Chapter 3 Point Estimate

BY YUEJIAN MO

May 21, 2018

Here two method to estimate parameters of pdf/pmf of r.v.: maximum likelihood estimation, method of moments and Bayesian estimation.

1 Maximum Likehood Estimator 最大似然估计

1.1 Point estimator and point estimate

Definition 3.1 (A statistic 统计量). A function of one or more r.v's that doest not depend on the unknown paramemter vector is called statistic.

对于统计量，一旦样本确定，统计量的值也就定下来，但其分布往往包含未知参数。Following show the mean of similar words.

- Estimation 估计(方法)
- Estimator 估计量
- Estimate 估计值
- population 母体， $\boldsymbol{X(r,v)}$
- sample 组 $x_1, x_2, ..., x_n$(iid)

1.2 Joint density and likelihood function

In formal contexts, "likelihood" is often used as a synonym "probalilty". In mathematical statistics, the two terms have different meaning. *Probability* in this technical context describes the plausibility of a future outcome, given a model parameter value, without reference to any observed data. *Likehood* describes the plausibility of a model parameter value, given specific observed data.

Since $\boldsymbol{x}$ has been observed and its components are therefore fixed real numbers, we regard $f(\boldsymbol{x};\boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$, and define

$$L(\boldsymbol{\theta}) = L(\boldsymbol{\theta};\boldsymbol{x}) = f(\boldsymbol{x};\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i;\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta},$$

as the *likehood function* of the random smaple $\boldsymbol{x}$. It also can be called: $L(\theta)$ is the likehood function of $\boldsymbol{\theta}$.

For avoid the operation of $\prod$, we has *log-likehood*

$$l(\boldsymbol{\theta}) \hat{=} \log\{L(\boldsymbol{\theta})\} = \sum_{i=1}^{n} \log\{f(x_i;\boldsymbol{\theta})\} \,\text{for}\, \boldsymbol{\theta} \in \boldsymbol{\Theta}$$

There is no loss of information in using $l(\boldsymbol{\theta})$ instead of $L(\theta)$ because log(.) is a monotonic increasing function.

1.3 Maximum likelihood estimator and maximum likelihood estimate

To get reasonable $\boldsymbol{\theta}$, we suppose that a statistic

$$\hat{\boldsymbol{\theta}} = \begin{pmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_n \end{pmatrix} = \begin{pmatrix} u_1(\boldsymbol{x}) \\ \vdots \\ u_n(\boldsymbol{x}) \end{pmatrix} \hat{=} \boldsymbol{u}(\boldsymbol{x})$$

statisfies

$$L(\hat{\boldsymbol{\theta}}) = \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\boldsymbol{\theta}).$$

We call $\hat{\boldsymbol{\theta}} = u(\boldsymbol{x})$ the *maximum likehood estimator*(MLE) of $\boldsymbol{\theta}$ and call $u(\boldsymbol{x})$ a *maximum likehood estimate estimate*(mle) of $\boldsymbol{\theta}$. There is no guarantee that the MLE exists or if does whether it is unique.

1.4 The invariance property of MLE

Theorem 3.1: (Invariance of MLE). Let $\hat{\boldsymbol{\theta}} = u(X_1, ..., X_N)$ be the MLE of $\theta_{p\times 1} \in \Theta$. If $\boldsymbol{\eta}_{p\times 1} = (h_1(\boldsymbol{\theta}), ..., h_p(\boldsymbol{\theta}))^T$ is a one-to-one transformation between $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$, then $\hat{\boldsymbol{\eta}} = h(\hat{\boldsymbol{\theta}})$ is the MLE of $\boldsymbol{\eta}$.

Theorem 3.2 (Extension of Theorem 3.1): Let $\hat{\boldsymbol{\theta}}$ be the MLE of $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)^T \in \Theta$. If $\boldsymbol{\eta}_{r\times 1} = h(\boldsymbol{\theta}) = (h_1(\theta), ..., h_r(\theta))^T$ for $1 \leqslant r \leqslant p$ is a many-to-few transformation between $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$, then $\hat{\boldsymbol{\eta}} = h(\hat{\boldsymbol{\theta}}) = (h_1(\hat{\theta}), ..., h_r(\hat{\theta}))^T$ is the MLE of $\boldsymbol{\eta}$.

This property of invariance of MLEs allows us in our discussin of maximum likelihood estimation to consider estimating $(\theta_1, ..., \theta_p)^T$ rather than the more general $h_1(\theta_1, ..., \theta_p), ..., h_r(\theta_1, ..., \theta_p)$.

2 Moment Estimator

Method of moments is proposed by the great Brith statistician Karl Pearson near the turn of the twentieth century. If H_0 is rejected, one way is to guess another population distribution. Alternatively, we can estimate the first and second moments of the unknown population distribution $F(\cdot)$ by using the *method of moments*.

The mothod of moments can be applied to both *parametric* and *nonparametric* statistics.

3 Beysian Estimator

4 Properties of Estimators

4.1 Unbiasedness

Definition 3.2 (Unbiased estimator and bias). An estimator $\varphi(x)$ is an *unbiased setimotor* of the parameter θ if $E\{\varphi(\boldsymbol{x})\} = \theta$ for $\theta \in \Theta$. Otherwise, the estimator is biased and the bias is defined by

$$b(\theta) = E\{\varphi(\boldsymbol{x})\} - \theta$$

where $\boldsymbol{x} = (X_1, ..., X_n)^T$.

