Probability and Statistics

Y-DATA School of Data Science

P&P 4

Due: 29.11.2022

PROBLEM 1. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} Geo(\theta)$, $\theta \in [0,1]$. Find the MLE for θ.

The likelihood function is given by

$$\begin{aligned} L(\theta; X) &= \prod_{i=1}^{n} (1-\theta)^{X_i - 1}\theta \\ &= (1-\theta)^{\sum_{i=1}^{n} X_i - n}\theta^n \end{aligned}$$

The log-likelihood is then

$$\ell(\theta; X) = \left(\sum_{i=1}^{n} X_i - n\right)\log(1-\theta) + n\log(\theta)$$

To find the maximum, we need to take the derivative w.r.t. θ and equate the derivative to 0:

$$\frac{d\ell(\theta; X)}{d\theta} = -\frac{\sum_{i=1}^{n} X_i - n}{1-\theta} + \frac{n}{\theta} = 0$$

The solution is then,

$$\hat{\theta}_{MLE} = \frac{1}{\bar{X}_n}$$

PROBLEM 2. Let $X_1, \ldots, X_n$ be an i.i.d. sample from the density

$$f_\theta(x) = \frac{1}{x} e^{-\pi(\log(x)-\theta)^2}$$

Compute the MLE for θ.

The likelihood function is

$$\begin{aligned} L(\theta; X) &= \prod_{i=1}^{n} \frac{1}{X_i} e^{-\pi(\log(X_i)-\theta)^2} \\ &= \frac{1}{\prod_{i=1}^{n} X_i} e^{-\pi\sum_{i=1}^{n}(\log(X_i)-\theta)^2} \end{aligned}$$

The log-likelihood is then

$$\ell(\theta; X) = -\log\left(\prod_{i=1}^{n} X_i\right) - \pi\sum_{i=1}^{n}(\log(X_i)-\theta)^2$$

and the derivative is

$$\frac{d\ell(\theta; X)}{d\theta} = 2\pi\sum_{i=1}^{n}(\log(X_i)-\theta)$$

Equating to 0 and solving for θ, we get that

$$\hat{\theta}_{MLE} = \frac{1}{n}\sum_{i=1}^{n}\log(X_i)$$

PROBLEM 3. Let $X_1,\ldots,X_n \overset{\text{i.i.d.}}{\sim} N(\mu,\sigma^2)$, where μ and σ^2 are the unknown parameters. Find the MLE of μ and σ^2.

Denote for convenience $\theta = (\mu,\sigma^2)$. The likelihood function is

$$L(\theta;X) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(X_i-\mu)^2}{2\sigma^2}} = \left(2\pi\sigma^2\right)^{-n/2}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2}$$

and the log-likelihood is

$$\ell(\theta;X) = -\frac{n}{2}log(2\pi) - \frac{n}{2}log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$$

Now, in order to maximize the log-likelihood w.r.t. μ and σ^2, we need to differentiate once with respect to each parameter, equate to zero these equations and solve them. To this end,

$$\frac{\partial\ell(\theta;X)}{\partial\mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu) = 0 \Rightarrow \hat{\mu}_{MLE} = \bar{X}_n$$

and

$$\frac{\partial\ell(\theta;X)}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(X_i-\mu)^2 = 0 \Rightarrow \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$$

but since $\hat{\mu}_{MLE} = \bar{X}_n$,

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X}_n)^2$$

PROBLEM 4. Let $X_1,\ldots,X_n \overset{\text{i.i.d.}}{\sim} U(\theta+2,\theta+10)$ (continuous).

(1) Find $\hat{\theta}_{MOM}$ (method of moments estimator for θ).
(2) Evaluate $\hat{\theta}_{MOM}$ for the sample

$$12.3, 17.5, 15.1, 14.7$$

(1) The MOM estimator is obtained by equating the theoretical mean (expected value) to the empirical mean and solving the equation for θ. In our case we need to solve the equation:

$$E_\theta(X_1) = \bar{X}_n$$

that is,

$$\frac{\theta+2+\theta+10}{2} = \bar{X}_n$$

Therefore,

$$\hat{\theta}_{MOM} = \bar{X}_n - 6$$

(2)

$$\hat{\theta}_{MOM} = \frac{1}{4}(12.3+17.5+15.1+14.7) - 6 = 8.9$$

PROBLEM 5. It is assumed that the daily amount of rain (in mm) that falls in London during January is distributed $N(\mu, 25)$. We are interested in estimating $P(X > 75)$. Two approaches were suggested:

A Estimate μ using the method of moments, and then estimate the probability using $\hat{\mu}_{MOM}$ instead of μ in the normal distribution.

B Don't assume normality. Estimate the probability by calculating the proportion of observations that are greater than 75.

In a random sample of 10 observations, the following results were received:

$$68.49, 63.61, 71.22, 76.38, 75.99, 78.66, 59.08, 68.82, 75.47, 64.56$$

(1) Estimate the required probability using both methods and compare the results.
(2) Estimate the probability $P(X > 72)$ using both methods and compare the results.

