Statistics with R

Chapter 3: Point estimation

Tabea Rebafka

October 2018

Master AIMS 2018–19

Outline

Setting

- In the whole chapter, let $(x_1, \ldots, x_n)$ be a realization of the distribution $\mathbb{P}_{\theta_0}$ and consider a **parametric** statistical model $\{\mathbb{P}_\theta, \theta \in \Theta\}$ with $\Theta \subset \mathbb{R}^d$ and $d < \infty$.
- We present two main approaches for the construction of an estimator of θ_0 based on the data $(x_1, \ldots, x_n)$:
  - the method of moments
  - the maximum likelihood estimator

Method of moments I

The method of moments of introduced by Karl Pearson in 1894.

Assumptions

- Let $(x_1, \ldots, x_n)$ be i.i.d. realizations from $\mathbb{P}_{\theta_0} \in \mathcal{P} = \{\mathbb{P}_\theta, \theta \in \Theta\}$ with $\Theta \subset \mathbb{R}^d$ and $d < \infty$.
- Suppose that $\mathbb{E}_\theta[|X|^d] < \infty$ for all $\theta \in \Theta$.
- Denote the r-th moment by $\mu_r(\theta) = \mathbb{E}_\theta[X^r]$.
  Suppose that the mappings $\theta \mapsto \mu_r(\theta)$ for $r = 1, \ldots, d$ are explicitly known.

Method of moments II

- If the "true values" $\mu_r^* = \mu_r(\theta_0)$ were known, one could solve the system of equations

$$\mu_r(\theta) = \mu_r^*, \quad r = 1, \ldots, d$$

to find θ_0.

- As the "true" theoretical moments μ_r^* are unknown, the values μ_r^* may be estimated by the sample moments computed on the data, that is by

$$m_r = \frac{1}{n} \sum_{i=1}^{n} x_i^r.$$

- That means, we hope that the solution of

$$\mu_r(\theta) = m_r, \quad r = 1, \ldots, d \tag{1}$$

is close to θ_0.

Method of moments III

Definition

Any statistic $\hat{\theta}^{MM}$ taking its values in Θ that is solution of (1) is called an **estimator by the method of moments** (EMM).

- The EMM may not exist.
- The EMM mai not be unique.

Method of moments IV

Generalization of the method of moments:

- Instead of using the first d moments, we can apply the same procedure with any moments of the form $\mathbb{E}_\theta[\varphi_r(X)]$, where φ_r are any arbitrary integrable functions.
- Denote $\tilde{\mu}_r(\theta) = \mathbb{E}_\theta[\varphi_r(X)]$ and suppose that these are explicitly known functions.

Definition

Any statistic $\hat{\theta}^{GMM}$ taking its values in Θ that is solution of the following system of equations

$$\tilde{\mu}_r(\theta) = \frac{1}{n}\sum_{i=1}^{n}\varphi_r(x_i), \quad r = 1, \ldots, d$$

is called an **estimator by the generalized method of moments** (GMM).

Method of moments V

Example: Uniform distribution

1. Let $x_1, \ldots, x_n$ be i.i.d realizations from the uniform distribution $U[0, \theta]$ with $\theta > 0$.
   Compute the EMM of θ.
2. Let $x_1, \ldots, x_n$ be i.i.d realizations from the uniform distribution $U[-\theta, \theta]$ with $\theta > 0$.
   - Apply the method of moments using the first moment.
   - Apply the method of moments using the second moment.
   - Apply the generalized method of moments to estimate θ.

Maximum likelihood estimator I

Tossing a coin

- Let $x_1, \ldots, x_n$ be i.i.d realizations from the Bernoulli distribution $B(p)$ with unknown parameter $p \in (0, 1)$.
- Compute the probability (or likelihood) of observing $x_1, \ldots, x_n$ depending on p:

$$\begin{aligned}\mathcal{L}(x_1, \ldots, x_n; p) &= \mathbb{P}_p(X_1 = x_1, \ldots, X_n = x_n) \\ &= p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}\end{aligned}$$

- Numerical example: Let $n = 5$ and $(x_1, \ldots, x_5) = (1, 1, 1, 1, 0)$:

$$\mathcal{L}((1,1,1,1,0); p) = p^4(1-p) = \begin{cases} 0.03125 & \text{if } \; p = 0.5 \\ 0.08192 & \text{if } \; p = 0.8 \end{cases}$$

