Emory University
MATH 362 Mathematical Statistics II
Learning Notes

Jiuru Lyu

February 1, 2024

Contents

1 Estimation

1.1 Introduction

Definition 1.1.1 (Model). A *model* is a distribution with certain parameters.

Example 1.1.2 The normal distribution: $N(\mu, \sigma^2)$.

Definition 1.1.3 (Population). The *population* is all the objects in the experiment.

Definition 1.1.4 (Data, Sample, and Random Sample). *Data* refers to observed value from sample. The *sample* is a subset of the population. A *random sample* is a sequence of independent, identical ($i.i.d.$) random variables.

Definition 1.1.5 (Statistics). *Statistics* refers to a function of the random sample.

Example 1.1.6 The sample mean is a function of the sample:

$$\overline{Y} = \frac{1}{n}(Y_1 + \cdots + Y_n).$$

Example 1.1.7 Central Limit Theorem

We randomly toss $n = 200$ fair coins on the table. Calculate, using the central limit theorem, the probability that at least 110 coins have turned on the same side.

$$\overline{X} = \frac{X_1 + \cdots + X_{200}}{200} \quad \overset{\text{CLT}}{\sim} \quad N\left(\mu, \sigma^2\right),$$

where

$$\mu = \mathbf{E}\left(\overline{X}\right) = \frac{\sum_{i=1}^{200} \mathbf{E}(X_i)}{200},$$

$$\sigma^2 = \mathbf{Var}\left(\overline{X}\right) = \mathbf{Var}\left(\frac{X_1 + \cdots + X_{200}}{200}\right) = \frac{\sum_{i=1}^{200} \mathbf{Var}(X_i)}{200^2}.$$

Definition 1.1.8 (Statistical Inference). The process of *statistical inference* is defined to be the process of using data from a sample to gain information about the population.

Example 1.1.9 Goals in statistical inference

1. **Definition 1.1.10 (Estimation).** To obtain values of the parameters from the data.
2. **Definition 1.1.11 (Hypothesis Testing).** To test a conjecture about the parameters.
3. **Definition 1.1.12 (Goodness of Fit).** How well does the data fit a given distribution.
4. Linear Regression

1.2 The Method of Maximum Likelihood and the Method of Moments

Example 1.2.1 Given an unfair coin, or p-coin, such that

$$X = \begin{cases} 1 & \text{head with probability } p, \\ 0 & \text{tail with probability } 1-p. \end{cases}$$

How can we determine the value p?

Solution 1.

1. Try to flip the coin several times, say, three times. Suppose we get HHT.
2. Draw a conclusion from the experiment.

Key idea: The choice of the parameter p should be the value that maximizes the probability of the sample.

$$\mathbf{P}(X_1 = 1, X_2 = 1, X_3 = 0) = \mathbf{P}(X_1 = 1)\mathbf{P}(X_2 = 1)\mathbf{P}(X_3 = 0) = p^2(1-p) := f(p).$$

Solving the optimization problem $\max\limits_{p>0} f(p)$, we find it is most likely that $p = \frac{2}{3}$. This method is called the *likelihood maximization method.* □

Definition 1.2.2 (Likelihood Function). For a random sample of size n from the discrete (or continuous) pdf $p_X(k;\theta)$ (or $f_Y(y;\theta)$), the *likelihood function*, $L(\theta)$, is the product of the pdf evaluated at $X_i = k_i$ (or $Y_i = y_i$). That is,

$$L(\theta) := \prod_{i=1}^{n} p_X(k_i;\theta) \quad \text{or} \quad L(\theta) := \prod_{i=1}^{n} f_Y(y_i;\theta).$$

Definition 1.2.3 (Maximum Likelihood Estimate). Let $L(\theta)$ be as defined in Definition 1.2.2. If θ_e is a value of the parameter such that $L(\theta_e) \geq L(\theta)$ for all possible values of θ, then we call θ_e the *maximum likelihood estimate* for θ.

