W1211 Introduction to Statistics
Lecture 20

Wei Wang

Nov 14, 2012

Methods of Point Estimation

- The definition of unbiasedness does not in general indicate how unbiased estimators can be derived.

- There are two commonly used “constructive” methods for obtaining point estimators: the method of moments and the method of maximum likelihood.

- Although maximum likelihood estimators are generally preferable to moment estimators because of certain efficiency properties, they often require significantly more computation than do moment estimators.

- It is NOT guaranteed that these two methods would yield unbiased estimators.

Population Moment and Sample Moment

- Let $X_1, \ldots, X_n$ be a random sample from a pmf or pdf $f(x)$. For $k = 1, 2, \ldots$, the kth population moment is $E(X^k)$. The kth sample moment is $(1/n)\sum_{i=1}^{n} X_i^k$.
- The essence of the Methods of Moment is to equate population moments with sample moments and solve the resulting equations.

Moment Estimators

- Definition:

Let X_1, X_2, …, X_n be an i.i.d. sample from a pmf or pdf $f(x)$. For $k = 1, 2, 3, \ldots$, the moment estimator for the kth population moment, is the kth sample moment, i.e.,

$$\widehat{\mathrm{E}(\mathrm{X}^k)} = \frac{\sum_{i=1}^{n} \mathrm{X}_i^k}{n}$$

Example

Ex. Show that the sample proportion is the moment estimator of the population probability.

Example

Ex. Let $X_1, X_2, \ldots, X_n$ be an i.i.d. normal sample, and assume that the underlying normal distribution is $N(\mu,\sigma^2)$ where μ,σ^2 are unknown. How can we construct moment estimators to estimate the two unknown parameters?

As we already know if $X \sim N(\mu,\sigma^2)$, then $E(X) = \mu$, and $E(X^2) = \mu^2+\sigma^2$.

Therefore, we have two equations:

$$\begin{cases} \hat{\mu} = \sum_{i=1}^{n} X_i/n \\ \hat{\mu}^2 + \hat{\sigma}^2 = \sum_{i=1}^{n} X_i^2/n \end{cases} \longrightarrow \begin{cases} \hat{\mu} = \sum_{i=1}^{n} X_i/n \\ \hat{\sigma}^2 = \sum_{i=1}^{n} X_i^2/n - \bar{X}^2 \end{cases}$$

Is the variance estimator unbiased?

Example

Ex. Let X_1, X_2, …, X_n be an i.i.d. sample from exponential distribution with parameter λ which is unknown. How do we estimate λ using moment estimator?

As we already know if $X \sim \mathrm{Exp}(\lambda)$, then $E(X) = 1/\lambda$.

Thus, we have equation $1/\hat{\lambda} = \bar{\mathrm{X}} \rightarrow \hat{\lambda} = 1/\bar{\mathrm{X}}$.

Is this estimator unbiased?

Maximum Likelihood Est.

- The method of maximum likelihood was first introduced by R.A. Fisher, a geneticist and statistician, in the 1920s. It is by far the most commonly used method to obtain estimators.

- Likelihood function is just another way of looking at the *joint pmf or the pdf*. In particular, let $X_1, X_2, \ldots, X_n$ (not necessarily i.i.d.) have joint pmf or pdf

$$f(x_1, x_2, \ldots, x_n; \theta_1, \ldots, \theta_m)$$

  where $\theta_1, \ldots, \theta_m$ are parameters whose values are unknown. When $x_1, x_2, \ldots, x_n$ are the observed sample values and $f(.)$ is then regarded as a function of $\theta_1, \ldots, \theta_m$, it is called the likelihood function.

Example

Ex. A biased coin has been flipped for 10 times. Let X_1, X_2, …, X_{10} denote the outcomes of the coin flips. Assume the probability of having a head is p (parameter of interest), and the sample we observed is {0,1,1,0,0,0,1,0,0,0}. Write down the likelihood function for p.

$$f(x_1, x_2, \ldots, x_n; p) = f(x_1; p)\, f(x_2; p) \ldots f(x_n; p) = (1-p)\, p\, p\, (1-p) \ldots (1-p) = p^3(1-p)^7$$

Idea of Maximum Likelihood: can we find a p that can maximize the above function?

MLE

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

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

  when the X_i's are substituted in place of the x_i's.

- Remark: 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.

- In practice, in stead of maximizing the likelihood itself, people usually choose to maximize the log-likelihood function.

Example

Ex. Let X_1, X_2, …, X_n be an i.i.d. sample from exponential distribution with parameter λ which is unknown. Write down the likelihood function for λ. What is the MLE of λ? Is the MLE unbiased?

Since we have an i.i.d. sample, it is easy to see that the likelihood function is a product of the individual pdf's:

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

$$\Downarrow$$

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

$$\Downarrow$$

$$\hat{\lambda} = n / \sum \mathrm{X}_i$$

Example with Normal

- Let $X_1, X_2, \ldots, X_n$ be an IID sample from normal distribution with mean μ and variance σ^2, what is the likelihood function?

