Methods in
Molecular Biology 930

Springer Protocols

Brad Reisfeld
Arthur N. Mayeno *Editors*

# Computational Toxicology

## Volume II

Humana Press

# Chapter 24

# Maximum Likelihood

Shuying Yang and Daniela De Angelis

## Abstract

The maximum likelihood method is a popular statistical inferential procedure widely used in many areas to obtain the estimates of the unknown parameters of a population of interest. This chapter gives a brief description of the important concepts underlying the maximum likelihood method, the definition of the key components, the basic theory of the method, and the properties of the resulting estimates. Confidence interval and likelihood ratio test are also introduced. Finally, a few examples of applications are given to illustrate how to derive maximum likelihood estimates in practice. A list of references to relevant papers and software for a further understanding of the method and its implementation is provided.

**Key words:** Likelihood, Maximum likelihood estimation, Censored data, Confidence interval, Likelihood ratio test, Logistic regression, Linear regression, Dose response

## 1. Introduction

The maximum likelihood method is, like the least squares method, a statistical inferential technique to obtain estimates of the unknown parameters of a population using the information from an observed sample. It was primarily introduced by RA Fisher between 1912 and 1920, though the idea has been traced back to the late nineteenth century (1, 2).

The principle of the maximum likelihood method is to find the value of the population parameter, the maximum likelihood estimate (MLE), that maximize the probability of observing the given data. The maximum likelihood method, by motivation, is different from the least squares method, but the MLEs coincide with the least squares estimates (LSEs) under certain assumptions, e.g., that residual errors follow a normal distribution.

While the maximum likelihood theory has its basis the point estimate of unknown parameters in a population described by a

Brad Reisfeld and Arthur N. Mayeno (eds.), *Computational Toxicology: Volume II*, Methods in Molecular Biology, vol. 930, DOI 10.1007/978-1-62703-059-5_24, © Springer Science+Business Media, LLC 2013

certain distribution (e.g., examples in Subheading 2 and example 1 in Subheading 3), its application extends far beyond the simple distributional forms to situations where the distribution of the random quantities or variables of interest ($y$) are determined by some other variables ($x$). This is the case in linear or nonlinear regression models and compartmental pharmacokinetics models such as those described in the previous chapters. In such situations, mathematical models are utilized to describe the relationship between $y$ and $x$ given some unknown parameters, which are referred to as model parameters. The most frequent use of the maximum likelihood method is to obtain the point estimates of these model parameters.

The maximum likelihood method has been widely applied for statistical estimation in various models as well as for model selection (3–9).

The likelihood and log-likelihood functions are the foundation of the maximum likelihood method. Definitions of the likelihood and log-likelihood are given in the next sections. The idea of likelihood is also at the basis of the Bayesian inferential approach, which will be explained in the next chapter in more detail.

The aim of this chapter is to introduce the concept of the maximum likelihood method, to explain how maximum likelihood estimates are obtained and to provide some examples of application of the maximum likelihood method in the estimation of population and model parameters. The practical examples are provided with details so that readers will have thorough understanding of the maximum likelihood method and be able to apply it at the same time.

## 2. Important Concepts

### *2.1. Likelihood and Log-Likelihood Function*

Suppose we have a sample $y = (y_1, \ldots y_n)$ where each $y_i$ is independently drawn from a population characterized by a distribution $f(y;\ \theta)$. Here $f(y;\ \theta)$ denotes the probability density function (PDF) (for continuous y) or the probability distribution function (for discrete $y$) of the population, and $\theta$ are unknown parameters. Depending on the distributions, $\theta$ can be a single scalar parameter or a parameter vector.

#### *2.1.1. Likelihood Function*

If $\theta$ were specified, the probability of observing $y_i$, given the population parameter $\theta$, can be written as $f(y_i;\ \theta)$, which is the probability density function or the probability function evaluated at $y_i$. Then the joint probability of observing $(y_1, \ldots, y_n)$ is $\prod_i f(y_i;\ \theta)$. This is the likelihood function. Throughout this chapter, we use interchangeably the notation $L(\theta)$ and $L(\theta;\ y)$ to describe the likelihood function, where $y = (y_1, \ldots, y_n)$. In practice, $\theta$ is unknown and it is our objective to infer the value of $\theta$ from the observed data, in order to describe the population of interest.

The likelihood function appears to be defined the same as the probability or probability density function. However, the likelihood function is a function of $\theta$. Specifically, a probability or probability density function is a function of the data given a particular set of population parameter values, whereas a likelihood is a function of the parameters assuming the observed data are fixed. It measures the relative possibility of different $\theta$ values representing the true population parameter value. For simplicity $L(\theta)$ has been expressed as a function of a single parameter $\theta$. In more general terms, however, the likelihood is a multidimensional function. For many commonly encountered problems, likelihood functions are unimodal; however, they can have multiple modes, particularly in complex models. In addition, a likelihood function may be analytically intractable. In that case, it may be difficult to express it in a simple mathematical form and some form of simplification or linearization of the likelihood may be required (8–10).

### 2.1.2. Log-Likelihood Function

The log-likelihood is defined as the natural logarithm of the likelihood. It is denoted as $LL(\theta) = LL(\theta; y) = \ln(L(\theta; y)) = \sum\limits_i \ln(f(y_i; \theta))$, where ln indicates the natural logarithm. The log-likelihood is a monotonic transformation of the likelihood function, so they both reach the maximum at the same value of $\theta$. In addition, for the frequently used distributions, the $LL(\theta)$ is a simpler function than $L(\theta)$ itself.

