Chapter 2

Fundamentals of Statistics

This chapter discusses some fundamental concepts of mathematical statistics. These concepts are essential for the material in later chapters.

2.1 Populations, Samples, and Models

A typical statistical problem can be described as follows. One or a series of random experiments is performed; some data from the experiment(s) are collected; and our task is to extract information from the data, interpret the results, and draw some conclusions. In this book we do not consider the problem of planning experiments and collecting data, but concentrate on statistical analysis of the data, assuming that the data are given.

A descriptive data analysis can be performed to obtain some summary measures of the data, such as the mean, median, range, standard deviation, etc., and some graphical displays, such as the histogram and box-and-whisker diagram, etc. (see, e.g., Hogg and Tanis (1993)). Although this kind of analysis is simple and requires almost no assumptions, it may not allow us to gain enough insight into the problem. We focus on more sophisticated methods of analyzing data: *statistical inference* and *decision theory*.

2.1.1 Populations and samples

In statistical inference and decision theory, the data set is viewed as a realization or observation of a random element defined on a probability space $(\Omega, \mathcal{F}, P)$ related to the random experiment. The probability measure P is called the *population*. The data set or the random element that produces

the data is called a *sample* from P. The size of the data set is called the *sample size*. A population P is *known* if and only if $P(A)$ is a known value for every event $A \in \mathcal{F}$. In a statistical problem, the population P is at least partially unknown and we would like to deduce some properties of P based on the available sample.

Example 2.1 (Measurement problems). To measure an unknown quantity θ (for example, a distance, weight, or temperature), n measurements, $x_1, ..., x_n$, are taken in an experiment of measuring θ. If θ can be measured without errors, then $x_i = \theta$ for all i; otherwise, each x_i has a possible measurement error. In descriptive data analysis, a few summary measures may be calculated, for example, the *sample mean*

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

and the *sample variance*

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

However, what is the relationship between $\bar{x}$ and θ? Are they close (if not equal) in some sense? The sample variance s^2 is clearly an average of squared deviations of x_i's from their mean. But, what kind of information does s^2 provide? Finally, is it enough to just look at $\bar{x}$ and s^2 for the purpose of measuring θ? These questions cannot be answered in descriptive data analysis.

In statistical inference and decision theory, the data set, $(x_1, ..., x_n)$, is viewed as an outcome of the experiment whose sample space is $\Omega = \mathcal{R}^n$. We usually assume that the n measurements are obtained in n *independent* trials of the experiment. Hence, we can define a random n-vector $X = (X_1, ..., X_n)$ on $\prod_{i=1}^{n} (\mathcal{R}, \mathcal{B}, P)$ whose realization is $(x_1, ..., x_n)$. The population in this problem is P (note that the product probability measure is determined by P) and is at least partially unknown. The random vector X is a sample and n is the sample size. Define

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{2.1}$$

and

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 . \tag{2.2}$$

Then $\bar{X}$ and S^2 are random variables that produce $\bar{x}$ and s^2, respectively. Questions raised previously can be answered if some assumptions are imposed on the population P, which are discussed later. ∎

When the sample $(X_1, ..., X_n)$ has i.i.d. components, which is often the case in applications, the population is determined by the marginal distribution of X_i.

Example 2.2 (Life-time testing problems). Let $x_1, ..., x_n$ be observed lifetimes of some electronic components. Again, in statistical inference and decision theory, $x_1, ..., x_n$ are viewed as realizations of independent random variables $X_1, ..., X_n$. Suppose that the components are of the same type so that it is reasonable to assume that $X_1, ..., X_n$ have a common marginal c.d.f. F. Then the population is F, which is often unknown. A quantity of interest in this problem is $1 - F(t)$ with a $t > 0$, which is the probability that a component does not fail at time t. It is possible that all x_i's are smaller (or larger) than t. Conclusions about $1 - F(t)$ can be drawn based on data $x_1, ..., x_n$ when certain assumptions on F are imposed. ∎

Example 2.3 (Survey problems). A survey is often conducted when one is not able to evaluate all elements in a collection $\mathcal{P} = \{y_1, ..., y_N\}$ containing N values in $\mathcal{R}^k$, where k and N are finite positive integers but N may be very large. Suppose that the quantity of interest is the *population total* $Y = \sum_{i=1}^{N} y_i$. In a survey, a subset $\boldsymbol{s}$ of n elements are selected from $\{1, ..., N\}$ and values y_i, $i \in \boldsymbol{s}$, are obtained. Can we draw some conclusion about Y based on data y_i, $i \in \boldsymbol{s}$?

How do we define some random variables that produce the survey data? First, we need to specify how $\boldsymbol{s}$ is selected. A commonly used probability sampling plan can be described as follows. Assume that every element in $\{1, ..., N\}$ can be selected at most once, i.e., we consider *sampling without replacement*. Let $\mathcal{S}$ be the collection of all subsets of n distinct elements from $\{1, ..., N\}$, $\mathcal{F}_s$ be the collection of all subsets of $\mathcal{S}$, and p be a probability measure on $(\mathcal{S}, \mathcal{F}_s)$. Any $\boldsymbol{s} \in \mathcal{S}$ is selected with probability $p(\boldsymbol{s})$. Note that $p(\boldsymbol{s})$ is a known value whenever $\boldsymbol{s}$ is given. Let $X_1, ..., X_n$ be random variables such that

$$P(X_1 = y_{i_1}, ..., X_n = y_{i_n}) = \frac{p(\boldsymbol{s})}{n!}, \qquad \boldsymbol{s} = \{i_1, ..., i_n\} \in \mathcal{S}. \tag{2.3}$$

Then $(y_i, i \in \boldsymbol{s})$ can be viewed as a realization of the sample $(X_1, ..., X_n)$. If $p(\boldsymbol{s})$ is constant, then the sampling plan is called the *simple random sampling (without replacement)* and $(X_1, ..., X_n)$ is called a *simple random sample*. Although $X_1, ..., X_n$ are identically distributed, they are *not* necessarily independent. Thus, unlike in the previous two examples, the population in this problem may not be specified by the marginal distributions of X_i's. The population is determined by $\mathcal{P}$ and the known selection probability measure p. For this reason, $\mathcal{P}$ is often treated as the population. Conclusions about Y and other characteristics of $\mathcal{P}$ can be drawn based on data y_i, $i \in \boldsymbol{s}$, which are discussed later. ∎

2.1.2 Parametric and nonparametric models

A *statistical model* (a set of assumptions) on the population P in a given problem is often postulated to make the analysis possible or easy. Although testing the correctness of postulated models is part of statistical inference and decision theory, postulated models are often based on knowledge of the problem under consideration.

Definition 2.1. A set of probability measures P_θ on $(\Omega, \mathcal{F})$ indexed by a *parameter* $\theta \in \Theta$ is said to be a *parametric family* if and only if $\Theta \subset \mathcal{R}^d$ for some fixed positive integer d *and* each P_θ is a *known* probability measure when θ is known. The set Θ is called the *parameter space* and d is called its *dimension*. ∎

A *parametric model* refers to the assumption that the population P is in a given parametric family. A parametric family $\{P_\theta : \theta \in \Theta\}$ is said to be *identifiable* if and only if $\theta_1 \neq \theta_2$ and $\theta_i \in \Theta$ imply $P_{\theta_1} \neq P_{\theta_2}$. In most cases an identifiable parametric family can be obtained through reparameterization. Hence, we assume in what follows that every parametric family is identifiable unless otherwise stated.

Let $\mathcal{P}$ be a family of populations and ν be a σ-finite measure on $(\Omega, \mathcal{F})$. If $P \ll \nu$ for all $P \in \mathcal{P}$, then $\mathcal{P}$ is said to be dominated by ν, in which case $\mathcal{P}$ can be identified by the family of densities $\{\frac{dP}{d\nu} : P \in \mathcal{P}\}$ (or $\{\frac{dP_\theta}{d\nu} : \theta \in \Theta\}$ for a parametric family).

Many examples of parametric families can be obtained from Tables 1.1 and 1.2 in §1.3.1. All parametric families from Tables 1.1 and 1.2 are dominated by the counting measure or the Lebesgue measure on $\mathcal{R}$.

Example 2.4 (The k-dimensional normal family). Consider the normal distribution $N_k(\mu, \Sigma)$ given by (1.24) for a fixed positive integer k. An important parametric family in statistics is the family of normal distributions

$$\mathcal{P} = \{N_k(\mu, \Sigma) : \ \mu \in \mathcal{R}^k, \ \Sigma \in \mathcal{M}_k\},$$

where $\mathcal{M}_k$ is a collection of $k \times k$ symmetric positive definite matrices. This family is dominated by the Lebesgue measure on $\mathcal{R}^k$.

In the measurement problem described in Example 2.1, X_i's are often i.i.d. from the $N(\mu, \sigma^2)$ distribution. Hence, we can impose a parametric model on the population, i.e., $P \in \mathcal{P} = \{N(\mu, \sigma^2) : \ \mu \in \mathcal{R}, \ \sigma^2 > 0\}$.

The normal parametric model is perhaps not a good model for the life-time testing problem described in Example 2.2, since clearly $X_i \geq 0$ for all i. In practice, the normal family $\{N(\mu, \sigma^2) : \ \mu \in \mathcal{R}, \ \sigma^2 > 0\}$ can be used for a life-time testing problem if one puts some restrictions on μ and σ so that $P(X_i < 0)$ is negligible. Common parametric models for

life-time testing problems are the exponential model (containing the exponential distributions $E(0, \theta)$ with an unknown parameter θ; see Table 1.2 in §1.3.1), the gamma model (containing the gamma distributions $\Gamma(\alpha, \gamma)$ with unknown parameters α and γ), the log-normal model (containing the log-normal distributions $LN(\mu, \sigma^2)$ with unknown parameters μ and σ), the Weibull model (containing the Weibull distributions $W(\alpha, \theta)$ with unknown parameters α and θ), and any subfamilies of these parametric families (e.g., a family containing the gamma distributions with one known parameter and one unknown parameter).

The normal family is often not a good choice for the survey problem discussed in Example 2.3. ∎

In a given problem, a parametric model is not useful if the dimension of Θ is very high. For example, the survey problem described in Example 2.3 has a natural parametric model, since the population $\mathcal{P}$ can be indexed by the parameter $\theta = (y_1, ..., y_N)$. If there is no restriction on the y-values, however, the dimension of the parameter space is kN, which is usually much larger than the sample size n. If there are some restrictions on the y-values (for example, y_i's are nonnegative integers no larger than a fixed integer m), then the dimension of the parameter space is at most $m + 1$ and the parametric model becomes useful.

A family of probability measures is said to be *nonparametric* if it is not parametric according to Definition 2.1. A *nonparametric model* refers to the assumption that the population P is in a given nonparametric family. There may be almost no assumption on a nonparametric family, for example, the family of all probability measures on $(\mathcal{R}^k, \mathcal{B}^k)$. But in many applications, we may use one or a combination of the following assumptions to form a nonparametric family on $(\mathcal{R}^k, \mathcal{B}^k)$:

(1) The joint c.d.f.'s are continuous.

