Lecture 6: Principle of Data Reduction

Ma Xuejun

School of Mathematical Sciences

Soochow University

https://xuejunma.github.io

Outline

Review

- The Delta Method and The Multivariate Delta Method
- The Edgeworth Expansion

Estimation

- WLLN and CLT shows that the sample average is a reasonable estimator for the expecation
  - Converges to the expectation
  - Rate $O(1/\sqrt{n})$
- Is sample average the best estimator for the expectation?
  - 'Best' in what sense?
  - If not,how to find the 'best' estimate have?
  - What performance will the 'best' estimate have?
- Estimations for other parameters,or function of parameters?
  - Example:$X \sim N(\mu, \sigma^2)$.What is the estimation for σ?
  - Example:$X \sim Gamma(\alpha, \beta)$.How to estimate α and β? How about $\alpha + \beta$?
  - Not all of them can be estimated from sample mean
  - What is a proper estimation?

Introduction

This section covers the section topic of our class.including:

- Parametric models
- Data reduction via statistics
- How to construct estimators

Evaluation of estimators will be covered in the third topic.

Population and Sample

- Population
  - The collection of measurements on a variable of interest:e.g,the condition of each light bulb of one manufactory.
  - Usually,hypothesize a model:e.g,Bernoulli(p)
- Sample

Definition 5.1.1: Random sample

The random variables $X_1, X_2, \cdots, X_n$ are called a random sample of size n from the population $f_X(x)$ if $X_1, X_2, \cdots, X_n$ are i.i.d. random vriables with PMF or PDF $f_X(x)$.

- Example:A sample from the light bulb manufactory:$X_1, X_2, \cdots, X_n \sim Ber(p)$
- The "i.i.d" condition can be relaxed
- If "i.i.d" condition holds,then the joint density of the random sample is

$$f_{X_1,\cdots,X_n}(x_1, \cdots, x_n) = \prod_{i=1}^{n} f_X(x_i)$$

Parameter Estimation

- Usually, f_X is not known to us.We draw samples to explore the properties of f_X,e.g.,expectation,variance,tails,ect.
- If prior information is known,say, f_X has an unknown finite dimentional parameter $\theta \in \Theta$,which characterizes f_X.Then the problem is to estimate θ
  - Joint distribution of the sample:

$$f_{X_1,\cdots,X_n}(x_1,\cdots,x_n;\theta)=\prod_{i=1}^{n} f_X(x_i;\theta)$$

  - Estimate θ
  - Construct some statistical tests for θ;say,model selection
  - asymptotic properties
- If no prior information is known, we cannot assume the distribution family for f_X.We call it as non-parametric statistics
  - Splines
  - Kernel estimation
  - etc

I.I.D. Normal Model

- A basic model statisticians usually use is the normal model
- Let $X_1, X_2, \cdots, X_n \sim N(\mu, \sigma^2)$.Here,the unkown parameters are $\theta = (\mu, \sigma^2)$.
- Given the observations $x_1, \cdots, x_n$,the joint density is

$$f_{X_{1:n}}(x_1, \cdots, x_n; \theta) = \prod_{i=1}^{n} \frac{\exp\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\}}{\sqrt{2\pi}\sigma} = \frac{\exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\}}{(\sqrt{2\pi}\sigma)^n}$$

- If the observations are given,then $f_{X_{1:n}}(x_1, \cdots, x_n; \theta)$ can be seen as a function about θ,

$$L(\theta; x_1, \cdots, x_n) = \frac{\exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\}}{(\sqrt{2\pi}\sigma)^n}$$

  $L(\theta; x_1, \cdots, x_n)$ is called the likelihood function for this models.
- One way to estimate the parameters $\theta = (\mu, \sigma^2)$ is to find the maximister of $L(\theta)$:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta; x_1, \cdots, x_n)$$

I.I.D. Normal Model

Maxinum likelihood function estimation for normal dist.(Quick review)

- Note that $l(\theta) = \log L(\theta)$ has the same maximister with $L(\theta)$.

$$l(\theta; x_1, \cdots, x_n) = \log L(\theta; x_1, \cdots, x_n) =$$
$$\exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\} - \frac{n}{2}\log(2\pi\sigma^2)$$

- Take the partial derivative,we have

$$(\frac{\partial l(\theta)}{\partial \mu}, \frac{\partial l(\theta)}{\partial \sigma^2}) = (\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu), -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2)$$

- Let the derivative to be 0(local extrema).The solution is

$$\tilde{\mu}_n \sum_{i=1}^{n} x_i, \qquad \tilde{\sigma^2}_n = \frac{1}{n}\sum_{i=1}^{n}(x_i - \tilde{\mu}_n)^2$$

- Since this is the only solution,this local extrema should be a global extrema.Check whether it is a maxima.We need the Hessian matrix to be a negative definite matrix.

