8 Parameter Estimations and Sufficient Statistics

A General Model for Statistics

- Many problems have the following common structure. A continuous signal $\{x(t) : t \in \mathbb{R}\}$ is measured at $t_1, \ldots, t_n$ producing vector $x = (x_1, \ldots, x_n)$, where $x_i = x(t_i)$. The vector x is a realization of a random vector or a random process $X = (X_1, \ldots, X_n)$ with a joint distribution which is **of known form but depends on some unknown parameters** $\theta = (\theta_1, \ldots, \theta_p)$. The estimation theory aims to *estimate* these unknown parameters θ based on the observed realization x.

- Formally, the above problem has the following ingredients:
  - $X = (X_1, \ldots, X_n)$ is a vector of random measurements or observations taken over the course of the experiment
  - $\mathcal{X}$ is sample or measurement space of realizations x of X, e.g., $\mathcal{X} = \mathbb{R} \times \cdots \times \mathbb{R}$
  - $\theta = (\theta_1, \ldots, \theta_p)$ is an unknown parameter vector of interest
  - Θ is parameter space for the experiment
  - $P_\theta : \mathcal{B}(\mathbb{R}^n) \to [0, 1]$ is a probability measure such that, for any Borel set or event B, we have

$$P_\theta(B) = \text{probability of event } B \subseteq \mathcal{X} = \begin{cases} \displaystyle\int_B f(x;\theta)dx & \text{if } X \text{ is continuous} \\ \displaystyle\sum_{x \in B} p(x;\theta) & \text{if } X \text{ is discrete} \end{cases}$$

  Such $\{P_\theta\}_{\theta \in \Theta}$ is called the **statistical model** of the experiment.

  The probability model also induces the joint C.D.F. associated with X

$$F(x;\theta) = P_\theta(X_1 \le x_1, \ldots, X_n \le x_n),$$

  which is assumed to be known for each $\theta \in \Theta$. We denote by $E_\theta(X)$ the expectation of random variable X given $\theta \in \Theta$.

Parametric Statistics (Estimation Theory)

- The basic estimation problem is as follows. The observations $X = (X_1, \ldots, X_n)$ is actually generated by a true parameter $\theta_0 \in \Theta$. In case X_i are i.i.d., we have $X_i \sim P_{\theta_0}(\cdot)$. Then we want to find an **estimator** $\hat{\theta} : X \to \Theta$ such that the **estimate** $\hat{\theta}(X_1, \ldots, X_n)$ approximates θ_0 "optimally".

- The question is that how to describe whether $\hat{\theta}$ is a good estimator. Depending on whether or not we have prior knowledge about the distribution θ, we will discuss two different approaches: **Bayesian estimation** and **non-random estimation**.

Definition of Sufficient Statistics

- Let us consider an i.i.d. observations $X = (X_1, \dots, X_n)$ with distribution P_θ from the family $\{P_\theta : \theta \in \Theta\}$. Imagine that there are two people A and B, and that
  - A observes the entire sample $(X_1, \dots, X_n)$;
  - B observes only a smaller vector $T = T(X_1, \dots, X_n)$ which is a function of the sample. In this case, function $T : \mathbb{R}^n \to \mathbb{R}^m, m \leq n$. is called a **statistic**.

  Clearly, A has more information about the distribution of the data and, in particular, about the unknown parameter θ. However, in some cases, for some choices of function T (called **sufficient statistics**) B will have as much information about θ as A has.

- To see this more clearly, for observations $X = (X_1, \dots, X_n)$ and statistic $T(X)$, the conditional probability

$$f_{X|T(X)}(x \mid t, \theta) = P_\theta(X_1 = x_1, \dots, X_n = x_n \mid T(X) = t)$$

  is, typically, a function of both t and θ. For some choices of statistic T, however, $f_{X|T(X)}(x \mid t, \theta)$ can be θ-independent.