(2) The joint c.d.f.'s have finite moments of order $\leq$ a fixed integer.

(3) The joint c.d.f.'s have p.d.f.'s (e.g., Lebesgue p.d.f.'s).

(4) $k = 1$ and the c.d.f.'s are symmetric.

For instance, in Example 2.1, we may assume a nonparametric model with symmetric and continuous c.d.f.'s. The symmetry assumption may not be suitable for the population in Example 2.2, but the continuity assumption seems to be reasonable.

In statistical inference and decision theory, methods designed for parametric models are called *parametric methods*, whereas methods designed for nonparametric models are called *nonparametric methods*. However, nonparametric methods are used in a parametric model when parametric methods are not effective, such as when the dimension of the parameter

space is too high (Example 2.3). On the other hand, parametric methods may be applied to a *semi-parametric model*, which is a nonparametric model having a parametric component. Some examples are provided in §5.1.4.

2.1.3 Exponential and location-scale families

In this section, we discuss two types of parametric families that are of special importance in statistical inference and decision theory.

Definition 2.2 (Exponential families). A parametric family $\{P_\theta : \theta \in \Theta\}$ dominated by a σ-finite measure ν on $(\Omega, \mathcal{F})$ is called an *exponential family* if and only if

$$\frac{dP_\theta}{d\nu}(\omega) = \exp\big\{[\eta(\theta)]^\tau T(\omega) - \xi(\theta)\big\}h(\omega), \quad \omega \in \Omega, \tag{2.4}$$

where $\exp\{x\} = e^x$, T is a random p-vector with a fixed positive integer p, η is a function from Θ to $\mathcal{R}^p$, h is a nonnegative Borel function on $(\Omega, \mathcal{F})$, and $\xi(\theta) = \log\big\{\int_\Omega \exp\{[\eta(\theta)]^\tau T(\omega)\}h(\omega)d\nu(\omega)\big\}$. ∎

In Definition 2.2, T and h are functions of ω only, whereas η and ξ are functions of θ only. Ω is usually $\mathcal{R}^k$. The representation (2.4) of an exponential family is not unique. In fact, any transformation $\tilde{\eta}(\theta) = D\eta(\theta)$ with a $p \times p$ nonsingular matrix D gives another representation (with T replaced by $\tilde{T} = (D^\tau)^{-1}T$). A change of the measure that dominates the family also changes the representation. For example, if we define $\lambda(A) = \int_A h d\nu$ for any $A \in \mathcal{F}$, then we obtain an exponential family with densities

$$\frac{dP_\theta}{d\lambda}(\omega) = \exp\big\{[\eta(\theta)]^\tau T(\omega) - \xi(\theta)\big\}. \tag{2.5}$$

In an exponential family, consider the reparameterization $\eta = \eta(\theta)$ and

$$f_\eta(\omega) = \exp\big\{\eta^\tau T(\omega) - \zeta(\eta)\big\}h(\omega), \quad \omega \in \Omega, \tag{2.6}$$

where $\zeta(\eta) = \log\big\{\int_\Omega \exp\{\eta^\tau T(\omega)\}h(\omega)d\nu(\omega)\big\}$. This is the *canonical form* for the family, which is not unique for the reasons discussed previously. The new parameter η is called the *natural parameter*. The new parameter space $\Xi = \{\eta(\theta) : \theta \in \Theta\}$, a subset of $\mathcal{R}^p$, is called the *natural parameter space*. An exponential family in canonical form is called a *natural exponential family*. If there is an open set contained in the natural parameter space of an exponential family, then the family is said to be of *full rank*.

Example 2.5. Let P_θ be the binomial distribution $Bi(\theta, n)$ with parameter θ, where n is a fixed positive integer. Then $\{P_\theta : \theta \in (0,1)\}$ is an

exponential family, since the p.d.f. of P_θ w.r.t. the counting measure is

$$f_\theta(x) = \exp\left\{x \log \tfrac{\theta}{1-\theta} + n \log(1-\theta)\right\} \binom{n}{x} I_{\{0,1,\ldots,n\}}(x)$$

($T(x)=x$, $\eta(\theta)=\log \frac{\theta}{1-\theta}$, $\xi(\theta)=-n\log(1-\theta)$, and $h(x)=\binom{n}{x} I_{\{0,1,\ldots,n\}}(x)$). If we let $\eta = \log \frac{\theta}{1-\theta}$, then $\Xi = \mathcal{R}$ and the family with p.d.f.'s

$$f_\eta(x) = \exp\{x\eta - n\log(1+e^\eta)\} \binom{n}{x} I_{\{0,1,\ldots,n\}}(x)$$

is a natural exponential family of full rank. ∎

Example 2.6. The normal family $\{N(\mu,\sigma^2) : \mu \in \mathcal{R}, \sigma > 0\}$ is an exponential family, since the Lebesgue p.d.f. of $N(\mu,\sigma^2)$ can be written as

$$\frac{1}{\sqrt{2\pi}} \exp\left\{\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2} - \log\sigma\right\}.$$

Hence, $T(x) = (x, -x^2)$, $\eta(\theta) = \left(\frac{\mu}{\sigma^2}, \frac{1}{2\sigma^2}\right)$, $\theta = (\mu,\sigma^2)$, $\xi(\theta) = \frac{\mu^2}{2\sigma^2} + \log\sigma$, and $h(x) = 1/\sqrt{2\pi}$. Let $\eta = (\eta_1,\eta_2) = \left(\frac{\mu}{\sigma^2}, \frac{1}{2\sigma^2}\right)$. Then $\Xi = \mathcal{R}\times(0,\infty)$ and we can obtain a natural exponential family of full rank with $\zeta(\eta) = \eta_1^2/(4\eta_2) + \log(1/\sqrt{2\eta_2})$.

A subfamily of the previous normal family, $\{N(\mu,\mu^2) : \mu \in \mathcal{R}, \mu \neq 0\}$, is also an exponential family with the natural parameter $\eta = \left(\frac{1}{\mu}, \frac{1}{2\mu^2}\right)$ and natural parameter space $\Xi = \{(x,y) : y = 2x^2,\ x \in \mathcal{R},\ y > 0\}$. This exponential family is not of full rank. ∎

For an exponential family, (2.5) implies that there is a nonzero measure λ such that

$$\frac{dP_\theta}{d\lambda}(\omega) > 0 \qquad \text{for all } \omega \text{ and } \theta. \tag{2.7}$$

We can use this fact to show that a family of distributions is not an exponential family. For example, consider the family of uniform distributions, i.e., P_θ is $U(0,\theta)$ with an unknown $\theta \in (0,\infty)$. If $\{P_\theta : \theta \in (0,\infty)\}$ is an exponential family, then from the previous discussion we have a nonzero measure λ such that (2.7) holds. For any $t > 0$, there is a $\theta < t$ such that $P_\theta([t,\infty)) = 0$, which with (2.7) implies that $\lambda([t,\infty)) = 0$. Also, for any $t \leq 0$, $P_\theta((-\infty,t]) = 0$, which with (2.7) implies that $\lambda((-\infty,t]) = 0$. Since t is arbitrary, $\lambda \equiv 0$. This contradiction implies that $\{P_\theta : \theta \in (0,\infty)\}$ cannot be an exponential family.

The reader may verify which of the parametric families from Tables 1.1 and 1.2 are exponential families. As another example, we consider an important exponential family containing multivariate discrete distributions.

Example 2.7 (The multinomial family). Consider an experiment having $k+1$ possible outcomes with p_i as the probability for the ith outcome, $i = 0, 1, ..., k$, $\sum_{i=0}^{k} p_i = 1$. In n independent trials of this experiment, let X_i be the number of trials resulting in the ith outcome, $i = 0, 1, ..., k$. Then the joint p.d.f. (w.r.t. counting measure) of $(X_0, X_1, ..., X_k)$ is

$$f_\theta(x_0, x_1, ..., x_k) = \frac{n!}{x_0! x_1! \cdots x_k!} p_0^{x_0} p_1^{x_1} \cdots p_k^{x_k} I_B(x_0, x_1, ..., x_k),$$

where $B = \{(x_0, x_1, ..., x_k) : \ x_i\text{'s are integers} \geq 0, \ \sum_{i=0}^{k} x_i = n\}$ and $\theta = (p_0, p_1, ..., p_k)$. The distribution of $(X_0, X_1, ..., X_k)$ is called the *multinomial* distribution, which is an extension of the binomial distribution. In fact, the marginal c.d.f. of each X_i is the binomial distribution $Bi(p_i, n)$. Let $\Theta = \{\theta \in \mathcal{R}^{k+1} : \ 0 < p_i < 1, \sum_{i=0}^{k} p_i = 1\}$. The parametric family $\{f_\theta : \ \theta \in \Theta\}$ is called the multinomial family. Let $x = (x_0, x_1, ..., x_k)$, $\eta = (\log p_0, \log p_1, ..., \log p_k)$, and $h(x) = [n!/(x_0! x_1! \cdots x_k!)] I_B(x)$. Then

$$f_\theta(x_0, x_1, ..., x_k) = \exp\{\eta^\tau x\} \, h(x), \qquad x \in \mathcal{R}^{k+1}. \tag{2.8}$$

Hence, the multinomial family is a natural exponential family with natural parameter η. However, representation (2.8) does not provide an exponential family of full rank, since there is no open set of $\mathcal{R}^{k+1}$ contained in the natural parameter space. A reparameterization leads to an exponential family with full rank. Using the fact that $\sum_{i=0}^{k} X_i = n$ and $\sum_{i=0}^{k} p_i = 1$, we obtain that

$$f_\theta(x_0, x_1, ..., x_k) = \exp\{\eta_*^\tau x_* - \zeta(\eta_*)\} \, h(x), \qquad x \in \mathcal{R}^{k+1}, \tag{2.9}$$

where $x_* = (x_1, ..., x_k)$, $\eta_* = (\log(p_1/p_0), ..., \log(p_k/p_0))$, and $\zeta(\eta_*) = -n \log p_0$. The η_*-parameter space is $\mathcal{R}^k$. Hence, the family of densities given by (2.9) is a natural exponential family of full rank. ∎

If $X_1, ..., X_m$ are independent random vectors with p.d.f.'s in exponential families, then the p.d.f. of $(X_1, ..., X_m)$ is again in an exponential family. The following result summarizes some other useful properties of exponential families. Its proof can be found in Lehmann (1986).

