Inferential Statistics
L3 - The likelihood function

Erlis Ruli (erlis.ruli@unipd.it)

Department of Statistics, University of Padova

Contents

1. The likelihood function
2. The information
3. The information matrix

Overview

The likelihood function is the cornerstone of statistical inference, in virtually all its varieties.

We are going to introduce the likelihood function descriptively, and illustrate it by means of practical examples.

We'll see how it's used for inferential purposes in the incoming lectures.

Definition

Let $X_1, \ldots, X_n$ be an iid sample with X_i having pdf $f(x; \theta)$. Then the likelihood function is $L(\theta) : \Theta \rightarrow R_{\geq 0}$ defined by

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

The likelihood, thus, maps θ to a non-negative real number, while holding the data fixed.

For an observed sample $x_i, \ldots, x_n$, it's defined as

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

The resemblance with the joint of the sample is remarkable, but they are different things!

- the joint pdf maps $x_1, \ldots, x_n$ to a non-negative real, holding θ fixed,
- $L(\theta)$ maps θ to a non-negative real, holding $x_1, \ldots, x_n$ fixed
- the joint pdf always integrates to 1
- $L(\theta)$ is not a density thus it may or may not integrate to 1.

Let's go through some examples...

Example 1

Let $X_1, \ldots, X_n$ be an iid sample with $X_i \sim \text{Ber}(\theta)$. Then

$$L(\theta) = \prod_{i=1}^{n} \theta^{X_i}(1-\theta)^{1-X_i}$$
$$= \theta^{\sum_i X_i}(1-\theta)^{n-\sum_i X_i}$$

Suppose now that $n = 10$ and let the list $0, 1, 1, 1, 0, 1, 1, 0, 1, 1$ be an observed sample.

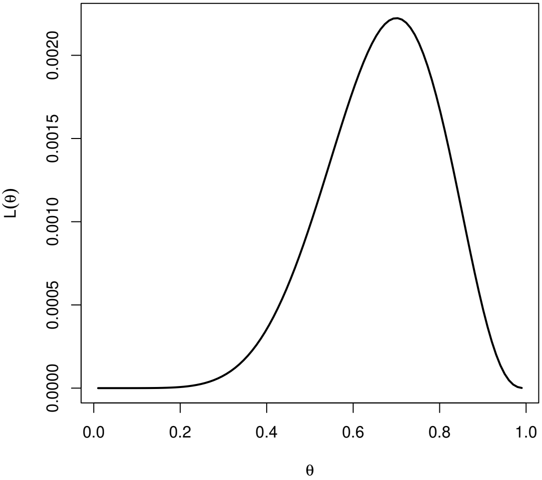
The likelihood function is

$$L(\theta) = \theta^7(1-\theta)^3$$

By the way, (check that) $\int_0^1 L(\theta)d\theta = \cancel{\frac{1}{320}}. \frac{1}{1320}$

Here is why $L(\theta)$ bears that weird name...

Example 1 (cont'd)



On the interpretation of L

For a sample $x_1, \dots, x_n$ of discrete rv's, $L(a)$ can be interpreted as:

> The probability of observing that sample under the chosen model when the parameter θ equals a.

If the rv's are continuous, this interpretation is not correct. In this case, we can only say that

> the higher the value of L the more likely is it to observe the sample.

In general, $L(a)$, gives us the likelihood of observing the sample when $\theta = a$.
It's natural then to look for θ with largest L. We call this maximum likelihood estimate, denoted $\widehat{\theta}$ and defined by

$$\widehat{\theta} = \arg\sup_{\theta \in \Theta} L(\theta).$$

Random L vs observed L

Example 2

Coins for the gambling industry come in three types U1, U2, F, and all have two faces: W (Win) and L (Loose).

U1-types have $P(W) = 1/3$, U2-types have $P(W) = 1/4$ and F-types have $P(W) = 1/2$.

Nature picks one at random from the three available, and tosses it three times. If we let $\theta = P(W)$, then

$$P(WWW) = \theta^3, \quad P(LLL) = (1-\theta)^3,$$
$$P(WWL) = \theta^2(1-\theta), \quad P(WLL) = \theta(1-\theta)^2.$$

The next table gives the sample points and the associated probabilities for this experiment, for each θ.

