IIT ROORKEE

FREE ONLINE EDUCATION swayam

NPTEL ONLINE CERTIFICATION COURSE

Maximum Likelihood Estimation - I

Dr. A. Ramesh

DEPARTMENT OF MANAGEMENT STUDIES

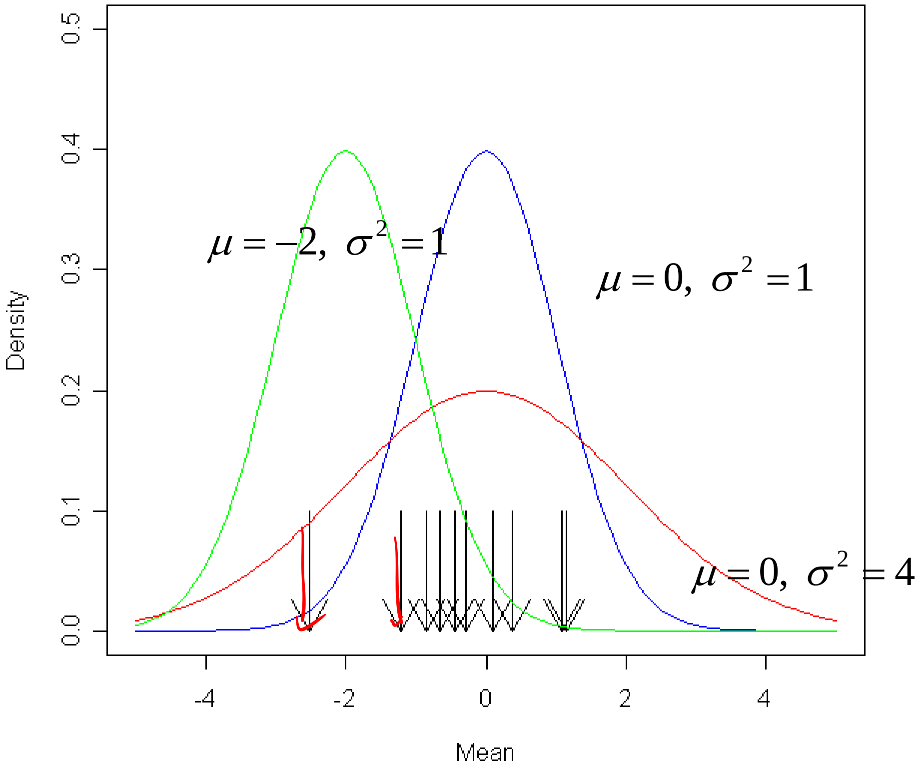
Agenda

- This lecture will provide intuition behind the MLE using Theory and examples.

Maximum Likelihood Estimation

- The method of maximum likelihood was first introduced by R. A. Fisher, a geneticist and statistician, in the 1920s.
- Most statisticians recommend this method, at least when the sample size is large, since the resulting estimators have certain desirable efficiency properties
- Maximum likelihood estimation(MLE) is a method to find most likely density function, that would have generated data.
- MLE requires one to make distribution assumption first.

An intuitive view on likelihood



Maximum Likelihood Estimation: Problem

- A sample of ten new bike helmets manufactured by a certain 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(flawed helmet), i.e., p is the proportion of all such helmets that are flawed.
- Define (Bernoulli) random variables $X_1, X_2, \ldots, X_{10}$ by

$$X_1 = \begin{cases} 1 \text{ if 1st helmet is flawed} \\ 0 \text{ if 1st helmet isn't flawed} \end{cases} \quad \cdots \quad X_{10} = \begin{cases} 1 \text{ if 10th helmet is flawed} \\ 0 \text{ if 10th helmet isn't flawed} \end{cases}$$

Source: Probability and Statistics for Engineering and the Sciences, Jay L Devore, 8th Ed, Cengage

Maximum Likelihood Estimation: Problem

- Then for the obtained sample, $X_1 = X_3 = X_{10} = 1$ and the other seven X_i's are all zero
- The probability mass function of any particular X_i is $p^{x_i}(1-p)^{1-x_i}$, which becomes p if $x_i = 1$ and $1 - p$ when $x_i = 0$
- Now suppose that the conditions of various helmets are independent of one another
- This implies that the X_i's are independent, so their joint probability mass function is the product of the individual pmf's.

