Likelihood Function

Back ground

Let's first set some notation and terminology. Observable data $X_1, \ldots, X_n$ has a specified model, say, a collection of distribution functions $\{F_\theta : \theta \in \Theta\}$ indexed by the parameter space Θ. Data is observed, but we don't know which of the models F_θ it came from. we shall assume that the model is correct, i.e., that there is a θ value such that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F_\theta$. The goal, then, is to identify the "best" model—the one that explain the data the best. This amounts to identifying the true but unknown θ value. Hence, our goal is to estimate the unknown θ.

Concept of Likelihood

Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F_\theta$, where θ is unknown. For the time being, we assume that θ resides in a subset Θ of $\mathbb{R}$. We further suppose that, for each θ, $F_\theta(x)$ admits a PMF/PDF $f_\theta(x)$. By the assumed independence, the joint distribution of $(X_1, \ldots, X_n)$ is characterized by

$$f_\theta(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_\theta(x_i),$$

i.e., "independence means multiply." we understand the above expression to be a function of $(x_1, \ldots, x_n)$ for fixed θ. **Now** we flip this around. That is, we will fix $(x_1, \ldots, x_n)$ at the observed $(X_1, \ldots, X_n)$, and imagine the above expression as a function of θ only.

Definition 1. If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_\theta$, then the *likelihood function* is

$$L(\theta) = f_\theta(X_1, \ldots, X_n) = \prod_{i=1}^{n} f_\theta(X_i), \tag{1}$$

treated as a function of θ. In what follows, I may occasionally add subscripts, i.e., $L_X(\theta)$ or $L_n(\theta)$, to indicate the dependence of the likelihood on data $X = (X_1, \ldots, X_n)$ or on sample size n. Also write

$$\ell(\theta) = \log L(\theta), \tag{2}$$

for the log-likelihood; the same subscript rules apply to $\ell(\theta)$.

So clearly $L(\theta)$ and $\ell(\theta)$ depend on data $X = (X_1, \ldots, X_n)$, but they're treated as functions of θ only. How can we interpret this function? The first thing to mention is a warning—*the likelihood function is NOT a PMF/PDF for θ!* So it doesn't make sense to integrate over θ values like you would a PDF We're mostly interested in the shape of the likelihood curve or, equivalently, the relative comparisons of the $L(\theta)$ for different θ's.

If $L(\theta_1) > L(\theta_2)$ (equivalently, if $\ell(\theta_1) > \ell(\theta_2)$), then θ_1 is more likely to have been responsible for producing the observed $X_1, \ldots, X_n$. In other words, F_{θ_1} is a better model than F_{θ_2} in terms of how well it fits the observed data.

So, we can understand likelihood (and log-likelihood) of providing a sort of *ranking* of the θ values in terms of how well they match with the observations.

A sensible way to estimate the parameter $\boldsymbol{\theta}$ given the data $\mathbf{y}$ is to maximize the likelihood (or equivalently the log-likelihood) function, choosing the parameter value that makes the data actually observed as likely as possible. Formally, we define the *maximum-likelihood estimator* (mle) as the value $\hat{\boldsymbol{\theta}}$ such that

$$\log \mathrm{L}(\hat{\boldsymbol{\theta}}; \mathbf{y}) \geq \log \mathrm{L}(\boldsymbol{\theta}; \mathbf{y}) \text{ for all } \boldsymbol{\theta}.$$

The **likelihood function** *is the density function regarded as a function of θ.*

$$L(\theta|x) = f(x|\theta),\ \theta \in \Theta.$$

- Where, $f(x|\theta)$ is the pdf corresponding to the random vector **X**.

Example

Example: The Log-Likelihood for the Geometric Distribution. Consider a series of independent Bernoulli trials with common probability of success π. The distribution of the number of *failures* Y_i before the first success has pdf

$$\Pr(Y_i = y_i) = (1-\pi)^{y_i}\pi. \tag{A.4}$$

for $y_i = 0, 1, \ldots$. Direct calculation shows that $\mathrm{E}(Y_i) = (1-\pi)/\pi$.

The log-likelihood function based on n observations $\mathbf{y}$ can be written as

$$\begin{aligned} \log \mathrm{L}(\pi; \mathbf{y}) &= \sum_{i=1}^{n} \{y_i \log(1-\pi) + \log \pi\} && \text{(A.5)} \\ &= n(\bar{y} \log(1-\pi) + \log \pi), && \text{(A.6)} \end{aligned}$$

where $\bar{y} = \sum y_i/n$ is the sample mean. The fact that the log-likelihood depends on the observations only through the sample mean shows that $\bar{y}$ is a *sufficient* statistic for the unknown probability π.

