Topic 2-1: Likelihood Construction & Estimation Univariate Models

Department of Experimental Statistics
Louisiana State University

Date

1. Introduction

- Constructing the likelihood of the data is the foundation of model-based statistical inference.
- It leads to essentially automatic methods of inference, including point and interval estimation.
- In this topic, we will focus on constructing the likelihood functions for various types of data, including discrete, continuous, mixture of discrete and continuous, and so on.

Definition of Likelihood Functions

- If random variables $\boldsymbol{Y} = (Y_1, \cdots, Y_n)^T$ has *joint density (or joint probaility mass function)* $f(\boldsymbol{Y}; \boldsymbol{\theta})$ with unknown parameters $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_b)^T$, then the function of $\boldsymbol{\theta}$ defined by

$$L(\boldsymbol{\theta}|\boldsymbol{Y} = \boldsymbol{y}) = f(\boldsymbol{Y} = \boldsymbol{y}; \boldsymbol{\theta}),$$

  where $\boldsymbol{y} = (y_1, \cdots, y_n)^T$ is the observed data points from $\boldsymbol{Y}$, is the **likelihood function**.
- That is, the likelihood function is just the joint density (or probability mass function) evaluated at the observed data points.

Likeliihood Functions for IID Data

- If the random variables $Y_1, \cdots, Y_n$ are indenpendent but each Y_i can have a different density $f_i(Y_i; \boldsymbol{\theta})$, then the likelihood function becomes

$$L(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{n} f_i(Y_i = y_i; \boldsymbol{\theta}).$$

- If $Y_1, \cdots, Y_n$ are **indenpendent and indentically distributed (iid)** random variables following density f, then the likelihood function becomes

$$L(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{n} f(Y_i = y_i; \boldsymbol{\theta}).$$

Discrete IID Random Variables: A Poisson Example

- Fetal lamb movements data (Example 2.1): Leroux and Puterman (1992) give data on counts of movements in 240 five-second intervals of one fetal lamb:

No. of movements	0	1	2	3	4	5	6	7
Widgets	182	41	12	2	2	0	0	1

- Suppose the counts of movements are from iid random variables $Y_1, \cdots, Y_n$ following the *Poisson probability mass function* with $\theta = \lambda$ and

$$f(y;\lambda) = \frac{\lambda^y e^{-y}}{y!}, y = 0, 1, ...$$

Then, the likelihood function is

$$L(\lambda|\mathbf{y}) = \prod_{i=1}^{n} f(y_i;\lambda) = \prod_{i=1}^{n} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} = \lambda^{n\bar{y}} e^{-n\lambda} \left(\prod_{i=1}^{n} y_i! \right)^{-1},$$

where $\bar{y} = \sum_{i=1}^{n} y_i / n$.

Maximum Likelihood Estimator (MLE)

- The value of $\boldsymbol{\theta}$ that maximizes the likelihood function $L(\boldsymbol{\theta}; \boldsymbol{y})$ is called the **maximum likelihood estimator (MLE)** denoted by $\hat{\boldsymbol{\theta}}_{MLE}$, i.e.,

$$\hat{\boldsymbol{\theta}}_{MLE} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\boldsymbol{y})$$

.

- Such estimator is generally *optimal or at least optimal* in large samples (Fisher 1922), which we will discuss more in a future lecture.

Finding the MLE

- In practice, $\hat{\boldsymbol{\theta}}_{MLE}$ is usually calculated by finding the optimizer of the **log likelihood function** $\log(L(\boldsymbol{\theta}|\boldsymbol{y}))$.
- If the likelihood function is differentiable in θ_i for $i = 1, \cdots, b$, then the possible candidates for the MLE are the values of $(\theta_1, \cdots, \theta_b)$ that solve

$$\frac{\partial}{\partial \theta_i} \log(L(\boldsymbol{\theta}|\boldsymbol{y})) = 0 \text{ for } i = 1, \cdots, b.$$

Solving the equations may give you *local or global minima, local or global maxima, or inflection points*. Then, some further arguments, such as checking that the second derivatives evaluated at the cadidate is less than 0, is needed for finding a global maximum.

