Tutorial 4: Solutions

February 7, 2018

Finding a MLE: Poisson

Suppose that the number of Legionella bacteria in a 1 litre sample of water follows a Poisson distribution with unknown parameter λ. Given a random sample $X_1, X_2, \ldots, X_n$

(a) Derive the MLE of λ. Is it biased or unbiased?

The pdf of the Poisson distribution is:

$$f(x;\lambda) = \begin{cases} \dfrac{e^{-\lambda}\lambda^x}{x!} & x = 0, 1, 2, \ldots, \qquad \lambda > 0 \\ 0 & \text{otherwise} \end{cases}$$

Our likelihood function is:

$$\mathcal{L}(\lambda \mid \vec{x}) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}$$

$$l(\lambda \mid \vec{x}) := \ln(\mathcal{L}(\lambda \mid \vec{x})) = -n\lambda + \sum_{i=1}^{n} x_i \ln(\lambda) - \ln\left(\prod_{i=1}^{n} x_i!\right)$$

$$\frac{dl}{d\lambda} = -n + \frac{\sum_{i=1}^{n} x_i}{\lambda}$$

$$\frac{d^2 l}{d\lambda^2} = -\frac{\sum_{i=1}^{n} x_i}{\lambda^2}$$

The MLE occurs when the first derivative equals zero:

$$\left.\frac{dl}{d\lambda}\right|_{\lambda=\widehat{\lambda}} = 0$$

$$-n + \frac{\sum_{i=1}^{n} x_i}{\widehat{\lambda}} = 0$$

$$\widehat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}$$

We notice that the second derivative is always less than zero so we don't need to plug in $\widehat{\lambda}$ to verify that it is a maximum. Therefore, the MLE for λ is $\widehat{\lambda} = \overline{x}$.

We know that if $X \sim \text{Poisson}(\lambda)$, then $\mathbf{E}(X) = \lambda$. So:

$$\begin{aligned}
\mathbf{E}\left(\overline{X}\right) &= \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}\left(X_i\right) \\
&= \frac{n\lambda}{n} \\
&= \lambda
\end{aligned}$$

Therefore, our MLE for λ is an unbiased estimator of λ.

(b) Suppose we are given the following observations:

232 225 249 233 242 203 223 229 224 230 235 217 217 192

Calculate a maximum likelihood estimate for λ.

$$\sum_{i=1}^{14} x_i = 3151, \qquad n = 14$$

$$\widehat{\lambda} = \overline{x} = \frac{3151}{14} = 225.07 \text{ Bacteria per litre}$$

Question 6.2.20, Page 273

A diagnostic test for a certain disease is applied to n individuals known to not have the disease. Let X be the number among the n test results that are positive (indicating presence of the disease, so X is the number of false positives) and p be the probability that a disease-free individual's test result is positive (i.e., p is the true proportion of test results from disease-free individuals that are positive). Assume that only X is available rather than the actual sequence of test results.

(a) Derive the maximum likelihood estimator of p. If $n = 20$ and $x = 3$, what is the estimate?

We can derive the likelihood function as a product of n independent Bernoulli trials or as a single Binomial experiment of n trials. Their likelihood functions will differ by a constant but that will not affect the maximization (see example 6.15 on page 266). I will consider a Binomial experiment for below.

$$\mathcal{L}(p \mid \vec{x}) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$l(p \mid \vec{x}) := \ln(\mathcal{L}(p \mid \vec{x})) = \ln\left(\binom{n}{x}\right) + x\ln(p) + (n-x)\ln(1-p)$$

$$\frac{dl}{dp} = \frac{x}{p} - \frac{n-x}{1-p} \qquad \text{(Don't forget about Chain Rule!)}$$

$$\frac{d^2 l}{dp^2} = -\frac{x}{p^2} - \frac{n-x}{(1-p)^2}$$

The MLE occurs when the first derivative equals zero.