Theorem 2.1. Let $\mathcal{P}$ be a natural exponential family given by (2.6).
(i) Let $T = (Y, U)$ and $\eta = (\vartheta, \varphi)$, where Y and ϑ have the same dimension. Then, Y has the p.d.f.

$$f_\eta(y) = \exp\{\vartheta^\tau y - \zeta(\eta)\}$$

w.r.t. a σ-finite measure depending on φ. In particular, T has a p.d.f. in a natural exponential family. Furthermore, the conditional distribution of Y given $U = u$ has the p.d.f. (w.r.t. a σ-finite measure depending on u)

$$f_{\vartheta,u}(y) = \exp\{\vartheta^\tau y - \zeta_u(\vartheta)\},$$

which is in a natural exponential family indexed by ϑ.
(ii) If η_0 is an interior point of the natural parameter space, then the m.g.f. ψ_{η_0} of $P_{\eta_0} \circ T^{-1}$ is finite in a neighborhood of 0 and is given by

$$\psi_{\eta_0}(t) = \exp\{\zeta(\eta_0 + t) - \zeta(\eta_0)\}.$$

Furthermore, if f is a Borel function satisfying $\int |f| dP_{\eta_0} < \infty$, then the function

$$\int f(\omega) \exp\{\eta^\tau T(\omega)\} h(\omega) d\nu(\omega)$$

is infinitely often differentiable in a neighborhood of η_0, and the derivatives may be computed by differentiation under the integral sign. ∎

Using Theorem 2.1(ii) and the result in Example 2.5, we obtain that the m.g.f. of the binomial distribution $Bi(p, n)$ is

$$\begin{aligned}\psi_\eta(t) &= \exp\{n \log(1 + e^{\eta+t}) - n \log(1 + e^\eta)\} \\ &= \left(\frac{1 + e^\eta e^t}{1 + e^\eta}\right)^n \\ &= (1 - p + pe^t)^n,\end{aligned}$$

since $p = e^\eta/(1 + e^\eta)$.

Definition 2.3 (Location-scale families). Let P be a known probability measure on $(\mathcal{R}^k, \mathcal{B}^k)$, $\mathcal{V} \subset \mathcal{R}^k$, and $\mathcal{M}_k$ be a collection of $k \times k$ symmetric positive definite matrices. The family

$$\{P_{(\mu,\Sigma)} : \mu \in \mathcal{V}, \Sigma \in \mathcal{M}_k\} \tag{2.10}$$

is called a *location-scale family* (on $\mathcal{R}^k$), where

$$P_{(\mu,\Sigma)}(B) = P\left(\Sigma^{-1/2}(B - \mu)\right), \quad B \in \mathcal{B}^k,$$

$\Sigma^{-1/2}(B - \mu) = \{\Sigma^{-1/2}(x - \mu) : x \in B\} \subset \mathcal{R}^k$, and $\Sigma^{-1/2}$ is the inverse of the "square root" matrix $\Sigma^{1/2}$ satisfying $\Sigma^{1/2}\Sigma^{1/2} = \Sigma$. The parameters μ and $\Sigma^{1/2}$ are called the location and scale parameters, respectively. ∎

The following are some important examples of location-scale families. The family $\{P_{(\mu,I_k)} : \mu \in \mathcal{R}^k\}$ is called a *location family*, where I_k is the $k \times k$ identity matrix. The family $\{P_{(0,\Sigma)} : \Sigma \in \mathcal{M}_k\}$ is called a *scale family*. In some cases, we consider a location-scale family of the form $\{P_{(\mu,\sigma^2 I_k)} : \mu \in \mathcal{R}^k, \sigma > 0\}$. If $X_1, ..., X_k$ are i.i.d. with a common distribution in the location-scale family $\{P_{(\mu,\sigma^2)} : \mu \in \mathcal{R}, \sigma > 0\}$, then the joint distribution of the vector $(X_1, ..., X_k)$ is in the location-scale family $\{P_{(\mu,\sigma^2 I_k)} : \mu \in \mathcal{V}, \sigma > 0\}$ with $\mathcal{V} = \{(x, ..., x) \in \mathcal{R}^k : x \in \mathcal{R}\}$.

A location-scale family can be generated as follows. Let X be a random k-vector having a distribution P. Then the distribution of $\Sigma^{1/2}X + \mu$ is $P_{(\mu,\Sigma)}$. On the other hand, if X is a random k-vector whose distribution is in the location-scale family (2.10), then the distribution $DX + c$ is also in the same family, provided that $D\mu + c \in \mathcal{V}$ and $D\Sigma D^\tau \in \mathcal{M}_k$.

Let F be the c.d.f. of P. Then the c.d.f. of $P_{(\mu,\Sigma)}$ is $F\left(\Sigma^{-1/2}(x-\mu)\right)$, $x \in \mathcal{R}^k$. If F has a Lebesgue p.d.f. f, then the Lebesgue p.d.f. of $P_{(\mu,\Sigma)}$ is $\mathrm{Det}(\Sigma^{-1/2}) f\left(\Sigma^{-1/2}(x-\mu)\right)$, $x \in \mathcal{R}^k$ (Proposition 1.8).

Many families of distributions in Table 1.2 (§1.3.1) are location, scale, or location-scale families. For example, the family of exponential distributions $E(a,\theta)$ is a location-scale family on $\mathcal{R}$ with location parameter a and scale parameter θ; the family of uniform distributions $U(0,\theta)$ is a scale family on $\mathcal{R}$ with a scale parameter θ. The k-dimensional normal family discussed in Example 2.4 is a location-scale family on $\mathcal{R}^k$.

2.2 Statistics, Sufficiency, and Completeness

Let us assume now that our data set is a realization of a sample X (a random vector) from an unknown population P on a probability space.

2.2.1 Statistics and their distributions

A measurable function of X, $T(X)$, is called a *statistic* if $T(X)$ is a known value whenever X is known, i.e., the function T is a known function. Statistical analyses are based on various statistics, for various purposes. Of course, X itself is a statistic, but it is a trivial statistic. The range of a nontrivial statistic $T(X)$ is usually simpler than that of X. For example, X may be a random n-vector and $T(X)$ may be a random p-vector with a p much smaller than n. This is desired since $T(X)$ simplifies the original data.

From a probabilistic point of view, the "information" within the statistic $T(X)$ concerning the unknown distribution of X is contained in the σ-field $\sigma(T(X))$. To see this, assume that S is any other statistic for which $\sigma(S(X)) = \sigma(T(X))$. Then, by Lemma 1.2, S is a measurable function of T, and T is a measurable function of S. Thus, once the value of S (or T) is known, so is the value of T (or S). That is, it is not the particular values of a statistic that contain the information, but the generated σ-field of the statistic. Values of a statistic may be important for other reasons.

Note that $\sigma(T(X)) \subset \sigma(X)$ and the two σ-fields are the same if and only if T is one-to-one. Usually $\sigma(T(X))$ simplifies $\sigma(X)$, i.e., a statistic provides a "reduction" of the σ-field.

Any $T(X)$ is a random element. If the distribution of X is unknown, then the distribution of T may also be unknown, although T is a known function. Finding the form of the distribution of T is one of the major problems in statistical inference and decision theory. Since T is a transformation of X, tools we learn in Chapter 1 for transformations may be useful in finding the distribution or an approximation to the distribution of $T(X)$.

Example 2.8. Let $X_1, ..., X_n$ be i.i.d. random variables having a common distribution P and $X = (X_1, ..., X_n)$. The sample mean $\bar{X}$ and sample variance S^2 defined in (2.1) and (2.2), respectively, are two commonly used statistics. Can we find the joint or the marginal distributions of $\bar{X}$ and S^2? It depends on how much we know about P.

First, let us consider the moments of $\bar{X}$ and S^2. Assume that P has a finite mean denoted by μ. Then

$$E\bar{X} = \mu.$$

If P is in a parametric family $\{P_\theta : \theta \in \Theta\}$, then $E\bar{X} = \int x dP_\theta = \mu(\theta)$ for some function $\mu(\cdot)$. Even if the form of μ is known, $\mu(\theta)$ may still be unknown when θ is unknown. Assume now that P has a finite variance denoted by σ^2. Then

$$\mathrm{Var}(\bar{X}) = \sigma^2/n,$$

which equals $\sigma^2(\theta)/n$ for some function $\sigma^2(\cdot)$ if P is in a parametric family. With a finite $\sigma^2 = \mathrm{Var}(X_1)$, we can also obtain that

$$ES^2 = \sigma^2.$$

With a finite $E|X_1|^3$, we can obtain $E(\bar{X})^3$ and $\mathrm{Cov}(\bar{X}, S^2)$, and with a finite $E|X_1|^4$, we can obtain $\mathrm{Var}(S^2)$ (exercise).

Next, consider the distribution of $\bar{X}$. If P is in a parametric family, we can often find the distribution of $\bar{X}$. See Example 1.20 and some exercises in §1.6. For example, $\bar{X}$ is $N(\mu, \sigma^2/n)$ if P is $N(\mu, \sigma^2)$; $n\bar{X}$ has the gamma distribution $\Gamma(n, \theta)$ if P is the exponential distribution $E(0, \theta)$. If P is not in a parametric family, then it is usually hard to find the exact form of the distribution of $\bar{X}$. One can, however, use the CLT (§1.5.4) to obtain an approximation to the distribution of $\bar{X}$. Applying Corollary 1.2 (for the case of $k = 1$), we obtain that

$$\sqrt{n}(\bar{X} - \mu) \to_d N(0, \sigma^2)$$

and, by (1.100), the distribution of $\bar{X}$ can be approximated by $N(\mu, \sigma^2/n)$, where μ and σ^2 are the mean and variance of P, respectively, and are assumed to be finite.

Compared to $\bar{X}$, the distribution of S^2 is harder to obtain. Assuming that P is $N(\mu, \sigma^2)$, one can show that $(n-1)S^2/\sigma^2$ has the chi-square

distribution χ^2_{n-1} (see Example 2.18). An approximate distribution for S^2 can be obtained from the approximate joint distribution of $\bar{X}$ and S^2 discussed next.