Maximum Likelihood Estimation: Binomial Distribution

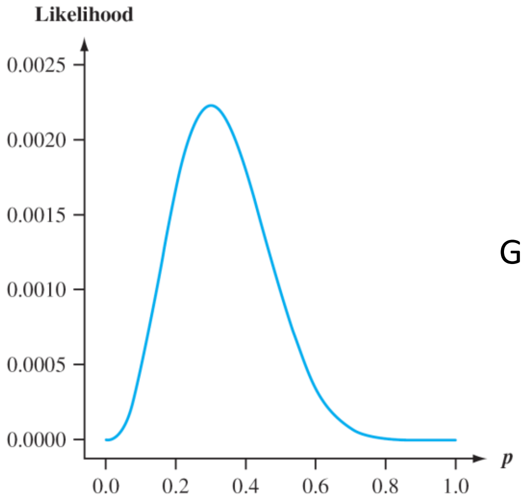
- Joint pmf evaluated at the observed X_i's is

  $f(x_1, . . . , x_{10}; p) = p(1 – p)p \cdots p = p^3(1 – p)^7$ - (1)

- Suppose that $p = .25$. Then the probability of observing the sample that we actually obtained is $(.25)^3(.75)^7 = .002086$.
- If instead $p = .50$, then this probability is $(.50)^3(.50)^7 = .000977$.
- For what value of p is the obtained sample most likely to have occurred?
- That is, for what value of p is the joint pmf (eq 1) as large as it can be?
- What value of p maximizes (eq 1)

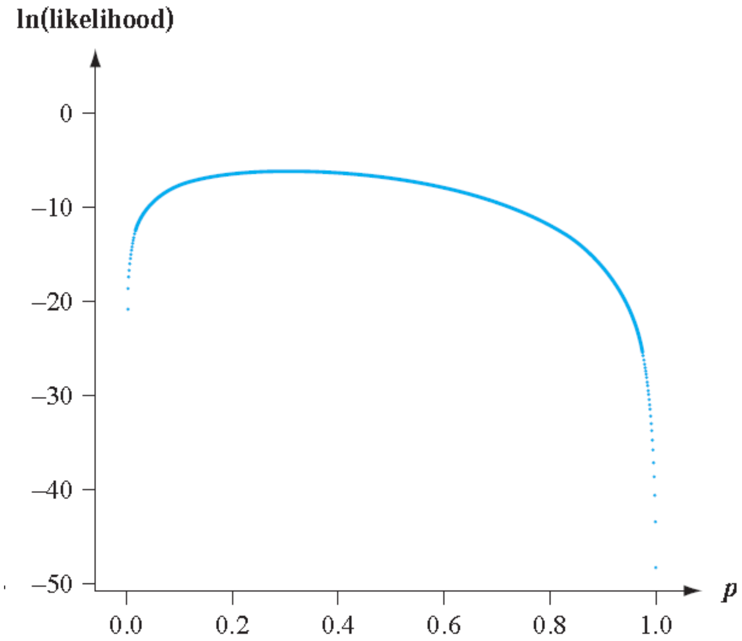
Maximum Likelihood Estimation: Binomial Distribution

- Figure shows a graph of the *likelihood* (eq 1) as a function of p.
- It appears that the graph reaches its peak above p = .3 = the proportion of flawed helmets in the sample.



Graph of the likelihood (joint pmf) (eq 1)

Graph of the natural logarithm of the likelihood



- Figure shows a graph of the natural logarithm of (eq 1)
- Since *ln*[g(u)] is a strictly increasing function of $g(u)$, finding u to maximize the function $g(u)$ is the same as finding u to maximize *ln*[g(u)].

Maximum Likelihood Estimation: Binomial Distribution

- We can verify our visual impression by using calculus to find the value of p that maximizes (eq 1).
- Working with the natural log of the joint pmf is often easier than working with the joint pmf itself, since the joint pmf is typically a product so its logarithm will be a sum.
- Here $\ln[f(x_1, \dots, x_{10}; p)] = \ln[p^3(1-p)^7]$
- $= 3\ln(p) + 7\ln(1-p)$

Maximum Likelihood Estimation: Binomial Distribution

Thus $$\frac{d}{dp}\{\ln[f(x_1, \ldots, x_{10}; p)]\} = \frac{d}{dp}\{3\ln(p) + 7\ln(1 - p)\}$$

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

$$= \frac{3}{p} - \frac{7}{1 - p}$$

Interpretation

- Equating this derivative to 0 and solving for p gives
  $3(1 - p) = 7p$, from which $3 = 10p$ and so $p = 3/10 = .30$ as conjectured

- That is, our point estimate is $p = .30$.
- It is called the *maximum likelihood estimate* because it is the parameter value that maximizes the likelihood (joint pmf) of the observed sample
- In general, the second derivative should be examined to make sure a maximum has been obtained, but here this is obvious from Figure

Maximum Likelihood Estimation: Binomial Distribution

- Suppose that rather than being told the condition of every helmet, we had only been informed that three of the ten were flawed.
- Then we would have the observed value of a binomial random variable X = the number of flawed helmets.
- The pmf of X is $\binom{10}{x} p^x (1 - p)^{10-x}$. For x = 3, this becomes $\binom{10}{3} p^3 (1 - p)^7$.
- The binomial coefficient $\binom{10}{3}$ is irrelevant to the maximization, so again p = 0.30.

