Maximum Likelihood & Method of Moments Estimation

Patrick Zheng

01/30/14

Introduction

- **Goal**: Find a good POINT estimation of population parameter
- **Data**: We begin with a random sample of size n taken from the totality of a population.
  - We shall estimate the parameter based on the sample
- **Distribution**: Initial step is to identify the probability distribution of the sample, which is characterized by the parameter.
  - The distribution is always easy to identify
  - The parameter is unknown.

Notations

- Sample: $X_1, X_2, \ldots, X_n$
- Distribution: X_i iid f(x, θ)
- Parameter: θ

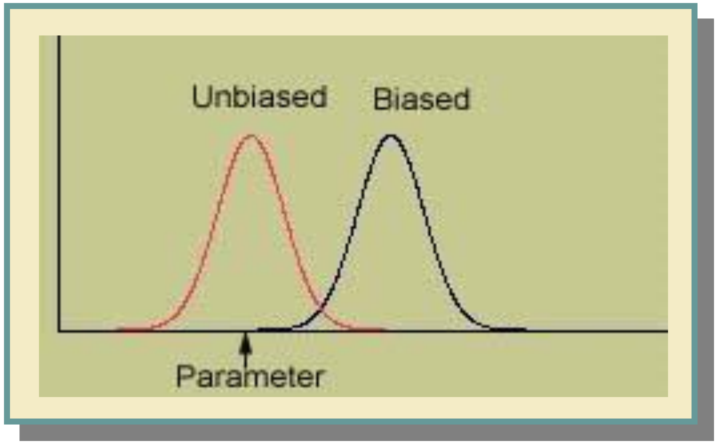
- Example
  - e.g., the distribution is normal (f=Normal) with unknown parameter μ and σ^2(θ=(μ, σ^2)).
  - e.g., the distribution is binomial (f=binomial) with unknown parameter *p* (θ= *p*).

It's important to have a good estimate!

- The importance of point estimates lies in the fact that many statistical formulas are based on them, such as confidence interval and formulas for hypothesis testing, etc..
- *A good estimate should*
  1. *Be unbiased*
  2. *Have small variance*
  3. *Be efficient*
  4. *Be consistent*

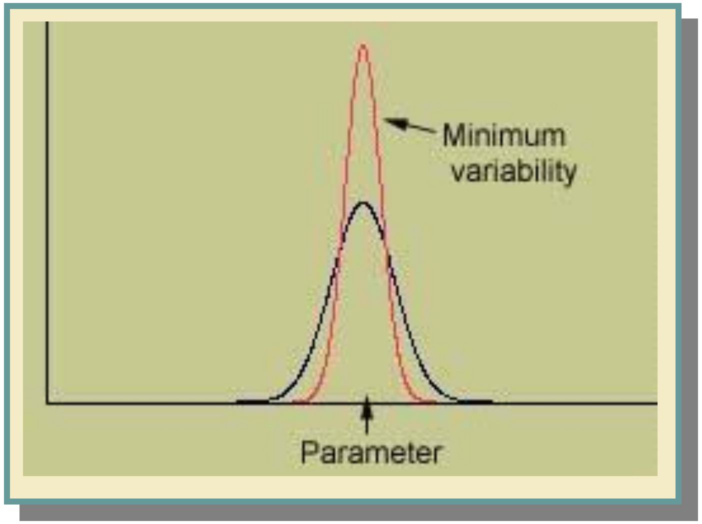
Unbiasedness

- An **estimator** is **unbiased** if its mean equals the parameter.
- It does not systematically overestimate or underestimate the target parameter.
- Sample mean($\overline{x}$)/proportion($\hat{p}$) is an unbiased estimator of population mean/proportion.



Small variance

- We also prefer the sampling distribution of the estimator has a **small spread** or **variability,** i.e. small standard deviation.



Efficiency

- An estimator $\hat{\theta}$ is said to be efficient if its Mean Square Error (MSE) is minimum among all competitors.

$$\mathrm{MSE}(\hat{\boldsymbol{\theta}}) = \mathrm{E}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2 = \mathrm{Bias}^2(\hat{\boldsymbol{\theta}}) + \mathrm{var}(\hat{\boldsymbol{\theta}}),$$

$$\text{where } \mathrm{Bias}(\hat{\boldsymbol{\theta}}) = \mathrm{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}.$$

- Relative Efficiency($\hat{\theta}_1$, $\hat{\theta}_2$) $= \frac{MSE(\hat{\theta}_2)}{MSE(\hat{\theta}_1)}$
  - If >1, $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$.
  - If <1, $\hat{\theta}_2$ is more efficient than $\hat{\theta}_1$.