The Poisson Example (Cont.)

- Recall: The likelihood function from the Poisson example (page 5) is $L(\lambda|\boldsymbol{y}) = \lambda^{n\bar{y}} e^{-n\lambda} \left(\prod_{i=1}^{n} y_i!\right)^{-1}$.
- The log likelihood function is

$$\log\left[L(\lambda|\boldsymbol{y})\right] = n\bar{y}\log\lambda - n\lambda + \sum_{i=1}^{n}\log y_i!.$$

- Equate the derivative of the log likelihood with respect to λ to zero, we obtain $\frac{n\bar{y}}{\lambda} - n = 0$, i.e., the MLE candidate of λ is $\bar{y} = \frac{86}{240} = .358$.
- Because the second derivative of the log likelihood function is $\frac{-n\bar{y}}{\lambda^2} < 0$ for all λ, the function is *concave* and $\bar{y}$ is the global maximum. That is,

$$\hat{\lambda}_{MLE} = \bar{y}.$$

Using R to find MLE: The Poisson Example

```
#data from the Poisson Example
count.v ← rep(c(0,1,2,3,4,5,6,7), c(182, 41, 12, 2,
    2, 0, 0, 1))

#Log-likelihood function
poisson.log.like ← function(lambda,data){
  -sum(dpois(data, lambda, log = TRUE))
}

#Find MLE
fit.result ← nlminb(start = 0.5, objective =
    poisson.log.like, data = count.v, lower = 0,
    upper = Inf)
fit.result$par
[1] 0.3583333

#theoretical MLE
mean(count.v)
[1] 0.3583333
```

A Multinomial Example

- Consider n independent trials, where each trail has $k \geq 2$ outcomes and has probability p_i to be i-th outcome for $i = 1, \cdots, k$.
- Denote N_i is the number of trials that are i-th outcome. Then, $(N_1, \cdots, N_k)$ are distributed as a *multinomial distribution* with parameter n and $\mathbf{p} = (p_1, \cdots, p_k)$ with constraint $\sum_{i=1}^{k} p_i = 1$ denoted by *multinomial*$(n; \mathbf{p})$, and the likelihood of the observed trails $N_1 = n_1, \cdots, N_k = n_k$ is

$$L(\mathbf{p}|n_1, \cdots, n_k) = \frac{n!}{n_1!n_2! \cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}.$$

- Binomial is a special case of the multinomial distribution with $k = 2$.
- See Example 2.2 for a k = 3 case.

Continuous IID Random Variables

Hurricane data: Larsen and Marx (2001) collect 36 hurricanes that had moved far inland on the East Coast of the U.S. in 1900-1969:

```
> Hurr.rain
[1] 31.00 2.82 3.98 4.02 9.50 4.50 11.40 10.71
[9] 6.31 4.95 5.64 5.51 13.40 9.72 6.47 10.16
[17] 4.21 11.60 4.75 6.85 6.25 3.42 11.80 0.80
[25] 3.69 3.10 22.22 7.43 5.00 4.58 4.46 8.00
[33] 3.73 3.50 6.20 0.67
```

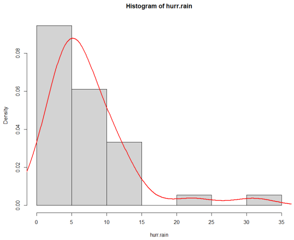

Figure 1: Hurricane data

Continuous IID Variables

- Consider iid sample data $y_1, \cdots, y_n$ from continuous random variable with density $f(y; \boldsymbol{\theta})$
- For example, if f is the *gamma density* $f(y; \boldsymbol{\theta} = (\alpha, \beta)) = \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-y/\beta}$, then the likelihood function is

$$\prod_{i=1}^{n} \frac{1}{\Gamma(\alpha)\beta^\alpha} y_i^{\alpha-1} e^{-y_i/\beta} = \{\Gamma(\alpha)\}^{-n} \beta^{-n\alpha} \{\prod_{i=1}^{n} y_i\}^{\alpha-1} e^{-\sum_{i=1}^{n} y_i/\beta}$$

and the log likelihood is

$$\ell(\boldsymbol{\theta}) = -n \log \Gamma(\alpha) - n\alpha \log \beta + (\alpha - 1) \sum_{i=1}^{n} \log y_i - \frac{\sum_{i=1}^{n} y_i}{\beta}.$$

