

Chapter 1

Statistical Spaces

Mathematical statistics is a science that studies the statistical regularity of random phenomena, essentially by some observation values of random variable (r.v.) X . Sometimes the observation values are also called sample points. The lower case x is used to denote a realization of the r.v. X . Furthermore, the calligraphic letter \mathcal{X} is used to denote the set from which X takes value, and \mathcal{A} is used to denote the σ -algebra generated by some subsets of \mathcal{X} . The measurable space $(\mathcal{X}, \mathcal{A})$ is used as a starting point for our further analysis, and will be called the sample space. Considering the need of the research, we define a measure in the measurable space $(\mathcal{X}, \mathcal{A})$ by three methods: $(\mathcal{X}, \mathcal{A}, \nu)$, where ν is a σ -finite measure, is called a **measure space**; $(\mathcal{X}, \mathcal{A}, P)$, where P is a probability measure, is called a **probability space**; and $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, where \mathcal{P} is a family of probability measures or a family of probability distributions, for the sake of convenience, is called a **statistical space**. In many situations, the family of probability distributions depends only on the parameters in probability distribution:

$$\mathcal{P} = \{P_\theta; \theta \in \Theta\},$$

where Θ is a parameter space. For example, the family of normal distributions with known variance depends only on the unique mean parameter, the parameter θ hereof is a real number; but when the variance is also unknown, the distribution depends on both the mean and the variance, the parameter θ hereof is a bivariate vector. The primary task of a statistical test is to infer whether the parameter from the distribution is equal to some given constant θ_0 , *i.e.* to test the hypothesis $H : \theta = \theta_0$; or to infer whether it belongs to some given subset Θ_0 of Θ , *i.e.*, to test $H : \theta \in \Theta_0$. Before discussing this problem, this chapter will review some basic concepts which will be involved in subsequent chapters. Firstly, analyze some properties of the statistical space; secondly, discuss conditional probability, sufficient

statistics and some characteristics of the exponential distribution family; at last, review some basic estimation methods. The readers who are familiar with these contents can skip the chapter.

1.1 Basic Properties of Statistical Spaces

The above-mentioned three types of metric space are consistent in form, but their connotations are quite different in nature: for $A \in \mathcal{A}$, the measure in the measure space is denoted as $\nu(A)$, which represents the size of the set A ; the measure in the probability space is denoted as $P(X \in A)$, which represents the possibility of r.v. X taking values in A ; the measure in the statistical space is denoted as $P_\theta(X \in A)$, is used to find a parameter θ which is the most “appropriate” to measure the possibility. About the “appropriateness”, we can give different criteria according to different problems. First of all, We will study some general properties about statistical spaces.

1.1.1 Measure in Statistical Spaces

Let \mathcal{P} and ν be the family of probability distributions and σ -finite measure in the sample space $(\mathcal{X}, \mathcal{A})$, respectively. If

$$A \in \mathcal{A}, \nu(A) = 0 \quad \Rightarrow \quad P(A) = 0, \forall P \in \mathcal{P}, \quad (1.1.1)$$

then \mathcal{P} is said to be absolutely continuous with respect to ν , denoted by $\mathcal{P} \ll \nu$. If the family of probability distributions depends only on the parameter in the probability distributions, then (1.1.1) can be rewritten as

$$A \in \mathcal{A}, \nu(A) = 0 \quad \Rightarrow \quad P_\theta(A) = 0, \forall \theta \in \Theta. \quad (1.1.2)$$

In the later discussion, we will focus on the definition given by (1.1.2). According to the Radon-Nikodym Theorem (cf. Halmos, 1957), if $\mathcal{P} \ll \nu$, then for $\forall \theta \in \Theta$, there exists an \mathcal{A} -measurable function f_θ , for $\forall A \in \mathcal{A}$ we have

$$P_\theta(A) = \int_A f_\theta(x) d\nu(x). \quad (1.1.3)$$

Furthermore, f_θ exists uniquely almost everywhere (a.e.) with respect to the measure ν . Then f_θ is called a density function of P_θ and denoted by

$$\frac{dP_\theta}{d\nu} = f_\theta \quad \text{or} \quad dP_\theta = f_\theta d\nu.$$

There are two equivalent formulations about uniqueness a.e. If there also exists f_θ^* s.t. $dP_\theta = f_\theta^* d\nu$, then

$$\nu(x; f_\theta(x) \neq f_\theta^*(x)) = 0;$$

or, for $\forall A \in \mathcal{A}$ we have

$$\int_A f_\theta(x) d\nu(x) = \int_A f_\theta^*(x) d\nu(x).$$

Therefore the uniqueness a.e. is not related to only the measure ν but also the σ -algebra \mathcal{A} .

Theorem 1.1.1. *Let \mathcal{P} and ν be the family of probability distributions and σ -finite measure in the sample space $(\mathcal{X}, \mathcal{A})$, respectively. Let g be an \mathcal{A} -measurable function. If $\mathcal{P} \ll \nu$, then for $\forall \theta \in \Theta$ we have*

$$\int_{\mathcal{X}} g(x) dP_\theta(x) = \int_{\mathcal{X}} g(x) \frac{dP_\theta}{d\nu} d\nu(x) = \int_{\mathcal{X}} g(x) f_\theta(x) d\nu(x). \quad (1.1.4)$$

Proof. From (1.1.3), (1.1.4) holds obviously when $g(x) = I_A(x)$, for $A \in \mathcal{A}$. When g is a nonnegative simple function, i.e., for $i = 1, \dots, n$, there exist $A_i \in \mathcal{A}$, satisfying $A_i \cap A_j = \emptyset$ for $i \neq j$, and $a_i > 0$, such that

$$g(x) = a_1 I_{A_1}(x) + \dots + a_n I_{A_n}(x),$$

then Eq. (1.1.4) holds, i.e.,

$$\sum_{i=1}^n a_i P_\theta(A_i) = \sum_{i=1}^n a_i \int_{\mathcal{X}} I_{A_i}(x) f_\theta(x) d\nu(x).$$

If g is a nonnegative measurable function, then there exists a series of nonnegative simple functions g_n satisfying

$$0 \leq g_1(x) \leq g_2(x) \leq \dots, \quad \text{and} \quad g_n(x) \rightarrow g(x).$$

By the monotone convergence theorem, we have

$$\begin{aligned} \int_{\mathcal{X}} g(x) dP_\theta(x) &= \lim_{n \rightarrow \infty} \int_{\mathcal{X}} g_n(x) dP_\theta(x) \\ &= \lim_{n \rightarrow \infty} \int_{\mathcal{X}} g_n(x) f_\theta(x) d\nu(x) \\ &= \int_{\mathcal{X}} g(x) f_\theta(x) d\nu(x). \end{aligned}$$

When g is a general measurable function, we can decompose g into a positive part and a negative part. Thus the theorem is proved. \square

The technique used in the proof of Theorem 1.1.1 is generally called **I-method**. The method indicates that a proposition still holds for a measurable function if it holds for a measurable set in statistical space. Generally Eq. (1.1.4) is called the **mean** of $g(x)$ when the parameter is θ , and denoted by

$$E_{\theta}g(X) = \int_{\mathcal{X}} g(x)dP_{\theta}(x) = \int_{\mathcal{X}} g(x)f_{\theta}(x)d\nu(x). \quad (1.1.5)$$

When Eq. (1.1.5) is finite,

$$V_{\theta}g(X) = E_{\theta}(g(X) - E_{\theta}g(X))^2 \quad (1.1.6)$$

is called the **variance** of $g(x)$ when the parameter is θ . Especially when $g(x) = x$, $E_{\theta}X$ and $V_{\theta}X$ are called the mean and variance of the r.v. X respectively when the parameter is θ . Let $h(x)$ is a measurable function as well. When $E_{\theta}h(X)$ is finite,

$$CV_{\theta}(g(X), h(X)) = E_{\theta}(g(X) - E_{\theta}g(X))(h(X) - E_{\theta}h(X)) \quad (1.1.7)$$

is said to be the **covariance** of $g(x)$ and $h(x)$ when the parameter is θ .

It can be seen that the measure ν plays a key role in statistical spaces. Generally we consider only two forms of ν , one is the Counting measure, and the other is the Lebesgue measure.

Let \mathbf{Z}_+ denote a set of zero and positive integers, and $\mathcal{A}_{\mathbf{Z}_+}$ denote a σ -algebra generated by some subsets of \mathbf{Z}_+ , then the Counting measure ν depends on indicator functions defined on $(\mathbf{Z}_+, \mathcal{A}_{\mathbf{Z}_+})$, i.e., for $A \in \mathcal{A}_{\mathbf{Z}_+}$,

$$\nu(A) = \sum_{x \in A} I_A(x),$$

which indicates the number of elements in A . Let \mathcal{P} be the family of probability distributions defined on $(\mathbf{Z}_+, \mathcal{A}_{\mathbf{Z}_+})$. If $\mathcal{P} \ll \nu$, where ν is a Counting measure, then \mathcal{P} is called a **discrete probability distribution**.

Example 1.1.1. (Binomial distribution $Bi(n, \theta)$, where $\theta \in (0, 1)$)
Let $A = \{x; x \leq n, x \in \mathbf{Z}_+\}$, then the probability density distribution (pdf) of the binomial distribution $Bi(n, \theta)$ is

$$P_{\theta}(X = x) = \binom{n}{x} \theta^x (1 - \theta)^{n-x} I_A(x).$$

Its mean and variance are

$$E_{\theta}X = n\theta, \text{ and } V_{\theta}X = n\theta(1 - \theta),$$

respectively.

Example 1.1.2. (Poisson distribution $P(\theta)$, where $\theta > 0$) The pdf of the Poisson distribution $P(\theta)$ is

$$P_\theta(X = x) = e^{-\theta} \frac{\theta^x}{x!} I_{\mathbf{Z}_+}(x).$$

Its mean and variance are

$$E_\theta X = \theta, \text{ and } V_\theta X = \theta,$$

respectively.

Example 1.1.3. (Multinomial distribution $M(n; \theta)$, where $\theta = (p_1, \dots, p_k)$, $p_i > 0$, and $\sum_{i=1}^k p_i = 1$) Let $A = \{\mathbf{x} = (x_1, \dots, x_k); x_i \in \mathbf{Z}_+, \sum_{i=1}^k x_i = n\}$, then the pdf of the multinomial distribution $M(n; \theta)$ is

$$P_\theta(X_1 = x_1, \dots, X_k = x_k) = \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k} I_A(\mathbf{x}).$$

Its mean and variance of X_i are

$$E_\theta X_i = np_i, \text{ and } V_\theta X_i = np_i(1 - p_i),$$

respectively. When $i \neq j$, the covariance of X_i and X_j is

$$CV_\theta(X_i, X_j) = -np_i p_j.$$

Let \mathbf{R} denote the real space, and \mathcal{B} be a Borel algebra generated by some subsets of \mathbf{R} . Then the Lebesgue measure ν depends on the length of interval defined on $(\mathbf{R}, \mathcal{B})$, i.e., for the half-open interval $(a, b] \in \mathcal{B}$,

$$\nu((a, b]) = b - a.$$

Let \mathcal{P} be a family of probability distributions defined on $(\mathbf{R}, \mathcal{B})$. If $\mathcal{P} \ll \nu$, where ν is the Lebesgue measure, then \mathcal{P} is a **continuous probability distribution**.

Example 1.1.4. (Uniform distribution $U(a, b)$, where $a < b$) The pdf of the uniform distribution $U(a, b)$ is

$$f_\theta(x) = \begin{cases} \frac{1}{b-a}, & \text{if } x \in [a, b] \\ 0, & \text{otherwise,} \end{cases}$$

where $\theta = (a, b)$. The mean and the variance of X are

$$E_\theta X = \frac{1}{2}(a + b), \text{ and } V_\theta X = \frac{1}{12}(b - a)^2,$$

respectively.

Example 1.1.5. (Normal distribution $N(\mu, \sigma^2)$, where $\sigma^2 > 0$) The pdf of the normal distribution $N(\mu, \sigma^2)$ is

$$f_{\theta}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{1}{2\sigma^2}(x - \mu)^2 \right\},$$

where $\theta = (\mu, \sigma^2)$. The mean and the variance of X are

$$E_{\theta}X = \mu, \quad \text{and} \quad V_{\theta}X = \sigma^2,$$

respectively. Especially when $\mu = 0$ and $\sigma = 1$, it is called a **standard normal distribution** with pdf

$$f_{\theta}(x) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2}x^2 \right\}.$$

Example 1.1.6. (Exponential distribution $E(\mu, \sigma)$, where $\sigma > 0$) The pdf of the exponential distribution $E(\mu, \sigma)$ is

$$f_{\theta}(x) = \begin{cases} \frac{1}{\sigma} \exp \left\{ -\frac{1}{\sigma}(x - \mu) \right\}, & \text{if } x \geq \mu, \\ 0, & \text{otherwise,} \end{cases}$$

where $\theta = (\mu, \sigma)$. The mean and the variance of X are

$$E_{\theta}X = \mu + \sigma, \quad \text{and} \quad V_{\theta}X = \sigma^2,$$

respectively.

Example 1.1.7. (Cauchy distribution $C(\mu, \sigma)$, where $\sigma > 0$) The pdf of the Cauchy distribution $C(\mu, \sigma)$ is

$$f_{\theta}(x) = \frac{1}{\pi} \frac{\sigma}{\sigma^2 + (x - \mu)^2},$$

where $\theta = (\mu, \sigma)$. The mean and the variance of the Cauchy distribution do not exist.

Example 1.1.8. (The χ^2 distribution $\chi^2(n)$ and noncentral χ^2 distribution $\chi^2(n, \alpha^2)$) Let X_1, \dots, X_n be mutually independent random variables, and $X_i \sim N(\mu_i, 1), i = 1, \dots, n$. Let $\alpha^2 = \sum_{i=1}^n \mu_i^2$, and

$X = \sum_{i=1}^n X_i^2$. Then the pdf of X is

$$f_{\theta}(x) = \begin{cases} \sum_{s=0}^{\infty} e^{-\alpha^2/2} \frac{(\alpha^2/2)^s}{s!} \frac{1}{2^{n/2+s}\Gamma(n/2+s)} x^{n/2+s-1} e^{-x/2}, & \text{if } x > 0, \\ 0, & \text{otherwise,} \end{cases}$$

where $\theta = (n, \alpha^2)$. Especially when $\alpha = 0$, we call it a χ^2 distribution with n degrees of freedom with pdf

$$f_{\theta}(x) = \begin{cases} \frac{1}{2^{n/2}\Gamma(n/2)}x^{n/2-1}e^{-x/2}, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Example 1.1.9. (Student’s t -distribution $t(n)$ and noncentral Student’s t -distribution $t(n, \alpha)$) Let Y and Z be mutually independent random variables, where $Y \sim N(\alpha, 1)$, and $Z \sim \chi^2(n)$. Let $X = Y/\sqrt{Z/n}$. Then X has a noncentral Student’s t -distribution with n degrees of freedom, and its pdf is

$$f_{\theta}(x) = \frac{n^{n/2}}{\sqrt{\pi}\Gamma(\frac{n}{2})} \frac{e^{-\alpha^2/2}}{(n+x^2)^{(n+1)/2}} \sum_{s=0}^{\infty} \Gamma\left(\frac{n+s+1}{2}\right) \frac{\alpha^s}{s!} \left(\frac{2x^2}{n+x^2}\right)^{s/2},$$

where $\theta = (n, \alpha)$. Especially when $\alpha = 0$, X has a Student’s t -distribution with n degrees of freedom, and its pdf is

$$f_{\theta}(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}.$$

Example 1.1.10. (F -distribution $F(m, n)$ and noncentral F -distribution $F(m, n, \alpha^2)$) Let Y and Z be mutually independent random variable, where $Y \sim \chi^2(m, \alpha^2)$, $Z \sim \chi^2(n)$. Let $X = m^{-1}Y/(n^{-1}Z)$. Then X has a noncentral F -distribution with m and n degrees of freedom, its pdf is

$$f_{\theta}(x) = \begin{cases} e^{-\frac{\alpha^2}{2}} \sum_{s=0}^{\infty} \frac{(\frac{\alpha^2}{2})^s}{s!} \frac{n^{\frac{m}{2}} m^{\frac{m}{2}+s}}{B(\frac{m}{2} + s, \frac{n}{2})} \frac{x^{\frac{m}{2}+s-1}}{(n+mx)^{\frac{m+n}{2}+s}}, & \text{if } x > 0 \\ 0, & \text{otherwise,} \end{cases}$$

where $\theta = (m, n, \alpha^2)$ and $B(\frac{m}{2} + s, \frac{n}{2}) = \frac{\Gamma(\frac{m}{2} + s)\Gamma(\frac{n}{2})}{\Gamma(\frac{m+n}{2} + s)}$. Especially when $\alpha = 0$, X has an F -distribution with m and n degrees of freedom, its pdf

$$f_{\theta}(x) = \begin{cases} \frac{n^{\frac{m}{2}} m^{\frac{m}{2}}}{B(\frac{m}{2}, \frac{n}{2})} \frac{x^{m/2-1}}{(n+mx)^{(m+n)/2}}, & \text{if } x > 0, \\ 0, & \text{otherwise,} \end{cases}$$

where $B(\frac{m}{2}, \frac{n}{2}) = \frac{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}{\Gamma(\frac{m+n}{2})}$.

Example 1.1.11. (Multidimensional normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ is an n -dimensional vector, and $\boldsymbol{\Sigma}$ is an $n \times n$ positive definite matrix) The pdf of the multidimensional normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = (2\pi)^{-n/2} \frac{1}{\sqrt{\det \boldsymbol{\Sigma}}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\},$$

where $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathbf{x} = (x_1, \dots, x_n)'$, $E_{\boldsymbol{\theta}} \mathbf{X} = \boldsymbol{\mu}$, $\boldsymbol{\Sigma} = (\sigma_{ij})$, and $\sigma_{ij} = E_{\boldsymbol{\theta}}(X_i - \mu_i)(X_j - \mu_j)$.

Example 1.1.12. (Beta distribution $Be(a, b)$, where $a > 0, b > 0$) The pdf of the beta distribution $Be(a, b)$ is

$$f_{\boldsymbol{\theta}}(x) = \begin{cases} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1}, & \text{if } 0 \leq x \leq 1 \\ 0, & \text{otherwise,} \end{cases}$$

where $\boldsymbol{\theta} = (a, b)$. The mean and variance of X are

$$E_{\boldsymbol{\theta}} X = \frac{a}{a+b}, \quad \text{and} \quad V_{\boldsymbol{\theta}} X = \frac{ab}{(a+b+1)(a+b)^2},$$

respectively.

It can be seen from the above discussion that it is important that the family of probability distributions is absolutely continuous with respect to a σ -finite measure in statistical spaces, since it forms the foundation of probability calculation which depends on parameters. In that way, how to judge whether there exists a σ -finite measure or not for a given family of probability distributions, such that it is absolutely continuous about the σ -finite measure? This will involve some notions of the family of probability distributions, such as divisibility.

1.1.2 Equivalence and Divisibility of Family of Probability Distributions

Now we discuss the relation between two measures. Let ν and λ be two σ -finite measures defined in the sample space $(\mathcal{X}, \mathcal{A})$. If $\nu \ll \lambda$ and $\lambda \ll \nu$, then ν and λ are said to be **mutually absolutely continuous**. The necessity of the assumption of σ -finiteness of measure can be seen in the following theorems.

Theorem 1.1.2. *Let $(\mathcal{X}, \mathcal{A}, \nu)$ be a measure space. If ν is σ -finite, then there exists a probability measure which is mutually absolutely continuous with respect to ν .*

Proof. Since \mathcal{A} is a σ -algebra, and ν is σ -finite, then there exists a series of sets $\{A_i\}$, such that

$$A_i \in \mathcal{A}, \quad \nu(A_i) < \infty, \quad A_1 \subset A_2 \subset \dots, \quad \text{and} \quad \bigcup_{i=1}^{\infty} A_i = \mathcal{X}.$$

Without loss of generality, let $\nu(A_1) > 0$. For $\forall A \in \mathcal{A}$, let

$$\lambda(A) = \sum_{i=1}^{\infty} \frac{1}{2^i} \frac{\nu(A \cap A_i)}{\nu(A_i)}.$$

Obviously λ is a probability measure, and we have $\nu \ll \lambda$ and $\lambda \ll \nu$.

Let \mathcal{P} and ν be a family of probability distributions and a σ -finite measure respectively defined in the sample space $(\mathcal{X}, \mathcal{A})$. If 1) $\mathcal{P} \ll \nu$; and 2) $\nu(A) = 0$ if $P_\theta(A) = 0$ for $\forall A \in \mathcal{A}$ and $\forall \theta \in \Theta$, then \mathcal{P} and ν are **equivalent**. □

Theorem 1.1.3. *Let \mathcal{P} be a family of probability distributions defined in the probability space $(\mathcal{X}, \mathcal{A})$. \mathcal{P} is absolutely continuous with respect to some σ -infinite measure, and its necessary and sufficient condition is that there exists a probability measure that is equivalent to \mathcal{P} with the form*

$$\lambda = \sum_{i=1}^{\infty} c_i P_{\theta_i} \tag{1.1.8}$$

where $\theta_i \in \Theta$, $c_i > 0$, and $\sum_{i=1}^{\infty} c_i = 1$.

Proof. The sufficiency is obviously given by the condition of equivalence, so it suffices to prove the necessity as below.