Example: efficiency

- Suppose $X_1, X_2, \ldots X_n$ iid$\sim N(\mu, \sigma^2)$.
- If $\hat{\mu}_1 = X_1$, then

$$MSE(\hat{\mu}_1) = Bias^2(\hat{\mu}_1) + var(\hat{\mu}_1) = 0 + \sigma^2.$$

- If $\hat{\mu}_2 = \overline{X} = \dfrac{X_1 + X_2 + \ldots + X_n}{n}$, then

$$MSE(\hat{\mu}_2) = Bias^2(\hat{\mu}_2) + var(\hat{\mu}_2) = 0 + \sigma^2 / n.$$

- Since $R.E.(\hat{\mu}_1, \hat{\mu}_2) = \dfrac{MSE(\hat{\mu}_2)}{MSE(\hat{\mu}_1)} = \dfrac{\sigma^2 / n}{\sigma^2} = \dfrac{1}{n} < 1$,

  $\hat{\mu}_2$ is more efficient than $\hat{\mu}_1$.

Consistency

- An estimator $\hat{\theta}$ is said to be consistent if sample size n goes to $+\infty$, $\hat{\theta}$ will converge in probability to θ.

$$\forall \boldsymbol{\varepsilon} > 0,\ \Pr(|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}| > \boldsymbol{\varepsilon}) \to 0 \quad \text{as} \ \ n \to +\infty$$

- Chebychev's rule

$$\forall \boldsymbol{\varepsilon} > 0,\ \Pr(|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}| \geq \boldsymbol{\varepsilon}) \leq \frac{E(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2}{\boldsymbol{\varepsilon}^2} = \frac{MSE(\hat{\boldsymbol{\theta}})}{\boldsymbol{\varepsilon}^2}$$

  - If one can prove MSE of $\hat{\theta}$ tends to 0 when n goes to $+\infty$, then $\hat{\theta}$ is consistent.

Example: Consistency

- Suppose $X_1, X_2, \ldots X_n$ iid$\sim N(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$.

- Estimator $\hat{\boldsymbol{\mu}} = \overline{X} = \dfrac{X_1 + X_2 + \ldots + X_n}{n}$ is consistent, since

$$\forall \boldsymbol{\varepsilon} > 0,\ \Pr(|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}| \geq \boldsymbol{\varepsilon}) \leq \frac{E(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^2}{\boldsymbol{\varepsilon}^2} = \frac{MSE(\hat{\boldsymbol{\mu}})}{\boldsymbol{\varepsilon}^2}$$

$$= \frac{\boldsymbol{\sigma}^2 / n}{\boldsymbol{\varepsilon}^2} \to 0 \quad \text{as } n \to +\infty$$

Point Estimation Methods

- There are many methods available for estimating the parameter(s) of interest.

- Three of the most popular methods of estimation are:
  - The method of moments (MM)
  - The method of maximum likelihood (ML)
  - Bayesian method

1, The Method of Moments

The Method of Moments

- One of the oldest methods; very simple procedure

- What is Moment?

- Based on the assumption that sample moments should provide GOOD ESTIMATES of the corresponding population moments.

How it works?

THE METHOD OF MOMENTS PROCEDURE

Suppose there are l parameters to be estimated, say $\theta = (\theta_1, \ldots, \theta_l)$.

1. Find l population moments, $\mu'_k, k = 1, 2, \ldots, l$. μ'_k will contain one or more parameters $\theta_1, \ldots, \theta_l$.
2. Find the corresponding l sample moments, $m'_k, k = 1, 2, \ldots, l$. The number of sample moments should equal the number of parameters to be estimated.
3. From the system of equations, $\mu'_k = m'_k, k = 1, 2, \ldots, l$, solve for the parameter $\theta = (\theta_1, \ldots, \theta_l)$; this will be a moment estimator of $\hat{\theta}$.

$$\mu'_k = E[X^k]$$

$$m'_k = (1/n) \sum_{i=1}^{n} X_i^k \qquad m'_1 = \bar{X}; \; m'_2 = (1/n) \sum_{i=1}^{n} X_i^2$$

$$\mu'_k = m'_k$$

Example: normal distribution

$X_1, X_2, \ldots X_n$ iid~ $N(\tau, \sigma^2)$.

step 1, $\mu_1' = E(X) = \tau;\ \mu_2' = E(X^2) = \tau^2 + \sigma^2$.

step 2, $m_1' = \bar{X};\ m_2' = (1/n)\sum_{i=1}^n X_i^2$.

step 3, Set $\mu_1' = m_1', \mu_2' = m_2'$, therefore,

$$\tau = \bar{X},$$

$$\tau^2 + \sigma^2 = (1/n)\sum_{i=1}^n X_i^2$$

Solving the two equations, we get $\hat{\tau} = \bar{X}, \hat{\sigma}^2 = (1/n)\sum_{i=1}^n X_i^2 - \bar{X}^2$

Example: Bernoulli Distribution

Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli population with parameter p.

