Statistical inference
Part 4
Likelihood, and related, estimators

Michael Goldstein
Durham University
APTS December 2023

Outline

The likelihood principle points us to the central importance of the likelihood function for statistical inference.

In this section, we will explore the statistical properties of the likelihood function and of the maximum likelihood and related estimators.

For simplicity, we will mainly restrict our discussion to problems with a single parameter but these results are true in much greater generality.

Framework

Our framework is as follows.

We want to make inferences about a population, based on a sample from that population.

Our model is that we observe data $\underline{x} = (x_1, ..., x_n)$.

The data is an iid sample of size n from distribution $f(x|\theta)$.

The probability distribution of x depends on the model parameter $\theta \in \Theta$.

We assume that the model is "true", so that only $\theta \in \Theta$ is unknown.

Likelihood

We wish to learn about θ from observations $\underline{x}$.

If we observe $\underline{x}$, then the likelihood function is

$$L(\theta) = L(\theta; \underline{x}) = f(\underline{x} \,|\, \theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

considered as a function of θ for fixed $\underline{x}$.

Log likelihood

We often work with the **log-likelihood**

$$l(\theta) = \log(L(\theta))$$

For a sample of size n, we may write this as

$$l_n(\theta) = l(\theta; \underline{x}) = \log(L(\theta; \underline{x})) = \sum_{i=1}^{n} \log(f(x_i|\theta))$$

so that

$$l_n(\theta) = l(\theta; x_1) + ... + l(\theta; x_n)$$

Maximum likelihood

The **maximum likelihood estimate** (MLE), $\hat{\theta}(x)$, is the value of θ which maximises $L(\theta)$, or equivalently which maximises $l(\theta)$.

When x comprises an independent sample, of size n, we often write the MLE as $\hat{\theta}_n$.

We will mainly be concerned with **regular** problems where $l_n(\theta)$ is differentiable, and the range of the sample space does not depend on θ.

In such cases, $\hat{\theta}_n(x)$ can usually be found as a solution of the **likelihood equation**

$$l_n'(\theta) = 0$$

Example

Suppose $X_1, ..., X_n$ are an independent, identically distributed sample, of size n from an exponential distribution, with parameter θ. The common density is

$$f(x \mid \theta) = \frac{\exp(-x/\theta)}{\theta}, \; 0 < x < \infty, \; 0 < \theta < \infty$$

If $X \sim \exp(\theta)$, then

$$\mathbb{E}(X) = \theta, \;\; \mathsf{V}ar(X) = \theta^2$$

The density for the sample, $\underline{x} = (x_1, ..., x_n)$, is

$$L(\theta) = f(\underline{x} \mid \theta) = \prod_{i=1}^{n} \frac{\exp(-x_i/\theta)}{\theta} = \frac{\exp(-s_n/\theta)}{\theta^n}$$

where $s_n = x_1 + ... + x_n$, the sufficient statistic for the sample.

Example: MLE

Therefore,

$$l_n(\theta) = -nlog(\theta) - \frac{s_n}{\theta}$$

so that

$$l_n'(\theta) = -\frac{n}{\theta} + \frac{s_n}{\theta^2}$$

Therefore, the likelihood equation, $l_n'(\theta) = 0$, has the unique solution

$$\hat{\theta} = \frac{s_n}{n}.$$

Check that this is indeed a maximum from the second derivative

$$l_n''(\theta) = \frac{n}{\theta^2} - 2\frac{s_n}{\theta^3}$$

Example: unbiasedness and consistency

Our estimator is

$$\hat{\theta}_n = \frac{X_1 + ... + X_n}{n}$$

Here are some sampling properties of $\hat{\theta}_n$. As

$$\mathbb{E}(X) = \theta$$

$\hat{\theta}_n$ is an unbiased estimator of θ for all n.

Further, from the strong law of large numbers, almost surely,

$$\hat{\theta}_n \to \theta$$

i.e. $\hat{\theta}_n$ is strongly consistent as an estimator of θ.

Some probability definitions

[See Essentials of Statistical Inference, Young and Smith (2005)]

A sequence of random variables $Y_1, Y_2, ...$ is said to **converge in probability** to real value a if the following:

given $\epsilon > 0, \delta > 0$, there exists an $n_0 = n_0(\delta, \epsilon)$ such that, for all $n > n_0$, $Pr(|Y_n - a| > \epsilon) < \delta$.

