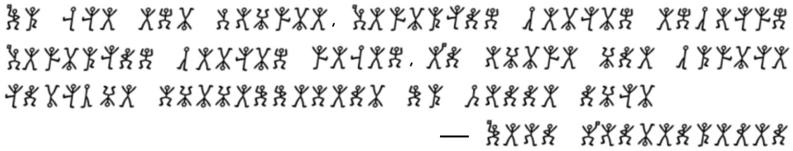
# Lecture 2. Data Reduction & Estimation Evaluation

[1]School of Data Science, The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen)

## 1. Data Reduction



Within the notes, we write $X = \{X_1, \cdots, X_n\}$ as a random sample from the population $F_\theta$ (associated with pdf or pmf $f_\theta$) unless otherwise specified.

*Definition* 1.1 (♣ **Statistic**). A statistic $T = T(X) = T(X_1, \cdots, X_n)$ is a function of the data $X$ that does not depent upon any unknown parameters.

Apparently, $T$ is a function of random variables, so $T = T(X)$ itself is a random variable. It will generally depend on the sample size $n$, and when we observed $X = x$, where $x = (x_1, \cdots, x_n)$ is a realization of $X = (X_1, \cdots, X_n)$, we would write the realization of $T(X)$ as $T(x)$.

### *1.1. Sufficiency*

#### *1.1.1. Definition of sufficient statistics*

When unheralded "Big Data" strike this world, one simple question we wanna answer ourselves is, can we find a statistic $T = T(X)$ for inferring the parameter of interests, say $\theta$, s.t., knowing $T$ is "equivalent" to knowing the whole sample $X = \{X_1, \cdots, X_n\}$, while $T$ is ideally costing less in storage? In other words, we are hoping to find a statistic $T$ such that all information about the parameter of interests $\theta$ within the sample $X = \{X_1, \cdots, X_n\}$ are stored in $T$. And this intuitive data reduction procedure is related to the following sufficiency principle and sufficient statistics.

*Definition* 1.2 (**Sufficiency Principle**). If $T(X)$ is a "sufficient statistic" for $\theta$, then any inference about $\theta$ should depend on the sample $X$ only through the value $T(X)$. That is , if $x$ and $y$ are two sample points, s.t. $T(x) = T(y)$, then the inference about $\theta$ should be the same whether $X = x$ or $X = y$ is observed.

*Definition* 1.3 (♣ **Sufficient Statistics**). A statistic $T(X)$ is a sufficient statistic for $\theta$ if the conditional distribution of $\big[X|T(X)\big]$ does not depend on $\theta$.

• *Example* 1.4 (♣ **Normal Sufficient Statistic**). Let $X_1, \cdots, X_n \sim_{i.i.d} N(\mu, \sigma^2)$. Show that the sample mean, $T(X) = \bar{X}$ is a sufficient statistic for $\mu$.

*Definition* 1.5 (♣ **Rank Statistics**). For a random sample $X = \{X_1, X_2, \cdots, X_n\}$ from a continuous population $F$ (i.e., we require $\mathbb{P}(X_i = X_j) = \mathbb{1}(i = j)$), define

$$R_i = \sum_{j=1}^{n} \mathbb{1}(X_j \le X_i), \quad i = 1, \cdots, n.$$

$R = (R_1, R_2, \cdots, R_n)$ is called the rank statistics of $X = \{X_1, X_2, \cdots, X_n\}$.

*Definition* 1.6 (♣ **Order Statistics**). For a random sample $X = \{X_1, X_2, \cdots, X_n\}$ from a continuous population $F_\theta$ (associated with pdf or pmf $f_\theta$) and denote its corresponding rank statistics as $R = (R_1, R_2, \cdots, R_n)$, define

$$X_{(i)} = \sum_{j=1}^{n} X_j \cdot \mathbb{1}(R_j = i), \quad i = 1, \cdots, n.$$

$X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ is called the order statistics of $X$.

**Theorem 1.7.** *Define*

$$\mathcal{R} = \Big\{(r_1, \cdots, r_n) \mid (r_1, \cdots, r_n) \text{ is a permutation of the integers } (1, 2, \cdots, n)\Big\}$$

*Obviously, $|\mathcal{R}| = n!$. Then the rank statistics $R = (R_1, \cdots, R_n)$ is uniformly distributed on $\mathcal{R}$, i.e., for $\forall r \in \mathcal{R}$, $\mathbb{P}(R = r) = 1/n!$.*

