STAT3602 Statistical Inference

Example Class 7: Sufficiency and Completeness

LIU Chen

Department of Statistics and Actuarial Science, HKU

04/11/2020

Outline

1. Chapter 4: Sufficiency and Likelihood
2. Exercise 1
3. Exercise 2
4. Exercise 3
5. Exercise 4

Chapter 4: Sufficiency and Likelihood I

1. **Likelihood and loglikelihood functions of θ.**

$$\ell_{\boldsymbol{X}}(\theta) \propto f(\boldsymbol{X} \mid \theta)$$

$$S_{\boldsymbol{X}}(\theta) = \ln \ell_{\boldsymbol{X}}(\theta)$$

2. **Sufficiency.**
   Statistic $T = T(X)$ is sufficient for θ if the conditional distribution of X given T is free of θ.

3. **Theorem 4.2.6.** The following three statements are equivalent:

- Statistic $T = T(\boldsymbol{X})$ is sufficient for θ.
- **Likelihood Ratio Criterion**.
  For any samples X, X', $T(X) = T\left(X'\right) \Rightarrow \ell_X(\theta) \propto \ell_{X'}(\theta)$.
- **Factorization Theorem.** There exists a function $g(\cdot)$ such that for any sample $\boldsymbol{X}, \ell_{\boldsymbol{X}}(\theta) \propto g(T(\boldsymbol{X}), \theta)$.

4. **Minimal Sufficiency.**
   1. **Definition:** $T(X)$ is minimal sufficient for θ if it is sufficient and is a function of every other sufficient statistic.
   2. **Theorem 4.3.2**
      $T(\boldsymbol{X})$ is minimal sufficient for θ if and only if

$$\text{for any samples } \boldsymbol{X}, \boldsymbol{X}', \quad T(\boldsymbol{X}) = T\left(\boldsymbol{X}'\right) \Leftrightarrow \ell_{\boldsymbol{X}}(\theta) \propto \ell_{\boldsymbol{X}'}(\theta)$$

5. **Sufficiency for Exponential Family**
   $\boldsymbol{X} \sim f(\boldsymbol{x} \mid \boldsymbol{\pi}) \propto h(\boldsymbol{x}) \exp\left\{\sum_{j=1}^{k} \pi_j t_j(\boldsymbol{x})\right\}$.
   Assume that the natural parameter space Π is not contained in an affine hyperplane of the form $\left\{\boldsymbol{\pi} : \sum_{i=1}^{k} c_i \pi_i = b\right\} \subset \mathbb{R}^k$ for c_i 's not all equal to 0.
   Then $T(\boldsymbol{X}) = (t_1(\boldsymbol{X}), \ldots, t_k(\boldsymbol{X}))$ is minimal sufficient for $\boldsymbol{\pi}$.

Chapter 4: Completeness I

1. **Definition.**
   A sufficient statistic $T = T(\boldsymbol{X})$ is complete for θ if and only if, for any real function $g(\cdot)$ not depending on θ,

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

   The latter condition says that $g(\cdot)$ is a zero function almost surely under any θ.

2. **Lehmann-Scheffé Theorem.**
   If $T(\boldsymbol{X})$ is complete sufficient for θ, then $T(\boldsymbol{X})$ is minimal sufficient.

3. **Theorem for (Exponential family)**
   If the natural parameter space Π contains an open rectangle, i.e. a nonempty set of the form $(a_1, b_1) \times \cdots \times (a_k, b_k) \subset \mathbb{R}^k$, then the natural statistic $T(\boldsymbol{X}) = (t_1(\boldsymbol{X}), \ldots, t_k(\boldsymbol{X}))$ is complete for the natural parameter $\pi = (\pi_1, \ldots, \pi_k)$.

Outline

1 Chapter 4: Sufficiency and Likelihood

2 Exercise 1

3 Exercise 2

4 Exercise 3

5 Exercise 4

Exercise 1 I

Let $X_1, \ldots, X_n$ be independent Poisson random variables with X_j having parameter $j\lambda$, where $\lambda > 0$ is an unknown parameter.

