Math 541: Statistical Theory II

Sufficient Statistics and Exponential Family

Lecturer: Songfeng Zheng

1 Statistics and Sufficient Statistics

Suppose we have a random sample $X_1, \cdots, X_n$ taken from a distribution $f(x|\theta)$ which relies on an unknown parameter θ in a parameter space Θ. The purpose of parameter estimation is to estimate the parameter θ from the random sample.

We have already studied three parameter estimation methods: method of moment, maximum likelihood, and Bayes estimation. We can see from the previous examples that the estimators can be expressed as a function of the random sample $X_1, \cdots, X_n$. Such a function is called a *statistic*.

Formally, any real-valued function $T = r(X_1, \cdots, X_n)$ of the observations in the sample is called a *statistic*. In this function, there should not be any unknown parameter. For example, suppose we have a random sample $X_1, \cdots, X_n$, then $\overline{X}$, $\max(X_1, \cdots, X_n)$, $\text{median}(X_1, \cdots, X_n)$, and $r(X_1, \cdots, X_n) = 4$ are statistics; however $X_1 + \mu$ is not statistic if μ is unknown.

For the parameter estimation problem, we know nothing about the parameter but the observations from such a distribution. Therefore, the observations $X_1, \cdots, X_n$ is our first hand of information source about the parameter, that is to say, all the available information about the parameter is contained in the observations. However, we know that the estimators we obtained are always functions of the observations, i.e., the estimators are statistics, e.g. sample mean, sample standard deviations, etc. In some sense, this process can be thought of as "compress" the original observation data: initially we have n numbers, but after this "compression", we only have 1 numbers. This "compression" always makes us lose information about the parameter, can never makes us obtain more information. The best case is that this "compression" result contains the same amount of information as the information contained in the n observations. We call such a statistic as *sufficient statistic*. From the above intuitive analysis, we can see that sufficient statistic "absorbs" all the available information about θ contained in the sample. This concept was introduced by R. A. Fisher in 1922.

If $T(X_1, \cdots, X_n)$ is a statistic and t is a particular value of T, then the conditional joint distribution of $X_1, \cdots, X_n$ given that $T = t$ can be calculated. In general, this joint conditional distribution will depend on the value of θ. Therefore, for each value of t, there

will be a family of possible conditional distributions corresponding to the different possible values of $\theta \in \Theta$. However, it may happen that for each possible value of t, the conditional joint distribution of $X_1, \cdots, X_n$ given that $T = t$ is the same for all the values of $\theta \in \Theta$ and therefore does not actually depend on the value of θ. In this case, we say that T is a sufficient statistic for the parameter θ.

Formally, a statistic $T(X_1, \cdots, X_n)$ is said to be **sufficient** for θ if the conditional distribution of $X_1, \cdots, X_n$, given $T = t$, does not depend on θ for any value of t.

In other words, given the value of T, we can gain no more knowledge about θ from knowing more about the probability distribution of $X_1, \cdots, X_n$. We could envision keeping only T and throwing away all the X_i without losing any information!

The concept of sufficiency arises as an attempt to answer the following question: Is there a statistic, i.e. a function $T(X_1, \cdots, X_n)$, 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. For example, consider a sequence of independent Bernoulli trials with unknown probability of success, θ. We may have the intuitive feeling that the total number of successes contains all the information about θ that is in the sample, that the order in which the successes occurred, for example, does not give any additional information about θ.

Example 1: Let $X_1, \cdots, X_n$ be a sequence of independent bernoulli trials with $P(X_i = 1) = \theta$. We will verify that $T = \sum_{i=1}^{n} X_i$ is sufficient for θ.

Proof: We have

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

Bearing in mind that the X_i can take on only the values 0s or 1s, the probability in the numerator is the probability that some particular set of t X_i are equal to 1s and the other $n - t$ are 0s. Since the X_i are independent, the probability of this is $\theta^t(1-\theta)^{n-t}$. To find the denominator, note that the distribution of T, the total number of ones, is binomial with n trials and probability of success θ. Therefore the ratio in the above equation is

$$\frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \frac{1}{\binom{n}{t}}$$

The conditional distribution thus does not involve θ at all. Given the total number of ones, the probability that they occur on any particular set of t trials is the same for any value of θ so that set of trials contains no additional information about θ.