For the case where $y$ follows a normal distribution, $f(y; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$, and $\ln(f(y; \theta)) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}$ $(y - \mu)^2$, where $\theta = (\mu, \sigma^2)$. $\mu$ and $\sigma^2$ represent the population mean and variance, respectively. The likelihood function based on data $y_1,\ldots,y_n$, is then $L(\theta) = \prod\limits_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2}$, and the log-likelihood is $LL(\theta) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum\limits_i (y_i - \mu)^2$. For illustration purpose, Fig. 1 shows the likelihood (left panel) and log-likelihood (right panel) functions based on a set of data ($n = 1{,}000$) randomly drawn from a standard normal distribution.

Suppose Y is a discrete variable taking two values, for example, success (1) or failure (0); presence of skin lesions (1) or no skin lesions (0). In statistical terms, $y$ is known to follow a Bernoulli distribution with probability $P(y = 1) = p$ and $P(y = 0) = 1 - p$, where $0 \leq = p \leq = 1$. The probability function of the Bernoulli random variable is $f(y; \theta) = p^y(1 - p)^{(1-y)}$. Note that here $\theta = p$.

Let $y_1,\ldots,y_n$ be $n$ observations from a Bernoulli distribution, where $y_i = 1$ or 0, $i = 1,2,\ldots n$. Of the $n$ observations, $k$ is the

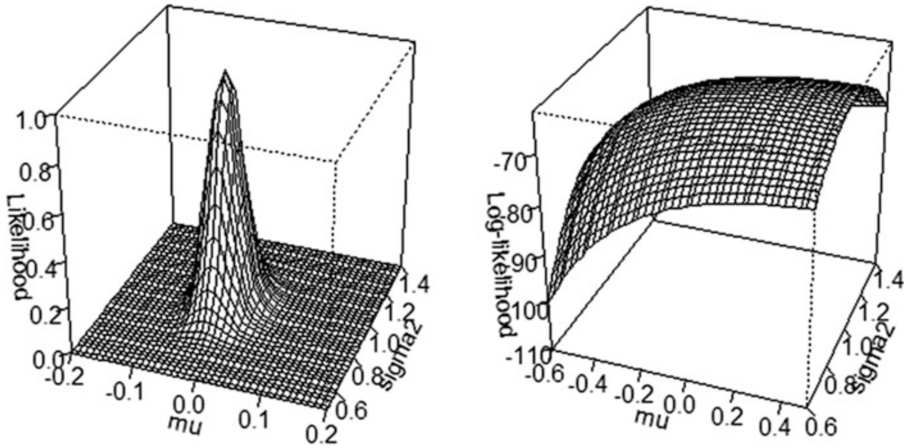

Fig. 1. The likelihood (*left*) and log-likelihood (*right*) functions based on $n = 1{,}000$ samples randomly selected from a standard normal distribution [*mu* denotes the mean, *sigma2* indicates $\sigma^2$].

number of 1s and $n - k$ is the number of 0s. The likelihood corresponding to these data is:

$$\begin{aligned} L(\theta; y) &= p^{y_1}(1-p)^{(1-y_1)} p^{y_2}(1-p)^{(1-y_2)} \ldots p^{y_n}(1-p)^{(1-y_n)} \\ &= p^{\sum_i y_i}(1-p)^{\sum_i (1-y_i)} \\ &= p^k(1-p)^{n-k} \end{aligned}$$

and the log-likelihood is:

$LL(\theta; y) = k\ln(p) + (n-k)ln(1-p)$. In Fig. 2, the top panel shows the likelihood and log-likelihood function of this example for $n = 10$ and $k = 2$.

### *2.1.3. Likelihood Function of Censored Data*

There are cases where a subset $y_{k+1},\ldots,y_n$ of data $y_1,\ldots,y_n$ may not be precisely observed, but the values are known to be either below or above a certain threshold. For example, many laboratory based measurements are censored due to the assay accuracy limit, usually referred to as the lower limit of quantification (LLQ). This happens when the bioanalysis system cannot accurately distinguish the level of component of interest from the system "noise". For such cases, the exact value for $y_i$, $i = k + 1,\ldots n$, is not available. However, it is known that the value is equal to or below the LLQ. Such data are referred to as left censored data.

In other cases, the data are ascertained to be above a certain threshold, with no specific value assigned. For example, in animal experiments, the animals are examined every day to monitor the appearance of particular features, e.g., skin lesions. The time to the appearance of lesions is then recorded and analyzed. For animals with no skin lesions by the end of the study (2 weeks for example),