(a) Find the moment estimator for p.

Solution

(a) *For the Bernoulli random variable,* $\mu'_k = E[X] = p$, *so we can use* m'_1 *to estimate* p. *Thus,*

$$m'_1 = \hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

X follows a Bernoulli distribution, if $P(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$

Example: Poisson distribution

Let $X_1, \ldots, X_n$ be a random sample from a Poisson distribution with parameter $\lambda > 0$. Show that both $(1/n)\sum_{i=1}^{n} X_i$ and $(1/n)\sum_{i=1}^{n} X_i^2 - \left((1/n)\sum_{i=1}^{n} X_i\right)^2$ are moment estimators of λ.

Solution

We know that $E(X) = \lambda$, from which we have a moment estimator of λ as $(1/n)\sum_{i=1}^{n} X_i$. Also, because we have $Var(X) = \lambda$, equating the second moments, we can see that

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\lambda = E(X^2) - (EX)^2,$$

so that

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2.$$

Both are moment estimators of λ. Thus, the moment estimators may not be unique. We generally choose $\overline{X}$ as an estimator of λ, for its simplicity.

Note

- MME may not be unique.

- In general, minimum number of moment conditions we need equals the number of parameters.

- Question: Can these two estimators be combined in some optimal way?

  Answer: Generalized method of moments.

Pros of Method of Moments

- Easy to compute and always work:
  - The method often provides estimators when other methods fail to do so or when estimators are hard to obtain (as in the case of gamma distribution).

- MME is consistent.

Cons of Method of Moments

- They are usually not the “best estimators” available. By best, we mean most efficient, i.e., achieving minimum MSE.

- Sometimes it may be meaningless.

(see next page for example)

Sometimes, MME is meaningless

- Suppose we observe 3,5,6,18 from a $U(0,\theta)$

- Since E(X)= θ /2,

  MME of θ is $2\bar{X}=2*\frac{3+5+6+18}{4}=16$, which is not acceptable, because we have already observed a value of 18.

2, The Method of Maximum Likelihood

The Method of Maximum Likelihood

- Proposed by geneticist/statistician:
  Sir Ronald A. Fisher in 1922

- **Idea**: We attempt to find the values of the parameters which would have most likely produced the data that we in fact observed.

What is likelihood?

- **Definition 5.3.1** *Let $f(x_1, \ldots, x_n; \theta), \theta \in \Theta \subseteq \mathbb{R}^k$, be the joint probability (or density) function of n random variables $X_1, \ldots, X_n$ with sample values $x_1, \ldots, x_n$. The* likelihood function *of the sample is given by*

$$L(\theta; x_1, \ldots, x_n) = f(x_1, \ldots, x_n; \theta), [= L(\theta), \text{ in a briefer notation}].$$

We emphasize that L is a function of θ for fixed sample values.

- E.g., Likelihood of θ=1 is the chance of observing $X_1, X_2, ..., X_n$ when θ=1.

How to compute Likelihood?

- If $X_1, \ldots, X_n$ are discrete iid random variables with probability function $p(x, \theta)$, then, the likelihood function is given by

$$\begin{aligned} L(\theta) &= P(X_1 = x_1, \ldots, X_n = x_n) \\ &= \prod_{i=1}^{n} P(X_i = x_i), \quad \text{(by multiplication rule for independent random variables)} \\ &= \prod_{i=1}^{n} p(x_i, \theta) \end{aligned}$$

- and in the continuous case, if the density is $f(x, \theta)$, then the likelihood function is

$$L(\theta) = \prod_{i=1}^{n} f(x_i, \theta).$$

Example of computing likelihood (discrete case)

Suppose $X_1, \ldots, X_n$ are a random sample from a geometric distribution with parameter p, $0 \le p \le 1$.