(1) According to the first approach we have

$$\hat{\mu}_{MOM} = \bar{X}_n$$

which in our case is $\bar{X}_n = 70.228$. Thus,

$$\hat{P}_A(X > 75) = 1 - \hat{P}_A(X \leq 75) = 1 - \Phi\left(\frac{75 - 70.228}{5}\right) = 0.17$$

Using the second approach, we get

$$\hat{P}_B(X > 75) = \frac{\#(X_i > 75)}{n} = \frac{4}{10} = 0.4$$

The difference between the estimators is quite large. We would probably need more observations to get precise results.

(2) Similarly to the first part,

$$\hat{P}_A(X > 72) = 1 - \hat{P}_A(X \leq 72) = 1 - \Phi\left(\frac{72 - 70.228}{5}\right) = 0.3615$$

and

$$\hat{P}_B(X > 72) = \frac{\#(X_i > 72)}{n} = \frac{4}{10} = 0.4$$

In this case, we see that the results are not as different as in the first part, but we also observe the lack of sensitivity of method B compared to method A. This, again, could be due to the small number of observations.

PROBLEM 6. Let $X_1, \ldots X_n \overset{\text{i.i.d.}}{\sim} Pois(\lambda)$.

(1) Compute the MSE of the MLE for λ.
(2) A researcher believes that λ is approximately 3, so he suggests to use the estimator which is the average between the MLE and 3: $T = \frac{\bar{X}_n + 3}{2}$. Compute the MSE of T.
(3) Compare the bias and the variance of the estimators as functions of λ.
(4) Compare the MSE of the estimators as a function of λ and find for which values of λ each estimator is better than the other. Note that the range of λ might depend on n.

(1) As proved in class, the MLE of λ in this case is $\hat{\lambda} = \bar{X}_n$. Recall that the formula for the MSE is given by

$$MSE(\hat{\lambda},\lambda) = E_\lambda\left(\hat{\lambda}-\lambda\right)^2 = Var(\hat{\lambda}) + \left(E(\hat{\lambda})-\lambda\right)^2$$

Since $E(\bar{X}_n) = E(X_1) = \lambda$, the bias is 0. It is left to compute the variance:

$$Var(\bar{X}_n) = \frac{1}{n}Var(X_1) = \frac{\lambda}{n}$$

To sum up,

$$MSE(\hat{\lambda},\lambda) = \frac{\lambda}{n}$$

(2) Again, we will use the bias-variance decomposition of the MSE.

$$E(T) = \frac{1}{2}(E(\bar{X}_n)+3) = \frac{1}{2}(\lambda+3)$$

where the first equality follows from linearity of expectation. The bias is then,

$$b(T,\lambda) = \frac{1}{2}(\lambda+3)-\lambda = 1.5-\lambda/2$$

The variance is,

$$Var(T) = \frac{1}{4}Var(\bar{X}_n) = \frac{\lambda}{4n}$$

Overall, the MSE of T is

$$MSE(T,\lambda) = \frac{\lambda}{4n} + (1.5-\lambda/2)^2$$

(3) The MLE $\hat{\lambda}$ has lower bias (in fact, no bias at all) than T. However, the variance of T is lower than the variance of $\hat{\lambda}$, as $Var(T) = Var(\hat{\lambda})/4$.

(4) We would prefer T over $\hat{\lambda}$ whenever

$$MSE(T,\lambda) < MSE(\hat{\lambda},\lambda)$$

That is,

$$\begin{aligned}
&\iff \quad \frac{\lambda}{4n} + \left(\frac{3-\lambda}{2}\right)^2 < \frac{\lambda}{n} \\
&\iff 0 < \frac{3\lambda}{4n} - \frac{9-6\lambda+\lambda^2}{4} \\
&\iff 0 < \frac{3\lambda-9n+6n\lambda-n\lambda^2}{4n} \\
&\iff 0 > n\lambda^2 - \lambda(3+6\text{n})+9n \\
&\iff \quad \lambda_{1,2} = \frac{3+6n\pm\sqrt{9+36n}}{2n}
\end{aligned}$$

Since this is a convex parabola, it will be negative for

$$\frac{3+6n-\sqrt{9+36n}}{2n} < \lambda < \frac{3+6n+\sqrt{9+36n}}{2n}$$

i.e. for the above values of λ we would prefer T over $\hat{\lambda}$. This makes sense, as these are values around 3. So, when the true (unknown) parameter is close to 3, the estimator that incorporates this belief is better.

PROBLEM 7. The weight of students in some university is normally distributed. A sample of 12 students is drawn with the following results (in Kg):

$$53.8, 67.34, 51.7, 52, 58.9, 74, 45.3, 53, 62.5, 48.87, 49, 55.6$$

(1) Assuming that the variance is known and equals 1.5 Kg, calculate the confidence interval for the expected value of the weight with confidence level 95%.
(2) Repeat part 1, this time for a confidence level of 90%. What can you say about the difference between the results?
(3) Assuming that the variance is known and equals 2 Kg, calculate the confidence interval for the expected value of the weight with confidence level 95%. What can you conclude from the result?
(4) Repeat part 1, assuming that the variance is unknown.