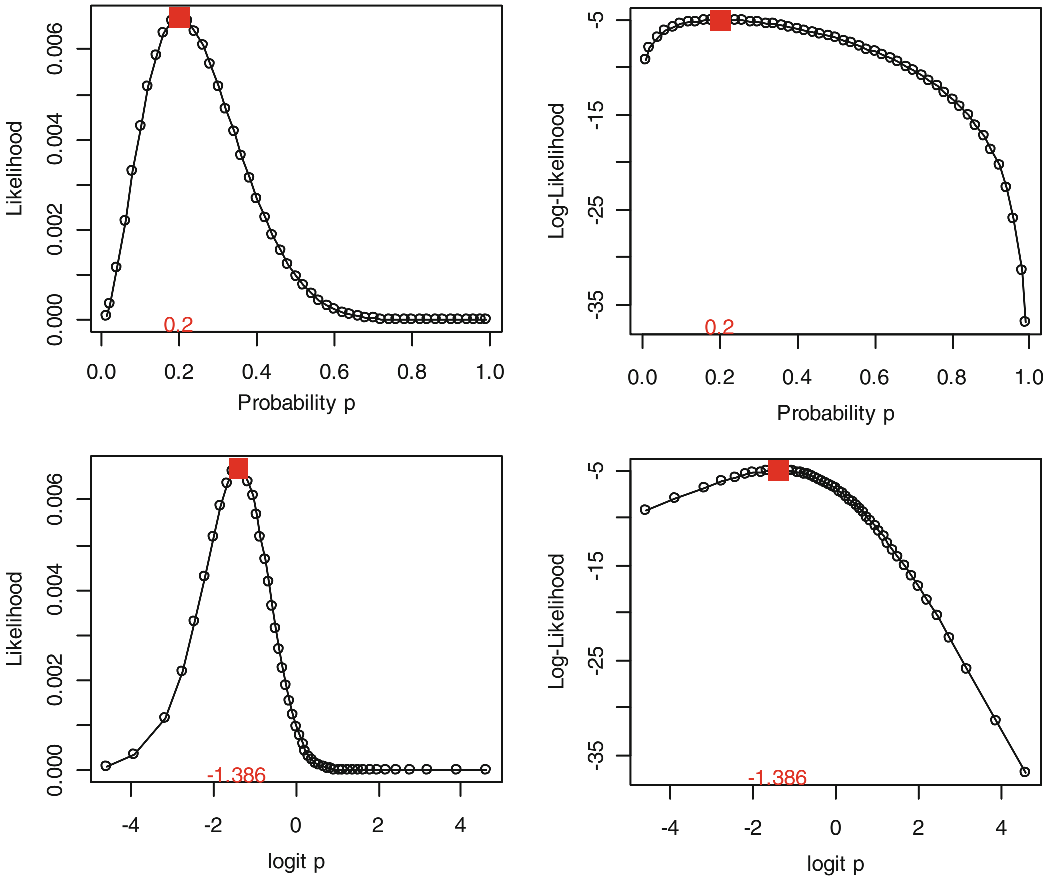

Fig. 2. Likelihood and log-likelihood functions with respect to *p* and *a*. Note: the *red solid* square points mark the maximum of the *L(p)* and *LL(p)*, the texts above the x-axis indicate the MLE of *p* (0.2) or a (logit of *p*) (-1.386).

the time to lesions will be recorded as > 2 weeks. These data are referred to as right censored data. As it is only known that an animal has no lesion at 2 weeks after the treatment, whether the animal will have and when it will have skin lesions is not known.

However, suppose examinations were not carried out between day 7 and 11 and at day 11 an animal was found to have lesions. Then the time to lesions will be between 7 and 11 days, although the exact time of lesions appearance is not known. In this case, the time to lesion for this animal will be interval censored, i.e., it is longer than 7 days, but shorter than 11 days.

When such cases arise in practice, ignoring the characteristics of the data in the analysis may cause biases (see ref. 11 and the references cites therein), so appropriate adjustments must be applied.

Let $y_1,\ldots,y_k$ represent the observed data, and $y_{k+1},\ldots,y_n$ those not precisely observed but known to be left censored (assumed to lie within interval $[-\infty,\text{LLQ}]$ or $[0,\text{LLQ}]$ for laboratory measurements that must be greater than or equal to 0). Assume that the observed or unobserved $y$ follow the same normal distribution $N(\mu, \sigma^2)$, then the likelihood for the precisely observed data $y_1,\ldots, y_k$ is: $L(\theta) = \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2}$, but the probability of the censored $y_i$ $(i = k + 1,\ldots,n)$ needs to be written as: $L_i(\theta) = \int_{-\infty}^{LLQ} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2} dy$ (replace $-\infty$ with 0 if $y_i$ must be greater or equal to 0), which is the cumulative probability up to LLQ of the normal distribution. Note, $\theta = \left(\mu, \sigma^2\right)$.

The full likelihood function of all data $y_1,\ldots,y_k, y_{k+1},\ldots,y_n$ is, therefore:

$$L(\theta; y) = \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2} \prod_{i=k+1}^{n} \int_{-\infty}^{LLQ} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2} dy$$

The likelihood function for interval censored and right censored data, can be written in the exact same way. Instead of integrating from $(-\infty$,LLQ) for the left censoring, integration from (LOW, $+\infty$) for right censored and [LOW,UPP] for interval censored data, where LOW and UPP are the threshold for the lower and upper limit of the observation, respectively.

### 2.2. Maximum Likelihood Estimation

The identification of the maximum likelihood estimation (MLE) is achieved by searching the parameter space (one or multidimensional), to find the parameter values that give the maximum of the likelihood function.

We show below how this is carried out in the case where $\theta$ is a scalar parameter.

*Maximization Process*

From mathematical theory, the maximum of any function is achieved at the point where the first derivative (if exist) is equal to zero. As defined above, $LL(\theta) = \sum_i \ln(f(y_i; \theta))$, so the MLE of $\theta$ satisfies the following equation:

$$\frac{dLL(\theta)}{d\theta} = LL'(\theta) = \sum_i \frac{\frac{df(y_i;\theta)}{d\theta}}{f(y_i;\theta)} = 0$$

where $\frac{dLL(\theta)}{d\theta}$ (or $LL'(\theta)$) and $\frac{df(y_i;\theta)}{d\theta}$ indicate the first derivative of $LL(\theta)$ and $f(y_i;\theta)$ with respect to (w.r.t) parameter $\theta$. Let $\hat{\theta}$ be the solution of the above equation.