1. Find a minimal sufficient statistic for λ.
2. What is its distribution?
3. Is it a complete sufficient statistic?

Exercise 1: Solution I

The likelihood function for λ is

$$\ell_{\mathcal{X}}(\lambda) = \prod_{j=1}^{n} \frac{e^{-j\lambda}(j\lambda)^{X_j}}{X_j!} = e^{-\lambda n(n+1)/2} \lambda^{\sum X_j} \prod_j \frac{j^{X_j}}{X_j!}$$

Consider

$$\begin{aligned} \Lambda_{\mathcal{X}}\left(\lambda, \lambda'\right) &= \frac{e^{-\lambda n(n+1)/2} \lambda^{\sum X_j}}{e^{-\lambda' n(n+1)/2} \lambda'^{\sum X_j}} \\ &= e^{-(\lambda-\lambda')n(n+1)/2} \left(\frac{\lambda}{\lambda'}\right)^{\sum X_j} \end{aligned}$$

Thus, $\Lambda_{\mathcal{X}}\left(\lambda, \lambda'\right) \equiv \Lambda_{\mathcal{X}'}\left(\lambda, \lambda'\right)$ if and only if $\sum X_j = \sum X_j'$. Therefore, $T(\mathcal{X}) = \sum_{j=1}^{n} X_j$ is minimal sufficient for λ.

Exercise 1: Solution II

Given $X_j \sim \text{Poisson}(j\lambda), \sum_{j=1}^{n} X_j \sim \text{Poisson}\left(\sum_{j=1}^{n} j\lambda = \frac{n(n+1)}{2}\lambda\right)$. To show T is complete let $\mu = \frac{n(n+1)}{2}\lambda$

$$E_\mu g(T) \equiv 0$$
$$\sum_{k=0}^{\infty} g(k)e^{-\mu \frac{\mu^k}{k!}} \equiv 0$$
$$\sum_{k=0}^{\infty} \frac{g(k)}{k!}\mu^k \equiv 0$$

It is well-known that two power series $\sum_0^\infty a_j x^j$ and $\sum_0^\infty b_j x^j$ have the same value for every x in some interval if and only if $a_0 = b_0, a_1 = b_1, \ldots$ since

$$\sum_{k=0}^{\infty} (0)\mu^k \equiv 0$$

Exercise 1: Solution III

we conclude that $q(k)/k! = 0$, i.e. $q(k) = 0$ for $k = 0, 1, 2, \ldots$
In fact, for distributions in exponential family form with the natural parameter space containing an open rectangle, the natural statistic is a complete sufficient statistic.
Therefore, in the question, as $\pi = \log\lambda$ is the natural parameter and the natural parameter space $\Pi = (-\infty, \infty)$ contains an open rectangle, we are ready to conclude that $T = \sum X_i$ is a complete sufficient statistic.

Outline

1. Chapter 4: Sufficiency and Likelihood
2. Exercise 1
3. Exercise 2
4. Exercise 3
5. Exercise 4

Exercise 2

Let $X_1, \ldots, X_n$ be i.i.d. with a common density function

$$f(x) = \begin{cases} \dfrac{2}{3\theta}, & 0 \leq x \leq \theta/2 \\ \dfrac{4}{3\theta}, & \theta/2 < x \leq \theta \end{cases}$$

Let Y_n be the maximum of $X_1, \ldots, X_n$, and N be the number of observations less than $Y_n/2$. Rewrite the X_i 's which are greater than or equal to $Y_n/2$ as $Y_{N+1} \leq Y_{N+2} \leq \cdots \leq Y_n$.

1. Show that the likelihood is

$$\ell_{\mathcal{X}}(\theta) = \left(\frac{4}{3\theta}\right)^n \left(\frac{1}{2}\right)^{N + \#\{i : Y_{N+i} \leq \theta/2, i=1,\ldots,n-N\}} 1\left\{Y_n \leq \theta\right\}$$

