MATHEMATICAL STATISTICS, Lessons 3-5.

Theory of point estimation. *Likelihood function*: for $\mathbf{x} = (x_1, \dots, x_n) \in \mathcal{X}$ and $\theta \in \Theta$ let $L_\theta(\mathbf{x}) = \mathbb{P}_\theta(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^n \mathbb{P}_\theta(X_i = x_i) = \prod_{i=1}^n p_\theta(x_i)$ in the discrete, and $L_\theta(\mathbf{x}) = \prod_{i=1}^n f_\theta(x_i)$ in the absolutely continuous case.

Definition: The statistic $T(\mathbf{X})$ is *sufficient* for θ if the distribution of $\mathbf{X}$ conditioned on $T(\mathbf{X})$ does not depend on θ.

Theorem (Neyman–Fisher factorization): The statistic $T(\mathbf{X})$ is *sufficient* for θ if and only if $L_\theta(\mathbf{x}) = g_\theta(T(\mathbf{x})) \cdot h(\mathbf{x})$, $\forall \theta \in \Theta, \quad \mathbf{x} \in \mathcal{X}$ with some measurable, nonnegative real functions g and h.

A sufficient statistic contains all the information for θ, and it is *minimal* if it is the function of any other sufficient statistic. Minimal sufficient statistic always exists, and it is unique up to equivalence.

Definition: The statistic $T = T(\mathbf{X})$ is *complete* for θ if $\mathbb{E}_\theta g(T) = 0$, $\forall \theta$ implies that $g(T(\mathbf{X})) = 0$ almost surely (with probability 1).

Theorem: If a statistic is sufficient and complete, than it is also minimal sufficient.

Definition: The $\mathbb{P}_\theta$ distribution belongs to the *exponential family* if its p.m.f. or p.d.f. has the form

$$c(\theta) \cdot \exp\left[\sum_{j=1}^k a_j(\theta) \cdot T_j(x)\right] \cdot h(x), \qquad \forall \theta \in \Theta,$$

where $k = \dim(\Theta)$, c and a_j's are measurable functions on Θ, whereas T_j's and h are real measurable functions.

Theorem: Let $\mathbf{X} = (X_1, \dots, X_n)$ be i.i.d. sample from the $\mathbb{P}_\theta$ distribution that belongs to the exponential family, $\theta \in \Theta \subset \mathbb{R}^k$. Then

$$T(\mathbf{X}) = \left(\sum_{i=1}^n T_1(X_i), \dots, \sum_{i=1}^n T_k(X_i)\right)$$

is sufficient for θ. (It is an easy consequence of the Neyman–Fisher factorization.)

When $a_j(\theta) = \theta_j$ $(j = 1, \dots, k)$, then $\theta = (\theta_1, \dots, \theta_k)$ is called *canonical parameter*, and the above $T(\mathbf{X})$ is called *canonical statistic*.

Theorem (Halmos: If the parameter space $\Theta \subset \mathbb{R}^k$ contains a k-dimensional parallelepiped, then the above statistic $T(\mathbf{X})$ is also complete, so it is minimal sufficient (in exponential family).

Let $(\Omega, \mathcal{A}, \mathcal{P})$ be parametric statistical space, $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$. We want to estimate θ, or its measurable function $\psi(\theta)$ by means of the statistic $T(\mathbf{X})$ on the basis of the i.i.d. sample $\mathbf{X} = (X_1, \dots, X_n)$. The point estimator is sometimes denoted by $\hat{\theta}$ or $\hat{\psi}$. Criteria for the „goodness" of a point estimator:

- $T(\mathbf{X})$ is an **unbiased** estimator of $\psi(\theta)$, if $\mathbb{E}_\theta(T(\mathbf{X})) = \psi(\theta), \quad \forall \theta \in \Theta$.
- $T(\mathbf{X}_n)$ is an **asymptotically unbiased** estimator of $\psi(\theta)$, if $\lim_{n\to\infty} \mathbb{E}_\theta(T(\mathbf{X}_n)) = \psi(\theta), \quad \forall \theta \in \Theta$.
- Let T_1 and T_2 be both unbiased estimators of $\psi(\theta)$. T_1 is **at least as efficient** than T_2, if $\mathbb{D}^2_\theta(T_1) \le \mathbb{D}^2_\theta(T_2)$, $\forall \theta \in \Theta$. An unbiased estimator is **efficient**, if it is at least as efficient than any other unbiased estimator. Efficient estimator does not always exist, but if yes, then it is unique (with probability 1).
- $T(\mathbf{X}_n)$ is a weakly/strongly **consistent** estimator of $\psi(\theta)$, if $\forall \theta \in \Theta$: $T(\mathbf{X}_n) \to \psi(\theta)$ in probability/almost surely as $n \to \infty$.