**Theorem 1.8.** *The joint density function of $(X_{(1)}, X_{(2)}, \cdots, X_{(n)})$, denoted as $p_\theta(\cdot)$, is given by*

$$p_\theta(x_{(1)}, \cdots, x_{(n)}) = n! \cdot \prod_{i=1}^{n} f_\theta(x_{(i)}) \cdot \mathbb{1}(x_{(1)} \le \cdots \le x_{(n)}). \tag{1.1}$$

- *Example* 1.9. (♣ **Sufficient statistics always exists!**)
  (i) (Whole Sample) Apparently, sample $X$ is a sufficient statistic since the conditional distribution of $\big[X|X\big]$ is irrelavant to $F_\theta$. This is an "useless" sufficient statistic cause it didn't really help with data reduction.
  (ii) (Order Statistics) Cause the condition density of $\Big[X|X_{(1)},\cdots,X_{(n)}\Big]$ is

$$\begin{aligned}&f_\theta(x_1,\cdots,x_n)/p_\theta(x_{(1)},\cdots,x_{(n)})\\=&(n!)^{-1}\cdot\mathbb{1}(x_1,\cdots,x_n\text{ is a permutation of }x_{(1)}<\cdots<x_{(n)}),\end{aligned}$$

  which does not depend on the unknown parameter $\theta$, so by definition order statistics $X_{(1)},\cdots,X_{(n)}$ is a sufficent statistic.

### *1.1.2. Methods of finding sufficient statistics*

**Theorem 1.10** (♣ **Neyman-Fisher Factorization Theorem**)**.** *Suppose $X_1$, $\cdots,X_n$ have joint pdf. or pmf. $f(x|\theta)$, $\theta\in\Theta$. Let $T=T(X)$ be a statistic. Then*

$$T\text{ is sufficient for }\theta\quad\Leftrightarrow\quad\forall\,\theta\in\Theta,x\in\mathcal{X}_n,\ f(x|\theta)=g(T(x),\theta)h(x),\tag{1.2}$$

*where $g(\cdot)$ and $h(\cdot)$ are functions such that $g(t,\theta)$ depends on $x$ only through $t=T(x)$, and $h(x)$ is invariant over $\theta$.*

*Proof.* (of Theorem.1.10: Neyman-Fisher Factorization Theorem) We prove the FactorizationTheorem for discrete random variables as an illustration, the general proof follow the same line.

- Suppose $T$ is sufficient. We want to prove the right hand side of (1.2). Let $t=T(x)$. The joint pmf. of $X$ is

$$\begin{aligned}f(x|\theta)&=\mathbb{P}_\theta(X=x)=\mathbb{P}_\theta(\{X=x\}\bigcap\{T=t\})\\&=\mathbb{P}_\theta(X=x|T=t)\mathbb{P}_\theta(T=t)=:h(x)g(t,\theta).\end{aligned}$$

- Suppose the right hand side of (1.2) holds. We want to prove $T$ is sufficient. Apparently, when $T(x)\neq t$, we have $\mathbb{P}_\theta(X=x|T=t)=0$ by definition, which is invariant over $\theta$. Meanwhile, when $T(x)=t$. Let $S=\{x'\in\mathcal{X}_n:T(x')=t\}$. Then

$$\begin{aligned}\mathbb{P}(X=x|T=t)&=\frac{\mathbb{P}_\theta(\{X=x\}\bigcap\{T=t\})}{\mathbb{P}_\theta(T=t)}=\frac{\mathbb{P}_\theta(\{X=x\})}{\mathbb{P}_\theta(T=t)}\\=&\frac{\mathbb{P}_\theta(\{X=x\})}{\{\sum_{x'\in S}\mathbb{P}_\theta(X=x')\}}=\frac{g(t,\theta)h(x)}{\{\sum_{x'\in S}g(t,\theta)h(x')\}}=\frac{h(x)}{\{\sum_{x'\in S}h(x')\}}\end{aligned}$$

  which is invariant over $\theta$, and conclude that $T$ is sufficient by definition.

□

**Lemma 1.11.** *Suppose $X_1, \cdots, X_n$ have joint pdf. or pmf. $f(x|\theta)$, $\theta \in \Theta$. Then, a statistic $T = T(X)$ is sufficient if $T(x) = T(x')$ implies that $f(x|\theta)/f(x'|\theta)$ is invariant over $\theta$ for arbitrary $x, x' \in \Omega$, i.e.,*

$$\left\{ T(x) = T(x') \;\Rightarrow\; \frac{f(x|\theta)}{f(x'|\theta)} \;\; \textit{is invariant over}\, \theta, \text{for } \forall x, x' \} \right\} \;\Rightarrow\; T \textit{ is sufficient for } \theta.$$

A proof of Lemma.1.11 is postponed to Supplement.