Let ν be a σ -finite measure, satisfying $\mathcal{P} \ll \nu$. From Theorem 1.1.2, we can consider ν as a probability measure. Let

$$\mathcal{Q} = \left\{ Q; Q = \sum_{i=1}^{\infty} c_i P_{\theta_i}, \theta_i \in \Theta, c_i > 0, \sum_{i=1}^{\infty} c_i = 1 \right\}.$$

Then \mathcal{Q} is also a family of probability distributions defined in $(\mathcal{X}, \mathcal{A})$, and we have $\mathcal{Q} \ll \nu$. Let

$$\sup_{Q \in \mathcal{Q}} \nu \left\{ x; \frac{dQ(x)}{d\nu(x)} > 0 \right\} = \alpha.$$

From the definition of supremum, there exists a series of probability distributions $\{Q_n\}$ in \mathcal{Q} , such that

$$\lim_{n \rightarrow \infty} \nu \left\{ x; \frac{dQ_n(x)}{d\nu(x)} > 0 \right\} = \alpha. \tag{1.1.9}$$

Furthermore, let $\lambda = \sum_{n=1}^{\infty} \frac{1}{2^n} Q_n$, then $\lambda \in \mathcal{Q}$. And from Eq. (1.1.9), we have

$$\nu \left\{ x; \frac{d\lambda(x)}{d\nu(x)} > 0 \right\} = \alpha.$$

Now we prove that λ and \mathcal{P} are equivalent. Since $\lambda \in \mathcal{Q}$, for $\forall A \in \mathcal{A}$, then we have $\lambda(A) = 0$ if $P_\theta(A) = 0$ for $\forall \theta \in \Theta$. We use the proof by contradiction method to show that $\mathcal{P} \ll \lambda$. Suppose that there exists $A \in \mathcal{A}$, and $\theta \in \Theta$, such that $\lambda(A) = 0$ and $P_\theta(A) > 0$. Let $Q_0 = \frac{1}{2}(\lambda + P_\theta)$. Obviously $Q_0 \in \mathcal{Q}$. From $\lambda(A) = 0$, $d\lambda(x)/d\nu(x) = 0$ holds almost everywhere in A . Therefore

$$\begin{aligned} \nu \left\{ x; \frac{dQ_0(x)}{d\nu(x)} > 0 \right\} &= \nu \left\{ x; \frac{d\lambda(x)}{d\nu(x)} > 0 \text{ or } \frac{dP_\theta(x)}{d\nu(x)} > 0 \right\} \\ &\geq \nu \left\{ x; \frac{d\lambda(x)}{d\nu(x)} > 0 \right\} + \nu \left\{ x \in A; \frac{dP_\theta(x)}{d\nu(x)} > 0 \right\} \\ &> \nu \left\{ x; \frac{d\lambda(x)}{d\nu(x)} > 0 \right\} \\ &= \alpha. \end{aligned}$$

This contradicts with the definition of α . □

Though the above theorem provides a judging criterion, the conditions themselves are very difficult to be verified. So it is necessary to make a further analysis of the family of probability distributions. At first we review the definition of distance.

Let d be a bivariate nonnegative function defined on \mathcal{X} . If for $\forall x, y, z \in \mathcal{X}$, d satisfies the following conditions

- (1) $d(x, x) = 0$,
- (2) $d(x, y) = d(y, x)$,
- (3) $d(x, z) \leq d(x, y) + d(y, z)$,

then d is called a **quasi-distance**, and \mathcal{X} a **quasi-distance space**. If an additional condition is attached to Condition (1), *i.e.*, $d(x, y) = 0$ iff $x = y$, then d is called a **distance**, and \mathcal{X} a **distance space**.

Example 1.1.13. Let $(\mathcal{X}, \mathcal{A}, P)$ be a probability space. For $A_1, A_2 \in \mathcal{A}$, let

$$\begin{aligned} A_1 \triangle A_2 &= (A_1 - A_2) \cup (A_2 - A_1), \\ d(A_1, A_2) &= P(A_1 \triangle A_2). \end{aligned}$$

Then the σ -algebra \mathcal{A} forms a quasi-distance space about d .

Example 1.1.14. Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be a statistical space. For $\theta_1, \theta_2 \in \Theta$, let

$$d(\theta_1, \theta_2) = \sup_{A \in \mathcal{A}} |P_{\theta_1}(A) - P_{\theta_2}(A)|.$$

Then the family of probability distributions \mathcal{P} forms a quasi-distance space about d . If θ and P_θ are one-to-one, then \mathcal{P} forms a quasi-distance space about d .

Example 1.1.15. Let \mathcal{P} and ν be the family of probability distributions and σ -finite measure in the sample space $(\mathcal{X}, \mathcal{A})$ respectively, and $\mathcal{P} \ll \nu$. Let \mathcal{F} be a set of all the pdfs, *i.e.*

$$\mathcal{F} = \left\{ f_\theta; f_\theta = \frac{dP_\theta}{d\nu}, \theta \in \Theta \right\}.$$

For $f_{\theta_1}, f_{\theta_2} \in \mathcal{F}$, let

$$d^*(f_{\theta_1}, f_{\theta_2}) = \int_{\mathcal{X}} |f_{\theta_1}(x) - f_{\theta_2}(x)| d\nu(x).$$

Then \mathcal{F} forms a quasi-distance space about d^* , and forms a distance space almost everywhere, *i.e.*,

$$d^*(f_{\theta_1}, f_{\theta_2}) = 0 \iff f_{\theta_1}(x) = f_{\theta_2}(x) \quad \text{a.e. } \nu.$$

It can be verified that, if θ and P_θ are one-to-one, then there exists an equivalent relation between the two distances defined in Examples 1.1.4 and 1.1.5, *i.e.*, for $\theta_1, \theta_2 \in \Theta$ we have

$$\begin{aligned} d(\theta_1, \theta_2) &= \sup_{A \in \mathcal{A}} |P_{\theta_1}(A) - P_{\theta_2}(A)| \\ &= \frac{1}{2} \int_{\mathcal{X}} |f_{\theta_1}(x) - f_{\theta_2}(x)| d\nu(x) \\ &= \frac{1}{2} d^*(f_{\theta_1}, f_{\theta_2}), \end{aligned}$$

where $f_{\theta_i} = dP_{\theta_i}/d\nu$, $i = 1, 2$.

Let \mathcal{X} form a quasi-distance space about d . If there exists a countable subset \mathcal{X}_0 of \mathcal{X} , we can find an $x_0 \in \mathcal{X}_0$, such that $d(x, x_0) < \varepsilon$, the \mathcal{X} is called **divisible** about d , and \mathcal{X}_0 a **countable dense subset** of \mathcal{X} .

Theorem 1.1.4. Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be a statistical space. Suppose that \mathcal{P} is divisible about the distance d defined in Example 1.1.14, and the corresponding countable dense subset is $\mathcal{P}_0 = \{P_{\theta_n}\}$. For $A \in \mathcal{A}$, let

$$\lambda(A) = \sum_{n=1}^{\infty} \frac{1}{2^n} P_{\theta_n}(A). \tag{1.1.10}$$

then \mathcal{P} is equivalent to λ .

Proof. From Eq. (1.1.10), λ is also a probability measure defined on $(\mathcal{X}, \mathcal{A})$, and for $A \in \mathcal{A}$, we must have $\lambda(A) = 0$ if $P_\theta(A) = 0$ for $\forall \theta \in \Theta$. Now we need to prove that $\mathcal{P} \ll \lambda$.

For $A \in \mathcal{A}$, if $\lambda(A) = 0$, from Eq. (1.1.10), $P_{\theta_n}(A) = 0$ holds for all $P_{\theta_n} \in \mathcal{P}_0$. We will Prove that $P_\theta(A) = 0$, for $\forall P_\theta \in \mathcal{P}$. Since \mathcal{P}_0 is a countable dense subset of \mathcal{P} , for $\forall \varepsilon > 0$, there exists $P_{\theta_n} \in \mathcal{P}_0$, such that $d(\theta, \theta_n) < \varepsilon$. Since $P_{\theta_n}(A) = 0$,

$$P_\theta(A) = |P_\theta(A) - P_{\theta_n}(A)| \leq d(\theta, \theta_n) < \varepsilon.$$

By the arbitrariness of $\varepsilon > 0$, $P_\theta(A) = 0$. □

Theorem 1.1.5. *Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be a statistical space. If \mathcal{P} is divisible about the distance d defined in Example 1.1.14, then there exists a σ -finite measure ν defined on $(\mathcal{X}, \mathcal{A})$, such that $\mathcal{P} \ll \nu$.*

The above theorem follows the results of Theorems 1.1.3 and 1.1.4 directly, and gives a condition that is easier to be verified. Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be a statistical space. If the parameter Θ is an at most countable set, then it obviously satisfies the conditions of Theorem 1.1.5; otherwise, it suffices to assume that Θ_0 be a subset formed by the rational numbers in the parameter space Θ , let $\mathcal{P}_0 = \{P_\theta; \theta \in \Theta_0\}$, and then verify whether \mathcal{P}_0 can form a countable dense subset of \mathcal{P} about the distance defined in Example 1.1.14 or not. As a result, we have a very clear insight into the measure in statistical spaces and the relation between the measure and the parameter.

1.2 Conditional Probability and Sufficient Statistics

First we will review the conditional probability in the classical probability model by a simple example. When tossing a coin, there are two possible outcomes: Head (H) or Tail (T). And then there are four possible cases when tossing two times: {HH, HT, TH, TT}. Let A denote the event {H occurs at least one time}, and B the event {H occurs exactly once}. Suppose that the coin is even, then $P(A) = 3/4$, $P(AB) = 1/2$; the conditional probability of B given A is

$$P(B|A) = \frac{2}{3} = \left(\frac{1}{2}\right) / \left(\frac{3}{4}\right) = P(AB)/P(A),$$

which is equivalent to the following product-form

$$P(AB) = P(B|A)P(A). \tag{1.2.1}$$

This section will discuss a general conditional probability which involves the concept of statistics.

1.2.1 Statistics and Random Variables

In the former discussion, we have used the terms of “Random Variable” and “Statistics” in a non-formal way, now we will define them more exactly. Let $(\mathcal{X}, \mathcal{A}, P)$ be a probability space, and $(\mathcal{Y}, \mathcal{B})$ a measurable space. Let t be a mapping from $(\mathcal{X}, \mathcal{A})$ to $(\mathcal{Y}, \mathcal{B})$. If the mapping is measurable, *i.e.*

$$t^{-1}(B) \in \mathcal{A}, \quad \forall B \in \mathcal{B},$$

and then t is called a **statistic**. Especially, when \mathcal{Y} is a subset of the real space, t is called a **random variable**. If t is a statistic, let

$$\mathcal{A}_t = \{t^{-1}(B); B \in \mathcal{B}\} \quad \text{and} \quad Q(B) = P(t^{-1}(B)), \quad \forall B \in \mathcal{B}. \quad (1.2.2)$$

Obviously \mathcal{A}_t is also a σ -algebra, and we have $\mathcal{A}_t \subset \mathcal{A}$; $Q(B)$ is a probability measure defined in $(\mathcal{Y}, \mathcal{B})$, which is called an **induced probability measure** by t . $(\mathcal{Y}, \mathcal{B}, Q)$ is called an **induced probability space** by t . Sometimes $(\mathcal{X}, \mathcal{A}_t, P)$ is also called an **induced probability space** by t .

Theorem 1.2.1. *For a given probability space $(\mathcal{X}, \mathcal{A}, P)$, let $(\mathcal{Y}, \mathcal{B}, Q)$ be the induced probability space by t . Then for any \mathcal{B} -measurable function h we have*

$$\int_{\mathcal{X}} h(t(x))dP(x) = \int_{\mathcal{Y}} h(y)dQ(y). \quad (1.2.3)$$

Proof. Obviously $h(t(x))$ is an \mathcal{A}_t -measurable function, thus the left-hand side of Eq. (1.2.3) is integrable. When $h(y) = I_B(y), B \in \mathcal{B}$, Eq. (1.2.3) is equivalent to $P(t^{-1}(B)) = Q(B)$, and by Eq. (1.2.2), it holds obviously. And then by the I-method, we can prove the conclusion. \square

From the above theorem, we sometimes write $Q(B) = P(t(X) \in B)$ for $B \in \mathcal{B}$. The situation with two statistics will be considered as below. Let t and u be statistics from $(\mathcal{X}, \mathcal{A}, P)$ to measurable space $(\mathcal{Y}, \mathcal{B})$ and $(\mathcal{Z}, \mathcal{C})$, respectively. Let \mathcal{A}_{tu} be a σ -algebra generated by $\{t^{-1}(B) \cap u^{-1}(C); B \in \mathcal{B}, C \in \mathcal{C}\}$. Obviously $\mathcal{A}_{tu} \subset \mathcal{A}$, $(\mathcal{X}, \mathcal{A}_{tu}, P)$ also forms a probability space. If for $\forall B \in \mathcal{B}, \forall C \in \mathcal{C}$ we have

$$P(t^{-1}(B) \cap u^{-1}(C)) = P(t^{-1}(B))P(u^{-1}(C)),$$

then t and u are said to be **mutually independent**. According to Theorem 1.2.1, the above equation can be also written as

$$P(t(X) \in B, u(X) \in C) = P(t(X) \in B)P(u(X) \in C). \quad (1.2.4)$$

Especially when \mathcal{Y} and \mathcal{Z} are both subsets of the real space, the r.v.'s $t(X)$ and $u(X)$ are said to be mutually independent. The definition can easily be extended to the general situation.

Let t_i be a statistic from $(\mathcal{X}, \mathcal{A}, P)$ to measurable space $(\mathcal{Y}_i, \mathcal{B}_i)$, $i = 1, \dots, n$. If for $\forall B_i \in \mathcal{B}_i, i = 1, \dots, n$, we have

$$P\left(\bigcap_{i=1}^n t_i^{-1}(B_i)\right) = \prod_{i=1}^n P(t_i^{-1}(B_i)),$$

then t_1, \dots, t_n are said to be **mutually independent**. Especially when $\mathcal{Y}_i, i = 1, \dots, n$, all are the subsets of the real space, the r.v.'s $t_1(X), \dots, t_n(X)$ are said to be mutually independent.

Now the statistical space $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ is considered, where $\mathcal{P} = \{P_\theta; \theta \in \Theta\}$. Similar to Eq. (1.2.2), for $\theta \in \Theta$, let $Q_\theta(B) = P_\theta(t^{-1}(B))$ for $B \in \mathcal{B}$, then $\mathcal{Q} = \{Q_\theta; \theta \in \Theta\}$ also forms a family of probability distributions, which will be called an **induced family of probability distributions** by t and $(\mathcal{Y}, \mathcal{B}, \mathcal{Q})$ an **induced statistical space** by t . If P_θ and θ are one-to-one, then Q_θ and θ are also one-to-one. Therefore, when we are making a statistical inference to the parameter θ , to simplify the problem, we usually find an appropriate statistic t , and then study the problem after transformed it into the induced statistical space by t .

Example 1.2.1. Let X_1, \dots, X_n be the mutually independent r.v.'s from a normal distribution with mean θ and variance 1, which is sometimes denoted as

$$X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} N(\theta, 1), \quad (1.2.5)$$

and then $\mathbf{X} = (X_1, \dots, X_n)'$ has an n -dimensional normal distribution with n -dimensional mean vector $(\theta, \dots, \theta)'$ and covariance matrix $\mathbf{\Sigma} = \mathbf{I}$. Since we only make a statistical inference to the parameter θ , the sample mean can be considered as a statistic, *i.e.*,

$$t(\mathbf{X}) = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

Then $t(\mathbf{X}) \sim N(\theta, 1/n)$, which is a univariate normal distribution. The parameter θ does not change, but the variance becomes smaller than that in Eq. (1.2.5). Considering intuitively, when we make statistical inference to θ we can obtain a higher precision by using $t(\mathbf{X})$ than by using X_i ; the same precision as using $\mathbf{X} = (X_1, \dots, X_n)'$, but the corresponding computation is more convenient. We will further discuss the problem later.

Let μ be a σ -finite measure defined on $(\mathcal{X}, \mathcal{A})$, satisfying $\mathcal{P} \ll \mu$. For $\theta \in \Theta$, the pdf is $dP_\theta/d\mu = f_\theta$. Similar to Eq. (1.2.2), for $B \in \mathcal{B}$, let

$$\nu(B) = \mu(t^{-1}(B)).$$

And then ν is a σ -finite measure defined on $(\mathcal{Y}, \mathcal{B})$. It is easy to verify that $\mathcal{Q} \ll \nu$. For $\theta \in \Theta$, let $dQ_\theta/d\nu = g_\theta$. And then Eq. (1.2.3) in Theorem 1.2.1 can be written as

$$\int_{\mathcal{X}} h(t(x))f_\theta(x)d\mu(x) = \int_{\mathcal{Y}} h(y)g_\theta(y)d\nu(y). \tag{1.2.6}$$

1.2.2 Conditional Probability

For given probability space $(\mathcal{X}, \mathcal{A}, P)$, let $(\mathcal{Y}, \mathcal{B}, Q)$ be the induced probability space by t . Let h be a measurable function defined in $(\mathcal{X}, \mathcal{A}, P)$, satisfying $Eh(X) < \infty$. For $\forall B \in \mathcal{B}$, let

$$R(B) = \int_{t^{-1}(B)} h(x)dP(x). \tag{1.2.7}$$

And then R is a measure defined on $(\mathcal{Y}, \mathcal{B})$. From Eq. (1.2.2), $R \ll Q$. Applying the Radon-Nikodym Theorem, there exists a \mathcal{B} -measurable function $g = dR/dQ$, s.t. for $B \in \mathcal{B}$,

$$R(B) = \int_B g(y)dQ(y). \tag{1.2.8}$$

Comparing Eq. (1.2.7) with Eq. (1.2.8), we can get that, for $\forall B \in \mathcal{B}$,

$$\int_{t^{-1}(B)} h(x)dP(x) = \int_B g(y)dQ(y). \tag{1.2.9}$$

g in Eq. (1.2.9) is unique almost everywhere with respect to the measure Q . The function $g(y)$ in Eq. (1.2.9) is called a **conditional mean** of $h(x)$ when $t(x) = y$ is given, which is denoted as

$$E(h(X)|y). \tag{1.2.10}$$

Obviously this is a function of y . Equation (1.2.9) can be rewritten as, for $\forall B \in \mathcal{B}$,

$$\int_{t^{-1}(B)} h(x)dP(x) = \int_B E(h(X)|y)dQ(y). \tag{1.2.11}$$

Especially, when $h(x) = I_A(x)$, for $A \in \mathcal{A}$, let $P(A|y) = E(I_A(X)|y)$, Eq. (1.2.11) can be rewritten as, for $\forall B \in \mathcal{B}$

$$P\{A \cap t^{-1}(B)\} = \int_B P(A|y)dQ(y). \tag{1.2.12}$$

$P(A|y)$ in Eq. (1.2.12) is called a **conditional probability** of A given $t(x) = y$.

The above result is applicable to the statistical space. Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be a given statistical space, where $\mathcal{P} = \{P_\theta; \theta \in \Theta\}$; let $(\mathcal{Y}, \mathcal{B}, \mathcal{Q})$ be the induced statistical space by t , then \mathcal{Q} can be denoted as $\mathcal{Q} = \{Q_\theta; \theta \in \Theta\}$. Equations (1.2.11) and (1.2.12) can be written as

$$\int_{t^{-1}(B)} h(x)dP_\theta(x) = \int_B E_\theta(h(X)|y)dQ_\theta(y), \tag{1.2.13}$$

and

$$P_\theta(A \cap t^{-1}(B)) = \int_B P_\theta(A|y)dQ_\theta(y), \tag{1.2.14}$$

respectively. Therefore, both the conditional mean and the conditional probability depend on θ in the statistical space.

Example 1.2.2. Let $(\mathcal{X} \times \mathcal{Y}, \mathcal{A} \times \mathcal{B}, \mu \times \nu)$ denote the direct product space of two σ -infinite measure spaces $(\mathcal{X}, \mathcal{A}, \mu)$ and $(\mathcal{Y}, \mathcal{B}, \nu)$, and $(\mathcal{X} \times \mathcal{Y}, \mathcal{A} \times \mathcal{B}, \mathcal{P})$ the statistical space corresponding to the direct product space, where $\mathcal{P} = \{P_\theta; \theta \in \Theta\}$, and $\mathcal{P} \ll \mu \times \nu$. Let $dP_\theta(x, y) = f_\theta(x, y)d\mu(x)d\nu(y)$. Let h be an $\mathcal{A} \times \mathcal{B}$ measurable function, $E_\theta h(X, Y) < \infty$.

