1. Introduction

Welcome to the series of E-learning modules on Method of Maximum Likelihood and their properties. In this module, we are going to cover the basic principle of method of maximum likelihood, advantages and properties of maximum likelihood estimators and estimation of certain population parameters by the method of maximum likelihood.

By the end of this session, you will be able to:

- Explain the basic concept of method of maximum likelihood
- Explain the principle of the method and properties of likelihood estimators
- Explain the advantages and disadvantages of the method of maximum likelihood
- Explain the procedure to estimate certain parameters of the population by the method of maximum likelihood

Selection of a random sample is very much essential for drawing valid inferences about the population. Methods that are developed for estimating the population parameters provide a theoretical basis of connection between sample information and population parameters, which assists in drawing efficient estimates from the sample, and draw inference about the population.

In statistics, Maximum-Likelihood Estimation (MLE) is a method of estimating the parameters of a statistical model, which is generally used in estimation. When applied to a data set and given a statistical model, maximum-likelihood estimation provides estimates for the model's parameters.

The method of maximum likelihood corresponds to many well-known estimation methods in statistics. For example, one may be interested in the intelligence of school going children. However, will be unable to measure the intelligence of every single child in a population due to cost or time constraints.
Assuming that the intelligence is normally distributed with some unknown mean and variance, the mean and variance can be estimated with MLE while only knowing the intelligence of some sample of the overall population.

MLE would accomplish this by taking the mean and variance as parameters and finding particular parametric values that make the observed results the most probable.
In general, for a fixed set of data and underlying statistical model, the method of maximum likelihood selects values of the model parameters that produce a distribution that gives the observed data the greatest probability.

2. Applications of the Method of Maximum Likelihood

Applications of the Method of Maximum Likelihood:
Maximum likelihood estimation is used for a wide range of statistical models, including:
Linear Models and generalized linear models
Exploratory and confirmatory factor analysis
Structural Equation Modelling
Many situations in the context of hypothesis testing and confidence interval formation

These uses arise across applications in widespread set of fields, including:

- Communication systems
- Psychometrics
- Econometrics
- Data modelling in nuclear and particle physics
- Magnetic resonance imaging
- Computational Phylogenetics
- Geographical satellite-image classification

On the other hand, MLE is not as widely recognized among modellers in psychology, but it is a standard approach to parameter estimation and inference in statistics. MLE has many optimal properties in estimation, such as sufficiency, consistency, efficiency etc.

Further, many of the inference methods in statistics are developed based on MLE. For example, MLE is a prerequisite for the chi-square test, the Gsquare test, Bayesian methods, inference with missing data, modelling of random effects, and many model selection criteria.

Once data has been collected and the likelihood function of a model is determined, we can make statistical inferences about the population, that is, the probability distribution that underlies the data. Given the different parameter value index and probability distributions, we are interested in finding the parameter value that corresponds to the desired probability distribution.

The principle of maximum likelihood estimation originally developed by R. A. Fisher in the nineteen twenty's. It states that the desired probability distribution is the one that makes the observed data "most likely," which means that one must seek the value of the parameter vector that maximizes the likelihood function. The resulting parameter vector, which is sought by searching the multi-dimensional parameter space, is called the MLE.

3. Likelihood Function

Likelihood Function:
A likelihood function of a number of sample observations is defined to be their joint density function. To use the method of maximum likelihood, we should first specify the joint density function for all observations. For an Independent and Identically Distributed (IID) sample, this joint density function is
f of (x one, x two, up to xn,theta) is equal to f of (x one, theta) into f of (x two, theta) into f of (x three, theta) up to f of (xn, theta)

Now, we look at this function from a different perspective by considering the observed values x one, x two, upto xn to be fixed "parameters" of this function, whereas theta will be the function's variable and allowed to vary freely. This function will be called the likelihood.

In case, the sample observations are independent, the likelihood function happens to be the product of the density functions of the random observations. If x one, x two, up to xn **are** *n* independent and identically distributed observations, from a population with an unknown parameter theta then the likelihood function of the random observations denoted by
L is equal to L of (x one, x two, up to xn,theta) **is given by**
L is equal to L of (x one, x two, up to xn,theta) **is equal to product of f of (xi,theta)**
Where, f of (xi, theta) is the p.d.f of the population.

L gives relative likelihood or chance that the random observations assume a particular set of values. For a particular sample (x one, x two, up to xn), L becomes a function of (x one, x two, up to xn) and an unknown parameter theta and p.d.f. f of (x, theta). It is desirable to find an estimator theta cap, which would be closest to the true value theta. Both the observed variables x_i and the parameter theta can be vectors.