- To see the above argument, let us consider consider the case $X = (X_1, \dots, X_n)$, a sequence of n Bernoulli trials with success probability parameter θ and the statistic $T(X) = X_1 + \cdots + X_n$ the total number of successes. Then

$$P_\theta(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^t(1-\theta)^{n-t},$$

  where $t = T(x_1, \dots, x_n) = x_1 + \cdots x_n$. Therefore, if $\sum_{i=1}^n x_i \neq t$, then we know that the statistic is incompatible with the observation. Otherwise, we have

$$f_{X|T(X)}(x \mid t, \theta) = \frac{f_X(x \mid \theta)}{f_{T(X)}(t \mid \theta)} = \frac{P_\theta(X_1 = x_1, \dots, X_n = x_n)}{P_\theta(T(X) = t)} = \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \binom{n}{t}^{-1}$$

  which does not depend on the parameter θ. This means that all information about θ in X has been summarized by $T(X)$. This motivates the following definition.

Definition: Sufficient Statistics

A statistic $T = T(X)$ is said to be sufficient for parameter θ if

$$P_\theta(X_1 \leq x_1, \dots, X_n \leq x_n \mid T(X) = t) = G(x, t)$$

where $G(\cdot, \cdot)$ is a function that does not depend on θ. Equivalent, we have

- $p(x \mid t, \theta) = P_\theta(X = x \mid T(X) = t) = G(x, t)$ if X is discrete;
- $f(x \mid t, \theta) = G(x, t)$ if X is continuous.

- Thus, by the law of total probability

$$P_\theta(X_1 \leq x_1, \dots, X_n \leq x_n) = P(X_1 \leq x_1, \dots, X_n \leq x_n \mid T(X) = T(x))P_\theta(T(X) = T(x))$$

  and once we know the value of the sufficient statistic, we cannot obtain any additional information about the value of θ from knowing the observed values.

Neyman-Fisher Factorization Theorem

- The above definition of sufficient statistics is often difficult to use since it involves derivation of the conditional distribution of X given T. However, when the random variable X is discrete or continuous a simpler way to verify sufficiency is through the **Neyman-Fisher factorization criterion**.

Theorem: Fisher Factorization Criterion

A statistic $T = T(X)$ is sufficient for θ if and only if functions g and h can be found such that

$$f_X(x \mid \theta) = g(T(x), \theta)h(x)$$

We only proof the case of discrete random variables, i.e., $f_X(x;\theta)$ is the PMF.

- ($\Rightarrow$) Because T is a function of x, we have

$$f_X(x \mid \theta) = f_{X,T(X)}(x, T(x) \mid \theta) = f_{X|T(X)}(x \mid T(x), \theta) f_{T(X)}(T(x) \mid \theta)$$

Since T is sufficient, then $f_{X|T(X)}(x \mid T(x), \theta)$ is not a function of θ and we can set it to be $h(X)$. The second term is a function of $T(x)$ and θ. We will write it $g(T(x), \theta)$.

- ($\Leftarrow$) Suppose that we have the factorization. By the definition of conditional expectation,

$$f_{X|T(X)}(x \mid t, \theta) = \frac{f_{X,T(X)}(x, t \mid \theta)}{f_{T(X)}(t \mid \theta)}$$

For the numerator, we have

$$f_{X,T(X)}(x, t \mid \theta) = \begin{cases} 0 & \text{if } T(x) \neq t \\ f_X(x \mid \theta) = g(t, \theta)h(x) & \text{otherwise} \end{cases}$$

Furthermore, for the denominator, we have

$$f_{T(X)}(t \mid \theta) = \sum_{\tilde{x}:T(\tilde{x})=t} f_X(\tilde{x} \mid \theta) = \sum_{\tilde{x}:T(\tilde{x})=t} g(t, \theta)h(\tilde{x})$$

Therefore, we have

$$f_{X|T(X)}(x \mid t, \theta) = \frac{g(t, \theta)h(x)}{\sum_{\tilde{x}:T(\tilde{x})=t} g(t, \theta)h(\tilde{x})} = \frac{h(x)}{\sum_{\tilde{x}:T(\tilde{x})=t} h(\tilde{x})},$$

which is independent of θ and, therefore, T is sufficient.