The Exponential Family

- Generalize the family of normal distribution
- Exponential family is a class of densities,which for a random variable X and parameter θ,the density function is

$$f_X(x;\theta) = h(x)\exp\{\eta(\theta)T(x) - A(\theta)\}$$

  - $h, , T, A$ are known functions
  - The density functions is a product of data-only part $h(x)$, parameter-only part $\exp\{-A(\theta)\}$,and the cross-term of data and parameters.
  - The cross-term can be expressed as exponential transformation of the product of parameter and data.
  - Joint density:

$$f(x_1, \cdots, x_n) = (\prod_{i=1}^{n} h(x_i))\exp\{\eta(\theta)\sum_{i=1}^{n} T(X_i) - nA(\theta)\}$$

The Exponential Family-Example

- The normal distribution belongs to the exponential Family.
  If $X \sim N(\mu, \sigma^2)$,the density function is

  $$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\} = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2 - \frac{\mu^2}{2\sigma^2} - \log\sigma}\}$$

  Let
  $h(x) = \frac{1}{\sqrt{2\pi}}, \eta(\theta) = (-\frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2}), T(x) = (x^2, x), A(\theta) = \frac{\mu^2}{2\sigma^2} + \log\sigma.$
  Then we have

  $$f_X(x;\theta) = h(x)\exp\{\eta(\theta)^T T(x) - A(\theta)\},$$

  which is an exponential family distribution.

Bayesian Models:Example

Example: Suppose that $X_i|\theta \sim \exp(\lambda)$. Also,we know that $\lambda \sim Gamma(a,b)$.Given the observations $x_1, x_2, \cdots, x_n$,what information can we get about λ?

Solution

The joint density for $x_1, x_2, \cdots, x_n$ and λ is

$$
\begin{aligned}
f(x_1, x_2, \cdots, x_n, \lambda) &= [\prod_{i=1}^{n} f_X(x;\lambda)]\pi(\lambda) \\
&= [\prod_{i=1}^{n} \lambda e^{\lambda x_i}]\frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda} = \frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{\lambda\sum x_i}\lambda^{a-1}e^{-b\lambda} \\
&= \frac{b^a}{\Gamma(a)}\lambda^{n+a-1}e^{-\lambda(b+\sum x_i)}
\end{aligned}
$$

Now,we are curious about λ,so we want to know the conditional distribution of λ given the observations.

According to the definition of conditional distribution,we have

$$
\pi(\lambda|x_1, x_2, \cdots, x_n) = \frac{f(x_1, x_2, \cdots, x_n, \lambda)}{f(x_1, x_2, \cdots, x_n)}
$$

Here,to differentiate the density function for X and λ,we always use π for the density function of λ,and f for the density function of $X_1, \cdots, X_n$.

Bayesian Models:Example

We want to solve

$$\pi(\lambda|x_1, x_2, \cdots, x_n) = \frac{f(x_1, x_2, \cdots, x_n, \lambda)}{f(x_1, x_2, \cdots, x_n)}$$

- The numerator is knowm, $f(x_1, x_2, \cdots, x_n, \lambda) = \frac{b^a}{\Gamma(a)} \lambda^{n+a-1} e^{-\lambda(b+\sum x_i)}$
- The denominator can be calculated :

$$\begin{aligned} f(x_1, x_2, \cdots, x_n) &= \int_\lambda f(x_1, x_2, \cdots, x_n, \lambda) d\lambda \\ &= \frac{b^a}{\Gamma(a)} \int_0^\infty \frac{b^a}{\Gamma(a)} \lambda^{n+a-1} e^{-\lambda(b+\sum x_i)} d\lambda \\ &= \frac{b^a}{\Gamma(a)} \times \frac{\Gamma(n+a)}{(b + \sum_{i=1}^n x_i)^{n+a}} \end{aligned}$$

- So,the conditional distribution for λ is

$$\begin{aligned} \pi(\lambda|x_1, x_2, \cdots, x_n) &= \frac{(b + \sum_{i=1}^n x_i)^{n+a}}{\Gamma(n+a)} \frac{b^a}{\Gamma(a)} \lambda^{n+a-1} e^{-\lambda(b+\sum x_i)} \\ &\sim Gamma(n+a, b + \sum x_i) \end{aligned}$$

We call it as posterior distribution.

Bayesian Models:Remarks

- In the previous example, if we have proir information about λ, say, the expectation and variance,then we can identify the values of a and b,and the posterior is totally known.
- In Bayesian statistics, the posterior function is the "final answer".For frequentist, the estimation is a value.
- Depend on the loss function,the posterior function can be further reduced to a value. For example,when the loss function is $L^2 - loss((\hat{\theta} - \theta)^2)$,then the Bayes estimator can be reduced as $E[\pi(\theta|x_1, \cdots, x_n)]$. Details discussed later.
- Therefore, there is no "confidence interval" in Bayesian statistics. A similar notion is "credible interval". Details later.

The Linear Model

- Consider a sequence of data in pairs:$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$,where $y_i \in \mathbb{R}$,and $x_i \in \mathbb{R}^p, p \geq 1, 1 \leq i \leq n$.

