Math 362: Mathematical Statistics II

Le Chen
le.chen@emory.edu

Emory University
Atlanta, GA

Last updated on December 20, 2020

2021 Spring

Chapter 5: Estimation

Chapter 5. Estimation

§ 5.1 Introduction

§ 5.2 Estimating parameters: MLE and MME

§ 5.1 Introduction

Motivating example: Given an unfair coin, or p-coin, such that

$$X = \begin{cases} 1 & \text{head with probability } p, \\ 0 & \text{tail with probability } 1 - p, \end{cases}$$

how would you determine the value p?

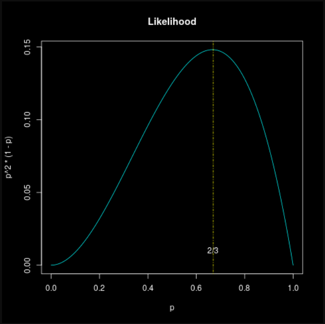
Solutions:

1. You need to try the coin several times, say, three times. What you obtain is "HHT".
2. Draw a conclusion from the experiment you just made.

Rationale: The choice of the parameter p should be the value that maximizes the probability of the sample.

$$\mathbb{P}(X_1 = 1, X_2 = 1, X_3 = 0) = P(X_1 = 1)P(X_2 = 1)P(X_3 = 0)$$
$$= p^2(1 - p).$$

```
# Hello, R.
p <- seq(0,1,0.01)
plot(p,p^2*(1-p),
      type=”l”,
      col=”red”)
title(”Likelihood”)
# add a vertical dotted (4) blue line
abline(v=0.67, col=”blue”, lty=4)
# add some text
text(0.67,0.01, ”2/3”)
```



Maximize $f(p) = p^2(1 - p)$ ....

A random sample of size n from the population – Bernoulli(p):

- $X_1, \cdots, X_n$ are i.i.d.[1] random variables, each following Bernoulli(p).
- Suppose the outcomes of the random sample are: $X_1 = k_1, \cdots, X_n = k_n$.
- What is your choice of p based on the above random sample?

$$p = \frac{1}{n} \sum_{i=1}^{n} k_i =: \bar{k}.$$

[1] independent and identically distributed

A random sample of size n from the population with given pdf:

- $X_1, \cdots, X_n$ are i.i.d. random variables, each following the same given pdf.

- a **statistic** or an **estimator** is a function of the random sample.

  Statistic/Estimator is a random variable!

  e.g.,

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

- The outcome of a statistic/estimator is called an **estimate**. e.g.,

$$p_e = \frac{1}{n}\sum_{i=1}^{n} k_i.$$

Chapter 5. Estimation

§ 5.1 Introduction

§ 5.2 Estimating parameters: MLE and MME

§ 5.2 Estimating parameters: MLE and MME

Two methods:	Corresponding estimator
1. Method of maximum likelihood.	MLE
2. Method of moments.	MME

Maximum Likelihood Estimation

Definition 5.2.1. For a random sample of size n from the discrete (resp. continuous) population/pdf $p_X(k;\theta)$ (resp. $f_Y(y;\theta)$), the **likelihood function**, $L(\theta)$, is the product of the pdf evaluated at $X_i = k_i$ (resp. $Y_i = y_i$), i.e.,

$$L(\theta) = \prod_{i=1}^{n} p_X(k_i;\theta) \qquad \left(\text{resp. } L(\theta) = \prod_{i=1}^{n} f_Y(y_i;\theta)\right).$$

Definition 5.2.2. Let $L(\theta)$ be as defined in Definition 5.2.1. If θ_e is a value of the parameter such that $L(\theta_e) \geq L(\theta)$ for all possible values of θ, then we call θ_e the **maximum likelihood estimate** for θ.