Theorem 1.2.4 The Method of Maximum Likelihood
Given random samples $X_1, \ldots, X_N$ and a density function $p_X(x)$ (or $f_X(x)$), then we have the likelihood function defined as

$$\begin{aligned} L(\theta) = p_X(X; \theta) &= \mathbf{P}(X_1, X_2, \ldots, X_N) \\ &= \mathbf{P}(X_1)\mathbf{P}(X_2) \cdots \mathbf{P}(X_N) && [independent] \\ &= \prod_{i=1}^{N} p_X(X_i; \theta) && [identical] \end{aligned}$$

Then, the maximum likelihood estimate for θ is given by

$$\theta^* = \arg\max_\theta L(\theta),$$

where

$$L\Big(\arg\max_\theta L(\theta)\Big) = L^*(\theta) = \max_\theta L(\theta).$$

Example 1.2.5 Consider the Poisson distribution $X = 0, 1, \ldots,$ with $\lambda > 0$. Then, the pdf is given by

$$p_X(k, \lambda) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, \ldots$$

Given data $k_1, \ldots, k_n$, we have the likelihood function

$$L(\lambda) = \prod_{i=1}^{n} p_X(X = k; \lambda) = \prod_{i=1}^{n} e^{-\lambda} = \frac{\lambda^{k_i}}{k_i!} = e^{-n\lambda} \frac{\lambda^{\sum k_i}}{k_1! \cdots k_n!}$$

Then, to find the maximum likelihood estimate of λ, we need to $\max_\lambda L(\lambda)$. That is to solve $\frac{\partial L(\lambda)}{\partial \lambda} = 0$ and $\frac{\partial^2 L(\lambda)}{\partial \lambda^2} < 0$.

Example 1.2.6 Waiting Time.

Consider the exponential distribution $f_Y(y) = \lambda e^{-\lambda y}$ for $y \geq 0$. Find the MLE λ_e of λ.
Solution 2.

The likelihood function of the exponential distribution is given by

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda y_i} = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} y_i\right).$$

Now, define

$$\ell(\lambda) = \ln L(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^{n} y_i.$$

To optimize $\ell(\lambda)$, we compute

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\ell(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} y_i \overset{set}{=} 0$$

So,

$$\frac{n}{\lambda} = \sum_{i=1}^{n} y_i \implies \lambda_e = \frac{n}{\sum_{i=1}^{n} y_i} =: \frac{1}{\overline{y}},$$

where $\overline{y}$ is the sample mean. □

Example 1.2.7 Given the exponential distribution $f_Y(y) = \lambda e^{-\lambda y}$ for $y \geq 0$. Find the MLE of λ^2.

Solution 3.

Define $\tau = \lambda^2$. Then, $\lambda = \sqrt{\tau}$, and so

$$f_Y(y) = \sqrt{\tau} e^{-\sqrt{\tau} y}, \quad y \geq 0.$$

Then, the likelihood function becomes

$$L(\tau) = \prod_{i=1}^{n} f_Y(y) = \tau^{\frac{n}{2}} \exp\left(-\sqrt{\tau} \sum_{i=1}^{n} y_i\right).$$

Similarly, after maximization, we find

$$\tau_e = \frac{1}{(\overline{y})^2}.$$

□

Theorem 1.2.8 Invariant Property for MLE
Suppose λ_e is the MLE of λ. Define $\tau := h(\lambda)$. Then, $\tau_e = h(\lambda_e)$.

Proof 4. In this proof, we will prove the case when h is a one-to-one function. The case of h being a many-to-one function is beyond the scope of this course.

Suppose $h(\cdot)$ is a one-to-one function. Then, $\lambda = h^{-1}(\tau)$ is well-defined. Then,

$$\max_{\lambda} L(\lambda; y_1, \ldots, y_n) = \max_{\tau} L\big(h^{-1}(\tau); y_1, \ldots, y_n\big) = \max_{\tau} L(\tau; y_1, \ldots, y_n).$$

■

Example 1.2.9 Waiting Time with an unknown Threshold.

Let $\lambda = 1$ in exponential but there is an unknown threshold θ, that, is $f_Y(y) = e^{-(y-\theta)}$ for $y \geq \theta,\ \theta > 0$.

Solution 5.