Maximum likelihood estimator II

Tossing a coin

- Maximum likelihood approach: The value of p that maximizes

$$p \mapsto \mathcal{L}((1,1,1,1,0); p)$$

  is the most likely parameter, as with this parameter the probability to observe $(1,1,1,1,0)$ is maximal.
- Hence, an estimator of p is given by

$$\hat{p}^{ML} = \arg \max_{p \in (0,1)} \mathcal{L}((1,1,1,1,0); p) = \bar{x}_n = 0.8$$

Maximum likelihood estimator III

General method

- Ronald Fisher (1922) introduced the general approach and gave its theoretical foundations.
- Assume that all $\mathbb{P}_\theta \in \{\mathbb{P}_\theta, \theta \in \Theta\}$ are of the same type: either all continuous or all discrete.
- Define the **likelihood function** by

$$\theta \mapsto \mathcal{L}(\mathbf{x};\theta) = \begin{cases} \mathbb{P}_\theta(X_1 = x_1, \ldots, X_n = x_n) & \text{in the discrete case} \\ f_\theta(x_1, \ldots, x_n) & \text{in the continuous case} \end{cases}$$

$$\overset{\text{if } x_i \text{ i.i.d.}}{=} \begin{cases} \prod_{i=1}^n \mathbb{P}_\theta(X = x_i) & \text{in the discrete case} \\ \prod_{i=1}^n f_\theta(x_i) & \text{in the continuous case} \end{cases}$$

Maximum likelihood estimator IV

Definition

Every statistic $\hat{\theta}^{ML} \in \Theta$ that satisfies

$$\mathcal{L}(\mathbf{x}; \hat{\theta}^{ML}) = \max_{\theta \in \Theta} \mathcal{L}(\mathbf{x}; \theta). \tag{2}$$

is called **maximum likelihood estimator (MLE)** of θ in the statistical model $\{P_\theta, \theta \in \Theta\}$. To put it differently, $\hat{\theta}^{ML}$ is given by

$$\hat{\theta}^{ML} = \arg\max_{\theta \in \Theta} \mathcal{L}(\mathbf{x}; \theta).$$

Maximum likelihood estimator V

- The MLE may not be unique.
- The MLE may not exist.
- If the support of the distributions $\mathbb{P}_\theta$ does not depend on θ, i.e. e.g. in the continuous case if $\{x : f_\theta(x) > 0\}$ is the same for all $\theta \in \Theta$, then define the **log-likelihood function** by

$$\ell(\theta) = \log(\mathcal{L}(\mathbf{x}; \theta)).$$

Then

$$\hat{\theta}^{ML} = \arg\max_{\theta \in \Theta} \ell(\theta).$$

Maximum likelihood estimator VI

Exercise: Uniform distribution

Let $x_1, \ldots, x_n$ be i.i.d realizations of X and compute the MLE in the following cases:

- when X has uniform distribution $U[0, \theta]$ with $\theta > 0$.
- when X has uniform distribution $U[\theta, \theta + 1]$ with $\theta \in \mathbb{R}$.

Properties of point estimators I

- Let $\mathbf{x}$ be an observation and $\mathcal{P} = \{\mathbb{P}_\theta, \theta \in \Theta\}$ a statistical model.
- Denote θ_0 the true parameter value such that $\mathbf{x}$ is a realization of $\mathbb{P}_{\theta_0}$.
- Denote $\hat{\theta} = \hat{\theta}(\mathbf{x})$ an estimator of θ_0.
- Question: **When is $\hat{\theta}$ a good estimator of θ_0?**

Properties of point estimators II

- Answer: When $\hat{\theta}(\mathbf{x})$ is "close" to θ_0.
- However, we want $\hat{\theta}(\mathbf{x}) \approx \theta_0$ not only for the current dataset $\mathbf{x}$, but **for any realization x from** $\mathbb{P}_{\theta_0}$.
- Hence, analyze the **random variable** $\hat{\theta}(\mathbf{X})$ with $\mathbf{X} \sim \mathbb{P}_{\theta_0}$.
- However, θ_0 is unknown.
- So, analyze the random variable $\hat{\theta}(\mathbf{X})$ with $\mathbf{X} \sim \mathbb{P}_{\theta}$ **for any** $\theta \in \Theta$.

$\hat{\theta}$ is a "good" estimator if $\hat{\theta}(\mathbf{X})$ with $\mathbf{X} \sim \mathbb{P}_{\theta}$ is "close" to θ for any $\theta \in \Theta$

Consistency I

- We expect that the more data, i.e. the more information on $\mathbb{P}_\theta$ we have, the better is the estimation.
- If we have an infinity of information, i.e. when the sample size n tends to infinity, then the estimation should be perfect.

