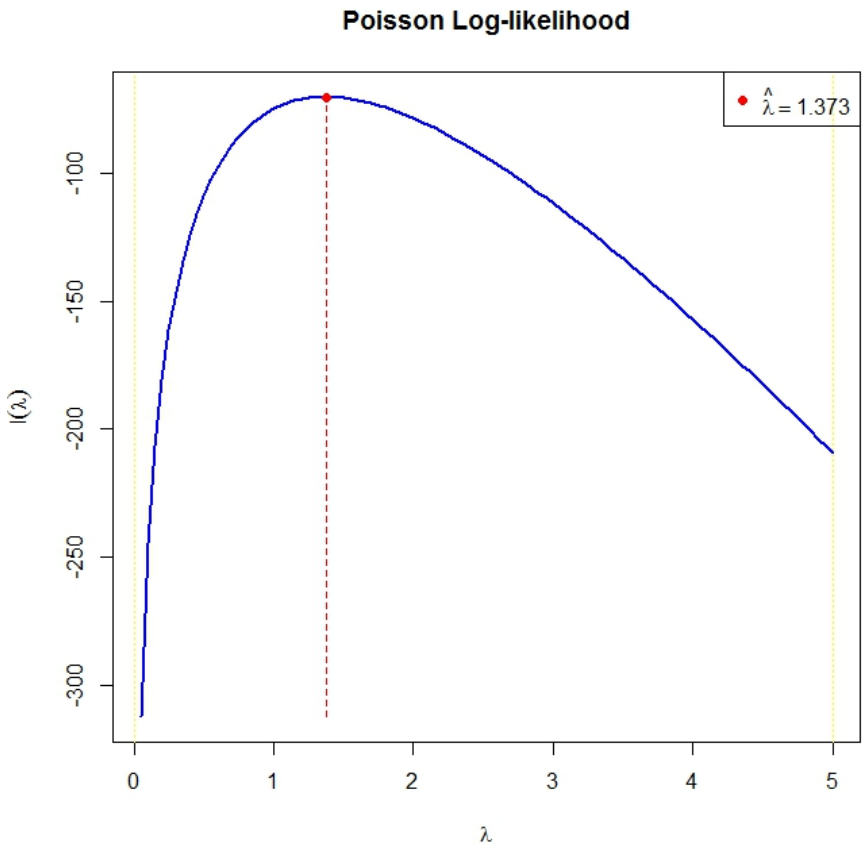
Then, we take the first derivative of $\ell(\lambda)$ with respect of λ in order to find the maximum of this function (where the derivative is equal to zero)

$$\frac{d}{d\lambda}\ell(\lambda) = -n + \frac{n\bar{x}}{\lambda} = 0. \tag{4}$$

By solving ecuation (4), we get that the MLE of λ is $\hat{\lambda} = \bar{x}$.

(b) *If the data are the values reported in PoissonCounts.xls, calculate the MLE for λ and plot the log-likelihood function from $\lambda = 0.1$ to $\lambda = 5$.*

Using the expression (4) and the dataset provided in PoissonCounts.xls, the corresponding log-likelihood function $\ell(\lambda)$ is shown in the next figure. The MLE is indicated with a red point over the curve and it's value is $\hat{\lambda} = 1.373$.



2. *Analyze the data contained in the "simple birds" tab of "Jack.xls". Each bird will have a series of events (where s/he flies to a location with a mysterious place name like "MTS" or "GREAT BEYOND"). First calculate the wait time between successive events for each bird. If a bird has five events, it will have four time intervals.*

   (a) *If the wait-times are independent (within and among birds) and governed by a common exponential distribution with parameter β, obtain the MLE for β.*