- The y_i's are called response variables and the y_i's are explanatory variables.It is hypothesised that there exist some functional relationship of the form

$$Y_i = g_\theta(x_i) + \epsilon$$

Therefore,we can use x_i to predict the responses y_i,Here,ϵ_i is interpreted as noise.

- A simple prediction function is linear function.Therefore,the prediction function is

$$Y_i = \theta_0 + \sum_{j=1}^{p-1} \theta_j x_{ij} + \epsilon_i,$$

where ϵ_i are i.i.d.zero mean random samples,usually assumed to be $N(0, \sigma^2)$.
This model is called the linear regression model.

Statistics

- Random sample:$X_1, X_2, \cdots, X_n$.
- Work on the random sapmle to achieve information.

Definition: Statistics

For a random sample $X_1, X_2, \cdots, X_n$, a statistics is a function of the random sample $T(X_1, X_2, \cdots, X_n)$.

- The statistic $T(X)$ is also a random variable.Most times,its distribution changes with n,and we denote the CDF as G_n,called the sampling distribution.
- With the observations $x_1, x_2, \cdots, x_n$, we have $T(x_1, x_2, \cdots, x_n)$,a realization of the statistic $T(X_1, X_2, \cdots, X_n)$.

Statistics:Examples

Some examples of Statistics:

- Single observation of the sample:X_1
- order statistics:$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$
- Sample mean:$\bar{X_n}$.
- Sample variance:$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X_n})^2$
- sample minimum:$X_{(1)}$
- sample maximum:$X_{(n)}$
- sample rande:$X_{(n)} - X_{(1)}$

What is a "good" statistics?

Properties of Statistics

- Recall that we are doing parametric inference,where the model is
$$f_X(X;\theta)$$
- We hope the statistics can be a summary of all the the data, relevant to the parameter.
- The process can be seen as a data reduction process
- Properties:
  - Sufficient statistics
  - Ancillary statistics
  - Complete statistics

Sufficient statistics

- Reduce the data,so that all the information relevant to the parameter can be summarized in one statistic

Sufficiency Principle

Let $X = (X_1, X_2, \cdots, X_n)$ be random sample from the distribution $f(x;\theta)$.If $T(x)$ is a sufficient statistic for θ,then any inference about θ should depend upon the sample X only through the value of $T(X)$

- We say,$T(X)$ is sufficient for the parameter θ
- We can replace X with $T(X)$ without losing information

Definition:Suffiency/sufficient statistics

A statistic $T(X)$ is a `sufficient` for θ if the conditional diistribution of the sample X given $T(X)$ does not depend on θ,i.e.,

$$f(x_1, x_2, \cdots, x_n|t;\theta) = f(x_1, x_2, , \cdots, x_n|t).$$

The above definition is not easy to check whether a statistic $T(\mathbf{X})$ is a sufficient statistic.

Theorem 6.2.2

If $p(\mathbf{x}|\theta)$ is the pdf or pmf of $\mathbf{X}$, and $q(\mathbf{t}|\theta)$ is the pdf or pmf of $T(\mathbf{X})$, then $T(\mathbf{X})$ is a sufficient statistic for θ if, for every $\mathbf{x}$,

$$\frac{p(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)} \equiv \text{constant in} \quad \theta$$

.

Example 1

Example.$X_1, X_2, \cdots, X_n \sim Poisson(\theta)$.Let $T = \sum_{i=1}^{n} X_i$.Since Poisson dstribution is a discrete distribution,we are working with the PMF.
The conditional distribution is

$$P(x_1, x_2, \cdots, x_n|t) = \frac{P(X_1=x_1, \cdots, X_n=x_n, T=t)}{P(T=t)}$$

Since $T = \sum_{i=1}^{n} X_i$,

$$P(X_1 = x_1, \cdots, X_n = x_n, T = t) = \begin{cases} 0, & T(x) \neq t \\ P(X_1 = x_1, \cdots, X_n = x_n), & T(x) = t \end{cases}$$

And,

$$P(X^n = x^n) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{\sum x_i}}{\prod(x_i!)}$$

Now,$T(x_1, \cdots, x_n) = \sum x_i = t$.According to the property of Poisson distribution,$T \sim Poission(n\theta)$,so

$$P(X^n = x^n)/P(T = t) = t!/[\prod(x_i!)n^t].$$

which does not not depend on θ.So,T is a sufficient statistic for θ.

