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)

sample	Type of coins		
	$\theta = 1/4$ (type U2)	$\theta = 1/3$ (type U1)	$\theta = 1/2$ (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?

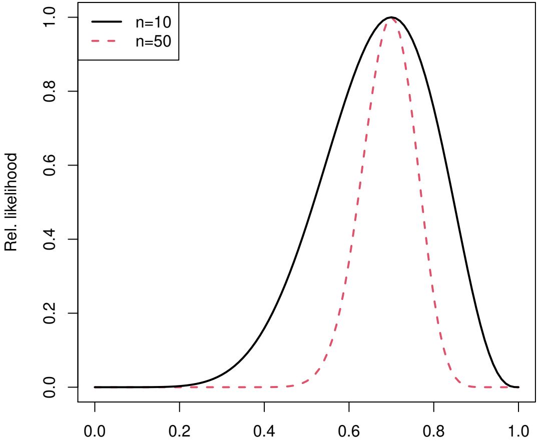
Most useful L are those with an infinite domain Θ, and infinite co-domain.

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.

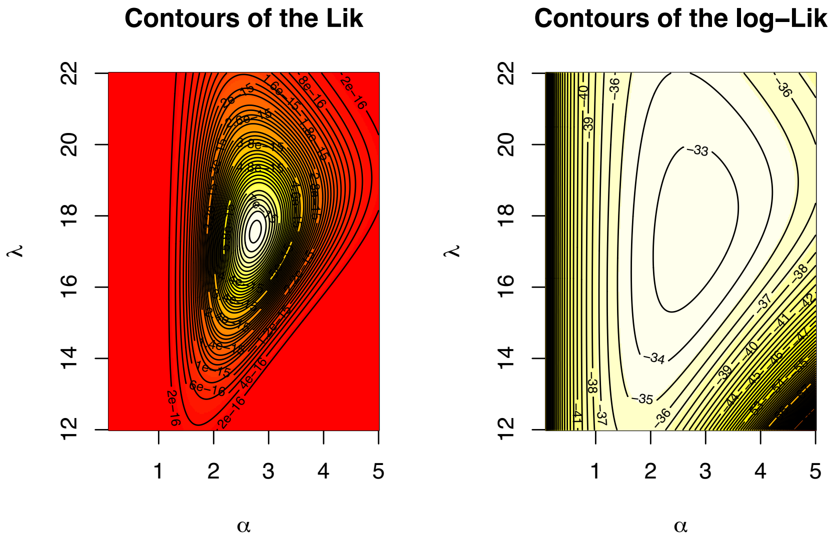
Example 3 (cont'd)

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.


Contours of the Lik
Contours of the log–Lik
λ
α
22
20
18
16
14
12
1
2
3
4
5
2e−16
6e−16
1e−15
1.4e−15
4e−16
2e−16
8e−16
1.6e−15
2e−15
2.6e−15
3.8e−15
4.8e−15
−33
−34
−35
−36
−37
−38
−39
−40
−41
−42

A nasty likelihood

Example 4

Let $X_1, \ldots, X_n$ 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
- $\widehat{J} = J(\widehat{\theta})$.

Inferential Statistics
L4 - Point estimation

Erlis Ruli (erlis.ruli@unipd.it)

Department of Statistics, University of Padova

Contents

Overview

Suppose $Y_1, \ldots, Y_n$ is a random sample with $Y_i \sim F_\theta$ and,

Nature picks $\theta = \theta_0$ (secretly) and uses it generate the observed sample $y_1, \ldots, y_n$ from the above random sample.

With this observed sample at hand, one of the aims of statistics is to guess θ_0.

Such a guess is called an estimate of the unknown parameter θ_0. In this lecture we'll study methods estimating a parameter.

In the first part of this lecture we will see what properties we wish our estimates should satisfy. In the second part we will see methods for building such estimates.

Statistics (dejavu')

Let $Y_1, \ldots, Y_n$ be a random sample with $Y_i \sim F_\theta$, with pdf f_θ and unknown parameter θ.

If $T_n = T(Y_1, \ldots, Y_n)$, with $T_n : \mathbb{R}^n \to \mathbb{R}^d$ doesn't depend on any unknown quantity, then its is called a statistic.

All summary statistics we saw in L2 are all examples of statistics.

In L2 we didn't pay much attention, but T_n is a rve, thus it has a df.

The point is that, when T_n is chosen with care, it reveals us something useful about θ.

Example 1

For the iid random sample $Y_1, \ldots, Y_n$, assume that $E(X_i) = \mu$, and $\text{var}(X_i) = \sigma^2$. Then, the sample average $\overline{Y} = (Y_1 + \cdots + Y_n)/n$ is a statistic and

$$E(\overline{Y}) = E((Y_1 + \cdots + Y_n)/n) = n^{-1}E(Y_1 + \cdots + Y_n) = \mu,$$

and

$$\text{var}(\overline{Y}) = \sigma^2/n$$

Thus if we are interested in learning the expected value of a population, i.e. $\theta = \mu$, then the sample average is a good candidate.

Furthermore, let $Y_i \sim N(\mu, \sigma^2)$, then we can show that

$$\overline{Y} \sim N(\mu, \sigma^2/n) \quad \text{or} \quad \frac{\sqrt{n}(\overline{Y}-\mu)}{\sigma^2} \sim N(0, 1).$$

Note that, because μ, σ^2 are unknown, $\frac{\sqrt{n}(\overline{Y}-\mu)}{\sigma^2}$ is not a statistic

If this doesn't make much sense to you, let's make it more concrete. Assume that F_θ is discrete and Y can assume values in $\{1, 2, 3\}$ with equal probability; so $\mu = 2$.

Let $n = 2$, and consider Y_1, Y_2 iid sample from F_θ. The possible observed samples are

$$1, 1; 1, 2; 1, 3; 2, 1; 2, 2; 2, 3, 3, 1; 3, 2; 3, 3.$$

Using the distribution of the sample averages (below) we find that the average of the sample averages is
$E(\overline{Y}) = 1 \cdot \frac{1}{9} + 1.5 \cdot \frac{2}{9} + 2 \cdot \frac{3}{9} + 2.5 \cdot \frac{2}{9} + 3 \cdot \frac{1}{9} = 2 = \mu.$

$\overline{Y}$	$P(\overline{Y} = k)$
1	1/9
1.5	2/9
2	3/9
2.5	2/9
3	1/9

Average of sample variance = population variance

Example 2

Under the assumptions of Example 1, let $\widehat{\sigma}^2 = n^{-1}\sum_i (Y_i - \overline{Y})^2$ be (a version of) the sample variance. Then

$$E(\widehat{\sigma}^2) = E[(Y_1 - \overline{Y})^2] = \tfrac{n-1}{n}\sigma^2.$$

On the other hand for the sample variance $S^2 = (n-1)^{-1}\sum_i (Y_i - \overline{Y})^2$, we have

$$E(S^2) = E[n\widehat{\sigma}^2/(n-1)] = \tfrac{n}{n-1}\tfrac{n-1}{n}\sigma^2 = \sigma^2.$$

This is the reason why we defined S^2 dividing by $n-1$. It can be show that

$$\text{var}(S^2) = \tfrac{\mu_4 - \mu_2^2}{n} - \tfrac{2(\mu_4 - 2\mu_2^2)}{n^2} + \tfrac{\mu_4 - 3\mu_2^2}{n^3},$$

where $\mu_k = E(Y_1^k)$, is the kth moment of Y_1.
In general $\overline{Y}$ and S^2 are not independent, except if $Y_i \sim N(\mu, \sigma^2)\ldots$