Mathematical Statistics

Anna Janicka

Lecture IV, 15.03.2021

POINT ESTIMATION

Plan for today

1. Estimation
2. Sample characteristics as estimators
3. Estimation techniques
   - method of moments
   - method of quantiles
   - maximum likelihood method

Point Estimation

- The choice, on the base of the data, of *the best* parameter θ, from the set of parameters which may describe P_θ
- An **Estimator** of parameter θ is any statistic $T = T(X_1, X_2, \ldots, X_n)$ with values in Θ (we interpret it as an approximation of θ). Usually denoted by $\hat{\theta}$
- Sometimes we estimate $g(\theta)$ rather than θ.

Estimation: an example
Empirical frequency

Quality control example:

0 1 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0

0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1

- Model: $\mathcal{X} = \{0,1, 2, ..., n\}$ (here n=50),

$$P_\theta(X = x) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \quad \text{for } \theta \in [0,1]$$

- parameter θ : probability of faulty element
- an obvious estimator: $\hat{\theta} = X/n = 6/50$

n – sample size

X – number of faulty elements in sample

For a different model (all outcomes) the estimator is an average

Problems with (frequency) estimators...

Example: three genotypes in a population, with frequencies $\theta^2 : 2\theta(1-\theta) : (1-\theta)^2$

In a population of size *n*, N_1 and N_2 and N_3 individuals of particular genotypes were observed. Which estimator should we use?

(1) $\hat{\theta} = \sqrt{N_1 / n}$?

(2) $\hat{\theta} = 1 - \sqrt{N_3 / n}$?

(3) $\hat{\theta} = N_1 / n + \frac{1}{2} N_2 / n$?

Maybe something else?

→ How do we choose the best one?

Estimation – sample characteristics

Sample characteristics:

estimators based on the empirical distribution (empirical CDF)

Empirical CDF

- Let $X_1, X_2, \ldots, X_n$ be a sample from a distribution given by F (modeled by $\{P_F\}$)

  (n-th) **empirical CDF**

$$\hat{F}_n(t) = \tfrac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,t]}(X_i) = \frac{\text{number of observations } X_i : X_i \le t}{n}$$

- For a given realization $\{X_i\}$ it is a function of t, the CDF of the empirical distribution (uniform over $x_1, x_2, \ldots, x_n$). For a given t it is a statistic with a distribution

$$P(\hat{F}(t) = \tfrac{k}{n}) = \binom{n}{k} F(t)^k (1 - F(t))^{n-k}, \quad k = 0, 1, \ldots, n$$

Empirical CDF: properties

1. $E_F \hat{F}_n(t) = F(t)$
2. $\mathrm{Var}\hat{F}_n(t) = \frac{1}{n}F(t)(1-F(t))$
3. from CLT: $$\frac{\hat{F}_n(t)-F(t)}{\sqrt{F(t)(1-F(t))}}\sqrt{n} \xrightarrow[n\to\infty]{} N(0,1)$$

   i.e., for any *z:* $$P\left(\frac{\hat{F}_n(t)-F(t)}{\sqrt{F(t)(1-F(t))}}\sqrt{n} \le z\right) \to \Phi(z)$$
4. Glivenko-Cantelli Theorem

$$\sup_{t\in\mathcal{R}} |\hat{F}_n(t)-F(t)| \xrightarrow{a.s.} 0$$

for $n \to \infty$

if sample size increases, we will approximate the unknown distribution with any given level of precision

Order statistics

- Let X_1, X_2, ..., X_n be a sample from a distribution with CDF F. If we organize the observations in ascending order:

  $X_{1:n}$, $X_{2:n}$, ..., $X_{n:n}$ $\leftarrow$ **order statistics**

  ($X_{1:n}$ = min, $X_{n:n}$ = max)
- An empirical CDF is a stair-like function, constant over intervals $[X_{i:n}, X_{i+1:n})$

Distribution of order statistics

- Let X_1, X_2, ..., X_n be independent random variables from a distribution with CDF F. Then $X_{k:n}$ has a CDF equal to

$$F_{k:n}(x) = P(X_{k:n} \leq x) = \sum_{i=k}^{n} \binom{n}{i} \left(F(x)\right)^i \left(1 - F(x)\right)^{n-i}$$

- If additionally the distribution is continuous with density f, then $X_{k:n}$ has density

$$f_{k:n}(x) = n \binom{n-1}{k-1} f(x) \left(F(x)\right)^{k-1} \left(1 - F(x)\right)^{n-k}$$

Sample moments and quantiles as estimators

Sample moments and quantiles are moments and quantiles of the empirical distribution, so they are estimators of the corresponding theoretical values.