2 Factorization Theorem

The preceding definition of sufficiency is hard to work with, because it does not indicate how to go about finding a sufficient statistic, and given a candidate statistic, T, it would typically be very hard to conclude whether it was sufficient statistic because of the difficulty in evaluating the conditional distribution.

We shall now present a simple method for finding a sufficient statistic which can be applied in many problems. This method is based on the following result, which was developed with increasing generality by R. A. Fisher in 1922, J. Neyman in 1935, and P. R. Halmos and L. J. Savage in 1949, and this result is know as the *Factorization Theorem.*

Factorization Theorem: *Let $X_1, \cdots, X_n$ form a random sample from either a continuous distribution or a discrete distribution for which the pdf or the point mass function is $f(x|\theta)$, where the value of θ is unknown and belongs to a given parameter space Θ. A statistic $T(X_1, \cdots, X_n)$ is a sufficient statistic for θ if and only if the joint pdf or the joint point mass function $f_n(\mathbf{x}|\theta)$ of $X_1, \cdots, X_n$ can be factorized as follows for all values of $\mathbf{x} = (x_1, \cdots, x_n) \in R^n$ and all values of $\theta \in \Theta$:*

$$f_n(\mathbf{x}|\theta) = u(\mathbf{x})v[T(\mathbf{x}), \theta].$$

Here, the function u and v are nonnegative, the function u may depend on $\mathbf{x}$ but does not depend on θ, and the function v depends on θ but will depend on the observed value $\mathbf{x}$ only through the value of the statistic $T(\mathbf{x})$.

Note: In this expression, we can see that the statistic $T(X_1, \cdots, X_n)$ is like an "interface" between the random sample $X_1, \cdots, X_n$ and the function v.

Proof: We give a proof for the discrete case. First, suppose the frequency function can be factored in the given form. We let $\mathbf{X} = (X_1, \cdots, X_n)$ and $\mathbf{x} = (x_1, \cdots, x_n)$, then

$$P(T = t) = \sum_{\mathbf{x}:T(\mathbf{x})=t} P(\mathbf{X} = \mathbf{x}) = \sum_{\mathbf{x}:T(\mathbf{x})=t} u(\mathbf{x})v(T(\mathbf{x}), \theta) = v(t, \theta) \sum_{\mathbf{x}:T(\mathbf{x})=t} u(\mathbf{x})$$

Then we can have

$$P(\mathbf{X} = \mathbf{x}|T = t) = \frac{P(\mathbf{X} = \mathbf{x}, T = t)}{P(T = t)} = \frac{u(\mathbf{x})}{\sum_{\mathbf{x}:T(\mathbf{x})=t} u(\mathbf{x})}$$

which does not depend on θ, and therefore T is a sufficient statistic.

Conversely, suppose the conditional distribution of $\mathbf{X}$ given T is independent of θ, that is T is a sufficient statistic. Then we can let $u(\mathbf{x}) = P(\mathbf{X} = \mathbf{x}|T = t, \theta)$, and let $v(t, \theta) = P(T = t|\theta)$. It follows that

$$\begin{aligned} P(\mathbf{X} = \mathbf{x}|\theta) &= P(\mathbf{X} = \mathbf{x}, T = t|\theta) \quad \text{this is true because } T(\mathbf{x}) = t, \text{it is redundant} \\ &= P(\mathbf{X} = \mathbf{x}|T = t, \theta)P(T = t|\theta) = u(\mathbf{x})v(T(x_1, \cdots, x_n)), \theta) \end{aligned}$$

which is of the desired form.

Example 2: Suppose that $X_1, \cdots, X_n$ form a random sample from a Poisson distribution for which the value of the mean θ is unknown ($\theta > 0$). Show that $T = \sum_{i=1}^{n} X_i$ is a sufficient statistic for θ.