Note that the likelihood function is given by

$$\begin{aligned} L(\theta; y_1, \ldots, y_n) = \prod_{i=1}^{n} f_Y(y_1) &= \exp\left(-\sum_{i=1}^{n}(y_i - \theta)\right), \quad y_i \geq \theta,\ \theta > 0 \\ &= \exp\left(-\sum_{i=1}^{n}(y_i - \theta)\right) \cdot \mathbb{1}_{[y_i \geq 0,\ \theta > 0]}, \end{aligned}$$

where

$$\mathbb{1}_{x \in A} = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A. \end{cases}$$

Using order statistics,

$$\begin{aligned} L(\theta) &= \exp\left(-\sum_{i=1}^{n}(y_i - \theta)\right) \cdot \mathbb{1}_{\left[y_{(n)} \geq y_{(n-1)} \geq \cdots \geq y_{(1)} \geq \theta,\ \theta > 0\right]} \\ &= \exp\left(-\sum_{i=1}^{n} y_i + n\theta\right) \mathbb{1}_{\left[y_{(n)} \geq \cdots \geq y_{(1)} \geq \theta,\ \theta > 0\right]}. \end{aligned}$$

So, we know $\theta \leq y_{(1)} = y_{\min}$.

To maximize the likelihood function, we want to maximize $-\sum y_i + n\theta$. That is, to maximize θ, as $\theta \leq y_{\min}$, it must be that $\theta_{\max} = y_{\min}$. Therefore, the MLE is $\theta^* = y_{\min}$. □

Example 1.2.10 Suppose $Y_1, \dots, Y_n \sim$ Uniform$[0, a]$. That is, $f_Y(y; a) = \dfrac{1}{a}$ for $y \in [0, a]$. Find MLE a_e of a.

Solution 6.

Note that

$$\begin{aligned} f_Y(y; a) &= \frac{1}{a} \cdot \mathbb{1}_{\{y \in [0,a]\}} \\ &= \frac{1}{a} \cdot \mathbb{1}_{\left\{0 \le y_{(1)} \le \cdots \le y_{(n)} \le a\right\}} \qquad \text{where } y_{(1)} = \min y_i \text{ and } y_{(n)} = \max y_i \end{aligned}$$

Then,

$$L(a) = \frac{1}{a^n} \mathbb{1}_{\left\{0 \le y_{(1)} \le \cdots \le y_{(n)} \le a\right\}}$$

To maximize $L(a)$, we want to minimize a^n. Since $a \ge y_{(n)}$, it must be that $a_e = y_{(n)}$. Here, we call $a_e = y_{(n)}$ an *estimate*, and $\widehat{a_{\text{MLE}}} = Y_{(n)}$ an *estimator*. □

Example 1.2.11 MLE that Does Not Esist

Suppose $f_Y(y; a) = \dfrac{1}{a}, \quad y \in [0, a)$. Find the MLE.

Solution 7.

The likelihood function is the same:

$$L(a) = \frac{1}{a^n} \mathbb{1}_{\left\{0 \le y_{(1)} \le \cdots \le y_{(n)} < a\right\}}.$$

However, since $[0, a)$ is not a closed set, the optimization problem $\max\limits_{a \in [0,a)} L(a)$ does not have a solution. Hence, the estimate does not exist. □

Remark 1.1 *MLE may not be unique all the time.*

Example 1.2.12 Multiple MLE Values

Suppose $X_1, \dots, X_n \sim$ Uniform$\left[a - \dfrac{1}{2}, a + \dfrac{1}{2}\right]$, where $f_X(x; a) = 1,\ x \in \left[a - \dfrac{1}{2}, a + \dfrac{1}{2}\right]$. Find the MLE.

Solution 8.

In the indicator function notation, we can rewrite the pdf to be

$$f_X(x; a) = \mathbb{1}_{\left\{a - \frac{1}{2} \le x \le a + \frac{1}{2}\right\}} = \mathbb{1}_{\left\{a - \frac{1}{2} \le x_{(1)} \le \cdots \le x_{(n)} \le a + \frac{1}{2}\right\}}.$$

So, the likelihood function will be

$$L(a) = \prod_{i=1}^{n} f_x(x_i; a) = \begin{cases} 1, & a \in \left[x_{(n)} - \frac{1}{2}, x_{(1)} + \frac{1}{2}\right] \\ 0, & \text{otherwise.} \end{cases}$$

So, the $L(a)$ will be maximized whenever $a \in \left[x_{(n)} - \frac{1}{2}, x_{(1)} + \frac{1}{2}\right]$. Therefore, MLE can be any value in the range $\left[x_{(n)} - \frac{1}{2}, x_{(1)} + \frac{1}{2}\right]$. Say,

$$a_e = x_{(n)} - \frac{1}{2} \quad \text{or} \quad a_e = x_{(1)} - \frac{1}{2} \quad \text{or} \quad a_e = \frac{x_{(n)} - \frac{1}{2} + x_{(1)} + \frac{1}{2}}{2} = \frac{x_{(n)} + x_{(1)}}{2}.$$