A sequence of random variables $Y_1, Y_2, ...$ is said to **converge almost surely** to real value a if the following:

given $\epsilon > 0, \delta > 0$, there exists an $n_0 = n_0(\delta, \epsilon)$ such that $Pr(|Y_n - a| > \epsilon, \text{ for some } n > n_0) < \delta$.

Laws of large numbers

Let $X_1, X_2, ...$ be independent, identically distributed random variables with finite mean μ.

The strong law of large numbers (SLLN)

The sequence of random variables

$$Y_n = \frac{X_1 + \ldots + Xn}{n}$$

converges almost surely to μ, if and only if $\mathbb{E}|X_i|$ is finite.

The weak law of large numbers (WLLN)

If the X_i have finite variance, then Y_n converges to μ in probability.

Asymptotic distribution of MLE 1

The consistency property of MLE in our example is true in great generality. We have the following theorem.

Theorem Under suitable regularity conditions,

$\hat{\theta}_n$ is strongly consistent as an estimator of θ.

[i.e. $\hat{\theta}_n \to \theta$ almost surely.]

This property is interesting in its own right and also as a key step to deriving further properties of the MLE.

Concave and convex functions

To prove this result, we need Jensen's inequality, which is a property of convex and concave functions.

$f(x)$ is a concave function if the chord between any two points on the curve lies below the curve.

Equivalently, f is concave if $f''(x) \leq 0$ over the range of x.

f is convex if the chord lies above the curve or equivalently if $f''(x) \geq 0$

so a straight line is concave and convex.

(Strict concave/convex replaces above inequalities with strict inequalities.)

Jensen's inequality

Jensen's inequality is as follows.

Suppose that $f(x)$ is a real function, X is a bounded random quantity.

If f is a concave function then $\mathbb{E}(f(X)) \leq f(\mathbb{E}(X))$

If f is a convex function then $\mathbb{E}(f(X)) \geq f(\mathbb{E}(X))$

(Strict concave/convex replaces above inequalities with strict inequalities when X is not a constant.)

Jensen's inequality: examples

Example 1

$$f(x) = x^2$$

$$f''(x) = 2 > 0$$

so $f(x)$ is convex so

$$\mathbb{E}(X^2) \geq (\mathbb{E}(X))^2$$

So

$$\mathsf{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \geq 0.$$

Jensen's inequality: examples

Example 2

In decision theory, a person is said to be risk averse if, for any money gamble, they have a preference for the expected money value of the gamble over the gamble itself.

If $U(.)$ is the utility function for random amounts of money X, then by Jensen's inequality, the person is risk averse if and only if U is a concave function so that we must have

$$\mathbb{E}(U(X)) \leq U(\mathbb{E}(X))$$

i.e. the utility of the expected value of the gamble is bigger than the utility of the gamble (which is equal to its expected utility)

Jensen's inequality: examples

Example 3

$$f(x) = \log(x),\ x > 0.$$

As

$$f''(x) = -\frac{1}{x^2} < 0,$$

log is a concave function and so

$$\mathbb{E}(\log(X) \leq \log(\mathbb{E}(X))$$

.

Comment Daniel Bernoulli (1738), discussed the St Petersburg paradox. A ticket promises to pay the bearer $£2^k$ where k is the number of times a fair coin is tossed until it first lands heads. How much is the ticket worth?
The expected payout is infinite, but it clearly is not worth infinity. Bernoulli, discussing decreasing marginal value, suggested maximising "moral expectation" of money, essentially log of money amount.

Outline proof of consistency of MLE

Suppose $f(x|\theta)$ is a family of probability densities or probability mass functions and let θ_0 denote the true value of the parameter θ.

Note that, for any θ, θ_0

$$\mathbb{E}_{\theta_0}(\frac{f(X|\theta)}{f(X|\theta_0)}) = \int (\frac{f(x|\theta)}{f(x|\theta_0)}) f(x|\theta_0) dx = \int f(x|\theta) dx = 1$$

Therefore, for any $\theta \neq \theta_0$, we have by Jensen's inequality

$$\mathbb{E}_{\theta_0} \log(\frac{f(X|\theta)}{f(X|\theta_0)}) \leq \log(\mathbb{E}_{\theta_0}(\frac{f(X|\theta)}{f(X|\theta_0)})) = 0,$$

The inequality is strict unless $\frac{f(X|\theta)}{f(X|\theta_0)} = 1$ (almost everywhere),as a function of X.