- For example, in the maximum likelihood estimation, we have to find the best estimate $\theta \in \Theta$ such that the *likelihood function*

$$L(\theta \mid x) = f_X(x \mid \theta)$$

is maximized for the observed sample $x = (x_1, \ldots, x_n)$. For sufficient statistics, since $f_X(x \mid \theta) = g(T(x), \theta)h(x)$, maximizing the likelihood is equivalent to maximizing $g(T(x), \theta)$ and the maximum likelihood estimator $\hat{\theta}(T(x))$ is a function of the sufficient statistic.

General Examples of Sufficient Statistics

- **Example 1: Entire Sample**
  $X = (X_1, \ldots, X_n)$ is clearly sufficient but not very interesting.

- **Example 2: Rank Ordered Sample**
  $X_{(1)}, \ldots, X_{(n)}$ is sufficient when X_i are i.i.d. This is because, under the i.i.d. setting,

$$f(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta) = \prod_{i=1}^{n} f(x_{(i)} \mid \theta)$$

- **Example 3: Binary Likelihood Ratios**
  Suppose that θ only takes two possible values $\Theta = \{\theta_0, \theta_1\}$, or simply $\theta \in \{0, 1\}$. This gives the binary decision problem: "decide between $\theta = 0$ versus $\theta = 1$. Then the "likelihood ratio" (assume it is finite)

$$\Lambda(X) = \frac{f_1(X)}{f_0(X)} = \frac{f(X \mid 1)}{f(X \mid 0)}$$

  is sufficient for θ, because we can write

$$f_\theta(X) = \theta f_1(X) + (1-\theta) f_0(X) = \Bigg(\underbrace{\theta \Lambda(X) + (1-\theta)}_{g(T,\theta)} \Bigg) \underbrace{f_0(X)}_{h(X)}$$

- **Example 4: Discrete Likelihood Ratios**
  Suppose that θ takes p possible values, i.e., $\Theta = (\theta_1, \ldots, \theta_p)$. Then the vector of $p-1$ likelihood ratios (assume it is finite)

$$\Lambda(X) = \left(\frac{f_{\theta_1}(X)}{f_{\theta_p}(X)}, \ldots, \frac{f_{\theta_{p-1}}(X)}{f_{\theta_p}(X)} \right) = (\Lambda_1(X), \ldots, \Lambda_{p-1}(X))$$

  is sufficient for θ. Try to prove this as a homework.

- **Example 5: Likelihood Ratio Trajectory**
  When Θ is a set of scalar parameters θ the likelihood ratio trajectory over Θ is

$$\Lambda(X) = \left\{ \frac{f_\theta(X)}{f_{\theta_0}(X)} \right\}_{\theta \in \Theta}$$

  is sufficient for θ. Here θ_0 is an arbitrary reference point in Θ for which the trajectory is finite for all X. When θ is not a scalar, this becomes a likelihood ratio surface, which is also a sufficient statistic.

- We say T_{min} is a **minimal sufficient statistic** if for any sufficient statistic T there exists a function q such that $T_{min} = q(T)$. Finding minimal sufficient statistic is in general difficult; the following provides a sufficient condition for $T(X)$ to be be minimal

$$\forall x, x' \in \mathcal{X} : \Lambda(T(x)) = \Lambda(T(x')) \Rightarrow T(x) = T(x')$$

  Note that $\Lambda(t)$ is well-defined because $\Lambda(x) = \Lambda(T(x))$ for any sufficient statistic T as we discussed above.