Under the assumption that P is $N(\mu, \sigma^2)$, it can be shown that $\bar{X}$ and S^2 are independent (Example 2.18). Hence, the joint distribution of $(\bar{X}, S^2)$ is the product of the marginal distributions of $\bar{X}$ and S^2 given in the previous discussion. Without the normality assumption, an approximate joint distribution can be obtained as follows. Assume again that $\mu = EX_1$, $\sigma^2 = \text{Var}(X_1)$, and $E|X_1|^4$ are finite. Let $Y_i = (X_i - \mu, (X_i - \mu)^2)$, $i = 1, ..., n$. Then $Y_1, ..., Y_n$ are i.i.d. random 2-vectors with $EY_1 = (0, \sigma^2)$ and variance-covariance matrix

$$\Sigma = \begin{pmatrix} \sigma^2 & E(X_1 - \mu)^3 \\ E(X_1 - \mu)^3 & E(X_1 - \mu)^4 - \sigma^4 \end{pmatrix}.$$

Note that $\bar{Y} = n^{-1} \sum_{i=1}^n Y_i = (\bar{X} - \mu, \tilde{S}^2)$, where $\tilde{S}^2 = n^{-1} \sum_{i=1}^n (X_i - \mu)^2$. Applying the CLT (Corollary 1.2) to Y_i's, we obtain that

$$\sqrt{n}(\bar{X} - \mu, \tilde{S}^2 - \sigma^2) \to_d N_2(0, \Sigma).$$

Since

$$S^2 = \frac{n}{n-1} \left[\tilde{S}^2 - (\bar{X} - \mu)^2 \right]$$

and $\bar{X} \to_{a.s.} \mu$ (the SLLN, Theorem 1.13), an application of Slutsky's theorem (Theorem 1.11) leads to

$$\sqrt{n}(\bar{X} - \mu, S^2 - \sigma^2) \to_d N_2(0, \Sigma). \quad \blacksquare$$

Example 2.9 (Order statistics). Let $X = (X_1, ..., X_n)$ with i.i.d. random components and let $X_{(i)}$ be the ith smallest value of $X_1, ..., X_n$. The statistics $X_{(1)}, ..., X_{(n)}$ are called the *order statistics*, which is a set of very useful statistics in addition to the sample mean and variance in the previous example. Suppose that X_i has a c.d.f. F having a Lebesgue p.d.f. f. Then the joint Lebesgue p.d.f. of $X_{(1)}, ..., X_{(n)}$ is

$$g(x_1, x_2, ..., x_n) = \begin{cases} n! f(x_1) f(x_2) \cdots f(x_n) & x_1 < x_2 < \cdots < x_n \\ 0 & \text{otherwise.} \end{cases}$$

The joint Lebesgue p.d.f. of $X_{(i)}$ and $X_{(j)}$, $1 \le i < j \le n$, is

$$g_{i,j}(x, y) = \begin{cases} \frac{n![F(x)]^{i-1}[F(y)-F(x)]^{j-i-1}[1-F(y)]^{n-j} f(x) f(y)}{(i-1)!(j-i-1)!(n-j)!} & x < y \\ 0 & \text{otherwise} \end{cases}$$

and the Lebesgue p.d.f. of $X_{(i)}$ is

$$g_i(x) = \frac{n!}{(i-1)!(n-i)!} [F(x)]^{i-1} [1 - F(x)]^{n-i} f(x). \quad \blacksquare$$

2.2.2 Sufficiency and minimal sufficiency

Having discussed the reduction of the σ-field $\sigma(X)$ by using a statistic $T(X)$, we now ask whether such a reduction results in any loss of information concerning the unknown population. If a statistic $T(X)$ is fully as informative as the original sample X, then statistical analyses can be done using $T(X)$ that is simpler than X. The next concept describes what we mean by fully informative.

Definition 2.4 (Sufficiency). Let X be a sample from an unknown population $P \in \mathcal{P}$, where $\mathcal{P}$ is a family of populations. A statistic $T(X)$ is said to be *sufficient* for $P \in \mathcal{P}$ (or for $\theta \in \Theta$ when $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is a parametric family) if and only if the conditional distribution of X given T is *known* (does not depend on P or θ). ∎

Definition 2.4 can be interpreted as follows. Once we observe X and compute a sufficient statistic $T(X)$, the original data X do not contain any further information concerning the unknown population P (since its conditional distribution is unrelated to P) and can be discarded. A sufficient statistic $T(X)$ contains all information about P contained in X (see Exercise 36 in §3.6 for an interpretation of this from another viewpoint) and provides a reduction of the data if T is not one-to-one. Thus, one of the questions raised in Example 2.1 can be answered as follows: it is enough to just look at $\bar{x}$ and s^2 for the problem of measuring θ if $(\bar{X}, S^2)$ is sufficient for P (or θ when θ is the only unknown parameter).

The concept of sufficiency depends on the given family $\mathcal{P}$. If T is sufficient for $P \in \mathcal{P}$, then T is also sufficient for $P \in \mathcal{P}_0 \subset \mathcal{P}$ but not necessarily sufficient for $P \in \mathcal{P}_1 \supset \mathcal{P}$.

Example 2.10. Suppose that $X = (X_1, ..., X_n)$ and $X_1, ..., X_n$ are i.i.d. from the binomial distribution with the p.d.f. (w.r.t. the counting measure)

$$f_\theta(z) = \theta^z(1-\theta)^{1-z} I_{\{0,1\}}(z), \quad z \in \mathcal{R}, \quad \theta \in (0,1).$$

For any realization x of X, x is a sequence of n ones and zeros. Consider the statistic $T(X) = \sum_{i=1}^n X_i$, which is the number of ones in X. Before showing that T is sufficient, we can intuitively argue that T contains all information about θ, since θ is the probability of an occurrence of a one in x. Given $T = t$ (the number of ones in x), what is left in the data set x is the redundant information about the positions of t ones. Since the random variables are discrete, it is not difficult to compute the conditional distribution of X given $T = t$. Note that

$$P(X = x|T = t) = \frac{P(X = x, T = t)}{P(T = t)}$$

and $P(T=t)=\binom{n}{t}\theta^t(1-\theta)^{n-t}I_{\{0,1,...,n\}}(t)$. Let x_i be the ith component of x. If $t\neq\sum_{i=1}^n x_i$, then $P(X=x,T=t)=0$. If $t=\sum_{i=1}^n x_i$, then

$$P(X=x,T=t)=\prod_{i=1}^n P(X_i=x_i)=\theta^t(1-\theta)^{n-t}\prod_{i=1}^n I_{\{0,1\}}(x_i).$$

Let $B_t=\{(x_1,...,x_n):\ x_i=0,1,\ \sum_{i=1}^n x_i=t\}$. Then

$$P(X=x|T=t)=\frac{1}{\binom{n}{t}}I_{B_t}(x)$$

is a known p.d.f. This shows that $T(X)$ is sufficient for $\theta\in(0,1)$, according to Definition 2.4 with the family $\{f_\theta:\theta\in(0,1)\}$. ∎

Finding a sufficient statistic by means of the definition is not convenient since it involves guessing a statistic T that might be sufficient and computing the conditional distribution of X given $T=t$. For families of populations having p.d.f.'s, a simple way of finding sufficient statistics is to use the factorization theorem. We first prove the following lemma.

Lemma 2.1. If a family $\mathcal{P}$ is dominated by a σ-finite measure, then $\mathcal{P}$ is dominated by a probability measure $Q=\sum_{i=1}^\infty c_iP_i$, where c_i's are nonnegative constants with $\sum_{i=1}^\infty c_i=1$ and $P_i\in\mathcal{P}$.

Proof. Assume that $\mathcal{P}$ is dominated by a finite measure ν (the case of σ-finite ν is left as an exercise). Let $\mathcal{P}_0$ be the family of all measures of the form $\sum_{i=1}^\infty c_iP_i$, where $P_i\in\mathcal{P}$, $c_i\geq 0$, and $\sum_{i=1}^\infty c_i=1$. Then, it suffices to show that there is a $Q\in\mathcal{P}_0$ such that $Q(A)=0$ implies $P(A)=0$ for all $P\in\mathcal{P}_0$. Let $\mathcal{C}$ be the class of events C for which there exists $P\in\mathcal{P}_0$ such that $P(C)>0$ and $dP/d\nu>0$ a.e. ν on C. Then there exists a sequence $\{C_i\}\subset\mathcal{C}$ such that $\nu(C_i)\to\sup_{C\in\mathcal{C}}\nu(C)$. Let C_0 be the union of all C_i's and $Q=\sum_{i=1}^\infty c_iP_i$, where P_i is the probability measure corresponding to C_i. Then $C_0\in\mathcal{C}$ (exercise). Suppose now that $Q(A)=0$. Let $P\in\mathcal{P}_0$ and $B=\{x:dP/d\nu>0\}$. Since $Q(A\cap C_0)=0$, $\nu(A\cap C_0)=0$ and $P(A\cap C_0)=0$. Then $P(A)=P(A\cap C_0^c\cap B)$. If $P(A\cap C_0^c\cap B)>0$, then $\nu(C_0\cup(A\cap C_0^c\cap B))>\nu(C_0)$, which contradicts $\nu(C_0)=\sup_{C\in\mathcal{C}}\nu(C)$ since $A\cap C_0^c\cap B$ and therefore $C_0\cup(A\cap C_0^c\cap B)$ is in $\mathcal{C}$. Thus, $P(A)=0$ for all $P\in\mathcal{P}_0$. ∎

Theorem 2.2 (The factorization theorem). Suppose that X is a sample from $P\in\mathcal{P}$ and $\mathcal{P}$ is a family of probability measures on $(\mathcal{R}^n,\mathcal{B}^n)$ dominated by a σ-finite measure ν. Then $T(X)$ is sufficient for $P\in\mathcal{P}$ if and only if there are nonnegative Borel functions h (which does not depend on P) on $(\mathcal{R}^n,\mathcal{B}^n)$ and $g_{_P}$ (which depends on P) on the range of T such that

$$\frac{dP}{d\nu}(x)=g_{_P}\big(T(x)\big)h(x). \tag{2.11}$$

Proof. (i) Suppose that T is sufficient for $P \in \mathcal{P}$. Then, for any $A \in \mathcal{B}^n$, $P(A|T)$ does not depend on P. Let Q be the probability measure in Lemma 2.1. By Fubini's theorem and the result in Exercise 35 of §1.6,

$$\begin{aligned} Q(A \cap B) &= \sum_{j=1}^{\infty} c_j P_j(A \cap B) \\ &= \sum_{j=1}^{\infty} c_j \int_B P(A|T) dP_j \\ &= \int_B \sum_{j=1}^{\infty} c_j P(A|T) dP_j \\ &= \int_B P(A|T) dQ \end{aligned}$$

for any $B \in \sigma(T)$. Hence, $P(A|T) = E_Q(I_A|T)$ a.s. Q, where $E_Q(I_A|T)$ denotes the conditional expectation of I_A given T w.r.t. Q. Let $g_P(T)$ be the Radon-Nikodym derivative dP/dQ on the space $(\mathcal{R}^n, \sigma(T), Q)$. From Propositions 1.7 and 1.10,

$$\begin{aligned} P(A) &= \int P(A|T) dP \\ &= \int E_Q(I_A|T) g_P(T) dQ \\ &= \int E_Q[I_A g_P(T)|T] dQ \\ &= \int_A g_P(T) \frac{dQ}{d\nu} d\nu \end{aligned}$$

for any $A \in \mathcal{B}^n$. Hence, (2.11) holds with $h = dQ/d\nu$.
(ii) Suppose that (2.11) holds. Then

$$\frac{dP}{dQ} = \frac{dP}{d\nu} \Big/ \sum_{i=1}^{\infty} c_i \frac{dP_i}{d\nu} = g_P(T) \Big/ \sum_{i=1}^{\infty} g_{P_i}(T) \quad \text{a.s. } Q, \tag{2.12}$$

where the second equality follows from the result in Exercise 35 of §1.6. Let $A \in \sigma(X)$ and $P \in \mathcal{P}$. The sufficiency of T follows from

$$P(A|T) = E_Q(I_A|T) \quad \text{a.s. } P, \tag{2.13}$$

where $E_Q(I_A|T)$ is given in part (i) of the proof. This is because $E_Q(I_A|T)$ does not vary with $P \in \mathcal{P}$, and result (2.13) and Theorem 1.7 imply that the conditional distribution of X given T is determined by $E_Q(I_A|T)$, $A \in \sigma(X)$. By the definition of conditional probability, (2.13) follows from

$$\int_B I_A dP = \int_B E_Q(I_A|T) dP \tag{2.14}$$

for any $B \in \sigma(T)$. Let $B \in \sigma(T)$. By (2.12), dP/dQ is a Borel function of T. Then, by Proposition 1.7(i), Proposition 1.10(vi), and the definition of the conditional expectation, the right-hand side of (2.14) is equal to

$$\int_B E_Q(I_A|T)\frac{dP}{dQ}dQ = \int_B E_Q\left(I_A\frac{dP}{dQ}\middle|T\right)dQ = \int_B I_A\frac{dP}{dQ}dQ,$$

which equals the left-hand side of (2.14). This proves (2.14) for any $B \in \sigma(T)$ and completes the proof. ∎