• *Example* 1.12 (♣ **Poisson Sufficient Statistic**). Let $X_1, \cdots, X_n \overset{i.i.d}{\sim} Poisson(\lambda)$, please find a sufficient statistic for $\lambda$.

• *Example* 1.13 (♣ **Bernoulli Sufficient Statistic**). Let $X_1, \cdots, X_n \sim_{i.i.d}$ $Bernoulli(p)$, where $p \in (0, 1)$, please find a sufficient statistic for $\lambda$.

• *Example* 1.14 (♣ **Normal Sufficient Statistic**). Let $X_1, \cdots, X_n \overset{i.i.d}{\sim} N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$. Please find a sufficient statistic for the unknown parameter $\theta$ for the following three cases separately.

(i) $\mu$ is unkown and $\sigma^2$ is known, $\theta = \mu$.

(ii) $\mu$ is kown and $\sigma^2$ is unknown, $\theta = \sigma^2$.
(iii) Both $\mu$ and $\sigma^2$ are unknown, $\theta = (\mu, \sigma^2)$.

• *Example* 1.15 (♣ **Uniform Sufficient Statistic**). Let $\{X_i\}_{i=1}^n \overset{i.i.d}{\sim}$ Uniform$(0, \theta)$, where $\theta \in \mathbb{R}^+$. please find a sufficient statistic for $\theta$.

**Theorem 1.16** (♣ **One-to-One Mapping preserves Sufficiency**). *Let $X_1$, $\cdots, X_n$ have joint pdf. or pmf. $f(x|\theta)$, $\theta \in \Theta$. Suppose $T = T(X) \in \mathbb{R}^{k_1}$ is a sufficient statistic for $\theta$. If $\psi : \mathbb{R}^{k_1} \to \mathbb{R}^{k_2}$ is a one-to-one function and free of $\theta$, then $S(X) = \psi(T(X))$ is also sufficient for $\theta$.*

A proof of Theorem.1.16 is postponed to Supplement.

### 1.1.3. Minimal Sufficiency

Notice that for a sample $X = \{X_1, \cdots, X_n\}$ and an estimating problem, there are actually many sufficient statistics. For example, any one-to-one function of a sufficient statistic is a sufficient statistic. Therefore, it is a natural question that whether one sufficient statistic is any better than another. Recall that the purpose of sufficient statistic is to achieve data reduction without loss of information about the parameter $\theta$. Thus, a statistic that achieves the most data reduction while retaining all the information about $\theta$ might be considered preferable, and the following definition serves the purpose,

*Definition* 1.17 (♣ **Minimal Sufficient Statistic**). A sufficient statistic $T(X)$ is called a minimal sufficient statistic if, for any other sufficient statistic $T'(X)$, $T(X)$ is a function of $T'(X)$, i.e., for arbitrary two sample points $x$ and $y$, whenever $T'(x) = T'(y)$ implies $T(x) = T(y)$.

**Theorem 1.18** (♣ **Lehmann-Scheffé's Theorem**). *Suppose $X_1, \cdots, X_n$, have joint pdf. or pmf. $f(x|\theta)$, $\theta \in \Theta$. Then, a statistic $T = T(X)$ is minimal sufficient if $T(x) = T(x')$ is equivalent to $f(x|\theta)/f(x'|\theta)$ being invariant over $\theta$ for arbitrary $x, x' \in \Omega$, i.e.,*

$$\left\{ T(x) = T(x') \Leftrightarrow \frac{f(x|\theta)}{f(x'|\theta)} \ \textit{is invariant over}\, \theta, \text{for } \forall x, x' \right\} \ \Rightarrow \ T \textit{ is minimal sufficient for } \theta.$$

A proof of Lehmann-Scheffé's Theorem.1.18 is postponed to Supplement. Notice that the statement of Lehmann-Scheffé's Theorem.1.18, is highly resemble to Lemma 1.11.

- *Example* 1.19 (♣ **Minimal Sufficient Statistic of Uniform Distribution**). Let $X = \{X_1, \cdots, X_n\}$ be a random sample from Uniform$(\theta, \theta+1)$, where $\theta \in \mathbb{R}$. Please find a minimal sufficient statistic for $\theta$.

- *Example* 1.20 (♣ **Minimal Sufficient Statistic for Cauchy Distribution**). Let $X_1, \cdots, X_n \overset{i.i.d}{\sim} Cauchy(\theta, 1)$, where $\theta \in \mathbb{R}$. Prove that $T = T(X) = (X_{(1)}, \cdots, X_{(n)})^T$ is a minimal sufficient statistics for $\theta$.

