Statistical models for non iid data

In many applications the iid assumption, especially the "identically" part, is unrealistic. Here is a common situation.

P1 produces washing machine (WM) motors. He claims that his new model (NM) is more efficient, while achieving the same speed as the old, version motors (OM). Which one should we buy? To answer this question, we have to run experiments with WM+NM and WM+OM and then analyse the data. Which statistical model should we use for this problem?

Two-sample normal model

Let the $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be the energy consumptions measured under the OM and NM respectively.

It's reasonable to assume that

- the measures within each motor independent (discussion?)
- the measures between motors are also independent
- the data-generating process under OM may differ from that of NM
- energy consumption is reasonably Gaussian (sum of many small energetically hungry components), thus

Joint distribution for OM: $f(x_1, \ldots, x_m, \theta_x) = \prod_{i=1}^{m} f(x_i; \theta_x)$
Joint distribution for NM: $f(y_1, \ldots, y_n) = \prod_{j=1}^{n} f(y_j; \theta_y)$.
Joint distribution for OM and NM:

$$f(x_1, \ldots, x_m, y_1, \ldots, y_n; \theta_x, \theta_y) = \left(\prod_i f(x_i; \theta_x) \right) \left(\prod_j f(y_j; \theta_y) \right),$$

where $\theta_x = (\mu_x, \sigma_x^2), \theta_y = (\mu_y, \sigma_y^2)$. The statistical model is the set of all joint distributions generated by the different parameter values

Linear regression

Suppose we have an arsenic remover device (which has negative health effects on human beings) from drinkable water and we suspect the effectiveness of the removal depends on the pH of water. Se we need to assess how arsenic removal is affected by water pH.

Let $x_1, \ldots, x_n$ be pH values of n samples of water and let $y_1, \ldots, y_n$ be the measured values of arsenic removed from each of the water samples.

Typically we assume $x_1, \ldots, x_n$ are fixed and the y_i are a realisation of the random sample $Y_1, \ldots, Y_n$.

Then, a possible model is:

given x_i → "simple" situation, fixed x_i

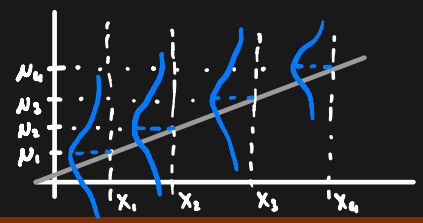
$$Y_i | x_i \sim N(\mu_i, \sigma^2)$$

$$\mu_i = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots$$

Y_i independent from Y_j for all i, j.

$$\left\{ \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2} : \beta_0, \beta_1 \in \mathbb{R},\ x_i, y_i \in \mathbb{R}\ \forall i \right\}$$

The unknown parameter is $(\beta_0, \beta_1, \sigma^2)$. → ϑ



Logistic regression

Suppose you want to predict chicken sex from its egg features.

For the ith egg, let $x_{i1}, \ldots, x_{ip}$ be p features (e.g. volume, color, etc.) and let $y_1, \ldots, y_n$ be the chicken sex (1=female, 0=male).
A simple model for this problem can be built as follows:

$$Y_i | x_{i1}, \ldots, x_{ip} \sim \text{Ber}(\theta_i)$$

$$\theta_i = \frac{1}{1+e^{-\mu_i}}$$ ← sigmoid function

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$ → not bounded [9:1]

$$Y_i, Y_j \quad \text{independent for all } i, j.$$

The joint distribution is

$$f(y_1, \ldots, y_n | \mathbf{X}, \theta) = \prod_{i=1}^{n} \left(\frac{1}{1+e^{-(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip})}} \right)^{y_i} \left(\frac{1}{1+e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}} \right)^{1-y_i}.$$

The statistical model is the set of joint distributions at all $\theta = (\beta_0, \ldots, \beta_p) \in \mathbb{R}^{p+1}$.

Inferential Statistics
L3 - The likelihood function

Erlis Ruli (erlis.ruli@unipd.it)

Department of Statistics, University of Padova

Contents

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 $L(\theta) : \Theta \to R_{\geq 0}$ 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.