If $\mathcal{P}$ is an exponential family with p.d.f.'s given by (2.4) and $X(\omega) = \omega$, then we can apply Theorem 2.2 with $g_\theta(t) = \exp\{[\eta(\theta)]^\tau t - \xi(\theta)\}$ and conclude that T is a sufficient statistic for $\theta \in \Theta$. In Example 2.10 the joint distribution of X is in an exponential family with $T(X) = \sum_{i=1}^n X_i$. Hence, we can conclude that T is sufficient for $\theta \in (0,1)$ without computing the conditional distribution of X given T.

Example 2.11 (Truncation families). Let $\phi(x)$ be a positive Borel function on $(\mathcal{R}, \mathcal{B})$ such that $\int_a^b \phi(x)dx < \infty$ for any a and b, $-\infty < a < b < \infty$. Let $\theta = (a,b)$, $\Theta = \{(a,b) \in \mathcal{R}^2 : a < b\}$, and

$$f_\theta(x) = c(\theta)\phi(x)I_{(a,b)}(x),$$

where $c(\theta) = \left[\int_a^b \phi(x)dx\right]^{-1}$. Then $\{f_\theta : \theta \in \Theta\}$, called a truncation family, is a parametric family dominated by the Lebesgue measure on $\mathcal{R}$. Let $X_1, ..., X_n$ be i.i.d. random variables having the p.d.f. f_θ. Then the joint p.d.f. of $X = (X_1, ..., X_n)$ is

$$\prod_{i=1}^n f_\theta(x_i) = [c(\theta)]^n I_{(a,\infty)}(x_{(1)})I_{(-\infty,b)}(x_{(n)})\prod_{i=1}^n \phi(x_i), \tag{2.15}$$

where $x_{(i)}$ is the ith smallest value of $x_1, ..., x_n$. Let $T(X) = (X_{(1)}, X_{(n)})$, $g_\theta(t_1, t_2) = [c(\theta)]^n I_{(a,\infty)}(t_1)I_{(-\infty,b)}(t_2)$, and $h(x) = \prod_{i=1}^n \phi(x_i)$. By (2.15) and Theorem 2.2, $T(X)$ is sufficient for $\theta \in \Theta$. ∎

Example 2.12 (Order statistics). Let $X = (X_1, ..., X_n)$ and $X_1, ..., X_n$ be i.i.d. random variables having a distribution $P \in \mathcal{P}$, where $\mathcal{P}$ is the family of distributions on $\mathcal{R}$ having Lebesgue p.d.f.'s. Let $X_{(1)}, ..., X_{(n)}$ be the order statistics given in Example 2.9. Note that the joint p.d.f. of X is

$$f(x_1)\cdots f(x_n) = f(x_{(1)})\cdots f(x_{(n)}).$$

Hence, $T(X) = (X_{(1)}, ..., X_{(n)})$ is sufficient for $P \in \mathcal{P}$. The order statistics can be shown to be sufficient even when $\mathcal{P}$ is not dominated by any σ-finite measure, but Theorem 2.2 is not applicable (see Exercise 31 in §2.6). ∎

There are many sufficient statistics for a given family $\mathcal{P}$. In fact, if T is a sufficient statistic and $T = \psi(S)$, where ψ is measurable and S is another statistic, then S is sufficient. This is obvious from Theorem 2.2 if the population has a p.d.f., but it can be proved directly from Definition 2.4 (Exercise 25). For instance, in Example 2.10, $(\sum_{i=1}^m X_i, \sum_{i=m+1}^n X_i)$ is sufficient for θ, where m is any fixed integer between 1 and n. If T is sufficient and $T = \psi(S)$ with a measurable ψ that is not one-to-one, then $\sigma(T) \subset \sigma(S)$ and T is more useful than S, since T provides a further reduction of the data (or σ-field) without loss of information. Is there a sufficient statistic that provides "maximal" reduction of the data?

Before introducing the next concept, we need the following notation. If a statement holds except for outcomes in an event A satisfying $P(A) = 0$ for all $P \in \mathcal{P}$, then we say that the statement holds a.s. $\mathcal{P}$.

Definition 2.5 (Minimal sufficiency). Let T be a sufficient statistic for $P \in \mathcal{P}$. T is called a *minimal sufficient* statistic if and only if, for any other statistic S sufficient for $P \in \mathcal{P}$, there is a measurable function ψ such that $T = \psi(S)$ a.s. $\mathcal{P}$. ∎

If both T and S are minimal sufficient statistics, then by definition there is a one-to-one measurable function ψ such that $T = \psi(S)$ a.s. $\mathcal{P}$. Hence, the minimal sufficient statistic is unique in the sense that two statistics that are one-to-one measurable functions of each other can be treated as one statistic.

Example 2.13. Let $X_1, ..., X_n$ be i.i.d. random variables from P_θ, the uniform distribution $U(\theta, \theta + 1)$, $\theta \in \mathcal{R}$. Suppose that $n > 1$. The joint Lebesgue p.d.f. of $(X_1, ..., X_n)$ is

$$f_\theta(x) = \prod_{i=1}^n I_{(\theta,\theta+1)}(x_i) = I_{(x_{(n)}-1, x_{(1)})}(\theta), \quad x = (x_1, ..., x_n) \in \mathcal{R}^n,$$

where $x_{(i)}$ denotes the ith smallest value of $x_1, ..., x_n$. By Theorem 2.2, $T = (X_{(1)}, X_{(n)})$ is sufficient for θ. Note that

$$x_{(1)} = \sup\{\theta : f_\theta(x) > 0\} \quad \text{and} \quad x_{(n)} = 1 + \inf\{\theta : f_\theta(x) > 0\}.$$

If $S(X)$ is a statistic sufficient for θ, then by Theorem 2.2, there are Borel functions h and g_θ such that $f_\theta(x) = g_\theta(S(x))h(x)$. For x with $h(x) > 0$,

$$x_{(1)} = \sup\{\theta : g_\theta(S(x)) > 0\} \quad \text{and} \quad x_{(n)} = 1 + \inf\{\theta : g_\theta(S(x)) > 0\}.$$

Hence, there is a measurable function ψ such that $T(x) = \psi(S(x))$ when $h(x) > 0$. Since $h > 0$ a.s. $\mathcal{P}$, we conclude that T is minimal sufficient. ∎

Minimal sufficient statistics exist under weak assumptions, e.g., $\mathcal{P}$ contains distributions on $\mathcal{R}^k$ dominated by a σ-finite measure (Bahadur, 1957). The next theorem provides some useful tools for finding minimal sufficient statistics.

Theorem 2.3. Let $\mathcal{P}$ be a family of distributions on $\mathcal{R}^k$.
(i) Suppose that $\mathcal{P}_0 \subset \mathcal{P}$ and a.s. $\mathcal{P}_0$ implies a.s. $\mathcal{P}$. If T is sufficient for $P \in \mathcal{P}$ and minimal sufficient for $P \in \mathcal{P}_0$, then T is minimal sufficient for $P \in \mathcal{P}$.
(ii) Suppose that $\mathcal{P}$ contains p.d.f.'s $f_0, f_1, f_2, ...$, w.r.t. a σ-finite measure. Let $f_\infty(x) = \sum_{i=0}^{\infty} c_i f_i(x)$, where $c_i > 0$ for all i and $\sum_{i=0}^{\infty} c_i = 1$, and let $T_i(X) = f_i(x)/f_\infty(x)$ when $f_\infty(x) > 0$, $i = 0, 1, 2, ...$ Then $T(X) = (T_0, T_1, T_2, ...)$ is minimal sufficient for $P \in \mathcal{P}$. Furthermore, if $\{x : f_i(x) > 0\} \subset \{x : f_0(x) > 0\}$ for all i, then we may replace f_∞ by f_0, in which case $T(X) = (T_1, T_2, ...)$ is minimal sufficient for $P \in \mathcal{P}$.
(iii) Suppose that $\mathcal{P}$ contains p.d.f.'s f_P w.r.t. a σ-finite measure and that there exists a sufficient statistic $T(X)$ such that, for any possible values x and y of X, $f_P(x) = f_P(y)\phi(x, y)$ for all P implies $T(x) = T(y)$, where ϕ is a measurable function. Then $T(X)$ is minimal sufficient for $P \in \mathcal{P}$.
Proof. (i) If S is sufficient for $P \in \mathcal{P}$, then it is also sufficient for $P \in \mathcal{P}_0$ and, therefore, $T = \psi(S)$ a.s. $\mathcal{P}_0$ holds for a measurable function ψ. The result follows from the assumption that a.s. $\mathcal{P}_0$ implies a.s. $\mathcal{P}$.
(ii) Note that $f_\infty > 0$ a.s. $\mathcal{P}$. Let $g_i(T) = T_i$, $i = 0, 1, 2, ...$ Then $f_i(x) = g_i(T(x))f_\infty(x)$ a.s. $\mathcal{P}$. By Theorem 2.2, T is sufficient for $P \in \mathcal{P}$. Suppose that $S(X)$ is another sufficient statistic. By Theorem 2.2, there are Borel functions h and $\tilde{g}_i$ such that $f_i(x) = \tilde{g}_i(S(x))h(x)$, $i = 0, 1, 2, ...$ Then $T_i(x) = \tilde{g}_i(S(x))/\sum_{j=0}^{\infty} c_j \tilde{g}_j(S(x))$ for x's satisfying $f_\infty(x) > 0$. By Definition 2.5, T is minimal sufficient for $P \in \mathcal{P}$. The proof for the case where f_∞ is replaced by f_0 is the same.
(iii) From Bahadur (1957), there exists a minimal sufficient statistic $S(X)$. The result follows if we can show that $T(X) = \psi(S(X))$ a.s. $\mathcal{P}$ for a measurable function ψ. By Theorem 2.2, there are Borel functions g_P and h such that $f_P(x) = g_P(S(x))h(x)$ for all P. Let $A = \{x : h(x) = 0\}$. Then $P(A) = 0$ for all P. For x and y such that $S(x) = S(y)$, $x \notin A$ and $y \notin A$,

$$\begin{aligned} f_P(x) &= g_P(S(x))h(x) \\ &= g_P(S(y))h(x)h(y)/h(y) \\ &= f_P(y)h(x)/h(y) \end{aligned}$$

for all P. Hence $T(x) = T(y)$. This shows that there is a function ψ such that $T(x) = \psi(S(x))$ except for $x \in A$. It remains to show that ψ is measurable. Since S is minimal sufficient, $g(T(X)) = S(X)$ a.s. $\mathcal{P}$ for a measurable function g. Hence g is one-to-one and $\psi = g^{-1}$. The measurability of ψ follows from Theorem 3.9 in Parthasarathy (1967). ∎