### 1.2. Ancillary

*Definition* 1.21 (♣ **Ancillary Statistic**). A statistic $S(X)$ is called an ancillary statistic of $\theta$ if the distribution of $S(X)$ does not depend on parameter $\theta$.

From the definition, one would think the ancillary statistics can be discarded since itself does not contain any information about the parameter of interests ($\theta$). However, this isn't the case. Sometimes, ancillary statistic can be an "ancillary" to other statistics in estimating $\theta$, e.g.,

• *Example* 1.22. Recall Example.1.19, for a random sample $X = \{X_1, \cdots, X_n\}$ from Uniform$(\theta, \theta+1)$, $\theta \in \mathbb{R}$, we have $T(X) = (X_{(1)}, X_{(n)})$ is a minimal sufficient statistic for $\theta$. Since there is a one-to-one mapping between $T(X) = (X_{(1)}, X_{(n)})$ and $T'(X) = (X_{(n)} - X_{(1)}, X_{(1)} + X_{(n)})$, so $T'(X)$ is also a minimal sufficient statistic for $\theta$.

However, if we look at $S(X) = X_{(n)} - X_{(1)}$. We may define $Z_i = X_i - \theta \sim$Uniform$(0,1)$, and $X_i \leq X_j \Leftrightarrow Z_i \leq Z_j$, so $Z_{(n)} = X_{(n)} - \theta$, $Z_{(1)} = X_{(1)} - \theta$. Hence $S(X) = X_{(n)} - X_{(1)} = Z_{(n)} - Z_{(1)}$. Since $\{Z_{(i)}, 1 \leq i \leq n\}$ are order statistics of $\{Z_i, 1 \leq i \leq n\}$ whose distribution does not depend on $\theta$, so the distribution of $S(X)$ also does not depend on $\theta$, i.e., $S(X)$ is an ancillary statistic.

It's transparent that subtract the ancillary statistic $S(X)$ from the minimal sufficient statistic $T'(X)$ (which leaves us with $X_{(1)} + X_{(n)}$) would not even leads to a sufficient statistic.

• *Example* 1.23. (♣ **Ancillary statistics in location scale family**) Consider a random sample $X = \{X_1, X_2, \cdots, X_n\}$ from the population $F\left(\frac{x-\theta}{\eta}\right)$ with unknown $\gamma = (\theta, \eta) \in \{\theta \in \mathbb{R}, \eta > 0\}$ and a known distribution function $F(\cdot)$. Please show that the statistic

$$T(X_1, \cdots, X_n) = \frac{X_{(2)} - X_{(1)}}{X_{(n)} - X_{(1)}},$$

is an ancillary statistic of $\gamma$.

### 1.3. Completeness

From a historical point of view, the study of the completeness of a statistic, is a derivant of the study of some property called the “first order ancillary” of a statistic. And from the name we would know that the “completeness” should not be something far away from the “ancillary”. Meanwhile, from a geometrical point of view, the “completeness” is deeply connected with orthogonality.

As a matter of fact, a sufficient statistic $T$ appears to be most successful in reducing the data if it is a sufficient complete statistic. And consider its extraordinary effectiveness in reducing the data, it is not surprising that a sufficient complete statistic is always minimal (provided that minimal sufficient statistics exists, see the following Bahadur's Thorem.1.24).

**Theorem 1.24** (♣ **Bahadur's Theorem**). *If a minimal sufficient statistic exists, then any sufficient complete statistic is also a minimal sufficient statistic.*

A proof of Thorem.1.24 is provided in Bahadur (1957).

Formally, the completeness of a statistic or the completeness of a distribution family is defined as

*Definition* 1.25 (♣ **Complete Statistics**). A statistic $T$ is call a complete statistic if arbitrary function $\phi$ statisfying $\mathbb{E}_\theta\left[\phi(T)\right] = 0$ for $\forall\theta \in \Theta$ could imply $\mathbb{P}_\theta\big(\phi(T) = 0\big) = 1$, i.e.,

$$\text{If } \left\{\phi : \mathbb{E}_\theta\left[\phi(T)\right] = 0,\ \text{for } \forall\theta \in \Theta\right\} \subset \left\{\phi : \mathbb{P}_\theta\big(\phi(T) = 0\big) = 1\right\},\ \text{Then } T \text{ is complete.}$$

More rigorously, instead of saying $T$ is complete, we say the distribution family $\{f_T(\cdot|\theta) : \theta \in \Theta\}$ is complete.

- *Example* 1.26 (♣ **Complete Statistic for Binomial Distribution**). Suppose $\{X_i\}_{i=1}^m$ are mutually independent sample with $X_i$ came from Binomial$(n_i, p)$ and $\sum_{i=1}^m n_i = n$ with $n_i$ and $n$ being known. Please show $T(X) = \sum_{i=1}^m X_i$ is a complete statistic of $p$.

