Maximum Likelihood Estimation

©Dennis L. Bricker
Dept of Mechanical & Industrial Engineering
University of Iowa
dennis-bricker@uiowa.edu

Suppose that we have observed values t_1, t_2, ... t_n of a random variable **T**.

Suppose also that the distribution of T is known to belong to a certain type (e.g., exponential, normal, etc.) but the vector $\theta = \left(\theta_1, \theta_2, \ldots \theta_p\right)$ of unknown parameters associated with it is *unknown* (where *p* is the number of unknown parameters).

Let the *density function* be written as $f(t;\theta)$.

For example, if **T** has Normal distribution,

$$f(t;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right\}$$

(where $\theta_1=\mu$ & $\theta_2=\sigma$ *have yet to be determined.)*

We want to estimate the unknown parameters by choosing those values of θ which make the *likelihood* of the observed values *as large as possible*.

Other alternative methods:

- *method of moments*: choose θ so that the moments of $f(t;\theta)$ are equal to those of the sample (e.g., match the sample mean and sample variance).

- use *regression analysis*, i.e., curve-fitting, to choose θ so as to minimize the sum of the squared errors in the nonlinear system of equations:

$$\begin{cases} 1/n = F(t_1;\theta) \\ 2/n = F(t_2;\theta) \\ \vdots \\ n/n = F(t_n;\theta) \end{cases}$$

where F is the CDF of the dist'n

(This is generally an unconstrained nonlinear minimization problem which must be solved by an iterative algorithm, although often transformations can be applied to obtain a linear system which can then be solved easily.)

Maximum Likelihood Estimation (MLE)

Consider first the case in which T is *discrete*.

A simple example, with a not-at-all surprising result:

Suppose that a Bernouilli random variable is sampled,

i.e., $t_i \in \{0,1\}$ for each i=1,2,...n.

The number of "successes" is known to have a *binomial* distribution with parameter p = probability of "success".

Suppose that the number of successes in the sample,

i.e., $\sum_{i=1}^{n} t_i$, be ***k***.

What then should be our estimate of p ?

The probability, or *likelihood*, of ***k*** successes in *n* trials, if *T* is a Bernouilli random variable, is

$$L(p) = \binom{n}{k} p^k (1-p)^{n-k}$$

(which has been written as a function of the unknown parameter p.)

The *maximum likelihood estimate* of *p* is the value which maximizes the function *L(p)*.

solution: consider the stationary points of L:

$$\frac{dL}{dp} = \binom{n}{k}\left[kp^{k-1}(1-p)^{n-k} + p^k(n-k)(1-p)^{n-k-1}(-1)\right] = 0$$

$$\frac{dL}{dp}=\binom{n}{k}p^{k-1}(1-p)^{n-k-1}\left[k(1-p)-p(n-k)\right]=0$$

One of the factors must be zero in the solution, so the three solutions are:

$$p=0$$

$$(1-p)=0 \quad \Rightarrow p=1$$

or

$$k(1-p)-p(n-k)=0 \quad \Rightarrow k-kp-np+kp=k-np=0 \Rightarrow p={}^{k}\!/_{n}$$

Obviously the first two solutions, i.e. $p=0$ & 1, do *not* maximize the function L, while the third solution is what we would have expected to be the MLE!

Consider now the case in which T does *not* have a discrete distribution, and $f(t;\theta)$ is its density function.

Since the observed values are independent, the **likelihood function** $L(t,\theta)$ is the **product** of the probability density function evaluated at each observed value:

$$L(t,\theta) = \prod_{i=1}^{n} f(t_i;\theta)$$

The **maximum likelihood estimator** $\hat{\theta}$ is found by maximizing $L(t,\theta)$ with respect to θ. Thus $\hat{\theta}$ corresponds to the distribution that is most likely to have yielded the observed data $t_1, t_2, \ldots t_n$.

The problem

$$\underset{\theta}{Maximize} \;\; L(t_1,\ldots t_n;\theta)$$

is a *nonlinear optimization problem* which might be solved by any appropriate NLP algorithm (Newton or quasi-Newton methods, the conjugate gradient method, etc.)