Let $t(x, y) = y$, then t is a statistic from $\mathcal{A} \times \mathcal{B}$ to \mathcal{B} , and the induced statistical space by t can be denoted as $(\mathcal{Y}, \mathcal{B}, \mathcal{Q})$. For $\forall B \in \mathcal{B}$, since $Q_\theta(B) = P_\theta(t^{-1}(B))$, we have

$$Q_\theta(B) = \int_{\mathcal{X} \times B} f_\theta(x, y)d\mu(x)d\nu(y) = \int_B \left[\int_{\mathcal{X}} f_\theta(x, y)d\mu(x) \right] d\nu(y). \tag{1.2.15}$$

Obviously $\mathcal{Q} \ll \nu$. For $\theta \in \Theta$, let $q_\theta = dQ_\theta/d\nu$, from the above equation we have

$$q_\theta(y) = \int_{\mathcal{X}} f_\theta(x, y)d\mu(x). \tag{1.2.16}$$

q_θ is called a **marginal density** of y . From Eqs. (1.2.13) and (1.2.15), we have

$$\int_B \left[\int_{\mathcal{X}} h(x, y)f_\theta(x, y)d\mu(x) \right] d\nu(y) = \int_B q_\theta(y)E_\theta(h(X, Y)|y)d\nu(y).$$

From Theorem 1.1.3, we can set that $\nu\{y; q_\theta(y) = 0\} = 0$. Since the above equation holds for any $B \in \mathcal{B}$, we have

$$E_\theta(h(X, Y)|y) = \begin{cases} \frac{1}{q_\theta(y)} \int_{\mathcal{X}} h(x, y)f_\theta(x, y)d\mu(x), & \text{if } q_\theta(y) > 0, \\ E_\theta h(X, Y), & \text{if } q_\theta(y) = 0. \end{cases} \tag{1.2.17}$$

a.e. with respect to the measure ν . Especially when $h(x, y) = I_C(x, y)$ for $C \in \mathcal{A} \times \mathcal{B}$, we have

$$P_\theta(C|y) = \begin{cases} \frac{1}{q_\theta(y)} \int_{C_y} f_\theta(x, y) d\mu(x), & \text{if } q_\theta(y) > 0, \\ P_\theta(C), & \text{if } q_\theta(y) = 0, \end{cases} \quad (1.2.18)$$

where $C_y = \{x; (x, y) \in C\}$, and $y \in \mathcal{Y}$. When $q_\theta(y) > 0$, obviously $P_\theta(\cdot|y) \ll \mu$. Let $f_\theta(x|y)$ denote its pdf, then Eq. (1.2.18) can be written as

$$\int_{C_y} f_\theta(x|y) d\mu(x) = \frac{1}{q_\theta(y)} \int_{C_y} f_\theta(x, y) d\mu(x).$$

Therefore, we have

$$f_\theta(x, y) = f_\theta(x|y)q_\theta(y) \quad (1.2.19)$$

a.e. with respect to the measure μ . That is corresponding to Eq. (1.2.1) under the classical probability model. Therefore, generally, the conditional mean of $h(x)$ can be rewritten as

$$E_\theta(h(X)|y) = \int_{\mathcal{X}} h(x) f_\theta(x|y) d\mu(x).$$

Taking the mean value with respect to y on both sides of the above equation, we have

$$\begin{aligned} E_\theta E_\theta(h(X)|Y) &= \int_{\mathcal{Y}} \left[\int_{\mathcal{X}} h(x) f_\theta(x|y) d\mu(x) \right] q_\theta(y) d\nu(y) \\ &= \int_{\mathcal{X}} h(x) \int_{\mathcal{Y}} f_\theta(x, y) d\nu(y) d\mu(x) \\ &= \int_{\mathcal{X}} h(x) s_\theta(x) d\mu(x) \\ &= E_\theta h(X), \end{aligned}$$

where $s_\theta(x) = \int f_\theta(x, y) d\nu(y)$ is the marginal density of x .

Sometimes, we need to consider the conditional probability in the same space, just as the conditional density defined in Eq. (1.2.1) under the classical probability model at the beginning of this section. For the given statistical space $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, let \mathcal{A}_0 be a sub σ -algebra of \mathcal{A} , i.e., $A_0 \in \mathcal{A}_0$, then we must have $A_0 \in \mathcal{A}$. For $\theta \in \Theta, \forall A_0 \in \mathcal{A}_0$, let

$$S_\theta(A_0) = \int_{A_0} h(x) dP_\theta(x), \quad (1.2.20)$$

then $S_\theta \ll P_\theta$. Let the pdf be $dS_\theta/dP_\theta = s_\theta$, which is an \mathcal{A}_0 measurable function. For $\forall A_0 \in \mathcal{A}_0$, we also have

$$S_\theta(A_0) = \int_{A_0} s_\theta(x) dP_\theta(x). \quad (1.2.21)$$

Combining Eq. (1.2.20) with Eq. (1.2.21), we have

$$\int_{A_0} h(x) dP_\theta(x) = \int_{A_0} s_\theta(x) dP_\theta(x) \quad (1.2.22)$$

for $\forall A_0 \in \mathcal{A}_0$. Note that h in the above equation is \mathcal{A} -measurable, while s_θ is \mathcal{A}_0 -measurable, which generally depends on the parameter θ . Generally, s_θ satisfying Eq. (1.2.22) is called the **conditional mean** of h given \mathcal{A}_0 , and denoted as

$$E_\theta(h(X)|\mathcal{A}_0, x).$$

Therefore Eq. (1.2.22) can be written as

$$\int_{A_0} h(x) dP_\theta(x) = \int_{A_0} E_\theta(h(X)|\mathcal{A}_0, x) dP_\theta(x) \quad (1.2.23)$$

for $\forall A_0 \in \mathcal{A}_0$. Especially when $h(x) = I_A(x)$, $A \in \mathcal{A}$,

$$P_\theta(A \cap A_0) = \int_{A_0} E_\theta(I_A(X)|\mathcal{A}_0, x) dP_\theta(x) = \int_{A_0} P_\theta(A|\mathcal{A}_0, x) dP_\theta(x), \quad (1.2.24)$$

of which $P_\theta(A|\mathcal{A}_0, x)$ is called the **conditional probability** of A given \mathcal{A}_0 .

In Eq. (1.2.23), let $A_0 = \mathcal{X}$, we can get the following theorem similar to Example 1.2.2.

Theorem 1.2.2. For the given statistical space $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, let \mathcal{A}_0 be a sub σ -algebra of \mathcal{A} , then for \mathcal{A} -integrable function h and $\forall \theta \in \Theta$, we have

$$E_\theta E_\theta(h(X)|\mathcal{A}_0, X) = E_\theta h(X).$$

Note that when $\mathcal{A}_0 = \mathcal{A}_t$, applying Theorem 1.2.1, from Eqs. (1.2.13) and (1.2.23) we have

$$\int_{A_0} E_\theta(h(X)|t(x)) dP_\theta(x) = \int_{A_0} E_\theta(h(X)|\mathcal{A}_0, x) dP_\theta(x)$$

for $\forall A_0 \in \mathcal{A}_t$. The above equation implies that $P_\theta(D) = 0$ for $D \in \mathcal{A}_t$ with $D = \{x; E_\theta(h(X)|t(x)) \neq E_\theta(h(X)|\mathcal{A}_0, x)\}$.

Example 1.2.3. For the given statistical space $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ with $\mathcal{P} = \{P_\theta; \theta \in \Theta\}$, let h be an \mathcal{A} -measurable function satisfying $E_\theta h(X) < \infty$ for $\forall \theta \in \Theta$.

- (i) If $\mathcal{A}_0 = \{\phi, \mathcal{X}\}$, then obviously

$$E_\theta(h(X)|\mathcal{A}_0, x) = E_\theta h(X).$$

That is the most common conditional mean.

- (ii) For $A_0 \in \mathcal{A}$, let $\mathcal{A}_0 = \{\phi, A_0, A_0^c, \mathcal{X}\}$, where A_0^c denotes the complement of A_0 in \mathcal{X} . Since the conditional mean must be an \mathcal{A}_0 -measurable function, the necessary and sufficient condition is that there exist two constants c_1 and c_2 , s.t.

$$E_\theta(h(X)|\mathcal{A}_0, x) = \begin{cases} c_1, & \text{if } x \in A_0, \\ c_2, & \text{if } x \in A_0^c. \end{cases}$$

Substituting into Eq. (1.2.23) yields

$$E_\theta(h(X)|\mathcal{A}_0, x) = \begin{cases} \frac{1}{P_\theta(A_0)} \int_{A_0} h(x) dP_\theta(x), & \text{if } x \in A_0 \\ \frac{1}{P_\theta(A_0^c)} \int_{A_0^c} h(x) dP_\theta(x), & \text{if } x \in A_0^c. \end{cases}$$

Especially when $h(x) = I_A(x)$, for $A \in \mathcal{A}$, we have

$$P_\theta(A|\mathcal{A}_0, x) = \begin{cases} \frac{1}{P_\theta(A_0)} P_\theta(A_0 \cap A), & \text{if } x \in A_0 \\ \frac{1}{P_\theta(A_0^c)} P_\theta(A_0^c \cap A), & \text{if } x \in A_0^c. \end{cases}$$

When $x \in A_0$, we will use $P_\theta(A|A_0)$ to denote $P_\theta(A|\mathcal{A}_0, x)$, we have

$$P_\theta(A \cap A_0) = P_\theta(A|A_0)P_\theta(A_0),$$

which coincides with Eq. (1.2.1) defined in the classical probability model.

From the definition of conditional mean, the conditional mean is an integral under the probability measure, so it is easy to prove the following properties.

- (1) For constants a and b ,

$$E_\theta(ah(X) + bg(X)|\mathcal{A}_0, x) = aE_\theta(h(X)|\mathcal{A}_0, x) + bE_\theta(g(X)|\mathcal{A}_0, x).$$

- (2) If $h(x) \geq 0$, then $E_\theta(h(X)|\mathcal{A}_0, x) \geq 0$.

- (3) **Monotone convergence theorem.** For an integrable function h and a series of measurable functions $\{h_n\}$, if $h_n \uparrow h$, then $E_\theta(h_n(X)|\mathcal{A}_0, x) \uparrow E_\theta(h(X)|\mathcal{A}_0, x)$.

- (4) **Dominated convergence theorem.** For a function h and a series of integrable functions $\{h_n\}$, if there exists a nonnegative integrable function g satisfying $|h_n(x)| \leq g(x)$, then $E_\theta(h_n(X)|\mathcal{A}_0, x) \rightarrow E_\theta(h(X)|\mathcal{A}_0, x)$ when $h_n \rightarrow h$.

Theorem 1.2.3. Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be a statistical space, $\mathcal{A}_0 \subset \mathcal{A}$ a sub σ -algebra, f an \mathcal{A} -measurable function, and g an \mathcal{A}_0 -measurable function. If $E_\theta g(X)h(X) < \infty$ for $\theta \in \Theta$, then we have

$$E_\theta(g(X)h(X)|\mathcal{A}_0, x) = g(x)E_\theta(f(X)|\mathcal{A}_0, x)$$

a.e. with respect to \mathcal{A}_0 and P_θ .

Proof. At first, we consider the case where g is a simple function, i.e. $g(x) = I_{A_0}(x)$ for $A_0 \in \mathcal{A}_0$. For $\forall A \in \mathcal{A}_0$, Note that $A \cap A_0 \in \mathcal{A}_0$, we have

$$\begin{aligned} \int_A I_{A_0}(x) E_\theta(h(X)|\mathcal{A}_0, x) dP_\theta(x) &= \int_{A \cap A_0} E_\theta(h(X)|\mathcal{A}_0, x) dP_\theta(x) \\ &= \int_{A \cap A_0} h(x) dP_\theta(x) \\ &= \int_A E_\theta(I_{A_0}(X)h(X)|\mathcal{A}_0, x) dP_\theta(x). \end{aligned}$$

From the above properties (1) and (3), based on the equivalent formulation about uniqueness a.e. discussed in the first section, we can prove the theorem by the I-method. \square

1.2.3 Sufficient Statistics

Let \mathbf{x} be a sample from the statistical space $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, where $\mathcal{P} = \{P_\theta; \theta \in \Theta\}$. Our aim is to make statistical inference to the parameter by using the sample \mathbf{x} . In practical inference, we need to find an appropriate statistic t , and then make inference to θ based on $t(\mathbf{x})$ and the induced statistical space by t . It is important to find “appropriate statistics”, since we can find many seemingly reasonable statistics in principle. Recall Example 1.2.1, let x_1, \dots, x_n be an independent sample from $N(\theta, 1)$, and $\mathbf{x} = (x_1, \dots, x_n)$. We have defined the sample mean, i.e. $t(\mathbf{x}) = \bar{x}$, to estimate θ . We can also define $t_1(\mathbf{x}) = x_1, t_{\max}(\mathbf{x}) = \max\{x_i\}$ and so on. And then which statistics are “appropriate”? What are the criteria for the appropriateness? Fisher proposed the concept of sufficient statistics in 1922. Halmos and Savage (1949), Bahadur (1954) gave an exact formulation about it and the proof of some equivalent propositions. The sufficiency of statistics has been one of the most important concepts in statistics.

Basically speaking, a sufficient statistic is a statistic without losing any information about the parameter θ , which can be characterized by a conditional probability as follows: for $\forall A \in \mathcal{A}$, if the conditional probability $P_\theta(A|t)$ is independent of θ , then t is called a sufficient statistic of the parameter θ or t is called a **sufficient statistic** about the family of distributions \mathcal{P} . This implies that the measurement of the statistical space is independent of the parameter θ given t , *i.e.*, the statistic t carries all the information about the parameter θ .

Example 1.2.4. (Continuation of Example 1.2.1) From Eq. (1.2.19), when $\bar{x} = t$, the conditional pdf of \mathbf{x} is

$$f_\theta(\mathbf{x}|t) = \begin{cases} \frac{\left(\frac{1}{2\pi}\right)^{n/2} \exp\left\{-\frac{1}{2} \sum_{i=1}^n (x_i - \theta)^2\right\}}{\left(\frac{n}{2\pi}\right)^{1/2} \exp\left\{-\frac{n}{2}(t - \theta)^2\right\}}, & \text{if } \bar{x} = t, \\ 0, & \text{otherwise,} \end{cases}$$

$$= \begin{cases} \frac{1}{\sqrt{n}} \left(\frac{1}{2\pi}\right)^{(n-1)/2} \exp\left\{-\frac{1}{2} \sum_{i=1}^n (x_i - t)^2\right\}, & \text{if } \bar{x} = t \\ 0, & \text{otherwise,} \end{cases}$$

which is independent of the mean θ , so $t(\mathbf{x}) = \bar{x}$ is a sufficient statistic for the mean independent of parameter θ .

Example 1.2.5. Let the statistical space be $(\mathcal{X}, \mathcal{A}, \mathcal{P}) = (\mathcal{R}^n, \mathcal{B}^n, \mathcal{P})$, where \mathcal{B}^n is a σ -algebra generated by the Borel sets in the n -dimensional Euclidean space \mathcal{R}^n . For the sample $\mathbf{x} = (x_1, \dots, x_n)$, let $y_1 \leq y_2 \leq \dots \leq y_n$ be the result of arranging x_1, \dots, x_n from the smallest to the largest. Let $\mathbf{y} = (y_1, \dots, y_n)$, $\mathbf{t}(\mathbf{x}) = \mathbf{y}$ is called the **order statistics** of \mathbf{x} . Then the range of \mathbf{y} is

$$\mathcal{Y} = \{\mathbf{y}; y_1 \leq y_2 \leq \dots \leq y_n\}.$$

Obviously $\mathcal{Y} \subset \mathcal{B}^n$. Let \mathcal{B} be a σ -algebra generated by all the Borel sets in \mathcal{Y} , and \mathcal{A}_t be the induced σ -algebra by \mathbf{t} in \mathcal{A} . For $A \in \mathcal{A}_t$, A is a Borel set symmetrical relative to the n coordinates, *i.e.*, $\forall \mathbf{x} = (x_1, \dots, x_n) \in A$ implying $(x_{i_1}, \dots, x_{i_n}) \in A$, where (i_1, \dots, i_n) is a permutation of $(1, \dots, n)$.

For $P_\theta \in \mathcal{P}$, if the probability keeps unchanged for different permutation, *i.e.*, for $\forall A \in \mathcal{A}$, we have

$$P_\theta((X_1, \dots, X_n) \in A) = P_\theta((X_{i_1}, \dots, X_{i_n}) \in A),$$

then P_θ is called a **symmetrical distribution**. Obviously, when x_1, \dots, x_n are i.i.d. r.v.s, the joint distribution of $\mathbf{x} = (x_1, \dots, x_n)$ is a symmetrical distribution. We will explain that when all the distributions in \mathcal{P} are symmetrical, the order statistic t is sufficient for \mathcal{P} .

For $\mathbf{x} \in \mathcal{X}$ and $A \in \mathcal{A}$, let $\#(A, \mathbf{x})$ denote the number of permutations satisfying $(x_{i_1}, \dots, x_{i_n}) \in A$, which is a symmetrical function about \mathbf{x} -coordinate. Let $\mathbf{y} = \mathbf{t}(\mathbf{x})$, and then $\#(A, \mathbf{x}) = \#(A, \mathbf{y})$. For $\forall A_0 \in \mathcal{A}_t$ and $\forall P_\theta \in \mathcal{P}$, from the symmetry we have

$$P_\theta(A \cap A_0) = \int_{A_0} I_A(x_1, \dots, x_n) dP_\theta(\mathbf{x}) = \int_{A_0} I_A(x_{i_1}, \dots, x_{i_n}) dP_\theta(\mathbf{x}). \quad (1.2.25)$$

Note that $\#(A, \mathbf{y}) = \sum I_A(x_{i_1}, \dots, x_{i_n})$, where the summation runs over the $n!$ permutations. After taking summation on both sides of Eq. (1.2.25), we have

$$P_\theta(A \cap A_0) = \int_{A_0} \frac{\#(A, \mathbf{y})}{n!} dP_\theta.$$

Then $P_\theta(A|\mathbf{y}) = \#(A, \mathbf{y})/n!$ is independent of θ .

From the above two examples we can see that, though we can calculate a sufficient statistic directly according to its definition, its calculation is quite complicated. Now we offer a decision theorem, which is convenient to use and usually called as the **Neyman's Factorization Theorem**.

Theorem 1.2.4. *Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be a statistical space, where exists a σ -infinite measure λ , s.t. $\mathcal{P} \ll \lambda$. Let \mathcal{A}_t be a sub σ -algebra of \mathcal{A} induced by t . The t is a sufficient statistic for \mathcal{P} if and only if for $\forall \theta \in \Theta$, there exists an \mathcal{A}_t -measurable function g_θ , s.t.*

$$\frac{dP_\theta}{d\lambda} = g_\theta, \quad (1.2.26)$$

a.e. with respect to \mathcal{A} and λ , where λ is given by Eq. (1.1.7) in Theorem 1.1.3.

Proof. From Theorem 1.1.2, we can consider both P_θ and λ as probability measures on $(\mathcal{X}, \mathcal{A}_t)$. From Theorem 1.1.3, $P_\theta \ll \lambda$. Then there exists a Radon-Nikodym derivative g_θ on $(\mathcal{X}, \mathcal{A}_t)$. Obviously g_θ is \mathcal{A}_t -measurable, and for any integrable function h on $(\mathcal{X}, \mathcal{A}_t, P_\theta)$ we have

$$\int_{\mathcal{X}} h(x) dP_\theta(x) = \int_{\mathcal{X}} h(x) g_\theta(x) d\lambda(x). \quad (1.2.27)$$

Necessity. Let t be a sufficient statistic for \mathcal{P} , i.e. for $\forall A \in \mathcal{A}$ there exists a conditional probability $P(A|\mathcal{A}_t, x)$ independent of θ . Then we

will prove Eq. (1.2.26). From the definition of λ and Theorem 1.1.3, for $\forall A_0 \in \mathcal{A}_t$ we have

$$\begin{aligned}\lambda(A \cap A_0) &= \sum c_i P_{\theta_i}(A \cap A_0) \\ &= \sum c_i \int_{A_0} P(A|\mathcal{A}_t, x) dP_{\theta_i}(x) \\ &= \int_{A_0} P(A|\mathcal{A}_t, x) d\lambda(x).\end{aligned}$$

This implies

$$\lambda(A|\mathcal{A}_t, x) = P(A|\mathcal{A}_t, x) \quad (1.2.28)$$

a.e. with respect to \mathcal{A}_t , and λ . We use E_λ to denote the mean when the probability measure is λ , and then for $\forall \theta \in \Theta$, and $\forall A \in \mathcal{A}$ we have

$$\begin{aligned}P_\theta(A) &= \int_{\mathcal{X}} P(A|\mathcal{A}_t, x) dP_\theta(x) \\ &= \int_{\mathcal{X}} \lambda(A|\mathcal{A}_t, x) g_\theta(x) d\lambda(x) \\ &= \int_{\mathcal{X}} E_\lambda(I_A(X)|\mathcal{A}_t, x) g_\theta(x) d\lambda(x) \\ &= \int_{\mathcal{X}} E_\lambda(I_A(X) g_\theta(X)|\mathcal{A}_t, x) d\lambda(x) \\ &= \int_{\mathcal{X}} I_A(x) g_\theta(x) d\lambda(x) \\ &= \int_A g_\theta(x) d\lambda(x),\end{aligned}$$

where the second equality follows from Eqs. (1.2.27) and (1.2.28); the fourth follows from Theorem 1.2.2; and the fifth follows from Theorem 1.2.1. Since $A \in \mathcal{A}$ is arbitrary, Eq. (1.2.26) holds.