Proof: For every set of nonnegative integers $x_1, \cdots, x_n$, the joint probability mass function $f_n(\mathbf{x}|\theta)$ of $X_1, \cdots, X_n$ is as follows:

$$f_n(\mathbf{x}|\theta) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \left(\prod_{i=1}^{n} \frac{1}{x_i!}\right) e^{-n\theta}\theta^{\sum_{i=1}^{n} x_i}$$

It can be seen that $f_n(\mathbf{x}|\theta)$ has been expressed as the product of a function that does not depend on θ and a function that depends on θ but depends on the observed vector $\mathbf{x}$ only through the value of $\sum_{i=1}^{n} x_i$. By factorization theorem, it follows that $T = \sum_{i=1}^{n} X_i$ is a sufficient statistic for θ.

Example 3: Applying the Factorization Theorem to a Continuous Distribution. Suppose that $X_1, \cdots, X_n$ form a random sample from a continuous distribution with the following p.d.f.:

$$f(x|\theta) = \begin{cases} \theta x^{\theta-1}, & \text{for} \quad 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

It is assumed that the value of the parameter θ is unknown ($\theta > 0$). We shall show that $T = \prod_{i=1}^{n} X_i$ is a sufficient statistic for θ.

Proof: For $0 < x_i < 1$ ($i = 1, \cdots, n$), the joint p.d.f. $f_n(\mathbf{x}|\theta)$ of $X_1, \cdots, X_n$ is as follows:

$$f_n(\mathbf{x}|\theta) = \theta^n \left(\prod_{i=1}^{n} x_i\right)^{\theta-1}.$$

Furthermore, if at least one value of x_i is outside the interval $0 < x_i < 1$, then $f_n(\mathbf{x}|\theta) = 0$ for every value of $\theta \in \Theta$. The right side of the above equation depends on $\mathbf{x}$ only through the value of the product $\prod_{i=1}^{n} x_i$. Therefore, if we let $u(\mathbf{x}) = 1$ and $r(\mathbf{x}) = \prod_{i=1}^{n} x_i$, then $f_n(\mathbf{x}|\theta)$ can be considered to be factored in the form specified by the factorization theorem. It follows from the factorization theorem that the statistic $T = \prod_{i=1}^{n} X_i$ is a sufficient statistic for θ.

Example 4: Suppose that $X_1, \cdots, X_n$ form a random sample from a normal distribution for which the mean μ is unknown but the variance σ^2 is known. Find a sufficient statistic for μ.

Solution: For $\mathbf{x} = (x_1, \cdots, x_n)$, the joint pdf of $X_1, \cdots, X_n$ is

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

This could be rewritten as

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

It can be seen that $f_n(\mathbf{x}|\mu)$ has now been expressed as the product of a function that does not depend on μ and a function the depends on $\mathbf{x}$ only through the value of $\sum_{i=1}^n x_i$. It follows from the factorization theorem that $T = \sum_{i=1}^n X_i$ is a sufficient statistic for μ.

Since $\sum_{i=1}^n = n\bar{x}$, we can state equivalently that the final expression depends on $\mathbf{x}$ only through the value of $\bar{x}$, therefore $\bar{X}$ is also a sufficient statistic for μ. More generally, every one to one function of $\bar{X}$ will be a sufficient statistic for μ.

Property of sufficient statistic: Suppose that $X_1, \cdots, X_n$ form a random sample from a distribution for which the p.d.f. is $f(x|\theta)$, where the value of the parameter θ belongs to a given parameter space Θ. Suppose that $T(X_1, \cdots, X_n)$ and $T'(X_1, \cdots, X_n)$ are two statistics, and there is a one-to-one map between T and T'; that is, the value of T' can be determined from the value of T without knowing the values of $X_1, \cdots, X_n$, and the value of T can be determined from the value of T' without knowing the values of $X_1, \cdots, X_n$. Then T' is a sufficient statistic for θ if and only if T is a sufficient statistic for θ.

Proof: Suppose the one-to-one mapping between T and T' is g, i.e. $T' = g(T)$ and $T = g^{-1}(T')$, and g^{-1} is also one-to-one. T is a sufficient statistic if and only if the joint pdf $f_n(\mathbf{X}|T = t)$ can be factorized as

$$f_n(\mathbf{x}|T = t) = u(\mathbf{x})v[T(\mathbf{x}), \theta]$$

and this can be written as

$$u(\mathbf{x})v[T(\mathbf{x}), \theta] = u(\mathbf{x})v[g^{-1}(T'(\mathbf{x})), \theta] = u(\mathbf{x})v'[T'(\mathbf{x}), \theta]$$

Therefore the joint pdf can be factorized as $u(\mathbf{x})v'[T'(\mathbf{x}), \theta]$, by factorization theorem, T' is a sufficient statistic.

