Inferential Statistics **First semester, 2021**

Lecture 3: The likelihood function

Instructor: Erlis Ruli (ruli@stat.unipd.it), Department of Statistical Sciences

The likelihood function is one of the most important tools of statistical inference, both in the frequentist and in the Bayesian setting. In this lecture we introduce the likelihood function from a pure descriptive perspective and we illustrate its use by means of practical examples. In the incoming lectures we will see how the likelihood function is used for conducting inference on the parameters of a statistical model. Near the end of this course we will see how the likelihood function is used in the Bayesian setting.

3.1 The likelihood function

Definition 3.1 *Let $Y_1, \ldots, Y_n$ be a random sample of size n, with $Y_i \sim F_\theta$, $\theta \in \Theta$ independently for each $i = 1, \ldots, n$ and let $f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \theta) = \prod_{i=1}^n f(y_i; \theta)$ be the joint probability distribution of the sample. For a fixed sample $y_1, \ldots, y_n$, the likelihood function, denoted $L(\theta)$, with $L(\theta) : \Theta \to \mathbb{R}_{\geq 0}$ is defined by*

$$L(\theta) = L(\theta; y_1, \ldots, y_n) = f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \theta) = \prod_{i=1}^n f(y_i; \theta).$$

Thus the likelihood function is obtained by holding fixed $y_1, \ldots, y_n$ and letting θ vary in Θ, the space of admissible values for θ. Θ is also called the parameter space.

For a given parameter value, say $\theta = \theta_0$, the likelihood function can be interpreted as the probability of observing a sample like the one actually observed. Different parameter values lead to different probabilities, i.e. likelihoods. The higher is this probability the more likely is to observe data such as those actually observed, under the model F_θ. That is, if $L(\theta_1) > L(\theta_2)$, then θ_1 is more likely to generated the observed data than θ_2. Thus, it is reasonable to look for the value of θ which maximise $L(\theta)$; note that the likelihood function typically takes on very small values.

We start illustrating the likelihood function by an example in which the parameter space is finite.

Example 3.1 *A manufacturer of coins for the gambling industry produces three types of coins U1, U2, F. All coins have two faces: L (Win) and L (Loose). It holds that for a coin of type U1, $P(W) = 1/3$, coins of type U2 are such that $P(W) = 1/4$ and coins of type F are such that $P(W) = 1/2$.*

unfair coin

Nature picks a coin at random from the three types, without revealing to us the type, and tosses the coin three times. If we denote by $\theta = P(W)$, then $P(WWW) = \theta^3$, $P(LLL) = (1-\theta)^3$, $P(WWL) = \theta^2(1-\theta)$ and

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(•)
WWL	0.0469	0.0741	0.125(•)
WLW	0.0469	0.0741	0.125(•)
LWW	0.0469	0.0741	0.125(•)
WLL	0.1406	0.1482(•)	0.125
LWL	0.1406	0.1482(•)	0.125
LLW	0.1406	0.1482(•)	0.125
LLL	0.4219(•)	0.2963	0.125

best Prob. Distr. in each

Table 3.1: Likelihood function of θ for the problem of tossing three coins.

$P(WLL) = \theta(1-\theta)^2$. Table 3.1 gives the sample points and the associated probabilities for this experiment, for each value of θ.

Reading the table column-wise we get the usual probability distribution of the simple events, i.e. summation of rows of a given column gives 1. Hence in the table we see three different probability distributions. The likelihood function is obtained by reading the table row-wise; thus in this case we have four different likelihood functions.

The likelihood functions are read as follows. Suppose, we observed the sample $\{WWW\}$. Then for $\theta = 1/4$ the likelihood, i.e. the probability, of observing this sample is 0.0156. For $\theta = 1/2$ such a probability is equal to 0.125. This observed sample is more likely to be observed if $\theta = 1/2$, so the coin chosen by Nature is more likely to be of type F.