Example 6.2.3

Let $X_1, X_2, \cdots, X_n$ be i.i.d. Bernoulli(p). Let

$$T(\mathbf{X}) = X_1 + X_2 + \cdots + X_n.$$

Then

$$\begin{aligned}
p(\mathbf{x}|p) &= p^{x_1+\cdots+x_n}(1-p)^{n-(x_1+\cdots+x_n)} \\
q(t|p) &= \binom{n}{t} p^t (1-p)^{n-t} \\
\frac{p(\mathbf{X}|p)}{q(T(\mathbf{x})|p)} &= \frac{p^{x_1+\cdots+x_n}(1-p)^{n-(x_1+\cdots+x_n)}}{\binom{n}{x_1+\cdots+x_n} p^{x_1+\cdots+x_n}(1-p)^{n-(x_1+\cdots+x_n)}} \\
&= \frac{1}{\binom{n}{T(\mathbf{x})}} \quad \text{does not depend on } \theta
\end{aligned}$$

Example 6.2.4

Let $X_1, X_2, \cdots, X_n$ be i.i.d.$N(\mu, \sigma^2)$, where σ is unknown.

$$T(\mathbf{X}) = (X_1 + X_2 + \cdots + X_n)/n$$

is sufficient for μ.

$$\begin{aligned} f(\mathbf{x}|\mu) &= \prod_{i=1}^{n} (2\pi)^{-1/2}\sigma^{-1} \exp\left(-(x_i-\mu)^2/(2\sigma^2)\right) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left[-\sum_{i=1}^{n}(x_i-\mu)^2/(2\sigma^2)\right] \\ &= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2 - \frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right] \end{aligned}$$

$\bar{\mathbf{X}} \sim N(\mu, \sigma^2/n)$

$$f_{\bar{\mathbf{X}}}(t|\mu) = (2\pi\sigma^2/n)^{-n/2} \exp\left[-\frac{n}{2\sigma^2}(t-\mu)^2\right]$$

So

$$\frac{f(\mathbf{x}|\mu)}{f_{\bar{\mathbf{X}}}(t|\mu)} = \frac{(2\pi)^{-n/2}}{(2\pi n^{-1})^{-1/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]$$

does not depend on μ.

Example 6.2.5 (Sufficient Order Statistic)

Suppose X_1, X_2, $\cdots$, X_n are i.i.d. $f(x)$. Then

$$(X_{(1)}, X_{(2)}, \cdots, X_{(n)})$$

is sufficient for $f(\cdot)$. $(X_{(1)}, X_{(2)}, \cdots, X_{(n)})$ is the order statistic of $X_1, X_2, \ldots, X_n$.

Sufficient Partition

- The sufficient can be viewed as a proper partition of the sample space.

Example. Let $X_1, X_2, X_3 \sim Bernoulli(p)$,Let $T = \sum x_i$.

(x_1, x_2, x_3)		t	$p(x\|t)$
$(0,0,0)$	$\rightarrow$	0	1
$(1,0,0)$	$\rightarrow$	1	1/3
$(0,1,0)$	$\rightarrow$	1	1/3
$(0,0,1)$	$\rightarrow$	1	1/3
$(1,1,0)$	$\rightarrow$	2	1/3
$(0,0,1)$	$\rightarrow$	2	1/3
$(0,1,1)$	$\rightarrow$	2	1/3
$(1,1,1)$	$\rightarrow$	3	1

According to different values of T,the original sample space Ω is partitioned onto 4 subsets.

$$\Omega = \{(0,0,0)\} \bigcup \{(1,0,0),(0,1,0),(0,0,1)\}$$
$$\bigcup \{(0,1,1),(1,0,1),(1,1,0)\} \bigcup \{(1,1,1)\}$$

Sufficient Partition:Remarks

$$\Omega = \{(0,0,0)\} \bigcup \{(1,0,0),(0,1,0),(0,0,1)\}$$
$$\bigcup \{(0,1,1),(1,0,1),(1,1,0)\} \bigcup \{(1,1,1)\}$$

- In each element of the partition (each of the four subset),the conditional probability of the data does not depend on θ
- We call such a partition as sufficient partition
- This partition is introduced by the statistic T.Any statistic T can introduce a partition.
- Diffierent statistic may introduce the same partition.For example,$10\sum x_i$,$(\sum x_i)^2$ introduce the same partition introduced is also sufficient.
- T is sufficient if and only if the partition introduced is also sufficient.

Sufficient Partition:One More Example

- How about the partition induced by other statistic?

Example. Let $X_1, X_2, X_3 \sim Bernoulli(p)$.Let $T = X_1 + X_2$.Then partition introduced is as following.

(x_1, x_2, x_3)		t	$p(x\|t)$
$(0,0,0)$	$\to$	0	$1-p$
$(0,0,1)$	$\to$	0	p
$(1,0,0)$	$\to$	1	$(1-p)/2$
$(0,1,0)$	$\to$	1	$(1-p)/2$
$(0,1,1)$	$\to$	1	$p/2$
$(1,0,1)$	$\to$	1	$p/2$
$(1,1,0)$	$\to$	2	$1-p$
$(1,1,1)$	$\to$	2	p

The sample space is decomposed into a 3-element partition.However,in this partition,the conditional dstribution still depends on p.This is not a sufficient partition,and T is not a sufficient statistic.

The Factorization Theorem

How to find a sufficient statistics?

