Topic III:
Principles of Data Reduction

Wei You

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Fall, 2023

Data Reduction

In statistics, one of the central tasks is to turn the large amount of data in a sample $\mathbf{X} = \{X_1, \dots, X_n\}$ into inferences about the world, e.g. an unknown parameter θ in a family of distributions.

- Do we have to keep ALL data in order to make a good inference?

Example: Consider a Uniform$([0, \theta])$ random sample, suppose we observed the following data

$$[2.16, 0.72, 9.75, 0.89, 2.21].$$

What can we say about θ?

Not all data are relevant to a particular statistical problem.

- A data reduction procedure that discards irrelevant data $\Rightarrow$ results in a simpler inference procedure.

Definition (Statistic)

A **statistic** $T(\mathbf{X})$ is a function of the data. It does not depend on any unknown parameters.

Example: In the Uniform$([0, \theta])$ example, $X_{(n)} = \max\{X_1, \dots, X_n\}$ is a statistic.

- If T is not one-to-one, it defines a form of data reduction.
- A "good" statistic should preserve information about the unknown parameter θ.

Key question: Is there a statistic that contains all the information in the sample about θ?

If so, a reduction or compression of the original data to this statistic without loss of information is possible.

Sufficient Statistics

Definition (Sufficient Statistics)

A statistic is sufficient[a] for a model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ if for any given t, the conditional distribution of X under $T(\mathbf{X}) = t$ does not depend on θ.

[a]This concept was introduced by R. A. Fisher in 1922.

- The concept of sufficiency depends on the model $\mathcal{P}$, i.e., the parameter θ.
- Intuitively speaking, if we know the value of a sufficient statistic T, then we can do just as good a job of estimating the unknown parameter θ as someone who knows the entire data.

To see this, we can consider the full data as “dummy” data generated using T.

- Instead of directly simulating $\mathbf{X}$, we are given an observation

$$T(\mathbf{X}) \sim \mathbb{P}_\theta(T(\boldsymbol{X}) = t).$$

- We then generate independent conditional r.v. $\mathbf{X}|T(\mathbf{X})$, which has the same distribution as the full data:

$$\mathbb{P}_\theta(\boldsymbol{X} = \boldsymbol{x}) = \mathbb{P}_\theta(\boldsymbol{X} = \boldsymbol{x} \mid T(\boldsymbol{X}) = T(\boldsymbol{x}))\mathbb{P}_\theta(T(\boldsymbol{X}) = T(\boldsymbol{x})).$$

- By sufficiency, $\mathbf{X}|T(\mathbf{X})$ does not depend on θ, all the information about θ is contained in $T(\mathbf{X})$.

Examples

Example: Bernoulli. Let $X_1, \dots, X_n$ be random sample from Bernoulli(θ). Is the number of heads $T = \sum X_i$ sufficient?

$$\mathbb{P}_\theta(\mathbf{X} = \mathbf{x}) = \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}$$

$$\begin{aligned}
\mathbb{P}_\theta(\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = t) &= \frac{\mathbb{P}_\theta(\mathbf{X} = \mathbf{x}, T(\mathbf{X}) = t)}{\mathbb{P}_\theta(T(\mathbf{X}) = t)} \\
&= \frac{\theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}\mathbb{1}(t = \sum x_i)}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\
&= \frac{\mathbb{1}(t = \sum x_i)}{\binom{n}{t}}, \quad \text{for all } x_i \in \{0, 1\}.
\end{aligned}$$

This does not depend on θ, by definition $\sum X_i$ is sufficient for θ.

How to understand $\mathbb{P}_\theta(\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = t)$ here?

Example: Uniform($[0, \theta]$). Conditioning on $T(\mathbf{X}) = X_{(n)} = t$, the remaining $n-1$ numbers behave like random sample from Uniform($[0, t]$), independent of θ.

$$
\begin{aligned}
&\mathbb{P}_\theta(X_1 \le x_1, \dots, X_{n-1} \le x_{n-1} \mid X_n = X_{(n)} = t) \\
&= \frac{\mathbb{P}_\theta(X_1 \le x_1, \dots, X_{n-1} \le x_{n-1}, X_n = X_{(n)} = t)}{\mathbb{P}_\theta(X_n = X_{(n)} = t)} \\
&= \frac{\mathbb{P}_\theta(X_1 \le x_1 \wedge t, \dots, X_{n-1} \le x_{n-1} \wedge t, X_n = t)}{\mathbb{P}_\theta(X_1 \le t, \dots, X_{n-1} \le t, X_n = t)} \\
&= \frac{\mathbb{P}_\theta(X_1 \le x_1 \wedge t) \dots \mathbb{P}_\theta(X_{n-1} \le x_{n-1} \wedge t)\mathbb{P}_\theta(X_n = t)}{\mathbb{P}_\theta(X_1 \le t) \dots \mathbb{P}_\theta(X_{n-1} \le t)\mathbb{P}_\theta(X_n = t)} \\
&= \prod_{i=1}^{n-1} \frac{x_i \wedge t}{t} \mathbb{1}(x_i \ge 0) \overset{iid}{\sim} \text{Uniform}([0, t]).
\end{aligned}
$$

Here $a \wedge b = \min\{a, b\}$. Recall that for the indicator function, we have $\mathbb{1}_A \mathbb{1}_B = \mathbb{1}_{A \cap B}$.

