Maximum likelihood estimation

Ermal Feleqi
mailto: ermal.feleqi@univlora.edu.al

Departamenti i Matematikës
Universiteti “I. Qemali”, Vloë

May 4th, 2018

Outline

1. Slang
2. The rationale behind statistical modelling
3. Maximum likelihood estimation
4. Examples
   - The German tank problem
   - The bias of a coin
   - Estimation of the average height

Outline

Slang I

population A set of similar items or events which is of interest for some question or experiment. E.g., all stars in the Milky Way, all hands in a poker game. A common aim of statistical analysis is to produce information about some chosen population

sample A set of data collected and/or selected from a statistical population by a defined procedure. The elements of a sample are known as sample points, sampling units or observationsc
In mathematical terms, given a random variable X with prob. distrb. F, a random sample of length n ($n \in 1, 2, 3, ...$) is a set of n of i.i.d. r.v.'s $x_1, \ldots, x_n$ with distrib. F.
A sample concretely represents n experiments in which the same quantity is measured.

In statistical inference, a subset of the population (a statistical sample) is chosen to represent the population in a statistical analysis. If a sample is

Slang II

chosen properly, characteristics of the entire population that the sample is drawn from can be estimated from corresponding characteristics of the sample

Statistical model A mathematical model that embodies a set of statistical assumptions concerning the generation of some sample data and similar data from a larger population. A statistical model represents, often in considerably idealized form, the data-generating process Formally, $\left(E, (\mathbb{P})_{\theta \in \Theta}\right)$, where E is the sample space, $(\mathbb{P})_{\theta \in \Theta}$ a collection of prob. measures on E, and Θ any set called the parameter set.

Parameter A quantity that indexes a family of probability distributions. It can be regarded as a numerical characteristic of a population or a statistical model

Slang III

Statistic A single measure of some attribute of a sample (e.g. its arithmetic mean value). It is calculated by applying a function (statistical algorithm) to the values of the items of the sample, which are known together as a set of data.

Estimator A rule for calculating an estimate of a given quantity based on observed data An estimator of θ is usually denoted by the symbol $\widehat{\theta}$.

Bias of an estimator The difference between the estimator's expected value and the true value of the parameter being estimated. Biased and unbiased estimators.

Likelihood A function of the parameters of a statistical model, given specific observed data: $=$ prob. of observed data given specific parameters.

Slang IV

Maximum likelihood estimator An estimator obtained by maximizing thw likelihood.

Outline

The rationale behind statistical modelling

- Let $X_1, \ldots, X_n$ be n independent copies of X.
- The goal of Statistics is to learn the distribution of X.
- Example 1: if $X \in \{0, 1\}$, easy! It's $\mathrm{Ber}(p)$ and only parameter p is to be determined.
- Example 2: Average height of a given population of animals is to be determined. Measurement of the height of each individual not feasible (time, costs). Measure heights only heights of a few randomly chosen individuals.
- Example 3: German tank problem: Cards, numbered from 1 to n placed inside a box. Number n of cards is to be estimated by picking one or more cards at random at looking at their number.

Outline

Maximum likelihood estimation

Given a statistical model, a collection of probability measures

$$\{f(\cdot\,;\theta) \mid \theta \in \Theta\}$$

depending on θ, a possibly multidimensional parameter.
Parameter θ is the unknown.
Observations: $x = (x_1, \ldots, x_n)$: e.g., i.i.d. copies of X
Likelihood = probability of observations x:

$$\mathcal{L}(\theta\,; x)$$

The maximum likelihood estimator (MLE) $\hat{\theta}$ of the model parameter θ is

$$\hat{\theta} \in \arg\max_{\theta \in \Theta} \mathcal{L}(\theta\,; x),$$

provided it exists. Thus, the method of MLE finds the value of the model parameter that maximize the likelihood function, $\mathcal{L}(\theta\,; x)$.
More convenient, log-likelihood

$$\ell(\theta\,; x) = \ln \mathcal{L}(\theta\,; x) \qquad \hat{\ell}(\theta\,; x) = \frac{1}{n} \ln \mathcal{L}(\theta\,; x)$$

Maximum likelihood estimation

Given a statistical model, a collection of probability measures

$$\{f(\cdot\,;\theta) \mid \theta \in \Theta\}$$

depending on θ, a possibly multidimensional parameter.
Parameter θ is the unknown.
Observations: $x = (x_1, \ldots, x_n)$: e.g., i.i.d. copies of X
Likelihood = probability of observations x:

$$\mathcal{L}(\theta\,; x)$$

The maximum likelihood estimator (MLE) $\hat{\theta}$ of the model parameter θ is

$$\hat{\theta} \in \arg\max_{\theta\in\Theta} \mathcal{L}(\theta\,; x),$$

provided it exists. Thus, the method of MLE finds the value of the model parameter that maximize the likelihood function, $\mathcal{L}(\theta\,; x)$.
More convenient, log-likelihood

$$\ell(\theta\,; x) = \ln \mathcal{L}(\theta\,; x) \qquad \hat{\ell}(\theta\,; x) = \frac{1}{n} \ln \mathcal{L}(\theta\,; x)$$

