class08

January 29, 2021 10:27 AM

Outline

§7.2 Methods of point estimation

Maximum likelihood estimation

Example 7.16

A sample of ten new bike helmets manufactured by a company is obtained. Upon testing, it is found that the first, third, and tenth helmets are flawed, whereas the others are not.

Let $p = P(\text{flawed helmet})$ and define $X_1, \ldots, X_{10}$ by $X_i = 1$ if the ith helmet is flawed and zero otherwise. Then the observed x_i's are $1, 0, 1, 0, 0, 0, 0, 0, 0, 1$, so the joint pmf of the sample is

$$f(x_1, x_2, \ldots, x_{10}; p) = p(1-p)p \cdots p = p^3(1-p)^7 \quad \textbf{(7.4)}$$

$$p^3(1-p)^7 = L(p)$$

Pmf of $\underline{X}$

$$f(x;p) = p^x(1-p)^{1-x}, \quad x = 0, 1.$$

$$= \begin{cases} p, & x = 1 \\ 1-p, & x = 0 \end{cases}$$

$x_1, x_2, \ldots, x_n$

$$f(x_1, x_2, \ldots, x_n; p) = f(x_1;p)f(x_2;p)\cdots f(x_n;p)$$

$$= \prod_{i=1}^{n} f(x_i;p)$$

$$= \prod_{i=1}^{n} \left[p^{x_i}(1-p)^{1-x_i} \right]$$

Two observations:

- For a given value of p, the joint pmf is the "probability of data".
- Regarding $x_1, \ldots, x_n$ as known numbers, the joint pmf is a function of parameter p, called the **likelihood function**.

"For what value of p is the observed sample most likely to have occurred?"

What value of p is the most compatible with the observed sample?

We wish to find the value of p for which the probability of the observed sample is the largest.

That is, we seek the value of p that maximizes (7.4).

Or, equivalently, find the value of p that maximizes the natural log of (7.4):

$$\ln[f(x_1, x_2, \ldots, x_{10}; p)] = 3\ln(p) + 7\ln(1-p) \qquad (7.5)$$

"ln" = "$\log_e$"

$\ln\left(p^3(1-p)^7\right)$

$= 3\ln p + 7\ln(1-p)$

"log" = "$\log_e$" in stats.

or ~~"$\log_{10}$"~~

Equating the derivative of (7.5) to zero gives the maximizing value:

$$\frac{d}{dp}\ln[f(x_1, x_2, \ldots, x_{10}; p)] = \frac{3}{p} - \frac{7}{1-p} = 0 \Rightarrow p = \frac{3}{10} = \frac{x}{n}$$

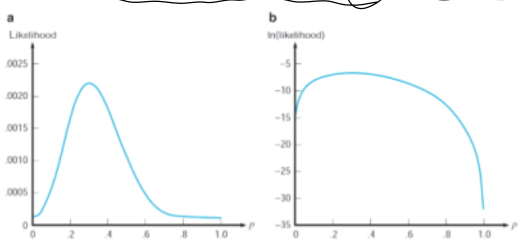


Figure 7.5 Likelihood and log likelihood plotted against p ■

$$\frac{d\ln p}{dp} = \frac{1}{p}$$

$$\frac{d\ln(1-p)}{dp} = \frac{1}{1-p}\cdot\frac{d(1-p)}{dp} = -\frac{1}{1-p}$$

$$\frac{3}{p} = \frac{7}{1-p} \qquad 3(1-p) = 7p$$

$$3 - 3p = 7p \Rightarrow p = \frac{3}{10}$$

Equating the derivative of (7.5) to zero gives the maximizing value:

$$\frac{d}{dp}\ln[f(x_1, x_2, \ldots, x_{10}; p)] = \frac{3}{p} - \frac{7}{1-p} = 0 \Rightarrow p = \frac{3}{10} = \frac{x}{n}$$

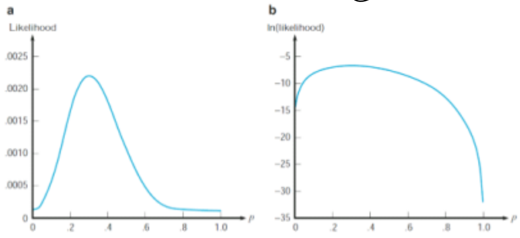


Figure 7.5 Likelihood and log likelihood plotted against p ■

$$\frac{d\ln(1-p)}{dp} = \frac{1}{1-p} \cdot \frac{d(1-p)}{dp} = -\frac{1}{1-p}$$

$$\frac{3}{p} = \frac{7}{1-p} \qquad 3(1-p) = 7p$$

$$3 - 3p = 7p \Rightarrow p = \frac{3}{10}$$

7

The estimate of p is now $\hat{p} = \frac{3}{10}$. It is called the **maximum likelihood estimate** because for fixed $x_1, \ldots, x_{10}$, it is the parameter value that maximizes the likelihood (joint pmf) of the observed sample, the parameter value that agrees most with the data.