We want to give a lower bound for the variance of an unbiased estimator if $dim(\Theta) = 1$.
Definition: The Fisher information contained in the i.i.d. sample $\mathbf{X} = (X_1, \dots, X_n) \sim \mathbb{P}_\theta$ is

$$I_n(\theta) = \mathbb{E}_\theta \left(\frac{\partial}{\partial \theta} \ln L_\theta(\mathbf{X}) \right)^2 \geq 0, \quad \theta \in \Theta.$$

Note that $I_n(\theta) = nI_1(\theta)$ under the regularity conditions below, and then we also have

$$I_1(\theta) = -\mathbb{E}_\theta \left(\frac{\partial^2}{\partial \theta^2} \ln p_\theta(X_1) \right), \quad I_1(\theta) = -\mathbb{E}_\theta \left(\frac{\partial^2}{\partial \theta^2} \ln f_\theta(X_1) \right)$$

that gives an interesting relation to the Shannon entropy.

Theorem (Cramér–Rao inequality) In the above setup let $T(\mathbf{X})$ be unbiased estimator of the differentiable parameter function $\psi(\theta)$, and suppose that $\mathrm{Var}_\theta(T) < \infty$ $(\forall \theta \in \Theta)$. Further, the following regularity conditions hold, $\forall \theta \in \Theta$:

$$\frac{\partial}{\partial \theta} \int L_\theta(\mathbf{x})\, d\mathbf{x} = \int \frac{\partial}{\partial \theta} L_\theta(\mathbf{x})\, d\mathbf{x} \quad \text{and} \quad \frac{\partial}{\partial \theta} \int T(\mathbf{x}) L_\theta(\mathbf{x})\, d\mathbf{x} = \int T(\mathbf{x}) \frac{\partial}{\partial \theta} L_\theta(\mathbf{x})\, d\mathbf{x}.$$

Then $\mathrm{Var}_\theta(T) \geq \frac{(\psi'(\theta))^2}{I_n(\theta)} = \frac{(\psi'(\theta))^2}{nI_1(\theta)}$, $\quad \forall \theta \in \Theta$.

The Cramér–Rao inequality can be extended to biased estimates. Let $\mathbf{X} = (X_1, \dots, X_n)$ be i.i.d. sample from the $\mathbb{P}_\theta$ distribution, and the *bias* of the statistic $T(\mathbf{X})$ with respect to the differentiable parameter function $\psi(\theta)$ is

$$b_T(\theta) = \mathbb{E}_\theta(T) - \psi(\theta), \qquad \theta \in \Theta, \quad \dim(\Theta) = 1$$

which is also differentiable with respect to θ. Let $\mathrm{Var}\,\theta^2(T) < +\infty$, $\forall \theta \in \Theta$. Then under the usual regularity conditions,

$$\mathrm{Var}\,\theta^2(T) \geq \frac{(\psi'(\theta) + b_T'(\theta))^2}{I_n(\theta)}, \qquad \forall \theta \in \Theta.$$

The statement easily follows if we apply the Cramér–Rao inequality for the parameter function $\psi(\theta) + b_T(\theta)$. Observe that T is an unbiased estimator of it.

The Cramér–Rao inequality for multidimensional parameter spaces. Under some regularity conditions, it gives unified lower bound for the covariance matrix of every unbiased estimator (for a given parameter function) based merely on a quantity, called Fisher information matrix, which can be calculated from the underlying distribution as a function of the parameter.

Let $(\Omega, \mathcal{A}, \mathcal{P})$ be dominated, identifiable, parametric statistical space, and $X_1, \dots, X_n$ be i.i.d. (univariate or multivariate) sample from the $\mathbb{P}_\theta$ distribution, where $\theta \in \Theta \subset \mathbb{R}^k$ is the parameter space. Suppose, we want to estimate the function $\psi(\theta) = (\psi_1(\theta), \dots, \psi_k(\theta))$ of the parameter, where $\psi : \mathbb{R}^k \to \mathbb{R}^k$ is one-to-one function (often it is the identity, when the parameter θ itself is to be estimated).

- Based on the above i.i.d. sample, construct the k-dimensional statistic $\mathbf{T}(X_1, \dots, X_n)$, briefly $\mathbf{T} = (T_1, \dots T_k)$ which is an unbiased estimator of $\psi(\theta)$ in the sense that

  $$\mathbb{E}_\theta \mathbf{T} = \psi(\theta), \quad \forall \theta \in \Theta,$$

  where the expectation of the random vector $\mathbf{T}$ is the vector $(\mathbb{E}_\theta T_1, \dots, \mathbb{E}_\theta T_k)'$ and $'$ denotes the transposition (the vectors are column vectors now).