If instead we observed $\{LLL\}$, then it is more likely that the coin is of type U2.

leave to read to us

Remark 3.1

(i) Whatever is our deduction, i.e. inference, from the observed sample and the likelihood function, we will never know Nature's choice. That is, we will never know the truth. However, as we will see below, with increasing amounts of data we can be increasingly more confident about our deduction.

(ii) In practice we have only one observed sample, or few observed samples if we have enough funds, but certainly we cannot observe the complete sample space. Thus, in practice, we can analyse only a single row (or few rows) of the above table, that is, we can only deal with one likelihood function.

(iii) With simulated data we do know the truth and in this case we can study precisely the behaviour of maximum of the likelihood function. For instance, in this example, we see that the the maximum of the likelihood function has a discrete distribution with support $\{1/4, 1/3, 1/2\}$.

(iv) Often the likelihood function can be factored as $L(\theta) = w(y_1, \ldots, y_n)q(\theta; y_1, \ldots, y_n)$, where $w(y_1, \ldots, y_n)$ is a function of the sample and $q(\theta; y_1, \ldots, y_n)$ is a function of both the parameter θ and the sample

$y_1, \ldots, y_n$ or a function of it. If this is the case, then we can discard $w(y_1, \ldots, y_n)$ and write $L(\theta) \propto q(\theta; y_1, \ldots, y_n)$, where the symbol "$\propto$" means "proportional to".

(v) When $L(\theta) \in \mathbb{R}_{>0}$, we can apply the natural logarithm and get the log-likelihood function $\ell(\theta) = \log L(\theta)$. If the remark (iv) applies, then we can write $\ell(\theta) = \log q(\theta; y_1, \ldots, y_n) + \text{const.}$, where const. means not a function of θ.

Exercise. *Let $\widehat{\theta}$ denote the value of θ at which the maximum of the likelihood function is achieved. Remark 3.1(iii) and Table 3.1 suggest that $\widehat{\theta}$ is a r.v.. Determine its distribution.*

In most practical problems the parameter space Θ is infinite and uncountable, i.e. $\Theta \subseteq \mathbb{R}^d$. Here are some examples of this kind.

Example 3.2 *Let $Y_1, \ldots, Y_n$ be a random sample as in Example 2.3 (Lecture 2). Then the likelihood function is*

$$L(\lambda) \propto e^{-n\lambda} \lambda^{\sum_{i=1}^n y_i},$$

and the log-likelihood is

$$\ell(\lambda) = -n\lambda + \sum_{i=1}^n y_i \log(\lambda) + \text{const.}.$$

Example 3.3 *Let $Y_1, \ldots, Y_n$ be a random sample as in Example 2.4 (Lecture 2). Then the likelihood function is*

$$L(\theta) \propto \theta^{\sum_{i=1}^n y_i} (1-\theta)^{n - \sum_{i=1}^n y_i},$$

and the log-likelihood is

$$\ell(\lambda) = \sum_{i=1}^n y_i \log(\theta) + \left(n - \sum_{i=1}^n y_i \right) \log(1-\theta) + \text{const.}.$$

Example 3.4 *Let $Y_1, \ldots, Y_n$ be a random sample as in Example 2.2 (Lecture 2). Then the likelihood function is*

$$L(\mu, \sigma^2) \propto \frac{1}{(\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2}$$

and the log-likelihood function is

$$\begin{aligned} \ell(\mu, \sigma^2) &= -\tfrac{n}{2} \log(\sigma^2) - \tfrac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 + \text{const.} \\ &= -\tfrac{n}{2} \log(\sigma^2) + \tfrac{n\mu^2}{2\sigma^2} - \tfrac{1}{2\sigma^2} \sum_{i=1}^n y_i^2 + \overline{y} \tfrac{n\mu}{\sigma^2} + \text{const.}. \end{aligned}$$

Note that here the likelihood function is $L(\mu, \sigma^2) : \mathbb{R} \times \mathbb{R}_{>2} \to \mathbb{R}_{>0}$.

