Ch 6. Principles of Data Reduction

Introduction

- Inference about the population usually means inference about $\boldsymbol{\theta}$.
- From the random sample $X_1, \cdots, X_n$, want to extract all information about $\boldsymbol{\theta}$
- Do we need all measurements of $X_1 = x_1, \cdots, X_n = x_n$ or just a couple of (or even one) numbers to provide the information about $\boldsymbol{\theta}$ that the sample has ?
- Do you want to have 20,000 cents or 2 one hundred dollar notes ?

Principle: If $T(\mathbf{X})$ (Vector or Scalar) is a sufficient statistic for $\boldsymbol{\theta}$, then any information about $\boldsymbol{\theta}$ should depend on the sample $\mathbf{X} = (X_1, \cdots, X_n)$ only through the value of $T(\mathbf{X})$. That is, if $\mathbf{x}$ and $\mathbf{y}$ are two points such that $T(\mathbf{x}) = T(\mathbf{y})$, then the inference should be the same whatever $\mathbf{X} = \mathbf{x}$ or $\mathbf{Y} = \mathbf{y}$ is observed.

Definition

A statistic $T(\mathbf{X})$ is a sufficient statistic for $\boldsymbol{\theta}$ if the conditional distribution of sample $\mathbf{X}$ given $T(\mathbf{X})$ does not depend on $\boldsymbol{\theta}$.

▷ Example: $X_1, \cdots, X_n \overset{iid}{\sim}$ Bernoulli(θ). $T(\mathbf{X}) = \sum_{i=1}^n X_i$.

Proof: See Example 6.2.3.

• Checking or defining a sufficient statistic using the definition is tricky especially for continuous distributions.

Theorem

If $f(\mathbf{x}|\boldsymbol{\theta})$ is the joint pdf or pmf of $\mathbf{X}$ and $q(\mathbf{t}|\boldsymbol{\theta})$ is the pdf or pmf of $T(\mathbf{X})$, then T is sufficient for $\boldsymbol{\theta}$ if and only if the ratio

$$\frac{f(\mathbf{x}|\boldsymbol{\theta})}{q(\mathbf{t}|\boldsymbol{\theta})}$$

is independent of $\boldsymbol{\theta}$.

$\triangleright$ Example 1: $X_1, \cdots, X_n \overset{iid}{\sim}$ Exponential(θ). $T(\mathbf{X}) = \sum_{i=1}^n X_i$.

$\triangleright$ Example 2: $X_1, \cdots, X_n \overset{iid}{\sim}$ Normal(μ, σ^2), where σ known. $T(\mathbf{X}) = \sum_{i=1}^n X_i$. [Example 6.2.4.]

▷ Example 1: $X_1, \cdots, X_n \overset{iid}{\sim} \text{Uniform}(0, \theta)$.

$$T(\mathbf{X}) = X_{(n)} = \max_i X_i$$

▷ Example 2:

$X_1, \cdots, X_n \overset{iid}{\sim}$ Cauchy distribution with $f(x|\theta) = \frac{1}{\pi(x-\theta)^2}$.

$$T(\mathbf{X}) = (X_{(1)}, \ldots, X_{(n)})$$

(Solutions: Find the ratio of)

$$\frac{f(\mathbf{x}|\boldsymbol{\theta})}{q(\mathbf{t}|\boldsymbol{\theta})}$$

Theorem (Factorization Theorem (Theorem 6.2.6))

Let $f(\mathbf{x}|\boldsymbol{\theta})$ denote the joint pdf or pmf of $\mathbf{X}$. $T(\mathbf{X})$ is sufficient for $\boldsymbol{\theta}$ if and only if there exist functions $g[\mathbf{t}(\mathbf{x})|\boldsymbol{\theta}]$ and $h(\mathbf{x})$ such that for all sample points $\mathbf{x}$ and all $\boldsymbol{\theta}$

$$f(\mathbf{x}|\boldsymbol{\theta}) = g[\mathbf{t}(\mathbf{x})|\boldsymbol{\theta}]h(\mathbf{x}).$$

▷ Note:

1. The choice of g and h is not unique.
2. A trivial sufficient statistic is $T(\mathbf{X}) = \mathbf{X}$.
3. If T is sufficient then any one to one mapping of T is also sufficient.

▷ Example: $X_1, \cdots, X_n \stackrel{iid}{\sim}$ Bernoulli(θ). $T(\mathbf{X}) = \sum_{i=1}^n X_i$

▷ Example: $X_1, \cdots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$. σ^2 is known.
Proof: See Example 6.2.9.

▷ Example: $X_1, \cdots, X_n \overset{iid}{\sim}$ Discrete Uniform on$(1, 2, \cdots, \theta)$, where θ is a positive integer.

▷ Example: $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Both μ and σ^2 are unknown.

Number of parameters = Number of sufficient statistics ?

▷ Example: $X_1, \cdots, X_n \overset{iid}{\sim}$ Uniform$(-\theta, \theta)$.

Theorem

$X_1, \cdots, X_n \overset{iid}{\sim} f(\mathbf{x}|\boldsymbol{\theta})$, where $f(\mathbf{x}|\boldsymbol{\theta})$ belongs to an exponential family. That is

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

Then

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

is sufficient for $\boldsymbol{\theta}$.