For computational convenience, it's usually preferable to maximize the *logarithm* of the maximum likelihood (which will yield the same maximizing $\hat{\theta}$):

$$\underset{\theta}{Maximize}\ \ \ln L\left(t_1,\ldots t_n;\theta\right)$$

i.e., because $\ln L\left(t;\theta\right) = \ln \prod_{i=1}^{n} f\left(t_i;\theta\right) = \sum_{i=1}^{n} \ln f\left(t_i;\theta\right)$

we solve the problem:

$$\underset{\theta}{Maximize}\ \sum_{\mathrm{i}=1}^{\mathrm{n}} \ln f\left(t_i;\theta\right)$$

Example: Exponential Distribution

(another not-so-surprising result)

The probability density function *(pdf)* of the exponential distribution with parameter λ is

$$f(t;\lambda) = \lambda e^{-\lambda t}$$

We have a set of *n* observations t_1, t_2, … t_n. What is the value of the parameter λ which makes this set of observations most likely?

Sample data: Times to failure of six electronic components are (in hours):

25, 75, 150, 230, 430, and 700.

Solution: The likelihood function is

$$L(t_1,\ldots t_n;\lambda)=\prod_{i=1}^{n}\lambda e^{-\lambda t_i}=\lambda^n \exp\left\{-\lambda\sum_{i=1}^{n}t_i\right\}$$

The *logarithm* of the likelihood is

$$\ln L(t;\lambda)=n\log\lambda-\lambda\sum_{i=1}^{n}t_i$$

which has *derivative*

$$\frac{d}{d\lambda}L(t;\lambda)=\frac{n}{\lambda}-\sum_{i=1}^{n}t_i$$

In the case, then, we can solve the nonlinear optimization problem (with one variable) by finding a stationary point, i.e., a value of λ for which the above derivative is zero.

$$\frac{d}{d\lambda}L(t;\lambda)=\frac{n}{\hat{\lambda}}-\sum_{i=1}^{n}t_i=0$$

$$\Rightarrow\frac{1}{\hat{\lambda}}=\frac{1}{n}\sum_{i=1}^{n}t_i$$

$$\Rightarrow\hat{\lambda}=\frac{n}{\sum_{i=1}^{n}t_i}$$

That is, in the case of the exponential distribution, the MLE is *(surprise!)* simply

the reciprocal of the average of the observed values.

That is, for the sample data,

$$\hat{\lambda}=\frac{6\,failures}{(25+75+150+230+430+700)\,hrs}=\frac{6\,failures}{1610hrs}=0.0037267\,failures/hr.$$

In the case of the normal distribution (with *two* parameters, μ & σ), the optimality conditions for maximum of the *log likelihood* is a pair of nonlinear equations, but *again* they can be solved in closed form, and the results are as one might expect:

- the MLE for μ is the average of the observations, and
- the MLE for σ is the square root of the sample variance.

In general, however, one cannot find a closed-form solution for the maximim likelihood estimator(s), requiring an ***iterative*** *algorithm. (For example, MLE for Weibull & Gumbel distributions.)*

Maximum Likelihood Estimation
with "censored" data

Suppose that an experiment was terminated at time τ after only r of the n units in a lifetest had failed. This is accounted for by defining the likelihood as

$$L(t,\theta) = \left[1 - F(\tau;\theta)\right]^{n-r} \times \prod_{i=1}^{r} f(t_i;\theta)$$

since

$\left[1 - F(\tau;\theta)\right]^{n-r}$ is the probability that the $n-r$ units survive until time τ.

Since $$L(t,\theta)=\left[1-F(\tau;\theta)\right]^{n-r}\times\prod_{i=1}^{r}f(t_i;\theta)$$

the log-likelihood function is therefore

$$\ln L(t;\theta)=(n-r)\ln\left[1-F(t;\theta)\right]+\sum_{i=1}^{r}\ln f(t_i;\theta)$$

Generally, this is maximized either

- by solving the optimality conditions

$$\frac{\partial}{\partial\theta_i}\ln L(t;\theta)=0 \quad \text{for} \quad i=1,2,\ldots p$$

- by an iterative optimization algorithm (e.g. Quasi-Newton)