(1) We got from our sample that $\bar{X}_n = 56$. Since we have a normal sample with known variance, the CI for μ is

$$\left[\bar{X}_n - z_{0.975}\frac{\sigma}{\sqrt{n}}, \bar{X}_n + z_{0.975}\frac{\sigma}{\sqrt{n}}\right] = \left[56 - 1.96\frac{\sqrt{1.5}}{\sqrt{12}}, 56 + 1.96\frac{\sqrt{1.5}}{\sqrt{12}}\right]$$
$$= [55.3, 56.69]$$

(2) In this case, $\alpha = 0.1$ and

$$\left[\bar{X}_n - z_{0.95}\frac{\sigma}{\sqrt{n}}, \bar{X}_n + z_{0.95}\frac{\sigma}{\sqrt{n}}\right] = \left[56 - 1.645\frac{\sqrt{1.5}}{\sqrt{12}}, 56 + 1.645\frac{\sqrt{1.5}}{\sqrt{12}}\right]$$
$$= [55.42, 56.58]$$

We see that when we decrease the confidence level, we get a shorter interval.

(3) In this case, $\sigma^2 = 2$ and the confidence interval is

$$\left[\bar{X}_n - z_{0.975}\frac{\sigma}{\sqrt{n}}, \bar{X}_n + z_{0.975}\frac{\sigma}{\sqrt{n}}\right] = \left[56 - 1.96\frac{\sqrt{2}}{\sqrt{12}}, 56 + 1.96\frac{\sqrt{2}}{\sqrt{12}}\right]$$
$$= [55.19, 56.8]$$

As expected, we got a larger interval when increasing the variance.

(4) In this case, the CI is

$$\left[\bar{X}_n - t_{0.975,11}\frac{S_n}{\sqrt{n}}, \bar{X}_n + t_{0.975,11}\frac{S_n}{\sqrt{n}}\right]$$

where

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

So,

$$\left[\bar{X}_n - t_{0.975,11}\frac{S_n}{\sqrt{n}}, \bar{X}_n + t_{0.975,11}\frac{S_n}{\sqrt{n}}\right] = \left[56 - 2.2\frac{\sqrt{69.64}}{\sqrt{12}}, 56 + 2.2\frac{\sqrt{69.64}}{\sqrt{12}}\right]$$
$$= [50.69, 61.3]$$

Of course, the interval is larger.

PROBLEM 8. Let $X \sim N(\mu, \sigma^2)$ (both parameters are unknown). In a random sample of 10 observations we received that

$$\sum_{i=1}^{n} x_i = 15, \sum_{i=1}^{n} x_i^2 = 27$$

and the CI for μ is $[1.09, 1.91]$. What is the confidence level of this confidence interval?

The given CI under this scenario in calculated using the formula

$$\left[\bar{X}_n \pm t_{1-\alpha/2,n-1}\frac{S_n}{\sqrt{n}}\right]$$

To find the confidence level, we need to equate one of the endpoints of the interval to the corresponding formula and solve for α. From the data at hand we know that $\bar{X}_n = 1.5$ and that

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n} X_i - \frac{n}{n-1}\bar{X}_n^2 = \frac{27}{9} - \frac{10}{9}1.5^2 = 0.5$$

Now,

$$1.09 = \bar{X}_n - t_{1-\alpha/2,n-1}\frac{S_n}{\sqrt{n}} \iff 1.09 = 1.5 - t_{1-\alpha/2,9}\frac{\sqrt{0.5}}{\sqrt{10}}$$

$\Rightarrow$

$$t_{1-\alpha/2,9} = 1.833$$

This means that

$$P(T_{n-1} \leq 1.833) = 1 - \alpha/2$$

so we need to calculate this probability (use R or Python), which in our case is 0.95. Thus,

$$0.95 = 1 - \alpha/2 \Rightarrow \alpha = 0.1$$

Equivalently, we can say that the confidence level is 90%.

PROBLEM 9. In a random sample of 100 students, it was found that 30 like Bamba.

(1) Compute an asymptotic confidence interval for the proportion of Bamba lovers among the students.
(2) Find the minimal sample size n for which the length of the CI will be at most 0.02.

(1) We have $X_1, ..., X_{100} \overset{\text{i.i.d.}}{\sim} Ber(p)$ where p is the probability to like Bamba, and we need to build a 95% asymptotic CI for p. The suitable formula in this case is

$$\left[\hat{p} \pm z_{1-\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right] = \left[0.3 \pm 1.96\frac{\sqrt{0.3 \cdot 0.7}}{\sqrt{100}}\right] = [0.21, 0.389]$$

(2) We need to solve for n the following inequality,

$$2 \cdot 1.96 \cdot \frac{\sqrt{0.21}}{\sqrt{n}} \leq 0.02$$

and we get that

$$n \geq 8068$$