Review: Likelihood

1 Likelihood

Consider a set of *iid* random variables $\mathbf{X} = \{X_i\},\ i = 1, ..., N$ that come from some probability model $\mathcal{M} = \{f(x;\theta); \theta \in \Theta\}$ where $f(x;\theta)$ is the probability density of each X_i and θ is an unknown but *fixed* parameter from some space $\Theta \subseteq \Re^d$. (Note: this would be a typical frequentist setting.) The **likelihood function** (Fisher, 1925, 1934) is defined as:

$$\mathcal{L}(\theta \mid \vec{X}) = \prod_{i=1}^{N} f(X_i;\theta), \tag{1}$$

or any other function of θ proportional to (1) (it does not need be normalized). More generally, the likelihood function is any function proportional to the joint density of the sample, $\vec{X} = \{X_i\}_{i=1}^N$, denoted as $f(\vec{X};\theta)$. Note you don't need to assume independence, nor identically distributed variables in the sample.

The likelihood function is sacred to most statisticians. Bayesians use it jointly with the prior distribution $\pi(\theta)$ to derive the posterior distribution $\pi(\theta \mid \mathbf{X})$ and draw all their inference based on that posterior. Frequentists may derive their estimates directly from the likelihood function using maximum likelihood estimation (MLE).

We say that a parameter is identifiable if for any two $\theta \neq \theta^*$ we have $\mathcal{L}(\theta \mid \vec{X}) \neq \mathcal{L}(\theta^* \mid \vec{X})$

2 Maximum likelihood estimators

One way of identifying (estimating) parameters is via the maximum likelihood estimators. Maximizing the likelihood function (or equivalently, the log likelihood function thanks to the monotonicity of the log transform), with respect to the parameter vector θ, yields the maximum likelihood estimator, denoted as $\widehat{\theta}$. The idea is to estimate the parameter by its most likely value given the data.

The maximization process relies on finding the critical points of the log-likelihood function, and checking that they are indeed the maxima:

1. differentiate the log-likelihood with respect to the parameter θ:
2. set the resulting expression to 0 and solve for θ
3. the solution (if it exists and it is unique) is the maximum likelihood estimator, $\widehat{\theta}$.
4. find the second derivative of the log-likelihood with respect to θ and evaluate it at $\widehat{\theta}$. If the result is negative, then $\widehat{\theta}$ is indeed the maximum.

Note that there might be more than one critical point of the likelihood function, and thus more than one solution in the step#2 above. This does not necessarily mean that the MLE is not unique! First check (if possible) which of the solutions actually yields the maximum value of $\mathcal{L}(\theta \mid \vec{X})$, by evaluating the likelihood function at each of the solutions found in step #2. The result will then be the MLE. If more than one of the solutions yield the same maximum value of the likelihood function, then we say the MLE is not unique.

3 Properties of the maximum likelihood estimators

MLEs have many nice properties. Let θ_0 denote the true unknown value of the parameter θ. Under certain regularity conditions (eg., that the first three derivatives of the log likelihood function are continuous and finite for all θ), we have the following:

1. MLEs are consistent: $\widehat{\theta} \to \theta_0$ in probability. In other words, the probability of an MLE deviating from the truth by any given fixed amount goes to 0 as the sample size increases.

2. MLEs are asymptotically normal: Regardless of what the probability density $f(\vec{X} \mid \theta)$ of the sample is, we have that as the sample size increases, the distribution of the MLE approaches the normal density:

$$\widehat{\theta} \sim \mathcal{N}(\theta_0, I(\theta_0)^{-1})$$

where $I(\theta_0)$ is the Fisher information matrix, defined as the negative expectation of the second derivative of the log likelihood. The expectation we refer to represents the average over all repeated samples of the same size, generated by the same model (with the same true parameter θ_0) as the sample at hand.

3. MLEs are asymptotically efficient: that means they have the lowest variance of all other "nice" estimators (all estimators that are also consistent and asymptotically normal).