- sample mean = estimator of the expected value
- sample variance = estimator of variance
- sample median = estimator of median
- sample quantiles = estimators of quantiles

Method of Moments Estimation (MM)

- We compare the theoretical moments (depending on unknown parameter(s)) to their empirical counterparts.
- Justification: limit theorems
- We need to solve a (system of) equation(s).

EMM – cont.

- If θ is single-dimensional, we use one equation, usually: $E_\theta X = \overline{X}$
- If θ is two-dimensional, we use two equations, usually:

$$\begin{cases} E_\theta X = \overline{X}, \\ \mathrm{Var}_\theta X = \hat{S}^2 \end{cases}$$

- If θ is k-dimensional, we use k equations, usually

$$\begin{cases} E_\theta X = \overline{X}, \\ \mathrm{Var}_\theta X = \hat{S}^2, \\ E_\theta (X - E_\theta X)^3 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^3, \\ \ldots .\ E_\theta (X - E_\theta X)^k = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^k \end{cases}$$

MME – Example 1.

- Exponential model: $X_1, X_2, \ldots, X_n$ are a sample from an exponential distr. Exp(λ).

we know: $$E_\lambda X = \frac{1}{\lambda}$$

equation: $$\frac{1}{\lambda} = \overline{X}$$

solution:

$$\hat{\lambda} = MME(\lambda) = \hat{\lambda}_{MM} = \frac{1}{\overline{X}}$$

MME – Example 2.

- Gamma model: $X_1, X_2, \ldots, X_n$ are a sample from distr. Gamma(α,λ).

We know: $$E_{\alpha,\lambda}X = \frac{\alpha}{\lambda}, \quad \text{Var}_{\alpha,\lambda}X = \frac{\alpha}{\lambda^2}$$

System of equations:

$$\frac{\alpha}{\lambda} = \overline{X}, \quad \frac{\alpha}{\lambda^2} = \hat{S}^2$$

Solution:

$$\hat{\lambda}_{MM} = \frac{\overline{X}}{\hat{S}^2}, \quad \hat{\alpha}_{MM} = \frac{\overline{X}^2}{\hat{S}^2}$$

$$f_{\alpha,\lambda}(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \text{ dla } x > 0$$

Method of Quantiles Estimation (MQ)

- If moments are hard to calculate or formulae are complicated, we can use quantiles instead of moments. We choose as many levels of *p* as we have parameters, and we put

$$q_p(\theta) = \widehat{q}_p$$

or equivalently

$$F_\theta(\widehat{q}_p) = p$$

MQE – Example 1.

- Exponential model: $X_1, X_2, \ldots, X_n$ are a sample from an exponential distr. Exp(λ).

  CDF: $F_\lambda = 1 - \exp(-\lambda x)$ for $\lambda > 0$

  one parameter → one equation, usually for the median

  $$1 - \exp(-\lambda \hat{q}_{1/2}) = \tfrac{1}{2}$$

  solution:

  $$MQE(\lambda) = \hat{\lambda}_{MQ} = -\frac{\ln \frac{1}{2}}{\hat{q}_{1/2}} = \frac{\ln 2}{\text{Med}}$$

MQE – Example 2.

