STAT 611-600

Theory of Inference
Lecture 6: MLE

Tiandong Wang

How to get estimators: Method of Maximum Likelihood

Suppose we observe

$$x_1, \ldots, x_n,$$

which we call what we saw.

If we have to choose a model from the statistical family of models

$$(S, \mathcal{A}, \{\mathbb{P}_\theta, \theta \in \Theta\})$$

to explain the data, choose the model which maximizes the likelihood of seeing what we saw.

Example: Suppose an urn contains b black and w white balls but we don't know if

$$b = 3w \qquad \text{(3 times as many black as white)} \tag{1}$$

or

$$w = 3b \qquad \text{(3 times as many white as black).} \tag{2}$$

Experiment: Sample with replacement 3 times and let

$$X = \# \text{ of black balls drawn.}$$

So

$$X \sim b(k; n = 3, p),$$

where

$$p = 3/4 \qquad \text{or} \qquad p = 1/4.$$

There are 2 possible mass functions corresponding to (1), (2):

Outcome x	0	1	2	3
$p_{3/4}(x)$	$1/64$	$9/64$	$27/64$	$27/64$
$p_{1/4}(x)$	$27/64$	$27/64$	$9/64$	$1/64$

Experiment: Sample with replacement 3 times and let

$$X = \# \text{ of black balls drawn.}$$

So

$$X \sim b(k; n = 3, p),$$

where

$$p = 3/4 \qquad \text{or} \qquad p = 1/4.$$

There are 2 possible mass functions corresponding to (1), (2):

Outcome x	0	1	2	3
$p_{3/4}(x)$	$1/64$	$9/64$	$27/64$	$27/64$
$p_{1/4}(x)$	$27/64$	$27/64$	$9/64$	$1/64$

If we observe $X = 0$, what would be a sensible estimate of p?

A. $p = 1/4$.

B. $p = 3/4$.

C. $p = 1/100$, because that minimizes nastiness.

D. $p = 1/2$ because symmetry is good.

Outcome x	0	1	2	3
$p_{3/4}(x)$	$1/64$	$9/64$	$27/64$	$27/64$
$p_{1/4}(x)$	$27/64$	$27/64$	$9/64$	$1/64$

If $X = 0$, $p = 3/4$ is unlikely and it is natural to estimate

$$\hat{p} = \frac{1}{4}$$

since when $x = 0$,

$$p_{1/4}(0) > p_{3/4}(0).$$

Likewise

$$\hat{p}(0) = \hat{p}(1) = \frac{1}{4}, \text{ but } \hat{p}(2) = \hat{p}(3) = \frac{3}{4}.$$

So if we observe $X = x$, so that x is what we saw, we choose $\hat{p}$ to satisfy

$$p_{\hat{p}}(x) = \max\{p_{1/4}(x), p_{3/4}(x)\}.$$

General: Finding the MLE.

Given a random sample $X_1, \ldots, X_n$ from a density

$$f_{X_1,\ldots,X_n}(u_1, \ldots, u_n; \theta) = \prod_{i=1}^{n} f(u_i; \theta),$$

or from a pmf

$$p_{X_1,\ldots,X_n}(u_1, \ldots, u_n; \theta) = \prod_{i=1}^{n} p(u_i; \theta) = \mathbb{P}_\theta[X_1 = u_1, \ldots, X_n = u_n].$$

Suppose we do the experiment and observe

$$X_1 = x_1, \ldots, X_n = x_n.$$

The likelihood is the joint density considered as a function of θ:

$$L(\theta; x_1, \ldots, x_n) = \begin{cases} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n; \theta), & \text{if continuous case;} \\ p_{X_1,\ldots,X_n}(x_1, \ldots, x_n; \theta), & \text{if discrete case.} \end{cases}$$

In the discrete case, the likelihood is the probability of seeing what we saw.

Definition. The maximum likelihood estimator of θ is the value of θ given by the function

$$\hat{\theta}_{mle} = \hat{\theta}_{mle}(x_1, \ldots, x_n)$$

which maximizes the likelihood.

How to compute the mle (θ is 1-dimensional).