2. Prove that $(N, Y_{N+1}, \ldots, Y_n)$ is a sufficient statistic for θ.

Exercise 2: Solution I

The likelihood function is

$$
\begin{aligned}
&\ell_{\mathcal{X}}(\theta) = f\left(X_1, \ldots, X_n \mid \theta\right) \\
&= \prod_{i=1}^{n}\left(\frac{2}{3\theta} 1\left\{0 \leq X_i \leq \theta/2\right\} + \frac{4}{3\theta} 1\left\{\theta/2 < X_i \leq \theta\right\}\right) \\
&= \left(\frac{4}{3\theta}\right) \prod_{i=1}^{n} \prod_{i=1}^{n}\left(\frac{1}{2} 1\left\{X_i \leq \theta/2\right\} + 1\left\{X_i > \theta/2\right\}\right) 1\left\{0 \leq X_i \leq \theta\right\} \\
&= \left(\frac{4}{3\theta}\right)^{n} 1\left\{\min X_i > 0\right\} 1\left\{\max X_i \leq \theta\right\} \prod_{i=1}^{n}\left(\frac{1}{2} 1\left\{X_i \leq \theta/2\right\} + 1\left\{X_i > \theta/2\right\}\right) \\
&= \left(\frac{4}{3\theta}\right)^{n} 1\left\{\min X_i > 0\right\} 1\left\{\max X_i \leq \theta\right\}\left(\frac{1}{2}\right)^{\sum_{i=1}^{n} 1\left\{X_i \leq \theta/2\right\}} \\
&= \left(\frac{4}{3\theta}\right)^{n} 1\left\{\min X_i > 0\right\} 1\left\{\max X_i \leq \theta\right\}\left(\frac{1}{2}\right)^{\sum_{i=1}^{n} 1\left\{X_i \leq \max X_i/2\right\} + 1\left\{\max X_i/2 < X_i \leq \theta/2\right\}} \\
&= \left(\frac{4}{3\theta}\right)^{n} 1\left\{Y_n \leq \theta\right\}\left(\frac{1}{2}\right)^{N + \#\left\{i: Y_{N+i} \leq \theta/2, i=1, \ldots, n-N\right\}}
\end{aligned}
$$

Exercise 2: Solution II

1. The last equality follows from the definition of N and $Y_{N+1}, \ldots, Y_n$ and the fact that all observations are larger than 0 (with probability 1).
2. The likelihood function in (a) takes the form of a function of θ and the statistic $(N, Y_{N+1}, \ldots Y_n)$ times 1, which can be regarded as a constant function of the data X. By the factorization criterion, $(N, Y_{N+1}, \ldots Y_n)$ is sufficient for θ.

Outline

1. Chapter 4: Sufficiency and Likelihood

2. Exercise 1

3. Exercise 2

4. Exercise 3

5. Exercise 4

Exercise 3 I

A married man who frequently talks on his mobile is well known to have conversations the lengths of which are independent, identically distributed random variables, distributed as exponential with mean $1/\lambda$.

His wife has long been irritated by his behavior and knows, from infinitely many observations, the exact value of λ.

In an argument with her husband, the woman produces $t_1, \ldots, t_n$, the times of n telephone conversations, to prove how excessive her husband is.

He suspects that she has randomly chosen the observations, conditional on their all being longer than the expected length of conversation. Assuming he is right in his suspicion, the husband wants to use the data he has been given to infer the value of λ.

1. What is the minimal sufficient statistic he should use?
2. Find the maximum likelihood estimator for λ.

Exercise 3: Solution I

1. The distribution of $T_1, \ldots, T_n$ is the same as $X \mid X \geq 1/\lambda$ where $X \sim \text{Exp}(\lambda)$. The density function of T_i is

$$\begin{aligned} f_T(t) &= \frac{f_X(t)}{P(X > 1/\lambda)} 1\{t \geq 1/\lambda\} \\ &= \frac{\lambda e^{-\lambda t}}{e^{-\lambda(1/\lambda)}} 1\{t \geq 1/\lambda\} \\ &= \lambda e^{-\lambda t+1} 1\{t \geq 1/\lambda\} \end{aligned}$$

2. The likelihood function is

$$\begin{aligned} \ell_{\mathcal{T}}(\lambda) &= \prod_{i=1}^{n} \lambda e^{-\lambda T_i+1} 1\left\{T_i \geq 1/\lambda\right\} \\ &= \lambda^n e^{-\lambda \sum T_i+n} 1\left\{\min T_i \geq 1/\lambda\right\} \end{aligned}$$

Exercise 3: Solution II

The likelihood ratio is

$$\Lambda_{\mathcal{T}}(\lambda_1, \lambda_2) = (\lambda_1/\lambda_2)^n e^{-(\lambda_1-\lambda_2)\sum T_i} \frac{1\{\min T_i \geq 1/\lambda_1\}}{1\{\min T_i \geq 1/\lambda_2\}}$$

$$\Lambda_{\mathcal{T}}(\lambda_1, \lambda_2) = \begin{cases} (\lambda_1/\lambda_2)^n e^{-(\lambda_1-\lambda_2)\sum T_i}, \min T_i \geq 1/\lambda_1, 1/\lambda_2 \\ \infty, \quad 1/\lambda_1 \leq \min T_i < 1/\lambda_2 \\ 0, \quad 1/\lambda_2 \leq \min T_i < 1/\lambda_1 \\ \text{ undefined }, \quad \min T_i < 1/\lambda_1, 1/\lambda_2 \end{cases}$$

Therefore, $\Lambda_{\mathcal{T}} = \Lambda_{\mathcal{T}''}$ for all $\lambda_1, \lambda_2 > 0$ if and only if $\min T_i = \min T_i'$ and $\sum T_i = \sum T_i'$ $(\sum T_i, \min T_i)$ is minimal sufficient for λ.