Definition

An estimator $\hat{\theta}_n$ of θ is said to be **consistent** if and only if

$$\hat{\theta}_n = \hat{\theta}_n(\mathbf{X}_n) \xrightarrow{P} \theta, \qquad \text{for all } \theta \in \Theta,$$

where $\mathbf{X}_n \sim \mathbb{P}_\theta$.

Consistency II

- Consistency is a rather weak property of an estimator.
- Only consistent estimators should be considered.
- Example of a non consistent estimator: Let $X_i \sim \mathcal{N}(\theta, 1)$ i.i.d. Consider the estimator $\hat{\theta}_n = (X_{n-1} + X_n)/2$. Then $\hat{\theta}_n$ does not converge in probability as $n \to \infty$.

Consistency III

- Both the EMM and the MLE are in general consistent estimators.
- To show consistency especially of the EMM, the law of large numbers and the following **continuous mapping theorem** can often be used.

Theorem (Continuous mapping)

- Let $g : D \subset \mathbb{R} \longrightarrow \mathbb{R}$ be a continusous mapping.
- Let $X, X_1, X_2, \ldots$ be random variables with values in D *a.s.*

Then

$$X_n \xrightarrow{P} X \quad \Longrightarrow \quad g(X_n) \xrightarrow{P} g(X) \quad (n \to \infty).$$

Likewise for convergence almost surely and convergence in distribution.

Consistency IV

Exercise.

Let $X_i, i = 1, 2, \ldots$ be i.i.d. r.v. from the exponential distribution $\mathcal{E}(\lambda)$ with $\lambda > 0$.

- Compute the EMM $\hat{\lambda}_n^{MM}$ of λ.
- Show that the EMM $\hat{\lambda}_n^{MM}$ is a consistent estimator of λ.

Mean squared error I

Defintion

The **mean squared error** (MSE) or **quadratic risk** of an estimator $\hat{\theta}$ of θ is defined as

$$\mathrm{MSE}(\theta, \hat{\theta}) = \mathbb{E}_\theta[\|\hat{\theta} - \theta\|^2],$$

where $\mathbb{E}_\theta$ means that $\hat{\theta} = \hat{\theta}(\mathbf{X})$ with $\mathbf{X} \sim \mathbb{P}_\theta$.

If we admit $+\infty$ as a valid value, then the MSE is well defined for any estimator $\hat{\theta}$.

Mean squared error II

Theorem

If

$$\lim_{n\to\infty} \mathrm{MSE}(\theta, \hat{\theta}_n) = 0 \quad \text{for all } \theta \in \Theta,$$

then $\hat{\theta}_n$ is a consistent estimator of θ.

Mean squared error III

Theorem

The MSE can be decomposed in two terms:

$$\begin{aligned}\mathrm{MSE}(\theta, \hat{\theta}) &= \left(\|\mathbb{E}_\theta[\hat{\theta}] - \theta\| \right)^2 + \mathbb{E}_\theta \left[\|\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}]\|^2 \right] \\ &=: b^2(\theta, \hat{\theta}) + \sigma^2(\theta, \hat{\theta}).\end{aligned}$$

- $b(\theta, \hat{\theta})$ is the bias of $\hat{\theta}$. It corresponds to the systematic deviation of the estimator from the target.
- $\sigma^2(\theta, \hat{\theta})$ is a measure of the stochastic variability of the estimator.
- If $b(\theta, \hat{\theta}) = 0$, the estimator is called **unbiased**.
- If $\lim_{n \to \infty} b(\theta, \hat{\theta}_n) = 0$, the estimator is called **asymptotically unbiased**.
- A good estimator has **both: small bias and small variance**.

Mean squared error IV

- The smaller the MSE, the better is the performance of the estimator.
- The MSE can be used to compare different estimators of θ in the same statistical model **at finite sample size** n.

Definition

Let $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$ be both estimators of θ. If

$$\mathrm{MSE}(\theta, \hat{\theta}^{(1)}) \leq \mathrm{MSE}(\theta, \hat{\theta}^{(2)}) \quad \text{for all } \theta \in \Theta,$$

and if there exists $\theta' \in \Theta$ such that the above inequality is sharp, then $\hat{\theta}^{(1)}$ is said to be **more efficient** than $\hat{\theta}^{(2)}$ and $\hat{\theta}^{(2)}$ is said to be **inadmissible**.