$$\begin{aligned}
\left.\frac{dl}{dp}\right|_{p=\widehat{p}} &= 0 \\
\frac{x}{\widehat{p}} - \frac{n-x}{1-\widehat{p}} &= 0 \\
(1-\widehat{p})x - \widehat{p}(n-x) &= 0 \\
x - x\widehat{p} - \widehat{p}n + x\widehat{p} &= 0 \\
\widehat{p}n &= x \\
\widehat{p} &= \frac{x}{n}
\end{aligned}$$

Now we need to plug it into the second derivative to verify that it is a maximum.

$$\begin{aligned}
\left.\frac{d^2 l}{dp^2}\right|_{p=\widehat{p}} &= -\frac{x}{\widehat{p}^2} - \frac{n-x}{(1-\widehat{p})^2} \\
&= -\frac{x}{\left(\frac{x}{n}\right)^2} - \frac{n-x}{\left(1-\frac{x}{n}\right)^2} \\
&= -\frac{xn^2}{x^2} - \frac{(n-x)n^2}{(n-x)^2} \\
&= -\frac{n^2}{x} - \frac{n^2}{n-x} \\
&= -n^2\left(\frac{1}{x} + \frac{1}{n-x}\right) < 0
\end{aligned}$$

Therefore $\widehat{p}$ is a MLE of p. Now, if $n = 20$ and $x = 3$, then $\widehat{p}$ is simply $3/20$.

(b) Is the estimator of part (a) unbiased?

Recall that if $X \sim \text{Bin}(n, p)$, then $\mathbf{E}(X) = np$.

Then $\mathbf{E}(\widehat{p}) = \mathbf{E}\left(\frac{X}{n}\right) = \frac{1}{n}\mathbf{E}(X) = \frac{np}{n} = p$.

Therefore, our MLE for p is an unbiased estimator.

(c) If $n = 20$ and $x = 3$, what is the MLE of the probability $(1-p)^5$ that none of the next five tests done on disease-free individuals are positive?

Recall the invariance property of MLEs that said that if $\widehat{\theta}$ was a MLE for θ, then $g(\widehat{\theta})$ is a MLE for $g(\theta)$. Thus a MLE for $(1-p)^5$ is $(1-\widehat{p})^5 = \left(1 - \frac{3}{20}\right)^5 = \left(\frac{17}{20}\right)^5 = 0.444$.

Question 6.2.28, Page 273

Let $X_1, X_2, \dots, X_n$ represent a random sample from a Rayleigh distribution with density function:

$$f(x;\theta) = \begin{cases} \frac{x}{\theta} e^{-x^2/2\theta} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

(a) Derive the maximum likelihood estimator of θ, and then calculate the estimate for the vibratory stress data given below.

16.88	10.23	4.59	6.66	13.68
14.23	19.87	9.40	6.51	10.95

$$\mathcal{L}(\theta \mid \vec{x}) = \prod_{i=1}^{n} \frac{x_i}{\theta} \exp\left\{-{x_i}^2/2\theta\right\}$$

$$= \frac{\prod_{i=1}^{n} x_i}{\theta^n} \exp\left\{-\sum_{i=1}^{n} {x_i}^2/2\theta\right\}$$

$$l(\theta \mid \vec{x}) = \ln(\mathcal{L}(\theta \mid \vec{x})) = \ln\left(\prod_{i=1}^{n} x_i\right) - n\ln(\theta) - \frac{1}{2\theta}\sum_{i=1}^{n} {x_i}^2$$

$$\frac{dl}{d\theta} = -\frac{n}{\theta} + \frac{\sum_{i=1}^{n} {x_i}^2}{2\theta^2}$$

$$\frac{d^2 l}{d\theta^2} = \frac{n}{\theta^2} - \frac{\sum_{i=1}^{n} {x_i}^2}{\theta^3}$$

The MLE occurs when the first derivative is zero.