Hence

$$
\begin{aligned}
&\mathbb{P}_\theta(X_1 \le x_1, \ldots, X_n \le x_n \mid X_{(n)} = t) \\
&= \sum_{i=1}^{n} \mathbb{P}_\theta(X_1 \le x_1, \ldots, X_n \le x_n, X_i = X_{(n)} \mid X_{(n)} = t) \\
&= \sum_{i=1}^{n} \mathbb{P}_\theta(X_1 \le x_1, \ldots, X_n \le x_n \mid X_i = X_{(n)} = t) \times \mathbb{P}_\theta(X_i = X_{(n)} \mid X_{(n)} = t) \\
&= \sum_{i=1}^{n} \prod_{j \ne i} \frac{x_j \wedge t}{t} \mathbb{1}(x_j \ge 0) \times \frac{1}{n}.
\end{aligned}
$$

This does not depend on θ, by definition $X_{(n)}$ is a sufficient statistic for θ.

Example: Normal. Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$ r.v.'s with known σ. Consider $T(\mathbf{X}) = \bar{X} = (X_1 + \ldots + X_n)/n$.

$$\mathbb{P}_\theta(\mathbf{X} = \boldsymbol{x} | T(\mathbf{X}) = T(\boldsymbol{x})) = \frac{\mathbb{P}_\theta(\mathbf{X} = \boldsymbol{x}, T(\mathbf{X}) = T(\boldsymbol{x}))}{\mathbb{P}_\theta(T(\mathbf{X}) = T(\boldsymbol{x}))} = \frac{\mathbb{P}_\theta(\mathbf{X} = \boldsymbol{x})}{\mathbb{P}_\theta(T(\mathbf{X}) = T(\boldsymbol{x}))} = \frac{f(\boldsymbol{x}|\mu)}{q(T(\boldsymbol{x})|\mu)}$$

where

$$\begin{aligned} f(\boldsymbol{x}|\mu) &= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{\sum_i (x_i - \mu)^2}{2\sigma^2}\right) \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{\sum_i (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2}{2\sigma^2}\right) \\ q(T(\boldsymbol{x})|\mu) = q(\bar{x}|\mu) &= \frac{1}{\sqrt{2\pi}\frac{\sigma}{\sqrt{n}}} \exp\left(-\frac{n(\bar{x} - \mu)^2}{2\sigma^2}\right) \end{aligned}$$

$\bar{X}$ is sufficient for μ, but not σ.

Example: Order statistic. Let $X_1, \dots, X_n$ be a random sample from pdf f. Consider $T(\mathbf{X}) = (X_{(1)}, \dots, X_{(n)})$.

$$\mathbb{P}_\theta(\mathbf{X} = \boldsymbol{x} | T(\boldsymbol{X}) = T(\boldsymbol{x})) = \frac{f(\boldsymbol{x})}{q(T(\boldsymbol{x}))} = \frac{\prod_i f(x_i)}{\prod_i n! f(x_{(i)})} = \frac{1}{n!},$$

for any $\boldsymbol{x} = \pi(\boldsymbol{X})$, i.e., a permutation of $X_1, \dots, X_n$.

- Non-parametric example: the "parameter" here is the distribution function f.
- Notice that there is not much data reduction.
- Outside the exponential family, it is rare to have sufficient statistics that are of lower dimension than the sample size.

Factorization Theorem

It can be complicated to use the definition to

- find a candidate sufficient statistic; and
- check if a statistic is sufficient.

Luckily, there is a theorem that makes both tasks easy.

Theorem (Factorization theorem)

Let $f(\boldsymbol{x}|\theta)$ denote the joint pdf/pmf of a sample $\boldsymbol{X}$. A statistic $T(\boldsymbol{X})$ is a sufficient statistic for θ if and only if there exist functions $g(t|\theta)$ and $h(\boldsymbol{x})$ such that, for all sample points $\boldsymbol{x}$ and all parameter θ,

$$f(\boldsymbol{x}|\theta) = h(\boldsymbol{x})g(T(\boldsymbol{x})|\theta).$$

Remarks

$T \xrightarrow{s} S = s(T(X))$.

$T = s^{-1}(S(X))$

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

$\quad = h(x)\, g(s^{-1}(S(x)) \mid \theta)$

- The function h can depend on the full random sample $\boldsymbol{x}$, but not on the unknown parameter θ.
- The function g can depend on θ, but can depend on the random sample only through the value of $t = T(\boldsymbol{x})$.
- It is easy to see that if $s(t)$ is a one to one function and T is a sufficient statistic, then $s(T)$ is a sufficient statistic.
  Example: In the order statistic example, equivalently, the empirical cdf $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \leq x)$ is sufficient.