- Weibull Model: $X_1, X_2, \ldots, X_n$ are a sample from a distribution with CDF

for b=1 exponential distr. with parameter c

$$F_{b,c} = 1 - \exp(-cx^b)$$

where b, c >0 are unknown parameters.

two parameters → two equations, usually quartiles

$$\begin{cases} 1 - \exp(-c\hat{q}_{1/4}^{b}) = \tfrac{1}{4} \\ 1 - \exp(-c\hat{q}_{3/4}^{b}) = \tfrac{3}{4} \end{cases}$$

solution:

$$MQE(b) = \hat{b}_{MQ} = \ln\left(\ln 4 / (\ln 4 - \ln 3)\right) / (\ln \hat{q}_{3/4} - \ln \hat{q}_{1/4}),$$

$$MQE(c) = \widehat{c}_{MQ} = \ln 4 \hat{q}_{3/4}^{-\hat{b}}$$

Properties of MME and MQE estimators

- Simple conceptually
- Not too complicated calculations
- BUT: sometimes not optimal (large errors, bad properties for small samples)
- Better method (usually): maximum likelihood

Maximum Likelihood Estimation (MLE)

We choose the value of θ for which the obtained results have the highest probability

Likelihood – describes the (joint) probability *f* (density or discrete probability) treated as a function of θ, for a given set of observations;
$L\colon \Theta \to \mathbb{R}$

$$L(\theta) = f(\theta; x_1, x_2, \ldots, x_n)$$

Maximum Likelihood Estimator

$\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ is the MLE of θ, if

$$f(\hat{\theta}(x_1, x_2, \ldots, x_n); x_1, x_2, \ldots, x_n) = \\ = \sup_{\theta \in \Theta} f(\theta; x_1, x_2, \ldots, x_n)$$

for any $x_1, x_2, \ldots, x_n$.

Denoted:

$$\hat{\theta} = \hat{\theta}_{ML} = MLE(\theta)$$

$$\mathrm{MLE}(g(\theta)) = g(\mathrm{MLE}(\theta))$$

independence of observations not required in the definition, but greatly simplifies calculations

MLE – practical problems

- Usually: sample of independent obs.

  Then:

$$L(\theta) = f_\theta(x_1) f_\theta(x_2) \ldots f_\theta(x_n)$$

- If $L(\theta)$ is differentiable, and θ is k-dimensional, then the maximum may be found by solving:

$$\frac{\partial L(\theta)}{\partial \theta_j} = 0, \quad j = 1, 2, \ldots, k$$

- very frequently: instead of max $L(\theta)$ we look for max $l(\theta) = \ln(L(\theta))$

MLE – Example 1.

- Quality control, cont. We maximize

$$L(\theta) = P_\theta(X = x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

or equivalently maximize

$$l(\theta) = \ln\binom{n}{x} + \ln(\theta^x) + \ln((1-\theta)^{n-x}) = \ln\binom{n}{x} + x\ln(\theta) + (n-x)\ln(1-\theta)$$

i.e. solve $\quad l'(\theta) = \dfrac{x}{\theta} - \dfrac{n-x}{1-\theta} = 0$

solution: $\quad MLE(\theta) = \hat{\theta}_{ML} = \dfrac{x}{n}$

MLE – Example 2.

- Exponential model: X_1, X_2, ..., X_n are a sample from Exp(λ), λ unknown.

We have: $L(\lambda) = f_\lambda(x_1, x_2, ..., x_n) = \lambda^n e^{-\lambda \Sigma x_i}$

we maximize

$$l(\lambda) = \ln L(\lambda) = n \ln \lambda - \lambda \Sigma x_i$$

we solve

$$l'(\lambda) = \frac{n}{\lambda} - \Sigma x_i = 0$$

we get

$$\hat{\lambda}_{ML} = \frac{1}{\overline{X}}$$

MLE – Example 3.

- Normal model: $X_1, X_2, \ldots, X_n$ are a sample from N(μ, σ^2). μ, σ unknown.

$$l(\mu,\sigma) = \ln\left(\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2\right)\right)$$

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

we solve

$$\begin{cases} \dfrac{\partial l}{\partial \sigma} = -\dfrac{n}{\sigma} + \dfrac{1}{\sigma^3}(\Sigma x_i^2 - 2\mu\Sigma x_i + n\mu^2) = 0 \\ \dfrac{\partial l}{\partial \mu} = \dfrac{1}{\sigma^2}\Sigma x_i - \dfrac{n\mu}{\sigma^2} = 0 \end{cases}$$

we get:

$$\hat{\mu}_{ML} = \overline{X}, \quad \hat{\sigma}^2_{ML} = \frac{1}{n}\sum(x_i - \overline{X})^2$$