It is known that the first derivative is zero at any minimum points as well. In order to get truly the maximum, the second derivative (if exists) evaluated at $\hat{\theta}$ must be negative, i.e.

$$\frac{d^2LL(\hat{\theta})}{d\theta^2} = \sum_i \left( \frac{\frac{d^2f(y_i;\hat{\theta})}{d\theta^2}}{f(y_i;\hat{\theta})} - \frac{(\frac{df(y_i;\hat{\theta})}{d\theta})^2}{f^2(y_i,\hat{\theta})} \right) < 0$$

with $\frac{d^2 LL(\hat{\theta})}{d\theta^2}$ and $\frac{d^2 f(y_i;\hat{\theta})}{d\theta^2}$ denoting the second derivative of $LL(\theta)$ and $f(y_i;\theta)$ w.r.t. $\theta$, and evaluated at $\hat{\theta}$.

For cases where more than one parameter is involved in the log-likelihood function, the partial derivatives of $LL(\theta)$ with respect to each of the parameter will be used (see example 3.3). It is not difficult to imagine that if log-likelihood has a very complicated form, the equations involving the derivatives or the partial derivatives may not be easily solved analytically. Fortunately, many algorithms have been developed to solve these equations, mostly iteratively—that is by repeating a sequence of calculations until the resulting values from subsequent iterations are similar. The Newton-Raphson Optimization Algorithm is one that most commonly used.

Starting from a plausible arbitrary point ($\theta_0$) in the parameter space, the iterative procedure search through the parameter space based on certain rules, which are proved mathematically to ensure that the process will identify the maximum of the log-likelihood function. For example, with the Newton's method, at iteration $n + 1$, $\theta_{n+1} = \theta_n - \frac{LL'(\theta_n)}{LL''(\theta_n)}$, where $\theta_n$ is the parameter value at the $n$-th step, $LL'(\theta_n)$ and $LL''(\theta_n)$ are the first and second derivative of the log-likelihood w.r.t $\theta$, evaluated at $\theta = \theta_n$. The algorithm stops when the process converges which means when $\theta_{n+1}$ is close enough (meets the predefined criteria, usually a small number say, 10E-8) to the value $\theta_n$ at the previous step (12).

*Difficulties*: The maximization process can encounter many problems. For example, when the log-likelihood function is flat with respect to some parameters, the algorithms may have difficulty to find the maximum point. This could indicate lack of information in the data to identity specific parameter values and that more data may be needed. There are also times where the algorithm seems to find a maximum point, but by choosing a different initial point $\theta_0$, the procedure may result in a different maximum. This happens when the log-likelihood function is multimodal. In practice, the general suggestion for such situations is to repeat the algorithm from several diverse starting points and/or modify the search steps (tolerance criteria) to ensure that similar results are achieved.

## 2.3. Properties of MLE

The MLE has several important properties making the maximum likelihood approach an attractive method for statistical inference. The MLEs are, for example, asymptotically normal, consistent, efficient, and parameterization invariant (13). What follows is focused on the properties more used in practice.

### 2.3.1. Asymptotically MLE Follows Normal or Multinormal Distribution

When $n$ (the number of observations) tends to infinity, the MLE of $\theta(\hat{\theta}_{\mathrm{MLE}})$ follows a normal distribution with mean the true parameter $\theta$, and variance (or variance-covariance matrix if $\theta$ is a vector) the inverse of the expectation of the second derivative of $LL(\theta)$ with respect to $\theta$, that is: $\hat{\theta}_{\mathrm{MLE}} \sim N(\theta, I(\theta)^{-1})$, where $I(\theta) = -E_{\theta}[\frac{d^2 LL(\theta)}{d\theta^2}]$. $I(\theta)$ is called the Fisher's information matrix.

Inferences can be made on the basis of this asymptotic distribution.

It should be noted that given the asymptotic nature of these properties, they are not guaranteed for small samples. For example, MLEs obtained from small samples can be biased.

### 2.3.2. Parameterization Invariance

This property states that if $\hat{\theta}_{\mathrm{MLE}}$ is the MLE for parameter $\theta$, then the MLE of any function of $\theta$, say, $g(\theta)$, is defined as $g(\hat{\theta}_{\mathrm{MLE}})$, that is the value of the function $g$ evaluated at $\theta = \hat{\theta}_{\mathrm{MLE}}$,

This parameterization invariance property of MLEs allows flexibility in choosing model parameterizations, which is important in cases where parameter transformation can make the maximization step simpler and easier. Example 3.2 illustrates how parameters can be transformed in practice in order to obtain appropriate estimate of the unknown parameters of interest.

## 2.4. Confidence Interval from MLE

It is important to understand and characterize the uncertainty of the MLE if only one experiment is done and the MLE of the population or model parameters is obtained from data which are generated from that particular experiment. This is usually achieved by specifying a confidence interval for the unknown $\theta$ around the MLE.

From the previous section, $\hat{\theta}_{\mathrm{MLE}} \sim N(\theta, I(\theta)^{-1})$. In fact, $I(\theta)$ is not known as $\theta$ is unknown. In practice, $I(\theta)$ is usually approximated by plugging in the estimated value $\hat{\theta}_{\mathrm{MLE}}$, then obtaining $I\left(\hat{\theta}_{\mathrm{MLE}}\right) = \frac{d^2 LL(\hat{\theta}_{\mathrm{MLE}})}{d\theta^2}$. This is used to construct a confidence interval around the MLE.

The approximate $(1 - 2\alpha)\%$ confidence interval for the unknown $\theta$ around the corresponding MLE is:

$$\hat{\theta}_{\mathrm{MLE}} \pm z_{\alpha} I\left(\hat{\theta}_{\mathrm{MLE}}\right)^{-1/2}$$

where $z_{\alpha}$is the critical value of the standard normal distribution corresponding to the chosen $\alpha$ level, for example, $z_{\alpha} = 1.96$ for $\alpha = 0.025$, and 1.64 for $\alpha = 0.05$.