Example 2.14. Let $\mathcal{P} = \{f_\theta : \theta \in \Theta\}$ be an exponential family with p.d.f.'s f_θ given by (2.4) and $X(\omega) = \omega$. Suppose that there exists $\Theta_0 = \{\theta_0, \theta_1, ..., \theta_p\} \subset \Theta$ such that the vectors $\eta_i = \eta(\theta_i) - \eta(\theta_0)$, $i = 1, ..., p$, are linearly independent in $\mathcal{R}^p$. (This is true if the family is of full rank.) We have shown that $T(X)$ is sufficient for $\theta \in \Theta$. We now show that T is in fact minimal sufficient for $\theta \in \Theta$. Let $\mathcal{P}_0 = \{f_\theta : \theta \in \Theta_0\}$. Note that the set $\{x : f_\theta(x) > 0\}$ does not depend on θ. It follows from Theorem 2.3(ii) with $f_\infty = f_{\theta_0}$ that

$$S(X) = \left(\exp\{\eta_1^\tau T(x) - \xi_1\}, ..., \exp\{\eta_p^\tau T(x) - \xi_p\}\right)$$

is minimal sufficient for $\theta \in \Theta_0$, where $\xi_i = \xi(\theta_i) - \xi(\theta_0)$. Since η_i's are linearly independent, there is a one-to-one measurable function ψ such that $T(X) = \psi(S(X))$ a.s. $\mathcal{P}_0$. Hence, T is minimal sufficient for $\theta \in \Theta_0$. It is easy to see that a.s. $\mathcal{P}_0$ implies a.s. $\mathcal{P}$. Thus, by Theorem 2.3(i), T is minimal sufficient for $\theta \in \Theta$. ∎

The results in Examples 2.13 and 2.14 can also be proved by using Theorem 2.3(iii) (Exercise 32).

The sufficiency (and minimal sufficiency) depends on the postulated family $\mathcal{P}$ of populations (statistical models). Hence, it may not be a useful concept if the proposed statistical model is wrong or at least one has some doubts about the correctness of the proposed model. From the examples in this section and some exercises in §2.6, one can find that for a wide variety of models, statistics such as $\bar{X}$ in (2.1), S^2 in (2.2), $(X_{(1)}, X_{(n)})$ in Example 2.11, and the order statistics in Example 2.9 are sufficient. Thus, using these statistics for data reduction and summarization does not lose any information when the true model is one of those models but we do not know exactly which model is correct.

2.2.3 Complete statistics

A statistic $V(X)$ is said to be *ancillary* if its distribution does not depend on the population P and *first-order ancillary* if $E[V(X)]$ is independent of P. A trivial ancillary statistic is the constant statistic $V(X) \equiv c \in \mathcal{R}$. If $V(X)$ is a nontrivial ancillary statistic, then $\sigma(V(X)) \subset \sigma(X)$ is a nontrivial σ-field that does not contain any information about P. Hence, if $S(X)$ is a statistic and $V(S(X))$ is a nontrivial ancillary statistic, it indicates that $\sigma(S(X))$ contains a nontrivial σ-field that does not contain any information about P and, hence, the "data" $S(X)$ may be further reduced. A sufficient statistic T appears to be most successful in reducing the data if no nonconstant function of T is ancillary or even first-order ancillary. This leads to the following concept of completeness.

Definition 2.6 (Completeness). A statistic $T(X)$ is said to be *complete* for $P \in \mathcal{P}$ if and only if, for any Borel f, $E[f(T)] = 0$ for all $P \in \mathcal{P}$ implies $f(T) = 0$ a.s. $\mathcal{P}$. T is said to be *boundedly complete* if and only if the previous statement holds for any bounded Borel f. ∎

A complete statistic is boundedly complete. If T is complete (or boundedly complete) and $S = \psi(T)$ for a measurable ψ, then S is complete (or boundedly complete). Intuitively, a complete and sufficient statistic should be minimal sufficient, which was shown by Lehmann and Scheffé (1950) and Bahadur (1957) (see Exercise 48). However, a minimal sufficient statistic is not necessarily complete; for example, the minimal sufficient statistic $(X_{(1)}, X_{(n)})$ in Example 2.13 is not complete (Exercise 47).

Proposition 2.1. If P is in an exponential family of full rank with p.d.f.'s given by (2.6), then $T(X)$ is complete and sufficient for $\eta \in \Xi$.
Proof. We have shown that T is sufficient. Suppose that there is a function f such that $E[f(T)] = 0$ for all $\eta \in \Xi$. By Theorem 2.1(i),

$$\int f(t) \exp\{\eta^\tau t - \zeta(\eta)\} d\lambda = 0 \quad \text{for all } \eta \in \Xi,$$

where λ is a measure on $(\mathcal{R}^p, \mathcal{B}^p)$. Let η_0 be an interior point of Ξ. Then

$$\int f_+(t) e^{\eta^\tau t} d\lambda = \int f_-(t) e^{\eta^\tau t} d\lambda \quad \text{for all } \eta \in N(\eta_0), \tag{2.16}$$

where $N(\eta_0) = \{\eta \in \mathcal{R}^p : \|\eta - \eta_0\| < \epsilon\}$ for some $\epsilon > 0$. In particular,

$$\int f_+(t) e^{\eta_0^\tau t} d\lambda = \int f_-(t) e^{\eta_0^\tau t} d\lambda = c.$$

If $c = 0$, then $f = 0$ a.e. λ. If $c > 0$, then $c^{-1} f_+(t) e^{\eta_0^\tau t}$ and $c^{-1} f_-(t) e^{\eta_0^\tau t}$ are p.d.f.'s w.r.t. λ and (2.16) implies that their m.g.f.'s are the same in a neighborhood of 0. By Theorem 1.6(ii), $c^{-1} f_+(t) e^{\eta_0^\tau t} = c^{-1} f_-(t) e^{\eta_0^\tau t}$, i.e., $f = f_+ - f_- = 0$ a.e. λ. Hence T is complete. ∎

Proposition 2.1 is useful for finding a complete and sufficient statistic when the family of distributions is an exponential family of full rank.

Example 2.15. Suppose that $X_1, ..., X_n$ are i.i.d. random variables having the $N(\mu, \sigma^2)$ distribution, $\mu \in \mathcal{R}$, $\sigma > 0$. From Example 2.6, the joint p.d.f. of $X_1, ..., X_n$ is $(2\pi)^{-n/2} \exp\{\eta_1 T_1 + \eta_2 T_2 - n\zeta(\eta)\}$, where $T_1 = \sum_{i=1}^n X_i$, $T_2 = -\sum_{i=1}^n X_i^2$, and $\eta = (\eta_1, \eta_2) = \left(\frac{\mu}{\sigma^2}, \frac{1}{2\sigma^2}\right)$. Hence, the family of distributions for $X = (X_1, ..., X_n)$ is a natural exponential family of full rank ($\Xi = \mathcal{R} \times (0, \infty)$). By Proposition 2.1, $T(X) = (T_1, T_2)$ is complete and sufficient for η. Since there is a one-to-one correspondence between η

and $\theta = (\mu, \sigma^2)$, T is also complete and sufficient for θ. It can be shown that any one-to-one measurable function of a complete and sufficient statistic is also complete and sufficient (exercise). Thus, $(\bar{X}, S^2)$ is complete and sufficient for θ, where $\bar{X}$ and S^2 are the sample mean and variance given by (2.1) and (2.2), respectively. ∎

The following examples show how to find a complete statistic for a non-exponential family.

Example 2.16. Let $X_1, ..., X_n$ be i.i.d. random variables from P_θ, the uniform distribution $U(0, \theta)$, $\theta > 0$. The largest order statistic, $X_{(n)}$, is complete and sufficient for $\theta \in (0, \infty)$. The sufficiency of $X_{(n)}$ follows from the fact that the joint Lebesgue p.d.f. of $X_1, ..., X_n$ is $\theta^{-n} I_{(0,\theta)}(x_{(n)})$. From Example 2.9, $X_{(n)}$ has the Lebesgue p.d.f. $(nx^{n-1}/\theta^n) I_{(0,\theta)}(x)$ on $\mathcal{R}$. Let f be a Borel function on $[0, \infty)$ such that $E[f(X_{(n)})] = 0$ for all $\theta > 0$. Then

$$\int_0^\theta f(x) x^{n-1} dx = 0 \quad \text{for all } \theta > 0.$$

Let $G(\theta)$ be the left-hand side of the previous equation. Applying the result of differentiation of an integral (see, e.g., Royden (1968, §5.3)), we obtain that $G'(\theta) = f(\theta)\theta^{n-1}$ a.e. m_+, where m_+ is the Lebesgue measure on $([0, \infty), \mathcal{B}_{[0,\infty)})$. Since $G(\theta) = 0$ for all $\theta > 0$, $f(\theta)\theta^{n-1} = 0$ a.e. m_+ and, hence, $f(x) = 0$ a.e. m_+. Therefore, $X_{(n)}$ is complete and sufficient for $\theta \in (0, \infty)$. ∎

Example 2.17. In Example 2.12, we showed that the order statistics $T(X) = (X_{(1)}, ..., X_{(n)})$ of i.i.d. random variables $X_1, ..., X_n$ is sufficient for $P \in \mathcal{P}$, where $\mathcal{P}$ is the family of distributions on $\mathcal{R}$ having Lebesgue p.d.f.'s. We now show that $T(X)$ is also complete for $P \in \mathcal{P}$. Let $\mathcal{P}_0$ be the family of Lebesgue p.d.f.'s of the form

$$f(x) = C(\theta_1, ..., \theta_n) \exp\{-x^{2n} + \theta_1 x + \theta_2 x^2 + \cdots + \theta_n x^n\},$$

where $\theta_j \in \mathcal{R}$ and $C(\theta_1, ..., \theta_n)$ is a normalizing constant such that $\int f(x)dx = 1$. Then $\mathcal{P}_0 \subset \mathcal{P}$ and $\mathcal{P}_0$ is an exponential family of full rank. Note that the joint distribution of $X = (X_1, ..., X_n)$ is also in an exponential family of full rank. Thus, by Proposition 2.1, $U = (U_1, ..., U_n)$ is a complete statistic for $P \in \mathcal{P}_0$, where $U_j = \sum_{i=1}^n X_i^j$. Since a.s. $\mathcal{P}_0$ implies a.s. $\mathcal{P}$, $U(X)$ is also complete for $P \in \mathcal{P}$.