Examples for MLE

Often but not always MLE can be obtained by setting the first derivative equal to zero:

E.g. 1. Poisson distribution: $p_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}$, $k = 0, 1, \cdots$.

$$L(\lambda) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{k_i}}{k_i!} = e^{-n\lambda}\lambda^{\sum_{i=1}^{k} k_i}\left(\prod_{i=1}^{n} k_i!\right)^{-1}.$$

$$\ln L(\lambda) = -n\lambda + \left(\sum_{i=1}^{n} k_i\right)\ln\lambda - \ln\left(\prod_{i=1}^{n} k_i!\right).$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\ln L(\lambda) = -n + \frac{1}{\lambda}\sum_{i=1}^{n} k_i.$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\ln L(\lambda) = 0 \quad\Longrightarrow\quad \boxed{\lambda_e = \frac{1}{n}\sum_{i=1}^{n} k_i =: \bar{k}}.$$

Comment: The critical point is indeed global maximum because

$$\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2}\ln L(\lambda) = -\frac{1}{\lambda^2}\sum_{i=1}^{n} k_i < 0.$$

The following two cases are related to waiting time:

E.g. 2. Exponential distribution: $f_Y(y) = \lambda e^{-\lambda y}$ for $y \geq 0$.

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda y_i} = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} y_i\right)$$

$$\ln L(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^{n} y_i.$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln L(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} y_i.$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln L(\lambda) = 0 \quad \Longrightarrow \quad \boxed{\lambda_e = \frac{n}{\sum_{i=1}^{n} y_i} =: \frac{1}{\bar{y}}}.$$

A random sample of size n from the following population:

E.g. 3. Gamma distribution: $f_Y(y;\lambda) = \frac{\lambda^r}{\Gamma(r)} y^{r-1} e^{-\lambda y}$ for $y \geq 0$ with $r > 1$ known.

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^r}{\Gamma(r)} y_i^{r-1} e^{-\lambda y_i} = \lambda^{r\,n} \Gamma(r)^{-n} \left(\prod_{i=1}^{n} y_i^{r-1} \right) \exp\left(-\lambda \sum_{i=1}^{n} y_i \right)$$

$$\ln L(\lambda) = r\,n \ln\lambda - n \ln\Gamma(r) + \ln\left(\prod_{i=1}^{n} y_i^{r-1} \right) - \lambda \sum_{i=1}^{n} y_i.$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln L(\lambda) = \frac{r\,n}{\lambda} - \sum_{i=1}^{n} y_i.$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln L(\lambda) = 0 \quad \Longrightarrow \quad \boxed{\lambda_e = \frac{r\,n}{\sum_{i=1}^{n} y_i} = \frac{r}{\bar{y}}}.$$

Comment:

- When $r = 1$, this reduces to the exponential distribution case.
- If r is also unknown, it will be much more complicated.
  No closed-form solution. One needs numerical solver[2].
  Try MME instead.

[2] [DW, Example 7.2.25]

A detailed study with data:

E.g. 4. Geometric distribution: $p_X(k;p) = (1-p)^{k-1}p$, $k = 1, 2, \cdots$.

$$L(p) = \prod_{i=1}^{n} (1-p)^{k_i - 1} p = (1-p)^{-n+\sum_{i=1}^{k} k_i} p^n.$$

$$\ln L(p) = \left(-n + \sum_{i=1}^{n} k_i\right) \ln(1-p) + n \ln p.$$

$$\frac{\mathrm{d}}{\mathrm{d}p} \ln L(p) = -\frac{-n + \sum_{i=1}^{n} k_i}{1-p} + \frac{n}{p}.$$

$$\frac{\mathrm{d}}{\mathrm{d}p} \ln L(p) = 0 \quad \Longrightarrow \quad \boxed{p_e = \frac{n}{\sum_{i=1}^{n} k_i} = \frac{1}{\bar{k}}}.$$

Comment: Its cousin distribution, the negative binomial distribution can be worked out similarly (See Ex 5.2.14).