As illustrated in Subheading 3.2, users should be cautious when using this method to calculate the confidence interval for a parameter of interest.

## 2.5. Likelihood Ratio Test

Often several different models can be found to describe a population from which the data are selected. The problem is then to identify which model is more appropriate to explain the data and represent the population. It is possible to use likelihood theory to test the appropriateness of a given model in comparison to a model from the same family but with a different number of parameters (nested models). For example, assume $L_{\mathrm{A}}$ is the maximum of the likelihood function corresponding to a particular model (model A) and $L_{\mathrm{B}}$ is the maximum of the same model but with smaller number of parameters (model B). Let $k$ be the difference in the number of parameters between the two models. Then the likelihood ratio $R$ is:

$$R = -2\ln\left(\frac{L_{\mathrm{B}}}{L_{\mathrm{A}}}\right) = -2(\ln(L_{\mathrm{B}}) - \ln(L_{\mathrm{A}}))$$
$$= -2\ln(L_{\mathrm{B}}) - (-2\ln(L_{\mathrm{A}}))$$

$R$ has approximately a Chi-squared distribution with $k$-degrees of freedom if model B is the true model. The calculated value of $R$ should be consistent with such a Chi-squared distribution. Large value of $R$ (with small $p$ value) constitutes evidence against the model with fewer parameters, B, in favor of model A.

It is noted that sometimes, −2 times the log-likelihood is minimized instead of maximizing the log-likelihood in many statistical software.

### 2.6. Further Reading and Statistical Software

Readers who are interested in the general theory and application of the maximum likelihood method are referred to refs. 13–18.

A number of statistical packages are available for obtaining maximum likelihood estimates of model parameters in both the linear or nonlinear modelling fields. Packages commonly used in academia and pharmaceutical industry include, SAS (19), Stata (20), and R (21). In addition, a few other packages are dedicated to the analysis of non linear mixed effects models, like NONMEM (10) and Monolix (22). R and Monolix are freely downloadable online.

## 3. Examples

### 3.1. Maximum Likelihood Estimation of Exponential Distribution

Suppose random variable $y$ represent the time to an event, e.g., a certain toxicity from an experiment, and it is assumed that $y$ follows an exponential distribution with density function: $f(y;\theta) = \frac{1}{\theta}e^{-\frac{y}{\theta}}$, where $y > 0$.

Let $y_1,\ldots,y_n$ be a random sample from this exponential distribution, then the likelihood and log-likelihood functions are given by:

$L(\theta) = \left(\frac{1}{\theta}e^{-\frac{y_1}{\theta}}\right)\left(\frac{1}{\theta}e^{-\frac{y_2}{\theta}}\right)\ldots\left(\frac{1}{\theta}e^{-\frac{y_n}{\theta}}\right) = \frac{1}{\theta^n}e^{-\sum\limits_i \frac{y_i}{\theta}}$, and $LL(\theta) = -n\ ln(\theta) - \frac{1}{\theta}\sum\limits_i y_i$.

The solution of the equation $\frac{d(LL(\theta))}{d\theta} = -\frac{n}{\theta} + \frac{\sum_i y_i}{\theta^2} = 0$ is: $\hat{\theta} = \frac{1}{n}\sum\limits_i y_i$, which is the mean of sample $y_1,\ldots,y_n$, and denoted by $\bar{y}$.

It is noted that: $\frac{d^2(LL(\theta))}{d^2\theta} = \frac{n}{\theta^2} - 2\frac{n\bar{y}}{\theta^3} = \frac{n\theta - 2n\bar{y}}{\theta^3}$, which evaluated at $\hat{\theta}$, is such that $\frac{d^2\left(LL(\hat{\theta})\right)}{d^2\theta} < 0$. Therefore $LL(\theta)$ has a maximum at $\bar{y}$, and $\hat{\theta}_{\mathrm{MLE}} = \frac{1}{n}\sum\limits_i y_i = \bar{y}$.

Note: $\frac{d(LL(\theta))}{d\theta}$ represents the derivative of $LL(\theta)$ with respective to $\theta$, and $\frac{d^2(LL(\theta))}{d^2\theta}$ is the derivative of $\frac{d(LL(\theta))}{d\theta}$ with respective of parameter $\theta$, also the second order derivative of $LL(\theta)$ with respective to $\theta$.

An alternative way of proving that $LL(\theta)$ has maximum at $\bar{y}$ is through the analysis of the behavior of $\frac{d(LL(\theta))}{d\theta}$:

rewrite $\frac{d(LL(\theta))}{d\theta}$ as: $\frac{d(LL(\theta))}{d\theta} = -\frac{n}{\theta} + \frac{n\bar{y}}{\theta^2} = -\frac{n\theta}{\theta^2} + \frac{n\bar{y}}{\theta^2} = \frac{n\bar{y}-n\theta}{\theta^2} = \frac{n(\bar{y}-\theta)}{\theta^2}$

So if $\theta < \bar{y}$, then $\frac{d(LL(\theta))}{d\theta} > 0$, indicating $LL(\theta)$ is increasing, and if $\theta > \bar{y}$, then $\frac{d(LL(\theta))}{d\theta} < 0$, indicating $LL(\theta)$ is decreasing and therefore $LL(\theta)$ has the maximum at $\bar{y}$.