The Factorization Theorem

Let $f_X(x;\theta)$ be the density of a random sample.A statistic $T(X)$ is sufficient for θ if and only if there exist functions $g(t;\theta)$ and $h(x)$,such that for any (x,θ),

$$f_X(x;\theta) = g(T(X);\theta)h(x)$$

- $f_X(x;\theta)$ is the joint density for the random sample $x_1, \cdots, x_n$,
- The density function can be seen as a product of function about T and θ,and function about x only.
- No need to calculate the conditional distribution.

This theory is most useful in finding out sufficient statistic

Proof

We prove it assuming X is discrete;the condinous case is similar.

■ "Only if":Let T be sufficient.Choose $g(t;\theta) = P(T(X) = t;\theta)$ and $h(x) = P(X = x|T(X) = T(x))$.Since T is sufficient,$h(x)$ does not depend on θ.

$$\begin{aligned} f_X(x;\theta) &= P(X = x;\theta) = P(X = x|T(X) = T(x);\theta) \\ &= P(X = x|T(X) = T(x);\theta)P(T(X) = T(x);\theta) \\ &= P(X = x|T(X) = T(x))P(T(X) = T(x);\theta) \\ &= h(x)g(T(x);\theta). \end{aligned}$$

■ "if":suppose the factorization holds,and we want to show T is sufficient for θ.Let $A_{T(X)} = y;T(y) = T(x)$,then consider

$$\begin{aligned} \frac{f_X(x;\theta)}{f_T(t;\theta)} &= \frac{h(x)g(T(x);\theta)}{f_T(t;\theta)} = \frac{h(x)g(T(x);\theta)}{\sum_{u\in A_{T(x)}}} h(u)g(T(u);\theta) \\ &= \frac{h(x)g(T(x);\theta)}{g(T(x);\theta)\sum_{u\in A_{T(x)}} h(u)} = \frac{h(x)}{\sum_{u\in A_{T(x)}} h(u)} \end{aligned}$$

The conditional distribution does not depend on θ,hence T is sufficient for θ.

- **Example 6.2.7** $X_1, X_2, \ldots, X_n$ i.i.d. $N(\mu, \sigma^2)$, σ known, we have

$$f(\mathbf{x}|\mu) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right] \exp\left[-\frac{n}{2\sigma^2}(\bar{x} - \mu)^2\right]$$

since $\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]$ does not involve μ,
$\bar{\mathbf{X}} = \frac{1}{n}(X_1 + \ldots + X_n)$ is a sufficient statistic for μ.

Example 6.2.8: Uniform Sufficient Statistic

Let $X_1, X_2, \ldots, X_n$ be i.i.d. observations from discrete Uniform distribution on 1, 2, $\cdots, \theta$.

$$f(x|\theta) = \begin{cases} 1/\theta, & x = 1, 2, \cdots, \theta \\ 0, & \text{otherwise} \end{cases}$$

Thus the joint pmf of $X_1, \ldots, X_n$ is

$$f(\mathbf{x}|\theta) = \begin{cases} \theta^{-n}, & x_i \in \{1, 2, \cdots, \theta\} \text{ for } i = 1, 2, \cdots, n \\ 0, & \text{otherwise} \end{cases}$$

Let

$$\begin{aligned} f(\mathbf{x}|\theta) &= \theta^{-n} I(x)_{\{1,2,\ldots,\theta\}} = \theta^{-n} I(\max\{x_i\})_{\{\max\{x_i\} \leq \theta\}} \\ g(t|\theta) &= \theta^{-n}, t \leq \theta \\ &= \theta^{-n} \cdot 1[t \leq \theta] \end{aligned}$$

Then

$$f(\mathbf{x}|\theta) = g\left(\max_{1 \leq i \leq n} \{x_i\}|\theta\right) \cdot h(\mathbf{x})$$

$\Longrightarrow T(\mathbf{X}) = \max_{1 \leq i \leq n} \{X_i\}$ is a sufficient statistic for θ.

Example 6.2.9

$X_1, \cdots, X_n \sim N(\mu, \sigma^2)$.

$$
\begin{aligned}
f(\mathbf{x}|\mu,\sigma^2) &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2 - \frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right\} \\
&= h(\mathbf{x})g(T_1(\mathbf{x}), T_2(\mathbf{x})|\mu,\sigma^2)
\end{aligned}
$$

Here,

$$
\begin{aligned}
&h(\mathbf{x}) \equiv 1 \\
&g(t_1,t_2|\mu,\sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{n-1}{2\sigma^2}\cdot t_2 - -\frac{n-1}{2\sigma^2}(t_1-\mu)^2\right\}
\end{aligned}
$$

Hence, $T_1(\mathbf{x}) = \bar{X}$, $T_2(\mathbf{x}) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$ are sufficient statistics.