$\hat{p} = \frac{X}{n}$ as a r v

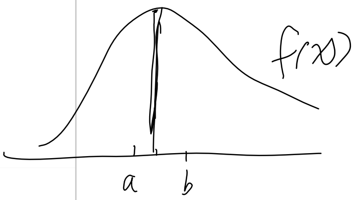
maximum likelihood estimator.

8

Note that if we had been told only that among the ten helmets there were three that were flawed, Equation (7.4) would be replaced by the binomial pmf

$$\binom{10}{3} p^3 (1-p)^7,$$

which is also maximized for $\hat{p} = \frac{3}{10}$.



$P(a < X < b) = \int_a^b f(x)\,dx$

If $a \approx b$, $P(a < X < b) \doteq f(x)(b-a)$

9

Let $X_1, \dots, X_n$ have joint pmf or pdf

$$f(x_1, \dots, x_n; \theta_1, \dots, \theta_m) \tag{7.6}$$

where the parameters $\theta_1, \dots, \theta_m$ have unknown values. When $x_1, \dots, x_n$ are the observed sample values and (7.6) is regarded as a function of $\theta_1, \dots, \theta_m$, it is called the **likelihood function**.

The **maximum likelihood estimates** $\hat{\theta}_1, \dots, \hat{\theta}_m$ are those values of the θ_i's that maximize the likelihood function, so that

$$f(x_1, \dots, x_n; \hat{\theta}_1, \dots, \hat{\theta}_m) \geq f(x_1, \dots, x_n; \theta_1, \dots, \theta_m)$$

for all $\theta_1, \dots, \theta_m$.

When the X_i's are substituted in place of the x_i's, the **maximum likelihood estimators** (mle's) result.

10

pdf for mass or density of X (discrete or continuous)

Example (Poisson).
Let X follow a Poisson distribution with parameter k (Will write $X \sim \text{Poisson}(k)$).

Pop'n dist'n: pdf of X

1) The pmf of X is

$$P(X = x) = f(x; k) = \frac{k^x e^{-k}}{x!}, x = 0, 1, 2, \dots$$

2) Suppose we have an observed sample $x_1, \dots, x_n$. The likelihood function is

$x_1, \cdots, x_n$ are regarded as known values.

Uppercase "L"

$$L(k) \equiv P(X_1 = x_1, \dots, X_n = x_n; k)$$
$$= P(X_1 = x_1; k) \dots P(X_n = x_n; k)$$
$$= \prod_{i=1}^{n} f(x_i; k) = \prod_{i=1}^{n} \left[\frac{k^{x_i} e^{-k}}{x_i!}\right].$$

function of K

11

$$l(k) = \ln\left(\prod_{i=1}^{n}\left[\frac{k^{x_i}}{x_i!} e^{-k}\right]\right)$$
$$= \sum_{i=1}^{n} \ln\left[\frac{k^{x_i}}{x_i!} e^{-k}\right]$$
$$= \sum_{i=1}^{n}\left[x_i \ln k - k - \ln(x_i!)\right]$$

3) It is equivalent to maximize the logarithm of the likelihood function

lowercase "l"

$$l(k) \equiv \ln(L(k)) = -nk + \sum_{x_i} \ln(k) - \ln(\prod_{i=1}^{n} x_i!)$$

Set the derivative equal to 0, and solve for k:

$$\frac{dl(k)}{dk} = -n + (\sum_{i=1}^{n} x_i)/k = 0 \Rightarrow \hat{k} = (\sum_{i=1}^{n} x_i)/n = \bar{x}.$$

Therefore, $\bar{x}$ is the maximum likelihood estimate of k and $\bar{X}$ is the maximum likelihood estimator.

12

$$\frac{dl(k)}{dk} = \sum_{i=1}^{n}\left[x_i \cdot \frac{1}{k} - 1 - 0\right]$$
$$= \frac{\sum x_i}{k} - n = 0 \Rightarrow \frac{\sum x_i}{k} = n$$
$$\Rightarrow \hat{k} = \frac{\sum x_i}{n} = \bar{x}$$

$\hat{k} = \frac{\sum X_i}{n} = \bar{X}$ is the ML estimator. (MLE)

Example (Normal).
Let X is normal with mean μ and variance σ^2 (write "$X \sim N(\mu, \sigma^2)$" hereafter).