### 3.2. Probability of Toxicity

In an animal toxicity study, ten cynomolgus monkeys were administered a 100 mg of compound X intravenously every week for 8 weeks. During the study, two out of the ten monkeys developed skin lesions. What is the probability of having the skin lesions in the entire population?

*Solution*:

Let $y = 1$ indicate the event of having skin lesions, and $y = 0$ indicate no skin lesions, where y following a Bernoulli distribution with the probability of having skin lesions $p$.

The data obtained from the monkey study were: $(y_1, \ldots, y_{10}) = (0,0,0,0,1,0,0,0,0,1)$.

Then the likelihood and log-likelihood of these data can be written as:

$$L(p) = p^k(1-p)^{n-k} \text{ and } LL(\theta) = k\ln(p) + (n-k)\ln(1-p),$$

where $p$ is the unknown parameter, $k = 2$ and $n = 10$ (see Fig. 2), so $LL(p) = 2\ln(p) + 8\ln(1-p)$.

Solving the equation $\frac{dLL(p)}{dp} = \frac{k}{p} - \frac{n-k}{1-p} = 0$ gives: $\hat{p} = \frac{k}{n} = \frac{2}{10} = 0.2$

To confirm that this is the maximum value of the log-likelihood function, the second derivative of the $LL(p)$ is calculated:

$\frac{d^2LL(p)}{dp^2} = -\frac{k}{p^2} + \frac{n-k}{(1-p)^2}$, substituting $\hat{p} = 0.2$ gives $\frac{d^2LL(\hat{p})}{dp^2} < 0$. Therefore, $\hat{p}$ is the maximum likelihood estimate of the population parameter $p$.

As illustrated in the above Subheading 2.4, the confidence interval of $p$ around $\hat{p}$, can be calculated using: $I(\hat{p}) = -\frac{d^2LL(\hat{p})}{dp^2} = \frac{n^3}{k(n-k)} = 62.5$, the standard error of $\hat{p}$ is 0.126, therefore, the 90% confidence interval of parameter $p$ around $\hat{p}$ would be $(-0.05, 0.41)$, assuming $p$ has an asymptotical normal distribution.

However, it is known that probability is between 0 and 1. In order to maintain this assumption throughout the calculation, a logit function is used to transform the probability into a variable that can take values between $-\infty$ and $\infty$, where $\text{logit}(p) = \ln(\frac{p}{1-p})$.

```
# y=(y1,...,yn), n=10, two of the yis are 1, eight of them are 0.


> res1 <-glm(y~1,family=binomial,data=d)
> summary(res1)
Call:
glm(formula = y ~ 1, family = binomial)
Deviance Residuals:
  Min      1Q   Median     3Q      Max
-0.668  -0.668  -0.668  -0.668   1.794
Coefficients:
           Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -1.3863     0.7906  -1.754   0.0795 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 10.008  on 9  degrees of freedom
Residual deviance: 10.008  on 9  degrees of freedom
AIC: 12.008
Number of Fisher Scoring iterations: 4
```

Fig. 3. R-code and outputs to obtain MLE of *p*.

Using this transformation, assume $a = \text{logit}(p)$, then $p = \frac{e^a}{1+e^a}$. Replace $p$ with $a$ in the above derivative functions, we then have:

$$LL(p) = k\ln(p) + (n-k)\ln(1-p) = ka - n\ln(1+e^a) = LL(a)$$

Solving equation $\frac{dLL(a)}{da} = k - n\frac{e^a}{1+e^a} = 0$, we have $\hat{a} = \ln(\frac{k}{n-k})$. As $I(\hat{a}) = -\frac{d^2LL(\hat{a})}{da^2} = \frac{k(n-k)}{n} = 1.6$, the standard error of $\hat{a}$ is approximately 0.79. The 90% confidence interval of a is then $(-2.68, -0.09)$.

According to the parameterization invariance property of the MLE, the MLE of parameter $p$ can be calculated by back transforming the logit function, thus $\hat{p} = \frac{e^{\hat{a}}}{1+e^{\hat{a}}} = \frac{k}{n} = 0.2$ and its 90% confidence interval is (0.06, 0.48). Note: $\frac{e^{-2.68}}{1+e^{-2.68}} = 0.06$, and $\frac{e^{-0.09}}{1+e^{-0.09}} = 0.48$. Figure 3 below gives the R code to obtain the MLEs of $a$ and $p$.

Figure 2 below gives the R code to obtain the MLEs of $a$ and $p$.

The readers are referred to the R manual (21) for details on how to set up models in R and the interpretation of the parameter estimates. Specifically for this example, the logit of probability $p$, i.e., $a = \log\text{it}(p) = \ln(\frac{p}{1-p})$ is estimated, and denoted as intercept in the R output.

Therefore $\hat{a} = -1.3863$ and its standard error is 0.7906. These values are similar to that calculated manually above. The 90% confidence interval of parameter a around its MLE is $(-2.6829, -0.0897)$. Back transform the logit function, we have $\hat{p} = \frac{\mathrm{e}^{\hat{a}}}{1+\mathrm{e}^{\hat{a}}} = 0.2$, and the 90% confidence interval of $p$ is (0.064, 0.4776).

### 3.3. Linear Regression

Assume that $y_i = b_0 + bx_i + \varepsilon_i$, where $i = 1,2,\ldots,n$, the $y_i$s are independently drawn from a population, the $x_i$s are independent variables, and that $\varepsilon_i \sim N(0, \sigma^2)$ is residual error.