More Examples of Sufficient Statistics

- **Example 1: Bernoulli Distribution**
Suppose that $X = (X_1, X_2, \dots, X_n)$ is i.i.d. and each X_i satisfies the Bernoulli distribution with unknown probability, i.e., $P_\theta(X_i = 1) = \theta$ and $P_\theta(X_i = 0) = 1 - \theta$. Then we claim that $T(X) = \sum_{i=1}^n X_i$ is a sufficient statistic. To see this, we write the joint PMF as

$$p_X(X;\theta) = \prod_{i=1}^n p_{X_i}(X_i;\theta) = \prod_{i=1}^n \theta^{X_i}(1-\theta)^{1-X_i} = \prod_{i=1}^n (1-\theta)(\frac{\theta}{1-\theta})^{X_i}$$
$$= \underbrace{(1-\theta)^n(\frac{\theta}{1-\theta})^{T(X)}}_{g(T(X,\theta))} \cdot \underbrace{1}_{h(X)}$$

Clearly, this sufficient statistic is minimal as it is already one-dimensional.

- **Example 2: Uniform Distribution**
Suppose that $X = (X_1, X_2, \dots, X_n)$ is i.i.d. and each X_i satisfies the uniform distribution over $[0, \theta]$ with unknown length θ. Then we claim that $T(X) = \max_{i=1}^n X_i$ is a sufficient statistic. To see this, we write

$$f_X(X;\theta) = \prod_{i=1}^n f_{X_i}(X_i;\theta) = \prod_{i=1}^n \frac{1}{\theta}\mathbf{1}_{[0,\theta]}(X_i) = \prod_{i=1}^n \frac{1}{\theta}\mathbf{1}_{[X_i,\infty)}(\theta) = \underbrace{\frac{1}{\theta^n}\mathbf{1}_{[T(X),\infty)}(\theta)}_{g(T(X,\theta))} \cdot \underbrace{1}_{h(X)}$$

Note that, the tricky part is $I_{[0,\theta]}(X_i) = I_{[X_i,\infty)}(\theta)$.

- **Example 3: Gaussian Distribution with Unknown Mean**
Suppose that $X = (X_1, X_2, \dots, X_n)$ is i.i.d. and each X_i satisfies the Gaussian distribution with unknown mean θ but the variance σ^2 is known. Then we claim that $T(X) = \sum_{i=1}^n X_i$ is a sufficient statistic. To see this, we have

$$\begin{aligned}
& f_X(X;\theta) \\
&= \prod_{i=1}^n f_{X_i}(X_i;\theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(X_i-\theta)^2}{2\sigma^2}\right) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\sum_{i=1}^n \frac{(X_i-\theta)^2}{2\sigma^2}\right) \\
&= \underbrace{\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(\frac{\theta T(X)}{\sigma^2}\right)\exp\left(\frac{-n\theta^2}{2\sigma^2}\right)}_{g(T(X,\theta))} \underbrace{\exp\left(\frac{-\sum_{i=1}^n X_i^2}{2\sigma^2}\right)}_{h(X)}
\end{aligned}$$

- **Example 4: Gaussian Distribution with Unknown Mean and Variance**
When the unknown mean is $\mu = \theta_1$ and the unknown variance is $\sigma^2 = \theta_2$, then we claim that $T(X) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is a sufficient statistic. To see this, we have

$$\begin{aligned}
& f_X(X;\theta) \\
&= \prod_{i=1}^n f_{X_i}(X_i;\theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(X_i-\theta)^2}{2\sigma^2}\right) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\sum_{i=1}^n \frac{(X_i-\theta)^2}{2\sigma^2}\right) \\
&= \underbrace{\left(\frac{1}{\sqrt{2\pi\theta_2}}\right)^n \exp\left(\frac{\theta_1}{\theta_2}T_1(X) - \frac{1}{2\theta_2}T_2(X)\right)\exp\left(\frac{-n\theta_1^2}{2\theta_2}\right)}_{g(T(X,\theta))} \cdot \underbrace{1}_{h(X)}
\end{aligned}$$