Maximum likelihood estimation

Given a statistical model, a collection of probability measures

$$\{f(\cdot\,;\theta) \mid \theta \in \Theta\}$$

depending on θ, a possibly multidimensional parameter.
Parameter θ is the unknown.
Observations: $x = (x_1, \ldots, x_n)$: e.g., i.i.d. copies of X
Likelihood = probability of observations x:

$$\mathcal{L}(\theta\,; x)$$

The maximum likelihood estimator (MLE) $\hat{\theta}$ of the model parameter θ is

$$\hat{\theta} \in \arg\max_{\theta\in\Theta} \mathcal{L}(\theta\,; x),$$

provided it exists. Thus, the method of MLE finds the value of the model parameter that maximize the likelihood function, $\mathcal{L}(\theta\,; x)$.
More convenient, log-likelihood

$$\ell(\theta\,; x) = \ln \mathcal{L}(\theta\,; x) \qquad \hat{\ell}(\theta\,; x) = \frac{1}{n} \ln \mathcal{L}(\theta\,; x)$$

Maximum likelihood estimation

Given a statistical model, a collection of probability measures

$$\{f(\cdot\,;\theta) \mid \theta \in \Theta\}$$

depending on θ, a possibly multidimensional parameter.
Parameter θ is the unknown.
Observations: $x = (x_1, \ldots, x_n)$: e.g., i.i.d. copies of X
Likelihood = probability of observations x:

$$\mathcal{L}(\theta\,; x)$$

The maximum likelihood estimator (MLE) $\hat{\theta}$ of the model parameter θ is

$$\hat{\theta} \in \arg\max_{\theta\in\Theta} \mathcal{L}(\theta\,; x),$$

provided it exists. Thus, the method of MLE finds the value of the model parameter that maximize the likelihood function, $\mathcal{L}(\theta\,; x)$.
More convenient, log-likelihood

$$\ell(\theta\,; x) = \ln \mathcal{L}(\theta\,; x) \qquad \hat{\ell}(\theta\,; x) = \frac{1}{n} \ln \mathcal{L}(\theta\,; x)$$

Maximum likelihood estimation

Given a statistical model, a collection of probability measures

$$\{f(\cdot\,;\theta) \mid \theta \in \Theta\}$$

depending on θ, a possibly multidimensional parameter.
Parameter θ is the unknown.
Observations: $x = (x_1, \ldots, x_n)$: e.g., i.i.d. copies of X
Likelihood = probability of observations x:

$$\mathcal{L}(\theta\,; x)$$

The maximum likelihood estimator (MLE) $\hat{\theta}$ of the model parameter θ is

$$\hat{\theta} \in \underset{\theta\in\Theta}{\arg\max}\ \mathcal{L}(\theta\,; x),$$

provided it exists. Thus, the method of MLE finds the value of the model parameter that maximize the likelihood function, $\mathcal{L}(\theta\,; x)$.
More convenient, log-likelihood

$$\ell(\theta\,; x) = \ln \mathcal{L}(\theta\,; x) \qquad \hat{\ell}(\theta\,; x) = \frac{1}{n} \ln \mathcal{L}(\theta\,; x)$$

Outline

German tank problem

Discrete uniform distribution, finite parameter set

Problem n tickets inside a box, numbered from 1 to n.
Pick a ticket at random; estimate the total munber n.

Solution: The MLE $\hat{n}$ of n is the number of the picked ticket.
Indeed, the likelihood function is

$$\mathcal{L}(n; m) = \begin{cases} 0 & \text{if } n < m \\ 1/n & \text{otherwise.} \end{cases}$$

Therefore $\hat{n} = m$.

$\mathrm{E}[m] = \sum_{k=1}^{n} \frac{1}{n} k = (n+1)/2$.

MLE $\hat{n}$ of n systematically underestimates n by $(n-1)/2$.

MLE is biased.

German tank problem

Discrete uniform distribution, finite parameter set

Problem n tickets inside a box, numbered from 1 to n.
Pick a ticket at random; estimate the total munber n.

Solution: The MLE $\hat{n}$ of n is the number of the picked ticket.
Indeed, the likelihood function is

$$\mathcal{L}(n; m) = \begin{cases} 0 & \text{if } n < m \\ 1/n & \text{otherwise.} \end{cases}$$

Therefore $\hat{n} = m$.

$\mathrm{E}[m] = \sum_{k=1}^{n} \frac{1}{n} k = (n+1)/2$.