   Let T be a random variable that represents the wait time between successive events. If T follows an exponential distribution with parameter β, its density function is given by

$$f(t;\beta) = \beta e^{-\beta t}, t > 0. \tag{5}$$

   If our data are the values $t_1, t_2, ..., t_n$ (wait-times within and among birds, n in total) and they are independent, then the likelihood function for β is

$$\mathcal{L}(\beta) \propto \prod_{i=1}^{n} f(t_i;\beta) = \prod_{i=1}^{n} \beta e^{-\beta t_i} = \beta^n e^{-\beta \sum_{i=1}^{n} t_i} = \beta^n e^{-n\beta \bar{t}}, \tag{6}$$

   where $\bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i$. And, the log-likelihood funtion for β has the following expression

$$\ell(\beta) = n\log(\beta) - n\beta\bar{t}. \tag{7}$$

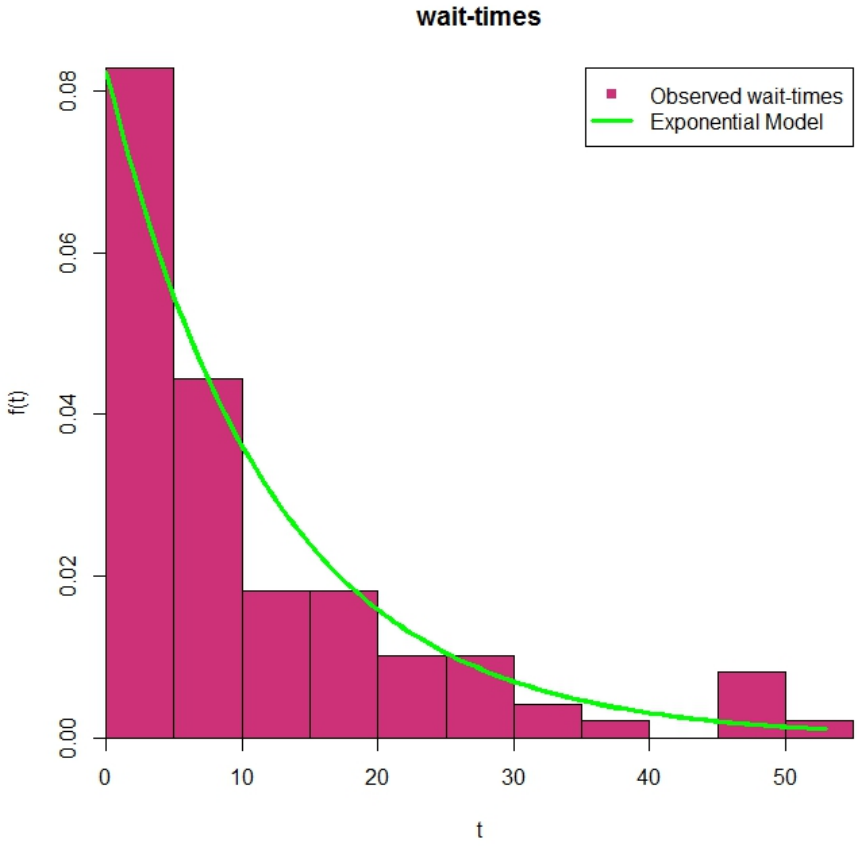
Now, in order to find the MLE, we need to solve the ecuation