Example 2 (cont'd)

	Type of coins		
	$\theta = 1/4$	$\theta = 1/3$	$\theta = 1/2$
sample	(type U2)	(type U1)	(Type F)
WWW	0.0156	0.0370	$0.125(\bullet)$
WWL	0.0469	0.0741	$0.125(\bullet)$
WLW	0.0469	0.0741	$0.125(\bullet)$
LWW	0.0469	0.0741	$0.125(\bullet)$
WLL	0.1406	$0.1482(\bullet)$	0.125
LWL	0.1406	$0.1482(\bullet)$	0.125
LLW	0.1406	$0.1482(\bullet)$	0.125
LLL	$0.4219(\bullet)$	0.2963	0.125

Each column is a probability distribution, each row is an observed likelihood function ($\bullet$ at its max), so here we can observed at most $2^3 = 8$ likelihood functions.

What's L useful for?

Most useful L are those with an infinite domain Θ, and infinite co-domain.

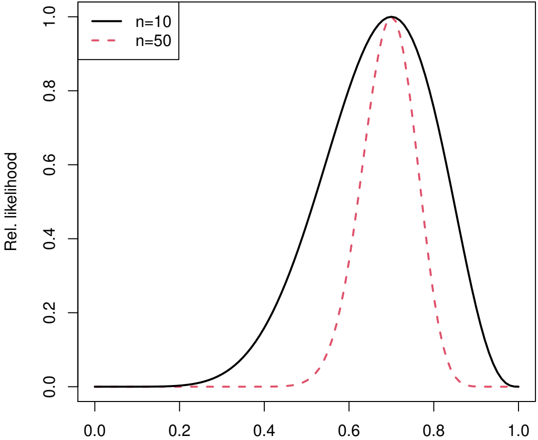
domain of log L != 0

Furthermore, if $L > 0$ for all $\theta \in \Theta$, then for all our purposes, working with $\log L(\theta) = \ell(\theta)$ will make our lives much easier.

A full answer to the title will be given in L4, L5 and L6, for the time being, here is a partial answer.

Example 3: two (scaled) L with different n

Suppose we have another observed sample as in Example 1, but with $n = 50$. The sample with $n = 50$ is more 'informative' since the interval of plausible values, (0.5, 0.85), is narrower.



The observed information

There is a more precise way quantify the informativeness of a likelihood function: the observed information.

This is denoted by $J(\theta)$ (or $J_n(\theta)$ when it's important to emphasise n) and is defined as

$$J(\theta) = -\frac{d^2\ell(\theta)}{d\theta^2}.$$

It turns out that $0 \leq J$ and the higher J the higher the peakedness of the likelihood.

For example, we saw in Example 1 we had $L(\theta) = \theta^7(1-\theta)^3$ and in Example 3, 1 is observed 35 times. In both cases, $\widehat{\theta} = 7/10$. It turns out that

$$J_{10}(\widehat{\theta}) = 47.6 < J_{50}(\widehat{\theta}) = 238.1.$$

Computation of $\widehat{\theta}$

In Example 1 we said $\widehat{\theta} = 0.7$. To compute it we

(i) compute gradient of the log-likelihood

(ii) find θ^* s.t. $\ell'(\theta^*) = 0$; this is also called likelihood equation

(iii) check that $J(\theta^*) > 0$, if so set $\widehat{\theta} = \theta^*$.

Step (iii) only guarantees that $\widehat{\theta}$ is a local maximum. To assess if θ is a global maximum further effort is required.

Following these steps we have $\ell'(\theta) = \frac{7}{\theta} - \frac{3}{(1-\theta)}$, with solution $\widehat{\theta} = 0.7$.

If analytical solution of the likelihood equation is not feasible, we can resort to numerical root-finding methods.

Among them, Newton-Raphson is perhaps the most widely known. The idea is to build a sequence $\tilde{\theta}_1, \tilde{\theta}_2, \dots$ s.t. it converges to the solution $\widehat{\theta}$.

In particular, given $\tilde{\theta}_m$, the next term in the sequence is defined recursively

$$\tilde{\theta}_{m+1} = \tilde{\theta}_m + \frac{\ell'(\tilde{\theta}_m)}{J(\tilde{\theta}_m)}, \quad m = 0, 1, 2, \dots,$$

and $\tilde{\theta}_0$ is a starting value.

A stopping condition must be imposed in order to arrive at a practical solution.

L for a vector-valued parameter

Example 3

The likelihood may be multivariate. For instance,

let $X_1, \dots, X_n$ be an iid random sample with $X_i \sim \text{Wei}(\alpha, 1/\beta)$; note $\theta = (\alpha, \beta)$. Suppose, for example, we have

$$5.1, 7.4, 10.9, 21.3, 12.3, 15.4, 25.4, 18.2, 17.4, 22.5,$$

a sample of waiting times on the Poste Italiane's Customers Serivce telephone exchange.