The result follows if we can show that there is a one-to-one correspondence between $T(X)$ and $U(X)$. Let $V_1 = \sum_{i=1}^n X_i$, $V_2 = \sum_{i<j} X_i X_j$, $V_3 = \sum_{i<j<k} X_i X_j X_k$,..., $V_n = X_1 \cdots X_n$. From the identities

$$U_k - V_1 U_{k-1} + V_2 U_{k-2} - \cdots + (-1)^{k-1} V_{k-1} U_1 + (-1)^k k V_k = 0,$$

$k = 1, ..., n$, there is a one-to-one correspondence between $U(X)$ and $V(X) = (V_1, ..., V_n)$. From the identity

$$(t - X_1) \cdots (t - X_n) = t^n - V_1 t^{n-1} + V_2 t^{n-2} - \cdots + (-1)^n V_n,$$

there is a one-to-one correspondence between $V(X)$ and $T(X)$. This completes the proof and, hence, $T(X)$ is sufficient and complete for $P \in \mathcal{P}$. In fact, both $U(X)$ and $V(X)$ are sufficient and complete for $P \in \mathcal{P}$. ∎

The relationship between an ancillary statistic and a complete and sufficient statistic is characterized in the following result.

Theorem 2.4 (Basu's theorem). Let V and T be two statistics of X from a population $P \in \mathcal{P}$. If V is ancillary and T is boundedly complete and sufficient for $P \in \mathcal{P}$, then V and T are independent w.r.t. any $P \in \mathcal{P}$.

Proof. Let B be an event on the range of V. Since V is ancillary, $P(V^{-1}(B))$ is a constant. Since T is sufficient, $E[I_B(V)|T]$ is a function of T (independent of P). Since $E\{E[I_B(V)|T] - P(V^{-1}(B))\} = 0$ for all $P \in \mathcal{P}$, $P(V^{-1}(B)|T) = E[I_B(V)|T] = P(V^{-1}(B))$ a.s. $\mathcal{P}$, by the bounded completeness of T. Let A be an event on the range of T. Then, $P(T^{-1}(A) \cap V^{-1}(B)) = E\{E[I_A(T)I_B(V)|T]\} = E\{I_A(T)E[I_B(V)|T]\} = E\{I_A(T)P(V^{-1}(B))\} = P(T^{-1}(A))P(V^{-1}(B))$. Hence T and V are independent w.r.t. any $P \in \mathcal{P}$. ∎

Basu's theorem is useful in proving the independence of two statistics.

Example 2.18. Suppose that $X_1, ..., X_n$ are i.i.d. random variables having the $N(\mu, \sigma^2)$ distribution, with $\mu \in \mathcal{R}$ and a known $\sigma > 0$. It can be easily shown that the family $\{N(\mu, \sigma^2) : \mu \in \mathcal{R}\}$ is an exponential family of full rank with natural parameter $\eta = \mu/\sigma^2$. By Proposition 2.1, the sample mean $\bar{X}$ in (2.1) is complete and sufficient for η (and μ). Let S^2 be the sample variance given by (2.2). Since $S^2 = (n-1)^{-1} \sum_{i=1}^n (Z_i - \bar{Z})^2$, where $Z_i = X_i - \mu$ is $N(0, \sigma^2)$ and $\bar{Z} = n^{-1} \sum_{i=1}^n Z_i$, S^2 is an ancillary statistic (σ^2 is known). By Basu's theorem, $\bar{X}$ and S^2 are independent w.r.t. $N(\mu, \sigma^2)$ with $\mu \in \mathcal{R}$. Since σ^2 is arbitrary, $\bar{X}$ and S^2 are independent w.r.t. $N(\mu, \sigma^2)$ for any $\mu \in \mathcal{R}$ and $\sigma^2 > 0$.

Using the independence of $\bar{X}$ and S^2, we now show that $(n-1)S^2/\sigma^2$ has the chi-square distribution χ^2_{n-1}. Note that

$$n \left(\frac{\bar{X} - \mu}{\sigma} \right)^2 + \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2.$$

From the properties of the normal distributions, $n(\bar{X} - \mu)^2/\sigma^2$ has the chi-square distribution χ^2_1 with the m.g.f. $(1 - 2t)^{-1/2}$ and $\sum_{i=1}^n (X_i - \mu)^2/\sigma^2$

has the chi-square distribution χ^2_n with the m.g.f. $(1-2t)^{-n/2}$, $t < 1/2$. By the independence of $\bar{X}$ and S^2, the m.g.f. of $(n-1)S^2/\sigma^2$ is

$$(1-2t)^{-n/2}/(1-2t)^{-1/2} = (1-2t)^{-(n-1)/2}$$

for $t < 1/2$. This is the m.g.f. of the chi-square distribution χ^2_{n-1} and, therefore, the result follows. ∎

2.3 Statistical Decision Theory

In this section, we describe some basic elements in statistical decision theory. More developments are given in later chapters.

2.3.1 Decision rules, loss functions, and risks

Let X be a sample from a population $P \in \mathcal{P}$. A *statistical decision* is an *action* that we take after we observe X, for example, a conclusion about P or a characteristic of P. Throughout this section, we use $\mathbb{A}$ to denote the set of allowable actions. Let $\mathcal{F}_{\mathbb{A}}$ be a σ-field on $\mathbb{A}$. Then the measurable space $(\mathbb{A}, \mathcal{F}_{\mathbb{A}})$ is called the *action space*. Let $\mathfrak{X}$ be the range of X and $\mathcal{F}_{\mathfrak{X}}$ be a σ-field on $\mathfrak{X}$. A *decision rule* is a measurable function (a statistic) T from $(\mathfrak{X}, \mathcal{F}_{\mathfrak{X}})$ to $(\mathbb{A}, \mathcal{F}_{\mathbb{A}})$. If a decision rule T is chosen, then we take the action $T(X) \in \mathbb{A}$ whence X is observed.

The construction or selection of decision rules cannot be done without any criterion about the performance of decision rules. In *statistical decision theory*, we set a criterion using a *loss function* L, which is a function from $\mathcal{P} \times \mathbb{A}$ to $[0, \infty)$ and is Borel on $(\mathbb{A}, \mathcal{F}_{\mathbb{A}})$ for each fixed $P \in \mathcal{P}$. If $X = x$ is observed and our decision rule is T, then our "loss" (in making a decision) is $L(P, T(x))$. The average loss for the decision rule T, which is called the *risk* of T, is defined to be

$$R_T(P) = E[L(P, T(X))] = \int_{\mathfrak{X}} L(P, T(x)) dP_X(x). \tag{2.17}$$

The loss and risk functions are denoted by $L(\theta, a)$ and $R_T(\theta)$ if $\mathcal{P}$ is a parametric family indexed by θ. A decision rule with small loss is preferred. But it is difficult to compare $L(P, T_1(X))$ and $L(P, T_2(X))$ for two decision rules, T_1 and T_2, since both of them are random. For this reason, the risk function (2.17) is introduced and we compare two decision rules by comparing their risks. A rule T_1 is *as good as* another rule T_2 if and only if

$$R_{T_1}(P) \le R_{T_2}(P) \quad \text{for any } P \in \mathcal{P}, \tag{2.18}$$

and is *better* than T_2 if and only if (2.18) holds and $R_{T_1}(P) < R_{T_2}(P)$ for at least one $P \in \mathcal{P}$. Two decision rules T_1 and T_2 are *equivalent* if and only

if $R_{T_1}(P) = R_{T_2}(P)$ for all $P \in \mathcal{P}$. If there is a decision rule T_* that is as good as any other rule in $\Im$, a class of allowable decision rules, then T_* is said to be $\Im$-*optimal* (or optimal if $\Im$ contains all possible rules).

Example 2.19. Consider the measurement problem in Example 2.1. Suppose that we need a decision on the value of $\theta \in \mathcal{R}$, based on the sample $X = (X_1, ..., X_n)$. If Θ is all possible values of θ, then it is reasonable to consider the action space $(\mathbb{A}, \mathcal{F}_{\mathbb{A}}) = (\Theta, \mathcal{B}_\Theta)$. An example of a decision rule is $T(X) = \bar{X}$, the sample mean defined by (2.1). A common loss function in this problem is the *squared error loss* $L(P, a) = (\theta - a)^2$, $a \in \mathbb{A}$. Then the loss for the decision rule $\bar{X}$ is the squared deviation between $\bar{X}$ and θ. Assuming that the population has mean μ and variance $\sigma^2 < \infty$, we obtain the following risk function for $\bar{X}$:

$$\begin{aligned} R_{\bar{X}}(P) &= E(\theta - \bar{X})^2 \\ &= (\theta - E\bar{X})^2 + E(E\bar{X} - \bar{X})^2 \\ &= (\theta - E\bar{X})^2 + \mathrm{Var}(\bar{X}) \qquad (2.19) \\ &= (\mu - \theta)^2 + \frac{\sigma^2}{n}, \qquad (2.20) \end{aligned}$$

where result (2.20) follows from the results for the moments of $\bar{X}$ in Example 2.8. If θ is in fact the mean of the population, then the first term on the right-hand side of (2.20) is 0 and the risk is an increasing function of the population variance σ^2 and a decreasing function of the sample size n.

Consider another decision rule $T_1(X) = (X_{(1)} + X_{(n)})/2$. However, $R_{T_1}(P)$ does not have an explicit form if there is no further assumption on the population P. Suppose that $P \in \mathcal{P}$. Then, for some $\mathcal{P}$, $\bar{X}$ (or T_1) is better than T_1 (or $\bar{X}$) (exercise), whereas for some $\mathcal{P}$, neither $\bar{X}$ nor T_1 is better than the other.

A different loss function may also be considered. For example, $L(P, a) = |\theta - a|$, which is called the *absolute error loss*. However, $R_{\bar{X}}(P)$ and $R_{T_1}(P)$ do not have explicit forms unless $\mathcal{P}$ is of some specific form. ∎

The problem in Example 2.19 is a special case of a general problem called *estimation*, in which the action space is the set of all possible values of a population characteristic ϑ to be estimated. In an estimation problem, a decision rule T is called an *estimator* and result (2.19) holds with $\theta = \vartheta$ and $\bar{X}$ replaced by any estimator with a finite variance. The following example describes another type of important problem called *hypothesis testing*.