Sufficiency. Suppose Eq. (1.2.26) holds, we will prove that for $\forall \theta \in \Theta$ and $A \in \mathcal{A}$, we have

$$P_\theta(A|\mathcal{A}_t, x) = \lambda(A|\mathcal{A}_t, x) \quad (1.2.29)$$

a.e. with respect to \mathcal{A}_t , and λ . For $\forall A_0 \in \mathcal{A}_t$, we have

$$\begin{aligned} \int_{A_0} \lambda(A|\mathcal{A}_t, x) dP_\theta(x) &= \int_{A_0} \lambda(A|\mathcal{A}_t, x) g_\theta(x) d\lambda(x) \\ &= \int_{A_0} E_\lambda(I_A(X)|\mathcal{A}_t, x) g_\theta(x) d\lambda(x) \\ &= \int_{A_0} E_\lambda(I_A(X) g_\theta(X)|\mathcal{A}_t, x) d\lambda(x) \\ &= \int_{A_0} I_A(x) g_\theta(x) d\lambda(x) \\ &= \int_{A \cap A_0} g_\theta(x) d\lambda(x) \\ &= P_\theta(A \cap A_0). \end{aligned}$$

Since $A_0 \in \mathcal{A}_t$ is arbitrary, from the definition of conditional probability, Eq. (1.2.29) holds. \square

Theorem 1.2.5. (Neyman's Factorization Theorem) Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be a statistical space, and ν be a σ -finite measure satisfying $\mathcal{P} \ll \nu$. Let \mathcal{A}_t be a sub σ -algebra of \mathcal{A} induced by t . The t is a sufficient statistic for \mathcal{P} if and only if for $\forall \theta \in \Theta$ there exists an \mathcal{A}_t measurable function g_θ and \mathcal{A} -measurable function h independent of θ , s.t.

$$\frac{dP_\theta(x)}{d\nu(x)} = g_\theta(x)h(x) \quad (1.2.30)$$

a.e. with respect to both \mathcal{A} and ν .

Proof. Let λ be the probability measure in Theorem 1.2.3, we will prove that Eq. (1.2.30) is equivalent to Eq. (1.2.26). From Theorem 1.1.3, \mathcal{P} and λ are equivalent, so $\lambda \ll \nu$. Let

$$\frac{d\lambda(x)}{d\nu(x)} = h(x),$$

then h is an \mathcal{A} -measurable function independent of θ . Hence Eq. (1.2.26) implies Eq. (1.2.30). On the contrary, from Eq. (1.2.30)

$$\frac{d\lambda(x)}{d\nu(x)} = \sum_{i=1}^{\infty} c_i \frac{dP_{\theta_i}(x)}{d\nu(x)} = \sum_{i=1}^{\infty} c_i g_{\theta_i}(x) h(x).$$

Let $s(x) = \sum_{i=1}^{\infty} c_i g_{\theta_i}(x)$, for $\forall \theta \in \Theta$ and $x \in \mathcal{X}$, let

$$g_\theta^*(x) = \begin{cases} \frac{g_\theta(x)}{s(x)}, & \text{when } 0 < s(x) < \infty \\ 0, & \text{otherwise.} \end{cases}$$

And then g_θ^* is an \mathcal{A}_t -measurable function. Since λ is a probability measure, we have $d\lambda(x)/d\nu(x) = s(x)h(x) > 0$ a.e. with respect to both \mathcal{A} and λ , thus

$$\begin{aligned} \frac{dP_\theta(x)}{d\lambda(x)} &= \frac{dP_\theta(x)}{d\nu(x)} \cdot \frac{d\nu(x)}{d\lambda(x)} \\ &= g_\theta(x)h(x) \cdot \frac{1}{s(x)h(x)} \\ &= g_\theta^*(x). \end{aligned}$$

The conclusion follows from Theorem 1.2.3. □

The above theorem shows that a statistic $t(x)$ from the statistical space $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ to the statistical space $(\mathcal{Y}, \mathcal{B}, \mathcal{Q})$ is sufficient for θ if there exists a \mathcal{B} -measurable function g_θ and an \mathcal{A} -measurable function h which is independent of θ , s.t.

$$f_\theta(x) = g_\theta(t(x))h(x), \tag{1.2.31}$$

where $f_\theta = dP_\theta/d\nu$ for $\forall \theta \in \Theta$. The example below shows that this theorem provides a convenient decision criterion.

Example 1.2.6. (Continuation of Examples 1.2.1 and 1.2.4) Let x_1, \dots, x_n be i.i.d. r.v.s from $N(\theta, 1)$, and $\mathbf{x} = (x_1, \dots, x_n)$. Let $t(\mathbf{x}) = \bar{x}$. And then the joint distribution can be factorized as

$$\begin{aligned} f_\theta(\mathbf{x}) &= \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2} \sum_{i=1}^n (x_i - \theta)^2\right\} \\ &= \exp\left\{-\frac{n}{2}(t - \theta)^2\right\} \cdot \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2} \sum_{i=1}^n (x_i - t)^2\right\}. \end{aligned}$$

From (1.2.31), t is a sufficient statistic for θ . However, both t_1 and t_{\max} mentioned at the beginning of this section are not sufficient for θ .

If x_1, \dots, x_n are i.i.d. r.v.s from $N(\mu, \sigma^2)$ with the mean μ known, then from the joint distribution, $S_\mu^2 = \sum_{i=1}^n (x_i - \mu)^2$ is a sufficient statistic for σ^2 . When μ is unknown, $\theta = (\mu, \sigma^2)$, the joint distribution can be factorized as

$$\begin{aligned} f_\theta(\mathbf{x}) &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right\} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2} \left[\sum_{i=1}^n x_i^2 - 2\mu \sum_{i=1}^n x_i + n\mu^2\right]\right\}. \end{aligned}$$

From Eq. (1.2.31), $(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2)$ is a sufficient statistic for θ . Let $\mathbf{t}(\mathbf{x}) = (\bar{x}, S_x^2)$. Since the one-to-one mapping of a sufficient statistic is still sufficient, \mathbf{t} is also a sufficient statistic for θ .

Example 1.2.7. (Continuation of Example 1.2.5) Let x_1, \dots, x_n be a set of samples from $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, where for $\forall \theta \in \Theta$, P_θ is a symmetric distribution, and f_θ denotes the pdf. Let $\mathbf{t}(\mathbf{x}) = \mathbf{y}$ be the order statistics of \mathbf{x} , then

$$f_\theta(\mathbf{x}) = g_\theta(\mathbf{y}) \cdot \frac{1}{n!},$$

where $g_\theta(\mathbf{y})$ is the pdf of \mathbf{t} , *i.e.*

$$g_\theta(\mathbf{y}) = \begin{cases} n! f_\theta(y_1, \dots, y_n), & \text{if } y_1 \leq \dots \leq y_n, \\ 0, & \text{otherwise.} \end{cases}$$

From Eq. (1.2.31) we know that \mathbf{t} is a sufficient statistic for θ . It can be seen that the decision method is easier to use than that of Example 1.2.5.

Example 1.2.8. Let x_1, \dots, x_n be a set of samples from uniform distribution $U(0, \theta)$, then the joint distribution is

$$f_\theta(\mathbf{x}) = \begin{cases} \frac{1}{\theta^n}, & \text{if } \max\{x_i\} \leq \theta, \\ 0, & \text{if otherwise.} \end{cases}$$

Let $x_{(n)} = \max\{x_i\}$, then the above equation can be written as

$$f_\theta(\mathbf{x}) = \frac{1}{\theta^n} I_{[0, \theta]}(x_{(n)}).$$

From Eq. (1.2.31), $x_{(n)}$ is a sufficient statistic for θ .

Minimal sufficient statistics. From Eq. (1.2.31), for a given statistical space, we can construct many sufficient statistics. Consider the problem discussed in the Example 1.2.1, Example 1.2.4 and Example 1.2.6. Let x_1, \dots, x_n be i.i.d. from $N(\theta, 1)$, and $\mathbf{x} = (x_1, \dots, x_n)$. Then the sample space is $(\mathcal{R}^n, \mathcal{B}^n)$, where \mathcal{B}^n denotes σ -algebra generated by the Borel sets of the n -dimensional Euclidean space \mathcal{R}^n . By Example 1.2.4 and Example 1.2.6, we know that $t(\mathbf{x}) = \bar{x}$ is a sufficient statistic. The complete data set itself, $t_1(\mathbf{x}) = \mathbf{x}$ is obviously a sufficient statistic, which is called a **trivial sufficient statistic** since it does not compress the data set. In fact, let

$$t_i(\mathbf{x}) = \left(\frac{1}{i} (x_1 + \dots + x_i), x_{i+1}, \dots, x_n \right),$$

where $i = 1, \dots, n$. Then from Eq. (1.2.31), $\mathbf{t}_i(\mathbf{x})$ for $i = 1, \dots, n$ are all sufficient statistics for θ . Which one is the best? What is the decision criterion?

Generally speaking, in all the sufficient statistics, the stronger the function of data compression a sufficient statistic has, the better it is, since its calculation is more convenient. Let \mathcal{A}_i be a σ -algebra generated by \mathbf{t}_i in \mathcal{B}^n , then essentially the \mathcal{A}_i is a σ -algebra generated by the Borel sets in $(n - i + 1)$ -dimensional Euclidean space, obviously we have

$$\mathcal{A}_1 \supset \mathcal{A}_2 \supset \dots \supset \mathcal{A}_n.$$

As far as the function of data compression is concerned, \mathbf{t}_{i+1} is superior to \mathbf{t}_i . Especially \mathbf{t}_{i+1} can be obtained by a function of \mathbf{t}_i , say let $\mathbf{t}_i(\mathbf{x}) = (y_1, \dots, y_{n-i+1})$, and then

$$\mathbf{t}_{i+1}(\mathbf{x}) = \left(\frac{1}{i+1}(iy_1 + y_2), y_3, \dots, y_{n-i+1} \right),$$

has $n - i$ components. The property can be abstracted as follows: let t be a sufficient statistic on $(\mathcal{A}, \mathcal{X}, \mathcal{P})$, if for any other sufficient statistic s , there exists a measurable mapping h , s.t.

$$t(x) = h(s(x)), \quad (1.2.32)$$

then t is called a **minimal sufficient statistic**. Roughly speaking, the minimal sufficient statistic is a sufficient statistic that cannot be compressed any more without losing information about the unknown parameter.

Bahadur (1954) proved that, for a given statistical space $(\mathcal{A}, \mathcal{X}, \mathcal{P})$, there exists a minimal sufficient statistic in \mathcal{P} if the parameter space Θ is divisible for the quasi-distance defined in Example 1.1.13. However, the problem how to decide whether a statistic is sufficient or not has not been solved. This problem will be discussed in the next section.

1.3 Exponential Family and Completeness

Many probability density functions can be written in a uniform form, which have the same properties. A collection of the functions forms a so-called exponential family. Before discussing it in detail, we will review the Laplace transform first. Let $(\mathcal{X}, \mathcal{A}, \nu)$ be a measure space, t_i and h be \mathcal{A} -measurable functions, where $i = 1, \dots, k$, if

$$f(u) = \int_{\mathcal{X}} h(x) \exp \left\{ \sum_{i=1}^k u_i t_i(x) \right\} d\nu(x) \quad (1.3.1)$$

is finite, (1.3.1) is called a **generalized Laplace transform** (cf. Widder, 1946).

Let $\mathcal{U} = \{u; f(u) < \infty\}$, from the property that the exponential function is convex we know \mathcal{U} is a convex set, i.e. for $\forall \alpha \in [0, 1]$ and $u, v \in \mathcal{U}$, we have

$$\left| h(x) \exp \left\{ \sum_{i=1}^k (\alpha u_i + (1 - \alpha) v_i) t_i(x) \right\} \right| \leq \alpha \left| h(x) \exp \left\{ \sum_{i=1}^k u_i t_i(x) \right\} \right| + (1 - \alpha) \left| h(x) \exp \left\{ \sum_{i=1}^k v_i t_i(x) \right\} \right|.$$

The integrability of the right side determines the integrability of the left. By the properties of the Laplace transform, we can get the following theorem.

Theorem 1.3.1. *The f is continuous at interior points in \mathcal{U} , and all its any-order partial derivatives with respect to u_i exist. Furthermore, the order of differentiation and integration can be interchanged, i.e.*

$$\frac{\partial f(u)}{\partial u_i} = \int_{\mathcal{X}} h(x) t_i(x) \exp \left\{ \sum_{i=1}^k u_i t_i(x) \right\} d\nu(x).$$

1.3.1 Exponential Family

Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be a statistical space, for a σ -finite measure ν on $(\mathcal{X}, \mathcal{A})$ we have $\mathcal{P} \ll \nu$. Let u, u_i be two functions in the parameter space Θ , $i = 1, \dots, k$, where $u(\theta) > 0$ holds for $\forall \theta \in \Theta$. If for $\forall \theta \in \Theta$,

$$\frac{dP_\theta}{d\nu} = u(\theta) \exp \left\{ \sum_{i=1}^k u_i(\theta) t_i(x) \right\} h(x), \tag{1.3.2}$$

and the support set $\{x; dP_\theta/d\nu > 0\}$ is independent of θ , then \mathcal{P} is called an **exponential family**. If $\Theta \subset R^k$, and for $\forall \theta \in \Theta$, we have $u_i(\theta) = \theta_i, i = 1, \dots, k$, then Eq. (1.3.2) can be written as

$$\frac{dP_\theta}{d\nu} = u(\theta) \exp \left\{ \sum_{i=1}^k \theta_i t_i(x) \right\} h(x), \tag{1.3.3}$$

\mathcal{P} is called a **natural exponential family**.

Example 1.3.1. Consider the normal distribution $N(\mu, \sigma^2)$, where $\theta = (\mu, \sigma^2)$. The probability density is

$$\begin{aligned} f_\theta(x) &= \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\} \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{\mu^2}{2\sigma^2} \right\} \exp \left\{ -\frac{1}{2\sigma^2} x^2 + \frac{\mu}{\sigma^2} x \right\}. \end{aligned} \tag{1.3.4}$$

$\mathcal{P} = \{P_\theta; \theta \in \mathbf{R} \times \mathbf{R}_+\}$ belongs to the exponential family with $t_1(x) = x, t_2(x) = x^2$. If let $\theta_1 = \mu/\sigma^2, \theta_2 = -1/(2\sigma^2)$, then (1.3.4) has the form (1.3.3), *i.e.*, the form of the natural exponential family. Solving inversely we have $\mu = -\theta_2/(2\theta_1), \sigma^2 = -1/(2\theta_1)$.

Example 1.3.2. Uniform distribution $U(0, \theta)$ does not belong to the exponential family, since its support set depends on θ . The density function of two-parameter exponential distribution is

$$f_\theta(x) = \begin{cases} \frac{1}{\sigma} \exp\left\{-\frac{1}{\sigma}(x - \mu)\right\}, & \text{if } x \geq \mu. \\ 0, & \text{otherwise.} \end{cases}$$

Obviously, it does not belong to the exponential family either, since its support set depends on θ , where $\theta = (\mu, \sigma)$. But when the parameter μ in the exponential distribution is known, then the above distribution becomes a one-parameter exponential distribution and belongs to the exponential family, with $t(x) = x - \mu$.

It is easy to verify that some well-known probability distributions (such as binomial, Poisson, multinomial, and multidimensional normal) all belong to the exponential family.

Since the natural exponential family is a probability measure, from Eq. (1.3.3) we have

$$\frac{1}{u(\theta)} = \int_{\mathcal{X}} h(x) \exp\left\{\sum_{i=1}^k \theta_i t_i(x)\right\} d\nu(x). \tag{1.3.5}$$

The set $S = \{\theta; 1/u(\theta) < \infty\}$ is called a **natural parameter space**.

Theorem 1.3.2. *Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be a statistical space. If \mathcal{P} belongs to the natural exponential family, then*

- (i) *The natural parameter space is a convex set.*
- (ii) *$u(\theta)$ is continuous at the interior points in S and its all partial derivatives with respect to θ_i exist.*
- (iii) *For the interior points in S we have*

$$E_\theta t_i(X) = -\frac{\partial}{\partial \theta_i} \ln u(\theta), \quad i = 1, \dots, k;$$

$$V_\theta t_i(X) = -\frac{\partial^2}{\partial \theta_i^2} \ln u(\theta), \quad i = 1, \dots, k;$$

$$CV_\theta(t_i(X), t_j(X)) = -\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln u(\theta), \quad i, j = 1, \dots, k; \quad i \neq j.$$

Proof. From Eq. (1.3.1), we know that Eq. (1.3.5) is a generalized Laplace transform. From Theorem 1.3.1, we know both (i) and (ii) hold. Now we prove (iii). Taking first partial derivative on both sides of Eq. (1.3.5), by Theorem 1.3.1, we have

$$-\frac{1}{u^2(\theta)} \cdot \frac{\partial u(\theta)}{\partial \theta_i} = \int_{\mathcal{X}} h(x) t_i(x) \exp \left\{ \sum_{i=1}^k \theta_i t_i(x) \right\} d\nu(x).$$

Thus we can get $E_{\theta} t_i(X)$. Similarly, taking second partial derivative yields the other two equalities. \square

Example 1.3.3. Consider a binomial distribution $X \sim Bi(n, \theta)$. The density function with respect to the Counting measure is

$$\begin{aligned} f_{\theta}(x) = P_{\theta}(X = x) &= \binom{n}{x} \theta^x (1 - \theta)^{n-x} \\ &= (1 - \theta)^n \exp \left[x \ln \frac{\theta}{1 - \theta} \right] \binom{n}{x}. \end{aligned} \quad (1.3.6)$$

This is a form of the exponential family, furthermore, let

$$\omega = \ln \frac{\theta}{1 - \theta}, \quad \text{and} \quad \theta = \frac{e^{\omega}}{1 + e^{\omega}}, \quad (1.3.7)$$

Eq. (1.3.6) becomes

$$f_{\omega}(x) = (1 + e^{\omega})^{-n} e^{\omega x} \binom{n}{x}.$$

This is a form of natural exponential family with $t(x) = x$ and $u(w) = (1 + e^w)^{-n}$. Applying Theorem 1.3.2, we have

$$E_w X = -\frac{\partial}{\partial w} u(w) = n \frac{e^w}{1 + e^w}.$$

By Eq. (1.3.7) and Theorem 1.2.1, it can be transformed into $E_{\theta} X = n\theta$. This coincides with Example 1.1.1.

1.3.2 Completeness

For a given statistical space $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, let $(\mathcal{Y}, \mathcal{B}, \mathcal{Q})$ be the induced statistical space by the statistic t . If t is a sufficient statistic for θ , from the discussion above, it suffices to consider measurable mappings and measurable functions based on t when we make inferences about θ . But now there are two problems that need to be considered further. The first one, if there are two integrable functions h_1 and h_2 , s.t. $E_{\theta} h_1(t(X)) = E_{\theta} h_2(t(X))$ for

$\forall \theta \in \Theta$, whether does $h_1(t) = h_2(t)$ holds almost everywhere? The second one, how to judge whether t is the minimal sufficient statistic or not? Both problems involve the concept of completeness.

For a \mathcal{B} -measurable function g , if

$$E_{\theta}g(t(X)) = 0 \text{ for } \forall \theta \in \Theta,$$

then we must have

$$g(t(x)) = 0$$

a.e., then t is said to be a **complete statistic**. Let \mathcal{A}_t be the induced sub σ -algebra of \mathcal{A} by the statistic t , and then \mathcal{A}_t is said to be **complete**.

Example 1.3.4. A Bernoulli experiment means that there are only two outcomes, success and failure. Let the probability of success be $\theta, 0 < \theta < 1$, and then the probability of failure is $1 - \theta$. If 1 denotes success, and 0 denotes failure, *i.e.* $P_{\theta}(X = 1) = \theta, P_{\theta}(X = 0) = 1 - \theta$. Repeat Bernoulli experiment for n times, then $t = x_1 + \dots + x_n$ is a sufficient statistic for θ , and from the binomial distribution $Bi(n, \theta)$ (cf. Examples 1.1.1 and 1.3.3). If $g(t)$ satisfies $E_{\theta}g(t(X)) = 0$ for $\forall \theta \in (0, 1)$, or equivalently,

$$\sum_{t=0}^n g(t) \binom{n}{t} w^t = 0, \quad \forall w \in (0, \infty)$$

where $w = (1 - \theta)^{-1}\theta$. Since the left side of the equation is a polynomial of w , then the coefficients of all the polynomials must be zero, so $g(t) = 0, t = 0, 1, \dots, n$. Since the statistic t is complete, t is then called a **complete sufficient statistic**.