$$\frac{d}{d\beta}\ell(\beta) = \frac{n}{\beta} - n\bar{t} = 0; \tag{8}$$

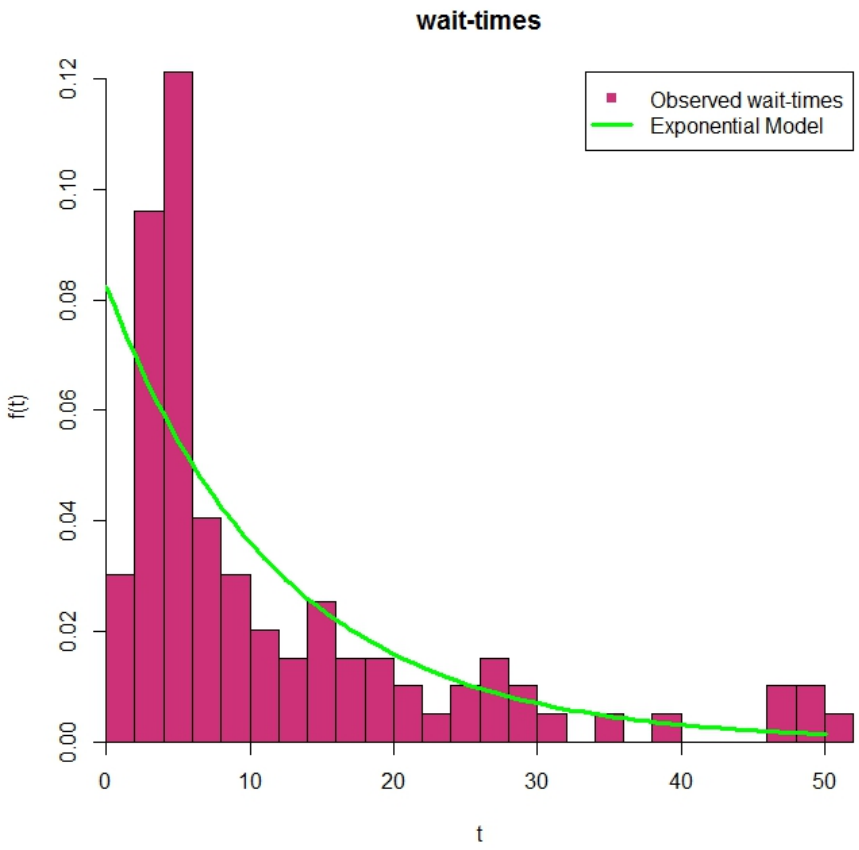
whose solution is $\hat{\beta} = \frac{1}{\bar{t}}$.

(b) *Compare the observed distribution of wait-times (pooling across birds) to the distribution predicted by your model. What do you conclude?*

Our dataset consists in time measurements of 20 different individuals, and for each individual we have a different number of observations. Once we calculate the corresponding wait-times, we end with $n = 99$ observations. For this observed values the MLE is $\hat{\beta} = 0.0824$. The next figure shows the histogram of the data together with the density function that we get after our likelihood estimation (green curve), that's to say the expression given in (5) for the particular case where $\beta = \hat{\beta}$. From the figure we



can see that the model seems to fit the data very well for small values of t; however, there are more large values of t than the model predicts. Moreover, if we change the histogram's bandwidth, we can also see the lack of fitness for small values of t (see next figure). Then, the exponential model doesn't seems to be the best model to represent this kind of data.



3. *Consider a bi-variate probability distribution $P[X, Y]$. The sample space is all positive integers for X and Y. The distribution function:*

$$P[X = i, Y = j] = (\frac{1}{2})^{i+j}, \forall i, j \in \mathbb{N}. \tag{9}$$

Find the probability that $X + Y \leq 3$.

Before calculating the probability of interest, lets calculate the density function of the random variable $X + Y$ whose sample space is $\{2, 3, 4, ...\}$. We are going to apply the law of total probability to express $P[X + Y = z]$ in terms of the random vector (X, Y) as follows

$$P[X + Y = z] = \sum_{j=1}^{\infty} P[X + Y = z, Y = j] = \sum_{j=1}^{\infty} P[X = z - j, Y = j]. \tag{10}$$

Using the joint distribution of (X, Y), (10) can be simplified in this way

$$\begin{aligned} P[X+Y=z] &= \sum_{j=1}^{\infty} P[X=z-j, Y=j] \\ &= \sum_{j=1}^{z-1} (\frac{1}{2})^{(z-j)+j} \\ &= \sum_{j=1}^{z-1} (\frac{1}{2})^{z} \\ &= (z-1)(\frac{1}{2})^{z}, z \in \{2,3,...\}. \end{aligned} \tag{11}$$

Now we can calculate the probability of interest by using (11)

$$P[X+Y \leq 3] = \sum_{z=2}^{3} (z-1)(\frac{1}{2})^{z} = \frac{1}{4} + 2(\frac{1}{8}) = \frac{1}{2}. \tag{12}$$