Proof

- "$\Rightarrow$" If sufficient, let $h(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{X} = \boldsymbol{x} | T(\boldsymbol{X}) = T(\boldsymbol{x}))$, then

$$f(\boldsymbol{x}|\theta) = \mathbb{P}_\theta(\boldsymbol{X} = \boldsymbol{x}) = \mathbb{P}(\boldsymbol{X} = \boldsymbol{x} | T(\boldsymbol{X}) = T(\boldsymbol{x}))\mathbb{P}_\theta(T(\boldsymbol{X}) = T(\boldsymbol{x}))$$

- "$\Leftarrow$" (in the case of discrete r.v.) If factorization, then let $q(T(\boldsymbol{x})|\theta)$ be the pmf of $T(\boldsymbol{X})$ and $A_{T(\boldsymbol{x})} = \{\mathbf{y} : T(\mathbf{y}) = T(\boldsymbol{x})\}$.

$$\mathbb{P}(\boldsymbol{X} = \boldsymbol{x} \mid T = t) = \frac{\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}, T = t)}{\mathbb{P}(T = t)} = \frac{f(\boldsymbol{x}|\theta)\mathbb{1}(T(\boldsymbol{x}) = t)}{\sum_{A_{T(\boldsymbol{x})}} f(\mathbf{y}|\theta)}$$
$$= \frac{g(T(\boldsymbol{x})|\theta)h(\boldsymbol{x})\mathbb{1}(T(\boldsymbol{x}) = t)}{\sum_{A_{T(\boldsymbol{x})}} g(T(\mathbf{y})|\theta)h(\mathbf{y})} = \frac{h(\boldsymbol{x})\mathbb{1}(T(\boldsymbol{x}) = t)}{\sum_{A_{T(\boldsymbol{x})}} h(\boldsymbol{y})}$$

- For a complete proof, see *Testing Statistical Hypothesis* (2015) by E. Lehmann and J. Romano, Section 2.6.

Examples Revisited

Example: Uniform($[0,\theta]$) revisited. We can write down the pdf of the full data as

$$f(\mathbf{x}) = 1/\theta^n \prod_i \mathbb{1}(0 \leq X_i \leq \theta) = 1/\theta^n \mathbb{1}(\max_i\{X_i\} \leq \theta) \prod_i \mathbb{1}(X_i \geq 0)$$

- By the factorization theorem, a sufficient statistic is $T(\boldsymbol{X}) = \max_i X_i = X_{(n)}$.
- The sample mean is not a sufficient statistic for $\mathbb{E}[X] = \theta/2$.

Example: Normal mean revisited. Consider a normal random sample $\mathcal{N}(\mu, \sigma^2)$ with known σ.

$$f(\boldsymbol{x}|\mu) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{\sum_i (x_i - \bar{x})^2}{2\sigma^2}\right) \exp\left(-\frac{n(\bar{x} - \mu)^2}{2\sigma^2}\right)$$

Then $T(\boldsymbol{X}) = \bar{X}$ and $g(t|\mu) = \exp\left(-\frac{n(t-\mu)^2}{2\sigma^2}\right)$.

Remarks

Definition (Multi-dimensional case)

The statistics $\boldsymbol{T} = (T_1, \ldots, T_k)$ are jointly sufficient if for each $\boldsymbol{t} = (t_1, \ldots, t_k)$, the conditional distribution of $\boldsymbol{X} = (X_1, \ldots, X_n)$ given $\boldsymbol{T}$ does not depend on $\boldsymbol{\theta}$.

- The factorization theorem applies to multi-dimensional parameter and statistic.

Example: Normal mean and variance. What if $\theta = (\mu, \sigma)$ is unknown?

Let $T_1(\boldsymbol{x}) = \bar{\boldsymbol{x}}$ and $T_2(\boldsymbol{x}) = s^2$.

$$g(\mathbf{t}|\theta) = g(t_1, t_2|\theta) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{(n-1)t_2}{2\sigma^2}\right) \exp\left(-\frac{n(t_1-\mu)^2}{2\sigma^2}\right)$$

Here $h(\boldsymbol{x}) = 1$. We see that $(T_1(\boldsymbol{x}), T_2(\boldsymbol{x}))$ is sufficient for (μ, σ).

If $\boldsymbol{\theta} \in \mathbb{R}^k$ and $\boldsymbol{f}(\boldsymbol{t})$ is a one to one function on $\mathbb{R}^k$ and $\boldsymbol{T}$ is a sufficient statistic for $\boldsymbol{\theta}$, then $\boldsymbol{f}(\boldsymbol{T})$ is also a sufficient statistic. More generally,

$S \xrightarrow{\psi} T = \psi(S)$

If T is sufficient and $T = \psi(S)$, where ψ is a (measurable) function and S is a statistic, then S is sufficient.

Example: Normal. $T_1 = \bar{X}, T_2 = \left(X_1, \sum_{i=2}^n X_i\right), T_3 = \boldsymbol{X}$ are all sufficient statistics for μ.

Remarks – Sufficiency

Any statistic T will induce a partition of the sample space according to the its value

$$\mathcal{A}(T) = \{A_t\}, \quad A_t = \{\boldsymbol{x} : T(\boldsymbol{x}) = t\}.$$

Example: Uniform$(0, \theta)$. Consider sample size of 2.

- For a statistic to be sufficient, the partition should be fine enough to distinguish information about different θ.
- If T is sufficient, we should draw identical statistical conclusions about θ inside each region.
- It is this partition, rather than the particular statistic inducing the partition, that is the fundamental object. This idea is formalized using σ-algebras in measure theory.

Exponential Families

A family of pdf or pmf is called a k-parameter exponential family if it can be written as

$$f(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^{k} w_i(\boldsymbol{\theta}) t_i(x)\right).$$

This special form is chosen for mathematical convenience.

Example: Binomial(n,p).