For instance, in Example 4, we showed that for normal distribution, both $T_1 = \sum_{i=1}^n X_i$ and $T_2 = \bar{X}$ are sufficient statistics, and there is a one-to-one mapping between T_1 and T_2: $T_2 = T_1/n$. Other statistics like $(\sum_{i=1}^n X_i)^3$, $\exp(\sum_{i=1}^n X_i)$ are also sufficient statistics.

Example 5: Suppose that $X_1, \cdots, X_n$ form a random sample from a beta distribution with parameters α and β, where the value of α is known and the value of β is unknown ($\beta > 0$). Show that the following statistic T is a sufficient statistic for the parameter β:

$$T = \frac{1}{n}\left(\sum_{i=1}^n \log\frac{1}{1 - X_i}\right)^3.$$

Proof: The p.d.f. $f(x|\beta)$ of each individual observation X_i is

$$f(x|\beta) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & \text{for} \quad 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

Therefore, the joint p.d.f. $f_n(\mathbf{x}|\beta)$ of $X_1, \cdots, X_n$ is

$$f_n(\mathbf{x}|\beta) = \prod_{i=1}^n \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x_i^{\alpha-1}(1-x_i)^{\beta-1} = \Gamma(\alpha)^{-n}\left(\prod_{i=1}^n x_i\right)^{\alpha-1}\left[\frac{\Gamma(\alpha+\beta)^n}{\Gamma(\beta)^n}\left(\prod_{i=1}^n (1-x_i)\right)^{\beta-1}\right]$$

We define $T'(X_1, \cdots, X_n) = \prod_{i=1}^{n}(1 - X_i)$, and because α is known, so we can define

$$u(\mathbf{x}) = \Gamma(\alpha)^{-n}\left(\prod_{i=1}^{n} x_i\right)^{\alpha-1} \qquad v(T', \beta) = \frac{\Gamma(\alpha+\beta)^n}{\Gamma(\beta)^n} T'(x_1, \cdots, x_n)^{\beta-1}$$

We can see that the function v depends on $\mathbf{x}$ only through T', therefore T' is a sufficient statistic.

It is easy to see that

$$T = g(T') = \frac{\log(-T')^3}{n},$$

and the function g is a one-to-one mapping. Therefore T is a sufficient statistic.

Example 6: Sampling for a Uniform distribution. Suppose that $X_1, \cdots, X_n$ form a random sample from a uniform distribution on the interval $[0, \theta]$, where the value of the parameter θ is unknown ($\theta > 0$). We shall show that $T = \max(X_1, \cdots, X_n)$ is a sufficient statistic for θ.

Proof: The p.d.f. $f(x|\theta)$ of each individual observation X_i is

$$f(x|\theta) = \begin{cases} \frac{1}{\theta}, & \text{for} \quad 0 \le x \le \theta \\ 0, & \text{otherwise} \end{cases}$$

Therefore, the joint p.d.f. $f_n(\mathbf{x}|\theta)$ of $X_1, \cdots, X_n$ is

$$f_n(\mathbf{x}|\theta) = \begin{cases} \frac{1}{\theta^n}, & \text{for} \quad 0 \le x_i \le \theta \\ 0, & \text{otherwise} \end{cases}$$

It can be seen that if $x_i < 0$ for at least one value of i ($i = 1, \cdots, n$), then $f_n(\mathbf{x}|\theta) = 0$ for every value of $\theta > 0$. Therefore it is only necessary to consider the factorization of $f_n(\mathbf{x}|\theta)$ for values of $x_i \ge 0$ ($i = 1, \cdots, n$).