Using R to find MLE: The Hurricane Example

```
#data from the Hurricane Example
hurr.rain ← c(31.00, 2.82, 3.98, 4.02, 9.50, 4.50,
    11.40, 10.71, 6.31, 4.95, 5.64, 5.51, 13.40, 9
    .72, 6.47, 10.16, 4.21, 11.60, 4.75, 6.85, 6.25
    , 3.42, 11.80, 0.80, 3.69, 3.10, 22.22, 7.43, 5
    .00, 4.58, 4.46, 8.00, 3.73, 3.50, 6.20, 0.67)

#Log-likelihood function
llik.gamma ← function(theta, dta=hurr.rain){
  −sum(dgamma(dta, shape=theta[1], scale=theta[2],
      log = TRUE))
}

#Find MLE
fit.result ← nlm(llik.gamma, c(1,2), dta=hurr.rain)
fit.result$estimate
[1] 2.187214 3.331863
```

References

- More examples for constructing likelihood associated with iid data can be found in Section 7.2.2 of Casella and Berger (2002).
  - Examples include Normal (7.2.5-6,7.2.11-12), Bernoulli (7.2.7), restricted MLE (7.2.8), Binomial (7.2.9) .
  - See how they argue the MLE is global maximum.
- See Example 2.1 of the textbook for a *zero inflate Poisson distribution* example, and Example 2.2 for a multinomial example.

2. Connection of Discrete and Continuous Likelihood

- The continuous-data likelihood $\prod_{i=1}^{n} f(y_i; \boldsymbol{\theta})$ looks the same as that for discrete data, but it is not a probability as it is for discrete data. (Note: It is from a density function, assigning zero probability to any single point)
- We can think the density evaluate at $Y = y$ as the limit probability on the small interval $[y - h, y + h]$, i.e.,

$$f(y) = \lim_{h \to 0^+} \frac{F(y - h) - F(y - h)}{2h} = \lim_{h \to 0^+} \frac{P(Y \in (y - h, y + h))}{2h}$$

- If Y now is a discrete random variable evaluated at y, then

$$\lim_{h \to 0^+} F(y + h) - F(y - h) = \lim_{h \to 0^+} F(y^+) - F(y^-) = f(y)$$

- More details can be found in section 2.2.3a of the textbook. This provides a unified perspective to connect discrete and continuous likelihood, and a broader definition of the likelihood is given in the next page.

A Working Definition of the Likelihood

- Suppose $Y_1, \cdots, Y_n$ are independent random variables, and Y_i has distribution function $F_{Y_i}(y_i; \boldsymbol{\theta})$.
- The likelihood of data $\boldsymbol{y} = (y_1, \cdots, y_n)$ observed from $(Y_1, \cdots, Y_n)$ is

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = \lim_{h \to 0^+} \left(\frac{1}{2h}\right)^m \prod_{i=1}^{n} \{F_i(y_i + h; \boldsymbol{\theta}) - F_i(y_i - h; \boldsymbol{\theta})\}, \quad (1)$$

  where $1 \leq m \leq n$ depends on the number of continuous component in the data.
- In following slides, this general **working definition of the likelihood function** is applied to more complicated examples.

3. Mixture of Discrete and Continuous components

- Some data $Y = y$, such as daily rainfall, often have a number of zeros (no rains), and the amounts greater than zero are best modeled by a continuous distribution.
- Such data are often modeled by a mixture of a point mass at zero $P(Y = 0) = 0$ and a continuous positive random variable T having distribution function $F_T(y; \boldsymbol{\theta})$. This means the distribution function of Y is

$$F_Y(y; p, \boldsymbol{\theta}) = pI(0 \leq y) + (1 - p)F_T(y; \boldsymbol{\theta}) \qquad (2)$$

Mixture of Discrete and Continuous components (Cont.)