4. MLEs are invariant under transformations. This is an amazing property often called the "invariance principle", and it states that an MLE of $g(\theta)$ is simply g of the MLE of θ:

$$\widehat{g(\theta)} = g(\widehat{\theta}).$$

To see why the last property holds (and why the simplifying assumption of g being one-to-one is not needed), define the likelihood induced by the g transform, $\mathcal{L}^g$, as:

$$\mathcal{L}^g(\lambda \mid \vec{X}) = \sup_{\theta : g(\theta) = \lambda} \mathcal{L}(\theta \mid \vec{X})$$

Then, let $\widehat{\lambda}$ be the MLE of that likelihood $\mathcal{L}^g(\lambda \mid \vec{X})$. We have:

$$\mathcal{L}^g(\widehat{\lambda} \mid \vec{X}) = \max_{\lambda} \mathcal{L}^g(\lambda \mid \vec{X}) = \max_{\lambda} \sup_{\theta : g(\theta) = \lambda} \mathcal{L}(\theta \mid \vec{X}) = \sup_{\theta} \mathcal{L}(\theta \mid \vec{X}) = L(\widehat{\theta} \mid \vec{X})$$

This allows us to easily find the asymptotic distribution of $\widehat{g(\theta)}$ as well (using also the first order Taylor expansion of g):

$$\widehat{g(\theta)} \sim \mathcal{N}(g(\theta_0), (g'(\theta_0))^T I(\theta_0)^{-1} (g'(\theta_0)))$$

where $g'(\theta_0)$ is the derivative of g with respect to θ, evaluated at θ_0.

The proof of the other properties and more can be found in any good mathematical statistics and econometrics book. I would recommend Casella and Berger's "Statistical Inference" (Chapter 7 in the 2nd edition) and William Greene's "Econometric Theory" (Chapter 14 in the 7th edition). Note that Greene does not present the general proof of the invariance principle.

4 The likelihood principle

The likelihood principle states that any two experiments that yield proportional likelihoods should yield identical inferences about θ. This principle is viewed as the absolute truth by some people and a dogma by others. It is important to realize that many classical (frequentist) inferential procedures, such as significance testing, violate this principle. Consider the following example:

An experiment is conducted in which we toss a coin for some time and at the end of the experiment the "score" is 3 heads and 9 tails. Suppose that we want to test the probability that the coin is fair (i.e. $H_0 : \theta = 0.5$ *vs.* $H_1 : \theta_1 > 0.5$ where θ is the probability of getting a tail). What is the likelihood for this experiment?

There are two standard scenarios that could have resulted in the same score: binomial and negative-binomial sampling scenario. Note that we were not told whether we explicitly wanted 12 coin flips and only recorded the number of heads in the end of the experiment, or whether we wanted 3 heads total and would have continued flipping the coin until we got those 3 heads, no matter how long that would take. But what if there was something else entirely, like we dropped the coin after the 12th toss and couldn't find it again?

The first standard scenario corresponds to the binomial sampling, and the likelihood function is:

$$\mathcal{L}_1(\theta) = \binom{12}{9} \theta^9 (1-\theta)^3 \tag{2}$$

while in the second standard scenario the negative-binomial likelihood function is:

$$\mathcal{L}_2(\theta) = \binom{11}{9}\theta^9(1-\theta)^3 \tag{3}$$

The third scenario – as long as the stopping rule is independent of the outcomes observed thus far in the experiment – should coincide with the Binomial likelihood.

These two likelihoods are the same, because they are both proportional to the single function of θ which is $\theta^9(1-\theta)^3$. The MLEs of the probabilities of "tails" in a single toss will be identical.

But the likelihood principle goes beyond the estimation and MLEs. According to the likelihood principle all **inference (and not just the point estimate)** about θ should also be the same, regardless which sampling framework was assumed.

Alas, a simple calculation of p-values for each of the experiments reveals a discrepancy:

$$p_1 = P(\text{at least 9 tails} \mid \theta = 0.5, \text{binomial with 12 flips}) = \sum_{i=9}^{12}\binom{12}{i}0.5^i 0.5^{12-i} = 0.075 \tag{4}$$

$$p_2 = P((\text{at least 9 tails } \mid \theta = 0.5, \text{negative-binomial with 3 heads}) = \sum_{i=9}^{\infty}\binom{(i+2)}{i}0.5^{i+3} = 0.0325 \tag{5}$$

which gives the opposite conclusions at 5% significance level.