MLE for Geom. Distr. of sample size n = 128

k	Observed frequency	Predicted frequency
1	72	74.14
2	35	31.2
3	11	13.13
4	6	5.52
5	2	2.32
6	2	0.98

Real p = unknown and MLE for p = 0.5792

```
# The example from the book.
library(pracma) # Load the library ”Practical Numerical Math Functions”
k<−c(72, 35, 11, 6, 2, 2) # observed freq.
a=1:6
pe=sum(k)/dot(k,a) # MLE for p.
f=a
for (i in 1:6) {
  f[i] = round((1−pe)ˆ(i−1) ∗ pe ∗ sum(k),2)
}
# Initialize the table
d <−matrix(1:18, nrow = 6, ncol = 3)
# Now adding the column names
colnames(d) <− c(”k”,
                 ”Observed freq.”,
                 ”Predicted freq.”)
d[1:6,1]<−a
d[1:6,2]<−k
d[1:6,3]<−f
grid.table(d) # Show the table
PlotResults(”unknown”, pe, d, ”Geometric.pdf”) # Output the results using a user
       defined function
```

MLE for Geom. Distr. of sample size n = 128

k	Observed frequency	Predicted frequency
1	42	40.96
2	31	27.85
3	15	18.94
4	11	12.88
5	9	8.76
6	5	5.96
7	7	4.05
8	2	2.75
9	1	1.87
10	2	1.27
11	1	0.87
13	1	0.59
14	1	0.4

Real p = 0.3333 and MLE for p = 0.32

```
# Now let’s generate random samples from a Geometric distribution with p=1/3 with the same size of the sample.
p = 1/3
n = 128
gdata<-rgeom(n, p)+1 # Generate random samples
g<- table(gdata) # Count frequency of your data.
g<- t(rbind(as.numeric(rownames(g)), g)) # Transpose and combine two columns.
pe=n/dot(g[,1],g[,2]) # MLE for p.
f <- g[,1] # Initialize f
for (i in 1:nrow(g)) {
  f[i] = round((1-pe)^(i-1) * pe * n,2)
} # Compute the expected frequency
g<-cbind(g,f) # Add one columns to your matrix.
colnames(g) <- c(”k”,
                 ”Observed freq.”,
                 ”Predicted freq.”) # Specify the column names.
d_df <- as.data.frame(d) # One can use data frame to store data
d_df # Show data on your terminal
PlotResults(p, pe, g, ”Geometric2.pdf”) # Output the results using a user defined function
```

MLE for Geom. Distr. of sample size n = 300

k	Observed frequency	Predicted frequency
1	99	105.88
2	69	68.51
3	47	44.33
4	28	28.69
5	27	18.56
6	9	12.01
7	8	7.77
8	5	5.03
9	5	3.25
10	3	2.11

Real p = 0.3333 and MLE for p = 0.3529

In case we have several parameters:

E.g. 5. Normal distribution: $f_Y(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$, $y \in \mathbb{R}$.

$$L(\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}} = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i-\mu)^2 \right)$$

$$\ln L(\mu,\sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2.$$

$$\begin{cases} \dfrac{\partial}{\partial\mu} \ln L(\mu,\sigma^2) = \dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}(y_i-\mu) \\ \dfrac{\partial}{\partial\sigma^2} \ln L(\mu,\sigma^2) = -\dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4}\displaystyle\sum_{i=1}^{n}(y_i-\mu)^2 \end{cases}$$

$$\begin{cases} \dfrac{\partial}{\partial\mu} \ln L(\mu,\sigma^2) = 0 \\ \dfrac{\partial}{\partial\sigma^2} \ln L(\mu,\sigma^2) = 0 \end{cases} \implies \boxed{\begin{cases} \mu_e = \bar{y} \\ \sigma_e^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(y_i-\bar{y})^2 \end{cases}}$$

In case when the parameters determine the support of the density: (Non regular case)

E.g. 6. Uniform distribution on $[a, b]$ with $a < b$: $f_Y(y; a, b) = \frac{1}{b-a}$ if $y \in [a, b]$.

$$L(a, b) = \begin{cases} \prod_{i=1}^{n} \frac{1}{b-a} = \frac{1}{(b-a)^n} & \text{if } a \le y_1, \cdots, y_n \le b, \\ 0 & \text{otherwise.} \end{cases}$$