Exercise 3: Solution III

3. To find the MLE for λ, we find $\hat{\lambda}$ such at $\ell_{\mathcal{T}}(\lambda)$ is maximized.

$$\ell_{\mathcal{T}}(\lambda) = \lambda^n e^{-\lambda \sum T_i + n} 1 \left\{ \min T_i \geq 1/\lambda \right\}$$

Without the indicator function, the expression is maximized by solving

$$\begin{aligned} \frac{\partial}{\partial \lambda} \log \left(\lambda^n e^{-\lambda \sum T_i + n} \right) &= 0 \\ \frac{\partial}{\partial \lambda} \left(n \log \lambda - \lambda \sum_i + n \right) &= 0 \\ \frac{n}{\lambda} - \sum T_i &= 0 \\ \lambda &= \frac{n}{\sum T_i} \end{aligned}$$

However, $n / \sum T_i \leq 1 / \min T_i$ always holds. The expression is

Exercise 3: Solution IV

decreasing for $\lambda > \frac{n}{\sum T_i}$. Therefore, the maximum likelihood estimator of λ is $\frac{1}{\min T_i}$

Outline

1. Chapter 4: Sufficiency and Likelihood

2. Exercise 1

3. Exercise 2

4. Exercise 3

5. Exercise 4

Exercise 4 I

Suppose we observe $(X_{1,\ldots,}X_n, Y_1, \ldots, Y_m)$ where $X_1, \ldots, X_n, Y_1 \ldots, Y_m$ are independent random variables, with $X_i \sim N\left(\mu, \sigma_1^2\right)$ and $Y_j \sim N\left(\mu, \sigma_2^2\right)$ for $0 \leq i \leq n$ and $0 \leq j \leq m$.
Assuming $\mu \in \mathbb{R}$ and $\sigma_1^2, \sigma_2^2 \in \mathbb{R}^+$,

1. Show that
$$T = \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2, \sum_{j=1}^{m} Y_i, \sum_{j=1}^{m} Y_i^2\right)$$
is a sufficient statistic.
2. Show that T is not a complete sufficient statistic.

Exercise 4: Solution I

The joint density of $\mathcal{X} = (X_1, \ldots, X_n)$ and $\mathcal{Y} = (Y_1, \ldots, Y_m)$ is

$$f_{\mathcal{X},\mathcal{Y}}(\boldsymbol{x}, \boldsymbol{y}) = \left(\frac{1}{\sqrt{2\pi\sigma_1^2}}\right)^n \left(\frac{1}{\sqrt{2\pi\sigma_2^2}}\right)^m \times$$

$$\exp\left\{-\frac{1}{2}\left(\sum_{i=1}^n \frac{(x_i-\mu)^2}{\sigma_1^2} + \sum_{j=1}^m \frac{(y_j-\mu)^2}{\sigma_2^2}\right)\right\}$$

$$\propto \exp\left\{\frac{-1}{2\sigma_1^2}\sum_{i=1}^n x_i^2 + \frac{\mu}{\sigma_1^2}\sum_{i=1}^n x_i + \frac{-1}{2\sigma_2^2}\sum_{j=1}^m y_j^2 + \frac{\mu}{\sigma_2^2}\sum_{j=1}^m y_j\right\}$$

It can be observed that $(\mathcal{X}, \mathcal{Y})$ belongs to the exponential family with natural statistic $T = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2, \sum_{j=1}^m Y_i, \sum_{j=1}^m Y_i^2\right)$ and natural parameter $\pi = \left(\frac{\mu}{\sigma_1^2}, \frac{-1}{2\sigma_1^2}, \frac{\mu}{\sigma_2^2}, \frac{-1}{2\sigma_2^2}\right)$.

Exercise 4: Solution II

We conclude that the natural statistic T is sufficient. As the natural parameter space does not contain an open rectangle, we find a function g such that $E(g(T)) = 0$ but $P(g(T) = 0) < 1$

Define a function g as $g : \mathbb{R}^4 \to \mathbb{R}$ with $(a, b, c, d) \mapsto \frac{a}{n} - \frac{c}{m}$

$$\begin{aligned} E(g(T)) &= E(\bar{X}) - E(\bar{Y}) \\ &= \mu - \mu = 0 \end{aligned}$$

However, $P(\bar{X} - \bar{Y} = 0) = 0$. Therefore, T is not a complete sufficient statistic.