Data Reduction

Feng-Chang Lin

Department of Biostatistics
University of North Carolina at Chapel Hill

flin@bios.unc.edu

(C&B §6)

Introduction

- Suppose that we are interested in estimating a parameter θ.
- If there is a random sample, X, whose pdf or pmf does not depend on θ, one would say "X does not contain any information about θ".
- On the other hand, it is possible to have a brief summary statistic that contains all the information about θ.
- We call this "data reduction", which summarizes a large number of observations into a small number of summary statistics.
- Our ultimate goal is to find the "smallest", most concise, summary statistics.

Sufficient Statistics

- Principle: If $T(X)$ is a sufficient statistic for θ, then it is sufficient to do any inference about θ through $T(X)$.
- That is, if x and y are two sample values such that $T(x) = T(y)$, then inference about θ should be the same whether $X = x$ or $X = y$ is observed.
- **Sufficient statistics**: A statistic $T(X)$ is a *sufficient statistic* for θ if the conditional distribution of the sample X given the value of $T(X)$ does not depend on θ.

Sufficient Statistics (cont'd)

- **Example** Let $X_1, \cdots, X_n$ be iid random variables distributed as bernoulli(θ), $0 < \theta < 1$. Show that $T(X) = \sum_{i=1}^{n} X_i$ a sufficient statistic for θ.
- **Proof** Since

$$P(X = x | T(X) = t) = \frac{P(X = x, T(X) = t)}{P(T(X) = t)},$$

where

$$P(T(x) = t) = \binom{n}{t} \theta^t (1-\theta)^{n-t},$$

and

$$P(X = x, T(X) = t) = P(X = x) = \prod_{i=1}^{n} P(X_i = x_i) = \theta^t (1-\theta)^{n-t}.$$

Sufficient Statistics (cont'd)

- Hence, $P(X = x|T(X) = t) = t!(n-t)!/n!$, for those $x_i's$ with $\sum_{i=1}^{n} x_i = t$, and $P(X = x|T(X) = t) = 0$, otherwise.

Sufficient Statistics (cont'd)

- For θ, the sufficiency statistics may not be unique.
- In this case, $\bar{X}$, $(X_1, \bar{X})$, $(X_1, \cdots, X_n)$ are all sufficient statistics.
- **Theorem 6.2.2** If $p(x|\theta)$ is the joint pdf or pmf of X and $q(t|\theta)$ is the pdf or pmf of $T(X)$. $T(X)$ is a sufficient statistic for θ if, for every x in the sample space, the ratio $p(x|\theta)/q(T(x)|\theta)$ does not depend on θ.

Finding Sufficient Statistics

- So far, we only show whether $T(X)$ is a sufficient statistic.
- The question here is "how to find one"?

Theorem (Factorization Theorem)

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

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

Finding Sufficient Statistics (cont'd)

- **Example** Let $X_1, \cdots, X_n$ be iid random variables distributed as Bernoulli(θ), $0 < \theta < 1$. Show that $T(x) = \sum_{i=1}^n x_i$ is a sufficient statistic using Factorization Theorem.
- **Proof** We first write the joint pmf

$$\begin{aligned}
P(X = x) &= \prod_{i=1}^n P(X_i = x_i) \\
&= \prod_{i=1}^n \theta^{x_i}(1-\theta)^{(1-x_i)} I(x_i \in \{0,1\}) \\
&= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} \prod_{i=1}^n I(x_i \in \{0,1\}).
\end{aligned}$$

- We can have $g(T(x)|\theta) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}$ as a function of $T(x) = \sum_{i=1}^n x_i$ and $h(x) = \prod_{i=1}^n I(x_i \in \{0,1\})$.

Finding Sufficient Statistics (cont'd)

- **Example** Let $X_1, \cdots, X_n$ be iid random variables distributed as Uniform$(0, \theta)$. Find a sufficient statistic for θ.
- **Solution** To apply the factorization theorem, we first write the joint pdf

$$f_X(x) = \theta^{-n} \prod_{i=1}^{n} I(0 < x_i < \theta) = \theta^{-n} I(0 < x_{(n)} < \theta) I(0 < x_{(1)})$$

- Take $T(x) = x_{(n)}$, $g(T(x)|\theta) = \theta^{-n} I(0 < T(x) < \theta)$, and $h(x) = I(0 < x_{(1)})$.
- We can conclude $T(X) = X_{(n)}$ is a sufficient statistic for θ.