Proof ctd.

Fix $\delta > 0$ and let

$$\mu_1 = \mathbb{E}_{\theta_0} log(\frac{f(X|\theta_0 - \delta)}{f(X|\theta_0)}) < 0$$

$$\mu_2 = \mathbb{E}_{\theta_0} log(\frac{f(X|\theta_0 + \delta)}{f(X|\theta_0)}) < 0$$

By the SLLN, as $l_n(\theta) = l(\theta; x_1) + ... + l(\theta; x_n)$,

$$\frac{l_n(\theta_0 - \delta) - l_n(\theta_0)}{n} \to \mu_1 < 0 \ \ (as)$$

Therefore, with probability 1, $l_n(\theta_0 - \delta) < l_n(\theta_0)$, for all n sufficiently large.

Similarly, with probability 1, $l_n(\theta_0 + \delta) < l_n(\theta_0)$, for all n sufficiently large.

Hence, for all n sufficiently large, there exists an estimator $\hat{\theta}_n$ which maximises the log-likelihood on $(\theta_0 - \delta, \theta_0 + \delta)$ for any $\delta > 0$.

Therefore, the MLE is a strongly consistent estimator. □

The score function

We define the **score function** as

$$u(\theta) = u(\theta; \underline{x}) = \frac{\partial}{\partial \theta} l(\theta; \underline{x}) = l'(\theta)$$

As $\underline{x} = (x_1, ..., x_n)$ is an independent sample, then

$$u(\theta; \underline{x}) = u(\theta; x_1) + ... + u(\theta; x_n)$$

The likelihood equation is

$$u(\hat{\theta}) = 0$$

Expectation of score function

We now look at the properties of the score function as X varies.

We write $U(\theta) = u(\theta; X)$. For regular problems, we have

$$\mathbb{E}_\theta(U(\theta)) = 0$$

Proof For each θ

$$\int f(x|\theta)dx = 1$$

Therefore

$$\frac{\partial}{\partial\theta}\int f(x|\theta)dx = \int \frac{f'(x|\theta)}{f(x|\theta)} f(x|\theta)dx = E(U(\theta)) = 0, \quad \square$$

[By regular, we mean here that we can differentiate through the integral.
In particular, we suppose that the range of integration does not depend on θ.]

Variance of the Score function

$$\mathbb{V}ar(U(\theta)) = \mathbb{E}_\theta((U(\theta))^2) = -\mathbb{E}_\theta(\frac{\partial^2 l(\theta; X)}{\partial \theta^2})$$

Proof

$$\begin{aligned}\mathbb{E}_\theta((\frac{\partial^2}{\partial \theta^2} log(f(X|\theta)))) &= \mathbb{E}_\theta(\frac{f(x|\theta)\frac{\partial^2}{\partial \theta^2} f(x|\theta) - (\frac{\partial}{\partial \theta} f(X|\theta))^2}{f^2(X|\theta)}) \\ &= -\mathbb{E}_\theta((\frac{\partial}{\partial \theta} log(f(X|\theta))^2)\end{aligned}$$

as

$$0 = \int \frac{\partial^2}{\partial \theta^2} f(x|\theta) dx = E_\theta(\frac{\frac{\partial^2}{\partial \theta^2} f(X|\theta)}{f(X|\theta)}))\ \square$$

Fisher's information

The variance of the score function, $(\text{Var}(U(\theta)))$ is often written as

$$i(\theta) = -\mathbb{E}(\frac{\partial^2 log(f(X|\theta)}{\partial \theta^2})$$

$i(\theta)$ is termed **Fisher's information**, in a sample of size 1.