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x} (1-p)^n \left(\frac{p}{1-p}\right)^x = \binom{n}{x} (1-p)^n \exp\left(\log\left(\frac{p}{1-p}\right) x\right)$$

Example: Normal $N(\mu, \sigma^2)$.

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\mu)^2}{2\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} x^2 + \frac{\mu}{\sigma^2} x\right)$$

Example: Counterexample. The shifted exponential distributions does not form an exponential family

$$f(x|\theta) = \frac{1}{\theta} \exp\left(\frac{\theta - x}{\theta}\right) \mathbb{1}(x \geq \theta)$$

Natural Parameters of the Exponential Family

A exponential family is sometimes reparametered as

$$f(x|\boldsymbol{\eta}) = h(x)c^*(\boldsymbol{\eta}) \exp\left(\sum_{i=1}^{k} \eta_i t_i(x)\right).$$

where $\boldsymbol{\eta} = (w_1(\theta), w_2(\theta), \ldots, w_k(\theta))$ is called the natural parameter. This is called the **canonical form** of the exponential family.

Natural Parameter Space

$$\mathcal{H} = \left\{\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k) : \int_{-\infty}^{\infty} h(x) \exp\left(\sum_{i=1}^{k} \eta_i t_i(x)\right) ds < \infty, \text{ or } \sum_{x} ... < \infty\right\}$$

Example: For exponential distribution, we have $\mathcal{H} = \{\eta > 0\}$.

Using Hölder's inequality, one can prove that $\mathcal{H}$ is a convex set.

Natural Sufficient Statistics

Suppose $X_1, \ldots, X_n$ is a random sample from

$$f(x|\boldsymbol{\eta}) = h(x)c^*(\boldsymbol{\eta}) \exp\left(\sum_{i=1}^{k} \eta_i t_i(x)\right).$$

Define statistics $\boldsymbol{T} = (T_1(\boldsymbol{X}), \ldots, T_k(\boldsymbol{X}))$, where

$$T_i(\boldsymbol{X}) = \sum_{j=1}^{n} t_i(X_j).$$

In matrix form

$$f_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\eta}) = \left(\prod_i h(x_i)\right) [c^*(\boldsymbol{\eta})]^n \exp\left(\boldsymbol{\eta}^T \boldsymbol{T}(\boldsymbol{x})\right)$$

Natural sufficient statistics

By the factorization theorem, the natural statistics are sufficient for $\boldsymbol{\eta}$

$$\boldsymbol{T}(\boldsymbol{X}) = \left(\sum_{j=1}^{n} t_1(X_j), \ldots, \sum_{j=1}^{n} t_k(X_j) \right)$$

Moreover, $\boldsymbol{T}$ still belongs to an exponential family

$$f_T(\boldsymbol{u}|\theta) = \tilde{h}(\boldsymbol{u})[c^*(\boldsymbol{\eta})]^n \exp\left(\boldsymbol{\eta}^T \boldsymbol{u}\right).$$

Example: Bernoulli(p).

$$f(x|p) = p^x(1-p)^{1-x} = (1-p)\exp\left(\log\left(\frac{p}{1-p}\right)x\right)$$

So $k = 1$, $t_1(x) = x$ and $\eta = \log\left(\frac{p}{1-p}\right)$.

$$T = T_1(X_1, \ldots, X_n) = X_1 + \ldots + X_n.$$

T is Binomial (n, p)

$$f(x|p) = \binom{n}{x}(1-p)^n \exp\left(\log\left(\frac{p}{1-p}\right)x\right) = \binom{n}{x}(1-p)^{n-x}p^x$$

Minimal Sufficient Statistics

For a given parameter, there are many sufficient statistics.

- The concept of sufficiency implies no loss of information for θ.
- The concept of sufficiency, by itself, does not imply data reduction. (E.g. the full data $\boldsymbol{X}$ is sufficient statistic for any parameter θ.)

Is there a sufficient statistic that provides "maximal" reduction of data?

Definition (Minimal Sufficient Statistics)

A sufficient statistic $T(\boldsymbol{X})$ is called a minimal sufficient statistic if, for any other sufficient statistic $S(\boldsymbol{X})$, there is a (measurable) function such that $T = \psi(S)$ (a.s. for any $\mathbb{P}_\theta$).

Recall the partition induced by a statistic

$$\mathcal{A}(T) = \{A_t\}, \quad A_t = \{\boldsymbol{x} : T(\boldsymbol{x}) = t\}$$

Minimal sufficiency of T implies that

For any sufficient statistic S, if $S(\boldsymbol{x}) = S(\mathbf{y})$, then $T(\boldsymbol{x}) = T(\mathbf{y})$.

Then the partition $\mathcal{A}(T)$ is coarser than $\mathcal{A}(S)$.

The simpler the partition is, the more data reduction we have.

While retaining all information of θ, minimal sufficiency identifies

- the maximal reduction of the data;
- the coarsest partition of the sample space; and
- *the coarsest σ-algebra.

Remarks

One-to-one mapping

Any one-to-one function of a minimal sufficient statistic is minimal sufficient.

This can be proved using factorization theorem and the definition of minimality.

Uniqueness

Minimal statistic is unique in the sense that two statistics that are one-to-one measurable functions of each other can be treated as the same.

- The partitions induced are the same.
- It is this partition that is the fundamental object.