Theorem 6.2.10

Let $X_1, X_2, \cdots, X_n$ be i.i.d. observations from a pdf or pmf $f(x|\boldsymbol{\theta})$ that belongs to an exponential family given by

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

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

$$T(\mathbf{X}) = \left(\sum_{j=1}^{n} t_1(X_j), \sum_{j=1}^{n} t_2(X_j), \cdots, \sum_{j=1}^{n} t_k(X_j)\right)$$

is a sufficient statistic for $\boldsymbol{\theta}$

- **Example** Let $X_1, X_2, \cdots, X_n$ be i.i.d. Gamma(α, β), then $T(\mathbf{X}) = \left(\sum_{j=1}^{n} \log X_j, \sum_{j=1}^{n} X_j\right)$ are sufficient for (α, β).
- **Example** Let $X_1, X_2, \cdots, X_n$ be i.i.d. Uniform(α, β), $\alpha < \beta$, then $(\min_{1 \leq i \leq n} X_i, \max_{1 \leq i \leq n} X_i)$ is sufficient for (α, β).

Minimal Sufficient Statistics(MSS)

- There are multiple sufficient statistics for one parameter
- Example:$X_1, X_2, X_3 \sim Bernoulli(p)$.For $p, \sum_{i=1}^{3} X_i, (\sum_{i=1}^{3} X_i)^2, (X_1 + X_2, X_3)$ are all sufficient statistics
- Which is the "best" one for us?

Recall:

- Sufficient statistics:data reduction
- Best:the sufficient statistics that maximal the data reduction
- The "best" statistics has minimal data but sufficient information.We call it the minimal sufficient statistics.

Minimal Sufficient Statistic

A statistic T is called a Minimal Sufficient Statistic if

- T is sufficient;
- For any other sufficient statistics U,$T = g(U)$ for some function g.

- For a fixed family of distribution, many sufficient statistics exist. We need to find the sufficient statistic which achieves the maximal data reduction.
- First, any one-to-one transformation of sufficient statistic is a sufficient statistic.

MSS:Example

Example:Let $X_1, X_2, X_3 \sim Bernoulli(p)$.Let $T = \sum_{i=1}^{3} X_i$, $U = 2X_1 + 3X_2 + 4X_3$.

(x_1, x_2, x_3)		t	$p(x\|t)$	u	$p(x\|u)$
$(0,0,0)$	$\rightarrow$	0	1	0	1
$(1,0,0)$	$\rightarrow$	1	$1/3$	2	1
$(0,1,0)$	$\rightarrow$	1	$1/3$	3	1
$(0,0,1)$	$\rightarrow$	1	$1/3$	4	1
$(0,1,1)$	$\rightarrow$	2	$1/3$	7	1
$(1,0,1)$	$\rightarrow$	2	$1/3$	6	1
$(1,1,0)$	$\rightarrow$	2	$1/3$	5	1
$(1,1,1)$	$\rightarrow$	2	1	9	1

Both T and U are sufficient statistics,but U is not minimal.

MSS

- How to check the minimal sufficiency?
- How to find a minimal sufficient statistic?

Theorem:Minimal Sufficient Statistics

Let $f_X(x;\theta)$ be the density of a random sample X,Let

$$R(x,y;\theta)=\frac{f_X(x;\theta)}{f_Y(y;\theta)}$$

For a statistic T,T is minimal sufficient if $R(x,y;\theta)$ does not depend on $\theta \Leftrightarrow T(x)=T(y)$.

- Here,x and y are two random samples with the same sample size
- Sometimes,it is hard to show the equivalence.

MSS:Example

■ Let $X_1, X_2, \cdots, X_n \sim Poisson(\theta), Y_1, Y_2, \cdots, Y_n \sim Poisson(\theta)$,then

$$p(x;\theta) = \frac{e^{-n\theta}\theta^{\sum x_i}}{\prod y_i!}, \qquad R(x,y;\theta) = \frac{\theta^{\sum y_i - \sum x_i}}{\prod y_i! / \prod x_i!}$$

It is independent with θ if and only if $(\sum y_i = \sum x_i)$,So $T = \sum x_i$ is a minimal sufficient statistic for θ.

■ Let $X_1, \cdots, X_n$ be a random sample with Cauchy distribution.Recall for Cauchy distribution,the PDF is $f(x;\theta) = \frac{1}{\pi(1+(x-\theta)^2)}$.So,the ratio is

$$R(x,y;\theta) = \frac{f(x;\theta)}{f(y;\theta)} = \frac{\prod 1/[\pi(1+(x_i-\theta)^2)]}{\prod 1/[\pi(1+(y_i-\theta)^2)]} = \frac{\prod 1/[1+(y_i-\theta)^2]}{\prod 1/[1+(x_i-\theta)^2]}$$

The result cannot be further reduced.However,note that the final result is not sffected by the order of the fata.Therefore,R does not depend on θ if and only if $(x_{(1)}, x_{(2)}, \cdots, x_{(n)}) = (y_{(1)}, y_{(2)}, \cdots, y_{(n)})$.The sufficient statistic is $T = (X_{(1)}, X_{(2)}, \cdots, X_{(n)})$.

- **Example 6.2.14** $X_1, X_2, \cdots, X_n$ be i.i.d. $N(\mu, \sigma^2)$. Then $\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$ and $S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ are minimal sufficient for μ, σ^2.