Example 1.3.5. We have discussed the sufficiency of order statistics in Example 1.2.5, now we will discuss their completeness. Let x_1, \dots, x_n be an i.i.d. sample from $(\mathcal{R}, \mathcal{B}, \mathcal{P})$, where \mathcal{B} is a σ -algebra generated by the Borel sets in Euclidean space, and \mathcal{P} is a family of all continuous distributions, *i.e.*, $\mathcal{P} \ll \nu$, where ν is the Lebesgue measure. For $P \in \mathcal{P}$, let $dP/d\nu = f$, and then the statistical space is $(\mathcal{R}^n, \mathcal{B}^n, \mathcal{P}^n)$, and we have $P^n \in \mathcal{P}^n$,

$$\frac{dP^n}{d\nu^n} = \prod_{i=1}^n f(x_i). \tag{1.3.8}$$

Let $t = (y_1, \dots, y_n)$ be the order statistics of x_1, \dots, x_n , and \mathcal{A}_t be a sub σ -algebra of \mathcal{B}^n induced by t . Obviously the measure given by (1.3.8) is symmetrical, and all \mathcal{A}_t -measurable functions are also symmetrical about

the coordinate axes. To verify the completeness of the order statistic t , we need to prove that, for an \mathcal{A}_t -measurable function g , if for $\forall P \in \mathcal{P}$ we have

$$\int_{\mathcal{R}^n} g(x_1, \dots, x_n) \prod_{i=1}^n f(x_i) d\nu(x_i) = 0, \quad (1.3.9)$$

then for $\forall B \in \mathcal{A}_t$, we must have

$$\int_B g(x_1, \dots, x_n) \prod_{i=1}^n d\nu(x_i) = 0. \quad (1.3.10)$$

In fact, we only need to prove the situation of $B = B_1 \times \dots \times B_n$, where $B_j \in \mathcal{B}$ satisfies $\nu(B_j) > 0, j = 1, \dots, n$. Define

$$f_j(x) = \frac{I_{B_j}(x)}{\nu(B_j)}.$$

This is also a continuous distribution. Let $f(x) = \sum \alpha_j f_j(x)$, where $\alpha_j > 0$ and $\sum \alpha_j = 1$. From (1.3.9) and the symmetry of g , we have

$$\begin{aligned} & \int_B g(x_1, \dots, x_n) \prod_{i=1}^n f(x_i) d\nu(x_i) \\ &= \int_{\mathcal{R}^n} g(x_1, \dots, x_n) \prod_{i=1}^n \left[\sum_{j=1}^n \alpha_j f_j(x_i) \right] d\nu(x_i) = 0. \end{aligned}$$

This is a homogeneous polynomial of α_j , and from the arbitrariness of α_j and the symmetry of g , we have

$$\int_{\mathcal{R}^n} g(x_1, \dots, x_n) \prod_{i=1}^n f_i(x_i) d\nu(x_i) = 0$$

i.e. Eq. (1.3.10) holds.

Example 1.3.6. Let x_1, \dots, x_n be a set of samples drawn from the uniform distribution $U(0, \theta)$, from Example 1.2.8, $t(x) = \max\{x_i\}$ is a sufficient statistic for θ . We will prove t is also a complete statistic. From Example 1.2.8, we know the density function of t is

$$h_\theta(t) = \begin{cases} \frac{1}{\theta^n} n t^{n-1}, & \text{if } 0 < t < \theta, \\ 0, & \text{otherwise.} \end{cases}$$

Let $g(t)$ be a measurable function satisfying $E_\theta(g(T)) = 0$ for any $\theta > 0$. Taking derivatives on both sides of the equation:

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta} \int_0^\theta g(t) \frac{1}{\theta^n} n t^{n-1} dt \\ &= \frac{1}{\theta^n} \frac{\partial}{\partial \theta} \int_0^\theta g(t) n t^{n-1} dt + \left(\frac{\partial}{\partial \theta} \left(\frac{1}{\theta^n} \right) \right) \int_0^\theta g(t) n t^{n-1} dt \\ &= \frac{\theta^{n-1}}{\theta^n} n g(\theta) + 0 \\ &= \frac{1}{\theta} n g(\theta). \end{aligned}$$

Since $n/\theta \neq 0$, $g(\theta) = 0$. Thus $t(x) = \max\{x_i\}$ is a complete sufficient statistic.

Now we will discuss the relation between the completeness and the minimal sufficient statistic.

Theorem 1.3.3. *Let t be a statistic in the statistical space $(\mathcal{X}, \mathcal{A}, \mathcal{P})$. If the sufficient statistic t is complete, then it must be a minimal sufficient statistic.*

Proof. Let \mathcal{A}_t be a sub σ -algebra of \mathcal{A} induced by t , and \mathcal{A}_1 be any sufficient region. From the definition of minimal sufficient statistic in (1.2.32), we need to prove for $\forall A_t \in \mathcal{A}_t$, there exists $A_1 \in \mathcal{A}_1$, s.t. we must have $P_\theta(A_t \triangle A_1) = 0$ for $\forall \theta \in \Theta$, where the definition has been given in Example 1.1.12.

Since both \mathcal{A}_t and \mathcal{A}_1 are sufficient regions, then there exist the following conditional means independent of θ :

$$\begin{aligned} f(x) &= E(I_{A_t} | \mathcal{A}_1, x) \\ g(x) &= E(f(X) | \mathcal{A}_t, x), \end{aligned}$$

where $f(x)$ and $g(x)$ are \mathcal{A}_1 and \mathcal{A}_t -measurable functions with the rang of $[0,1]$, respectively. From Theorem 1.2.1, for $\forall \theta \in \Theta$

$$E_\theta I_{A_t}(X) = E_\theta f(X) = E_\theta g(X). \tag{1.3.11}$$

Since t is a complete statistic, from $E_\theta [I_{A_t}(X) - g(X)] = 0$, we have

$$I_{A_t}(x) = g(x)$$

a.e. From Theorem 1.2.2, we have

$$I_{A_t}(x) = I_{A_t}(x) \cdot I_{A_t}(x) = I_{A_t}(x)g(x) = E(I_{A_t}(X)f(X) | \mathcal{A}_t, x)$$

almost everywhere. From Theorem 1.2.1 again, we have

$$E_{\theta}I_{A_t}(X) = E_{\theta}I_{A_t}(X)f(X) \tag{1.3.12}$$

for $\forall \theta \in \Theta$. Combining (1.3.11) with (1.3.12), we have

$$E_{\theta}I_{A_t}(X)(1 - f(X)) = E_{\theta}f(X)(1 - I_{A_t}(X)) = 0.$$

Since the above integrands are nonnegative, then we have

$$I_{A_t}(x)(1 - f(x)) = f(x)(1 - I_{A_t}(x))$$

a.e., i.e. we have $I_{A_t}(x) = f(x)$ a.e. It suffices to let

$$A_1 = \{x; f(x) = 1\}$$

then we have $P_{\theta}(A_t \triangle A_1) = 0$ for $\forall \theta \in \Theta$. □

1.3.3 Sufficiency and Completeness

From Theorem 1.3.3, we can easily judge the minimal sufficiency of a statistic, which is favorable to make statistical inferences. Now we analyze the exponential family.

Let x_1, \dots, x_n be i.i.d. from a natural exponential family with the pdf form as (1.3.3). The statistical space is $(\mathcal{X}^n, \mathcal{A}^n, \mathcal{P}^n)$, where $\mathcal{P}^n \ll \nu^n$, and for $P_{\theta}^n \in \mathcal{P}^n$ we have

$$\frac{dP_{\theta}^n}{d\nu^n} = c^n(\theta) \exp \left\{ \sum_{i=1}^k \theta_i \sum_{j=1}^n t_i(x_j) \right\} \prod_{j=1}^n h(x_j). \tag{1.3.13}$$

Theorem 1.3.4. *In Eq. (1.3.13), let*

$$\mathbf{t} = \left(\sum_{j=1}^n t_1(x_j), \dots, \sum_{j=1}^n t_k(x_j) \right), \tag{1.3.14}$$

then the statistic \mathbf{t} is sufficient for θ , and is complete with respect to \mathcal{P}^n .

Proof. From the Neyman’s Factorization Theorem, the statistic \mathbf{t} given by (1.3.14) is sufficient. Considering that an \mathcal{A}^n -measurable function g satisfying for $\forall \theta \in \Theta$, we have

$$\int_{\mathcal{X}^n} g(\mathbf{x})c^n(\theta) \exp \left\{ \sum_{i=1}^k \theta_i t_i^*(\mathbf{x}) \right\} h^*(\mathbf{x})d\nu^n(\mathbf{x}) = 0,$$

where $\mathbf{x} = (x_1, \dots, x_n)$, $t_i^*(\mathbf{x}) = \sum_{j=1}^n t_i(x_j)$, and $h^*(\mathbf{x}) = \prod_{j=1}^n h(x_j)$. Compared with Eq. (1.3.1), the equation above is also a generalized Laplace transform, then we have $g(\mathbf{x}) = 0$ a.e. Thus \mathbf{t} is a complete statistic. □

Applying Theorems 1.3.4 and 1.3.3, it is easy to get the minimal statistics of a series of probability distributions.

1.3.4 Ancillary Statistics

We have known that sufficient statistics are quite important, since they include all the information of the parameters. On the contrary, there exist some statistics with distributions independent of parameters, called ancillary statistics. Let $s(x)$ be a statistic from $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ to $(\mathcal{S}, \mathcal{B}, \mathcal{Q})$, where $\mathcal{P} = \{P_\theta; \theta \in \Theta\}$, and $\mathcal{Q} = \{Q_\theta; \theta \in \Theta\}$. From the construction of statistical space, we know that $Q_\theta(B) = P_\theta(s^{-1}(B))$ for $\forall B \in \mathcal{B}, \forall \theta \in \Theta$. If for any θ and $B \in \mathcal{B}$, $Q_\theta(B)$ is independent of θ , then $s(x)$ is called an **ancillary statistic**. Although an ancillary statistic contains no information about the parameter, it has at least two advantages, one is its invariance to the parameter, and the other is its independence from sufficient statistics.

Example 1.3.7. (Ancillary statistics of location parameter) For given probability measure P_θ , its corresponding **Cumulative Distribution Function** is defined as $F_\theta(x) = P_\theta(X \leq x)$. If $\mathcal{P} \ll \nu$, and $dP_\theta/d\nu = f_\theta$, then

$$F_\theta(x) = \int_{-\infty}^x f_\theta(y) d\nu(y). \tag{1.3.15}$$

Let $F(x)$ be a cdf, if for $\forall \theta \in \Theta$ we have $F_\theta(x) = F(x - \theta)$, then \mathcal{P} is called a **location distribution family** and θ a **location parameter**.

Let x_1, \dots, x_n be a set of samples from $F(x - \theta)$, we will prove that $s(x) = \max\{x_i\} - \min\{x_i\}$ is an ancillary statistic. Let z_1, \dots, z_n denote a set of samples from $F(x)$, then $x_1 = z_1 + \theta, \dots, x_n = z_n + \theta$. For any s we have

$$\begin{aligned} P_\theta(S(X) \leq s) &= P_\theta(\max\{X_i\} - \min\{X_i\} \leq s) \\ &= P_\theta(\max\{Z_i + \theta\} - \min\{Z_i + \theta\} \leq s) \\ &= P_\theta(\max\{Z_i\} - \min\{Z_i\} \leq s). \end{aligned}$$

Obviously the probability distribution does not depend on the parameter θ . From the arbitrariness of s , we know that $s(x)$ is an ancillary statistic. Usually, $s(x)$ is called the **sample range**.

Example 1.3.8. (Ancillary statistics of scale parameter) \mathcal{P} with cdf in the form of $F(x/\sigma)$ is called a **scale distribution family**, and σ a **scale parameter**, where $\sigma > 0$. Let x_1, \dots, x_n be a set of samples, we will prove that $\mathbf{s}(x) = (x_1/x_n, \dots, x_{n-1}/x_n)$ is an ancillary statistic.

Let z_1, \dots, z_n denote a set of samples from $F(x)$, then we have $x_1 = \sigma z_1, \dots, x_n = \sigma z_n$. For any $\mathbf{s} = (s_1, \dots, s_{n-1})$ we have

$$\begin{aligned} P_\sigma(\mathbf{S}(\mathbf{X}) \leq \mathbf{s}) &= P_\sigma(X_1/X_n \leq s_1, \dots, X_{n-1}/X_n \leq s_{n-1}) \\ &= P_\sigma(Z_1/Z_n \leq s_1, \dots, Z_{n-1}/Z_n \leq s_{n-1}), \end{aligned}$$

which is independent of σ .

The theorem below shows the relationship between ancillary statistics and complete sufficient statistics.

Theorem 1.3.5. *Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be a statistical space, where $\mathcal{P} = \{P_\theta; \theta \in \Theta\}$. Let $T(x)$ be a complete sufficient statistic, and $S(x)$ be an ancillary statistic, then $T(X)$ and $S(X)$ are mutually independent (denoted as $T(X) \perp S(X)$).*

Proof. Let \mathcal{A}_t and \mathcal{A}_s be sub σ -algebras of \mathcal{A} induced by $T(x)$ and $S(x)$, respectively. We need to prove that for any $\theta \in \Theta$, $A_t \in \mathcal{A}_t$, and $A_s \in \mathcal{A}_s$ we have

$$P_\theta(A_t \cap A_s) = P_\theta(A_t)P_\theta(A_s). \quad (1.3.16)$$

From the condition, $P_\theta(A_s)$ is independent of θ , let $P_\theta(A_s) = a$. Let

$$f(x) = P(A_s | \mathcal{A}_t, x) = E(I_{A_s}(X) | \mathcal{A}_t, x).$$

Since $T(x)$ is a sufficient statistic, and the above conditional probability is independent of the parameter θ . From Theorem 1.2.1, for $\forall \theta \in \Theta$,

$$E_\theta f(X) = E_\theta E(I_{A_s}(X) | \mathcal{A}_t, X) = E_\theta(I_{A_s}(X)) = a.$$

In other words, we have $E(f(X) - a) = 0$ for $\forall \theta \in \Theta$. From the completeness of \mathcal{A}_t , we have $f(x) = a$ a.e. with respect to \mathcal{A}_t and \mathcal{P} . Then from (1.2.4)

$$P_\theta(A_t \cap A_s) = \int_{A_t} P(A_s | \mathcal{A}_t, x) dP_\theta(x) = aP_\theta(A_t) = P_\theta(A_s)P_\theta(A_t). \quad \square$$

Theorem 1.3.5 is usually called as Basu's Lemma. For further discussion about ancillary statistics and Basu's Lemma, see Basu (1958, 1959), Koehn and Thomas (1975), and Lehmann (1980, 1986). In some situations, it is convenient to prove the independence by using Basu's Lemma, here is an example.

Example 1.3.9. (Independence of the normal sample mean, sample range and sample variance) Let x_1, \dots, x_n be a set of samples from $N(\mu, \sigma^2)$. Let \bar{X} , L and S^2 respectively denote the sample mean, the sample range and the sample variance. For any given σ_0^2 , \bar{X} is a complete sufficient statistic for μ , and both L and S^2 are ancillary statistics for μ , by Basu's Lemma, \bar{X} , L , and S^2 are mutually independent. By the arbitrariness, we know the result holds for any σ^2 .

1.4 Estimation Methods Based on Statistical Space

Although we mainly focus on the study of theory and methods of parameter tests, we will review some estimation methods at first, since the methods of tests are usually based on the estimation of parameter. Furthermore, we will analyze some basic properties of estimation procedures.

Essentially, estimation methods can be classified into two categories: one is that the form of distribution is partly or completely unknown, thus we can only estimate some typical characteristics of the distributions (such as the mean, the variance, and the median), or estimate the cdf itself and so on; the other is that the form of distribution is known, thus we can estimate the parameters in the distribution. Generally speaking, there is no advantage without disadvantage. In principle, we may grasp more information in the second category, and thus we can obtain more accurate estimates. However, if the estimation method depends on the distribution too heavily, then the estimate may behave badly once the information about the the distribution is unreliable (*i.e.*, there are some deviations in the form of distribution). In other words, the estimation method is not robust enough. We will discuss the problem in this section.

During the following discussion, it is usual to suppose that x_1, \dots, x_n is an i.i.d. sample from $(\mathcal{X}, \mathcal{A}, \mathcal{P})$. If we let $x = (x_1, \dots, x_n)$, then the statistical space has the form $(\mathcal{X}^n, \mathcal{A}^n, \mathcal{P}^n)$, where $\mathcal{P}^n = \{P^n; P \in \mathcal{P}\}$. Now, we will discuss the first category of estimation methods.

1.4.1 Moment Estimation and Median Estimation

Since moments are very important tools for judging the characteristics of distribution of random variable, it is very reasonable to use the sample moments to estimate population moments when the information about the distribution of random variable is not sufficient. Let x_1, \dots, x_n be a set of samples, and M_s and m_s denote the s -th population moments and sample moments, respectively, *i.e.*

$$M_s = EX_1^s, \quad \text{and} \quad m_s = \frac{1}{n} \sum_{i=1}^n x_i^s. \quad (1.4.1)$$

m_s is called a **moment estimator** of M_s . More generally, if some parameter θ can be denoted as a function of moments, say

$$\theta = g(M_1, \dots, M_k), \quad (1.4.2)$$

then $\hat{\theta} = g(m_1, \dots, m_k)$ is called a moment estimator of θ .

Example 1.4.1. The moment estimator of the mean μ is

$$\hat{\mu} = m_1 = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i. \quad (1.4.3)$$

Since the variance σ^2 can be denoted as $\sigma^2 = M_2 - M_1^2$, from (1.4.2), the moment estimator of the variance is

$$\hat{\sigma}^2 = m_2 - m_1^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2. \quad (1.4.4)$$

Since the moment estimator is also a random variable, we usually need to calculate the mean and variance of the moment estimator to test the effect of the estimator. Recall (1.1.5) and (1.1.6), we have

$$E\bar{X} = \frac{1}{n} \sum_{i=1}^n EX_i = \mu, \quad (1.4.5)$$

and

$$V\bar{X} = E(\bar{X} - \mu)^2 = \frac{1}{n^2} \sum_{i=1}^n E(X_i - \mu)^2 = \frac{\sigma^2}{n}. \quad (1.4.6)$$

Compared with Example 1.2.1, the above result coincides with that of normal distribution. From Eq. (1.4.5), we know it is always reasonable to estimate the population mean by the sample mean whatever the distribution is, since the mean of the estimator is consistent to the population mean. This property is called **unbiasedness**, which will be discussed further later. Equation (1.4.6) shows that it is more accurate to estimate the population mean by the sample mean than by using each observation value alone.

Example 1.4.2. Let x_1, \dots, x_n be i.i.d. Bernoulli trials with success probability p . From Example 1.3.4, we know that $x = x_1 + \dots + x_n$ has the binomial distribution $Bi(n, p)$. From Example 1.1.1, $E_p X = np$, the moment estimator of p is $\hat{p} = x/n$. In many practical problems, we usually need to consider the ratio of success to failure, *i.e.*

$$\omega = \frac{p}{1-p}, \quad (1.4.7)$$

which is called **odds**. From Eq. (1.4.2), the estimator of ω is

$$\hat{\omega} = \frac{\hat{p}}{1-\hat{p}} = \frac{x}{n-x}.$$

If we let y_1, \dots, y_m be i.i.d. Bernoulli trials with success probability q , then $y = y_1 + \dots + y_m$ has the binomial distribution $Bi(m, q)$. Its odds is

$$\nu = \frac{q}{1 - q},$$

and the estimator of ν is $\hat{\nu} = y/(m - y)$. $\theta = \omega/\nu$, the ratio of one odds to another, is a powerful statistic to measure the size of the two success probabilities, which will be called **odds ratio**. From Eq. (1.4.2), its estimator can be written as

$$\hat{\theta} = \frac{\hat{\omega}}{\hat{\nu}} = \frac{x(m - y)}{(n - x)y}. \tag{1.4.8}$$

Obviously we can claim that $q = p$ when $\hat{\theta} = 1$, and that $q > p$ when $\hat{\theta} > 1$. However, since $\hat{\theta}$ is a random variable, the probability of the event $\{\hat{\theta} = 1\}$ occurring is very small even if the true parameter is exactly 1. Then how to compare the size relation of p to q by using $\hat{\theta}$? This is a problem that will be discussed in the next chapter.

Besides the mean and the variance, the median is also an important statistical index of centrality. Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be a statistical space, where $\mathcal{P} = \{P_\theta; \theta \in \Theta\}$ is a continuous distribution, *i.e.* $\mathcal{P} \ll \nu$ with ν a Lebesgue measure. If θ satisfies the following equation

$$F_\theta(\theta) = P_\theta(X \leq \theta) = \frac{1}{2}, \tag{1.4.9}$$

then θ is called a **median**, in other words, the median is the solution of $F_\theta(\theta) = 1/2$. Since it is a continuous distribution, the solution exists uniquely with probability one. Especially when the distribution is symmetrical, the median coincides with the mean.

Let $\mathbf{x} = (x_1, \dots, x_n)$ be an i.i.d. sample from the statistical space $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, and $x_{(1)} < \dots < x_{(n)}$ denote the order statistics, and the **sample median** is defined as

$$m_{\mathbf{x}} = \begin{cases} x_{(n+1)/2}, & \text{when } n \text{ is odd,} \\ \frac{1}{2}[x_{(n/2)} + x_{(n/2+1)}], & \text{when } n \text{ is even.} \end{cases} \tag{1.4.10}$$

We can estimate the population median by the sample median, *i.e.* $\hat{\theta} = m_{\mathbf{x}}$.

For the given $\theta \in \Theta$, let $S(\theta) = \#\{x_i \leq \theta\} = \#\{x_{(i)} \leq \theta\}$ denote the the number of x_i 's that are less than or equal to θ , which is a non-decreasing step function of θ . Since θ is the median, based on the idea of moment estimation, we can regard the estimator of θ as the solution of

$$S(\theta) = n - S(\theta), \tag{1.4.11}$$

i.e. $S(\hat{\theta}) = n/2$. We can get that $\hat{\theta} = x_{((n+1)/2)}$ when n is odd, and that $\hat{\theta}$ can be any number between $x_{(n/2)}$ and $x_{(n/2+1)}$ when n is even, which coincides with the sample median m_x defined by (1.4.10).

We have known that the mean coincides with the median when the distribution is symmetrical, such as normal distribution. So, logically, we can estimate the mean by the sample median. And then which estimator is better? We will discuss this problem in the last part of this chapter.

1.4.2 Maximum Likelihood Estimators

Essentially, the maximum likelihood method discusses the problem of parameter estimation when the pdf is given. The method is generally credited to Fisher, although its roots date back as far as Lambert, Daniel Bernoulli, and Lagrange in the eighteenth century (Scholz, 1985). It is by far the most popular general method of estimation in statistics.

Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be a statistical space, where $\mathcal{P} = \{P_\theta; \theta \in \Theta\}$ and $\mathcal{P} \ll \nu$. For $\theta \in \Theta$, let $dP_\theta/d\nu = f_\theta$. Furthermore, let X_1, \dots, X_n be an i.i.d. sample from the distribution with pdf $f_\theta(x)$, and its corresponding observation is $\mathbf{x} = (x_1, \dots, x_n)$. Thus we can get the values of $f_\theta(x_1), \dots, f_\theta(x_n)$, where θ is unknown. The basic idea of maximum likelihood method is to find the parameter θ such that the values of the density functions attain their maxima. Here, the joint pdf of $\mathbf{x} = (x_1, \dots, x_n)$ is

$$L(\mathbf{x}; \theta) = \prod_{i=1}^n f_\theta(x_i), \quad (1.4.12)$$

or by the monotonicity of log function,

$$l(\mathbf{x}; \theta) = \ln L(\mathbf{x}; \theta) = \sum_{i=1}^n \ln f_\theta(x_i). \quad (1.4.13)$$

$L(\mathbf{x}; \theta)$ and $l(\mathbf{x}; \theta)$ are called a **likelihood function** and **log-likelihood Function** of the sample respectively. $\hat{\theta}$ is called a **maximum likelihood estimator** (MLE) of the parameter θ , if $\hat{\theta} \in \Theta$ and satisfies

$$L(\mathbf{x}; \hat{\theta}) = \sup_{\theta \in \Theta} L(\mathbf{x}; \theta) \quad \text{or} \quad l(\mathbf{x}; \hat{\theta}) = \sup_{\theta \in \Theta} l(\mathbf{x}; \theta). \quad (1.4.14)$$

Obviously, if the pdf is derivable with respect to θ_i , and the solution of

$$\frac{\partial}{\partial \theta_i} l(\mathbf{x}; \theta) = 0 \quad (1.4.15)$$

lies in the parameter space Θ , then the MLE of θ is the solution of Eq. (1.4.15).

Example 1.4.3. Let x_1, \dots, x_n be a set of samples from the normal distribution $N(\mu, \sigma^2)$ to calculate the MLE of the parameters. It is easy to get that

$$l(\mathbf{x}; \boldsymbol{\theta}) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2,$$

where $\boldsymbol{\theta} = (\mu, \sigma^2)$. Taking partial derivatives with respect to μ and σ^2 , we have

$$\frac{\partial}{\partial \mu} l(\mathbf{x}; \mu, \sigma^2) = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0, \tag{1.4.16}$$

and

$$\frac{\partial}{\partial \sigma^2} l(\mathbf{x}; \mu, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 = 0. \tag{1.4.17}$$

Notice that Eq. (1.4.16) implies that we always have $\hat{\mu} = \bar{x}$ irrespective of the variance. Substituting it into Eq. (1.4.17) yields $\hat{\sigma}^2 = 1/n \sum_{i=1}^n (x_i - \bar{x})^2$.

However it needs to prove that $(\bar{x}, \hat{\sigma}^2)$ is the MLE of $\boldsymbol{\theta}$, since the log-likelihood function is not a concave function of $\boldsymbol{\theta}$. Recall Example 1.3.1, we take the following transformation

$$\begin{aligned} \theta_1 &= \frac{\mu}{\sigma^2}, & t_1(\mathbf{x}) &= \sum_{i=1}^n x_i; \\ \theta_2 &= -\frac{1}{2\sigma^2}, & t_2(\mathbf{x}) &= \sum_{i=1}^n x_i^2. \end{aligned}$$

Then the likelihood function can be written as a natural exponential family,

$$L(\mathbf{x}; \boldsymbol{\theta}^*) = u(\boldsymbol{\theta}^*) \exp\{\theta_1 t_1(\mathbf{x}) + \theta_2 t_2(\mathbf{x})\},$$

where $\boldsymbol{\theta}^* = (\theta_1, \theta_2)$ and $u(\boldsymbol{\theta}^*) = (-\theta_2/\pi)^{n/2} \exp\{n\theta_1^2/(4\theta_2)\}$. Then the log-likelihood function is

$$l(\mathbf{x}; \boldsymbol{\theta}^*) = \ln u(\boldsymbol{\theta}^*) + \theta_1 t_1(\mathbf{x}) + \theta_2 t_2(\mathbf{x}).$$

From Theorem 1.3.2 we know $l(\mathbf{x}; \boldsymbol{\theta}^*)$ is a concave function of $\boldsymbol{\theta}^*$, so the MLE of $\boldsymbol{\theta}^*$ exists uniquely, which is the solution of the equations below,

$$\begin{cases} E_{\boldsymbol{\theta}^*} t_1(\mathbf{X}) = n\mu = t_1(\mathbf{x}), \\ E_{\boldsymbol{\theta}^*} t_2(\mathbf{X}) = n(\sigma^2 + \mu^2) = t_2(\mathbf{x}). \end{cases}$$

We have $\hat{\mu} = \bar{x}$ and $\hat{\sigma}^2 = \sum_{i=1}^n (x_i - \bar{x})^2/n$. It can be seen that this coincides with the moment estimators. It is necessary to point out that when parameter spaces are different, estimation methods may be different too. Take the consideration of the parameter space with the restriction of the mean being nonnegative as an example.

$$\Theta_1 = \{(\mu, \sigma^2); 0 \leq \mu < +\infty, 0 < \sigma^2 < \infty\}$$

then computing the MLE of μ is equivalent to computing

$$\min_{\mu \geq 0} (\bar{x} - \mu)^2.$$

Then the MLE of μ is $\hat{\mu}^* = \max\{\bar{x}, 0\}$. Notice that

$$\begin{aligned} (\hat{\mu} - \mu)^2 &= (\bar{x} - \mu)^2 \\ &= (\bar{x} - \hat{\mu}^*)^2 + 2(\bar{x} - \hat{\mu}^*)(\hat{\mu}^* - \mu) + (\hat{\mu}^* - \mu)^2 \end{aligned}$$

for $\forall \mu \geq 0$, where the cross-product term satisfies

$$(\bar{x} - \hat{\mu}^*)\hat{\mu}^* = 0, \quad (1.4.18)$$

and

$$(\bar{x} - \hat{\mu}^*)\mu \leq 0. \quad (1.4.19)$$

For $\forall \mu \geq 0$ we have $(\hat{\mu} - \mu)^2 \geq (\hat{\mu}^* - \mu)^2$, thus

$$E_{\theta}(\hat{\mu} - \mu)^2 > E_{\theta}(\hat{\mu}^* - \mu)^2$$

holds for $\forall \theta \in \Theta_1$. This result shows that when the restriction condition is true, the deviation of the MLE is smaller than that without any restriction. Lee (1981) had ever calculated the deviation.

Example 1.4.4. Let $\mathbf{x} = (x_1, \dots, x_k)$ be a set of samples from a multinomial distribution $M(n; p_1, \dots, p_k)$. There are two methods to get the MLE of the parameter $\boldsymbol{\theta} = (p_1, \dots, p_k)$.

(i) Let $p_k = 1 - p_1 - \dots - p_{k-1}$. From Example 1.1.3,

$$l(\mathbf{x}; \boldsymbol{\theta}) = \sum_{i=1}^{k-1} x_i \ln p_i + x_k \ln(1 - p_1 - \dots - p_{k-1}) + c,$$

where c is a constant independent of the parameter. Solving Eq. (1.4.15) yields

$$\frac{x_i}{p_i} = \frac{x_k}{1 - p_1 - \dots - p_{k-1}}, \quad i = 1, \dots, k-1.$$

Notice that $x_1 + \dots + x_k = n$, then the MLE of p_i is

$$\hat{p}_i = \frac{x_i}{n}, i = 1, \dots, k.$$

(ii) Applying the Lagrange multiplier method. The objective function is

$$H(\mathbf{x}; \boldsymbol{\theta}, \lambda) = \sum_{i=1}^k x_i \ln p_i + \lambda(1 - p_1 - \dots - p_k),$$

where λ is the Lagrange multiplier. Our aim is to solve the equations

$$\begin{cases} \frac{\partial}{\partial p_i} H(\mathbf{x}; \boldsymbol{\theta}, \lambda) = \frac{x_i}{p_i} - \lambda = 0, & i = 1, \dots, k, \\ \frac{\partial}{\partial \lambda} H(\mathbf{x}; \boldsymbol{\theta}, \lambda) = 1 - p_1 - \dots - p_k = 0. \end{cases}$$

After summation, we have $\sum_{i=1}^k x_i = \lambda \sum_{i=1}^k p_i$, *i.e.* $\lambda = n$. Then the MLE of p_i is $\hat{p}_i = \frac{x_i}{n}$, $i = 1, \dots, k$.

Then the reasonability about the idea of maximum likelihood method will be discussed. Let $L_n(\mathbf{x}; \theta)$ denote the likelihood function when the sample size is n , then we have the following theorem.

Theorem 1.4.1. *Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be a statistical space, $\mathcal{P} \ll \nu$ and $dP_\theta/d\nu = f_\theta$. If θ and f_θ are one-to-one, and the true value of the parameter θ_0 is an interior point of Θ , then for $\forall \theta \neq \theta_0$ we have*

$$\lim_{n \rightarrow \infty} P_{\theta_0} \{L_n(\mathbf{X}; \theta_0) > L_n(\mathbf{X}; \theta)\} = 1.$$

Proof. From the definition of likelihood function, we know that

$$L_n(\mathbf{x}; \theta_0) > L_n(\mathbf{x}; \theta) \quad \text{if and only if} \quad \frac{1}{n} \sum_{i=1}^n \ln \frac{f_\theta(x_i)}{f_{\theta_0}(x_i)} < 0.$$

Since θ_0 is an interior point of Θ , the above equation can be written as a form of integration when $n \rightarrow \infty$, *i.e.*

$$E_{\theta_0} \ln \frac{f_\theta(X)}{f_{\theta_0}(X)} < 0.$$

Since θ and f_θ are one-to-one, and $\ln x$ is a concave function, by Jensen inequality the left side of the above equation is strictly smaller than

$$\ln E_{\theta_0} \frac{f_\theta(X)}{f_{\theta_0}(X)} = 0.$$

Therefore this completes the proof of the theorem. □

Above theorem shows that when n is quite large, the likelihood function attains its maximum at the true value of parameter with probability one. Though we do not know the true parameter in the estimation problem, it is generally reasonable to obtain the MLE based on the sample just like the idea of moment method. For the same estimation method, generally speaking, the bigger the sample size, the better the estimator. We will exemplify the proposition by applying the maximum likelihood method. Let $\hat{\theta}$ be an estimator of θ . For $\theta \in \Theta$, $E_{\theta}(\hat{\theta} - \theta)^2$ is called the **mean squared error** of $\hat{\theta}$. Obviously the mean squared error is an important index to measure the difference between the estimator and the true value. The smaller the mean squared error, the better the estimator.

Theorem 1.4.2. *Let x_{n+m} denote a set of samples in size $n + m$ from the statistical space $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, which can be divided into two non-empty parts $x_{n+m} = (x_n, x_m)$. If \mathcal{P} belongs to the natural exponential family (cf. (1.3.3)), and $E_{\theta}t_i(X)$ is a linear function of θ , say $h(\theta)$, then for any interior point θ of Θ we have*

$$E_{\theta}(\hat{\theta}_{n+m} - \theta)^2 < \min\{E_{\theta}(\hat{\theta}_n - \theta)^2, E_{\theta}(\hat{\theta}_m - \theta)^2\},$$

where $\hat{\theta}_s$ denotes the MLE of θ when the sample size is s .

Proof. Let $l(x; \theta)$ denote the log-likelihood function of the sample x , then we have

$$l(x_{n+m}; \theta) = l(x_n; \theta) + l(x_m; \theta).$$

From Theorem 1.3.2, $\hat{\theta}_s$ is a solution of $E_{\theta}t_i(X_s) = t_i(x_s)$. Let $h(\theta)$ be the linear function as given in the assumption, then

$$\begin{aligned} h(\hat{\theta}_{n+m}) &= \frac{1}{n+m} E_{\hat{\theta}_{n+m}} t_i(X_{n+m}) \\ &= \frac{1}{n+m} t_i(x_{n+m}) \\ &= \frac{1}{n+m} t_i(x_n) + \frac{1}{n+m} t_i(x_m) \\ &= \frac{1}{n+m} E_{\hat{\theta}_n} t_i(X_n) + \frac{1}{n+m} E_{\hat{\theta}_m} t_i(X_m) \\ &= \frac{n}{n+m} h(\hat{\theta}_n) + \frac{m}{n+m} h(\hat{\theta}_m). \end{aligned}$$

According to the assumption of linearity of h , we have

$$\hat{\theta}_{n+m} = \frac{n}{n+m} \hat{\theta}_n + \frac{m}{n+m} \hat{\theta}_m. \quad (1.4.20)$$

Without the loss of generality, suppose that $\hat{\theta}_1$ is not a constant. Then we will apply the mathematical induction method to prove

$$E_{\theta}(\hat{\theta}_s - \theta)^2 = \frac{1}{s}E_{\theta}(\hat{\theta}_1 - \theta)^2 + \frac{s-1}{s}[E_{\theta}(\hat{\theta}_1 - \theta)]^2. \tag{1.4.21}$$

Notice that when $E_{\theta}\hat{\theta}_s = \theta$, the last term in the above equation is zero. At first, we will prove that (1.4.21) holds for $s = 2$.

$$\begin{aligned} E_{\theta}(\hat{\theta}_2 - \theta)^2 &= E_{\theta} \left(\frac{1}{2}\hat{\theta}_1 + \frac{1}{2}\hat{\theta}_1^* - \theta \right)^2 \\ &= E_{\theta} \left[\frac{1}{4}(\hat{\theta}_1 - \theta)^2 + \frac{1}{2}(\hat{\theta}_1 - \theta)(\hat{\theta}_1^* - \theta) + \frac{1}{4}(\hat{\theta}_1^* - \theta)^2 \right] \\ &= \frac{1}{2}E_{\theta}(\hat{\theta}_1 - \theta)^2 + \frac{1}{2}[E_{\theta}(\hat{\theta}_1 - \theta)]^2, \end{aligned}$$

where $\hat{\theta}_1^*$ and $\hat{\theta}_1$ are the two MLEs using different samples, obviously they are i.i.d. And hence when $s = 2$, Eq. (1.4.21) holds. We will prove the case for $s = p + 1$ assuming that Eq. (1.4.21) holds for $s = p$. By Eq. (1.4.20) and the induction assumption, we have

$$\begin{aligned} E_{\theta}(\hat{\theta}_{p+1} - \theta)^2 &= E_{\theta} \left(\frac{p}{p+1}\hat{\theta}_p + \frac{1}{p+1}\hat{\theta}_1 \right)^2 \\ &= \left(\frac{p}{p+1} \right)^2 E_{\theta}(\hat{\theta}_p - \theta)^2 + \frac{2p}{(p+1)^2}E_{\theta}(\hat{\theta}_p - \theta)E_{\theta}(\hat{\theta}_1 - \theta) \\ &\quad + \left(\frac{1}{p+1} \right)^2 E_{\theta}(\hat{\theta}_1 - \theta)^2 \\ &= \frac{p+1}{(p+1)^2}E_{\theta}(\hat{\theta}_1 - \theta)^2 + \left(\frac{p}{p+1} \right)^2 \cdot \frac{p-1}{p}[E_{\theta}(\hat{\theta}_1 - \theta)]^2 \\ &\quad + \frac{2p}{(p+1)^2}[E_{\theta}(\hat{\theta}_1 - \theta)]^2 \\ &= \frac{1}{p+1}E_{\theta}(\hat{\theta}_1 - \theta)^2 + \frac{p}{p+1}[E_{\theta}(\hat{\theta}_1 - \theta)]^2. \end{aligned}$$

Then Eq. (1.4.21) holds. From Eq. (1.4.21) we can get

$$E_{\theta}(\hat{\theta}_p - \theta)^2 - E_{\theta}(\hat{\theta}_{p+1} - \theta)^2 = \frac{1}{p(p+1)}[E_{\theta}(\hat{\theta}_1 - \theta)^2 - (E_{\theta}(\hat{\theta}_1 - \theta))^2].$$

By the Cauchy-Schwarz Inequality, the proof is completed. □

Remark 1.4.1. Let x_1, \dots, x_n be an independent and identically distributed random sample with density $f(\cdot, \theta)$, where θ is an unknown parameter, and let $\hat{\theta}_n$ denote the maximum likelihood estimator of θ based

on the sample. Now, suppose that we obtain an additional observation, say x_{n+1} , from the same distribution. In this case, $\hat{\theta}_{n+1}$ denotes the maximum likelihood estimator of θ based on the $n + 1$ observations. Theorem 1.4.2 tells us that $\hat{\theta}_{n+1}$ is preferred to $\hat{\theta}_n$ as estimation of θ under some given conditions. Here, we conjecture that this result still holds for *any* maximum likelihood estimator under some regular conditions such as $\text{MSE}(\hat{\theta})$ is finite (Shi, 2008).

1.4.3 Quality of Estimators

From the above discussion, in principle, we can construct many estimators, then what properties should a good estimator have? Furthermore, how to measure the quality of estimator? Obviously, a good estimator must be a function of sufficient statistics to guarantee that no information about the parameters is lost. Besides that, it should also possess the following properties.

(1) **Unbiasedness.** Let $\mathbf{x} = (x_1, \dots, x_n)$ be a set of samples from the statistical space $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, and $T_n(\mathbf{x})$ an estimator of θ . We know that the probability of $T_n(\mathbf{x}) = \theta$ is quite small almost for any distribution, which motivates us to consider the average situations. If for $\forall \theta \in \Theta$ we have

$$E_{\theta} T_n(\mathbf{X}) = \theta,$$

then $T_n(\mathbf{x})$ is called an **unbiased estimator** of θ . For a function of parameter, we can give a similar definition. Let g be a function of θ , if $E_{\theta} T_n(\mathbf{X}) = g(\theta)$ for $\forall \theta \in \Theta$, then $T_n(\mathbf{x})$ is called an unbiased estimator of $g(\theta)$. Consider the limiting behavior of the mean, if for $\forall \theta \in \Theta$ we have

$$\lim_{n \rightarrow \infty} E_{\theta} T_n(\mathbf{X}) = \theta,$$

then $T_n(\mathbf{x})$ is called an **asymptotic unbiased estimator** of θ .

(2) **Consistency.** As we have mentioned above, the probability of $T_n(\mathbf{x}) = \theta$ is quite small. However, we may hope that $T_n(\mathbf{x})$ approximates to θ with a higher probability when n is large enough, otherwise the estimator will become meaningless. For this we introduce the following criterion based on the convergence in probability: if for $\forall \varepsilon > 0$, we have

$$\lim_{n \rightarrow \infty} P_{\theta}(|T_n(\mathbf{X}) - \theta| \geq \varepsilon) = 0, \quad \forall \theta \in \Theta,$$

then $T_n(\mathbf{x})$ is called a **consistent estimator** of θ , and denoted as $T_n \xrightarrow{P_{\theta}} \theta$. Furthermore, we have the following theorem.

Theorem 1.4.3. *Let $T_n(\mathbf{x})$ be a consistent estimator of θ . If g is a continuous function of θ , then $g(T_n)$ is a consistent estimator of $g(\theta)$.*

Proof. By the continuity of function, we know that for $\forall \varepsilon > 0$, there exists $\delta > 0$, s.t. $|g(T_n) - g(\theta)| < \varepsilon$ when $|T_n(\mathbf{x}) - \theta| < \delta$. Then when $n \rightarrow \infty$,

$$\begin{aligned} 1 &\geq P_\theta(|g(T_n) - g(\theta)| \leq \varepsilon) \\ &\geq P_\theta(|T_n(X) - \theta| \leq \delta) \rightarrow 1, \quad \forall \theta \in \Theta, \end{aligned}$$

or equivalently, $g(T_n) \xrightarrow{P_\theta} g(\theta)$. □

(3) **Asymptotic normality.** For an estimator $T_n(\mathbf{x})$, besides the consideration of its approximation to the true parameter θ , we should take the distribution of the deviation $T_n(\mathbf{x}) - \theta$ into account as a basic idea in statistics. There are two main reasons. One is that we can determine the speed of convergence about the consistent estimator, and the other is that the limiting distribution should be a normal with mean zero if the deviate is induced by random errors (refer to Problem 1.2). Therefore the property of convergence in distribution is called asymptotic normality. If for any sample size n , there exists a function $\sigma_n^2(\theta)$ of θ , s.t. for any x when $n \rightarrow \infty$ we have

$$F_n(x) \rightarrow \Phi(x), \tag{1.4.22}$$

where $\Phi(x)$ is the cdf of the standard normal distribution $N(0, 1)$ and $F_n(x)$ is the cdf of the following r.v.

$$Z_n = \frac{T_n - \theta}{\sigma_n(\theta)},$$

then $T_n(\mathbf{x})$ is called an **asymptotic normal estimator** of θ , and $\sigma_n^2(\theta)$ the **asymptotic variance** of $T_n(\mathbf{x})$. If Z denotes an r.v. from $N(0, 1)$, then (1.4.22) can also be denoted as $Z_n \xrightarrow{L} Z$.