- *Example* 1.27 (♣ **Complete Statistic for Uniform Distribution**). Assume $X = \{X_1, \cdots, X_n\}$ is a random sample from $U(0, \theta)$. Please show $T(X) = X_{(n)} = \max_{\{1 \le i \le n\}} X_i$ is a complete statistic.

### *1.4. Connections between Sufficiency, Ancillary, Completeness and Exponential family*

Recall the canonical form of an exponential family is

$$f(x|\theta) = h(x)b(\omega)\exp\left(\sum_{i=1}^{k} \omega_i T_i(x)\right), \quad \Theta = \Big\{(\omega_1, \cdots, \omega_k) : b(\omega) \ge 0\Big\}. \tag{1.3}$$

**Theorem 1.28** (♣ **Minimal Sufficient and Complete Statistics in Exponential Family**). *If $X = \{X_1, \cdots, X_n\}$ is distributed according to the exponential family (1.3) and the family is of full rank, i.e., dim$(\Theta) = k$, then $T(X) = (T_1(X), \cdots, T_k(X))$ is minial sufficient and complete.*

A proof of Thorem.1.28 is provided in Barndorff-Nielsen (2014).

Using Theorem 1.28, we may look at some of the examples discussed before.

- *Example* 1.29 (♣ **Minimal Sufficient and Complete Statistic of Normal Distribution**). Now, if we revisit Example.1.14, please find a minimal sufficient and complete statistic for each of the cases separately.

vskip25e'm

- *Example* 1.30. Let $X_1, \cdots, X_n \overset{i.i.d}{\sim} f(x|\theta)$, where

$$f(x|\theta) = \frac{\theta}{(1+x)^{1+\theta}} \mathbb{1}(x \in \mathbb{R}^+),$$

and $\theta \in \mathbb{R}^+$. Please find a minimal sufficient and complete statistic for $\theta$.

* We are, not only been able to find minimal sufficient and complete statistics from an exponential family, but also been able to identify exponential families (for a family of continuous distributions indexed by $\theta$, a parameter of interests, who have a $\theta$-irrelevant support) if there exists a continuous sufficient statistic, i.e., see the following Barndorff-Nielson & Pedersen's Representation Theorem 1.31, that is saying, exponential family are the only ones that permit dimensional reduction of the sample through sufficiency.

*** Theorem 1.31** (Barndorff-Nielson & Pedersen's Representation Theorem)**.** *Suppose $X_1, \cdots, X_n$ are real-valued i.i.d according to a distribution with density $f(x|\theta)$ with respect to Lebesgue measure, which is continuous in $x$ and whose support for all $\theta$ is an interval $I$. Suppose that for the joint density of $X = (X_1, \cdots, X_n)$ is*

$$p(x|\theta) = \prod_{i=1}^{n} f(x_i|\theta),$$

*there exists a continuous $k-$dimensional sufficient statistic. Then*

*(i) If $k = 1$, then exist functions $\omega_1(\theta)$, $b(\cdot)$ and $h(\cdot)$ such that (1.3) holds;*

*(ii) If $k > 1$, and if the densities of $f(x_i|\theta)$ have continuous partial derivatives with respect to $x_i$, then there exist functions $\omega_i(\theta)$, $b(\cdot)$ and $h(\cdot)$ such that (1.3) holds with some $s \leq k$.*

A proof of Thorem.1.31 is provided in Barndorff-Nielsen and Pedersen (1968).

### *1.5. Applications of Sufficient, Ancillary, Complete Statistics*

Since the "completeness" is deeply connected with orthogonality from a geometrical point of view, so a useful result implementing this idea who reveals the relationship between an ancillary statistic and a sufficient complete statistic is the following Basu's Theorem.

**Theorem 1.32** (♣ **Basu's Theorem**)**.** *If $T$ is a sufficient complete statistic for the family $\mathcal{F} = \{f(\cdot|\theta), \theta \in \Theta\}$, then any ancillary statistic $V$ of $\theta$ is independent of $T$.*

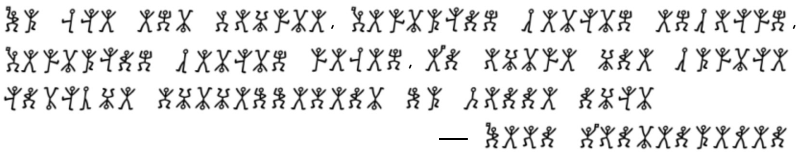
A proof of Basu's Thorem.1.32 is postponed to Supplement.