1) The density function for X is

$$f(x; \mu, \sigma) = f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

2) Given a random sample $x_1, \dots, x_n$, the likelihood function is

$$L(\mu, \sigma) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right]$$

$$l(\mu, \sigma) = \sum_{i=1}^{n} \ln[f(x_i)]$$
$$= \sum_{i=1}^{n}\left[-\ln(2\pi) - \ln\sigma - \frac{(x_i - \mu)^2}{2\sigma^2}\right]$$
$$= \sum_{i=1}^{n} \frac{(x_i - \mu)}{\sigma^2} = 0 \Rightarrow \sum x_i - n\mu = 0$$

3) The logarithm of the likelihood function is

$$l(\mu, \sigma) = -n\ln(\sqrt{2\pi}) - n\ln(\sigma) - \sum^{n}(x_i - \mu)^2/(2\sigma^2)$$

3) The logarithm of the likelihood function is

$$l(\mu, \sigma) = -n \ln(\sqrt{2\pi}) - n \ln(\sigma) - \sum_{i=1}^{n} (x_i - \mu)^2/(2\sigma^2)$$

13

To maximize this function, we take the partial derivatives wrt μ and σ, set them equal to 0 and solve the equation for μ and σ:

$$\begin{cases} \frac{\partial l(\mu,\sigma)}{\partial \mu} = -\sum_{i=1}^{n} 2(x_i - \mu)(-1)/(2\sigma^2) = 0 \\ \frac{\partial l(\mu,\sigma)}{\partial \sigma} = -\frac{n}{\sigma} - (-2)\sum_{i=1}^{n} (x_i - \mu)^2/(2\sigma^3) = 0 \end{cases}$$

$$\begin{cases} \sum_{i=1}^{n} (x_i - \mu) = 0 \\ -\frac{n}{\sigma} + \sum_{i=1}^{n} (x_i - \mu)^2/\sigma^3 = 0 \end{cases}$$

$$\begin{cases} \sum_{i=1}^{n} x_i - n\mu = 0 \\ -n\sigma^2 + \sum_{i=1}^{n} (x_i - \mu)^2 = 0 \end{cases}$$

$$\begin{cases} \hat{\mu} = \bar{x} \\ \hat{\sigma}^2 = \sum_{i=1}^{n} (x_i - \mu)^2/n \end{cases}$$

The maximum likelihood estimators for μ and σ^2 are $\hat{\mu} = \bar{X}$ and $\hat{\sigma}^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2/n$.

14

The method of moments and the method of maximum likelihood gives the same estimator for σ^2: $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2$, which is biased.

Rigorously, we need check the second-order derivatives to confirm that the likelihoods are maximized in the above examples. *Not a saddle point, log likelihood not convex.*

15

$$= \sum_{i=1} [-\ln(2\pi) - \ln \sigma \quad 2\sigma^2]$$

$$= \sum_{i=1}^{n} \frac{(x_i - \mu)}{\sigma^2} = 0 \Rightarrow \sum x_i - n\mu = 0$$

$$\begin{cases} \frac{\partial l(\mu,\sigma)}{\partial \mu} = \sum_{i=1}^{n} \left[- \frac{2(x_i - \mu)(-1)}{2\sigma^2} \right] = 0 \quad ① \\ \Rightarrow \mu = \bar{x} \\ \frac{d(x_i - \mu)^2}{d\mu} = 2(x_i - \mu)\frac{d(x_i - \mu)}{d\mu} = 2(x_i - \mu)(-1) \\ \frac{\partial l(\mu,\sigma)}{\partial \sigma} = \sum_{i=1}^{n} \left(0 - \frac{1}{\sigma} - \frac{(x_i - \mu)^2}{2}(-2)\sigma^{-3} \right) = 0 \quad ② \end{cases}$$

$$\frac{d(1/\sigma^2)}{d\sigma} = \frac{d(\sigma^{-2})}{d\sigma} = (-2)\sigma^{-3}$$

Plugging $\mu = \bar{x}$ in ②

$$\sum \left[-\frac{1}{\sigma} - \frac{(x_i - \bar{x})^2}{2}(-2)\sigma^{-3} \right] = 0$$

$$\Rightarrow \sigma^2 = \frac{1}{n}(x_i - \bar{x})^2$$

$$\Rightarrow \hat{\mu} = \bar{x}, \quad \hat{\sigma}^2 = \frac{1}{n}(x_i - \bar{x})^2 = \frac{n-1}{n}S^2$$

the same as the method of moments results.