- With sufficient regularity, often can get $\hat{\theta}_{mle}$ as the solution of

$$\frac{\partial}{\partial\theta} L(\theta; x_1, \ldots, x_n) = 0.$$

- Since $\log(\cdot)$ is continuous and strictly increasing, $L(\theta; \boldsymbol{x})$ and $\log L(\theta; \boldsymbol{x})$ (called the *log-likelihood*) have their maxima at the same θ. It may be easier to solve

$$\frac{\partial}{\partial\theta} \log L(\theta; x_1, \ldots, x_n) = 0.$$

- The solutions to likelihood equations, $\hat{\theta}$, are called the *extreme points in the interior* of Θ.
- These extreme points can be local minima, local maxima, or inflection points, so they provide possible candidates for the MLE.
- In order to find a global maximum, we need to first find out local maxima from the extreme points, and compare their likelihood values with those points at the boundary of Θ.

Second-order Criteria:

Assume $\hat{\theta}$ is the solution to the likelihood equation. Consider the second-order derivative condition

$$\frac{d^2}{d\theta^2} L(\theta; \boldsymbol{x})\Big|_{\theta=\hat{\theta}} < 0, (*)$$

- If (*) holds, then $\hat{\theta}$ is a local maxima. If $L(\hat{\theta}; \boldsymbol{x})$ is larger than the likelihood of all the boundary points, then $\hat{\theta}$ is MLE.
- (Special Case) If (*) holds and $\hat{\theta}$ is the unique solution from the likelihood equation, then $\hat{\theta}$ is MLE.

Example: If

$$f(x;\lambda) = \lambda e^{-\lambda x} 1_{[0,\infty)}(x),$$

then based on n iid samples giving observed values $x_1, \ldots, x_n$, the likelihood is

$$L(\lambda) = \lambda^n e^{-\lambda \sum_{i=1}^n x_i} \prod_{i=1}^n 1_{[0,\infty)}(x_i),$$

and

$$\log L(\lambda) = n \log \lambda - \lambda \sum_{i=1}^n x_i, \quad x_i > 0,$$

so differentiating with respect to λ and setting the result to 0 gives

$$\frac{n}{\lambda} - \sum_i x_i = 0,$$

and the solution is

$$\hat{\lambda}_{mle} = \frac{1}{\bar{X}}.$$

(This was also the MOME.)

Note: Cannot always rely on differentiation and the MLE may not be unique (and in fact, may not exist).

Note: Cannot always rely on differentiation and the MLE may not be unique (and in fact, may not exist).
Example. Suppose

$$f(x, \theta) = 1_{[\theta - \frac{1}{2}, \theta + \frac{1}{2}]}(x), \qquad \theta > 0.$$

Then

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} 1_{[\theta - \frac{1}{2}, \theta + \frac{1}{2}]}(x_i).$$

Let $x_{(1)} = \min\{x_1, \ldots, x_n\}$ and $x_{(n)} = \max\{x_1, \ldots, x_n\}$. The biggest the likelihood can be is 1. We have

$$\begin{aligned}
L(\theta) = 1 \Leftrightarrow \quad & \text{for all } i: \ \theta - \frac{1}{2} \leq x_i \leq \theta + \frac{1}{2}, \\
\Leftrightarrow \quad & \text{for all } i: \ -\frac{1}{2} - x_i \leq -\theta \leq \frac{1}{2} - x_i \\
\Leftrightarrow \quad & \text{for all } i: \ x_i - \frac{1}{2} \leq \theta \leq x_i + \frac{1}{2}, \\
\Leftrightarrow \quad & x_{(n)} - \frac{1}{2} \leq \theta \leq x_{(1)} + \frac{1}{2}.
\end{aligned}$$

So

$$L(\theta) = 1_{[x_{(n)} - \frac{1}{2}, x_{(1)} + \frac{1}{2}]}(\theta).$$

This is maximized by any θ in the interval

$$[x_{(n)} - \frac{1}{2}, x_{(1)} + \frac{1}{2}].$$

So the following are all mle's:

- $\hat{\theta} = x_{(n)} - \frac{1}{2}$.
- $\hat{\theta} = x_{(1)} + \frac{1}{2}$.
- $\hat{\theta} = \frac{1}{2}(x_{(1)} + x_{(n)})$=mid-range.
- etc