$$\left.\frac{dl}{d\theta}\right|_{\theta=\widehat{\theta}} = 0$$

$$-\frac{n}{\widehat{\theta}} + \frac{\sum_{i=1}^{n} {x_i}^2}{2\widehat{\theta}^2} = 0$$

$$\frac{n}{\widehat{\theta}} = \frac{\sum_{i=1}^{n} {x_i}^2}{2\widehat{\theta}^2}$$

$$2\widehat{\theta}n = \sum_{i=1}^{n} {x_i}^2$$

$$\widehat{\theta} = \frac{\sum_{i=1}^{n} {x_i}^2}{2n}$$

Now we plug our MLE into the second derivative to verify it is a maximum.

$$\left.\frac{d^2 l}{d\theta^2}\right|_{\theta=\widehat{\theta}} = \frac{n}{\widehat{\theta}^2} - \frac{\sum_{i=1}^{n} {x_i}^2}{\widehat{\theta}^3}$$

$$= \frac{n(4n^2)}{\left(\sum_{i=1}^{n} {x_i}^2\right)^2} - \frac{\left(\sum_{i=1}^{n} {x_i}^2\right) 8n^3}{\left(\sum_{i=1}^{n} {x_i}^2\right)^3}$$

$$= \frac{4n^3 - 8n^3}{\left(\sum_{i=1}^{n} {x_i}^2\right)^2}$$

$$= \frac{-4n^3}{\left(\sum_{i=1}^{n} {x_i}^2\right)^2} < 0$$

Therefore, $\widehat{\theta}$ is a MLE for θ. Using the given data, we get that $\sum {x_i}^2 = 1490.1058$ so $\widehat{\theta} = \dfrac{1490.1058}{2 \cdot 10} = 74.51$.

(b) Derive the MLE of the median of the vibratory stress distribution. [Hint: First express the median in terms of θ.]

We first need to express the median in terms of θ. Let m be the median. We want $\mathbf{P}(X \leq m) = 1/2$.

In order to get this probability, we need the CDF. To obtain this, we integrate the given PDF.

$$\int_0^x \frac{t}{\theta} \exp\left\{-t^2/2\theta\right\} dt \qquad\qquad d(-t^2/2\theta) = \frac{-2t}{2\theta}dt = \frac{-t}{\theta}dt$$

$$= -\int_0^x \exp\left\{-t^2/2\theta\right\} d(-t^2/2\theta)$$

$$= -\exp\left\{-t^2/2\theta\right\}\Big|_{t=0}^{t=x}$$

$$= 1 - e^{-x^2/2\theta}$$

$$\mathbf{P}(X \leq m) = 1 - e^{-m/2\theta} = 1/2$$

$$e^{-m^2/2\theta} = 1/2$$

$$-m^2/2\theta = \ln(1/2)$$

$$m^2/2\theta = \ln(2)$$

$$m = \sqrt{(2\ln(2))\theta}$$

Now we have expressed m in terms of θ. Using the invariance property, a MLE for m is

$$\widehat{m} = \sqrt{(2\ln(2))\widehat{\theta}} = \sqrt{\frac{2\ln(2)\sum_{i=1}^{n} {x_i}^2}{2n}} = \sqrt{\frac{\ln(2)\sum_{i=1}^{n} {x_i}^2}{n}}.$$

Question 6.S.32, Page 274

Let X_1, X_2, …, $X_n \overset{\text{iid}}{\sim} \text{Unif}(0,\, \theta)$.

(a) Show that the MLE for θ is $\widehat{\theta} = \max(X_i)$.