- Suppose there are iid data $y_1, \cdots, y_n$ from the mixture density (2), and n_0 observations of the data are 0.
- By the working definition of the likelihood (1), the likelihood function of $y_1, \cdots, y_n$ is

$$
\begin{aligned}
L(\boldsymbol{\theta}|\boldsymbol{y}) &= \lim_{h\to 0^+} \left(\frac{1}{2h}\right)^m \prod_{i=1}^{n} \{F_Y(y_i+h;p,\boldsymbol{\theta}) - F_Y(y_i-h;p,\boldsymbol{\theta})\} \\
&= \lim_{h\to 0^+} \{F_Y(h;p,\boldsymbol{\theta}) - F_Y(h;p,\boldsymbol{\theta})\}^{n_0} \times \\
&\quad \lim_{h\to 0^+} \prod_{Y_i>0}^{n} \left\{ \frac{F_Y(y_i+h;p,\boldsymbol{\theta}) - F_Y(y_i-h;p,\boldsymbol{\theta})}{2h} \right\} \\
&= \lim_{h\to 0^+} \{p+(1-p)F_T(h;\boldsymbol{\theta})\}^{n_0} \times \\
&\quad \prod_{Y_i>0}^{n} \left\{ \frac{(1-p)F_T(y_i+h;\boldsymbol{\theta}) - (1-p)F_T(y_i-h;\boldsymbol{\theta})}{2h} \right\} \\
&= p^{n_0}(1-p)^{n-n_0} \prod_{y_i>0}^{n} f_T(y_i:\boldsymbol{\theta})
\end{aligned}
$$

4. Proportional Likelihoods

- Suppose there are data $y_1, \cdots, y_n$ from a continuous distribution with density $f_Y(y; \boldsymbol{\theta})$.
- We are interested in constructing the likelihood of the transformed data $x_i = g(y_i)$ for $i = 1, \cdots, n$, where g is a known, increasing, continuously differentiable function.
- Note: The assumptions of g implies g is one-to-one function and its inverse function g^{-1} exists.
- The likelihood inference based on one dataset $(x_1, \cdots, x_n)$ should be identical to inference based on dataset $(y_1, \cdots, y_n)$. (why?)

Proportional Likelihoods (Cont.)

- Denote $g^{-1} = h$, the likelihood of $(x_1, \cdots, x_n)$

$$\begin{aligned} L(\boldsymbol{\theta}|\boldsymbol{x}) &= \prod_{i=1}^{n} f_Y(h(x_i;\boldsymbol{\theta}))h'(x_i) \\ &= \prod_{i=1}^{n} f_Y(h(x_i;\boldsymbol{\theta}))h'(g(y_i)) \\ &= \prod_{i=1}^{n} f_Y(h(x_i;\boldsymbol{\theta}))\frac{1}{g'(y_i)} \qquad (3) \\ & \quad (since \frac{dh(x)}{dx} = \frac{dg^{-1}(x)}{dx} = \frac{1}{g'(g^{-1}(x))}) \\ &= L(\boldsymbol{\theta}|\boldsymbol{y})\frac{1}{g'(y_i)}. \end{aligned}$$

- The two likelihoods are proportional as functions of $\boldsymbol{\theta}$ for all y_i,
- This implies that maximum likelihood estimates and likelihood ratio tests are identical whether derived from $L(\boldsymbol{\theta}|\boldsymbol{x})$ or $L(\boldsymbol{\theta}|\boldsymbol{y})$.

5. The Empirical Distribution Function as an MLE

- In some situation, we do not know how to model the data with a parametric distribution.
- We only know data $y_1, \cdots, y_n$ are iid from a continuous but unknown distribution whose distribution function is $F(y)$, i.e., the parameter space is the set of all distribution functions.
- Ignoring the factor $(2h)^{-m}$ in the working definition of the likelihood, an approximate likelihood for F is

  $$L_h(F|\boldsymbol{y}) = \prod_{i=1}^{n} \{F(y_i + h) - F(y_i - h)\},$$

  where h is assumed to be a small positive constant.