$\underline{X} = (X_1, ..., X_n)$, where $X_1, ..., X_n$ are iid from $f(x; \theta)$. Therefore,

$$l_n(\theta) = l(\theta; x_1) + ... + l(\theta; x_n)$$

Therefore $i_n(\theta)$, the information from a sample of size n, is

$$i_n(\theta) = n i_1(\theta)$$

Example: score function

In our example, $X_1, ..., X_n$ are an independent, identically distributed sample, of size n from an exponential distribution, with parameter θ.

We have found that,with $s_n = x_1 + ... + x_n$, the log-likelihood is

$$l_n(\theta) = -nlog(\theta) - \frac{s_n}{\theta}$$

Thefore, the score function is

$$u_n(\theta) = \frac{\partial}{\partial\theta} l_n(\theta) = -\frac{n}{\theta} + \frac{s_n}{\theta^2}$$

As $\mathbb{E}(X_i) = \theta$, we have

$$\mathbb{E}(U_n(\theta)) = 0$$

agreeing with our general result.

Example: Fisher's information

$$u_n(\theta) = \frac{\partial}{\partial\theta} l_n(\theta) = -\frac{n}{\theta} + \frac{s_n}{\theta^2}$$

As $\mathsf{Var}(X_i) = \theta^2$, the variance of the score function, found directly, is

$$\mathsf{Var}(U_n(\theta)) = \mathsf{Var}(\frac{S_n}{\theta^2}) = \frac{n}{\theta^2}$$

Compare the evaluation of the variance using Fisher' information as

$$i_n(\theta) = -\mathbb{E}(l_n''(\theta)) = -(\frac{n}{\theta^2}) + \mathbb{E}(\frac{2S_n}{\theta^3}) = \frac{n}{\theta^2}$$

Example: large sample distribution

In our example

$$\hat{\theta}_n = \frac{X_1 + ... + X_n}{n}$$

So, as $\mathbb{E}(X) = \theta$, $\mathsf{Var}(X) = \theta^2$, with $\theta = \theta_0$, we have

$$\mathbb{E}(\hat{\theta}_n) = \theta_0$$

$$\mathsf{Var}(\hat{\theta}_n) = \frac{\theta_0^2}{n} = \frac{1}{n i_1(\theta_0)}$$

and the large sample distribution of $\hat{\theta}_n$ is approximately Gaussian, from the central limit theorem, with the above mean and variance.

Asymptotic distribution of MLE 2

The large sample properties of MLE in our example are true in great generality.

We have the following theorem describing the large sample behaviour of the MLE

Theorem Under suitable regularity conditions, for given value θ_0, for large n,

the distribution of $\hat{\theta}_n$ is approximately normal,

with mean θ_0

and variance

$$\frac{1}{i_n(\theta_0)} = \frac{1}{ni_1(\theta_0)}$$

Reminder: Taylor expansions

The **Taylor expansion** for a function $f(x)$ of a single real variable about $x = a$ is given by

$$f(x) = f(a) + f^{(1)}(a)(x-a) + \tfrac{1}{2!}f^{(2)}(a)(x-a)^2 + .. + \tfrac{1}{n!}f^{(n)}(a)(x-a)^n + R_n,$$

where $f^{(k)}(a)$ is the k^{th} derivative of $f(x)$ evaluated at $x = a$, and the remainder is

$$R_n = \frac{1}{(n+1)!} f^{(n+1)}(c)(x-a)^{n+1}$$

for some $c \in [a, x]$.

Outline proof of asymptotic normality of MLE

We assume that $l_n(\theta)$ is twice differentiable on a neighbourhood of θ_0.

From the proof of consistency of the MLE, there exists a sequence of local maxima $\hat{\theta}_n$ such that $l_n'(\hat{\theta}_n) = 0$, and $\hat{\theta}_n \to \theta_0$, almost surely.