Visualization

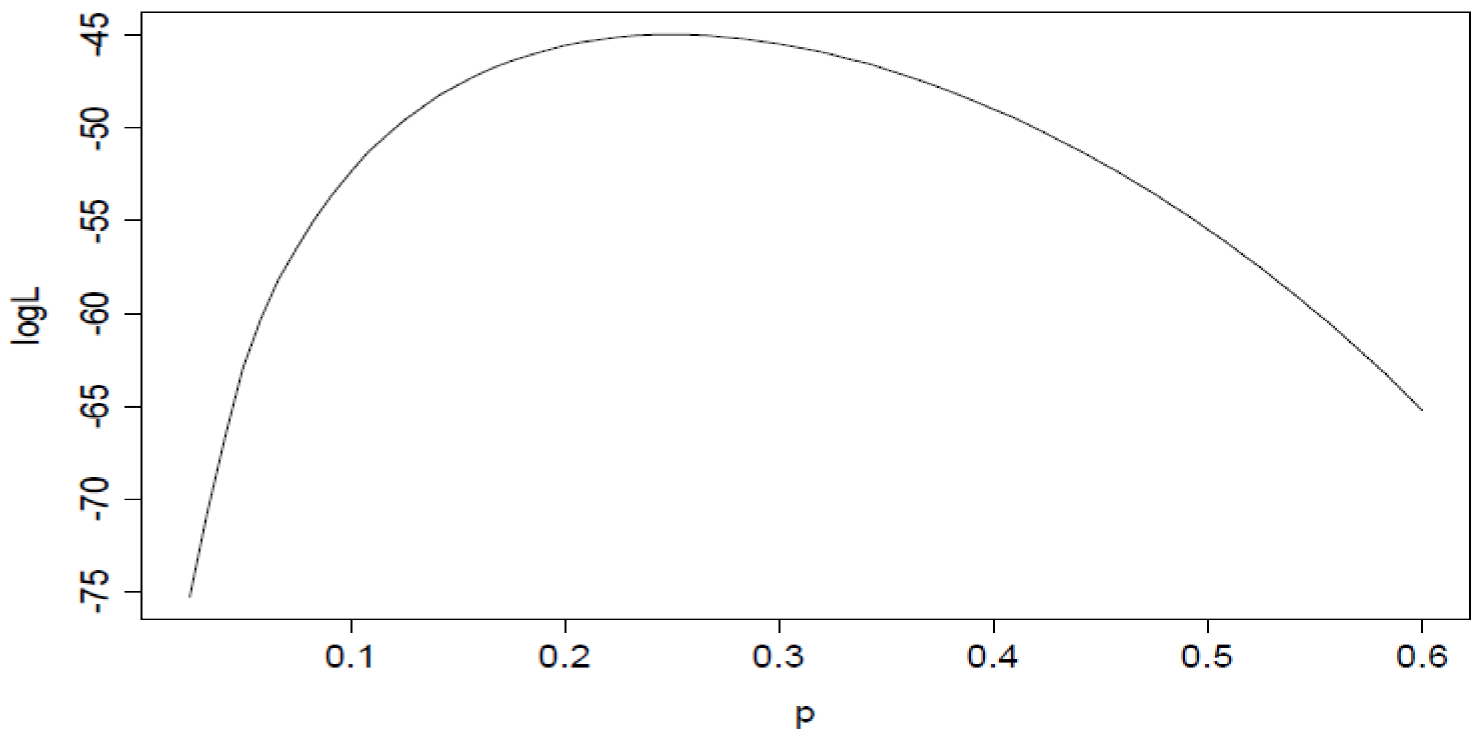

FIGURE A.1: The Geometric Log-Likelihood for $n = 20$ and $\bar{y} = 3$

Figure A.1 shows the log-likelihood function for a sample of $n = 20$ observations from a geometric distribution when the observed sample mean is $\bar{y} = 3$.$\square$

Properties

- The likelihood function is <u>not</u> a probability density function.
- It is an important component of both frequentist and Bayesian analyses
- It measures the support provided by the data for each possible value of the parameter. If we compare the likelihood function at two parameter points and find that $L(\theta_1|x) > L(\theta_2|x)$ then the sample we actually observed is more likely to have occurred if $\theta = \theta_1$ than if $\theta = \theta_2$. This can be interpreted as θ_1 is a more plausible value for θ than θ_2.

Likelihood Principle

If x and y are two sample points such that $L(\theta|x) \propto L(\theta|y)\ \forall\ \theta$ then the conclusions drawn from x and y should be identical.

Thus the likelihood principle implies that likelihood function can be used to compare the plausibility of various parameter values. For example, if $L(\theta_2|x) = 2L(\theta_1|x)$ and $L(\theta|x) \propto L(\theta|y)\ \forall\ \theta$, then $L(\theta_2|y) = 2L(\theta_1|y)$. Therefore, whether we observed x or y we would come to the conclusion that θ_2 is twice as plausible as θ_1.