Sufficiency in Exponential Family

- **Theorem 6.2.10** Let $X_1, \cdots, X_n$ be iid random variables from a pdf or pmf $f(x|\theta)$ that belongs to the exponential family given by

$$f(x|\theta) = h(x)c(\theta)\exp\left(\sum_{j=1}^{k} w_j(\theta)t_j(x)\right),$$

where $\theta = (\theta_1, \cdots, \theta_d)$, $d \leq k$. Then,

$$T(X) = \left(\sum_{i=1}^{n} t_1(X_i), \cdots, \sum_{i=1}^{n} t_k(X_i)\right)$$

is a sufficient statistic for θ.

Sufficiency in Exponential Family (cont'd)

- **Example** Let $X_1, \cdots, X_n$ be iid random variables distributed as Bernoulli(θ), $0 < \theta < 1$. Show that $T(X) = \sum_{i=1}^n X_i$ is a sufficient statistic for θ.
- **Solution** The pmf for one observation is

$$\begin{aligned} P(X_1 = x) &= \theta^x (1-\theta)^{1-x} I(x \in \{0,1\}) \\ &= I(x \in \{0,1\})(1-\theta) \exp\left(x \log \frac{\theta}{1-\theta}\right). \end{aligned}$$

- Take $h(x) = I(x \in \{0,1\})$, $c(\theta) = (1-\theta)$, $w_1(\theta) = \log \frac{\theta}{1-\theta}$, $t_1(x) = x$.
- By the sufficiency theorem in exponential family, one can conclude $T(X) = \sum_{i=1}^n t_1(X_i) = \sum_{i=1}^n X_i$ is a sufficient statistic for θ.

Minimal Sufficient Statistics

- In the Bernoulli example, there is a large number of sufficient statistics: $\sum_{i=1}^{n} X_i$, $\bar{X}$, $(X_1, \bar{X})$,$\cdots$,$(X_1, \cdots, X_n)$.
- Apparently, some of these can be reduced to a simpler form that is still sufficient for θ.
- **Minimal Sufficient Statistics**: A sufficient statistic is a minimal sufficient statistic if it is a function of every other sufficient statistic.
- Any one-to-one transformation of a minimal sufficient statistic is also a minimal sufficient statistic (still not unique).

Minimal Sufficient Statistics (cont'd)

- **Theorem 6.2.13** Let $f(x|\theta)$ be the joint pdf or pmf of X. Suppose that there exists a function $T(X)$ such that, for every two sample points x and y, the ratio $f(x|\theta)/f(y|\theta)$ does not depend on θ if and only if $T(x) = T(y)$. Then $T(X)$ is a minimal sufficient statistic for θ.

Minimal Sufficient Statistics (cont'd)

- **Example** Let $X_1, \cdots, X_n$ be iid random variables distributed as Bernoulli(θ), $0 < \theta < 1$. Show that $T(x) = \sum_{i=1}^n x_i$ is a minimal sufficient statistic.
- **Proof** To apply the above theorem, we first write the joint pmf

$$P(X = x) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} \prod_{i=1}^n I(x_i \in \{0, 1\}).$$

- If $T(x) = \sum_{i=1}^n x_i$, one can have

$$P(X = x) = \left(\frac{\theta}{1-\theta}\right)^{T(x)} (1-\theta)^n \prod_{i=1}^n I(x_i \in \{0, 1\}).$$

Minimal Sufficient Statistics (cont'd)

- Taking two points, x and y, in the sample space for X. One has

$$\frac{P(X=x)}{P(X=y)} = \left(\frac{\theta}{1-\theta}\right)^{T(x)-T(y)}.$$

- The ratio does not depend on θ if and only of $T(x) = T(y)$.

Ancillary Statistics

- Sample values may contain some additional information that is redundant of θ.
- For example, suppose that X_1, X_2 are iid as $N(\theta, 1)$. The random variable $X_1 - X_2$ is distributed as $N(0, 2)$.
- Is $X_1 - X_2$ expected to provide any information about θ?
- How about $(X_1 - X_2, X_2)$?
- **Ancillary Statistics**: A statistic whose distribution does not depend on the parameter θ is called an *ancillary statistic* (for θ).