Example with Normal

- Let $X_1, X_2, \ldots, X_n$ be an IID sample from normal distribution with mean μ and variance σ^2, what is the likelihood function?
- $$f(x_1, x_2, \ldots, x_n; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

  or in logarithm

  $$-\frac{n}{2}\log(2\pi\sigma^2) + \sum_{i=1}^{n}[-(x_i - \mu)^2/\sigma^2]$$

Example with Normal

- Let $X_1, X_2, \ldots, X_n$ be an IID sample from normal distribution with mean μ and variance σ^2, what is the likelihood function?
- $$f(x_1, x_2, \ldots, x_n; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

  or in logarithm

  $$-\frac{n}{2}\log(2\pi\sigma^2) + \sum_{i=1}^{n}[-(x_i - \mu)^2/\sigma^2]$$

- Take derivative with respect to μ and σ^2 and solve the resulting equations

  $$\hat{\mu} = \bar{X}, \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$$

Some complications

- The following is an example that MLE's can't be calculated analytically.

Ex. Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from Weibull distribution with parameters α and β and pdf

$$f(x; \alpha, \beta) = \begin{cases} \frac{\alpha}{\beta^\alpha} \cdot x^{\alpha-1} \cdot e^{-(x/\beta)^\alpha} & x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

by solving equations $$\frac{\partial \log(f)}{\partial \alpha} = 0 \quad \frac{\partial \log(f)}{\partial \beta} = 0$$

$$\downarrow$$

$$\hat{\alpha} = \left[\frac{\sum x_i^{\hat{\alpha}} \cdot \log(x_i)}{\sum x_i^{\hat{\alpha}}} - \frac{\sum \log(x_i)}{n} \right]^{-1} \quad \hat{\beta} = \left(\frac{\sum x_i^{\hat{\alpha}}}{n} \right)^{1/\hat{\alpha}}$$

Some Complications

- Also, sometimes we cannot use calculus to get the MLE, such as when the density is not differentiable.
- Read Example 6.22 on textbook P.262.

The Invariance Principle

- One of the nice features of MLE's is that, the MLE of a function of parameters, is the function of the MLE's of the parameters.

- More specifically, we have

> Let $\hat{\theta}_1, \ldots, \hat{\theta}_m$ be the MLE's of the parameters $\theta_1, \ldots, \theta_m$. Then the MLE of any function $h(\theta_1, \ldots, \theta_m)$ of these parameters is $h(\hat{\theta}_1, \ldots, \hat{\theta}_m)$.

<u>Ex.</u> In the normal example, what is the MLE of σ?

Large Sample Behavior

- The following proposition says, for large samples, it is "optimal" to use MLE's, because it is asymptotically unbiased and has the minimal variance among all unbiased estimators.

- Proposition:

  > Under very general conditions on the joint distribution of the sample,
  > When the sample size n is large, the maximum likelihood estimator is
  > Approximately the MVUE of the parameter.

Confidence Intervals

- A point estimate, because it is a single number, by itself provides no information about the precision and reliability of estimation (the reason why we need standard error).

- An alternative to reporting a single sensible value for the parameter being estimated is to calculate and report an entire interval of plausible values – an *interval estimate* or *confidence interval* (*CI*).

- A confidence interval is always calculated by first selecting a *confidence level*, which is a measure of the degree of reliability of the interval.

- Construct a confidence interval for a standard normal random variable.

Illustration

- Let's first consider a simple, somewhat unrealistic problem situation.
  1. We are interested in the population mean parameter μ.
  2. The population distribution is normal.
  3. The value of the population standard deviation σ is known. (unlikely!)

- Suppose we have a random sample $X_1, X_2, \ldots, X_n$ from a normal distribution with mean value μ and standard deviation σ. As we know, $\bar{X}$ also follows a normal distribution with mean value μ and standard deviation $\sigma/\sqrt{n}$. Thus, we could get a standard normal distribution by normalizing $\bar{X}$.

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Construction

- The smallest interval that contains 95% of the possible outcomes of Z is (-1.96, 1.96).

$$-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96$$

$$\Downarrow$$

$$-1.96 \cdot \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

$$\Downarrow$$

$$\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

Interpretation

- Thus we have $\mathrm{P}\left(\bar{\mathrm{X}} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{\mathrm{X}} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) = 0.95$.
- Some people interpreted this as: the true parameter *μ* has 95% chance of falling in the interval of $(\bar{\mathrm{X}} - 1.96 \cdot \sigma/\sqrt{n}, \bar{\mathrm{X}} + 1.96 \cdot \sigma/\sqrt{n})$. Is it right?
- In fact, the two boundaries of the interval given above are random! Thus every time we sample n observations from the same population, we will get a different confidence interval!

Random Interval

- By constructing a confidence interval like this, we never be sure whether *μ* actually lies in our confidence interval. However, we know that about 95 out of 100 times intervals constructed using this method will capture the true parameter.

- Interpreted as: "*the probability is .95 that the random interval includes or covers the true value of μ*."