MLE: Two-parameter case $\boldsymbol{\theta} = (\theta_1, \theta_2)$

- Simultaneous maximization: solve

$$\max_{\theta_1, \theta_2} L(\theta_1, \theta_2; \boldsymbol{x})$$

At the extreme point $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$, we need to check the negative definiteness of the Hessian matrix

$$H = \begin{pmatrix} \frac{\partial^2}{\partial \theta_1^2} L(\theta; \boldsymbol{x}) & \frac{\partial^2}{\partial \theta_1 \theta_2} L(\theta; \boldsymbol{x}) \\ \frac{\partial^2}{\partial \theta_2 \theta_1} L(\theta; \boldsymbol{x}) & \frac{\partial^2}{\partial \theta_2^2} L(\theta; \boldsymbol{x}) \end{pmatrix}$$

- Two-stage maximization (profile method)

$$\max_{\theta_1} \max_{\theta_2} L(\theta_1, \theta_2; \boldsymbol{x})$$

Examples:

1. $N(\mu, \sigma^2)$ and $\boldsymbol{\theta} = (\mu, \sigma^2)$.
2. Location-scale exponential family, with pdf

$$f(x; \mu, \beta) = \frac{1}{\beta} e^{-(x-\mu)/\beta} \quad \text{if} \quad x \geq \mu.$$

See board for the blood and guts.

Further properties of the mle.

1. **Invariance.** If

$$\hat{\theta}_{mle} = \hat{\theta}_{mle}(X_1, \ldots, X_n)$$

is the maximum likelihood estimator of θ, then the mle of $h(\theta)$ is $h(\hat{\theta}_{mle})$. This may seem obvious but think about the definitions; it isn't obvious.

Examples.

- The mle of σ is

$$\sqrt{\hat{\sigma}^2_{mle}}.$$

So for instance, for $N(\mu, \sigma^2)$,

$$\hat{\sigma}_{mle} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})}.$$

- For repeated Bernoulli trials, $B_1, \ldots, B_n$, we have the mass function

B_1	0	1
$P_p[B_1 = x]$	q	p

  The mle of p is

$$\hat{p}_{mle} = \bar{B} = \frac{1}{n}\sum_{i=1}^{n} B_i.$$

  The mle of the variance $p(1-p)$ is

$$\bar{B}(1-\bar{B}).$$

- If $X_1, \ldots, X_n$ are a random sample from $N(\mu, \sigma^2)$, then the mle of

$$\mathbb{E}_{\mu,\sigma^2}(X_1^2) = \mu^2 + \sigma^2$$

  is

$$\hat{\mu}_{mle}^2 + \sigma_{mle}^2.$$

2. Typically the mle is CAN=Consistent Asymptotically Normal.
   - Consistent: For larger and larger samples sizes, the estimator is more and more accurate in the sense that for all $\theta \in \Theta$

$$\lim_{n\to\infty} \mathbb{P}_\theta[|\hat{\theta}_{mle}(X_1,\ldots,X_n) - \theta| \le \delta] = 1,$$

     for any small $\delta > 0$.
   - Asymptotically normal: For large n, $\hat{\theta}_{mle}$ is approximately normally distributed with mean θ, and some variance $\sigma^2(\theta)$. For all θ:

$$\lim_{n\to\infty} \mathbb{P}_\theta[\sqrt{n}\big(\hat{\theta}_{mle}(X_1,\ldots,X_n) - \theta\big) \le x] = N\big(x; 0, \sigma^2(\theta)\big),$$

     for any $x \in \mathbb{R}$.
3. Typically the MLE has minimal asymptotic variance $\sigma^2(\theta)$. Small variance is good.

Doing maximum likelihood with R (Optional)

There are several R-packages for MLE fitting. A simple one written by Brian Ripley (one of the R watchdogs) is in a package called `MASS` containing a routine called

```
fitdistr
```

that fits univariate densities using maximum likelihood. It was intended for instructional purposes and is easy to use.

Doing maximum likelihood with R (Optional)

There are several R-packages for MLE fitting. A simple one written by Brian Ripley (one of the R watchdogs) is in a package called `MASS` containing a routine called

`fitdistr`

that fits univariate densities using maximum likelihood. It was intended for instructional purposes and is easy to use.

Syntax:

```
fitdistr(x,densfun, start)
```

where

- `densfun`=beta, cauchy, chi-squared, gamma, exponential, normal, etc.
- `start`= named list giving starting values for parameters in the numerical optimization. Can be skipped but better results achieved with sensible (eg, mome) starting values.

Ouput = object of class `fitdistr` with attributes *estimate, sd, and loglik*.