Minimal Sufficient Statistics

Example: Normal sufficient statistic for μ with known σ. Consider two sufficient statistics:

$$T(X) = \bar{X}, \quad T'(X) = (\bar{X}, S^2)$$

Recall that $\bar{X}$ and S^2 are independent. So $T'(X)$ can not be written as function of $T(X)$, and hence is not minimal.

How to check if $\bar{X}$ is minimal?

Minimal Sufficient Statistics

Theorem (Checking Rule)

Let $f(x|\theta)$ be the pmf/pdf of a sample $\boldsymbol{X}$. Suppose there exists a statistic $T(\cdot)$ such that, for every two sample realizations $\boldsymbol{x}$ and $\mathbf{y}$,

$$\frac{f(\boldsymbol{x}|\theta)}{f(\mathbf{y}|\theta)} \textit{ does not depend on } \theta \Leftrightarrow T(\boldsymbol{x}) = T(\mathbf{y}).$$

Then $T(\boldsymbol{X})$ is a minimal sufficient statistic for θ.

Example: Normal minimal sufficient statistic for $\theta = \mu, \sigma^2, (\mu, \sigma^2)$.

$$\frac{f(\boldsymbol{x}|\mu,\sigma)}{f(\mathbf{y}|\mu,\sigma)} = \frac{\frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{(n-1)s_{\boldsymbol{x}}^2}{2\sigma^2}\right) \exp\left(-\frac{n(\bar{x}-\mu)^2}{2\sigma^2}\right)}{\frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{(n-1)s_{\mathbf{y}}^2}{2\sigma^2}\right) \exp\left(-\frac{n(\bar{y}-\mu)^2}{2\sigma^2}\right)}$$

Proof of the Checking Rule

- Sufficiency. Given $T(\mathbf{x}) = T(\mathbf{y})$, then $f(\mathbf{x}|\theta)/f(\mathbf{y}|\theta)$ does not depend on θ. Hence $f(\mathbf{x}|\theta)$ is a constant function in θ on $A_t = \{\boldsymbol{x} : T(\boldsymbol{x}) = t\}$. Let $g(t|\theta) = f(\mathbf{x}|\theta)$ for any $\boldsymbol{x} \in A_t$. Let $h(\boldsymbol{x}) = f(\boldsymbol{x}|\theta)/g(T(\boldsymbol{x})|\theta)$. Note that h does not depend on θ. Sufficiency follows from the factorization theorem.
- Minimality. Let $T'(\boldsymbol{X})$ be a sufficient statistic, which has factorization $f(\boldsymbol{x}|\theta) = g'(T'(\boldsymbol{x})|\theta)h'(\boldsymbol{x})$. If $\boldsymbol{x}$ and $\mathbf{y}$ are two sample realizations such that $T'(\boldsymbol{x}) = T'(\mathbf{y})$, then

$$\frac{f(\boldsymbol{x}|\theta)}{f(\mathbf{y}|\theta)} = \frac{g'(T'(\boldsymbol{x})|\theta)h'(\boldsymbol{x})}{g'(T'(\mathbf{y})|\theta)h'(\mathbf{y})} = \frac{h'(\boldsymbol{x})}{h'(\mathbf{y})} \Rightarrow T(\boldsymbol{x}) = T(\mathbf{y})$$

  $T(\boldsymbol{x})$ is a function of $T'(\boldsymbol{x})$.

Uniform Distribution

Example: Consider $X_1, \ldots, X_n$ are uniform distributed r.v. in $[\theta, \theta+1]$.

- Remember the indicator function:

$$\frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)} = \frac{\mathbb{1}_{\theta \le x_{(1)} \le x_{(n)} \le \theta+1}}{\mathbb{1}_{\theta \le y_{(1)} \le y_{(n)} \le \theta+1}}$$

- By the checking rule, $T(\boldsymbol{X}) = (X_{(1)}, X_{(n)})$ is minimal sufficient.
- Notice that the sufficient static is of dimension 2, compared to the parameter $(1D)$ and the sample (nD).

Ancillary Statistics

Definition (Ancillary Statistic)

A statistic $V(\boldsymbol{X})$ whose distribution does not depend on the parameter θ is called an ancillary statistic.

Example: (Location family ancillary statistic) Suppose $F(\cdot)$ is a cdf, let $X_1, \ldots, X_n$ be a sample from $F(x-\theta)$. $R = X_{(n)} - X_{(1)}$ is ancillary.

$$P_\theta(R \leq r) = P_\theta(\max_i X_i - \min_i X_i \leq r) = P_\theta(\max_i (X_i - \theta) - \min_i (X_i - \theta) \leq r)$$

Recall the previous uniform example.

Example: (Scale family ancillary statistic) let $X_1, \ldots, X_n$ be a sample from $F(x/\sigma)$. $X_1/X_n, \ldots, X_{n-1}/X_n$ is ancillary. $X_i = \sigma Z_i$ with $Z_i \sim F$.

Remarks

- The simplest ancillary statistic is the constant statistic $V(\boldsymbol{X}) \equiv c$.
- A non-trivial ancillary statistic $V(\boldsymbol{X})$ identifies a partition $\mathcal{A}(V) = \{\{\boldsymbol{x} : V(\boldsymbol{x}) = v\} : v\}$ that does not contain any information about θ.
- Suppose that $T(\boldsymbol{X})$ is a statistic and $V(T(\boldsymbol{X}))$ is a non-trivial ancillary statistic, then the partition $\mathcal{A}(T)$ contains a coarser partition that does not contain any information about θ.
- This indicates that we may need further data reduction than T.
- A sufficient statistic seems to be the most "successful" in data reduction if <u>no</u> nonconstant function of it is ancillary.