The likelihood of $y = (y_1,\ldots,y_n)$ is: $L(\theta) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - b_0 - bx_i)^2}$, where $\theta = (b_0, b, \sigma^2)$,

$$LL(\theta) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_i (y_i - b_0 - bx_i)^2$$

or

$$-2LL(\theta) = n(\ln(2\pi) + \ln(\sigma^2)) + \frac{1}{\sigma^2}\sum_i (y_i - b_0 - bx_i)^2$$

The MLE of $b_0$, $b$ as well as $\sigma^2$ are obtained by maximizing the $LL(\theta)$ or minimizing the $-2LL(\theta)$. The minimization of $-LL(\theta)$ is illustrated below. The minimization is achieved by solving the following equations simultaneously.

$$\frac{\partial(-2LL(\theta))}{\partial b_0} = -2\frac{1}{\sigma^2}\sum_i \left(y_i - b_0 - bx_i\right) = 0 \qquad (1)$$

$$\frac{\partial(-2LL(\theta))}{\partial b} = -2\frac{1}{\sigma^2}\sum_i x_i\left(y_i - b_0 - bx_i\right) = 0 \qquad (2)$$

$$\frac{\partial(-2LL(\theta))}{\partial \sigma^2} = \frac{n}{\sigma^2} - \frac{1}{\sigma^4}\sum_i (y_i - b_0 - bx_i)^2 = 0 \qquad (3)$$

where $\frac{\partial(-2LL(\theta))}{\partial b_0}$, $\frac{\partial(-2LL(\theta))}{\partial b}$ and $\frac{\partial(-2LL(\theta))}{\partial \sigma^2}$ represent the first order partial derivation of $-2LL(\theta)$ with respect to $b_0$, $b$ and $\sigma^2$, respectively.

Solving the above three equations, we have:

$\hat{b}_{\text{MLE}} = \frac{\sum_i x_i y_i - \sum_i x_i \sum_i y_i}{\sum_i x_i^2 + \left(\sum_i x_i\right)^2}$, $\quad \hat{b}_{0\text{MLE}} = \frac{1}{n}\sum_i (y_i - bx_i)$ and $\hat{\sigma}^2{}_{\text{MLE}} = \frac{\sum_i (y_i - b_0 - bx_i)^2}{n}$

It is noted that the MLEs of $b$, $b_0$ are equivalent to their corresponding least squares estimates.

### 3.4. Dose Response Model

In the early drug development, compound X was tested in monkeys to assess its toxicological effects. Three doses (10, 100, 300 mg) of compound X and placebo were given to 40 monkeys, 10 monkeys in each dose group, every week for 8 weeks. During the 8 weeks of study, the appearance of skin lesion was observed and recorded. The question is whether the probability of skin lesion is associated with the dosage given. The number of skin lesion in each dose group was: 0, 1, 5, 9 for placebo, 10 mg, 100 mg and 300 mg group, respectively.

*Solutions*:

Let $y = 1$ indicate the presence of skin lesion, and $y = 0$ indicate no skin lesion. The question can then be rephrased by asking whether

$p = P(y = 1)$ is related to the dosage. Logistic regression analysis is a technique to analyze this type of data, where the dichotomous depend variable is specified in terms of several independent variables.

The basis of logistic regression is to model the logit transformation of the probability $p$ as a linear function of the independent variables. For this example, $p$ is the probability of skin lesion, and the independent variable is the dose.

Assume $D$ represents the dose administered. The logistic regression can be written as: $\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = a + b\ln(D + 1)$, where $a$ and $b$ are parameters of the model.

For the $i$-th animal, the corresponding probability of having skin lesions is described as $p_i$, and $\text{logit}(p_i) = a + b\ln(D_i + 1)$. Then the likelihood of observing the data as described above is:

$L(\theta) = \prod_i p_i^{y_i}(1 - p_i)^{1-y_i}$ and $LL(\theta) = \sum_i [y_i\ln(p_i) + (1 - y_i)$ $\ln(1 - p_i)]$. Given $p_i$ is a function of unknown parameters $a$ and $b$, then $LL(\theta)$ is a function of $a$ and $b$, and $\theta = (a, b)$.

The partial derivatives of $LL(\theta)$ with respect to $a$ and $b$ are:

$$\frac{\partial LL(\theta)}{\partial a} = \sum_i \left(\frac{y_i}{p_i} - \frac{1-y_i}{1-p_i}\right) \text{ and } \frac{\partial LL(\theta)}{\partial b} = \sum_i \left[\left(\frac{y_i}{p_i} - \frac{1-y_i}{1-p_i}\right)\ln(D_i + 1)\right]$$

The MLE of $a$ and $b$ can be obtained by solving the equations $\frac{\partial LL(\theta)}{\partial a} = 0$ and $\frac{\partial LL(\theta)}{\partial b} = 0$. Although these equations do not look complicated, solving them analytically is not easy and a numerical solution is required. In the following (Fig. 4), the results using ***glm*** in R (21) are presented. The MLE of $a$ (denoted as *Intercept* ) and $b$ (denoted as *log*(*dose + 1*)) are −5.7448 and 1.3118, with standard error of 2.0339 and 0.4324, respectively.

To test if increasing the dose has significantly increased the probability of skin lesions statistically, we can use the likelihood ratio test as described in Subheading 2.5. In the R output above, the Null deviance and Residual deviance are given, where Null deviance is the deviance of a null model where only intercept is fitted, and the Residual deviance is the deviance of the specified model. Note: the deviance in this case is defined as the minus twice the maximized log-likelihood evaluated at MLE of $\theta$ (i.e., $-2LL(\hat{\theta})$). The likelihood ratio is $R = 25.4$, with one degree of freedom. On the basis of a Chi-squared distribution with one degree of freedom, this corresponds to a $p$-value of 4.62e-07, indicative of evidence against the null model. The conclusion is that the probability of skin lesions is statistically significantly related to the dosage and it is increased with increasing dose. Figure 5 depicts the model predicted probability of having skin lesions and their 95% confidence intervals versus the amount of drug administered.