$L(a, b)$ is monotone increasing in a and decreasing in b. Hence, in order to maximize $L(a, b)$, one needs to choose

$$a_e = y_{min} \quad \text{and} \quad b_e = y_{max}.$$

E.g. 7. $f_Y(y; \theta) = \frac{2y}{\theta^2}$ for $y \in [0, \theta]$.

$$L(\theta) = \begin{cases} \prod_{i=1}^{n} \frac{2y_i}{\theta^2} = 2^n \theta^{-2n} \prod_{i=1}^{n} y_i & \text{if } 0 \le y_1, \cdots, y_n \le \theta, \\ 0 & \text{otherwise.} \end{cases}$$

$$\Downarrow$$

$$\theta_e = y_{max}.$$

In case of discrete parameter:

E.g. 8. **Wildlife sampling.** Capture-tag-recapture.... In the history, a tags have been put. In order to estimate the population size N, one randomly captures n animals, and there are k tagged. Find the MLE for N.

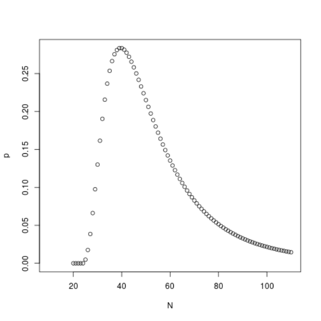
Sol. The population follows hypergeometric distr.:

$$p_X(k;N) = \frac{\binom{a}{k}\binom{N-a}{n-k}}{\binom{N}{n}}.$$

$$L(N) = \frac{\binom{a}{k}\binom{N-a}{n-k}}{\binom{N}{n}}$$

How to maximize $L(N)$?

```
> a=10
> k=5
> n=20
> N=seq(a,a+100)
> p=choose(a,k)*choose(N−a,n−k)/choose(N,n)
> plot(N,p,type = ”p”)
> print(paste(”The MLE is”, n*a/k))
[1] ”The MLE is 40”
```



The graph suggests to sudty the following quantity:

$$r(N) := \frac{L(N)}{L(N-1)} = \frac{N-n}{N} \times \frac{N-a}{N-a-n+k}$$

$$r(N) < 1 \quad \Longleftrightarrow \quad na < Nk \quad \text{i.e., } N > \frac{na}{k}$$

$$\boxed{N_e = \arg\max\left\{L(N) : N = \left\lfloor \frac{na}{k} \right\rfloor, \left\lceil \frac{na}{k} \right\rceil\right\}}.$$

□

Method of Moments Estimation

Rationale: The population moments should be close to the sample moments, i.e.,

$$E(Y^k) \approx \frac{1}{n}\sum_{i=1}^{n} y_i^k, \quad k = 1, 2, 3, \cdots .$$

Definition 5.2.3. For a random sample of size n from the discrete (resp. continuous) population/pdf $p_X(k;\theta_1,\cdots,\theta_s)$ (resp. $f_Y(y;\theta_1,\cdots,\theta_s)$), solutions to

$$\begin{cases} \mathbb{E}(Y) = \frac{1}{n}\sum_{i=1}^{n} y_i \\ \qquad \vdots \\ \mathbb{E}(Y^s) = \frac{1}{n}\sum_{i=1}^{n} y_i^s \end{cases}$$

which are denoted by $\theta_{1e},\cdots,\theta_{se}$, are called the **method of moments estimates** of $\theta_1,\cdots,\theta_s$.

Examples for MME

MME is often the same as MLE:

E.g. 1. Normal distribution: $f_Y(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$, $y \in \mathbb{R}$.

$$\begin{cases} \mu = \mathbb{E}(Y) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} y_i = \bar{y} \\ \sigma^2 + \mu^2 = \mathbb{E}(Y^2) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} y_i^2 \end{cases} \quad \Rightarrow \quad \begin{cases} \mu_e = \bar{y} \\ \sigma_e^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} y_i^2 - \mu_e^2 \\ \quad = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \end{cases}$$

More examples when MLE coincides with MME: Poisson, Exponential, Geometric.