Other options:

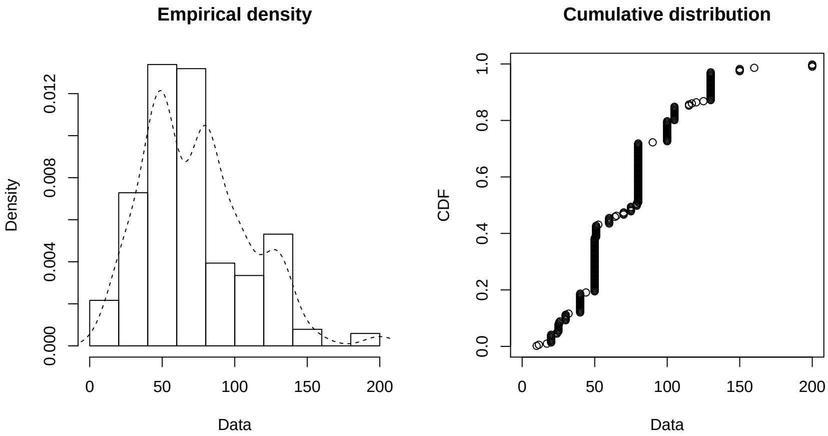
- `fitdist` in package `fitdistrplus`
  This is a steroidal version of `fitdistr`; it is not quite as simple but allows you to pick the estimation method. For example, you could start with *mme* to get mome's and then use the momes with "start" in the MLE estimation.

  Syntax

  ```
  fitdist(data, distr, method = c("mle", "mme", "qme", "mge"
      start=NULL, fix.arg=NULL, discrete, keepdata = TRUE, k
  ```
- `optim` function to optimize the log-likelihood numerically.

Example: `groundbeef` data in package `fitdistrplus`.

Plot the empirical distribution using `plotdist`:



Then we try to fit three different distributions to data:

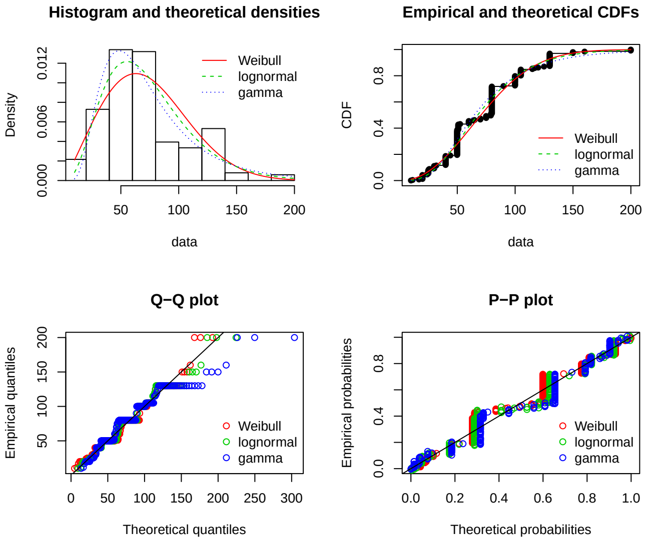
- Weibull, gamma and log-normal.

Parameters estimated using MLE:

```
> fit_w  <- fitdist(serving, "weibull")
> fit_g  <- fitdist(serving, "gamma")
> fit_ln <- fitdist(serving, "lnorm")
> rbind(fit_w$estimate, fit_g$estimate, fit_ln$estimate)
        shape       scale
[1,] 2.185885 83.34767905
[2,] 4.008253  0.05441911
[3,] 4.169370  0.53660951
```

Take a look at the summary:

```
> summary(fit_ln)
Fitting of the distribution ' lnorm ' by maximum likelihood
Parameters :
         estimate Std. Error
meanlog 4.1693701 0.03366988
sdlog   0.5366095 0.02380783
Loglikelihood:  -1261.319   AIC:  2526.639   BIC:  2533.713
Correlation matrix:
```


Histogram and theoretical densities
Density
data
Weibull
lognormal
gamma
Empirical and theoretical CDFs
CDF
data
Weibull
lognormal
gamma
Q−Q plot
Empirical quantiles
Theoretical quantiles
Weibull
lognormal
gamma
P−P plot
Empirical probabilities
Theoretical probabilities
Weibull
lognormal
gamma