In the normal example, both $\mathbf{X} = (X_1, \cdots, X_n)$ and $(\bar{X}, \sum X_i^2)$ are sufficient for $\boldsymbol{\theta} = (\mu, \sigma^2)$. Which one do we use and why ?

Definition

A sufficient statistic $T(\mathbf{X})$ is *minimal sufficient* if it can be written as a function of every other sufficient statistic.

▷ Note:

1. A one to one transformation of a minimal sufficient statistic is also a minimal sufficient statistic. (Not unique.)
2. A minimal sufficient statistic reduces data to the greatest extent w/o losing useful information for making inference about $\boldsymbol{\theta}$.

How to show the given statistic is minimal sufficient ?

Theorem

Let $f(\mathbf{x}|\boldsymbol{\theta})$ denote the joint pdf or pmf of $\mathbf{X}$. Suppose there exist a function $T(\mathbf{x})$ such that for any two sample points $\mathbf{x}$ and $\mathbf{y}$, the ratio

$$\frac{f(\mathbf{x}|\boldsymbol{\theta})}{f(\mathbf{y}|\boldsymbol{\theta})}$$

is a constant as a function of $\boldsymbol{\theta}$ if and only if $T(\mathbf{x}) = T(\mathbf{y})$. Then $T(\mathbf{X})$ is minimal sufficient for $\boldsymbol{\theta}$.

- ▷ Example: $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$
- ▷ Example: $X_1, \cdots, X_n \overset{iid}{\sim} \text{Uniform}(\theta, \theta + 1)$

Definition

A statistic $S(\mathbf{X})$ whose **distribution** does not depend on the parameter $\boldsymbol{\theta}$ is called *ancillary* statistic.

▷ Note:

1. Ancillary statistic can be thought as having no information about $\boldsymbol{\theta}$ itself.
2. But if it is used in conjunct with other statistics, sometimes it gives a valuable information.

▷ Example 1: $X_1, \cdots, X_n \overset{iid}{\sim} f(x|\theta) = h(x-\theta)$, $h(\cdot)$ is known. $-\infty < \theta < \infty$. [Location family] Then

$$f(\mathbf{x}|\theta) = \prod_{i=1}^{n} h(x_i - \theta).$$

Let $Z_i = X_i - \theta$. Then $f_Z(z) = h(z)$ which is independent from θ. [See Theorem 3.5.6]

▷ Example 2: $X_1, \cdots, X_n \overset{iid}{\sim} U(\theta, \theta+1), -\infty < \theta < \infty$. Are $R = X_{(n)} - X_{(1)}$ and $S_X^2 = \sum (X_i - \bar{X}_n)^2/(n-1)$ ancillary statistics?

Definition

Let $f(t|\theta)$ be a family of pdfs or pmfs for a statistic $T(\mathbf{X})$. **The family of probability distributions** is called complete if $E_\theta[g(T)] = 0$ for all θ implies $P_\theta[g(T) = 0] = 1$ for all θ. Equivalently, $T(\mathbf{X})$ is called a complete statistic.

▷ Note:

1. Completeness is a property of a family of probability distributions, not of a particular distribution.
2. The family of $N(\theta, 1), -\infty < \theta < \infty$ is complete.

▷ Example 6.2.22: $X_1, \cdots, X_n \overset{iid}{\sim}$ Bernoulli(p), $0 < p < 1$. $T(\mathbf{X}) = \sum X_i$.

▷ Example 6.2.23: $X_1, \cdots, X_n \overset{iid}{\sim} U(0, \theta)$, $0 < \theta < \infty$. $T(\mathbf{X}) = \max_i X_i$.

▷ Example: $X_1, \cdots, X_n \overset{iid}{\sim}$ Poisson(λ), $0 < \lambda$. $T(\mathbf{X}) = \sum X_i$.

So far, we have investigated *sufficient*, *ancillary* and *complete* statistics. We know that sufficient statistic has a whole information about the parameter $\boldsymbol{\theta}$ but ancillary has none. Can we say they are independent each other ?

Theorem (Basu's Theorem)

If $T(\mathbf{X})$ is complete, minimal sufficient then $T(\mathbf{X})$ is independent of every ancillary statistics.

▷ Example: $X_1, \cdots, X_n \overset{iid}{\sim}$ Uniform$(\theta, \theta + 1)$. $T(\mathbf{X}) = (X_{(1)}, X_{(n)})$.

Completeness is a property that is often hard to prove. But, we have a nice theorem for an exponential family.

Theorem

$X_1, \cdots, X_n \overset{iid}{\sim} f(\mathbf{x}|\boldsymbol{\theta})$, where $f(\mathbf{x}|\boldsymbol{\theta})$ belongs to an exponential family. That is

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

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

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

is complete.

Ch 6. Principles of Data Reduction

▷ Example: $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. σ^2 is fixed. Then

$\bar{X}$ is (sufficient and complete.)

The distribution of S^2 is independent of μ.

(Or, S^2 is an ancillary statistic.)

Thus, $\bar{X}$ is independent of S^2.

Theorem

Any complete statistic is also a minimal sufficient statistic if a minimal sufficient statistic exists.

[We cover Chapter 6.1 and 6.2 of *Casella and Berger* in STA513.]