Question: Recall that minimal sufficient statistics indicates a maximal data reduction while keeping the information of θ. Is minimal sufficient statistics "successful" in the above sense?

Ancillary statistics may be a component of the minimal sufficient statistic.

Example: Let $X_1, \ldots, X_n$ be a sample from Uniform $(\theta, \theta + 1)$.

- $R = X_{(n)} - X_{(1)}$ is ancillary.
- We know that $(X_{(1)}, X_{(n)})$ is minimal sufficient, hence

$$(X_{(n)} - X_{(1)}, X_{(n)} + X_{(1)}) \text{ is a minimal sufficient statistic.}$$

Therefore,

- There exist nonconstant function of minimal sufficient statistic that is ancillary $\Rightarrow$ minimal sufficient statistics is not "successful."
- Ancillary statistics is not always independent of minimal sufficient statistic.

This inspires the definition of **completeness**.

Completeness

Definition (Complete statistic)

Let $\boldsymbol{X}$ be i.i.d. from pdf/pmf $f(\cdot|\theta)$. A statistic $T(\boldsymbol{X})$ is said to be complete for θ, if any (measurable) function g <u>not depending on θ</u> satisfies that

$$\mathbb{E}_\theta[g(T(\boldsymbol{X}))] = 0 \text{ for all } \theta \Rightarrow \mathbb{P}_\theta(g(T(\boldsymbol{X})) = 0) = 1 \text{ for all } \theta.$$

- Complete if there is no non-trivial unbiased estimator for 0 based on $T(\boldsymbol{X})$.
- Completeness implies that unbiased estimator of θ based on T is unique.
- A minimal sufficient statistic is not necessarily complete. E.g. Uniform$([\theta, \theta + 1])$.
- Complete statistic is not necessarily sufficient. See example below.

Complete Statistics – Examples

Example: Uniform$(\theta, \theta+1)$ re-visited. $T(X) = (X_{(1)}, X_{(n)})$ is a minimal sufficient statistic. However, $X_{(n)} - X_{(1)} - \mathbb{E}[X_{(n)} - X_{(1)}]$ has mean 0, but is not 0 a.s., thus both $T(X)$ and $X_{(n)} - X_{(1)}$ are not complete. It is important here that $g(\cdot)$ does not depend on θ.

The range $R = X_{(n)} - X_{(1)}$ itself does not contain any information about θ, but combined with the sufficient statistics, it does! (See C-B Example 6.2.20.)

Example: Normal$(0, \sigma^2)$ with $\theta = \sigma$ and $T = \bar{X}$. Let $g(x) = x$, then $\mathbb{E}[g(\bar{X})] = 0$ but $\mathbb{P}(g(\bar{X}) = 0) \neq 1$. Not complete.

Example: Normal$(\mu, 1)$ with $\theta = \mu$ and $T = \bar{X}$. If $\mathbb{E}_\theta[g(\bar{X})] = 0$ for all θ, then $g \equiv 0$ with probability 1. Complete.

Example: Normal(μ, σ^2) and $T = \bar{X}$. T is complete for $\theta = \mu$ and $\theta = (\mu, \sigma^2)$, but not $\theta = \sigma^2$. Completeness does not imply sufficiency.

Example: Bernoulli complete statistic. For a Bernoulli random sample with success probability $\theta = p$, $0 < p < 1$. A minimal sufficient statistic is $T(\boldsymbol{X}) = \sum X_i \sim$Binomial$(n, p)$. Suppose

$$0 = \mathbb{E}_p[g(T)] = \sum_{t=0}^{n} g(t)\binom{n}{t} p^t (1-p)^{n-t} = (1-p)^n \sum_{t=0}^{n} g(t)\binom{n}{t}\phi^t.$$

where $\phi = \frac{p}{1-p}$, so $\phi \in (0, \infty)$. Note that $(1-p)^n > 0$. For a polynomial (in ϕ) to be a constant 0, every coefficient has to be 0. Complete!

Completeness Implies Minimality

Theorem

- *If a minimal sufficient statistic exists, then any complete sufficient statistic is also a minimal sufficient statistic.*
- *A finite dimensional complete sufficient statistic is also minimal sufficient.*

The theorem states that under mild conditions, a complete sufficient statistic is all you need. It implies minimal sufficiency.

Converse is not true: in the Uniform$(\theta, \theta+1)$ example, $(X_{(n)} - X_{(1)}, X_{(n)} + X_{(1)})$ is a minimal sufficient statistic for θ, but not complete.

If a minimal sufficient statistic T is not complete, then there does not exist any complete statistic.

Order Statistic Re-Visited

Example: We have argued that the order statistics $\boldsymbol{T}$ are sufficient for $\theta = f \in \Theta = \{\text{All dist. with a density.}\}$. We now show that it is also complete for $\theta \in \Theta$.