The method of maximum likelihood is the one in which for a given set of values x one, x two, up to xn, an estimator of theta is found that maximizes L. Thus, if there exists theta cap, a function of x one, x two, up to xn for which L is maximum for variations in theta. Then, theta cap is the M.L.E of theta if
Delta L by delta theta is equal to zero and Delta square L by delta theta square is less than zero

In practice, it is often more convenient to work with the logarithm of the likelihood function, called the log-likelihood.
Since logL attains maximum, when L attains maximum the solution of the equation.
Log L of (x one, x two, up to xn,theta) is equal to summation log f of (xi, theta)
Delta log L by delta theta is equal to zero and Delta square log L by delta theta square is less than zero also gives the M.L.E of θ
An MLE estimate is the same regardless of whether we maximize the likelihood or the log-likelihood function, since log is a monotone transformation.

4. Principle, Properties and Advantages of Maximum Likelihood Function

Principle:
The likelihood function L of (x one, x two, up to xn,theta) in a way gives the probability of the random sample x one, x two, up to xn, when the parameter of the distribution is theta. For different values of theta, L gives the different probabilities. Since the event of getting a sample x one, x two, up to xn, has occurred, the probability of the event must be high or else the value of L should be large.

So the value of theta, which gives the maximum probability for the sample is found and it is taken as the maximum likelihood estimate of theta. In fact, the function of the sample observations, which maximizes L, is determined and that is the MLE of theta.
For many models, a maximum likelihood estimator can be found as an explicit function of the observed data x one, x two, up to xn. For many other models, no closed-form solution to the maximization problem is known or available and an MLE has to be found numerically using optimization methods.

For some problems, there may be multiple estimates that maximize the likelihood. For other problems, no maximum likelihood estimate exists (meaning that the log-likelihood function increases without attaining the supremum value).
In the exposition above, it is assumed that the data are independent and identically distributed. The method can be applied however to a broader setting, as long as it is possible to write the joint density function *ƒ of (*x one, x two, up to xn, theta).

In a simpler extension, an allowance can be made for data heterogeneity, so that the joint density is equal to *ƒ* one of (x one,theta) into *ƒ* two of (x two,theta) up to *ƒ*n of (x n,theta). In the more complicated case of time series models, the independence assumption may have to be dropped as well.

Properties
A maximum-likelihood estimator is an extremum estimator obtained by maximizing, as a function of theta.
Maximum-likelihood estimators have no optimum properties for finite samples.
However, like other estimation methods, maximum-likelihood estimation possesses a number of attractive limiting properties.

As the sample-size increases to infinity, sequences of maximum-likelihood estimators have these properties:
Consistency: A subsequence of the sequence of MLEs converges in probability to the value being estimated.
Asymptotic normality: As the sample size increases, the distribution of the MLE tends to the Gaussian distribution.

Efficiency: MLEs are most efficient among the class of all consistent estimators. That is among the class of consistent estimators MLE has the minimum variance.
Sufficiency: Complete information about the parameter of interest is contained in its MLE.

Maximum-likelihood estimators can lack asymptotic normality and can be inconsistent if there is a failure of one (or more) of the below regularity conditions:

Estimate on boundary

Sometimes the maximum likelihood estimate lies on the boundary of the set of possible parameters. Standard asymptotic theory needs the assumption that the true parameter value lies away from the boundary.

Data boundary parameter-dependent

For the theory to apply in a simple way, the set of data values, which has positive probability (or positive probability density), should not depend on the unknown parameter.

A simple example where such parameter-dependence does hold is the case of estimating theta from a set of IID variables is when the common distribution is uniform on the range (zero, theta). For estimation purposes, the relevant range of theta is such that theta cannot be less than the largest observation. Because the interval (zero, theta) is not compact, there exists no maximum for the likelihood function.

Nuisance parameters

For maximum likelihood estimations, a model may have a number of nuisance parameters.

Increasing information

For the asymptotic to hold in cases where the assumption of IID observations does not hold, a basic requirement is that the amount of information in the data increases indefinitely as the sample size increases. Either which may not be met because of too much dependence or new independent observations are subject to an increasing observation error.

Advantages:

i. The method has a good intuitive foundation. The underlying concept is that the best estimate of a parameter is giving the highest probability that the observed set of measurements will be obtained.

ii. The least-squares method and various approaches for combining errors or calculating weighted averages, etc. can be derived or justified in terms of the maximum likelihood approach.

iii. The method is of sufficient generality that most problems are amenable to a straightforward application of this method, even in cases where other techniques become difficult. Inelegant but conceptually simple approaches often provide useful results where there is no easy alternative.

