MATHEMATICAL STATISTICS, Lesson 6.

Methods of point estimation

- *Maximum likelihood (ML) principle*: maximize the likelihood or log-likelihood function in θ! The ML-estimator is asymptotically unbiased, efficient, and strongly consistent as guaranteed by the forthcoming theorem.

  Theorem (Cramér–Dugué) Assume that the dominated, identifiable statistical space $(\Omega, \mathcal{A}, \mathcal{P})$, $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ satisfies the regularity conditions below (they are stated for 1-dimensional parameter space and for absolutely continuous distributions). Let $X_1, \dots, X_n$ be i.i.d. sample from the $\mathbb{P}_{\theta^*}$ distribution.

  1. In the open neighborhood of the „true" parameter θ^* there exist the first three partial derivatives (w.r.p. θ) of $f_\theta(x)$.
  2. $\exists F_1, F_2, F_3$ real functions s.t.

  $$\left|\frac{\partial}{\partial\theta} f_\theta(x)\right| < F_1(x), \qquad \left|\frac{\partial^2}{\partial\theta^2} f_\theta(x)\right| < F_2(x),$$

  $$\left|\frac{\partial^3}{\partial\theta^3} \ln f_\theta(x)\right| < F_3(x),$$

  $\forall\theta \in U$ and $x \in supp(f_\theta)$ where F_1 and F_2 are integrable functions and $\mathbb{E}_{\theta^*}(F_3(X)) < \infty$.
  3. $I_1(\theta) < \infty$ $(\forall\theta \in U)$ and $I_1(\theta^*) > 0$.

  Then the likelihood equation corresponding to the above sample has a root $\hat{\theta}_n$ for which the following asymptotics hold if $n \to \infty$:

  1. $\mathbb{P}_{\theta^*}(\hat{\theta}_n \to \theta^*) = 1$,
  2. $\sqrt{n}(\hat{\theta}_n - \theta^*) \to \mathcal{N}(0, 1/I_1(\theta^*))$ in distribution,
  3. all finite moments of $\sqrt{n}(\hat{\theta}_n - \theta^*)$ tend to the same moments of the $\mathcal{N}(0, 1/I_1(\theta^*))$-distribution.

- *Method of moments*: $\dim(\Theta) = k$ and find the first k moments as the function of $\theta_1, \dots, \theta_k$. The moment estimator $\hat{\theta}_j$ is the inverse function of the empirical moments.

  Let $X_1, \dots, X_n \sim \mathbb{P}_{\underline{\theta}}$ i.i.d. sample, $\underline{\theta} = (\theta_1, \dots, \theta_k)$. Consider the first k moments of the $\mathbb{P}_{\underline{\theta}}$-distribution:

  $$m_j = \mathbb{E}_{\underline{\theta}}(X^j) = g_j(\theta_1, \dots, \theta_k), \qquad j = 1, \dots, k.$$

  Assume that $(g_1, \dots, g_k) : \mathbb{R}^k \to \mathbb{R}^k$ has an inverse $(h_1, \dots, h_k) : \mathbb{R}^k \to \mathbb{R}^k$ (in terms of the Jacobian), i.e., $h_i(m_1, \dots, m_k) = \theta_i$.

  The *moment estimator* of θ_i is

  $$\hat{\theta}_i = h_i(\hat{m}_1, \dots, \hat{m}_k), \qquad i = 1, \dots, k$$

  where

  $$\hat{m}_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j$$

  is the j-th empirical moment of the sample. The moment estimator is usually strongly consistent.

Proposition Under the usual regularity conditions, in exponential families the likelihood equation boils down to solving

$$\mathbb{E}_{\underline{\theta}}(t(\mathbf{X})) = t(\mathbf{X}),$$

where $\underline{\theta}$ is canonical parameter and t is the canonical statistic.

Proof. The likelihood-function has the following form:

$$L_\theta(\mathbf{X}) = c^n(\underline{\theta}) \cdot e^{\sum_{j=1}^k \theta_j \sum_{i=1}^n t_j(X_i)} \cdot \prod_{i=1}^n h(x_i) = \frac{1}{a(\underline{\theta})} \cdot e^{\underline{\theta} \cdot t^T(\mathbf{X})} \cdot b(\mathbf{X}),$$

where the vectors are rows, ^{T} denotes the transposition, and

$$a(\underline{\theta}) = \int_{\mathcal{X}} e^{\underline{\theta} \cdot t^T(\mathbf{x})} \cdot b(\mathbf{x})\, d\mathbf{x}. \tag{1}$$

is the normalizing constant, while $\mathcal{X} \subset \mathbb{R}^n$ is the sample space. This formula will play a crucial role in our subsequent calculations.