- **Example 6.2.15** $X_1, X_2, \cdots, X_n$ be i.i.d. Uniform(θ, $\theta + 1$). Then

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = \begin{cases} 1, & \max_i x_i - 1 < \theta < \min_i x_i \\ 0, & \text{otherwise} \end{cases}$$

This implies (Theorem 6.2.13) that $\left(\max_i X_i, \min_i X_i\right)$ is minimal sufficient for θ.

Remark 1 The above is an example of two-dimensional minimal sufficient statistic for one-dimensional parameter.

Remark 1 Any one-to-one function of minimal sufficient statistic is also a minimal sufficient statistic.

Ancillary Statistics

- Sufficient statistic:the statistics that contain all information about θ.
- Ancillary statistics:the statistics does not depend on θ.

Definition:Ancillary Statistics

a statistic $S(X)$ of a random sample whose distribution does not depend on θ is called an ancillary statistics.

Example Let $X_1, X_2, \cdots, X_n$ be i.i.d. Uniform(θ, $\theta+1$), we see that (from Example 6.1.15) $X_{(n)}, X_{(1)}$ are minimal sufficient for θ. Therefore $\left(X_{(n)} - X_{(1)}, \frac{X_{(n)}+X_{(1)}}{2}\right)$ are minimal sufficient for θ. But Example 6.1.17 shows that $X_{(n)} - X_{(1)}$ is ancillary for θ.

Remark An ancillary statistic by itself may contain no information on θ, but when combine with other statistics, it may offer very important information. It is certainly not true that ancillary statistics are independent of minimal sufficient statistics.

- **Example 6.2.17** $X_1, X_2, \cdots, X_n$ are i.i.d. Uniform(θ, $\theta+1$), $-\infty < \theta < \infty$. Then $R = X_{(n)} - X_{(1)}$ is ancillary.

  Answer The joint pdf of $(X_{(n)}, X_{(1)})$ is

$$g\left(x_{(1)}, x_{(n)}|\theta\right) = n(n-1)\left(x_{(n)} - x_{(1)}\right)^{n-2}, \quad \theta < x_{(1)} < x_{(n)} < \theta+1.$$

Let

$$\begin{cases} R = X_{(n)} - X_{(1)}, \\ M = (X_{(1)} + X_{(n)})/2, \end{cases}$$

then

$f_{R,M}(r,m) = n(n-1)r^{n-2}, \;\; 0 < r < 1, \theta+(r/2) < m < \theta+1-(r/2).$

So the marginal distribution of R is

$$f_R(r) = \int_{\theta+r/2}^{\theta+1-r/2} n(n-1)r^{n-2}dm = n(n-1)r^{n-2}(1-r), \;\; 0 < r < 1.$$

$\Longrightarrow$ The pdf of R does not depend on θ. So R is ancillary for θ.

- Example 6.2.18 (Location Family Ancillary Statistic)
  $X_1, X_2, \cdots, X_n$ are i.i.d with cdf $F(x-\theta)$, $-\infty < \theta < \infty$. F is a known distribution function. In this case $R = X_{(n)} - X_{(1)}$ is ancillary for θ.
- Example 6.2.19 (Scale Family Ancillary Statistic)
  Let $X_1, X_2, \cdots, X_n$ be i.i.d from $F(x/\sigma)$, $\sigma > 0$. Then any statistic that depends on the sample through the $n-1$ values $X_1/X_n, \cdots, X_{n-1}/X_n$ is an ancillary statistic. For example, $(X_1 + \cdots + X_n)/X_n$ is ancillary. The joint distribution of $X_1/X_n, \cdots, X_{n-1}/X_n$ are

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

  does not depend on σ. $Z_1, \cdots, Z_n$ are i.i.d. from $F(x)$.
  Remark Ancillary statistic may still useful in estimation of θ. One example is that $X_1, X_2, \cdots, X_n$ i.i.d $N(\mu, \sigma^2)$ with σ^2 unknown. Then $T_1(\mathbf{X}) = \frac{1}{n}(X_1 + \cdots + X_n)$ is minimal sufficient for μ. But the variance estimate of $T_1(\mathbf{X})$ depends on S_n^2, which is ancillary for μ.

Complete Statistics

Definition:Complete Statistics

Let X be a random sample with density $f_X(x;\theta)$ and T a statistic with density $f_T(t;\theta)$.The collection of densities f_X is called complete if

$$E_\theta[g(T)] = 0 \Rightarrow P_\theta[g(T) = 0] = 1 \qquad g : T \to \mathbb{R}, \theta \in \Theta.$$

T is called a Complete Statistics.

Remark.

- g is a fixed function.Say,$g(x) = x$.There is no randomness for g.The randomness of $g(T)$ comes from T.
- g does not depend on θ.
- g:a function so that $E_\theta[g(T)] = 0$ for any $\theta \in \Theta$.For any g statisfying such condition,$g(T) = 0$ with probability 1 for any θ.
- The statistic is the statistic which ensures θ is identifiable.