We first define $\mathbb{I}(\max(x_i) \leq \theta)$ as an indicator function where:

$$\mathbb{I}(\max(x_i) \leq \theta) = \begin{cases} 1 & \text{if } \max(x_1, x_2, \ldots, x_n) \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

$$\mathcal{L}(\theta \mid \vec{x}) = \frac{1}{\theta^n}\mathbb{I}(\max(x_i) \leq \theta)$$

This likelihood function takes on the value of $1/\theta^n$ if all the x_is are less than θ and takes on a value of 0 otherwise. However, we cannot actually use the classical method of taking the derivative and solving for the MLE since in this case, θ is dependent on the x_is, whereas in previous examples, the unknown parameter was independent of the sample values. We can however take the derivative of the (log) likelihood just to observe other characteristics. Assuming that our likelihood is non-zero:

$$l(\theta \mid \vec{x}) = -n\ln(\theta)$$

$$\frac{dl}{d\theta} = -\frac{n}{\theta} < 0$$

We observe that the likelihood function is a non-negative function as it is the product of pdfs, and pdfs are non-negative by definition. From the derivative, we see that the likelihood is a decreasing function with respect to θ. Lastly, we observe that the likelihood function takes on a value of zero when $\theta < \max(x_i)$. From these three observations, we deduce that $\widehat{\theta} = \max(x_i)$ will maximize the likelihood function and is therefore our MLE. (See example 6.22 on page 271 for an illustration of the likelihood function).

(b) Find the CDF and PDF for $\widehat{\theta} = \max(X_i)$. Show that the estimator in (a) is biased.

Let $F_{\max}(x)$ be the CDF of $\widehat{\theta} = \max(X_i)$. Then,

$$\begin{aligned}
F_{\max}(x) &= \mathbf{P}\left(\max(X_i) \leq x\right) \\
&= \mathbf{P}\left(X_1 \leq x \text{ and } X_2 \leq x \text{ and } \ldots \text{ and } X_n \leq x\right) \\
&= \mathbf{P}\left(X_1 \leq x\right) \cdot \mathbf{P}\left(X_2 \leq x\right) \cdots \mathbf{P}\left(X_n \leq x\right) && \text{(Due to independence)} \\
&= F_1(x) \cdot F_2(x) \cdots F_n(x) && \text{(Definition of CDF)} \\
&= F(x) \cdot F(x) \cdots F(x) && \text{(Identically distributed)} \\
&= (F(x))^n \\
&= \left(\frac{x}{\theta}\right)^n = \frac{x^n}{\theta^n} \qquad 0 \leq x \leq \theta
\end{aligned}$$

We obtain the PDF by taking the derivative of the CDF:

$$\begin{aligned}
f_{\max}(x) &= \frac{d}{dx} F_{\max}(x) \\
&= \begin{cases} \dfrac{nx^{n-1}}{\theta^n} & 0 \leq x \leq \theta \\ 0 & \text{otherwise} \end{cases}
\end{aligned}$$

Now we can verify that our MLE from (a) is biased by taking the expected value:

$$\begin{aligned}
\mathbf{E}\left(\max(X_i)\right) &= \int_0^\theta x \cdot \frac{nx^{n-1}}{\theta^n} dx \\
&= \frac{n}{\theta^n} \int_0^\theta x^n dx \\
&= \frac{n}{n+1} \frac{1}{\theta^n} x^{n+1} \Big|_0^\theta \\
&= \frac{n}{n+1} \frac{\theta^{n+1}}{\theta^n} \\
&= \frac{n}{n+1} \theta
\end{aligned}$$

Since this value is not identically θ, it is a biased estimator. In addition, we can see that $\dfrac{n}{n+1}$ is always less than 1 and we can say that our MLE on average will underestimate θ.

(c) Find an unbiased estimator for θ.

We saw that $\mathbf{E}\left(\max(X_i)\right)$ was $\dfrac{n}{n+1}\theta$. For an unbiased estimator, we can simply take $\widehat{\theta}_{unb}$ to be $\dfrac{n+1}{n}\max(X_i)$.

$$\begin{aligned}\mathbf{E}\left(\frac{n+1}{n}\max(X_i)\right) &= \frac{n+1}{n}\cdot\mathbf{E}\left(\max(X_i)\right)\\ &= \frac{n+1}{n}\cdot\frac{n}{n+1}\cdot\theta\\ &= \theta\end{aligned}$$