Ancillary Statistics (cont'd)

- Let $X_1, \cdots, X_n$ be iid from a *scale* parameter family with cdf $F(x/\sigma)$, $\sigma > 0$.
- Any statistic that depends on $X_1/X_n, \cdots, X_{n-1}/X_n$ is an ancillary statistic.
- For example, $(X_1 + \cdots + X_n)/X_n = X_1/X_n + \cdots + X_{n-1}/X_n + 1$ is an ancillary statistic.
- Let $Z_i = X_i/\sigma$. We know that Z_i does not depend on σ.
- Since the joint cdf of $X_1/X_n, \cdots, X_{n-1}/X_n$ is

$$\begin{aligned} F(y_1, \ldots, y_{n-1}|\sigma) &= P(X_1/X_n \le y_1, \ldots, X_{n-1}/X_n \le y_{n-1}) \\ &= P(\sigma Z_1/(\sigma Z_n) \le y_1, \ldots, \sigma Z_{n-1}/(\sigma Z_n) \le y_{n-1}) \\ &= P(Z_1/Z_n \le y_1, \ldots, Z_{n-1}/Z_n \le y_{n-1}) \end{aligned}$$

- The last line shows the cdf does not depend on σ and $(X_1 + \cdots + X_n)/X_n$ is an ancillary statistic of σ.

Complete Statistics

- **Complete Statistics**: Let $\{f(t|\theta) : \theta \in \Theta\}$ be a family of pdfs or pmfs for $T(X)$. The family is called complete if $E_\theta g(T) = 0$ for all $\theta \in \Theta$ implies that $P_\theta(g(T) = 0) = 1$ for all $\theta \in \Theta$.
- Completeness means that the only function of T with mean 0 is the 0 function.
- **Example** Let $X_1, \cdots, X_n$ be iid random variables distributed as $N(\theta, \theta^2)$, $-\infty < \theta < \infty$. Is $T = (\bar{X}, S^2)$ complete? Since $E_\theta \bar{X}^2 = \theta^2 + \theta^2/n = (1 + 1/n)\theta^2$ and $E_\theta S^2 = \theta^2$, one can have $g(T) = \bar{X}^2 - (1 + 1/n)S^2$ and $E_\theta g(T) = 0$ for all $\theta \in \Theta$.
- Here $g(T)$ is not a zero function (with probability 1) and does not involve θ. Hence T is NOT complete.

Complete Statistics (cont'd)

- **Example** Let $X \sim \text{Bernoulli}(\theta)$, $\theta \in (0, 1)$. Take $T(X) = X$. Is T complete? This is equivalent to find out if $g = 0$ is the only function that has $E_\theta g(T) = 0$ for all $\theta \in (0, 1)$.
- **Solution** Since X follows Bernoulli, one only has $g(0)$ and $g(1)$ for $g(T)$. Then, if

$$E_\theta g(T) = g(0)(1 - \theta) + g(1)\theta = g(0) + \{g(1) - g(0)\}\theta = 0,$$

the only solution for g function is $g(0) = g(1) = 0$ for $\theta \in (0, 1)$.

Complete Statistics (cont'd)

- **Example** Similarly, let $X \sim \text{Binomial}(2, \theta)$, $\theta \in \Theta$, where $\Theta = \{1/3, 2/3\}$. Take $T(X) = X$. Is T complete? One can see $X = 0, 1, 2$. Follow the same approach,

$$E_\theta g(T) = (4/9)g(0) + (4/9)g(1) + (1/9)g(2), \text{ if } \theta = 1/3,$$
$$E_\theta g(T) = (1/9)g(0) + (4/9)g(1) + (4/9)g(2), \text{ if } \theta = 2/3.$$

  If $E_\theta g(T) = 0$, one can find $g(0) = g(2) = 4$, $g(1) = -5$ as a solution, which shows g function can be non-zero
- **Example** Let $X \sim \text{Binomial}(2, \theta)$, $\theta \in \Theta$, where $\Theta = \{1/3, 1/2, 2/3\}$. Take $T(X) = X$. Is T complete? Yes.
- That tells you the completeness highly depends on the parameter space.

Completeness in Exponential Families

- Let $X_1, \cdots, X_n$ be iid random variables from a pdf or pmf $f(x|\theta)$ that belongs to the exponential family given by

$$f(x|\theta) = h(x)c(\theta)\exp\left(\sum_{j=1}^{k} w_j(\theta)t_j(x)\right),$$

where $\theta = (\theta_1, \cdots, \theta_k)$. Then

$$T(X) = \left(\sum_{i=1}^{n} t_1(X_i), \cdots, \sum_{i=1}^{n} t_k(X_i)\right)$$

is complete if $\{(w_1(\theta), \cdots, w_k(\theta)) : \theta \in \Theta\}$ contains an open set in R^k.

- **Example**: The family $\{N(\mu, \sigma^2) : -\infty < \mu < \infty\}$ with a fixed $\sigma^2 < \infty$ is complete.