5. Illustrations

Illustrations:

For a random sampling from a normal population N (theta square), find the MLE for

i. Theta when sigma square is known
ii. Sigma square when theta is known
iii. Theta and sigma square when both are unknown

Let x one, x two, up to xn be a random sample of size n drawn from a given population
L is equal to L of (x one, x two, up to xn,theta) is equal to product of f of (xi,theta)
Which is equal to product of 1 by sigma into root 2 phi into exponential to the power (minus 1 by 2 sigma square into (xi minus theta) whole square
Which is equal to (1 by sigma into root 2 phi) to the power n into exponential to the power (minus 1 by 2 sigma square summation (xi minus theta) whole square

Ln L is equal to minus n by 2 into ln sigma square minus n into ln root 2 phi minus 1 by 2 sigma square into summation (xi minus theta) whole square
i) Maximum likelihood estimate of theta when sigma square is known
delta ln L by delta theta is equal to zero implies summation (xi minus theta) by sigma square is equal to zero
which implies summation xi minus n into theta is equal to zero, which implies theta cap is equal to summation xi by n which is equal to x bar
Therefore, MLE of theta θ is theta cap which is equal to summation xi by n which is equal to x bar with delta square ln L by delta theta square less than zero

ii) Maximum likelihood estimate of sigma square when theta is known
delta ln L by delta sigma square is equal to zero implies minus n by 2 sigma square plus summation (xi minus theta) whole square by 2 sigma to the power 4 is equal to zero implies minus n plus summation (xi minus theta) whole square by sigma square is equal to zero
which implies sigma cap square is equal to summation (xi minus theta) whole square by n
Therefore, MLE of sigma square when theta is known as theta not is sigma cap square which is equal to summation (xi minus theta not) whole square by n

iii) Maximum likelihood estimate of theta and sigma square when both are unknown
Delta ln L by delta theta is equal to zero implies summation (xi minus theta) by sigma square is equal to zero implies theta cap is equal to summation xi by n which is x bar. Call this as equation 1
delta ln L by delta sigma square is equal to zero implies minus n plus summation (xi minus theta) whole square by sigma square is equal to zero
implies minus n plus summation (xi minus theta cap) whole square by sigma square is equal to minus n plus summation (xi minus x bar) whole square by sigma square which is equal to zero, from equation (1)
Which implies sigma cap square is equal to summation (xi minus x bar) whole square by n.
Therefore, MLE of theta is theta cap is equal to x bar and the MLE of sigma square when theta is unknown is summation (xi minus x bar) whole square by n

Where we compare the above example with the method of moments, we see that the method

of moment estimators and the maximum likelihood estimators for theta **and** sigma square are the same. However, it happens rarely, if they are not the same.

Due to the fact that the maximum likelihood estimator of **x bar** has an approximate normal distribution with mean **theta** and a variance sigma square by n that is equal to a certain lower bound, thus at least approximately, it is unbiased minimum variance estimator. Accordingly, most statisticians prefer the maximum likelihood estimators than estimators found using the method of moments.

2)Find the MLE of theta in the following uniform distribution U of (0,theta)
Let x one, x two, up to xn be a random sample of size n drawn from an uniform population with probability density function
f of (x, theta) is equal to 1 by theta, 0 less than x less than theta
L is equal to L of (x one, x two, up to xn,theta) is equal to product of f of (xi,theta)
Is equal to (1 by theta) to the power n
Log L is equal to minus n into log theta

Delta log L by delta theta equal to zero implies minus n by theta is equal to zero which implies theta cap is equal to infinity or n is equal to zero, which is meaningless.
Hence, we use the basic principle of the maximum likelihood estimation, where the value of theta for which L is maximum
L is maximum when 1 by theta to the power n is maximum. That is when theta to the power n is minimum, when theta is minimum

But zero less than or equal to x less than or equal to theta
Implies, zero less than or equal to (x one) less than or equal to (x two) less than or equal to up to less than or equal to (xn) less than or equal to theta
Where, (x one) (x two) up to (xn) are order statistics.
Minimum value of theta is the maximum of the observations
Therefore, MLE of theta is theta cap is equal to xn, the maximum of the observations

Here's a summary of our learning in this session, where we understood:

- The principle of method of maximum likelihood estimation
- The applications and advantages of the method
- The properties of the maximum likelihood estimators
- The illustrative examples to obtain MLE