- First, note that δ is a function of $\boldsymbol{T}$ iff it is symmetric in its arguements, i.e. $\delta(\boldsymbol{x}) = \delta(\pi\boldsymbol{x})$ for any permutation π.
- Consider a family of distributions $f = \sum_{i=1}^{n} \alpha_i f_i \in \Theta$ for $\alpha_i > 0$, $\sum_i \alpha_i = 1$ and f_i to be specified. $\mathbb{E}_F[h(T(\boldsymbol{X}))] \equiv \mathbb{E}_F[\delta(\boldsymbol{X})] = 0$ implies that

$$0 = \int \cdots \int \delta(\boldsymbol{x}) \prod_{j=1}^{n} f(x_j) d\boldsymbol{x} = \int \cdots \int \delta(\boldsymbol{x}) \prod_{j=1}^{n} \left(\sum_{i=1}^{n} \alpha_i f_i(x_j) \right) d\boldsymbol{x}$$

- The RHS is a polynomial of $\boldsymbol{\alpha}$. Hence all coefficients must be zero.

- Consider the coefficient of $\prod_i \alpha_i$

$$
\begin{aligned}
0 &= \sum_{\pi} \int \cdots \int \delta(\boldsymbol{x}) \prod_{i=1}^{n} f_i(x_{\pi(i)}) d\boldsymbol{x} \\
&= \sum_{\pi} \int \cdots \int \delta(\pi^{-1}\boldsymbol{x}) \prod_{i=1}^{n} f_i(x_i) d\boldsymbol{x} = \sum_{\pi} \int \cdots \int \delta(\boldsymbol{x}) \prod_{i=1}^{n} f_i(x_i) d\boldsymbol{x} \\
&= n! \int \cdots \int \delta(\boldsymbol{x}) \prod_{i=1}^{n} f_i(x_i) d\boldsymbol{x}
\end{aligned}
$$

- Now, let f_i be uniform on interval $[a_i, b_i]$, then

$$
\int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} \delta(\boldsymbol{x}) d\boldsymbol{x} = 0 \Rightarrow \delta(\boldsymbol{x}) = 0, a.s.
$$

Complete!

Order Statistic Cont.

Example: Now that the ordered statistic $\boldsymbol{T}$ is complete and sufficient, consider the ranks of the observations

$$\boldsymbol{R} = (R_1, \ldots, R_n)$$

where $R_i \equiv \{\#\text{ of } X_j\text{'s} \leq X_i\}$. Then $\mathbb{P}(\boldsymbol{R} = \pi(1, \ldots, n)) = 1/n!$, $\boldsymbol{R}$ is ancillary!

In fact, $\boldsymbol{T}$ and $\boldsymbol{R}$ are independent

$$\mathbb{P}(\boldsymbol{T} = \boldsymbol{t}, \boldsymbol{R} = \boldsymbol{r}) = \frac{1}{n!} \times n! \prod_i f(t_i).$$

Complete Statistics

Theorem (Basu's)

If $T(\boldsymbol{X})$ is a complete and sufficient for $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Omega\}$, if $V(\boldsymbol{X})$ is ancillary, then $T(\boldsymbol{X})$ and $V(\boldsymbol{X})$ are independent under $\mathbb{P}_\theta$ for any θ.

Proof. Define $q_A(t) = \mathbb{P}_\theta(V \in A | T(\boldsymbol{X}) = t)$ and $p_A = P_\theta(V \in A)$. Let $g(t) = q_A(t) - p_A$. We have $g(T(\boldsymbol{X}))$ non-trivial and $g(t)$ does not depend on θ (by sufficiency and ancillarity). Now, note that

$$\mathbb{E}_\theta[g(T(\boldsymbol{X}))] = \mathbb{E}_\theta[\mathbb{P}_\theta(V \in A | T(\boldsymbol{X}))] - p_A = \mathbb{P}_\theta(V \in A) - p_A = 0.$$

By completeness, $q_A(T) = p_A, a.s.$

$$\begin{aligned}
\mathbb{P}_\theta(T \in A, V \in B) &= \mathbb{E}_\theta[\mathbb{1}_A(T)\mathbb{1}_B(V)] \\
&= \mathbb{E}_\theta[\mathbb{E}_\theta[\mathbb{1}_A(T)\mathbb{1}_B(V)|T]] \\
&= \mathbb{E}_\theta[\mathbb{1}_A(T)\mathbb{E}_\theta[\mathbb{1}_B(V)|T]] \\
&= \mathbb{E}_\theta[\mathbb{1}_A(T)q_A(T)] \\
&= \mathbb{E}_\theta[\mathbb{1}_A(T)p_A] \\
&= \mathbb{P}_\theta(T \in A)\mathbb{P}_\theta(V \in B)
\end{aligned}$$

Hence, T and V are independent.

Complete Statistics for Exponential Family

- Basu's theorem allows us to deduce the independence of two statistics. But how to show completeness?
- Luckily, for exponential family, we know how to do it.
- Let $X_1, \ldots, X_n$ a sample from a pdf/pmf that belongs to an k-parameter canonical exponential family given by

$$f(x|\boldsymbol{\eta}) = h(x)c^*(\boldsymbol{\eta}) \exp\left(\sum_{i=1}^{k} \eta_i t_i(x)\right),$$

where $\eta \in \Xi \subset \mathcal{H}$ is the parameter set.