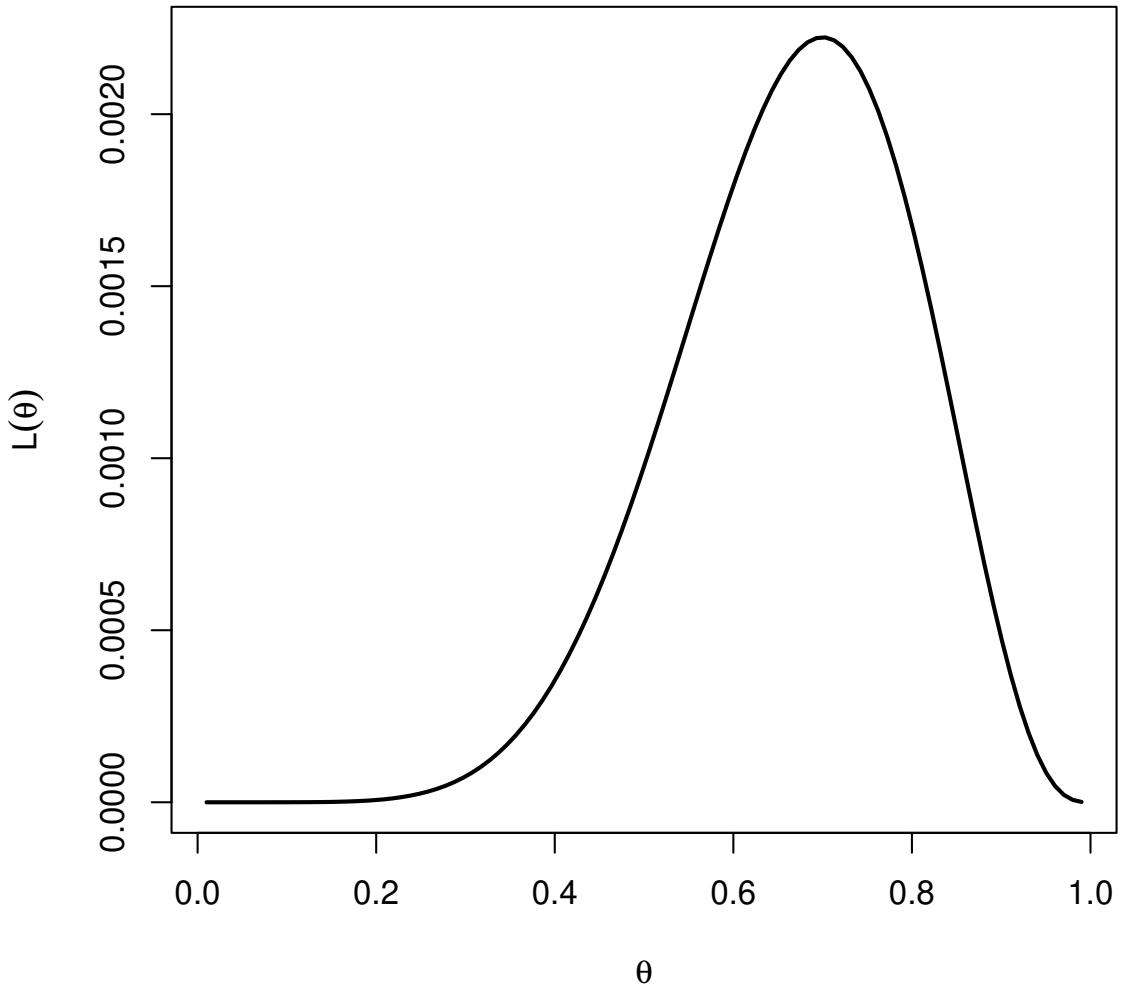
The likelihood function is

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

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

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

Example 1 (cont'd)



On the interpretation of L

For a sample $x_1, \ldots, x_n$ 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

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

in a continuous space it's more correct than saying argmax