MME is often much more tractable than MLE:

E.g. 2. Gamma distribution[3]: $f_Y(y; r, \lambda) = \frac{\lambda^r}{\Gamma(r)} y^{r-1} e^{-\lambda y}$ for $y \geq 0$.

$$\begin{cases} \dfrac{r}{\lambda} = \mathbb{E}(Y) = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} y_i = \bar{y} \\ \dfrac{r}{\lambda^2} + \dfrac{r^2}{\lambda^2} = \mathbb{E}(Y^2) = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} y_i^2 \end{cases} \quad \Rightarrow \quad \begin{cases} r_e = \dfrac{\bar{y}^2}{\hat{\sigma}^2} \\ \lambda_e = \dfrac{\bar{y}}{\hat{\sigma}^2} = \dfrac{r_e}{\bar{y}} \end{cases}$$

where $\bar{y}$ is the sample mean and $\hat{\sigma}^2$ is the sample variance: $\hat{\sigma}^2 := \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2$.

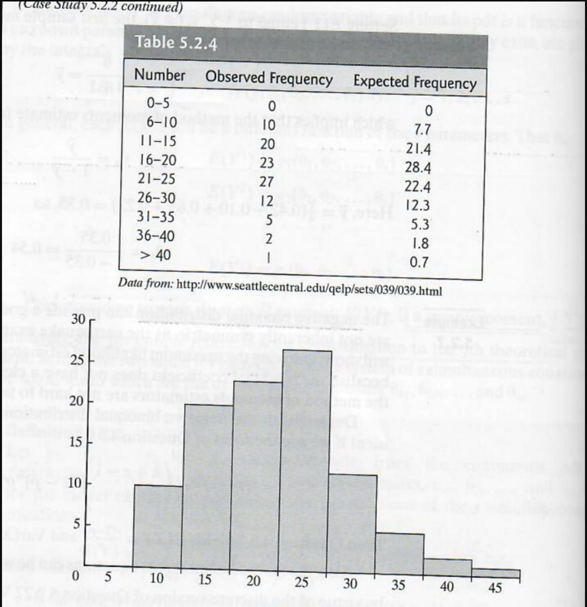
Comments: MME for λ is consistent with MLE when r is known.

[3] Check Theorem 4.6.3 on p. 269 for mean and variance

Another tractable example for MME, while less tractable for MLE:

E.g. 3. Neg. binomial distribution: $p_X(k;p,r) = \binom{k+r-1}{k}(1-p)^k p^r$, $k = 0, 1, \cdots$.

$$\begin{cases} \dfrac{r(1-p)}{p} = \mathbb{E}(X) = \bar{k} \\ \dfrac{r(1-p)}{p^2} = \mathrm{Var}(X) = \hat{\sigma}^2 \end{cases} \quad \Rightarrow \quad \begin{cases} p_e = \dfrac{\bar{k}}{\hat{\sigma}^2} \\ \\ r_e = \dfrac{\bar{k}^2}{\hat{\sigma}^2 - \bar{k}} \end{cases}$$



(Case Study 5.2.2 continued)

Table 5.2.4

Number	Observed Frequency	Expected Frequency
0–5	0	0
6–10	10	7.7
11–15	20	21.4
16–20	23	28.4
21–25	27	22.4
26–30	12	12.3
31–35	5	5.3
36–40	2	1.8
> 40	1	0.7

Data from: http://www.seattlecentral.edu/qelp/sets/039/039.html

$r_e = 12.74$ and $p_e = 0.391$.

E.g. 4. $f_Y(y;\theta) = \frac{2y}{\theta^2}$ for $y \in [0,\theta]$.

$$\overline{y} = \mathbb{E}[Y] = \int_0^\theta \frac{2y^2}{\theta^2}\mathrm{d}y = \frac{2}{3}\frac{y^3}{\theta^2}\bigg|_{y=0}^{y=\theta} = \frac{2}{3}\theta.$$

$$\Downarrow$$

$$\boxed{\theta_e = \frac{3}{2}\overline{y}.}$$