Example 1.4.5. (Discussing the properties of moment estimators)

From Example 1.4.1 we know that the moment estimators of the mean μ and the variance σ^2 are given by \bar{x} in (1.4.3) and $\hat{\sigma}^2$ in (1.4.4) respectively. Let $\theta = (\mu, \sigma^2)$ and $\hat{\theta} = (\bar{x}, \hat{\sigma}^2)$. From the Law of Large Numbers, $\hat{\theta}$ is a consistent estimator of θ . From (1.4.5), (1.4.6) and the Central Limit Theorem we know that, for $\forall \theta \in \Theta$, when $n \rightarrow \infty$ we have

$$P_\theta \left\{ \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \leq x \right\} \rightarrow \Phi(x).$$

Thus \bar{x} is an asymptotic normal estimator of μ . When σ^2 is unknown, since $\hat{\sigma}^2/\sigma^2 \rightarrow 1$ when $n \rightarrow \infty$, the above equation can be written as,

$$P_\theta \left\{ \frac{\sqrt{n}(\bar{X} - \mu)}{\hat{\sigma}} \leq x \right\} \rightarrow \Phi(x). \tag{1.4.23}$$

Since $E_{\theta}\hat{\sigma}^2 = (n-1)\sigma^2/n$ for $\forall\theta$, then $\hat{\sigma}^2$ is not an unbiased estimator of σ^2 . An unbiased estimator of σ^2 should be

$$\frac{n}{n-1}\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^n(x_i - \bar{x})^2. \quad (1.4.24)$$

Example 1.4.6. (Discussing the properties of median estimators)

We will utilize the sample median $\hat{\theta}_n$ to estimate the median θ . Since $\hat{\theta}_n$ is a function of order statistics, from Example 1.3.5, it is a function of sufficient statistics. From the Law of Large Numbers we know that, $\hat{\theta}_n$ is a consistent estimator of θ . Then we will discuss its asymptotic normality.

Let $Z_n = \sqrt{n}(\hat{\theta}_n - \theta)$. For $\forall x \in \mathcal{X}$ we have

$$\begin{aligned} P_{\theta}(Z_n \leq x) &= P_{\theta}\left\{\sqrt{n}(\hat{\theta}_n - \theta) \leq x\right\} \\ &= P_{\theta}\left\{\hat{\theta}_n \leq \theta + \frac{x}{\sqrt{n}}\right\} \\ &= P_{\theta}\left\{S\left(\theta + \frac{x}{\sqrt{n}}\right) \geq \frac{n}{2}\right\}, \end{aligned}$$

where $S(y) = \#\{x_i \leq y\}$ is given by (1.4.11). Notice that $S(\theta + x/\sqrt{n}) = \sum_{i=1}^n I_{\{x_i \leq \theta + x/\sqrt{n}\}}$, let

$$Y_{n_i} = I_{\{x_i \leq \theta + x/\sqrt{n}\}} - F\left(\theta + \frac{x}{\sqrt{n}}\right),$$

and

$$t_n = \frac{1}{\sqrt{n}}\left[\frac{n}{2} - nF\left(\theta + \frac{x}{\sqrt{n}}\right)\right],$$

then we have

$$\begin{aligned} S\left(\theta + \frac{x}{\sqrt{n}}\right) \geq \frac{n}{2} &\Leftrightarrow \sum_{i=1}^n I_{\{x_i \leq \theta + \frac{x}{\sqrt{n}}\}} - nF\left(\theta + \frac{x}{\sqrt{n}}\right) \geq \frac{n}{2} - nF\left(\theta + \frac{x}{\sqrt{n}}\right) \\ &\Leftrightarrow \sum_{i=1}^n Y_{n_i} \geq \sqrt{n}t_n. \end{aligned}$$

Notice that $S(y)$ has a binomial distribution $Bi(n, p)$ with $p = F(y)$, it is easy to verify that

$$E_{\theta}Y_{n_i} = F\left(\theta + \frac{x}{\sqrt{n}}\right) - F\left(\theta + \frac{x}{\sqrt{n}}\right) = 0,$$

and

$$\begin{aligned}
 V_{\theta}Y_{n_i} &= E_{\theta}Y_{n_i}^2 \\
 &= F\left(\theta + \frac{x}{\sqrt{n}}\right) \left[1 - F\left(\theta + \frac{x}{\sqrt{n}}\right)\right] \\
 &= F(\theta)(1 - F(\theta)) + o\left(\frac{1}{\sqrt{n}}\right) \\
 &= \frac{1}{4} + o\left(\frac{1}{\sqrt{n}}\right),
 \end{aligned}$$

and

$$\begin{aligned}
 t_n &= \frac{1}{\sqrt{n}} \left\{ \frac{n}{2} - n \left[F(\theta) + F'(\theta) \frac{x}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right) \right] \right\} \\
 &= \frac{1}{\sqrt{n}} \left\{ \frac{n}{2} - \frac{n}{2} - \sqrt{n}xf(\theta) - n \cdot o\left(\frac{1}{\sqrt{n}}\right) \right\} \\
 &= -xf(\theta) + o(1).
 \end{aligned}$$

From the Central Limit Theorem, we have

$$\begin{aligned}
 \lim_{n \rightarrow \infty} P_{\theta}(Z_n \leq x) &= \lim_{n \rightarrow \infty} P_{\theta} \left\{ \frac{1}{\sqrt{n/4}} \sum_{i=1}^n Y_{n_i} \geq \frac{1}{\sqrt{1/4}} t_n \right\} \\
 &= 1 - \Phi(-2xf(\theta)) \\
 &= \Phi(2xf(\theta)).
 \end{aligned}$$

Therefore $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{L} N\left(0, \frac{1}{4f^2(\theta)}\right)$.

Example 1.4.7. (Discussing the properties of MLE) From the Neyman's Factorization Theorem we know that the MLE is a function of sufficient statistics. Especially for an exponential family, from Theorem 1.3.4, the MLE is a function of complete sufficient statistics. From Theorem 1.4.1, we may infer that the MLE $\hat{\theta}_n$ is a consistent estimator of the parameter θ . Now we discuss its asymptotic normality. For $\theta \in \Theta$, let

$$I(\theta) = E_{\theta} \left(\frac{\partial \ln f_{\theta}(X)}{\partial \theta} \right)^2. \quad (1.4.25)$$

Usually $I(\theta)$ is called the Fisher information, since it is related to the vari-

ance of statistics. From Eq. (1.4.13), we can calculate that

$$\begin{aligned} E_{\theta} \frac{\partial \ln L(\mathbf{X}; \theta)}{\partial \theta} &= \sum_{i=1}^n \int \frac{\partial \ln f_{\theta}(x_i)}{\partial \theta} f_{\theta}(x_i) d\mu(\mathbf{x}) \\ &= \sum_{i=1}^n \int \frac{\partial f_{\theta}(x_i)}{\partial \theta} d\mu(\mathbf{x}) \\ &= \sum_{i=1}^n \frac{\partial}{\partial \theta} \int f_{\theta}(x_i) d\mu(\mathbf{x}) = 0. \end{aligned}$$

From the independence we can get,

$$\begin{aligned} V_{\theta} \frac{\partial \ln L(\mathbf{X}; \theta)}{\partial \theta} &= E_{\theta} \left(\frac{\partial \ln L(\mathbf{X}; \theta)}{\partial \theta} \right)^2 \\ &= E_{\theta} \left(\sum_{i=1}^n \frac{\partial \ln f_{\theta}(X_i)}{\partial \theta} \right)^2 = nI(\theta). \end{aligned} \quad (1.4.26)$$

If $f_{\theta}(x)$ has the third-order derivative with respect to θ , then we have

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{L} N(0, I^{-1}(\theta_0)), \quad (1.4.27)$$

where θ_0 , denoting the true value of parameter, is an interior point of Θ . Then we will prove Eq. (1.4.27). Taking a Taylor series expansion of

$h(\mathbf{x}; \theta) = \frac{\partial}{\partial \theta} \ln L(\mathbf{x}; \theta)$ at θ_0 , we have

$$h(\mathbf{x}; \theta) = h(\mathbf{x}; \theta_0) + (\theta - \theta_0) \frac{\partial}{\partial \theta} h(\mathbf{x}; \theta_0) + o(|\theta - \theta_0|).$$

Substituting $\hat{\theta}_n$ for θ in the above equation yields

$$0 = h(\mathbf{x}; \hat{\theta}_n) = h(\mathbf{x}; \theta_0) + (\hat{\theta}_n - \theta_0) \frac{\partial}{\partial \theta} h(\mathbf{x}; \theta_0) + o(|\hat{\theta}_n - \theta_0|).$$

Since $\hat{\theta}_n$ is a consistent estimator of θ_0 , by neglecting the higher-order infinitesimal term, and solving the above equation, we can get

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\sqrt{n} \left[\frac{\partial}{\partial \theta} h(\mathbf{x}; \theta_0) \right]^{-1} h(\mathbf{x}; \theta_0).$$

Then we will calculate the limiting distribution of the right side of the above equation. Since

$$\begin{aligned} \frac{1}{n} \left[\frac{\partial}{\partial \theta} h(\mathbf{x}; \theta) \right] &= \frac{\partial^2}{\partial \theta^2} \left[\frac{1}{n} \sum_{i=1}^n \ln f_{\theta}(x_i) \right] \\ &\xrightarrow{P_{\theta}} \frac{\partial^2}{\partial \theta^2} E_{\theta} \ln f_{\theta}(X) \\ &= E_{\theta} \frac{\partial^2 \ln f_{\theta}(X)}{\partial \theta^2} \\ &= -E_{\theta} \left(\frac{\partial \ln f_{\theta}(X)}{\partial \theta} \right)^2, \end{aligned}$$

by Eq. (1.4.25), we have

$$-\frac{1}{n} \left[\frac{\partial}{\partial \theta} h(\mathbf{x}; \theta_0) \right] \xrightarrow{P_{\theta_0}} I(\theta_0). \tag{1.4.28}$$

Similar to Eq. (1.4.26), we can get that

$$E_{\theta_0} h(X; \theta_0) = 0,$$

and

$$V_{\theta_0} h(X; \theta_0) = nI(\theta_0).$$

By the Central Limit Theorem,

$$\sqrt{n} \cdot \frac{1}{n} h(\mathbf{x}; \theta_0) \xrightarrow{L} N(0, I(\theta_0)).$$

Then we complete the proof of (1.4.27) by (1.4.28).

Similarly, we can get the corresponding result when the parameter is a multidimensional vector. Let the parameter space Θ be a subset of \mathbf{R}^k , and the true value θ_0 is an interior point of Θ , then we also have

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{L} N(\mathbf{0}, \mathbf{I}_k^{-1}(\theta_0)), \tag{1.4.29}$$

where $\mathbf{I}_k(\theta_0)$ is a $k \times k$ positive definite matrix with (i, j) -element

$$I_{ij}(\theta_0) = E_{\theta_0} \left(\frac{\partial \ln f_{\theta}(X)}{\partial \theta_i} \right) \left(\frac{\partial \ln f_{\theta}(X)}{\partial \theta_j} \right), \quad i, j = 1, \dots, k.$$

Usually $\mathbf{I}_k(\theta_0)$ is called the **Fisher information matrix**.

1.4.4 Comparison of Estimators

In previous subsection, we have discussed three properties that a good estimator should have. In fact, many estimators satisfy the three properties. Consequently, it is important to make a comparison of estimators. Essentially, the properties are mainly related to the means of estimators. Now, we will pay much attention to the variances of estimators when we make a comparison among them.

Let \mathcal{T} denote the set of all unbiased estimators. For T_1 and $T_2 \in \mathcal{T}$, if $V_{\theta} T_1(X) \leq V_{\theta} T_2(X)$ for $\forall \theta \in \Theta$, then we say that T_1 is superior to T_2 . If $V_{\theta} T^*(X) \leq V_{\theta} T(X)$ for $\forall T \in \mathcal{T}$ and $\forall \theta \in \Theta$, then $T^* \in \mathcal{T}$ is called a **uniformly minimal variance unbiased estimator (UMVUE)**.

Theorem 1.4.4. *If the UMVUE exists, then it uniquely exists with probability one.*

Proof. Let T_1 and T_2 be UMVUEs, then for $\forall \theta \in \Theta$, we have

$$E_{\theta}T_1 = E_{\theta}T_2 = \theta,$$

$$V_{\theta}T_1 = V_{\theta}T_2.$$

For $\forall \theta \in \Theta$, we have $E_{\theta}(T_1 - T_2) = 0$. By the Chebychev inequality, it suffices to prove that

$$\begin{aligned} 0 &= V_{\theta}(T_1 - T_2) = E_{\theta}(T_1 - T_2)^2 \\ &= E_{\theta}T_1(T_1 - T_2) - E_{\theta}T_2(T_1 - T_2). \end{aligned}$$

Thus, it suffices to prove that for $\forall \theta \in \Theta$, we have $E_{\theta}T_1(T_1 - T_2) = 0$. If there exists $\theta_0 \in \Theta$ s.t. $E_{\theta_0}T_1(T_1 - T_2) \neq 0$, let $\lambda = E_{\theta_0}T_1(T_1 - T_2)/E_{\theta_0}(T_1 - T_2)^2$ and $T_{\lambda} = T_1 - \lambda(T_1 - T_2)$. It is easy to verify that $T_{\lambda} \in \mathcal{T}$, and

$$\begin{aligned} E_{\theta_0}T_{\lambda}^2 &= E_{\theta_0}T_1^2 - 2\lambda E_{\theta_0}T_1(T_1 - T_2) + \lambda^2 E_{\theta_0}(T_1 - T_2)^2 \\ &< E_{\theta_0}T_1^2. \end{aligned}$$

This contradicts with that T_1 is a UMVUE. \square

We have ever discussed that when constructing estimators, sufficient statistics, especially complete sufficient statistics should be taken into consideration. The following two theorems will explain that from another point of view.

Theorem 1.4.5. Let $S(x)$ be a sufficient statistic for θ . For $\forall T \in \mathcal{T}$, let

$$T^*(x) = E(T(X)|S(x)), \quad (1.4.30)$$

then $T^* \in \mathcal{T}$, and $V_{\theta}T^* \leq V_{\theta}T$ for $\forall \theta \in \Theta$.

Proof. Since $S(x)$ is a sufficient statistic for θ , the conditional mean in (1.4.30) is independent of the parameter θ . From Theorem 1.2.1, for $\forall \theta \in \Theta$ we have

$$E_{\theta}T^*(X) = E_{\theta}E(T(X)|S(X)) = E_{\theta}T(X) = \theta.$$

Thus $T^* \in \mathcal{T}$. From Theorems 1.2.1 and 1.2.2, for $\forall \theta \in \Theta$

$$\begin{aligned} E_{\theta}(T(X) - T^*(X))(T^*(X) - \theta) &= E_{\theta}E_{\theta}[(T(X) - T^*(X))(T^*(X) - \theta)|S(X)] \\ &= E_{\theta}\{(T^*(X) - \theta)E[(T(X) - T^*(X))|S(X)]\} \\ &= 0. \end{aligned}$$

Then

$$\begin{aligned} V_{\theta}T(X) &= E_{\theta}(T(X) - \theta)^2 \\ &= E_{\theta}(T^*(X) - \theta)^2 + E_{\theta}((T(X) - T^*(X))^2) \\ &\geq V_{\theta}T^*(X). \end{aligned} \quad \square$$

Theorem 1.4.6. *If a complete sufficient statistic exists for θ , then there exists a UMVUE for θ that is a function of the complete sufficient statistic.*

Proof. Let $S(x)$ be a complete sufficient statistic for θ . For $T \in \mathcal{T}$, let

$$T^*(x) = E(T(X)|S(x)).$$

For $\forall T_1 \in \mathcal{T}$, from Theorem 1.4.4, it suffices to prove that $V_\theta T^*(X) \leq V_\theta T_1^*(X)$ for $\forall \theta \in \Theta$, where $T_1^*(x) = E(T_1(X)|S(x))$.

Since $E_\theta[T^*(X) - T_1^*(X)] = \theta - \theta = 0$, by the completeness of statistical space, we have

$$T^*(x) = T_1^*(x)$$

almost everywhere (a.e.). This completes the proof of the theorem. □

From the above theorem and Theorem 1.3.4, there always exists a UMVUE for the exponential family, and it is a function of complete sufficient statistics. We will discuss a more general class of comparison problems of estimators. Let \mathcal{T}^* denote the set of all the asymptotic unbiased estimators, *i.e.*

$$\mathcal{T}^* = \{T_n(x); \lim_{n \rightarrow \infty} E_\theta T_n(x) = \theta, V_\theta T_n(X) < \infty, \forall \theta \in \Theta\}.$$

Let $\sigma_n^2(\theta) = V_\theta T_n(X)$. From the Central Limit Theorem,

$$\frac{1}{\sigma_n(\theta)}(T_n - \theta) \xrightarrow{L} N(0, 1). \tag{1.4.31}$$

Thus $\sigma_n^2(\theta)$ is called an asymptotic variance, and T_n an asymptotic normal estimator of θ . We know from the discussion in the previous section that, $\sigma_n(\theta)$ satisfying (1.4.31) may be not unique, but if $\sigma'_n(\theta)$ also satisfies (1.4.31), we must have

$$\frac{\sigma_n(\theta)}{\sigma'_n(\theta)} \rightarrow 1 \quad \text{as } n \rightarrow \infty. \tag{1.4.32}$$

Thus, in the sense of (1.4.32), we can regard the asymptotic variance of the asymptotic normal estimator to be unique. So, we can compare estimators in virtue of the asymptotic variance. Let $T_{1n}(x)$ and $T_{2n}(x)$ be two asymptotic normal estimators for θ , and their asymptotic variances be σ_{1n}^2 and σ_{2n}^2 respectively. Then

$$e = e(T_{1n}, T_{2n}) = \lim_{n \rightarrow \infty} \frac{1/\sigma_{1n}^2}{1/\sigma_{2n}^2} = \lim_{n \rightarrow \infty} \frac{\sigma_{2n}^2}{\sigma_{1n}^2} \tag{1.4.33}$$

is called the **asymptotic relative efficiency** of T_{1n} with respect to T_{2n} . Obviously, when $e(T_{1n}, T_{2n}) > 1$, the estimator T_{1n} is superior to T_{2n} . If x_1, \dots, x_n is an i.i.d. sample satisfying $E_\theta X_i = \theta$ and $V_\theta X_i = \sigma^2(\theta) < \infty$. From the Central Limit Theorem,

$$\frac{\sqrt{n}(\bar{X} - \theta)}{\sigma(\theta)} \xrightarrow{L} N(0, 1).$$

This shows that the asymptotic variance has the order $1/n$, *i.e.*

$$n\sigma_n^2(\theta) \rightarrow \sigma^2(\theta), \quad \text{as } n \rightarrow \infty.$$

Therefore, when the sample sizes of the two estimators are different (say, n_1 and n_2 , respectively), the relation of the two sample sizes is approximately

$$e = \frac{n_1}{n_2} \implies n_1 = en_2$$

in order to reach the same asymptotic relative efficiency. When $e < 1$, the sample size of T_{2n} should be larger than that of T_{1n} in order to reach the same efficiency.

Example 1.4.8. Let x_1, \dots, x_n be a set of samples from $N(\theta, 1)$ to estimate the parameter θ . It can be seen that both the sample mean \bar{x}_n and the sample median m_n are reasonable estimators for θ , and belong to \mathcal{T}^* . The asymptotic variance of \bar{x}_n is $1/n$. From Example 1.4.6, when $\theta = 0$, the asymptotic variance of m_n is

$$\frac{1}{4nf^2(0)} = \frac{2\pi}{4n} = \frac{\pi}{2n}.$$

Then $e = e(\bar{x}_n, m_n) = 2/\pi < 1$, and thus \bar{x}_n is superior to m_n . To reach the same efficiency, we must have

$$n_1 = \frac{2}{\pi}n_2,$$

i.e., the sample size of m_n is $\pi/2 \approx 1.57$ times that of \bar{x}_n . This shows that \bar{x}_n is superior to m_n as far as estimation of the mean θ is concerned. But when considering from the point of view of **robustness**, we may get a different result. Consider

$$X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} (1 - \alpha)\Phi(x - \theta) + \alpha\Phi\left(\frac{x - \theta}{3}\right),$$

where $\alpha \in (0, 1)$. This is a mixture normal distribution, which is still symmetrical about θ . Its pdf is

$$f_\alpha(x) = (1 - \alpha)\phi(x - \theta) + \alpha\frac{1}{3}\phi\left(\frac{x - \theta}{3}\right),$$

where $\phi(x)$ is the pdf of the standard normal distribution. The asymptotic variance of \bar{x}_n is

$$\sigma_{1n}^2 = \frac{1}{n}[(1 - \alpha) + 9\alpha] = \frac{1}{n}(1 + 8\alpha),$$

and the asymptotic variance of m_n is

$$\begin{aligned} \sigma_{2n}^2 &= \frac{1}{4nf_\alpha^2(0)} \\ &= \frac{1}{4n} \left[\frac{1 - \alpha}{\sqrt{2\pi}} + \frac{\alpha}{9\sqrt{2\pi}} \right]^{-2} \\ &= \frac{\pi}{2n} \left[\frac{9}{9 - 8\alpha} \right]^2. \end{aligned}$$

It is easy to verify that, when $\alpha = 1/8$,

$$\sigma_{1n}^2 = \frac{2}{n} > \sigma_{2n}^2 \approx \frac{1.9}{n}.$$

This shows that m_n is more robust than \bar{x}_n when the sample distribution has some fluctuations.

