# CSE 312: Foundations of Computing II

## Section 8: MLE, MoM, Beta Solutions

## 1. 312 Grades

Suppose Professor Alex loses everyones grades for 312 and decides to make it up by assigning grades randomly according to the following probability distribution, and hoping the $n$ students wont notice: give an A with probability $0.5$, a B with probability $\theta$, a C with probability $2\theta$, and an $F$ with probability $0.5-3\theta$. Each student is assigned a grade independently. Let $x_A$ be the number of people who received an A, $x_B$ the number of people who received a B, etc, where $x_A + x_B + x_C + x_F = n$. Find the MLE for $\theta$.

**Solution:**

The data tells us, for each student in the class, what their grade was. We begin by computing the likelihood of seeing the given data given our parameter $\theta$. Because each student is assigned a grade independently, the likelihood is equal to the product over students of the chance they got the particular grade they got, which gives us:

$$L(x|\theta) = 0.5^{x_A}\theta^{x_B}(2\theta)^{x_C}(0.5-3\theta)^{x_F}$$

From there, we just use the MLE process to get the log-likelihood, take the first derivative, set it equal to 0, and solve for $\hat{\theta}$.

$$\ln L(x|\theta) = x_A \ln(0.5) + x_B \ln(\theta) + x_C \ln(2\theta) + x_F \ln(0.5-3\theta)$$

$$\frac{\partial}{\partial\theta} \ln L(x|\theta) = \frac{x_B}{\theta} + \frac{x_C}{\theta} - \frac{3x_F}{0.5-3\theta} = 0$$

Solving yields $\hat{\theta} = \frac{x_B+x_C}{6(x_B+x_C+x_F)}$.

## 2. A Red Poisson

Suppose that $x_1, \ldots, x_n$ are i.i.d. samples from a Poisson($\theta$) random variable, where $\theta$ is unknown. Find the MLE of $\theta$.

**Solution:**

Because each Poisson RV is i.i.d., the likelihood of seeing that data is just the PMF of the Poisson distribution multiplied together for every $x_i$. From there, take the log-likelihood, then the first derivative, set it equal to 0 and solve for for $\hat{\theta}$.

$$\begin{aligned}
L\left(x_1, \ldots, x_n \mid \theta\right) &= \prod_{i=1}^{n} e^{-\theta}\frac{\theta^{x_i}}{x_i!} \\
\ln L\left(x_1, \ldots, x_n \mid \theta\right) &= \sum_{i=1}^{n}\left[-\theta - \ln(x_i!) + x_i \ln(\theta)\right] \\
\frac{\partial}{\partial\theta} \ln L\left(x_1, \ldots, x_n \mid \theta\right) &= \sum_{i=1}^{n}\left[-1 + \frac{x_i}{\theta}\right] = 0 \\
-n + \frac{\Sigma_{i=1}^{n} x_i}{\hat{\theta}} &= 0 \\
\hat{\theta} = \frac{\Sigma_{i=1}^{n} x_i}{n}
\end{aligned}$$

## 3. Independent Shreds, You Say?

You are given 100 independent samples $x_1, x_2, \ldots, x_{100}$ from Bernoulli($\theta$), where $\theta$ is unknown. (Each sample is either a 0 or a 1). These 100 samples sum to 30. You would like to estimate the distribution's parameter $\theta$. Give all answers to 3 significant digits. What is the maximum likelihood estimator $\hat{\theta}$ of $\theta$?

**Solution:**

Note that $\Sigma_{i \in [n]} x_i = 30$, as given in the problem spec. Therefore, there are 30 **1**s and 70 **0**s. (Note that they come in some specific order.) Therefore, we can setup $L$ as follows, because there is a $\theta$ chance of getting a 1, and a $(1-\theta)$ chance of getting a 0 and they are each i.i.d. From there, take the log-likelihood, then the first derivative, set it equal to 0 and solve for for $\hat{\theta}$.

$$\begin{aligned} L(x_1, \ldots, x_n \mid \theta) &= (1-\theta)^{70}\theta^{30} \\ \ln L(x_1, \ldots, x_n \mid \theta) &= 70 \ln(1-\theta) + 30 \ln \theta \\ \frac{\partial}{\partial \theta} \ln L(x_1, \ldots, x_n \mid \theta) &= -\frac{70}{1-\theta} + \frac{30}{\theta} = 0 \\ \frac{30}{\hat{\theta}} &= \frac{70}{1-\hat{\theta}} \\ 30 - 30\hat{\theta} &= 70\hat{\theta} \\ \hat{\theta} &= \frac{30}{100} \end{aligned}$$

## 4. Y Me?

Let $y_1, y_2, ... y_n$ be i.i.d. samples of a random variable with density function

$$f_Y(y|\theta) = \frac{1}{2\theta} \exp(-\frac{|y|}{\theta})$$

.