□

Theorem 1.2.13 MLE for Multiple Parameters
In general, we have the likelihood function $L(\theta)$, where $\theta = (\theta_1, \ldots, \theta_p)$. To find the MLE, we need

$$\frac{\partial L(\theta)}{\partial \theta_i} = 0 \quad i = 1, \ldots, p,$$

and the Hessian matrix

$$\left(\frac{\partial^2 L(\theta)}{\partial \theta_i \partial \theta_j}\right)_{i,j=1,\ldots,p} := \begin{pmatrix} \frac{\partial^2 L(\theta)}{\partial \theta_1^2} & \cdots & \frac{\partial^2 L(\theta)}{\partial \theta_1 \partial \theta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 L(\theta)}{\partial \theta_p \partial \theta_1} & \cdots & \frac{\partial^2 L(\theta)}{\partial \theta_p^2} \end{pmatrix}$$

should be negative dfinite.

Example 1.2.14 MLE for Multiple Parameters: Normal Distribution
Suppose $Y_1, \ldots, Y_n \sim N(\mu, \sigma)$. Then,

$$f_{Y_i}(u; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y_i - \mu)^2 / \left(2\sigma^2\right)}.$$

Find the MLE for μ and σ.

Solution 9.

The likelihood function will be

$$L(\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(y_i-\mu)^2/\left(2\sigma^2\right)}.$$

Then, we define

$$\ell(\mu, \sigma) = \ln L(\mu, \sigma) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \sigma^2 - \frac{1}{2}\left(\sigma^2\right)^{-}1\sum_{i=1}^{n}(y_i-\mu)^2.$$

Set

$$\begin{cases} \dfrac{\partial \ell(\mu,\sigma)}{\partial \mu} = 0 & \text{①} \\ \dfrac{\partial \ell(\mu,\sigma)}{\partial \sigma} = 0 & \text{②} \end{cases}$$

From ①, we have

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_1-\mu) = 0$$

$$\sum_{i=1}^{n} y_i = n\mu \implies \boxed{\mu_e = \frac{\sum y_i}{n} = \overline{y}}$$

From ②, by the invariant property of MLE, we instead set

$$\frac{\partial \ell(\mu,\sigma)}{\partial \sigma^2} = 0$$

$$-\frac{n}{2}\cdot\frac{1}{\sigma^2} + \frac{1}{2}\left(\frac{1}{\sigma^2}\right)^2\sum_{i=1}^{n}(y_i-\mu)^2 = 0$$

$$\frac{1}{2\sigma^2}\left(-n + \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right) = 0$$

$$-n\sigma^2 + \sum_{i=1}^{n}(y_i-\mu)^2 = 0 \qquad (\mu_e = \overline{y})$$

$$\sum_{i=1}^{n}(y_i-\overline{y})^2 = n\sigma^2$$

$$\sigma_e^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2 \implies \boxed{\sigma_e = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2}}$$

□

1.3 The Method of Moment

Definition 1.3.1 (Moment Generating Function). The *Moment Generating Function (MGF)* is defined as

$$\mathbf{M}_X(t) = \mathbf{E}\left[e^{tX}\right],$$

and it uniquely determines a probability distribution.

Definition 1.3.2 (Moment). The k*-th order moment* of X is $\mathbf{E}\left[X^k\right]$.

Example 1.3.3 Meaning of Different Moments

- $\mathbf{E}[X]$: location of a distribution
- $\mathbf{E}[X^2] = \mathbf{Var}(X) - \mathbf{E}[X]^2$: width of a distribution
- $\mathbf{E}[X^3]$: skewness – positively skewed / negatively skewed
- $\mathbf{E}[X^4]$: kurtosis / tailedness – speed decaying to 0.