• *Example* 1.33 (♣ **Independence of Sample Mean and Sample Variance in Normal Distribution**). We again revisit Example.1.14, please prove that, for a normal random sample, $\bar{X} = \sum_{i=1}^{n} X_i/n$ is independent of $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$.

• *Example* 1.34 (♣ **Basu Theorem in Moment Calculation**). Let $\{X_i\}_{i=1}^{n+1} \overset{i.i.d}{\sim}$ exponential($\lambda$), with density $f(x_i|\lambda) = \lambda e^{-\lambda x_i} \cdot \mathbb{1}(x_i \geq 0)$. Please calculate the mean of $g(X) = (\sum_{i=j}^{k} X_i)/(\sum_{i=1}^{n+1} X_i)$.

• *Example* 1.35. Let's revisit Example 1.19, where we have $X = \{X_1, \cdots, X_n\}$ ia a random sample from Uniform $(\theta, \theta+1)$., and we have found that $T(X) = (X_{(1)}, X_{(n)})$ is a minimal sufficient statistic for $\theta$. Please show that $T$ is not complete, i.e., the inverse Bahadur's Theorem.1.24 does not holds.

## 2. Estimation Evaluation



The problem of point estimation aims to specify a plausible value for $\theta$.

*Definition* 2.1 (♣ **Estimand, Estimator and Estimate**). For a sample $X$ from $f(\cdot|\theta)$, let $\hat{\theta} = T(X)$ be a sataitstic. Then the parameter of interests $\theta$ is referred to as estimand; the statistic $\hat{\theta} = T(X)$ we used as the plausible value for $\theta$ is referred to as estimator; the realization of the estimator, i.e., when $X = x$ is observed, the realized value $T(x)$ is referred to as estimate.

In the remain of the section, we always consider $X$ is a sample from $f(\cdot|\theta)$, we have a estimator $\hat{\theta}$ for the estimand $\theta$ unless otherwise specified.

### *2.1. Evaluation of the Estimation: Bias and Unbiasedness*

*Definition* 2.2 (♣ **Estimand, Estimator and Estimate**). For a sample $X \sim f(x|\theta)$, denote $\widehat{\omega(\theta)}$ as our estimator for the estimand $\omega(\theta)$ with some known function $\omega(\cdot)$. The bias of $\widehat{\omega(\theta)}$ with respect to $\omega(\theta)$ is

$$Bias(\widehat{\omega(\theta)}) = \mathbb{E}\left[\widehat{\omega(\theta)}\right] - \omega(\theta),$$

and we say $\widehat{\omega(\theta)}$ is unbiased with respect to $\omega(\theta)$ if $Bias(\widehat{\omega(\theta)}) = 0$.

*Remark* 2.3. Bias ($Bias(\hat{\theta})$) of a estimator ($\hat{\theta}$) has a sign, despite the fact that the sign could could change along with the change of estimand value ( $\theta$). If it's always negative (i.e., $Bias(\hat{\theta}) < 0$ for all $\theta \in \Theta$), then we say this estimator ($\hat{\theta}$) underestimates the estimand ($\theta$). If it's always negative (i.e., $Bias(\hat{\theta}) > 0$ for all $\theta \in \Theta$), then we say this estimator ($\hat{\theta}$) overestimates the estimand ($\theta$).

### *2.2. Evaluation of the Estimation: Loss Funtion and Risk Function*

*Definition* 2.4 (♣ **Loss Function and Risk Function**). For a estimate $\hat{\theta}(x) = T(x)$ of $\theta$, the loss function is a function $L(\theta, T(x))$, such that

$$L(\theta, T(x)) \geq 0 \text{ for all } \theta \in \Theta, x \in \mathcal{X}, \text{ and } L(\theta, \theta) = 0 \text{ for all } \theta \in \Theta.$$

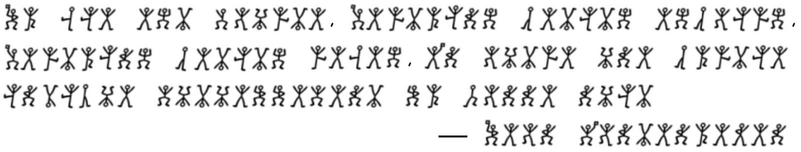
And the corresponding risk function is defined as the function

$$R(\theta, \hat{\theta}) = \left[R(\theta, \hat{\theta}(\cdot))|F_X\right] = \mathbb{E}_\theta\left[L(\theta, T(X))\right].$$

• *Example* 2.5. The mean square error (MSE) of an estimator, i.e., $\mathbb{E}(\hat{\theta}-\theta)^2$, is the risk function corresponding to the quadratic loss $L(a,b)=(a-b)^2$. Other commonly used loss such as absolute error loss $L(a,b)=|a-b|$, cross-entropy loss $L(f,g)=\int\big\{\log[f(x)/g(x)]\big\}f(x)dx$, $L_p$ loss $L(a,b)=|a-b|^p$, Huber loss, elastic net loss, etc.

