3. Another Example of Maximum Likelihood Estimator

MLE for a Loaded Die: Likelihood

1/1 point (graded)

You have a loaded (i.e. possibly unfair) six-sided die with the probability that it shows a "3" equal to η^* and the probability that it shows any other number equal to $(1-\eta^*)/5$.

Let X be a random variable representing a roll of this die. You roll this die n times, and record your data set, consisting of the values of the faces as $X_1, X_2, X_3, \ldots X_n$.

Let the outcome of a set of n rolls of the die be modeled by the i.i.d. random variable sequence $(X_1, \ldots, X_n)$. We model the i'th roll as X_i where $X_i = j$ if the top face of the die shows a "j".

You roll the die n times and you see k "3"s. What is the likelihood function $L_n(x_1, \ldots, x_n, \eta)$?

(Enter **eta** for η.)

eta^k*((1-eta)/5)^(n-k)

Answer: eta^k*((1-eta)/5)^(n-k)

$\eta^k \cdot \left(\frac{1-\eta}{5}\right)^{n-k}$

STANDARD NOTATION

Solution:

Denote by $p_\eta(x)$ the pmf of X_i. Then, the likelihood function is

$$\begin{aligned} L_n(x_1, \ldots, x_n, \eta) &= \prod_{i=1}^{n} p_\eta(x_i) \\ &= \eta^k \left(\frac{1-\eta}{5}\right)^{n-k}. \end{aligned}$$

提交 你已经尝试了1次（总共可以尝试3次）

Answers are displayed within the problem

MLE for a Loaded Die: MLE

0/1 point (graded)

Find the ML estimator $\hat{\eta}_n^{\text{MLE}}$.

n*k

Answer: k/n

$n \cdot k$

STANDARD NOTATION

Solution:

Since we are looking for the $\mathbf{argmax}_{\eta\in[0,1]} L_n(x_1, \ldots, x_n, \eta)$, we can ignoring any scaling constant in $L_n(x_1, \ldots, x_n, \eta)$. Hence, we will maximize $\widetilde{L}_n(x_1, \ldots, x_n, \eta) = \eta^k (1-\eta)^{n-k}$.

Taking the derivative of $\widetilde{L}_n\left(x_1,\ldots,x_n,\eta\right)$ with respect to η and setting it to 0, we get

$$\begin{aligned} k\left(1-\eta\right) &= \left(n-k\right)\eta \\ \Longrightarrow\ \hat{\eta}_n^{\text{MLE}} &= \frac{k}{n}. \end{aligned}$$

Remark: The function $\widetilde{L}_n\left(x_1,\ldots,x_n,\eta\right)=\eta^k\left(1-\eta\right)^{n-k}$ whose maximizer is $\hat{\eta}_n^{\text{MLE}}$ is the same as the likelihood function for a Bernoulli experiment with parameter η, even though each roll of a die has 6 potential outcomes.

提交 你已经尝试了3次（总共可以尝试3次）

Answers are displayed within the problem

(Optional) Generalization of the Loaded Die Problem

Question : What if we try to generalize the loaded die estimation problem? Say we observe $k_i, i=1,\ldots,6$ outcomes of result i out of a total of n rolls of a loaded die with probabilities $\eta_i^*, i=1,\ldots,6$. How do we obtain the ML estimate of $\eta_i, i=1,\ldots,6$? (This problem will also be presented in Recitation 6 on *MLE for Multinomials*.)

Solution: The likelihood function for this case, denoted by $L\left(x_1,\ldots,x_n,\eta_1,\ldots,\eta_6\right)$, ignoring constant terms, can be computed as

$$L_n\left(x_1,\ldots,x_n,\eta_1,\ldots,\eta_6\right)=\prod_{i=1}^{6}\left(\eta_i\right)^{k_i}.$$

Finding out $\left\{\hat{\eta}_{i,n}^{\text{MLE}}, i=i,\ldots,6\right\}$ involves maximizing $L_n\left(x_1,\ldots,x_n,\eta_1,\ldots,\eta_6\right)=\prod_{i=1}^{6}\left(\eta_i\right)^{k_i}$ with the following two constraints: $\sum_{i=1}^{6}\eta_i=1$ and $\eta_i\geq 0, i=1,\ldots,6$.