Define $h[\max(x_1, \cdots, x_n), \theta]$ as

$$h[\max(x_1, \cdots, x_n), \theta] = \begin{cases} 1, & \text{if} \quad \max(x_1, \cdots, x_n) \le \theta \\ 0, & \text{if} \quad \max(x_1, \cdots, x_n) > \theta \end{cases}$$

Also, $x_i \le \theta$ for $i = 1, \cdots, n$ if and only if $\max(x_1, \cdots, x_n) \le \theta$. Therefore, for $x_i \ge 0$ ($i = 1, \cdots, n$), we can rewrite $f_n(\mathbf{x}|\theta)$ as follows:

$$f_n(\mathbf{x}|\theta) = \frac{1}{\theta^n} h[\max(x_1, \cdots, x_n), \theta].$$

Since the right side depends on $\mathbf{x}$ only through the value of $\max(x_1, \cdots, x_n)$, it follows that $T = \max(X_1, \cdots, X_n)$ is a sufficient statistic for θ. According to the property of sufficient statistic, any one-to-one function of T is a sufficient statistic as well.

Example 7: Suppose that $X_1, X_2, \cdots, X_n$ are i.i.d. random variables on the interval $[0, 1]$ with the density function

$$f(x|\alpha) = \frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2}[x(1-x)]^{\alpha-1}$$

where $\alpha > 0$ is a parameter to be estimated from the sample. Find a sufficient statistic for α.

Solution: The joint density function of $x_1, \cdots, x_n$ is

$$f(x_1, \cdots, x_n|\alpha) = \prod_{i=1}^{n} \frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2}[x_i(1-x_i)]^{\alpha-1} = \frac{\Gamma(2\alpha)^n}{\Gamma(\alpha)^{2n}}\left[\prod_{i=1}^{n} x_i(1-x_i)\right]^{\alpha-1}$$

Comparing with the form in factorization theorem,

$$f(x_1, \cdots, x_n|\theta) = u(x_1, \cdots, x_n)v[T(x_1, \cdots, x_n), \theta]$$

we see that $T = \prod_{i=1}^{n} X_i(1-X_i)$, $v(t, \theta) = \frac{\Gamma(2\alpha)^n}{\Gamma(\alpha)^{2n}}t^{\alpha-1}$, $u(x_1, \cdots, x_n) = 1$, i.e. v depends on $x_1, \cdots, x_n$ only through t. By the factorization theorem, $T = \prod_{i=1}^{n} X_i(1-X_i)$ is a sufficient statistic. According to the property of sufficient statistic, any one-to-one function of T is a sufficient statistic as well.

3 Sufficient Statistics and Estimators

We know estimators are statistics, in particular, we want the obtained estimator to be sufficient statistic, since we want the estimator absorbs all the available information contained in the sample.

Suppose that $X_1, \cdots, X_n$ form a random sample from a distribution for which the pdf or point mass function is $f(x|\theta)$, where the value of the parameter θ is unknown. And we assume there is a sufficient statistic for θ, which is $T(X_1, \cdots, X_n)$. We will show that the MLE of θ, $\hat{\theta}$, depends on $X_1, \cdots, X_n$ only through the statistic T.

It follows from the factorization theorem that the likelihood function $f_n(\mathbf{x}|\theta)$ can be written as

$$f_n(\mathbf{x}|\theta) = u(\mathbf{x})v[T(\mathbf{x}), \theta].$$

We know that the MLE $\hat{\theta}$ is the value of θ for which $f_n(\mathbf{x}|\theta)$ is maximized. We also know that both u and v are positive. Therefore, it follows that $\hat{\theta}$ will be the value of θ for which $v[T(\mathbf{x}), \theta]$ is maximized. Since $v[T(\mathbf{x}), \theta]$ depends on $\mathbf{x}$ only through the function $T(\mathbf{x})$, it follows that $\hat{\theta}$ will depend on $\mathbf{x}$ only through the function $T(\mathbf{x})$. Thus the MLE estimator $\hat{\theta}$ is a function of the sufficient statistic $T(X_1, \cdots, X_n)$.

In many problems, the MLE $\hat{\theta}$ is actually a sufficient statistic. For instance, Example 1 shows a sufficient statistic for the success probability in Bernoulli trial is $\sum_{i=1}^{n} X_i$, and we

know the MLE for θ is $\bar{X}$; Example 4 shows a sufficient statistic for μ in normal distribution is $\bar{X}$, and this is the MLE for μ; Example 6 shows a sufficient statistic of θ for the uniform distribution on $(0, \theta)$ is $\max(X_1, \cdots, X_n)$, and this is the MLE for θ.