The Empirical Distribution Function as an MLE (Cont.)

- Assume there are no ties in the sample and h is small enough to ensure that $[Y_i + h, Y_i + h]$ does not contain Y_j for any $j \neq i$.
- Denote $p_{i,h} = F(y_i + h) - F(y_i - h)$. Then, $L_h(F|\boldsymbol{y})$ becomes $\prod_{i=1}^n p_{i,h}$.
- Since increasing $p_{i,h}$ increases $L_h(F|\boldsymbol{y})$, we want $p_{i,h}$ to be as large as possible while still satisfying $\sum_{i=1}^n p_{i,h} \leq 1$. This implies $p_{i,h} > 0$ and $\sum_{i=1}^n p_{i,h} = 1$.

The Empirical Distribution Function as an MLE (Cont.)

- To maximize the likelihood subject to $p_{i,h} > 0$ and $\sum_{i=1}^{n} p_{i,h} = 1$, we can use the method of Lagrage multipliers to find the stationary points of

$$g(p_{1,h}, \cdots, p_{n,h}, \lambda) = \sum_{i=1}^{n} \log(p_{i,h}) + \lambda(\sum_{i=1}^{n} p_{i,h} - 1).$$

- The stationary points satisfies

$$\begin{aligned} \frac{\partial g}{\partial p_{i,h}} &= \frac{1}{p_{i,h}} + \lambda = 0, i = 1, \cdots, n. \\ \frac{\partial g}{\partial \lambda} &= \sum_{i=1}^{n} p_{i,h} - 1 = 0. \end{aligned}$$

The Empirical Distribution Function as an MLE (Cont.)

- The first n equations implies $p_{i,h} = -1/\lambda$, which upon substitution into the last equation yields $\lambda = -n$. Thus, the MLE of $p_{i,h} = F(y_i + h) - F(y_i - h)$ satisfies

$$\hat{F}_h(y_i + h) - \hat{F}_h(y_i - h) = \frac{1}{n}.$$

  This means the MLE puts the equal probability mass $1/n$ on the n observed values.

- By Problem 2.10 of the textbook, $\hat{F}_h(y) \to \frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq y)$, which is called the empirical distribution function.
- Thus, we take

$$\hat{F}_{MLE}(y) = \frac{1}{n}\sum_{i=1}^{n} I(y_i \leq y).$$

  as the MLE of $F(y)$.

6. Likelihood for Type I Censoring

- Lawless (1982) gives data on pieces of equipment that are started at different times and later regularly checked for failure.
- When the study is ended, there are three of the items had not failed. For the three items, we can only know their failure time is greater than the ended time of the study but not exact failure time. Such data is called **right censored data**.
- The data in days is recorded below

Observed time	2	72	51	60	33	27	14	24	4	21
Failure (1) or Right Censored (0)	1	0	1	0	1	1	1	1	1	0

Likelihood for Type I Censoring (Cont.)

- Given a random variable X, we might observe it if $X \leq R_0$ (Right censoring) or $X \geq L_0$ (Left censoring), i.e. we observe all values of X in some specified time period. Such type of censoring is called **Type I Censoring**.
- Suppose X has density $f(x;\boldsymbol{\theta})$ with distribution function $F(x;\boldsymbol{\theta})$, and that y_i is independently observed data from

  $$Y = \begin{cases} L_0, X_i \leq L_i, \\ X_i, L_0 < X_i < R_0, \\ R_0, X_i \geq R_0. \end{cases}$$

  for $i = 1, \cdots, n$.
- The likelihood function of $y_1, y_2, \cdots, y_n$ is

  $$\{F(L_0;\boldsymbol{\theta})\}^{n_L}\{\prod_{L_0<y_i<R_0} f(y_i;\boldsymbol{\theta})\}\{1 - F(R_0;\boldsymbol{\theta})\}^{n_R}$$

  if there are n_L left censoring data and n_R right censoring data.