Maximum Likelihood Function Definition

- Let $X_1, X_2,\ldots, X_n$ have joint pmf or pdf

$$f(x_1, x_2, \ldots, x_n; \theta_1, \ldots, \theta_m) \quad \text{(a)}$$

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

- The maximum likelihood estimates (mle's) $\hat{\theta}_1,\ldots,\hat{\theta}_m$ are those values of the *i*'s that maximize the likelihood function, so that

$$f(x_1, x_2, \ldots, x_n; \hat{\theta}_1,\ldots,\hat{\theta}_m) \geq f(x_1, x_2, \ldots, x_n; \theta_1, \ldots, \theta_m) \text{ for all } \theta_1, \ldots, \theta_m$$

- When the $X_i's$ are substituted in place of the $x_i's$, the **maximum likelihood estimators** result.

Interpretation

- The likelihood function tells us how likely the observed sample is as a function of the possible parameter values.
- Maximizing the likelihood gives the parameter values for which the observed sample is most likely to have been generated—that is, the parameter values that “agree most closely” with the observed data.

Estimation of Poisson Parameter

- Suppose we have data generated from a Poisson distribution. We want to estimate the parameter of the distribution
- The probability of observing a particular random variable is $P(X;\mu)=\dfrac{e^{-\mu}\mu^{X}}{X!}$
- Joint likelihood by multiplying the individual probabilities together

$$P(X_1, X_2, \ldots, X_n;\mu)=\frac{e^{-\mu}\mu^{X_1}}{X_1!}\times\frac{e^{-\mu}\mu^{X_2}}{X_2!}\times\ldots\times\frac{e^{-\mu}\mu^{X_n}}{X_n!}$$

$$L(\mu;\mathbf{X})=\prod_i e^{-\mu}\mu^{X_i}$$

$$L(\mu;\mathbf{X})=e^{-n\mu}\mu^{n\overline{X}}$$

Estimation of Poisson Parameter

- Note in the likelihood function the factorials have disappeared.
- This is because they provide a constant that does not influence the relative likelihood of different values of the parameter
- It is usual to work with the **log likelihood** rather than the likelihood.
- Note that maximising the log likelihood is equivalent to maximising the likelihood.

$$L(\mu;\mathbf{X}) = e^{-n\mu}\mu^{n\bar{X}}$$

Take the natural log of the likelihood function

$$\ell(\mu;\mathbf{X}) = -n\mu + n\bar{X}\log\mu$$

Find where the derivative of the log likelihood is zero

$$\frac{d\ell}{d\mu} = -n + \frac{n\bar{X}}{\mu}$$

$$\hat{\mu} = \bar{X}$$

Note that here the MLE is the same as the moment estimator

Estimation of exponential distribution Parameter

- Suppose $X_1, X_2, \ldots, X_n$ is a random sample from an exponential distribution with parameter λ. Because of independence, the likelihood function is a product of the individual pdf's:

$$f(x_1, \ldots, x_n; \lambda) = (\lambda e^{-\lambda x_1}) \cdot \cdots \cdot (\lambda e^{-\lambda x_n})$$
$$= \lambda^n e^{-\lambda \Sigma x_i}$$

- The natural logarithm of the likelihood function is

$$\ln[f(x_1, \ldots, x_n; \lambda)] = n \ln(\lambda) - \lambda \Sigma x_i$$

Estimation of exponential distribution Parameter

- Equating $(d/d\lambda)[\ln(\text{likelihood})]$ to zero results in $n/\lambda - \Sigma x_i = 0$, or $\lambda = n/\Sigma x_i = 1/\bar{x}$.

- Thus the MLE is $\hat{\lambda} = 1/\bar{X}$;

Estimation of parameters of Normal Distribution

- Let $X_1, \ldots, X_n$ be a random sample from a normal distribution.
- The likelihood function is

$$f(x_1, \ldots, x_n; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_1-\mu)^2/(2\sigma^2)} \cdot \cdots \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_n-\mu)^2/(2\sigma^2)}$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\Sigma(x_i-\mu)^2/(2\sigma^2)}$$

- so

$$\ln[f(x_1, \ldots, x_n; \mu, \sigma^2)] = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum (x_i - \mu)^2$$

Estimation of parameters of normal distribution

- To find the maximizing values of μ and σ^2, we must take the partial derivatives of $\ln(f)$ with respect to μ and σ^2, equate them to zero, and solve the resulting two equations.

- Omitting the details, the resulting MLE's are

$$\hat{\mu} = \overline{X} \qquad \hat{\sigma}^2 = \frac{\sum (X_i - \overline{X})^2}{n}$$

- The MLE of σ^2 is not the unbiased estimator, so two different principles of estimation (unbiasedness and maximum likelihood) yield two different estimators

Thank you