Definition 3.3 (MSE). Given an estimator $Y = \varphi(\boldsymbol{x})$ of θ, the *mean square error* (MSE) of the estimator if defined by

$$\mathrm{MSE} = E\{\varphi(\boldsymbol{x}) - \theta\}^2 = \mathrm{Var}\{\varphi(\boldsymbol{x})\} + b^2(\theta)$$

If the an estimator $\varphi(\boldsymbol{x})$ is unbiased, then $\mathrm{MSE} = \mathrm{Var}(\varphi(\boldsymbol{x}))$.

4.2 Efficiency

Maybe two estimator share same bias for the same unknow parameter, so we notion of efficiency to choose the unbiased estimator with the *smaller* variance.

Definition 3.4 (Relative efficiency). Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two unbiased estimators for a parameter θ. If

$$\mathrm{Var}(\hat{\theta}_1) < \mathrm{Var}(\hat{\theta}_2),$$

we say that $\hat{\theta}_1$ is *more efficent* than $\hat{\theta}_2$. The *relative efficiency* of $\hat{\theta}_1$ to $\hat{\theta}_2$ is defined by the ratio

$$\frac{\mathrm{Eff}_{\hat{\theta}_1}}{\mathrm{Eff}_{\hat{\theta}_2}} = \frac{\mathrm{Var}(\hat{\theta}_2)}{\mathrm{Var}(\hat{\theta}_1)}$$

Let $u = \{\hat{\theta} : E(\hat{\theta}) = \theta\}$ denote the family of unbiased estimators of θ. We try to find the $\hat{\theta}^* \in u$ with the smallest variance. Here are found by Cramer and Rao, if we could find a constant c_0 satisfying

$$\mathrm{Var}(\hat{\theta}) \geqslant c_0, \forall \hat{\theta} \in u,$$

thus, the $\hat{\theta}^*$ is equivalent to finding the lower bound c_0.

Theorem 3.3 (The general CR inequality). Let $\tau(\theta)$ be an arbitrary function of the unknown θ. If (i) $\theta = T(\boldsymbol{x})$ is an unbiased estimator of $\tau(\theta)$, and (ii) the support of the population density $f(x;\theta)$ does not depend on the parameter θ, then

$$\mathrm{Var}(\hat{\theta}) \geqslant \frac{\{\tau'(\theta)\}^2}{I_n(\theta)},$$

where $I_n(\theta)$ is the Fisher information. The right hand side is called the *Cramer–Rao lower bound*.

Theorem 3.4 (Alternative expression). Let $I_n(\theta)$ denote the information, If $E\{S(\theta)\} = 0$, then

$$I_n(\theta) = E\left\{-\frac{d^2 \log \mathrm{L}(\theta;\boldsymbol{x})}{d\theta^2}\right\} = n I(\theta),$$

where

$$I(\theta) = E\left[\left\{\frac{d \log f(X;\theta)}{d\theta}\right\}^2\right] = E\left\{-\frac{d^2 \log f(X;\theta)}{d\theta^2}\right\}$$

denote the Fisher inoformation for a single sample.

Definition 3.5 (UMVUE). An estimator θ^* is called a UMVUE of θ if it is unbiased and has the smallest variance among all unbiased estimators.

Definition 3.6 (Efficient estimator). If an unbiased estimator $\theta = T(\boldsymbol{x})$ for $\tau(\theta)$ has variance equal to the Cramer-Rao lower bound, then θ is called an *efficent estimator* for $\tau(\theta)$.

Chi-square distribution

Notation: $X - \chi^2(n)$

4.3 Sufficiency

Definition 3.7(Sufficent statistc). A statistic $T(x)$ is said to be a sufficient statistic of θ if the conditional distribution of $\boldsymbol{x}$, given $T(\boldsymbol{x}) = t$, does not depend on θ for any value of t. In discrete case, this mean that

$$\Pr\{X_1 = x_1, ..., X_n = x_n; \theta | T(\boldsymbol{x}) = t\} = h(\boldsymbol{x})$$

Thm 3.5 (Factorization theorem) A statistic $T(\boldsymbol{x})$ is a sufficient statistc of the unknow parameter θ iff the joint pdf(or pmd) can be written in the form

$$f(x_1, ..., x_n; \theta) = f(\boldsymbol{x}; \theta) = g(T(\boldsymbol{x}); \theta) \times h(\boldsymbol{x}),$$

Defination 3.8 (Joint sufficient statistics). Let $X_1, ..., X_n \sim \text{iid}\, f(x; \boldsymbol{\theta})$. The statistics $T_1(\boldsymbol{x}), ..., T_r(\boldsymbol{x})$ are said to be jointly sufficient if the conditional distribution of $\boldsymbol{x}$, given

4.4 Completeness

Defniton 3.9 (Completeneness). Let $X_1, ..., X_n$ denote a random sample from the pdf (or pmf) $f(x; \theta)$ with parameter space and let

Theorem 3.7 (Lehamann-Scheffe Theorem). Let $T(\boldsymbol{x})$ is a complete sufficient statistic for θ. If $g(T)$ is an unbiased estimator of $\tau(\theta)$, then $g(T)$ is the unique UMVUE for $\tau(\theta)$.

Fisher 信息量

5 Reference

- https://en.wikipedia.org/wiki/Likelihood_function
- https://www.zhihu.com/question/33567579