One of the problems associated with the above experiment is that this frequentist procedure uses the unobserved data for inference. In other words, the probabilities of the data that have not occurred in the experiment (all the repeated samples we never actually collected) were used in the calculation of p-values. Is that ok? Should the "how confident we are about θ" depend on the evidence and data we have, or on the hypothetical data which may or may not ever be collected?

Bayesian inference only looks at the posterior given the observed data, which is a function of only θ, and finds the actual probability of each hypothesis being true. So in the Bayesian hypothesis testing the likelihood principle is satisfied – that is, at least as long as the prior does not depend on the experiment. We'll see that Jeffreys' priors for example are not coherent with the likelihood principle because they depend on the experiment just as much as frequentist procedures.

References/Further Reading:

1. Birnbaum (1962): "On the Foundations of Statistical Inference", JASA 57, 269–306.

2. Berger and Wolpert (1984): "The Likelihood Principle", Hayward, CA (IMS)

3. Greene (2011): "Econometric Analysis" – see "Maximum Likelihood Estimation" (Chapter 14 in the 7th edition)

4. Casella and Berger (2001): "Statistical Inference" (Chapter 7 in the 2nd edition)

4.1 Examples

1. Problem:
   Let $X_1, X_2, ..., X_5$ be a random sample from $N(\mu, 1)$. What are the likelihood and log-likelihood functions of μ?

   Solution:
   Due to the fact that X_i form a random sample, they are *iid*, and so:

$$L(\mu) = \prod_1^5 f(X_i; \mu) = \prod_1^5 \frac{1}{\sqrt{2\pi}} e^{-(\mu - x_i)^2/2} \propto \prod_1^5 e^{-(\mu - x_i)^2/2}$$

$$l(\mu) \propto \sum_1^5 -(\mu - x_i)^2/2 = \sum_1^5 -(\mu^2 - 2x_i\mu + x_i^2)/2 = -(5\mu^2 - 2\mu \sum_1^5 x_i + \sum_1^5 x_i^2)/2$$

   which is proportional to

$$-\frac{\mu^2 - 2\mu \sum_1^5 x_i/5 + \sum_1^5 x_i^2/5}{2/5}$$

   This is proportional to another Normal distribution, $\mu|\vec{x} \sim N(\sum_1^5 x_i/5, 1/5)$, i.e., the likelihood is proportional to a Normal distribution with mean equal to the sample mean, $\bar{x}$, and variance equal to σ^2/n.

2. Problem:
   Suppose that you have 10 components of type A and 5 components of type B in a particular lab machine. Assume that the lifetimes of all components are independent and exponentially distributed. Furthermore, assume that each A component has a mean lifetime of ω hours, and that each component B has a mean lifetime of 2ω hours. You turn the machine on, and leave it operating alone for 24 hours. Upon return you notice that 1 of the 10 components A have failed, and 1 of the 5 B components have failed.

   (a) What is the exact probability that 1 out of 10 A components and 1 out of 5 B component fail within 24 hours?

   (b) What are the likelihood and log likelihood functions of ω?

   (c) What value of ω maximizes the likelihood function? (This may be time consuming)

Hint:

1. The exact probability is given by the product of binomial pmfs:

$$P(\text{1 out of 10 components A failing}) = \binom{10}{1} p_A^1 (1 - p_A)^{10-1}$$

$$P(\text{1 out of 5 components B failing}) = \binom{5}{1} p_B^1 (1 - p_B)^{5-1}$$

   where p_A and p_B are probabilities of a component A and component B failing within 24 hours, respectively. They are obtained via cumulative probability distribution of an exponentially distributed lifetime:

$$p_A = \int_0^{24} \frac{1}{\omega} exp(\frac{-t}{\omega}) dt = 1 - exp(-24/\omega)$$

$$p_B = \int_0^{24} \frac{1}{2\omega} exp(\frac{-t}{2\omega}) dt = 1 - exp(-12/\omega)$$

   (b) The likelihood function of ω is thus:

$$p_A(1 - p_A)^9 p_B(1 - p_B)^4 = (1 - exp(-24/\omega)) exp(-24/\omega)^9 (1 - exp(-12/\omega)) exp(-12/\omega)^4$$

   We can optimize this likelihood by finding the critical points of its logarithm (exercise!)