The above discussion for MLE also holds for Bayes estimator. Let θ be a parameter with parameter space Θ equal to an interval of real numbers (possibly unbounded), and we assume that the prior distribution for θ is $p(\theta)$. Let X have p.d.f. $f(x|\theta)$ conditional on θ. Suppose we have a random sample $X_1, \cdots, X_n$ from $f(x|\theta)$. Let $T(X_1, \cdots, X_n)$ be a sufficient statistic. We first show that the posterior p.d.f. of θ given $\mathbf{X} = \mathbf{x}$ depends on $\mathbf{x}$ only through $T(\mathbf{x})$.

The likelihood term is $f_n(\mathbf{x}|\theta)$, according to Bayes formula, we have the posterior distribution for θ is

$$f(\theta|\mathbf{x}) = \frac{f_n(\mathbf{x}|\theta)p(\theta)}{\int_\Theta f_n(\mathbf{x}|\theta)p(\theta)d\theta} = \frac{u(\mathbf{x})v(T(\mathbf{x}), \theta)p(\theta)}{\int_\Theta u(\mathbf{x})v(T(\mathbf{x}), \theta)p(\theta)d\theta} = \frac{v(T(\mathbf{x}), \theta)p(\theta)}{\int_\Theta v(T(\mathbf{x}), \theta)p(\theta)d\theta}$$

where the second step uses the factorization theorem. We can see that the posterior p.d.f. of θ given $\mathbf{X} = \mathbf{x}$ depends on $\mathbf{x}$ only through $T(\mathbf{x})$.

Since the Bayes estimator of θ with respect to a specified loss function is calculated from this posterior p.d.f., the estimator also will depend on the observed vector $\mathbf{x}$ only through the value of $T(\mathbf{x})$. In other words, the Bayes estimator is a function of the sufficient statistic $T(X_1, \cdots, X_n)$.

Summarizing our discussion above, both the MLE estimator and Bayes estimator are functions of sufficient statistic, therefore they absorb all the available information contained in the sample at hand.

4 Exponential Family of Probability Distribution

A study of the properties of probability distributions that have sufficient statistics of the same dimension as the parameter space regardless of the sample size led to the development of what is called the *exponential family of probability distributions*. Many common distributions, including the normal, the binomial, the Poisson, and the gamma, are members of this family.

One-parameter members of the exponential family have density or mass function of the form

$$f(x|\theta) = \exp[c(\theta)T(x) + d(\theta) + S(x)]$$

Suppose that $X_1, \cdots, X_n$ are i.i.d. samples from a member of the exponential family, then the joint probability function is

$$\begin{aligned} f(\mathbf{x}|\theta) &= \prod_{i=1}^{n} \exp[c(\theta)T(x_i) + d(\theta) + S(x_i)] \\ &= \exp\left[c(\theta)\sum_{i=1}^{n} T(x_i) + nd(\theta)\right] \exp\left[\sum_{i=1}^{n} S(x_i)\right] \end{aligned}$$

From this result, it is apparent by the factorization theorem that $\sum_{i=1}^{n} T(x_i)$ is a sufficient statistic.

Example 8: The frequency function of Bernoulli distribution is

$$\begin{aligned} P(X=x) &= \theta^x(1-\theta)^{1-x} \quad x=0 \; or \; x=1 \\ &= \exp\left[x\log\left(\frac{\theta}{1-\theta}\right)+\log(1-\theta)\right] \end{aligned} \tag{1}$$

It can be seen that this is a member of the exponential family with $T(x)=x$, and we can also see that $\sum_{i=1}^{n} X_i$ is a sufficient statistic, which is the same as in example 1.

Example 9: Suppose that $X_1, X_2, \cdots, X_n$ are i.i.d. random variables on the interval $[0,1]$ with the density function

$$f(x|\alpha) = \frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2}[x(1-x)]^{\alpha-1}$$

where $\alpha > 0$ is a parameter to be estimated from the sample. Find a sufficient statistic for α by verifying that this distribution belongs to exponential family.