Minimal Exponential Families

Minimal exponential family

An exponential family parameterize by its natural parameters $\mathcal{P} = \{\mathbb{P}_\eta : \eta \in \mathcal{H}\}$ is **minimal** if

1. there is no set of coefficients $\boldsymbol{\lambda} \in \mathbb{R}^{k+1}, \boldsymbol{\lambda} \neq \mathbf{0}$, such that $\sum_i \lambda_i \eta_i = \lambda_0$;
2. there is no set of coefficients $\boldsymbol{\lambda} \in \mathbb{R}^{k+1}, \boldsymbol{\lambda} \neq \mathbf{0}$, such that $\sum_i \lambda_i T_i(\boldsymbol{x}) = \lambda_0$.

- The first condition rules out possibility to transform the k-dimensional exponential family into an exponential family of smaller dimension.
- The second condition rules out cases where the model is **unidentifiable** (i.e., exist $\eta_1 \neq \eta_2$ such that $\mathbb{P}_{\eta_1} = \mathbb{P}_{\eta_2}$.) **Example:** $X \sim \text{Exp}(\eta_1, \eta_2)$, where $p(x, \eta_1, \eta_2) = \exp(-\eta_1 x - \eta_2 x + \log(\eta_1 + \eta_2))\mathbb{1}(x \geq 0)$.

Curved Exponential Family

Curved exponential family

Suppose $\mathcal{P} = \{\mathbb{P}_\eta : \eta \in \Xi\}$ is an k-parameter **minimal** canonical exponential family. If Ξ contains an k-dimensional open set, then $\mathcal{P}$ is called **full-rank**. Otherwise, $\mathcal{P}$ is **curved**.

In curved exponential family, the η_i's are related in a non-linear way.

Examples

Consider Normal (μ, σ^2),

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\mu)^2}{2\sigma^2}} \exp\left(\frac{\mu}{\sigma^2} x - \frac{1}{2\sigma^2} x^2\right)$$

- **Example:** Minimal and full-rank. $\eta_1 = \frac{\mu}{\sigma^2}$ and $\eta_2 = -\frac{1}{2\sigma^2}$.
- **Example:** Non-minimal. When we restrict $\mu = \sigma^2 = \theta$, then $\eta_1 = 1$ and $\eta_2 = -\frac{1}{2\theta}$.
- **Example:** Minimal and curved. When we restrict $\mu = \sigma = \theta$, then $\eta_1 = 1/\theta$ and $\eta_2 = -1/(2\theta^2)$.
    - $T = (\bar{X}, S^2)$ is a sufficient statistic for θ, but it is not complete for θ.
    - To show it is not complete, need to find a nonzero function of T such that $\mathbb{E}[g(\bar{X}, S^2)] = 0$ for all θ. Let $g(\bar{X}, S^2) = n\bar{X}^2/(n+1) - S^2$.

Completeness and Exponential Family

Theorem

Suppose $\mathcal{P} = \{\mathbb{P}_\eta : \eta \in \Xi\}$ is an k-parameter minimal canonical exponential family of full-rank, then

$$T(\boldsymbol{X}) = \left(\sum_{j=1}^{n} t_1(X_j), \ldots, \sum_{j=1}^{n} t_k(X_j)\right)$$

is complete.

Application of Basu's Theorem

Example: Let $X_1, \ldots, X_n$ be a sample from Exponential(θ). The following two are independent

$$g(\boldsymbol{X}) = \frac{X_n}{X_1 + \ldots + X_n}, \quad T(\boldsymbol{X}) = X_1 + \ldots + X_n.$$

- Exponential is a scale family, so $g(\boldsymbol{X})$ is ancillary.
- Exponential is a minimal 1-parameter exponential family of full-rank, so $T(\boldsymbol{X})$ is complete and sufficient.
- Easy to verify minimality using the checking rule, but unnecessarily!
- So $\mathbb{E}_\theta[g(\boldsymbol{X})] = 1/n$.

Example: If we consider $N(\mu, \sigma^2)$ for a known σ, $\bar{X}$ and S^2 are independent.

Example: For $N(\mu, \sigma^2)$ with known σ. $\bar{X}$ is a sufficient and complete statistic and $med(\boldsymbol{X}) - \bar{X}$ is ancillary. So $\mathrm{Cov}(\bar{X}, med(\boldsymbol{X})) = \sigma^2/n$.

Summary

Consider two experiments

- Observe $X \sim \mathbb{P}_{X|\theta}$.
- Observe $T \sim \mathbb{P}_{T|\theta}$, then $X|T = t \sim \mathbb{P}_{X|t}$.

Then

- X in both experiments share the same dist., thus inference about θ should be the same in both cases.
- If T is sufficient, only the experiment of observing T is informative about θ.
  - T induce partitions on which identical statistical conclusions are drawn.
  - It is this partition (or σ-algebra), rather than the particular statistic inducing the partition, that is the fundamental object.
- If no coarser partition of the sample space that retains sufficiency is possible, then T is called minimal sufficient.

We will learn more about completeness next week.

Reading Materials

Same level

- Robert W. Keener, Theoretical Statistics, Chapter 2 and 3.
- (Not recommended) Casella and Berger, Statistical Inference, Section 6.2.

Measure theoretic

- Jun Shao, Mathematical Statistics, Section 2.2.
- Lehmann and Romano, Testing Statistical Hypothesis, Section 1.9 and 2.6.