In the discussion about the parameter estimation in a statistical space, we can see that it is usual to analyze and compare estimators in different angles. Especially in practical application, we should make a concrete analysis of concrete problems.

1.4.5 Nonparametric MLE for Population cdf

Now we will study how to estimate a cdf when its form is completely unknown. Recall the definition of cdf in Example 1.3.7, that is, for the given probability measure P , the corresponding cdf is $F(x) = P(X \leq x)$. Obviously, the cdf is a nondecreasing right-continuous function, *i.e.* the left-hand limit $F(x - 0)$ at x may not be equal to the function value $F(x)$ itself. Let \mathcal{F} denote the set of all the cdfs. Let x_1, \dots, x_n be a set of samples from F , where $F \in \mathcal{F}$, we would like to estimate F by the sample. Recall the definition of likelihood function in (1.4.12), we may define a nonparametric likelihood function as

$$L(F) = \prod_{i=1}^n [F(x_i) - F(x_i - 0)]. \tag{1.4.34}$$

If $\hat{F} \in \mathcal{F}$ and satisfies

$$L(\hat{F}) = \sup_{F \in \mathcal{F}} L(F), \tag{1.4.35}$$

then \hat{F} is called a **nonparametric maximum likelihood estimator (NMLE)** of F .

Now we discuss how to solve Eq. (1.4.35). Recall the discussion about the order statistics in Examples 1.2.5 and 1.3.5, let $y_1 \leq \dots \leq y_n$ be the order statistics of x_1, \dots, x_n , define

$$F_n(x) = \begin{cases} 0, & \text{if } x < y_1, \\ \frac{k}{n}, & \text{if } y_k \leq x < y_{k+1}, \\ 1, & \text{if } x \geq y_n. \end{cases} \tag{1.4.36}$$

Obviously, $0 \leq F_n(x) \leq 1$, and it is a non-decreasing right-continuous function about x , hence $F_n \in \mathcal{F}$. F_n is called an **empirical distribution function**, which is a function of complete sufficient statistics. The following theorem shows that F_n is the NMLE of F , i.e. F_n is a solution of Eq. (1.4.35).

Theorem 1.4.7. *For any $F \in \mathcal{F}$, if $F \neq F_n$, then*

$$L(F) < L(F_n). \tag{1.4.37}$$

Proof. Suppose that there are m different values in the sample $\{x_1, \dots, x_n\}$, and their corresponding order statistics are $y_1^* < \dots < y_m^*$. Let n_j denote the number of the samples that are equal to y_j^* , i.e. $n_j = \#\{x_i = y_j^*\}$, $i = 1, \dots, n$ and $j = 1, \dots, m$. Let $p_j = F(y_j^*) - F(y_j^* - 0)$ and $\hat{p}_j = n_j/n$. If there exists j s.t. $p_j = 0$, by the definition of Eq. (1.4.34), Equation (1.4.37) holds. So we can assume that all $p_j > 0$, then obviously $\sum_{j=1}^m p_j \leq 1$.

Since for any $z \geq 0$ we have $\ln z \leq z - 1$, where the equality holds if and only if $z = 1$. From the given conditions we know that, there exists at least one j s.t. $p_j \neq \hat{p}_j$, thus we can get

$$\begin{aligned} \ln[L(F)/L(F_n)] &= \sum_{j=1}^m n_j \ln(p_j/\hat{p}_j) \\ &< \sum_{j=1}^m n_j (p_j/\hat{p}_j - 1) \\ &= n \sum_{j=1}^m \hat{p}_j (p_j/\hat{p}_j - 1) \\ &\leq 0. \end{aligned}$$

This completes the proof of the theorem. □

By the definition we can see that the empirical distribution function $F_n(x)$ denotes the frequency of the samples not exceeding x , and hence for the given x , $nF_n(x)$ has a binomial distribution $Bi(n, p)$, where $p = F(x)$ and F denotes the true cdf. Thus

$$EF_n(x) = F(x), \tag{1.4.38}$$

$$VF_n(x) = \frac{1}{n}F(x)[1 - F(x)]. \tag{1.4.39}$$

Notice that Eq. (1.4.38) shows that the NMLE of F is unbiased. From the Law of Large Numbers and the Central Limit Theorem, we can obtain the consistency and asymptotic normality of F_n , *i.e.*

$$P\left\{\lim_{n \rightarrow \infty} F_n(x) = F(x)\right\} = 1, \tag{1.4.40}$$

and

$$\frac{\sqrt{n}[F_n(x) - F(x)]}{\sqrt{F(x)[1 - F(x)]}} \longrightarrow N(0, 1). \tag{1.4.41}$$

In fact, we can obtain a stronger version than Eq. (1.4.40). Let

$$D_n = \sup_x |F_n(x) - F(x)|$$

then we have

$$P\left\{\lim_{n \rightarrow \infty} D_n = 0\right\} = 1. \tag{1.4.42}$$

Notice that Eq. (1.4.42) states that the empirical distribution function $F_n(x)$ converges to the true cdf $F(x)$ with probability one. Usually, the result is called the Glivenko-Cantelli Theorem. For more detailed discussion, see Loève (1963).

By the empirical distribution function, we can make a further analysis of the moment estimators and the median estimators. Recall Eq. (1.4.1), by substituting the empirical distribution function for the cdf, we get

$$\int x^s dF_n(x) = \frac{1}{n} \sum_{i=1}^n x_i^s = m_s.$$

So the moment estimator can also be regarded as an estimator obtained by the empirical distribution function. For the median, the function in Eq. (1.4.11) is equivalent to

$$S(x) = nF_n(x).$$

Therefore the solution of $S(\theta) = n/2$ is exactly the sample median.

1.5 Problems

- 1.1. For a given statistical space $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ with $\mathcal{P} = \{P_\theta; \theta \in \Theta\}$, let \mathcal{A}_1 and \mathcal{A}_2 be two sub σ -algebras of \mathcal{A} satisfying $\mathcal{A}_1 \subset \mathcal{A}_2$, and $h(x)$ an \mathcal{A} -measurable function. Prove that for $\forall \theta \in \Theta$ we have

$$E_\theta[E_\theta(h(X)|\mathcal{A}_2, x)|\mathcal{A}_1, x] = E_\theta(h(X)|\mathcal{A}_1, x)$$

almost everywhere. Furthermore, explain the minimal sufficient statistics using the above result.

- 1.2. (**Theory of errors based on the normal distribution**) Let x denote the true but known length of an object. Let x_i denote the i -th measurement result with measurement error $\varepsilon_i = x_i - x$, $i = 1, 2, \dots, n$. Here, both the ε_i 's and the x_i 's are r.v.'s. Assume that
- There has no systematic error, *i.e.*, the mean of the n measurement results, $\bar{x} = \sum_{i=1}^n x_i/n$, is equal to the true length x ;
 - The ε_i 's are mutually independent;
 - The ε_i 's have a common distribution with density function $f(\cdot)$.

Verify that ε_i has a normal distribution with mean 0 based on the idea the maximum likelihood method.

- 1.3. (**The distribution of shoot deviation**) Consider a shooting contest where one aims a bullet at a target. The coordinate system is set up on the target plane where the bull's-eye is the origin O . Let the point where the bullet hits the target be (X, Y) . Here, the deviations X and Y are two random variables. Suppose that the following conditions are satisfied:

- The pdfs $p(x)$ and $q(y)$ of X and Y are continuous, and $p(0)q(0) > 0$;
- X and Y are mutually independent;
- The value of joint distribution of X and Y at (x, y) depends only on the distance $r = \sqrt{(x^2 + y^2)}$ between this point and the origin point.

Then both X and Y have a normal distribution with standard deviation $\sigma > 0$, *i.e.* they have the same pdf as follows:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2\sigma^2}\right\}.$$

- 1.4. (**Poisson process**) Consider the number of radioactive particles emitted in a unit of time. Assume that it satisfies the following three properties:

- (a) **(Stationary increments)** The number of particles emitted in $[t_0, t_0 + t)$ depends only on the length t but is independent of the starting time t_0 . If $P_k(t)$ denotes the probability that there are k particles emitted in a given interval of length t , then

$$\sum_{k=0}^{\infty} P_k(t) = 1$$

holds for any t . This property shows the probability distribution does not change with time.

- (b) **(Independent increments or without after-effects)** The event of k particles arriving at the given region in $[t_0, t_0 + t)$ is independent of the event occurring before t_0 . The property shows that the numbers of particles in two disjoint intervals are independent. Independent Increment indicates the processes are independent in the disjoint time intervals.

- (c) **(Orderliness)** In a sufficiently small interval, exactly one particle arrives at the given region at most. If

$$\psi(t) = \sum_{k=2}^{\infty} P_k(t) = 1 - P_0(t) - P_1(t)$$

then we must have $\psi(t) = o(t)$, *i.e.*

$$\lim_{t \rightarrow 0} \frac{\psi(t)}{t} = 0.$$

This property shows that, in practice, two or more particles are impossible to arrive at the given region simultaneously.

Prove that $P_k(t)$ has the Poisson distribution for the given t .

- 1.5. **(The relationship among normal distribution, Poisson distribution, and noncentral χ^2 distribution with 1 degree of freedom)** Prove that the noncentral χ^2 distribution with 1 degree of freedom and non-centrality parameter θ^2 can be factorized into an infinite sum of the product of Poisson distribution with parameter $\theta^2/2$ and the χ^2 with $2i + 1$ degrees of freedom, *i.e.*, if the random variable $X - \theta$ has the standard normal distribution, then the pdf of $Y = X^2$ is as follows

$$p_Y(y) = \frac{1}{2\sqrt{2\pi y}} \exp\left\{-\frac{y + \theta^2}{2}\right\} (e^{\sqrt{y}\theta} + e^{-\sqrt{y}\theta}), \quad y > 0,$$

and the above pdf can be transformed into

$$p_Y(y) = \sum_{i=0}^{\infty} P(R = i) f_{2i+1}(y),$$

where R has the Poisson distribution with parameter $\theta^2/2$, and f_m is the pdf of the χ_m^2 distribution.

- 1.6. Let x_1, \dots, x_n be a set of samples, and let \bar{x} and s^2 denote the sample mean and the variance respectively. Prove that

$$s^2 = \frac{1}{2n(n-1)} \sum_{i=1}^n \sum_{j=1}^n (x_i - x_j)^2.$$

- 1.7. Exemplify the following three situations respectively:

- (a) MLE exists and is unique;
- (b) MLE exists but is not unique;
- (c) MLE does not exist.

- 1.8. Verify the Lagrange's identity: for real numbers a_1, \dots, a_n and b_1, \dots, b_n we have

$$\left(\sum_{i=1}^n a_i^2 \right) \left(\sum_{i=1}^n b_i^2 \right) - \left(\sum_{i=1}^n a_i b_i \right)^2 = \sum_{i=1}^{n-1} \sum_{j=i+1}^n (a_i b_j - a_j b_i)^2.$$

Furthermore, prove that the correlation coefficient is equal to 1 if and only if all the sample points lies in a straight line (Wright, 1992).

- 1.9. Let X_1, \dots, X_n be a set of samples from the uniform distribution $U(0, \theta)$, where $\Theta = \{\theta : \theta > 0\}$. Let Y_n be the largest order statistic (cf. Example 1.2.5). Prove that Y_n is complete.
- 1.10. Let X_1, \dots, X_n be a set of samples from the uniform distribution $U(\theta, \theta + 1]$. For the given $0 \leq p < 1$, let $Z = g(X_1 - p) + p$, where the function $g(y)$ is defined as the maximal integer less than or equal to y . Prove that Z is an unbiased estimator for θ , but the UMVUE does not exist.
- 1.11. Let X_1, \dots, X_n be i.i.d r.v.s from $U(0, 1)$. For $1 \leq i \leq n$, let Y_i be the product of the first i variables, i.e. $Y_i = X_1 \cdots X_i$. Prove that the distribution of X_{k+1} given that $X_1 = x_1, \dots, X_k = x_k$ is a uniform distribution $U(0, x_k)$. Furthermore, prove that $E(Y_n) = 1/2^n$.
- 1.12. Let Y_2 and Y_4 denote the second and the 4-th order statistics of a random sample of size 5 from a distribution of the continuous type having distribution $F(x)$. Compute $P[F(Y_4) - F(Y_2) \geq 0.5]$.
- 1.13. Let X_1, X_2, \dots, X_n be a random sample from a uniform distribution $U(0, \theta)$, $\theta > 0$.
- (a) Find the MLE of θ .
 - (b) Find the UMVUE of θ .
 - (c) Find the method of moments estimate for $\theta(1 - \theta)$.

- 1.14.** X_1, X_2, \dots, X_n is a random sample from $f(x, \theta) = \theta e^{-\theta x} I(x > 0)$.
- (a) Use the factorization theorem to find a sufficient statistic for $\theta \in (0, \infty)$.
 - (b) Find the unbiased minimum variance estimator of $1/\theta$.
- 1.15.** $X = X_1, X_2, \dots, X_n$ is a random sample from $U(0, \theta)$.
- (a) Find the moment estimator and the maximum likelihood estimator for θ .
 - (b) Show that one of these estimators is sufficient for θ .
 - (c) Calculate both estimates for the sample $x = 0.1, 0.2, 0.4, 0.9$, and comment.

1.16. What is meant by an (m, m) exponential family of distributions? What is a curved exponential family? Write a brief account of data reduction by sufficiency in exponential families.

What is meant by an ancillary statistics? What is the conditionality principle of statistical inference?

Let Y_1, \dots, Y_n be independent, identically distributed $N(\mu, \mu^2)$, $\mu > 0$.

Show that $(T_1, T_2) = (\sum_{i=1}^n Y_i/n, \sqrt{\sum_{i=1}^n Y_i^2/n})$ is minimal sufficient and $Z = T_1/T_2$ is ancillary. Explain why inference about μ should be based on the conditional distribution of $V = \sqrt{n} T_2$, given Z , and show that this conditional distribution has density

$$f(v|z; \mu) = \frac{k}{\mu^n} v^{n-1} \exp \left\{ -\frac{1}{2} \left[\frac{v}{\mu} - z\sqrt{n} \right]^2 \right\},$$

for a constant k , not depending on μ .

- 1.17.** Suppose we have the year 2001 results for tennis matches between the 5 top women players. Let r_{ij} be the number of matches in which player i beat player j and let n_{ij} be the number of matches of player i against player j , for $1 \leq i < j \leq 5$. Assume that the (r_{ij}) are independent random variables, and assume

$$r_{ij} \sim Bi(n_{ij}, p_{ij}),$$

and

$$\ln(p_{ij}/(1 - p_{ij})) = \alpha_i - \alpha_j, \quad 1 \leq i < j \leq 5,$$

with $\alpha_5 = 0$.

- Write down the log-likelihood for the unknown parameters, and explain why we need a constraint on $(\alpha_1, \dots, \alpha_5)$.
- How would you find a confidence interval for the probability that player 1 beats player 5?
- How would you find a confidence interval for the probability that player 2 beats player 3?
- How might you extend the model to allow for a grass court/clay court factor?

[You are not expected to find explicit expressions for the maximum likelihood estimators $\hat{\alpha}_i$.]

- 1.18.** (a) Suppose $(Y|U = u)$ has a Poisson distribution, with mean μu , and U has probability density function $f(u)$, where

$$f(u) = \theta^\theta u^{\theta-1} e^{-\theta u} / \Gamma(\theta), \quad \text{for } u \geq 0.$$

Show that

- $E(Y) = \mu, \quad V(Y) = \mu + \mu^2/\theta,$
- Y has frequency function

$$g(y|\mu) = \frac{\Gamma(\theta + y)\mu^y\theta^\theta}{\Gamma(\theta)y!(\mu + \theta)^{\theta+y}}, \quad \text{for } y = 0, 1, 2, \dots.$$

- If (Y_1, \dots, Y_n) are independent observations, and Y_i has frequency function $g(y_i|\mu_i)$, where $\ln \mu_i = \beta x_i$ and x_1, \dots, x_n are given, describe how to estimate β in the case where θ is a known parameter, and derive the asymptotic distribution of your estimator.

- 1.19.** Suppose x_1, \dots, x_n are drawn independently from a mixture normal distribution with the pdf

$$f(x|\boldsymbol{\theta}) = \alpha f_1(x) + (1 - \alpha) f_2(x),$$

where $f_j(x)$ denotes the density for a normal distribution with mean μ_j and common variance σ^2 , and $\boldsymbol{\theta} = (\alpha, \mu_1, \mu_2, \sigma^2)$.

Suppose that we now introduce auxiliary variables Z_{ij} such that

$$Z_{ij} = \begin{cases} 1, & \text{if } x_i \sim N(\mu_j, \sigma^2), \\ 0, & \text{otherwise.} \end{cases}$$

Show that the likelihood function can be written as

$$L(\mathbf{x}|\boldsymbol{\theta}, \mathbf{Z}) = \prod_{i=1}^n \prod_{j=1}^2 (\alpha_j f_j(x_i))^{z_{ij}}.$$

- 1.20.** Let (Ω, \mathcal{A}) be a measurable space, and let μ be a σ -finite measure on \mathcal{A} . Try to show that there must exist a probability measure P on \mathcal{A} , such that $\mu \ll P$ and $P \ll \mu$.
- 1.21.** Let the distribution function $F(x)$ of an r.v. X be right-continuous. Try to show that $E(F(X)) \geq 1/2$, with equality if and only if $F(x)$ is continuous everywhere.
- 1.22.** Let X_1, X_2, \dots, X_n be i.i.d. r.v.s. Let $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ denote the order statistics of X_1, X_2, \dots, X_n . Suppose that $\varphi(y)$ is a Borel measurable function on the real line, and $E(\varphi(X_1))$ is finite. Prove that

$$E(\varphi(X_1) | X_{(1)}) = \frac{1}{n}\varphi(X_{(1)}) + \frac{1}{n} \sum_{i=2}^n E(\varphi(X_{(i)}) | X_{(1)}).$$

- 1.23.** Let the joint density function of X and Y be

$$[\Gamma(\alpha_1)\Gamma(\alpha_2)\theta_1^{\alpha_1}\theta_2^{\alpha_2}]^{-1}x^{\alpha_1-1}y^{\alpha_2-1}\exp\{-\theta_1^{-1}x - \theta_2^{-1}y\}$$

for $x > 0, y > 0, \alpha_1 > 0, \alpha_2 > 0, \theta_1 > 0$, and $\theta_2 > 0$, where α_1 and α_2 are known, θ_1 and θ_2 are parameters.

- (a) Find the UMVUE of $\theta_2^2 - \theta_1$.
 (b) For $\alpha_1 > 1$, find the UMVUE of θ_1^{-1} .

- 1.24.** (a) Let X_1, X_2, \dots, X_n be i.i.d. r.v.s ($n \geq 2$) with the common pdf $\sigma^{-1} \exp\{-\sigma^{-1}(x - \mu)\}$ for $x \geq \mu, -\infty < \mu < \infty$, and $\sigma > 0$, where μ and σ are parameters. Find the MLEs of μ and σ .
 (b) Let $(\mathcal{X}, \mathcal{B}_{\mathcal{X}}, \mathcal{P})$ be the statistical space of X with $\mathcal{P} = \{P_{\theta}; \theta \in \Theta\}$. Both μ_1 and μ_2 are σ -finite measures on $\mathcal{B}_{\mathcal{X}}$, satisfying $dP_{\theta}/d\mu_1 = f_1(x, \theta)$ and $dP_{\theta}/d\mu_2 = f_2(x, \theta)$. Try to show that the two MLEs of θ based on $f_1(x, \theta)$ and $f_2(x, \theta)$, respectively, are the same.

- 1.25.** Suppose that the joint probability density function of (X, Y) is given by

$$P(X = m, Y = n) = \binom{n}{m} p^m (1-p)^{n-m} \lambda^n e^{-\lambda} / n!$$

for $m = 0, 1, \dots, n$ and $n = 0, 1, 2, \dots$, where $0 < p < 1$ and $\lambda > 0$. Find the marginal probability density functions of X and Y .

- 1.26.** Suppose that the r.v. X is symmetric about the zero point, i.e., X and $-X$ have the same distribution with the cdf $F(x)$. Furthermore, suppose that the variance of X is finite. Prove that the variance of X is

$$\int_0^{+\infty} 4x[1 - F(x)]dx.$$

1.27. Suppose that the joint pdf of (X, Y) is given by

$$[2\pi(1-\rho^2)^{1/2}]^{-1} \exp \left\{ -[2(1-\rho^2)]^{-1}(x^2 - 2\rho xy + y^2) \right\} \quad \text{for } -1 < \rho < 1.$$

Let $T = X + Y$. Find the conditional expectation of X given $T = t$, $E(X|T = t)$.

1.28. Suppose that X_1, \dots, X_n are i.i.d. r.v.s with the pdf $e^{-x}(x > 0)$. Let $X_{(1)} = \min\{X_i; 1 \leq i \leq n\}$, and $T_n = \sum_{i=1}^n X_i/n - X_{(1)}$. Show that $X_{(1)}$ is independent of T_n .

1.29. Suppose X_1, \dots, X_n are i.i.d. r.v.s with the Weibull pdf

$$m\eta^{-m}x^{m-1} \exp \left\{ -(x/\eta)^m \right\} \quad \text{for } x > 0,$$

where $m > 0$ and $\eta > 0$ are parameters.

(a) Find the pdf of $\ln X_1$.

(b) Find the moment estimator of $p \hat{=} P(X_1 < L)$ ($L > 0$ is known).