- Based on the $\mathbb{P}_\theta$ distribution itself, calculate the Fisher information matrix of the 1-element sample:

$$\mathbf{I}_1(\theta) = \mathbb{E}_\theta \left(\frac{\partial}{\partial \theta} \ln f_\theta(X) \right) \left(\frac{\partial}{\partial \theta} \ln f_\theta(X) \right)' = \mathrm{Var}_\theta \left(\frac{\partial}{\partial \theta} \ln f_\theta(X) \right),$$

where $X \sim \mathbb{P}_\theta$ and f_θ is the p.d.f. of the $\mathbb{P}_\theta$ distribution (if it is absolutely continuous, and use p_θ for the p.m.f. if $\mathbb{P}_\theta$ is discrete). Here under Var of a random vector its covariance matrix is understood, whereas under the derivative with respect to the vector θ of the scalar valued function g the column vector of the derivatives with respect to the components of θ, i.e., the gradient vector is understood, that is $\frac{\partial}{\partial \theta} g = (\frac{\partial}{\partial \theta_1} g, \dots, \frac{\partial}{\partial \theta_k} g)'$. The information matrix of the n-element sample is defined by $\mathbf{I}_n(\theta) = n\mathbf{I}_1(\theta)$, provided the regularity condition

$$\mathbb{E}_\theta \left(\frac{\partial}{\partial \theta} \ln f_\theta(\mathbf{X}) \right) = \mathbf{0}$$

holds, where $\mathbf{0}$ is the zero vector. In fact, this condition follows from some simpler ones (we can „differentiate through" the $\int$ or $\sum$). In this case, $\mathbf{I}_n(\theta)$ is the $k \times k$ covariance matrix of the k-dimensional random vector $\frac{\partial}{\partial \theta} \ln \prod_{i=1}^n f_\theta(X_i)$.

- The Cramér–Rao inequality states the following relation between 1 and 2. Denote by $\mathbf{S} = (s_{ij})$ the $k \times k$ matrix of entries $s_{ij} = \frac{\partial}{\partial \theta_j} \psi_i(\theta)$. Then for the covariance matrix of any unbiased estimator (for $\psi(\theta)$), the inequality

$$\mathrm{Var}_\theta \mathbf{T} \geq \frac{1}{n} \mathbf{S} \mathbf{I}_1^{-1}(\theta) \mathbf{S}' = \mathbf{S} \mathbf{I}_n^{-1}(\theta) \mathbf{S}'$$

holds, where the inequality means that the difference of the left and right hand side matrices is positive semidefinite. Here $\mathbf{S}$ also depends on θ, however, if θ itself is estimated, then $\mathbf{S}$ is the identity matrix and does not appear in the formula above.

Rao–Blackwell–Kolmogorov Theorem: Let $\mathbf{X} = (X_1, \dots, \mathbf{X}_n)$ be an i.i.d. sample from the $\mathbb{P}_\theta$ distribution, $\theta \in \Theta \subset \mathbb{R}^k$ ($k > 1$ can be). Let $T(\mathbf{X})$ be a sufficient statistic and $S(\mathbf{X})$ be an unbiased estimator for $\psi(\theta)$. Then one can construct an unbiased estimator $U = g(T)$ for $\psi(\theta)$, that is at least as efficient as S. The construction of U („blackwellization"): $U := \mathbb{E}_\theta(S|T) = g(T(\mathbf{X}))$, $\forall \theta \in \Theta$.

Note that if T is also complete (so minimal sufficient), then blackwellizing any unbiased S (for the same parameter function) with it results in the same (unique) U. Consequently, this U will be the efficient estimator for the given parameter function. The message of the R-B-K theorem is: find the efficient estimator among the functions of the minimal sufficient statistic.

The Rao–Blackwell–Kolmogorov theorem can also be extended to biased estimators as follows. Let S be a biased estimator of $\psi(\theta)$ and blackwellize it with the sufficient statistic T. Then the so obtained U has the same bias as S and

$$R_\theta(U) \leq R_\theta(S), \quad \forall \theta \in \Theta,$$

where $R_\theta(T) = \mathbb{E}_\theta(T - \psi(\theta))^2$ is the *squared risk* of estimating the parameter function $\psi(\theta)$ with the statistic T. Observe that by the Steiner's equality

$$R_\theta(T) = \mathrm{Var}_\theta^2(T) + b_T^2(\theta).$$