Example 3.5 *Let* $Y_1, \ldots, Y_n$ *be an i.i.d. random sample with* $Y_i \sim \text{Wei}(\alpha, \beta)$. *The joint distribution of this random sample is given by the product of their marginal distributions, thus the statistical model is*

$$\left\{ \frac{\alpha^n}{\beta^n} \left(\prod_{i=1}^n y_i \right)^{\alpha-1} e^{-\sum_{i=1}^n y_i^\alpha/\beta} : (\alpha, \beta) \in \mathbb{R}^2_{>0} \right\}.$$

It follows that the likelihood function is

$$L(\alpha, \beta) \propto \frac{\alpha^n}{\beta^n} \left(\prod_{i=1}^n y_i \right)^{\alpha} e^{-\sum_{i=1}^n y_i^\alpha/\beta}$$

and the log-likelihood function is

$$\ell(\alpha, \beta) = n(\log \alpha - \log \beta) + \sum_{i=1}^n \log(y_i)\alpha - \sum_{i=1}^n y_i^\alpha/\beta + \text{const.}.$$

Example 3.6 *All the above examples deal with models which lead to a smooth likelihood function. As an example of non smooth likelihood function, let* $Y_1, \ldots, Y_n$ *be an i.i.d. random sample with* $Y_i \sim \text{Unif}(0, \theta)$, $\theta \in \mathbb{R}_{>0}$.

The joint distribution of this random sample is given by the product of their marginal distributions, thus the statistical model is

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

where $1_{(0,\theta)}(y_i)$ *is a function that takes value 1 if* $y_i \in [0, \theta]$ *and 0 otherwise.*

It follows that the likelihood function is

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

Note that in this case we cannot compute the log-likelihood since $L(\theta)$ *may be zero.*

We may wish to compare two different likelihood functions, perhaps based on different models or perhaps using different sets of data. In any case, when plotting the likelihood function it is useful to plot a scaled version of it called *relative likelihood* defined by

$$RL(\theta) = \frac{L(\theta)}{L(\hat{\theta})}.$$

where $\widehat{\theta}$ is the value of θ at which $L(\theta)$ achieves its maximum. Note that $0 < RL(\theta) \leq 1$ by definition since $L(\theta) \leq L(\widehat{\theta})$.

3.2 The observed information

Assume we wish to evaluate the conformity of the products of a production line with respect to quality standards and n products are taken at random. Then the statistical formulation of this problem is to consider $Y_1, \ldots, Y_n$ an i.i.d. random sample with $Y_i \sim \text{Ber}(\theta)$, with θ unknown. Owing to time constraints, it was decided to take only $n = 10$ samples, and the observed sample $(y_1, \ldots, y_n)$ resulted

$$(0, 1, 1, 1, 0, 1, 1, 0, 1, 1),$$

where 1 indicates that the product conforms with the quality standards and 0 indicates that the product does not conform. With this observed sample, the likelihood and the log-likelihood functions are respectively

$$L(\theta) = \theta^7(1-\theta)^3, \quad \text{and} \quad \ell(\theta) = 7\log\theta + 3\log(1-\theta).$$

The relative likelihood function is shown in Figure 3.1 by the thick black curve. From the plot of the relative likelihood we note that there is a θ such that the observed sample has the highest probability of being observed, that is, the likelihood function has a maximum. Let us locate this maximum. Taking the first derivative of the log-likelihood we get

$$\ell'(\theta) = \tfrac{d\ell(\theta)}{d\theta} = \tfrac{7}{\theta} - \tfrac{3}{1-\theta}.$$

Solving $\ell'(\theta) = 0$ in θ we obtain the solution to this equation being equal to $\widehat{\theta} = 7/10$. This solution is a global maximum since the function is strictly concave (the second derivative is negative everywhere).

Now suppose we increase the sample size to 50 and the observed sample is such that there are thirty five 1's and fifteen zeros. The relative likelihood function for this second sample is also shown in Figure 3.1 in red dashed.