Exponential Families

- **Example**: Let $f(x|\mu, \sigma^2)$ be the $N(\mu, \sigma^2)$ family of pdfs where $\theta = (\mu, \sigma^2)$, $-\infty < \mu < \infty$, $\sigma > 0$. Then

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

- Take $h(x) = 1$ for all x,

$$
\begin{aligned}
&c(\theta) = c(\mu, \sigma) = (\sqrt{2\pi}\sigma)^{-1} \exp(-\mu^2/(2\sigma^2)), -\infty < \mu < \infty, \sigma > 0, \\
&w_1(\mu, \sigma) = \sigma^{-2}, \sigma > 0, w_2(\mu, \sigma) = \mu/\sigma^{-2}, \sigma > 0, \\
&t_1(x) = -x^2/2, \text{ and } t_2(x) = x.
\end{aligned}
$$

Exponential Families (cont'd)

- **Example** If $f(x|\theta) = \theta^{-1} \exp(1 - (x/\theta))$, $0 < \theta < x < \infty$, it is not an exponential family since

$$f(x|\theta) = \theta^{-1} \exp\left(1 - \left(\frac{x}{\theta}\right)\right) I_{[\theta,\infty)}(x).$$

- The indicator function is not a function of x alone, and cannot be expressed as an exponential.

Basu's theorem

Theorem (Basu's Theorem)

If $T(X)$ is a complete and minimal sufficient statistic, then $T(X)$ is independent of every ancillary statistic.

Proof: (only for discrete distributions) Let $S(X)$ be any ancillary statistic, so $P(S(X) = s)$ does not depend on θ. Since $T(X)$ is a sufficient statistic,

$$P(S(X) = s|T(X) = t) = P(X \in \{x : S(x) = s\}|T(X) = t),$$

does not depend on θ. For independence, we owe to show

$$P(S(X) = s|T(X) = t) = P(S(X) = s)$$

for all possible values of $t \in \mathcal{T}$.

Basu's theorem (cont'd)

- Marginalizing the joint probability of $S(X)$ and $T(X)$, one can have

$$\begin{aligned} P(S(X)=s) &= \sum_{t\in\mathcal{T}} P(S(X)=s, T(X)=t) \\ &= \sum_{t\in\mathcal{T}} P(S(X)=s|T(X)=t)P_\theta(T(X)=t). \qquad (1) \end{aligned}$$

- Since $\sum_{t\in\mathcal{T}} P_\theta(T(X)=t)=1$, one can also write

$$\begin{aligned} P(S(X)=s) &= P(S(X)=s)\sum_{t\in\mathcal{T}} P_\theta(T(X)=t) \\ &= \sum_{t\in\mathcal{T}} P(S(X)=s)P_\theta(T(X)=t). \qquad (2) \end{aligned}$$

Basu's theorem (cont'd)

- By (1) and (2), we can have

$$\begin{aligned} 0 &= P(S(X)=s) - P(S(X)=s) \\ &= \sum_{t\in\mathcal{T}} \{P(S(X)=s|T(X)=t) - P(S(X)=s)\} P_\theta(T(X)=t) \end{aligned}$$

- If we let $g(t) = P(S(X)=s|T(X)=t) - P(S(X)=s)$, then

$$0 = \sum_{t\in\mathcal{T}} g(t) P_\theta(T(X)=t) = E_\theta g(T), \quad \text{for all} \quad \theta.$$

- Since $T(X)$ is a complete statistic, the equation above implies that $g(t) = 0$ for all possible values of $t \in \mathcal{T}$.
- Hence, we can claim $P(S(X)=s|T(X)=t) = P(S(X)=s)$.

Basu's theorem (cont'd)

- Did we use "minimality" of the sufficient statistics in the proof?
- For the problems we will consider, a sufficient statistic will be complete only if it is minimal.
- **Theorem 6.2.28** If a minimal sufficient statistic exists, then any complete statistic is also a minimal sufficient statistics.

Practical Use of Basu's theorem

- **Example** Let $X_1, \cdots, X_n$ be iid Exponential(θ). Compute the expected value of

$$S(X) = \frac{X_n}{X_1 + \cdots + X_n}.$$

- We can show that $S(X)$ is an ancillary statistic (How?)
- Since Exponential(θ) belongs to the exponential family (homework) with $t(x) = x$, so $T(X) = \sum_{i=1}^n X_i$ is a (minimal) sufficient statistic.
- Hence by Basu's theorem, $T(X)$ and $S(X)$ are independent and

$$\theta = E_\theta X_n = E_\theta T(X)S(X) = E_\theta T(X) E_\theta S(X) = n\theta E_\theta S(X).$$

One has $E_\theta S(X) = 1/n$.