Solution

For the geometric distribution, the pmf is given $p(1-p)^{x-1}, \quad 0 \le p \le 1, \quad x = 1, 2, 3, \ldots.$

Hence, the likelihood function is

$$L(p) = \prod_{i=1}^{n} \left[p\,(1-p)^{x-1} \right] = p^n\,(1-p)^{-n + \sum_{i=1}^{n} x_i}.$$

Example of computing likelihood (continuous case)

Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$ random variables. Let $x_1, \ldots, x_n$ be the sample values. Find the likelihood function.

Solution

The density function for the normal variable is given by $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$. *Hence, the likelihood function is*

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

Definition of MLE

- Definition 5.3.2 *The* maximum likelihood estimators (MLEs) *are those values of the parameters that maximize the likelihood function with respect to the parameter θ. That is,*

$$L\left(\hat{\theta}; x_1, \ldots, x_n\right) = \max_{\theta \in \Theta} L\left(\theta; x_1, \ldots, x_n\right)$$

where Θ is the set of possible values of the parameter θ.

- In general, the method of ML results in the problem of maximizing a function of single or several parameters. One way to do the maximization is to take derivative.

Procedure to find MLE

1. Define the likelihood function, $L(\theta)$.
2. Often it is easier to take the natural logarithm (ln) of $L(\theta)$.
3. When applicable, differentiate $\ln L(\theta)$ with respect to θ, and then equate the derivative to zero.
4. Solve for the parameter θ, and we will obtain $\hat{\theta}$.
5. Check whether it is a maximizer or global maximizer.

Example: Poisson Distribution

Suppose $X_1, \ldots, X_n$ are random samples from a Poisson distribution with parameter λ. Find MLE $\hat{\lambda}$.

Solution

We have the probability mass function

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \ldots, \quad \lambda > 0.$$

Hence, the likelihood function is

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}.$$

Then, taking the natural logarithm, we have

$$\ln L(\lambda) = \sum_{i=1}^{n} x_i \ln \lambda - n\lambda - \sum_{i=1}^{n} \ln(x_i!)$$

Example cont'd

and differentiating with respect to λ results in

$$\frac{d \ln L(\lambda)}{d\lambda} = \frac{\sum_{i=1}^{n} x_i}{\lambda} - n$$

and

$$\frac{d \ln L(\lambda)}{d\lambda} = 0, \text{ implies } \frac{\sum_{i=1}^{n} x_i}{\lambda} - n = 0.$$

That is,

$$\lambda = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}.$$

Hence, the MLE of λ is

$$\hat{\lambda} = \overline{X}.$$

Example: Uniform Distribution

Let $X_1, \ldots, X_n$ be a random sample from $U(0, \theta), \theta > 0$. Find the MLE of θ.

Solution

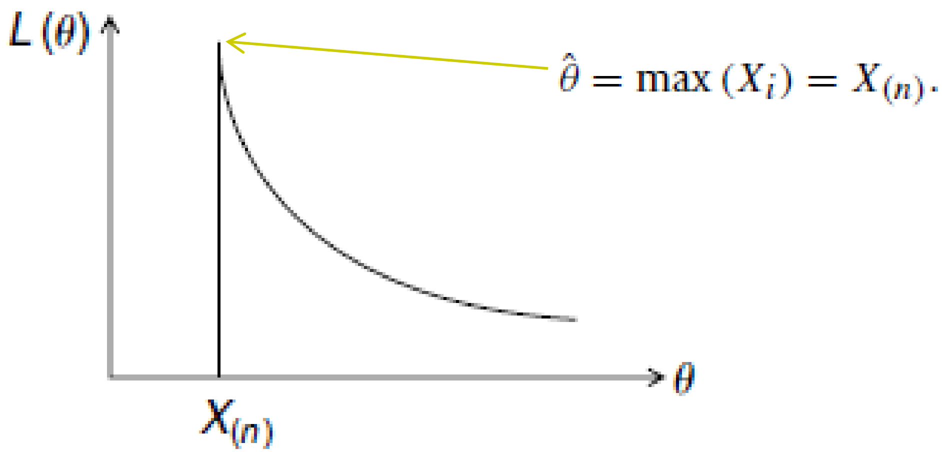
Note that the pdf of the uniform distribution is

$$f(x) = \begin{cases} \frac{1}{\theta}, & 0 \leq x \leq \theta \\ 0, & \text{otherwise.} \end{cases}$$

Hence, the likelihood function is given by

$$L(\theta, x_1, x_2, \ldots, x_n) = \begin{cases} \frac{1}{\theta^n}, & 0 \leq x_1, x_2, \ldots, x_n \leq \theta \\ 0, & \text{otherwise.} \end{cases}$$

Example cont’d



■ FIGURE 5.1 Likelihood function for uniform probability distribution.