The likelihood equation is

$$\nabla_{\underline{\theta}} \ln L_{\underline{\theta}}(\mathbf{X}) = \mathbf{0},$$

that is

$$-\nabla_{\underline{\theta}} \ln a(\underline{\theta}) + \nabla_{\underline{\theta}}(t(\mathbf{X})\underline{\theta}^T) = \mathbf{0}. \tag{2}$$

Under certain regularity conditions, by (1) we get that

$$\nabla_{\underline{\theta}} \ln a(\underline{\theta}) = \frac{1}{a(\underline{\theta})} \int_{\mathcal{X}} t(\mathbf{x}) e^{t(\mathbf{x})\underline{\theta}^T} \cdot b(\mathbf{x})\, d\mathbf{x} = \mathbb{E}_{\underline{\theta}}(t(\mathbf{X})).$$

Therefore, (2) is equivalent to

$$-\mathbb{E}_{\underline{\theta}}(t(\mathbf{X})) + t(\mathbf{X}) = \mathbf{0}$$

that finishes the proof. □

Note that this resembles the idea of the moment estimation. Indeed, if θ_i's are the true moments and $t_1(\mathbf{X}) = \sum_{i=1}^n X_i$, …, $t_k(\mathbf{X}) = \sum_{i=1}^n X_i^k$, then the ML-estimator of the canonical parameter is the same as the moment-estimator. This is the case, e.g., when our underlying distribution is Poisson, exponential, or Gaussian. However, this is not the case when the population distribution is $\mathcal{U}[a, b]$.

- *Bayes estimation*

  Assume that θ has a *prior* distribution $q(t)$ ($t \in \Theta$) known before the actual sampling (e.g., based on former experience).

  Let $\mathbf{X} = (X_1, \ldots, X_n)$ be i.i.d. sample with likelihood function $L_t(\mathbf{x})$ that is the joint distribution of the sample entries conditioned on $\theta = t$: $L_t(\mathbf{x}) = f(\mathbf{x}|t)$. Consequently, the joint density of $\mathbf{X}$ and θ is

  $$f(\mathbf{x}, t) = f(\mathbf{x}|t) \cdot q(t), \qquad \mathbf{x} \in \mathcal{X}, \quad t \in \Theta.$$

  Then, by the Bayes rule, we can find the *posterior* distribution of θ based on the sample realization $\mathbf{X} = \mathbf{x}$:

  $$q(t|\mathbf{x}) = \frac{f(\mathbf{x}|t) \cdot q(t)}{f(\mathbf{x})}, \quad \text{where} \quad f(\mathbf{x}) = \int_{t \in \Theta} f(\mathbf{x}|t) \cdot q(t)\, dt.$$

The conditional expectation of θ on the same condition is

$$\mathbb{E}(\theta|\mathbf{X}=\mathbf{x}) = \int_{t\in\Theta} t\cdot q(t|\mathbf{x})\,dt = \frac{\int_{t\in\Theta} t\cdot f(\mathbf{x}|t)\cdot q(t)\,dt}{\int_{t\in\Theta} f(\mathbf{x}|t)\cdot q(t)\,dt} = T(\mathbf{x}).$$

The Bayes estimator of θ is $\mathbb{E}(\theta|\mathbf{X}) = T(\mathbf{X})$, which is the function of any sufficient statistic.

To estimate the parameter function $\psi(\theta)$, we have

$$\mathbb{E}(\psi(\theta)|\mathbf{X}=\mathbf{x}) = \int_{t\in\Theta} \psi(t)\cdot q(t|\mathbf{x})\,dt = S(\mathbf{x})$$

and the Bayes estimator is $S(\mathbf{X})$.

By the properties of the conditional expectation, the Bayes estimator (even if not unbiased) minimizes the squared risk

$$\mathbb{E}(\psi(\theta) - V(\mathbf{X}))^2$$

in all statistics V of $\mathbf{X}$.

Interval estimation. The random interval $(T_1(\mathbf{X}), T_2(\mathbf{X}))$ is a *confidence interval* of level at least (exactly) $1-\varepsilon$ for $\psi(\theta)$, if $\mathbb{P}_\theta(T_1 < \psi(\theta) < T_2) \geq (=) 1-\varepsilon$ $(\forall\theta\in\Theta)$.