Find the MLE for $\theta$ in terms of $|y_i|$ and $n$.

**Solution:**

Since the samples are i.i.d., the likelihood of seeing n samples of them is just their PDFs multiplied together. From there, take the log-likelihood, then the first derivative, set it equal to 0 and solve for for $\hat{\theta}$.

$$\begin{aligned} L(y_1, \ldots, y_n \mid \theta) &= \prod_{i=1}^{n} \frac{1}{2\theta} \exp(-\frac{|y_i|}{\theta}) \\ \ln L(y_1, \ldots, y_n \mid \theta) &= \sum_{i=1}^{n} \left[ -\ln 2 - \ln \theta - \frac{|y_i|}{\theta} \right] \\ \frac{\partial}{\partial \theta} \ln L(y_1, \ldots, y_n \mid \theta) &= \sum_{i=1}^{n} \left[ -\frac{1}{\theta} + \frac{|y_i|}{\theta^2} \right] = 0 \\ -\frac{n}{\hat{\theta}} + \frac{\Sigma_{i=1}^{n} |y_i|}{\hat{\theta}^2} &= 0 \\ \hat{\theta} = \frac{\Sigma_{i=1}^{n} |y_i|}{n} \end{aligned}$$

## 5. Pareto

The Pareto distribution was discovered by Vilfredo Pareto and is used in a wide array of fields but particularly social sciences and economics. It is a density function with a slowly decaying tail, for example it can describe the wealth distribution (a small group at the top holds most of the wealth). The PDF is given by:

$$f_X(x; m, \alpha) = \frac{\alpha m^{\alpha}}{x^{\alpha+1}}$$

where $x \geq m$ and real $\alpha, m > 0$. $m$ describes the minimum value that $X$ takes on (scale) and $\alpha$ is the shape. So the range of $X$ is $\Omega_X = [m, \infty)$. Assume that $m$ is given and that $x_1, x_2, \ldots, x_n$ are i.i.d. samples from the Pareto distribution. Find the MLE estimation of $\alpha$.

**Solution:**

We first need to solve for the likelihood function for which we have:

$$L(x_1, \ldots, x_n; \alpha) = \prod_{i=1}^{n} \frac{\alpha m^\alpha}{x_i^{\alpha+1}}$$

So, for the log-likelihood function we have:

$$\begin{aligned}
l(\alpha) &= \sum_{i=1}^{n} (\ln(\frac{\alpha m^\alpha}{x_i^{\alpha+1}})) \\
&= \sum_{i=1}^{n} (\ln(\alpha m^\alpha) - \ln(x_i^{\alpha+1})) \\
&= \sum_{i=1}^{n} (\ln(\alpha) + \alpha \ln(m) - (\alpha+1)\ln(x_i)) \\
&= n\ln(\alpha) + n\alpha \ln(m) - (\alpha+1)\sum_{i=1}^{n} \ln(x_i)
\end{aligned}$$

So, for the derivative with respect to $\alpha$ we have:

$$\frac{\partial l(\alpha)}{\partial \alpha} = \frac{n}{\alpha} + n\ln(m) - \sum_{i=1}^{n} \ln(x_i)$$

And then by setting to zero we get:

$$\begin{aligned}
\frac{n}{\hat{\alpha}} + n\ln(m) - \sum_{i=1}^{n} \ln(x_i) = 0 \\
\frac{n}{\hat{\alpha}} &= \sum_{i=1}^{n} \ln(x_i) - n\ln(m) \\
\hat{\alpha} &= \frac{n}{\sum_{i=1}^{n} \ln(x_i) - n\ln(m)} \\
&= \frac{1}{\frac{1}{n}\sum_{i=1}^{n} \ln(x_i) - \ln(m)} \\
&= \frac{1}{\frac{1}{n}\sum_{i=1}^{n} \ln(x_i) - \ln(m)} \\
&= \frac{1}{\overline{\ln(x)} - \ln(m)}
\end{aligned}$$

Now, let's (optionally) do a second derivative test to prove this is in fact a maximum. We have:

$$\frac{\partial^2 l(\alpha)}{\partial \alpha^2} = -\frac{n}{\alpha^2} < 0$$

So this is a maximum!

## 6. MOM Practice

Let $X_1, ..., X_n$ be a random sample from the distribution with PDF $f_X(x \mid \theta) = (\theta^2 + \theta)x^{\theta-1}(1-x)$ for $0 < x < 1$ and $\theta > 0$. What is the MOM estimator for $\theta$?