## 3. *Supplement

Here we list some of the proofs of theorems listed before.



*Proof.* (of Theorem.1.8) Since $X_i = X_{(R_i)}$, $1\le i\le n$ and $\cup_{r\in\mathcal{R}}\{R=r\}$ is a partition of the sample space, therefore, for some $0<\epsilon<\frac{1}{2}\min\{x_{(i+1)}-x_{(i)}: i=1,\cdots,n\}$,

$$\begin{aligned}
&\mathbb{P}\Big(x_{(1)}-\epsilon\le X_{(1)}\le x_{(1)},\cdots,x_{(n)}-\epsilon\le X_{(n)}\le x_{(n)}\Big)\\
=&\sum_{r\in\mathcal{R}}\mathbb{P}\Big(x_{(1)}-\epsilon\le X_{(1)}\le x_{(1)},\cdots,x_{(n)}-\epsilon\le X_{(n)}\le x_{(n)},R=r\Big)\\
=&\sum_{r\in\mathcal{R}}\mathbb{P}\Big(x_{(r_1)}-\epsilon\le X_1\le x_{(r_1)},\cdots,x_{(r_n)}-\epsilon\le X_n\le x_{(r_n)}\Big)\cdot\mathbb{1}(x_{(1)}<\cdots<x_{(n)})\\
=&\sum_{r\in\mathcal{R}}\prod_{i=1}^{n}\big[F_\theta(x_{(r_i)})-F_\theta(x_{(r_i)}-\epsilon)\big]\cdot\mathbb{1}(x_{(1)}<\cdots<x_{(n)})\\
=&\sum_{r\in\mathcal{R}}\prod_{i=1}^{n}\big[F_\theta(x_{(i)})-F_\theta(x_{(i)}-\epsilon)\big]\cdot\mathbb{1}(x_{(1)}<\cdots<x_{(n)})\\
=&n!\cdot\prod_{i=1}^{n}\big[F_\theta(x_{(i)})-F_\theta(x_{(i)}-\epsilon)\big]\cdot\mathbb{1}(x_{(1)}<\cdots<x_{(n)})
\end{aligned}$$

Therefore,

$$\begin{aligned}
p_\theta(x_{(1)},\cdots,x_{(n)})&=\lim_{\epsilon\to 0}\frac{1}{\epsilon}\mathbb{P}\Big(x_{(1)}-\epsilon\le X_{(1)}\le x_{(1)},\cdots,x_{(n)}-\epsilon\le X_{(n)}\le x_{(n)}\Big)\\
&=n!\cdot\prod_{i=1}^{n}f_\theta(x_{(i)})\cdot\mathbb{1}(x_{(1)}<\cdots<x_{(n)}),
\end{aligned}$$

holds for arbitrary $x_{(1)}<\cdots<x_{(n)}$. Notice that since $F_\theta$ is continuous, so the order statistics has probability 0 in the set $\{x_{(i)}=x_{(j)},\text{ for some } i\ne j\}$. So (1.1) holds. □

*Proof.* (of Lemma.1.11) To simplify the proof, we assume $f(x|\theta) > 0$ for all $x \in \mathcal{X}$ and $\theta$. Define $\mathcal{T} = \{t : t = T(x), x \in \mathcal{X}\}$, so we may partition the sample space $\mathcal{X}$ to

$$\mathcal{X} = \bigcup_{t\in\mathcal{T}} \{x : T(x) = t\} \triangleq \bigcup_{t\in\mathcal{T}} A_t.$$

Now, for each $A_t$, $t \in \mathcal{T}$, we may choose one element from $A_t$ to be the representative element of $A_t$. Since the representative element is choosed for each $A_t$, the set indexed by $t \in \mathcal{T}$, so we can denote the representative element of $A_t$ as $\phi(t)$, $t \in \mathcal{T}$, and by definition $\phi(t) \in A_t$, so $T(\phi(t)) = t$.