For any given dose of the compound, the probability of skin lesions can be calculated by back transform the logit function, i.e.,

```
#  y=(y1,...,yn), and doseis a vector of doses given to each of the n animals

> Res.logit <-glm(y~log(dose+1),family=binomiall,data=d)

> Summary(res.logit)

Call:
glm(formula = y ~ log(dose + 1), family = "binomial", data = d)

Deviance Residuals:
     Min        1Q     Median        3Q       Max
-1.95101  -0.37867  -0.07993   0.56824   2.31126

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)     -5.7448      2.0339 -2.825  0.00473 **
log(dose + 1)    1.3118      0.4324   3.034  0.00242 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

   Null deviance: 52.925  on 39  degrees of freedom
Residual deviance: 27.510  on 38  degrees of freedom
AIC: 31.51

Number of Fisher Scoring iterations: 6
```

Fig. 4. The R-code and outputs to obtain the MLE of *a* and *b*.

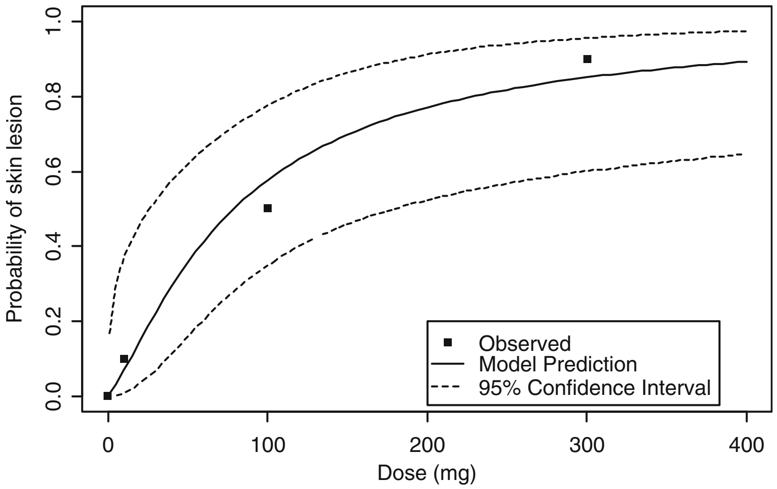

Fig. 5. Observed and model predicted probability of skin lesion (*black dots* are the observed proportion of monkeys having skin lesion for the corresponding dose group; *solid line* is the model predicted probability of skin lesion; *dashed lines* are 95 % confidence interval of probability (p); Dose = 1 represents placebo).

$p = \frac{\exp(a+b\ln(D+1))}{1+\exp(a+b\ln(D+1))}$. For example, when no drug is given, i.e., $D = 0$, the probability of having skin lesions is $\frac{e^{\hat{a}}}{1+e^{\hat{a}}} = 0.003$ and the 95% confidence interval is (0,0.15). When $D = 200$ mg, the probability of skin lesions is 0.77 with 95% confidence interval of (0.52, 0.91). Note that the confidence interval is calculated using the formula as described in Subheading 2.4.

## References

1. Hald A (1999) On the history of maximum likelihood in relation to inverse probability and least squares. Statist Sci 14(2):214–222
2. Aldrich J (1997) R.A. Fisher and the making of maximum likelihood 1912–1922. Statist Sci 12 (3):162–176
3. Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrox BN, Caski F (eds) Second international symposium on information theory. Akademiai Kiado, Budapest, pp 267–281
4. Schwarz G (1978) Estimating the dimension of a model. Ann Statist 6:461–464
5. McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman and Hall, New York
6. Cox DR (1970) The analysis of binary data. Chapman and Hall, London
7. Cox DR (1972) Regression models and life tables. J Roy Statist Soc 34:187–220
8. Lindsey JK (2001) Nonlinear models in medical statistics. Oxford University Press, Oxford, UK
9. Wu L (2010) Mixed effects models for complex data. Chapman and Hall, London
10. Beal SL, Sheiner LB, Boeckmann AJ (eds) (1989–2009) NONMEM users guides. Icon development solutions. Ellicott City
11. Yang S, Roger J (2010) Evaluations of Bayesian and maximum likelihood methods in PK models with below-quantification-limit data. Pharm Stat 9(4):313–330
12. Fletcher R (1987) Practical methods of optimization, 2nd edn. Wiley, New York
13. Young GA, Smith RL (2005) Essentials of statistical inference, chapter 8. Cambridge University Press, Cambridge, UK
14. Bickel PJ, Doksum KA (1977) Mathematical statistics. Holden-day, Inc., Oakland, CA
15. Casella G, Berger RL (2002) Statistical inference, 2nd edn. Pacific Grove, Duxberry, CA
16. DeGroot MH, Schervish MJ (2002) Probability and statistics, 3rd edn. Addison-Wesley, Boston, MA
17. Spanos A (1999) Probability theory and statistical inference. Cambridge University Press, Cambridge, UK
18. Pawitan Y (2001) In all likelihood: statistical modelling and inference using likelihood. Cambridge University Press, Cambridge, UK
19. SAS Institute Inc. (2009) SAS manuals. http://support.sas.com/documentation/index.html
20. STATA Data analysis and statistical software. http://www.stata.com/
21. The R project for statistical computing. http://www.r-project.org/
22. The Monolix software. http://www.monolix.org/