More than one parameter

As mentioned earlier, if the unknown parameter θ represents a vector of parameters, say $\theta = (\theta_1, \ldots, \theta_l)$, then the MLEs can be obtained from solutions of the system of equations

$$\frac{\partial}{\partial\theta} \ln L(\theta_1, \ldots, \theta_n) = 0, \quad \text{for} \quad i = 1, \ldots, l.$$

These are called the *maximum likelihood equations* and the solutions are denoted by $(\hat{\theta}_1, \ldots, \hat{\theta}_l)$.

Pros of Method of ML

- When sample size n is large (n>30), MLE is unbiased, consistent, normally distributed, and efficient (“regularity conditions”)
  - “Efficient” means it produces the minimum MSE than other methods including Method of Moments
- More useful in statistical inference.

Cons of Method of ML

- MLE can be highly biased for small samples.
- Sometimes, MLE has no closed-form solution.
- MLE can be sensitive to starting values, which might not give a global optimum.
  - Common when θ is of high dimension

How to maximize Likelihood

1. Take derivative and solve analytically (as aforementioned)

2. Apply maximization techniques including Newton's method, quasi-Newton method (*Broyden 1970*), direct search method (*Nelder and Mead 1965*), etc.
   - These methods can be implemented by R function optimize(), optim()

Newton's Method

- a method for finding successively better approximations to the roots (or zeroes) of a real-valued function.
  - Pick an x close to the root of a continuous function $f(x)$
  - Take the derivative of $f(x)$ to get $f'(x)$
  - Plug into $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$, $f'(x_n) \neq 0$
  - Repeat until converges where $x_{n+1} \approx x_n$

Example

- Solve $e^x - 1 = 0$
  - Denote $f(x)$= $e^x - 1$; let starting point $x_0 = 0.1$
  - $f'(x)$=e^x
  - $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$:
    - $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = 0.1 - \frac{e^{0.1}-1}{e^{0.1}} = 0.0048374$
    - $x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} =\ldots$
  - Repeat until $|x_{n+1} - x_n| < 0.00001$, $x_{n+1} = 7.106 * 10^{-17}$

Example: find MLE by Newton's Method

- In Poisson Distribution, find $\hat{\lambda}$ is equivalent to
  - maximizing $\ln L(\lambda)$
  - finding the root of $\frac{d \ln L(\lambda)}{d\, \lambda} = \frac{\sum x}{\lambda} - n$
- Implement Newton's method here,
  - define $f(\lambda) = \frac{d \ln L(\lambda)}{d\, \lambda} = \frac{\sum x}{\lambda} - n$
  - $f'(\lambda) = \frac{-\sum x}{\lambda^2}$
  - $\lambda_{n+1} = \lambda_n - \frac{f(\lambda_n)}{f'(\lambda_n)}$
  - Given $x_1, x_2, \ldots, x_m$ and λ_0, we can find $\hat{\lambda}$.

Example cont'd

- Suppose we collected a sample from Poi(λ):

  18,10,8,13,7,17,11,6,7,7,10,10,12,4,12,4,12,10,7,14,13,7

- Implement Newton's method in R:

$$\lambda_{n+1} = \lambda_n - \frac{f(\lambda_n)}{f'(\lambda_n)}$$

```
#use newton method to find lamda mle of poisson
#x here is data, l here is lamda
x<-c(18,10,8,13,7,17,11,6,7,7,10,10,12,4,12,4,12,10,7,14,13,7)
n<-length(x)
l<-NULL
l[1]<-8 # give initial value of lamda
i<-1
repeat{
l[i+1]<-l[i]-(-n+sum(x)/l[i])/(-sum(x)/(l[i]^2)) # iterative equation
diff<-abs(l[i+1]-l[i])                        # set up stopping criteria
i<-i+1
if ( diff < 0.0001) { break
     }
      }
> l
[1] 8.000000 9.570776 9.939750 9.954523 9.954545
```

Use R function optim()

$$f(\lambda) = \frac{\sum x}{\lambda} - n$$

Typo! This should be -lnL(lamda).

```
poi<-function(l){
x<-c(18,10,8,13,7,17,11,6,7,7,10,10,12,4,12,4,12,10,7,14,13,7)
n<-length(x)
-(-n*l+sum(x)*log(l))   # as optim can only minimize a function
}                        # so we add a minus sign to the target function
optim(7,poi,lower=0.1,upper=Inf,method="L-BFGS-B")
$par
[1] 9.954545
```

- The End!
- Thank you!