Example 1.3.4 Moment Estimate: Moments of Population and Sample

Population	**Sample,** $X_1, \ldots, X_n$
$\mathbf{E}[X] = \mu$	$\widehat{\mu} = \overline{X} = \dfrac{X_1 + \cdots + X_n}{n}$
$\mathbf{E}[X^2] = \mu^2 + \sigma^2$	$\widehat{\mu}^2 + \widehat{\sigma}^2 = \dfrac{X_1^2 + \cdots + X_n^2}{n}$
$\vdots$	$\vdots$
$\mathbf{E}\left[X^k\right]$	$\dfrac{X_1^k + \cdots + X_n^k}{n}$

Rationale: The population moments should be close to the sample moments.

Example 1.3.5

- Consider $N(\mu, \sigma^2)$, where σ is given. Estimate μ.

  By the method of moment estimate, we have $\mu_e = \overline{X}$.

- Consider $N(\mu, \sigma^2)$. Estimate μ and σ.

  We have $\mu_e = \overline{X}$ and $\mu_e^2 + \sigma_e^2 = \dfrac{X_1^2 + \cdots + X_n^2}{n}$.

- Consider $N(\theta, \sigma^2)$. Given $\mathbf{E}(X^4) = 3\sigma^4$, estimate μ and σ.

  We have $\mu_e = \overline{X}$, $\mu_e^2 + \sigma_e^2 = \dfrac{X_1^2 + \cdots + X_n^2}{n}$, and $3\sigma^4 = \dfrac{X_1^4 + \cdots + X_n^4}{n}$. We have three equations but only two unknowns, then a solution is not guaranteed. So, we need some restrictions on this method (see Remark 1.2).

Theorem 1.3.6 Method of Moments Estimates
For a random sample of size n from the discrete (or continuous) population/pdf $p_X(k; \theta_1, \ldots, \theta_s)$ (or $f_Y(y; \theta_1, \ldots, \theta_s)$), solutions to the system

$$\begin{cases} \mathbf{E}(Y) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} y_i \\ \qquad\vdots \\ \mathbf{E}(Y^s) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} y_i^s \end{cases}$$

which are denoted by $\theta_{1e}, \ldots, \theta_{se}$, are called the **method of moments estimates** of $\theta_1, \ldots, \theta_s$.

Remark 1.2 *To estimate k parameters with the method of moments estimates, we will only match the first k orders of moments.*

Example 1.3.7 Consider the Gamma distribution:

$$f_Y(y; r, \lambda) = \frac{\lambda^r}{\Gamma(r)} y^{r-1} e^{-\lambda y} \quad \text{for } y \geq 0.$$

Given $\mathbf{E}(Y) = \dfrac{r}{\lambda}$ and $\mathbf{E}(Y^2) = \dfrac{r}{\lambda^2} + \dfrac{r^2}{\lambda^2}$. Estimate r and λ.

Solution 1.

$$\mathbf{E}(Y) = \frac{r}{\lambda} \implies \frac{r_e}{\lambda_e} = \frac{y_1 + \cdots + y_n}{n} = \overline{y} \qquad ①$$

$$\mathbf{E}\big(Y^2\big) = \frac{r}{\lambda^2} + \frac{r^2}{\lambda^2} \implies \frac{r_e}{\lambda_e^2} + \frac{r_e^2}{\lambda_e^2} = \frac{y_1^2 + \cdots + y_n^2}{n} \qquad ②$$

Substitute ① into ②, we have

$$\frac{\overline{y}}{\lambda_e} + (\overline{y})^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 \implies \boxed{\lambda_e = \frac{\overline{y}}{\frac{1}{n}\sum y_i^2 - \overline{y}^2}} \qquad ③$$

Substitute ③ into ①, we have

$$r_e = \overline{y}\lambda_e = \boxed{\frac{\overline{y}^2}{\frac{1}{n}\sum y_i^2 - \overline{y}^2}}.$$

□

Remark 1.3 *The* sample variance *is defined as*

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i^2 - 2y_i\overline{y} + \overline{y}^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i^2 - 2\overline{y}\cdot\frac{\sum y_i}{n} + \frac{1}{n}\cdot n\overline{y}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i^2 - 2\overline{y}^2 + \overline{y}^2 \qquad \overline{y} = \frac{\sum y_i}{n} \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{y}^2.
\end{aligned}$$

So, in Example 1.3.7, if we define $\widehat{\sigma}^2$ *to be the sample variance, we can further simply our estimate as follows:*

$$\lambda_e = \frac{\overline{y}}{\widehat{\sigma}^2}, \qquad r_e = \frac{\overline{y}^2}{\widehat{\sigma}^2}.$$