We want to plot the likelihood function of these observed data.

Example 3 (cont'd)

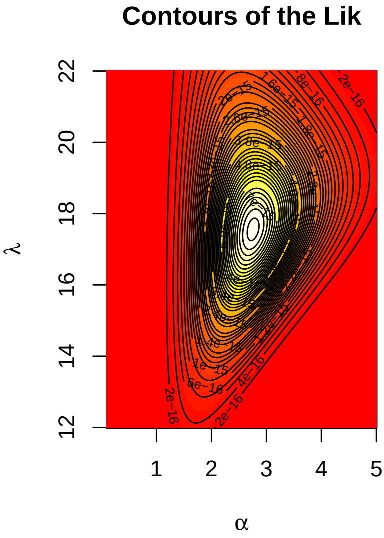
The likelihood function is $L(\alpha, \beta) : \mathbb{R}_{>0} \times \mathbb{R}_{>0} \to \mathbb{R}_{>0}$.

This is 3-d surface, thus we need a different plotting strategy. Below we see the contours of this surface.

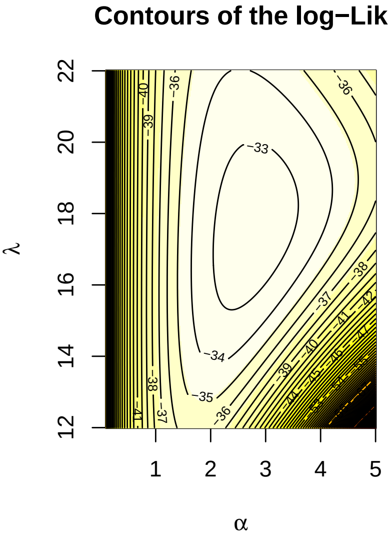
The contours are obtained by "cutting" the likelihood surface horizontally at some pre-specified points. The cut is then projected on the horizontal plane.

Sometimes it may be easier to visualize the log-likelihood surface instead.

Example 3 (cont'd)



Contours of the Lik



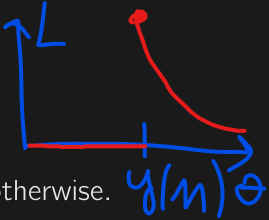
Contours of the log−Lik

A nasty likelihood

Example 4

Let $X_1, \dots, X_n$ be an iid random sample with $X_i \sim \mathrm{Unif}(0, \theta)$, $\theta \in \mathbb{R}_{>0}$. The joint pdf is the product of the marginals, thus the statistical model is

$$\left\{ \prod_{i=1}^{n} \tfrac{1}{\theta} \mathbf{1}_{[0,\theta]}(x_i) : \theta \in \mathbb{R}_{>0} \right\},$$



where $\mathbf{1}_{(0,\theta)}(x_i)$ takes on value 1 if $x_i \in [0, \theta]$ and 0 otherwise.

The likelihood function is

$$L(\theta) = \begin{cases} 1/\theta^n & \text{if } x_{(n)} \leq \theta \\ 0 & \text{otherwise.} \end{cases}$$

In this case we cannot compute the log-likelihood since $L(\theta)$ may be zero. Graph?

Observed information for vector-valued θ

For vector-valued θ, the observed information is the matrix

$$J(\theta) = (-1) \begin{pmatrix} \partial^2 \ell(\theta)/\partial\theta_1^2 & \partial^2 \ell(\theta)/\partial\theta_1\partial\theta_2 & \cdots & \partial^2 \ell(\theta)/\partial\theta_1\partial\theta_p \\ \partial^2 \ell(\theta)/\partial\theta_2\partial 1 & \partial^2 \ell(\theta)/\partial\theta_2^2 & \cdots & \partial^2 \ell(\theta)/\partial\theta_2\partial\theta_p \\ \vdots & \vdots & \vdots & \vdots \\ \partial^2 \ell(\theta)/\partial\theta_p\partial\theta_1 & \partial^2 \ell(\theta)/\partial\theta_p\partial\theta_2 & \cdots & \partial^2 \ell(\theta)/\partial\theta_p^2 \end{pmatrix},$$

It's clear that J is symmetric; alternate notation is $J(\theta) = [J(\theta)_{ij}] = [-\partial^2 \ell(\theta)/(\partial\theta_i\partial\theta_j)]$

In the sequel we'll denote:

- by $J(\theta)_{ij}$ the cell i, j of J,
- by $J(\theta)^{ij}$ the i, j cell of J^{-1} and
- $\widehat{J} = J(\widehat{\theta})$.