This constrained optimization problem has an explicit solution that can be obtained by analyzing what are called the **Karush-Kuhn-Tucker (KKT) conditions** in optimization theory. For a detailed explanation of what these conditions are and how they are obtained, we refer the reader to the textbook *Convex Optimization* by Stephen Boyd and Lieven Vandenberghe (Cambridge University Press). This textbook is also available online from the authors here: http://web.stanford.edu/ boyd/cvxbook/.

First, we set up the optimization problem (call this OP1) from a minimization perspective and using the log likelihoods:

$$\begin{aligned} \min_{\eta_1,\eta_2,\ldots,\eta_6} \quad & -\sum_{i=1}^{6}k_i\ln\left(\eta_i\right) \\ \text{constraints:} \quad & \sum_{i=1}^{6}\eta_i=1,\quad \eta_i\geq 0, i=1,\ldots,6 \end{aligned}$$

In order to be precise about what we state as the relevant KKT conditions for this problem, let us introduce some additional notation:

$$\begin{aligned} \min_{\eta_1,\eta_2,\ldots,\eta_6} \quad & f_0\left(\eta_1,\ldots,\eta_6\right)\triangleq -\sum_{i=1}^{6}k_i\ln\left(\eta_i\right) \\ \text{constraints:} \quad & h\left(\eta_1,\ldots,\eta_6\right)\triangleq\sum_{i=1}^{6}\eta_i-1=0, \\ & f_i\left(\eta_i\right)\triangleq -\eta_i\leq 0, i=1,\ldots,6 \end{aligned}$$

Let the set of all $\left(\eta_1,\eta_2,\eta_3,\eta_4,\eta_5,\eta_6\right)$ values where we can evaluate $f_0\left(\cdot\right)$, $f_i\left(\cdot\right), i=1,\ldots,6$, and $h\left(\cdot\right)$ be called the domain $\mathcal{D}$ of the optimization problem. In this case, $\mathcal{D}=\left\{\left(\eta_1,\eta_2,\eta_3,\eta_4,\eta_5,\eta_6\right)|\eta_i\in\left(0,\infty\right), i=1,\ldots,6\right\}$.

To derive the KKT conditions, we need what is called the **Lagrange dual** problem, which is the following optimization problem (call this OP2) for this specific case:

$$\max_{\lambda_1,\lambda_2,\ldots\lambda_6,\mu} g(\lambda_1,\lambda_2,\ldots,\lambda_6,\mu) \triangleq \min_{(\eta_1,\eta_2,\eta_3,\eta_4,\eta_5,\eta_6)\in\mathcal{D}} \left[f(\eta_1,\ldots,\eta_6,\lambda_1,\ldots,\lambda_6,\mu) \triangleq -\sum_{i=1}^{6} k_i \ln(\eta_i) - \sum_{i=1}^{6} \lambda_i\eta_i + \mu\left(\sum_{i=1}^{6}\eta_i - 1\right)\right]$$

$$\text{constraints:} \quad \mu\in\mathbb{R},\quad \lambda_i \geq 0, i=1,\ldots,6$$

One important property of the Lagrange dual problem, in general, is that the objective function $g(\cdot)$ is concave in its arguments λ_i and μ (we have only one equality constraint in OP1, and therefore have only one μ, but in general we have λ_i's and μ_j's in the Lagrange dual).

Another important property of the Lagrange dual is that with the constraint that $\lambda_i \geq 0, \forall i$, the Lagrange dual problem, in general, provides a lower bound on the optimal value of the original optimization problem (assuming the original optimization problem is always written as a minimization problem, which is the standard form in the aforementioned textbook).

To recap, there are two optimization problems: the original minimization problem, OP1, and the Lagrange dual problem, OP2. For this specific case, it turns out that OP1 and OP2 are equivalent in the sense that minimizing $f_0(\eta_1,\ldots,\eta_6)$ with its constraints in OP1 provides the same optimal value as maximizing $g(\lambda_1,\lambda_2,\ldots,\lambda_6,\mu)$ in OP2 with its constraints. The KKT conditions when the primal (original) optimization problem and the Lagrange dual yield the same optimal value are as follows (specialized for this problem).