Likelihood for Random Censoring

- In previous situation, the censoring times L_0 and R_0 were considered fixed. In medical studies, however, patients often enter the studies at different times that are modeled as random variables.
- For illustration, we only consider the random right censoring times $R_1, \cdots, R_n$.
- Define the random observing time $Y_i = min(X_i, R_i)$ and $\Delta_i = I(X_i \leq R_i)$, and assume the censoring times are independent of $X_1, \cdots, X_n$ and are iid with distribution $G(t)$ and density $g(t)$.

Likelihood for Random Censoring

- The likelihood due to $(Y_i = y_i, \delta = 1)$ is $\frac{P(Y_i \in (y_i - h, y_i + h], \delta_i = 1)}{2h}$

$$
\begin{aligned}
&= \frac{P(Y_i \in (y_i - h, y_i + h], X_i \leq R_i)}{2h} \\
&= \frac{1}{2h} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [I(y - h < t \leq y + h, t \leq r) f(t, \boldsymbol{\theta}) g(r)]\, dtdr \\
&= \frac{1}{2h} \int_{-y_i - h}^{y_i + h} \left[\int_{-\infty}^{\infty} I(t \leq r) g(r) dr \right] f(t, \boldsymbol{\theta}) dt \\
&= \frac{1}{2h} \int_{y_i - h}^{y_i + h} (1 - G(t))\, f(t, \boldsymbol{\theta}) dt \\
&\to [1 - G(y_i)] f(y_i; \boldsymbol{\theta})
\end{aligned}
$$

as $h \to 0$. Note the last line is by the Fundamental Theorem of Calculus.

Likelihood for Random Censoring

- A analogous argument can be used to argue the likelihood contributed from data $(Y_i = y_i, \delta = 0)$ is

$$\frac{P(Y_i \in (y_i - h, y_i + h], \delta_i = 1)}{2h} \to [1 - F(y; \boldsymbol{\theta})]g(y).$$

(See pages 49-50 of the textbook and checked by yourself.)

- Put the two types of likelihood together, the likelihood for the iid data $\boldsymbol{y} = (y_1, \cdots, y_n)$ and $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_n)$ is

$$\begin{aligned} L(\boldsymbol{\delta}|\boldsymbol{y}, \boldsymbol{\delta}) &= \left\{ \prod_{i=1}^{n} f(y_i; \boldsymbol{\delta})^{\delta_i} [1 - F(y_i; \boldsymbol{\delta})]^{1-\delta_i} \right\} \\ &\quad \times \prod_{i=1}^{n} \{g(y_i)^{1-\delta_i} [1 - G(Y_i)]^{\delta_i}\} \\ &= \prod_{i=1}^{n} \left\{ f(y_i; \boldsymbol{\delta})^{\delta_i} [1 - F(y_i; \boldsymbol{\delta})]^{1-\delta_i} g(y_i)^{1-\delta_i} [1 - G(Y_i)]^{\delta_i} \right\} \end{aligned}$$

- Note that the unknown censoring distribution G is not needed to estimate $\boldsymbol{\theta}$.

Revisit Lawless (1982) data

- There are $n = 10$ data points with $n_R = 3$ right censoring data.

$\boldsymbol{y}$: Observed time	2	72	51	60	33	27	14	24	4	21
$\boldsymbol{\delta}$: Failure (1) or Right Censored (0)	1	0	1	0	1	1	1	1	1	0

- Find MLE

$$\begin{aligned} L(\sigma|\boldsymbol{y},\boldsymbol{\delta}) &= \prod_{i=1}^{n}\left[\frac{1}{\sigma}\exp(-Y_i/\sigma)\right]^{\delta_i}[\exp(-y_i/\sigma)]^{1-\delta_i} \\ &= \left(\frac{1}{\sigma}\right)^{n-n_R}\exp(-n\bar{y}/\sigma) \\ \ell(\sigma) &= \log(L(\sigma|\boldsymbol{y})) = -(n-n_R)\log\sigma - \frac{n\bar{y}}{\sigma} \\ \hat{\sigma}_{MLE} &= \left(\frac{n}{n-n_R}\right)\bar{y} = 44.0 \end{aligned}$$

Main Reference and Homework

- HW1 has been posted on ... and the due day is ...