Complete Statistics:Example

Example.Let $X_1, X_2, X_3 \sim Bernoulli(p), \theta \in (0,1)$.Prove that $T = \sum X_i$ is complete.

Proof: Suppose that $T3 \sim Bernoulli(n, \theta), \theta \in (0,1)$ and g be such that $E_\theta[g(T)] = 0$.Then we must have

$$
\begin{aligned}
0 = E_\theta[g(T)] &= \sum_{t=0}^{n} g(t) \binom{n}{t} \theta^t (1-\theta)^{n-t} \\
&= (1-\theta)^n \sum_{t=0}^{n} g(t) \binom{n}{t} (\frac{\theta}{1-\theta})^t \\
&= (1-\theta)^n \sum_{t=0}^{n} g(t) \binom{n}{t} r^t
\end{aligned}
$$

where $r = \theta/(1-\theta)$.Let r be a very small number so that $g(0)\binom{n}{0}r^0$ term be the giant component,then since the summation is 0,obviously $g(0) = 0$.Similarly,we show that $g(t) = 0$ for each $t \in 0, \cdots, n$ must hold.Hence,T is complete.

Complete Statistics:Example 6.2.23

Let $X_i \sim Unif(0,\theta), i \in 1,\cdots,n$,for $\theta > 0$.Recall that $T = X_{(n)}$(the maximum of the sample) is sufficient for θ.Now,we want to prove that T is also complete.

Proof.The CDF of t is

$$F_T(t) = P(T \leq t) = P()\max X_1, X_2, \cdots, X_n \leq t) = (\tfrac{t}{\theta})^{\theta}$$

so the PDF of t is the derivative of F_T,which is $\frac{nt^{n-1}}{\theta^n}, 0 < t < \theta$.Suppose that $g(t)$ statifies that $E_\theta[g(T)] = 0$,then $\int_0^\theta g(t)\frac{nt^{n-1}}{\theta^n}dt = 0$.Since it stands for all θ,the dericative of $E_\theta[g(T)]$ also equals to 0.

$$0 = \frac{d}{d\theta}\int_0^\theta g(t)\frac{nt^{n-1}}{\theta^n}dt =$$
$$\frac{d}{d\theta}(\theta^{-n})\int_0^\theta g(t)nt^{n-1}dt + \frac{d}{d\theta}(\int_0^\theta g(t)nt^{n-1}dt)(\theta^{-n})$$

The first part equals to 0,since $\int_0^\theta g(t)\frac{nt^{n-1}}{\theta^n}dt = 0$..So we have

$$0 = \frac{d}{d\theta}(\int_0^\theta g(t)nt^{n-1}dt)(\theta^{-n}) = g(\theta)n\theta^{n-1}.$$

So,$g(\theta) = 0$ for any $\theta > 0$,which means that $g(x) = 0$ when $x > 0$.Recall that $T > 0$ with probability 1,so $g(T) = 0$ with probability 1,for any θ.

- Theorem 6.2.24 (Basu's Theorem) If $T(\mathbf{X})$ is a complete and minimal sufficient statistic, then $T(\mathbf{X})$ is independent of every ancillary statistic.
- Theorem 6.2.25 (Complete Statistics in the exponential Family) Let $X_1, X_2, \cdots, X_n$ be i.i.d. observations from an exponential family with pdf or pmf of the form

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

where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k)$. Then the statistic

$$T(\mathbf{X}) = \left(\sum_{i=1}^{k} t_1(X_i), \sum_{i=1}^{k} t_2(X_i), \cdots, \sum_{i=1}^{k} t_k(X_i)\right)$$

is complete as long as the parameter space Θ contains an open set in R^k.

- Theorem 6.2.28 If a minimal sufficient statistic exists, then any complete statistic is also a minimal sufficient statistic.
- Example 6.2.26 (Using Basu's Theorem I) To show $g(\mathbf{X}) = \frac{X_n}{X_1+\cdots+X_n}$. and $T(\mathbf{X}) = X_1 + \cdots + X_n$ are independent when $X_1, X_2, \cdots, X_n$ are i.i.d. exp(mean=θ).
- Example 6.2.27 (Using Basu's Theorem II) To show $\bar{X}_n$ and S^2 are independent if $X_1, X_2, \cdots, X_n$ are i.i.d. $N(\mu, \sigma^2)$.

Remarks

■ Sufficient statistics,ancillary statistics,and complete statistics are the statistics for data reduction

■ In past days,when the space is not enough

■ Sufficient statistics is to reduce data so that estimation through likelihood is doable.

■ Ancillary statistics is to figure out the part that not related to θ

■ Complete statistics is to make sure that θ is identifiable (no two θ with exactly the same model)

■ Reduce the samples to be only these statistics

■ Currently,thanks to the techonology development,saving the data is not that difficult.These statistics are used to help understand the model and the data and accelerate the algorithm.

■ Comlpete statistics and ancillary statistics are not popular now.