From this figure we notice that the dashed (red) curve is narrower than the thick (black) one. For instance, while with a sample size equal to 10, values of $\theta \in (0.2, 1)$ are plausible since the likelihood is much higher than for $\theta \notin (0.2, 1)$, for the red curve, plausible values of θ are found in much narrow intervals, say $(0.5, 0.85)$. Thus, although the most plausible value of θ is identical in both cases, the range of plausible values of θ with $n = 10$ is wider than that with $n = 50$.

To capture this idea it is useful to introduce the concept of observed information, which we define below.

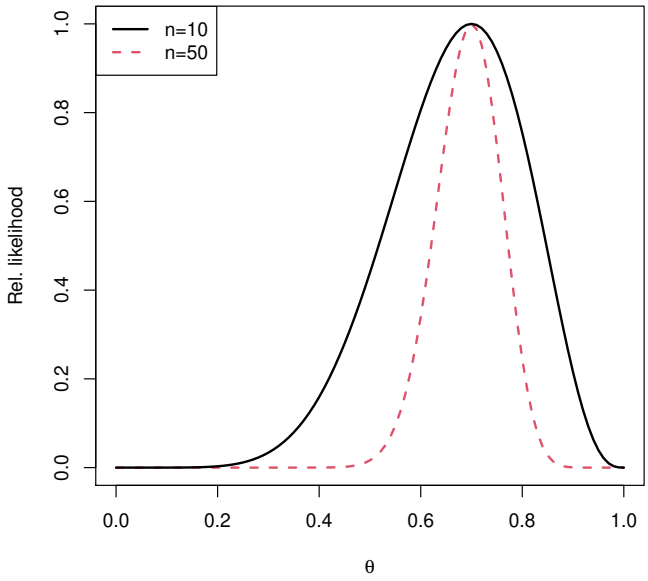

Figure 3.1: Relative likelihood functions for two observed Bernoulli random samples.

Definition 3.2 *Let $\ell(\theta)$ be the log-likelihood then the observed information is a function of θ defined by*

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

If θ is a vector-valued parameter with $\theta \in \mathbb{R}^d$, then the information is a $d \times d$ matrix and is defined by

$$J_n(\theta) = -\frac{d^2\ell(\theta)}{d\theta d\theta^{\mathrm{T}}}.$$

$J_n(\theta)$ is a function of θ and it quantifies the amount of data in a given likelihood function. Typically, the higher is the sample size, the higher is the observed information, i.e. if n_1 and n_2 are sizes of two samples and $1 < n_1 < n_2$ then $J_{n_1} \leq J_{n_2}$. For instance, in the above example with $n = 10$ the information is $\frac{3}{(1-\theta)^2} + \frac{7}{x^2}$, which is lower than $\frac{15}{(1-\theta)^2} + \frac{35}{x^2}$, the observed information with $n = 50$.

3.3 Some computational issues

Assuming that $L(\theta)$ is differentiable, to locate its maximum we proceed by finding $\widehat{\theta}$, the solution in θ to the equation

$$\frac{d\ell(\theta)}{d\theta} = 0.$$

This is the *first-order condition*. Then we check that $J_n(\widehat{\theta}) > 0$ or that $J_n(\widehat{\theta})$ is positive definite; this is

the *second-order condition*. With this two conditions we make sure that $\widehat{\theta}$ is at least a *local* maximum. In the example of Section 3.2 we were able to compute $\widehat{\theta}$ by this procedure analytically. However, the analytic solution of $\frac{d\ell(\theta)}{d\theta} = 0$ is not always feasible.

For instance, consider the observed sample of $n = 10$ waiting times in minutes at a regional telephone exchange for informations about infection by the SARS-COV-2 virus:

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

A distribution that is often used for modelling lifetime, duration or survival data is $\text{Wei}(\alpha, \beta)$. With the Weibull model and assuming that the observed sample is i.i.d., we have the log-likelihood function shown in Figure 3.2 by means of contours (see also Example 3.5). Also shown in the figure is the point in the parameter space with maximum likelihood, i.e. $\widehat{\theta} = (\widehat{\alpha}, \widehat{\beta}) = (2.8, 17.6)$.