Mean squared error V

Exercise

Let $X_i, i = 1, 2, \ldots$ be i.i.d. r.v. from the exponential distribution $\mathcal{E}(\lambda)$ with $\lambda > 0$.

- Compute the MSE of the EMM $\hat{\lambda}_n^{MM} = 1/\bar{X}_n$ and determine the rate of convergence to 0 of the MSE.

Mean squared error VI

Hint

- The density of the **Gamma distribution**, denoted by $\Gamma(p, \theta)$, with parameters $p > 0$ and $\theta > 0$ is given by

$$f(x) = \frac{\theta^p}{\Gamma(p)} x^{p-1} \mathrm{e}^{-\theta x} \mathbb{1}_{\{x>0\}},$$

  where $\Gamma(\alpha)$ denotes the Gamma function such that

$$\Gamma(\alpha) \stackrel{\text{def}}{=} \int_0^\infty x^{\alpha-1} \mathrm{e}^{-x} \mathrm{d}x, \quad \Gamma(\alpha+1) = \alpha\Gamma(\alpha), \quad \Gamma(1) = 1.$$

- Let $X \sim \Gamma(p, \theta)$. The characteristic function is given by

$$\varphi(t) = \mathbb{E}[\mathrm{e}^{itX}] = \frac{1}{(1 - it/\theta)^p}.$$

Limit distribution and rate of convergence I

Definition

Let $\hat{\theta}_n$ be a consistent estimator of θ. The distribution function G_θ is called the **limit distribution** of $\hat{\theta}_n$, if there exists a sequence $(r_n)_{n\geq 1}$ of positive real numbers such that $r_n \to \infty$ and

$$r_n(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \eta \sim G_\theta, \qquad \text{pour tout } \theta \in \Theta, \tag{3}$$

as $n \to \infty$. We say that $1/r_n$ is the **rate of convergence** of $\hat{\theta}_n$ and $\mathrm{Var}(\eta)$ is the **limit variance**.

- In parametric models (i.e. $\Theta \subset \mathbb{R}^d$), $1/r_n = n^{-1/2}$ is the most frequent rate of convergence.
- In non parametric models, rates of convergence are generally slower.

Limit distribution and rate of convergence II

The estimator $\hat{\theta}_n$ is called **asymptotically normal** if

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma_\theta^2), \qquad \text{for all } \theta \in \Theta,$$

as $n \to \infty$.

Limit distribution and rate of convergence III

Comparison of two estimators:

- We prefer the one with the faster rate of convergence.
- If both have the same rate of convergence, we prefer the one with the smaller limit variance.

Limit distribution and rate of convergence IV

To determine the limit distribution, we often use the central limit theorem and the following **delta method**.

Thoerem (Delta method)

- Let $g : D \subset \mathbb{R} \to \mathbb{R}$ be a continuously differentiable function.
- Let $\mathbf{X}_1, \mathbf{X}_2, \ldots$ and $\mathbf{X}$ be random variables taking their values in D *a.s.*.
- Assume that $r_n(\mathbf{X}_n - m) \xrightarrow{d} \mathbf{X}$, where $r_n \to \infty$ is a sequence of positive real numbers such that $r_n \to \infty$ and $m \in \mathbb{R}$ is a constant.

Then

$$r_n \left(g(\mathbf{X}_n) - g(m)\right) \xrightarrow{d} g'(m)\mathbf{X}, \quad n \to \infty.$$

Limit distribution and rate of convergence V

Corollary

- Let $g(\cdot)$ be a continuously differentiable function.
- Let $X_1, X_2, \ldots$ be i.i.d. r.v. such that $\mathbb{E}[X_1^2] < \infty$ and $\sigma^2 = \mathrm{Var}(X_1) > 0$.

Then

$$\sqrt{n}\left(g(\bar{X}_n) - g(\mathbb{E}[X_1])\right) \xrightarrow{d} \mathcal{N}(0, (g'(\mathbb{E}[X_1]))^2\sigma^2) \quad \text{as } n \to \infty.$$

Limit distribution and rate of convergence VI

Exercise.

Let $X_i, i = 1, 2, \ldots$ be i.i.d. r.v. from the exponential distribution $\mathcal{E}(\lambda)$ with $\lambda > 0$.

- Determine the limit distribution and the rate of convergence of the EMM $\hat{\lambda}_n^{MM} = 1/\bar{X}_n$.