We carry out a Taylor expansion of the score function

$$u(\theta, \underline{x}) = \frac{\partial}{\partial \theta} l(\theta; \underline{x})$$

around θ_0, evaluated at $\hat{\theta}$

$$0 = u(\hat{\theta}) \approx u(\theta_0) + (\hat{\theta} - \theta_0) u'(\theta_0)$$

So

$$\hat{\theta} - \theta_0 \approx -\frac{u(\theta_0)}{u'(\theta_0)}$$

Proof continued

$$\hat{\theta} - \theta_0 \approx -\frac{u(\theta_0)}{u'(\theta_0)}$$

As

$$u(\theta; \underline{x}) = u(\theta; x_1) + ... + u(\theta; x_n)$$

by the CLT, approximately,

$$u(\theta_0) \sim N(0, ni_1(\theta_0))$$

By SLLN,

$$u'(\theta_0) \approx \mathbb{E}(u'(\theta_0)) = -ni_1(\theta_0)$$

and the result follows. □

Cramer-Rao lower bound

Let $W(X)$ be any estimator of θ. Let

$$m(\theta) = \mathbb{E}_\theta(W(X)).$$

The Cramer-Rao lower bound (CRLB) for the variance of $W(X)$ for a regular likelihood is

$$\mathsf{Var}(W(X)) \geq \frac{(m'(\theta))^2}{i(\theta)}$$

In particular, if $W(X)$ is an unbiased estimator for θ, so that $m(\theta) = \theta$, then the CRLB is

$$\mathsf{Var}(W(X)) \geq \frac{1}{i(\theta)}$$

Unbiased estimators which achieve this lower bound are termed **efficient.**

Proof of bound

$$\begin{aligned}\mathsf{Cov}(W(X), U(\theta, X)) &= \int w(x)\frac{\partial}{\partial\theta} log(f(x;\theta)f(x;\theta)dx \\ &= \frac{\partial}{\partial\theta}\int w(x)f(x;\theta)dx = m'(\theta)\end{aligned}$$

Therefore as

$$(\mathsf{Cov}(W(X), U(\theta, X)))^2 \leq \mathsf{Var}(W(X)), \mathsf{Var}(U(\theta, X))$$

and

$$\mathsf{Var}(U(\theta, X)) = i(\theta)$$

it follows that

$$\mathsf{Var}(W(X) \geq \frac{(m'(\theta))^2}{i(\theta)} \quad \square$$

Example: efficiency

In our example, we have already shown that, for all n, our estimator $\hat{\theta}$ is unbiased with variance equal to $\frac{1}{i(\theta)}$.

Therefore $\hat{\theta}$ is efficient, for all sample sizes.

Comment If we had parametrised the density as

$$f(x|\theta) = \theta exp(-\theta x),\ x > 0$$

,
then the MLE would not attain the Cramer Rao lower bound. [Check!]

[For the Cramer Rao lower bound to be attained, we need to be sampling from the general exponential family of distributions with a particular, natural parametrisation.]

Discussion: large samples

We have shown that, under suitable regularity conditions, for large n, the distribution of $\hat{\theta}_n$ is approximately normal, with mean θ_0 and variance $\frac{1}{i(\theta_0)}$.

Therefore, asymptotically, $\hat{\theta}_n$ is an unbiased estimator which achieves the Cramer-Rao lower bound, and so is "asymptotically efficient".

Therefore, we have that

(i) $\hat{\theta}_n$ is asymptotically consistent

(ii) $\hat{\theta}_n$ is asymptotically unbiased

(iii) $\hat{\theta}_n$ is asymptotically efficient.

And, we have more (and better!) large sample properties to come!

All this explains why usually maximum likelihood is a strong method for large samples.

Discussion: small samples

For small samples, the situation is much more complicated.

Plotting the likelihood function should give you some idea as to how much evidence is provided by your sample.

[Unimodal or multimodal? Concentrated around the peak or diffuse? maximised in a scientifically plausible region?]

Simulation experiments can help you to decide whether your sample size is big enough to rely on the large sample approximations and whether your sample outcomes are typical reflections of the large sample properties.

[Simulate repeatedly from the model, assumed true, for a range of different parameter values.]

[While you are doing these simulations, add in some simulations from "noise-modified" versions of the model to see how robust your conclusions are.]

Observed information

Given a sample $\underline{x}$, we may estimate the information $i(\theta)$ by $i(\hat{\theta})$.

An alternative estimate is based on the idea of observed information.

The quantity

$$j(\theta) = -\frac{\partial^2 log(f(\underline{x}|\theta))}{\partial \theta^2}$$

for the observed sample $\underline{x}$ is called the **observed information**.

Note that

$$i(\theta) = \mathbb{E}_\theta(j(\theta))$$

We estimate $j(\theta)$ from sample $\underline{x}$ as $j(\hat{\theta})$.