MLE $\hat{n}$ of n systematically underestimates n by $(n-1)/2$.

MLE is biased.

German tank problem

Discrete uniform distribution, finite parameter set

Problem n tickets inside a box, numbered from 1 to n.
Pick a ticket at random; estimate the total munber n.

Solution: The MLE $\hat{n}$ of n is the number of the picked ticket.
Indeed, the likelihood function is

$$\mathcal{L}(n; m) = \begin{cases} 0 & \text{if } n < m \\ 1/n & \text{otherwise.} \end{cases}$$

Therefore $\hat{n} = m$.

$\mathrm{E}[m] = \sum_{k=1}^{n} \frac{1}{n} k = (n+1)/2$.

MLE $\hat{n}$ of n systematically underestimates n by $(n-1)/2$.

MLE is biased.

Discrete distribution, continuous parameter space

How biased an unfair coin is?

p= prob. of tossing head; $p =$?.

Observation: The coin is tossed 80 times and we observe 49 heads.

Goal: Find $\hat{p}$, the MLE of p.

Solution.

$$\mathcal{L}(p;41) = f_D(\mathrm{H} = 49 \mid p) = \binom{80}{49} p^{49}(1-p)^{31}$$

Maximize this function:

$$0 = \frac{\partial}{\partial p}\left(\binom{80}{49} p^{49}(1-p)^{31}\right),$$

$$0 = 49p^{48}(1-p)^{31} - 31p^{49}(1-p)^{30}$$

$$= p^{48}(1-p)^{30}\left[49(1-p) - 31p\right]$$

$$= p^{48}(1-p)^{30}\left[49 - 80p\right],$$

which has solutions $p = 0$, $p = 49/31$, $p = 1$; the maximum is clearly attained at $p = 40/31$. Generalises to any Bernoulli trial.

$\hat{p} = s/n$, where s = no. of successes, n = tot. no. of trials.

Discrete distribution, continuous parameter space

How biased an unfair coin is?

p= prob. of tossing head; $p = ?$.

Observation: The coin is tossed 80 times and we observe 49 heads.

Goal: Find $\hat{p}$, the MLE of p.

Solution.

$$\mathcal{L}(p; 41) = f_D(\mathrm{H} = 49 \mid p) = \binom{80}{49} p^{49}(1-p)^{31}$$

Maximize this function:

$$0 = \frac{\partial}{\partial p}\left(\binom{80}{49} p^{49}(1-p)^{31}\right),$$

$$\begin{aligned} 0 &= 49p^{48}(1-p)^{31} - 31p^{49}(1-p)^{30} \\ &= p^{48}(1-p)^{30}\left[49(1-p) - 31p\right] \\ &= p^{48}(1-p)^{30}\left[49 - 80p\right], \end{aligned}$$

which has solutions $p = 0$, $p = 49/31$, $p = 1$; the maximum is clearly attained at $p = 40/31$. Generalises to any Bernoulli trial.

$\hat{p} = s/n$, where $s =$ no. of successes, $n =$ tot. no. of trials.

Discrete distribution, continuous parameter space

How biased an unfair coin is?

p= prob. of tossing head; $p = ?$.

Observation: The coin is tossed 80 times and we observe 49 heads.

Goal: Find $\hat{p}$, the MLE of p.

Solution.

$$\mathcal{L}(p; 41) = f_D(\mathrm{H} = 49 \mid p) = \binom{80}{49} p^{49}(1-p)^{31}$$

Maximize this function:

$$0 = \frac{\partial}{\partial p}\left(\binom{80}{49} p^{49}(1-p)^{31}\right),$$

$$0 = 49p^{48}(1-p)^{31} - 31p^{49}(1-p)^{30}$$

$$= p^{48}(1-p)^{30}\left[49(1-p) - 31p\right]$$

$$= p^{48}(1-p)^{30}\left[49 - 80p\right],$$

which has solutions $p = 0$, $p = 49/31$, $p = 1$; the maximum is clearly attained at $p = 40/31$. Generalises to any Bernoulli trial.

$\hat{p} = s/n$, where $s =$ no. of successes, $n =$ tot. no. of trials.