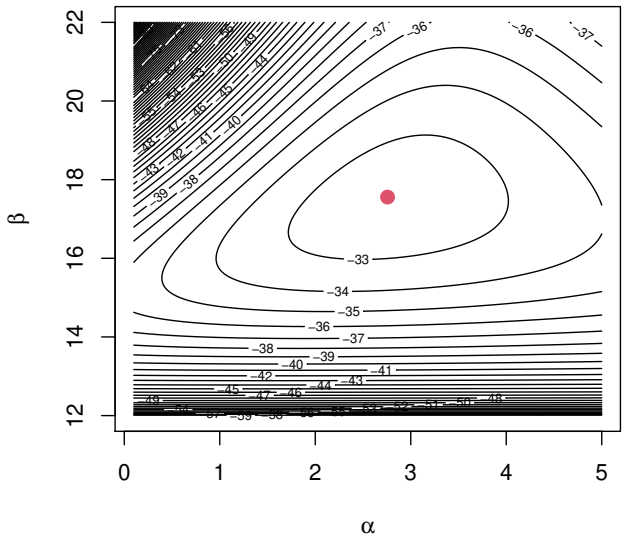

Figure 3.2: Contour plot of the likelihood function under the Weibull model based on an observed sample of size $n = 10$. The red dot indicates the maximum of the likelihood with coordinates $(\widehat{\alpha}, \widehat{\beta}) = (2.8, 17.6)$.

To find the point on the parameter space at which the maximum of the likelihood is achieved, that is $\widehat{\theta} = (\widehat{\alpha}, \widehat{\beta})$, we need to solve the first-order condition. The latter leads to the non nonlinear system:

$$\begin{cases} \frac{\partial \ell(\alpha,\beta)}{\partial \alpha} = \frac{n}{\alpha} + \sum_{i=1}^{n} \log y_i - \frac{\sum_{i=1}^{n} y_i^{\alpha} \log y_i}{\beta} = 0, \\ \frac{\partial \ell(\alpha,\beta)}{\partial \beta} = -\frac{n}{\beta} + \frac{\sum_{i=1}^{n} y_i^{\alpha}}{\beta^2} = 0. \end{cases}$$

From the second equation we get

$$\widehat{\beta}(\alpha) = \tfrac{1}{n} \sum_{i=1}^{n} y_i^{\alpha}.$$

Replacing β by $\widehat{\beta}(\alpha)$ in the first equation we obtain

$$g(\alpha) = \sum_{i=1}^{n} \log y_i + \tfrac{n}{\alpha} - \tfrac{n \sum_{i=1}^{n} y_i^{\alpha} \log y_i}{\sum_{i=1}^{n} y_i^{\alpha}} = 0.$$

The equation $g(\alpha) = 0$ cannot be solved explicitly, but it is possible to solve it by an iterative numerical method such as Newton-Raphson method. Specifically, we can define a sequence $\widehat{\alpha}_1, \widehat{\alpha}_2, \ldots$ such that

$$\widehat{\alpha}_m = \widehat{\alpha}_{m-1} - \tfrac{g(\widehat{\alpha}_{m-1})}{g'(\widehat{\alpha}_{m-1})},$$

where $\alpha_0 > 0$ is an initial value, $g'(\alpha)$ is the derivative of $g(\alpha)$ and $\widehat{\alpha}_m \to \widehat{\alpha}$ as $m \to \infty$. Once we have $\widehat{\alpha}$, then we set $\widehat{\beta} = \frac{1}{n} \sum_{i=1}^{n} y_i^{\widehat{\alpha}}$ and so $\widehat{\theta} = (\widehat{\alpha}, \widehat{\beta})$.

This approach is used in order to compute the coordinates of the red point in Figure 3.2. Note that, to assure that $\widehat{\theta}$ is a local maximum we also need to check the second-order condition, i.e. the matrix $J_n(\widehat{\theta})$ must be positive definite.

References

[HMC20] Hogg, R. V., McKeen, J. W. and Craig, A. T. (2018) *Introduction to Mathematical Statistics* (8th edition, global ed.), Pearson Education, Chapp. 4.