Example ctd: estimating information

In our exponential distribution example, we found that $i(\theta) = \frac{n}{\theta^2}$

Given a sample, we may estimate the information using the MLE, $\hat{\theta} = \frac{s_n}{n}$ as

$$i(\hat{\theta}) = \frac{n^3}{s_n^2}$$

Compare the observed information, which we evaluate from

$$j(\theta) = -l''(\theta) = -(\frac{n}{\theta^2} - 2\frac{s_n}{\theta^3})$$

as

$$j(\hat{\theta}) = -\frac{n^3}{s_n^2} + 2\frac{n^3 s_n}{s_n^3} = \frac{n^3}{s_n^2}$$

Comparing observed and expected information

There is an extensive literature on which is the better choice. In many situations observed information is more reliable.

“ At first sight it may seem that it would be preferable to use the theoretical Fisher information matrix, if it were easy to calculate, but in fact an extensive body of theory and practice suggests that in most cases the inverse of the observed information matrix gives a better approximation to the true covariance matrix of the estimators, and this is therefore the preferred method in most applications. An especially important reference here is Efron and Hinkley (1978).”

[See Essentials of Statistical Inference, Young and Smith]

MLE summary

The MLE, $\hat{\theta}$, is the solution of the equation

$$u(\theta) = 0$$

where $u(\theta) = l'(\theta; \underline{x})$.

Note that

$$\mathbb{E}(U(\theta)) = 0, \ \forall \theta$$

$\hat{\theta}$ is a consistent estimator with, asymptotically,

$$\hat{\theta} \sim \mathsf{N}(\theta, \frac{1}{\mathbb{E}(U^2)})$$

MLE issues

In some problems, either

(i) we do not have access to the whole likelihood function, but only to some properties of it, or

(ii) we have a likelihood function, but we are concerned with the robustness of inferences based on assuming that the likelihood form is exactly correct.

For such problems we may consider estimates based on a wider class of **estimating equations** which share many of the large sample properties of the MLE.

Estimating equations

For data $\underline{x}$ and parameter θ, we define the estimator θ^* as the solution to the estimating equation

$$h(\theta; \underline{x}) = 0$$

We often restrict attention to **unbiased** estimating equations, that is

$$\mathbb{E}(h(\theta; \underline{X})) = 0, \forall \theta$$

Large sample properties of estimating equations

Under the same regularity assumptions as before, θ^* is a consistent estimator with, asymptotically,

$$\theta^* \sim \mathsf{N}(\theta, \frac{\mathbb{E}(h^2)}{(\mathbb{E}(h'))^2})$$

where

$$\frac{\mathbb{E}(h^2)}{(\mathbb{E}(h'))^2} \geq \frac{1}{\mathbb{E}(U^2)} = \frac{1}{i(\theta)}$$

(equality if $h = U$)

Estimating equations: example

Suppose that X_i is a count of events occuring in an interval of length t_i.

We might model each X_i as Poisson parameter $t_i\theta$.

However, we might consider that the counts could be over-dispersed relative to the Poisson.

A simple modification would be to suppose that

(i) $\mathbb{E}(X_i) = t_i\theta$

(ii) $\mathsf{Var}(X_i) = \delta t_i\theta$

where δ is an "overdispersion" parameter.

We will use (i) to set up our estimating equation - there will be many possible choices.

Use (ii) to choose between these choices.

Example continued

Possible unbiased estimating equations, using $\mathbb{E}(X_i) = t_i\theta$, are of form

$$h(\theta, \underline{X}) = \sum_{i=1}^{n} c_i(X_i - t_i\theta)$$

We want to minimise the large sample variance of θ^*, which given (i) and (ii) is

$$\frac{\mathbb{E}(h^2)}{(\mathbb{E}(h'))^2} = \frac{\sum_i c_i^2 \delta t_i \theta}{(\sum_i c_i t_i)^2}$$

This is minimised when all of the c_i are equal, so that the estimator is

$$\theta^* = \frac{\sum_i X_i}{\sum_i t_i}$$

This is the same estimator as if we had assumed a Poisson distribution and used MLE.

However the variance will differ (by a factor of δ) and the properties of the estimator will not rely on the Poisson assumption.