**Solution:**

First, we need to determine the first moment of X, E[X]:

$$\begin{aligned}
E[X] &= \int_0^1 x(\theta^2 + \theta)x^{\theta-1}(1-x)\,dx \\
&= \int_0^1 (\theta^2 + \theta)x^{\theta}(1-x)dx \\
&= (\theta^2 + \theta)\int_0^1 x^{\theta} - x^{\theta+1}dx \\
&= (\theta^2 + \theta)\left[\frac{x^{\theta+1}}{\theta+1} - \frac{x^{\theta+2}}{\theta+2}\right]_0^1 \\
&= \frac{\theta(\theta+1)}{(\theta+1)(\theta+2)} \\
&= \frac{\theta}{\theta+2}
\end{aligned}$$

We then set the first true moment to the first sample moment as follows:

$$\frac{\theta}{\theta+2} = \bar{x}$$

Solving for $\theta$, we get

$$\theta = (\theta + 2)\bar{x}$$

$$\theta - \theta\bar{x} = 2\bar{x}$$

$$\theta = \frac{2\bar{x}}{1-\bar{x}}$$

(Notice, though, that the original PDF looks a lot like the beta distribution PDF.
In fact, $X \sim Beta(\alpha, \beta)$ with $\alpha = \theta$ and $\beta = 2$, for which we know $E[X] = \frac{\alpha}{\alpha+\beta} = \frac{\theta}{\theta+2}$.)

## 7. Laplace

Suppose $x_1, \ldots, x_{2n}$ are iid realizations from the Laplace density (double exponential density)

$$f_X(x \mid \theta) = \frac{1}{2}e^{-|x-\theta|}$$

Find the MLE for $\theta$. For this problem, you need not verify that the MLE is indeed a maximizer. You may find the **sign** function useful:

$$\operatorname{sgn}(x) = \begin{cases} +1, & x > 0 \\ -1, & x < 0 \end{cases}$$

(in our case undefined at 0)

**Solution:**

$$
\begin{aligned}
L(x_1, \ldots, x_{2n} \mid \theta) &= \prod_{i=1}^{2n} \frac{1}{2} e^{-|x_i - \theta|} \\
\ln L(x_1, \ldots, x_{2n} \mid \theta) &= \sum_{i=1}^{2n} \left[-\ln 2 - |x_i - \theta|\right] \\
\frac{\partial}{\partial \theta} \ln L(x_1, \ldots, x_{2n} \mid \theta) &= \sum_{i=1}^{2n} \operatorname{sgn}(x_i - \theta) = 0 \\
\hat{\theta} &= \text{any value in } \left[x_n', x_{n+1}'\right]
\end{aligned}
$$

where $x_i'$ is the $i^{\text{th}}$ order statistic: the $i^{\text{th}}$ smallest observation.
Intuitively (ignoring the edge cases) this is because if $\theta \in [x_n', x_{n+1}']$, for $i \in \{1, \ldots, n\}$, $\operatorname{sgn}(x_i' - \theta) = -1$ and for $i \in \{n+1, \ldots, 2n\}$, $\operatorname{sgn}(x_i' - \theta) = 1$. So the sum of these will be zero.
If you want to argue that this is a global maximizer, note that the log likelihood is the sum of concave functions (negative absolute value), so every critical point is a global maximizer.
However, if you want to argue this more rigorously considering edge cases, we need to show that it is a maximizer, but the second derivative test is inconclusive because the second derivative is $0$ except at $x_1, x_2, \ldots, x_{2n}$, where it is undefined. We inspect the log likelihood $\ln L(x_1, \ldots, x_{2n} \mid \theta)$ directly, ignoring the constant $-\ln 2$ terms:

$$
S = -\sum_{i=1}^{2n} |x_i - \theta|
$$

If $\theta \in [x_n', x_{n+1}']$, $S = -\sum_{i=1}^{n} (x_{n+i}' - x_{n+1-i}')$. When $\theta$ crosses an endpoint of $[x_n', x_{n+1}']$, the term $|x_{n+1}' - x_n'|$ in this sum is replaced by something greater, so $S$ decreases. Therefore, the log likelihood is maximized when $\theta \in [x_n', x_{n+1}']$ . This is also why we need to include the end points in our interval.

## 8. Beta

(a) Suppose you have a coin where you have no prior belief on its true probability of heads $p$. How can you model this belief as a Beta distribution?

(b) Suppose you have a coin which you believe is fair, with strength $\alpha$. That is, pretend youve seen $\alpha$ heads and $\alpha$ tails. How can you model this belief as a Beta distribution?

(c) Now suppose you take the coin from the previous part and flip it 10 times. You see 8 heads and 2 tails. How can you model your posterior belief of the coins probability of heads?

**Solution:**

(a) Beta(1, 1) is a uniform prior, meaning that prior to seeing the experiment, all probabilities of heads are equally likely.

(b) Beta($\alpha$ + 1, $\alpha$ + 1). This is our prior belief about the distribution.

(c) Beta($\alpha$ + 9, $\alpha$ + 3).