Let $\eta_i^*, i=1,\ldots,6$ denote a set of minimizers of OP1 and let $\lambda_i^*, i=1,\ldots,6$, and μ^* denote a set of maximizers of OP2. Then, the KKT conditions are

$$\begin{aligned}
f_i(\eta_i^*) &\leq 0,\ i=1,\ldots,6 \\
h(\eta_1^*,\ldots,\eta_6^*) &= 0 \\
\lambda_i^* &\geq 0,\ i=1,\ldots,6 \\
\lambda_i^* f_i(\eta_i^*) &= 0,\ i=1,\ldots,6 \\
\frac{\mathrm{d}f_0}{\mathrm{d}\eta_i}(\eta_i^*) + \lambda_i^*\frac{\mathrm{d}f_i}{\mathrm{d}\eta_i}(\eta_i^*) + \mu^*\frac{\mathrm{d}h}{\mathrm{d}\eta_i}(\eta_i^*) &= 0,\ i=1,\ldots,6.
\end{aligned}$$

Writing out these conditions explicitly, we get

$$\eta_i^* \geq 0, i=1,\ldots,6 \qquad \sum_{i=1}^{6}\eta_i^* = 1 \qquad \lambda_i^* \geq 0, i=1,\ldots,6$$

$$\lambda_i^*\eta_i^* = 0, 1=1,\ldots,6 \qquad -\frac{k_i}{\eta_i^*} - \lambda_i^* + \mu^* = 0, i=1,\ldots,6$$

From the equations in the second line above, we can obtain

$$\eta_i^* = \frac{k_i}{\mu^*}, i=1,\ldots,6.$$

Using the equation $\sum_{i=1}^{6}\eta_i^* = 1$ and the above, we can obtain that $\mu^* = \sum_{i=1}^{6} k_i = n$. Hence,

$$\eta_i^* = \hat{\eta}_{i,n}^{\mathrm{MLE}} = \frac{k_i}{n}, i=1,\ldots,6.$$

Remark: The "i.i.d. die outcomes" with "6" sides can be replaced by any "i.i.d. discrete statistical experiment" with "ℓ" mass points and the entire derivation remains the same. The MLE solution has the property that it is the same as the frequency estimate.

Another proof of optimal values for the loaded die problem: Recall from Lecture 8 that the KL divergence between two distributions $\mathbf{P}$ and $\mathbf{Q}$ can only take on non-negative values. That is,

$$\mathrm{KL}(\mathbf{P},\mathbf{Q}) \geq 0.$$

Also,

$$\mathrm{KL}(\mathbf{P},\mathbf{Q}) = 0 \iff \mathbf{P} = \mathbf{Q}.$$

We can use these properties of KL divergence to prove that $\hat{\eta}_{i,n}^{\text{MLE}} = \frac{k_i}{n}, i = 1, \ldots, 6$.

Let a distribution $\mathbf{P}$ be defined by the pmf $p_i \triangleq \frac{k_i}{n}, i = 1, \ldots, 6$, where k_i is the number of observations of outcome i and n is the total number of rolls of the die. Let a distribution $\mathbf{Q}$ be defined by the pmf $q_i = \eta_i, i = 1, \ldots, 6$. Now, the above properties of KL divergence mean that

$$\sum_{i=1}^{6} p_i \ln(p_i) \geq \sum_{i=1}^{6} p_i \ln(q_i = \eta_i),$$

with equality if and only if $q_i = p_i = \frac{k_i}{n}$. Since the optimization problem is exactly the same as maximizing the right-hand side of the above inequality with respect to $\eta_i, i = 1, \ldots, 6$, the upper bound specified by the left-hand side is attained at $q_i = \frac{k_i}{n}, i = 1, \ldots, 6$.

Therefore, the above one-line proof (which used properties of KL divergence) shows that $\hat{\eta}_{i,n}^{\text{MLE}} = \frac{k_i}{n}, i = 1, \ldots, 6$.

Hide

讨论

显示讨论

主题： Unit 3 Methods of Estimation:Lecture 10: Consistency of MLE, Covariance Matrices, and Multivariate Statistics / 3. Another Example of Maximum Likelihood Estimator