Continuous distribution, continuous parameter space I

Estimate the average height of a population of animals.
Model: Height X normally distributed, i.e., $X \sim \mathcal{N}(\mu, \sigma^2)$. with density probability function

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Thus $\mu, \sigma = ?$.
Observations: i.i.d. $x_1, \ldots, x_n$.
Solution: prob. density function of the i.i.d sample is

$$f(x_1, \ldots, x_n \mid \mu, \sigma^2) = \prod_{i=1}^{n} f(x_i \mid \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right),$$

Continuous distribution, continuous parameter space II

or more conveniently,

$$f(x_1, \ldots, x_n \mid \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2}{2\sigma^2}\right),$$

where $\bar{x}$ is the sample mean.
Likelihood: $\mathcal{L}(\mu, \sigma) = f(x_1, \ldots, x_n \mid \mu, \sigma)$.
The log-likelihood:

$$\log\Big(\mathcal{L}(\mu, \sigma)\Big) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\, x_i - \mu\,)^2.$$

We compute derivatives

$$0 = \frac{\partial}{\partial\mu}\log\Big(\mathcal{L}(\mu, \sigma)\Big) = 0 - \frac{-2n(\bar{x} - \mu)}{2\sigma^2}.$$

Continuous distribution, continuous parameter space III

This is solved by

$$\hat{\mu} = \bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}.$$

$$\mathrm{E}\left[\,\widehat{\mu}\,\right] = \mu$$

thus the MLE $\widehat{\mu}$ is unbiased.

Similarly we differentiate the log-likelihood with respect to σ and equate to zero

$$\begin{aligned} 0 &= \frac{\partial}{\partial\sigma} \log\left[\left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2}{2\sigma^2}\right)\right] \\ &= \frac{\partial}{\partial\sigma}\left[\frac{n}{2}\log\left(\frac{1}{2\pi\sigma^2}\right) - \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2}{2\sigma^2}\right] \end{aligned}$$

Continuous distribution, continuous parameter space IV

$$= -\frac{n}{\sigma} + \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2}{\sigma^3}$$

which is solved by

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Inserting the estimate $\mu = \widehat{\mu}$ we obtain

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}x_i x_j.$$

We calculate its expected value, $\delta_i \equiv \mu - x_i$.

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(\mu - \delta_i)^2 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\mu - \delta_i)(\mu - \delta_j).$$

Continuous distribution, continuous parameter space V

Simplifying the expression above, utilizing the facts that $\mathrm{E}\left[\ \delta_i\ \right] = 0$ and $\mathrm{E}\left[\ \delta_i^2\ \right] = \sigma^2$, we obtain

$$\mathrm{E}\left[\ \widehat{\sigma}^2\ \right] = \frac{n-1}{n}\sigma^2.$$

This means that the estimator $\widehat{\sigma}$ is biased. However, $\widehat{\sigma}$ is consistent.
The MLE $\theta = (\mu, \sigma^2)$ is

$$\widehat{\theta} = \left(\widehat{\mu}, \widehat{\sigma}^2\right)$$

The normal log-likelihood at its maximum takes a particularly simple form:

$$\log\Big(\mathcal{L}(\hat{\mu}, \hat{\sigma})\Big) = \frac{-n}{2}\Big(\ \log(2\pi\hat{\sigma}^2) + 1\ \Big).$$

Summary of exmaples

Examples

- Bernoulli trials: $\hat{p}_n^{MLE} = \bar{x}_n$.
- Exponential model: $x_1, \ldots, x_n \overset{iid}{\sim} \text{Poiss}(\lambda)$, $\widehat{\lambda}_n^{MLE} = \bar{x}_n$.
- Gaussian model: $\left(\widehat{\mu}_n^{MLE}, (\widehat{\sigma}_n^{MLE})^2\right) = \left(\bar{x}_n, \bar{S}_n^2\right)$

Properties I

Consistency: If θ_0 is the true parameter, i.e., if obseved data were generated by $f(\cdot\,;\theta_0)$, under some reasonable assumptions

$$\hat{\theta}_{\rm n} \xrightarrow{\rm p} \theta_0.$$

Under slightly stronger conditions:

$$\hat{\theta}_{\rm mle} \xrightarrow{\rm a.s.} \theta_0.$$

Suff. conditions for consistency: identifiability, compactness, continuity in θ for a.e. x,

Functional invariance: If $Y = g(x)$, for some g one-to-one $f_Y(y) = \frac{f_X(x)}{|g'(x)|}$, hence likelihood functions for x and y depend only on a factor that does not depend on parameters. The MLE param. of log-normal distribution are the same as those of the normal distribution fitted to the logarithm of the data.

Properties II

Efficiency: The maximum likelihood estimator is $\sqrt{n}$-consistent and asymptotically efficient, meaning that it reaches the CramérŰRao bound:

$$\sqrt{n}(\hat{\theta}_{\mathrm{mle}} - \theta_0) \xrightarrow{d} \mathcal{N}(0,\ I^{-1}),$$

where I is the Fisher information matrix:

$$I_{jk} = \mathrm{E}_X\left[-\frac{\partial^2 \ln f_{\theta_0}(X_t)}{\partial\theta_j\,\partial\theta_k} \right].$$

Slang

population

sample

Statistical inference

Statistical model

Parameter

Statistic

Estimator

Likelihood

Maximum likelihood estimator

Bias of an estimator

Thank You for Your Attention!