Solution: The density function

$$f(x|\alpha) = \frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2}[x(1-x)]^{\alpha-1} = \exp\left\{\log\Gamma(2\alpha) - 2\log\Gamma(\alpha) + \alpha\log[x(1-x)] - \log[x(1-x)]\right\}$$

Comparing to the form of exponential family,

$$T(x) = \log[x(1-x)]; \quad c(\alpha) = \alpha; \quad S(x) = -\log[x(1-x)]; \quad d(\alpha) = \log\Gamma(2\alpha) - 2\log\Gamma(\alpha)$$

Therefore, $f(x|\alpha)$ belongs to exponential family. Then the sufficient statistic is

$$\sum_{i=1}^{n} T(X_i) = \sum_{i=1}^{n} \log[X_i(1-X_i)] = \log\left[\prod_{i=1}^{n} X_i(1-X_i)\right]$$

In example 6, we got the sufficient statistic was $\prod_{i=1}^{n} X_i(1-X_i)$, which is different from the result here. But both of them are sufficient statistics because of the functional relationship between them.

A *k-parameter* member of the exponential family has a density or frequency function of the form

$$f(x|\theta) = \exp\left[\sum_{i=1}^{k} c_i(\theta)T_i(x) + d(\theta) + S(x)\right]$$

For example, the normal distribution, gamma distribution (Example below), beta distribution are of this form.

Example 10: Show that the gamma distribution belongs to the exponential family.

Proof: Gamma distribution has density function

$$f(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad 0 \le x < \infty$$

which can be written as

$$f(x|\alpha,\beta) = \exp\{-\beta x + (\alpha-1)\log x + \alpha \log\beta - \log\Gamma(\alpha)\}$$

Comparing with the form of exponential family

$$\exp\left\{\sum_{i=1}^{n} c_i(\theta)T_i(x) + d(\theta) + S(x)\right\}$$

We see that Gamma distribution has the form of exponential distribution with $c_1(\alpha,\beta) = -\beta$, $c_2(\alpha,\beta) = \alpha - 1$, $T_1(x) = x$, $T_2(x) = \log x$, $d(\alpha,\beta) = \alpha\log\beta - \log\Gamma(\alpha)$, and $S(x) = 0$.

Therefore, gamma distribution belongs to the exponential family.

5 Exercises

Instructions for Exercises 1 to 4: In each of these exercises, assume that the random variables $X_1, \cdots, X_n$ form a random sample of size n form the distribution specified in that exercise, and show that the statistic T specified in the exercise is a sufficient statistic for the parameter:

Exercise 1: A normal distribution for which the mean μ is known and the variance σ^2 is unknown; $T = \sum_{i=1}^n (X_i - \mu)^2$.

Exercise 2: A gamma distribution with parameters α and β, where the value of β is known and the value of α is unknown ($\alpha > 0$); $T = \prod_{i=1}^n X_i$.

Exercise 3: A uniform distribution on the interval $[a, b]$, where the value of a is known and the value of b is unknown ($b > a$); $T = \max(X_1, \cdots, X_n)$.

Exercise 4: A uniform distribution on the interval $[a, b]$, where the value of b is known and the value of a is unknown ($b > a$); $T = \min(X_1, \cdots, X_n)$.

Exercise 5: Suppose that $X_1, \cdots, X_n$ form a random sample from a gamma distribution with parameters $\alpha > 0$ and $\beta > 0$, and the value of β is known. Show that the statistic $T = \sum_{i=1}^n \log X_i$ is a sufficient statistic for the parameter α.

Exercise 6: The Pareto distribution has density function:

$$f(x|x_0,\theta) = \theta x_0^\theta x^{-\theta-1}, \quad x \ge x_0, \quad \theta > 1$$

Assume that $x_0 > 0$ is given and that $X_1, X_2, \cdots, X_n$ is an i.i.d. sample. Find a sufficient statistic for θ by (a) using factorization theorem, (b) using the property of exponential family. Are they the same? If not, why are both of them sufficient?

Exercise 7: Verify that following are members of exponential family:

a) Geometric distribution $p(x) = p^{x-1}(1-p)$;

b) Poisson distribution $p(x) = e^{-\lambda}\frac{\lambda^x}{x!}$;

c) Normal distribution $N(\mu, \sigma^2)$;

d) Beta distribution.