Now, for arbitrary $x \in \mathcal{X}$, apparently, there exists some $t$ such that $T(x) = t$. So $T(x) = T(\phi(t))$, hence

$$\frac{f(x|\theta)}{f(\phi(t)|\theta)} = \frac{f(x|\theta)}{f(\phi(T(x))|\theta)} \text{ is invariant over } \theta, \text{ and we denote it as } h(x).$$

Thus, by define $g(t, \theta) = f(\phi(t)|\theta)$, we conclude

$$f(x|\theta) = f(\phi(t)|\theta) \cdot \frac{f(x|\theta)}{f(\phi(t)|\theta)} = g(t, \theta) \cdot h(x),$$

by Factorization theorem, we conclude that $T$ is a sufficient statistic for $\theta$. □

*Proof.* (of Theorem.1.18: Lehmann-Scheffé's Theorem) Again, to simplify the proof, we assume $f(x|\theta) > 0$ for all $x \in \mathcal{X}$ and $\theta$.

Due to the highly resembles between Lemma.1.11 and the Lehmann-Scheffé's Theorem.1.18. We have already proved in Lemma.1.11 that

$$\left\{T(x) = T(x') \;\Rightarrow\; \frac{f(x|\theta)}{f(x'|\theta)} \text{ is invariant over } \theta, \text{for } \forall x, x'\}\right\} \;\Rightarrow\; T \text{ is sufficient for } \theta.$$

Therefore, here, we only have to investigate the event set

$$\left\{T(x) = T(x') \;\Leftarrow\; \frac{f(x|\theta)}{f(x'|\theta)} \text{ is invariant over } \theta, \text{for } \forall x, x'\}\right\} \bigcap \left\{T \text{ is sufficient for } \theta\right\},$$

and see whether it would imply that $T$ is a minimal sufficient statistic of $\theta$.

Now, say we have another sufficient statistic $S(X)$,

(i) if there doesn't exists two sample points $x, x' \in \mathcal{X}$ fow which $S(x) = S(x')$, then $x$ is a function of $S(x)$, hence $S(x) = x$, i.e., $S(X) = X$. In which case, we have $T = T(S(X))$ obviously.

(ii) If there exists two sample points $x, x' \in \mathcal{X}$ such that $S(x) = S(x')$. Then by Factorization theorem, there exist some function $g(s, \theta)$ and $h(x)$ such that

$$\frac{f(x|\theta)}{f(x'|\theta)} = \frac{g(S(x), \theta)h(x)}{g(S(x'), \theta)h(x')} = \frac{h(x)}{h(x')}, \text{ which is invariant over } \theta.$$

which means that under the interested event set, we also have $T(x) = T(x')$, i.e.,

$$S(x) = S(x') \Rightarrow T(x) = T(x'), \text{ for } \forall x, x' \in \mathcal{X}.$$

That is saying $T$ is a function of $S(X)$, i.e., $T = T(S(X))$.

Since $T$ is also a sufficient statistic, therefore, combine these two case concludes that $T$ is a minimal sufficient statistic for $\theta$. □

*Proof.* (of Theorem.1.16) Let $t = T(x)$ and $s = S(x) = \psi(t)$. Since $\psi$ is one-to-one, there exists an inverse function $\psi^{-1}(s)$. Using the sufficiency of $T$, we have, by the Factorization Theorem, that

$$f(x|\theta) = g(t, \theta)h(x) = g(\psi^{-1}(s), \theta)h(x) = \tilde{g}(s, \theta)h(x),$$

where $\tilde{g}(s, \theta) := g(\psi^{-1}(s), \theta)$. By the Factorization Theorem again, we know $S$ is also sufficient for $\theta$, hence sufficiency is preserved under one-to-one mapping. □

*Proof.* (of Theorem.1.32: Basu's Theorem) Since $V$ is ancillary, the probability $p_A = \mathbb{P}(V \in A)$ is independent of $\theta$ for all $A \in \mathcal{F}$. Define

$$\phi_A(t) = \mathbb{P}(V \in A|T = t) - p_A,$$

since $T$ is sufficient, so $\phi_A(\cdot)$ still does not depend on $\theta$, and

$$\mathbb{E}\phi_A(T) = 0$$

for all $\theta \in \Theta$. Notice that $T$ is also a complete statistic, therefore,

$$\mathbb{P}(\phi_A(T) = p_A) = 1, \ \ i.e., \ \ \mathbb{P}(V \in A|T) = \mathbb{P}(V \in A), \ a.s.,$$

Hence the independence of $V$ and $T$. □

## References

Bahadur, R. R. (1957). On unbiased estimates of uniformly minimum variance. Sankhyā: The Indian Journal of Statistics (1933-1960), 18(3/4), 211-224.

Barndorff-Nielsen, O. (2014). Information and exponential families: in statistical theory. John Wiley & Sons.

Barndorff-Nielsen, O. L. E., & Pedersen, K. (1968). Sufficient data reduction and exponential families. Mathematica Scandinavica, 22(1), 197-202.