Example 2.20. Let $\mathcal{P}$ be a family of distributions, $\mathcal{P}_0 \subset \mathcal{P}$, and $\mathcal{P}_1 = \{P \in \mathcal{P} : P \notin \mathcal{P}_0\}$. A hypothesis testing problem can be formulated as that of deciding which of the following two statements is true:

$$H_0 : \ P \in \mathcal{P}_0 \qquad \text{versus} \qquad H_1 : \ P \in \mathcal{P}_1. \qquad (2.21)$$

Here, H_0 is called the *null hypothesis* and H_1 is called the *alternative hypothesis*. The action space for this problem contains only two elements, i.e., $\mathbb{A} = \{0, 1\}$, where 0 is the action of accepting H_0 and 1 is the action of rejecting H_0. A decision rule is called a *test*. Since a test $T(X)$ is a function from $\mathfrak{X}$ to $\{0, 1\}$, $T(X)$ must have the form $I_C(X)$, where $C \in \mathcal{F}_{\mathfrak{X}}$ is called the *rejection region* or *critical region* for testing H_0 versus H_1.

A simple loss function for this problem is the 0-1 loss: $L(P, a) = 0$ if a correct decision is made and 1 if an incorrect decision is made, i.e., $L(P, j) = 0$ for $P \in \mathcal{P}_j$ and $L(P, j) = 1$ otherwise, $j = 0, 1$. Under this loss, the risk is

$$R_T(P) = \begin{cases} P(T(X) = 1) = P(X \in C) & P \in \mathcal{P}_0 \\ P(T(X) = 0) = P(X \notin C) & P \in \mathcal{P}_1. \end{cases}$$

See Figure 2.2 on page 127 for an example of a graph of $R_T(\theta)$ for some T and P in a parametric family.

The 0-1 loss implies that the loss for two types of incorrect decisions (accepting H_0 when $P \in \mathcal{P}_1$ and rejecting H_0 when $P \in \mathcal{P}_0$) are the same. In some cases, one might assume unequal losses: $L(P, j) = 0$ for $P \in \mathcal{P}_j$, $L(P, 0) = c_0$ when $P \in \mathcal{P}_1$, and $L(P, 1) = c_1$ when $P \in \mathcal{P}_0$. ∎

In the following example the decision problem is neither an estimation nor a testing problem. Another example is given in Exercise 93 in §2.6.

Example 2.21. A hazardous toxic waste site requires clean-up when the true chemical concentration θ in the contaminated soil is higher than a given level $\theta_0 \geq 0$. Because of the limitation in resources, we would like to spend our money and efforts more in those areas that pose high risk to public health. In a particular area where soil samples are obtained, we would like to take one of these three actions: a complete clean-up (a_1), a partial clean-up (a_2), and no clean-up (a_3). Then $\mathbb{A} = \{a_1, a_2, a_3\}$. Suppose that the cost for a complete clean-up is c_1 and for a partial clean-up is $c_2 < c_1$; the risk to public health is $c_3(\theta - \theta_0)$ if $\theta > \theta_0$ and 0 if $\theta \leq \theta_0$; a complete clean-up can reduce the toxic concentration to an amount $\leq \theta_0$, whereas a partial clean-up can only reduce a fixed amount of the toxic concentration, i.e., the chemical concentration becomes $\theta - t$ after a partial clean-up, where t is a known constant. Then the loss function is given by

$L(\theta, a)$	a_1	a_2	a_3
$\theta \leq \theta_0$	c_1	c_2	0
$\theta_0 < \theta \leq \theta_0 + t$	c_1	c_2	$c_3(\theta - \theta_0)$
$\theta > \theta_0 + t$	c_1	$c_2 + c_3(\theta - \theta_0 - t)$	$c_3(\theta - \theta_0)$

The risk function can be calculated once the decision rule is specified. We discuss this example again in Chapter 4. ∎

Sometimes it is useful to consider *randomized decision rules*. Examples are given in §2.3.2, Chapters 4 and 6. A randomized decision rule is a function δ on $\mathcal{X} \times \mathcal{F}_{\mathbb{A}}$ such that, for every $A \in \mathcal{F}_{\mathbb{A}}$, $\delta(\cdot, A)$ is a Borel function and, for every $x \in \mathcal{X}$, $\delta(x, \cdot)$ is a probability measure on $(\mathbb{A}, \mathcal{F}_{\mathbb{A}})$. To choose an action in $\mathbb{A}$ when a randomized rule δ is used, we need to simulate a pseudorandom element of $\mathbb{A}$ according to $\delta(x, \cdot)$. Thus, an alternative way to describe a randomized rule is to specify the method of simulating the action from $\mathbb{A}$ for each $x \in \mathcal{X}$. If $\mathbb{A}$ is a subset of a Euclidean space, for example, then the result in Theorem 1.7(ii) can be applied. Also, see §7.2.3.

A nonrandomized decision rule T previously discussed can be viewed as a special randomized decision rule with $\delta(x, \{a\}) = I_{\{a\}}(T(x))$, $a \in \mathbb{A}$, $x \in \mathcal{X}$. Another example of a randomized rule is a discrete distribution $\delta(x, \cdot)$ assigning probability $p_j(x)$ to a nonrandomized decision rule $T_j(x)$, $j = 1, 2, ...$, in which case the rule δ can be equivalently defined as a rule taking value $T_j(x)$ with probability $p_j(x)$. See Exercise 64 for an example.

The loss function for a randomized rule δ is defined as

$$L(P, \delta, x) = \int_{\mathbb{A}} L(P, a) d\delta(x, a),$$

which reduces to the same loss function we discussed when δ is a nonrandomized rule. The risk of a randomized rule δ is then

$$R_\delta(P) = E[L(P, \delta, X)] = \int_{\mathcal{X}} \int_{\mathbb{A}} L(P, a) d\delta(x, a) dP_X(x). \tag{2.22}$$

2.3.2 Admissibility and optimality

Consider a given decision problem with a given loss $L(P, a)$.

Definition 2.7 (Admissibility). Let $\Im$ be a class of decision rules (randomized or nonrandomized). A decision rule $T \in \Im$ is called $\Im$-*admissible* (or admissible when $\Im$ contains all possible rules) if and only if there does not exist any $S \in \Im$ that is better than T (in terms of the risk). ∎

If a decision rule T is inadmissible, then there exists a rule better than T. Thus, T should not be used in principle. However, an admissible decision rule is not necessarily good. For example, in an estimation problem a silly estimator $T(X) \equiv$ a constant may be admissible (Exercise 71).

The relationship between the admissibility and the optimality defined in §2.3.1 can be described as follows. If T_* is $\Im$-optimal, then it is $\Im$-admissible; if T_* is $\Im$-optimal and T_0 is $\Im$-admissible, then T_0 is also $\Im$-optimal and is equivalent to T_*; if there are two $\Im$-admissible rules that are not equivalent, then there does not exist any $\Im$-optimal rule.

Suppose that we have a sufficient statistic $T(X)$ for $P \in \mathcal{P}$. Intuitively, our decision rule should be a function of T, based on the discussion in §2.2.2. This is not true in general, but the following result indicates that this is true if randomized decision rules are allowed.

Proposition 2.2. Suppose that $\mathbb{A}$ is a subset of $\mathcal{R}^k$. Let $T(X)$ be a sufficient statistic for $P \in \mathcal{P}$ and let δ_0 be a decision rule. Then

$$\delta_1(t, A) = E[\delta_0(X, A)|T = t], \tag{2.23}$$

which is a randomized decision rule depending only on T, is equivalent to δ_0 if $R_{\delta_0}(P) < \infty$ for any $P \in \mathcal{P}$.

Proof. Note that δ_1 defined by (2.23) is a decision rule since δ_1 does not depend on the unknown P by the sufficiency of T. From (2.22),

$$\begin{aligned} R_{\delta_1}(P) &= E\left\{\int_{\mathbb{A}} L(P,a)d\delta_1(X,a)\right\} \\ &= E\left\{E\left[\int_{\mathbb{A}} L(P,a)d\delta_0(X,a)\middle|T\right]\right\} \\ &= E\left\{\int_{\mathbb{A}} L(P,a)d\delta_0(X,a)\right\} \\ &= R_{\delta_0}(P), \end{aligned}$$

where the proof of the second equality is left to the reader. ∎

Note that Proposition 2.2 does not imply that δ_0 is inadmissible. Also, if δ_0 is a nonrandomized rule,

$$\delta_1(t, A) = E[I_A(\delta_0(X))|T = t] = P(\delta_0(X) \in A|T = t)$$

is still a randomized rule, unless $\delta_0(X) = h(T(X))$ a.s. P for some Borel function h (Exercise 75). Hence, Proposition 2.2 does not apply to situations where randomized rules are not allowed.

The following result tells us when nonrandomized rules are all we need and when decision rules that are not functions of sufficient statistics are inadmissible.

Theorem 2.5. Suppose that $\mathbb{A}$ is a convex subset of $\mathcal{R}^k$ and that for any $P \in \mathcal{P}$, $L(P, a)$ is a convex function of a.
(i) Let δ be a randomized rule satisfying $\int_{\mathbb{A}} \|a\| d\delta(x, a) < \infty$ for any $x \in \mathcal{X}$ and let $T_1(x) = \int_{\mathbb{A}} a d\delta(x, a)$. Then $L(P, T_1(x)) \le L(P, \delta, x)$ (or $L(P, T_1(x)) < L(P, \delta, x)$ if L is strictly convex in a) for any $x \in \mathcal{X}$ and $P \in \mathcal{P}$.
(ii) (Rao-Blackwell theorem). Let T be a sufficient statistic for $P \in \mathcal{P}$, $T_0 \in \mathcal{R}^k$ be a nonrandomized rule satisfying $E\|T_0\| < \infty$, and $T_1 = E[T_0(X)|T]$. Then $R_{T_1}(P) \le R_{T_0}(P)$ for any $P \in \mathcal{P}$. If L is strictly convex in a and T_0 is not a function of T, then T_0 is inadmissible. ∎

The proof of Theorem 2.5 is an application of Jensen's inequality (1.47) and is left to the reader.

The concept of admissibility helps us to eliminate some decision rules. However, usually there are still too many rules left after the elimination of some rules according to admissibility and sufficiency. Although one is typically interested in a $\Im$-optimal rule, frequently it does not exist, if $\Im$ is either too large or too small. The following examples are illustrations.

Example 2.22. Let $X_1, ..., X_n$ be i.i.d. random variables from a population $P \in \mathcal{P}$ that is the family of populations having finite mean μ and variance σ^2. Consider the estimation of μ ($\mathbb{A} = \mathcal{R}$) under the squared error loss. It can be shown that if we let $\Im$ be the class of all possible estimators, then there is no $\Im$-optimal rule (exercise). Next, let $\Im_1$ be the class of all linear functions in $X = (X_1, ..., X_n)$, i.e., $T(X) = \sum_{i=1}^n c_i X_i$ with known $c_i \in \mathcal{R}$, $i = 1, ..., n$. It follows from (2.19) and the discussion after Example 2.19 that

$$R_T(P) = \mu^2 \left(\sum_{i=1}^n c_i - 1 \right)^2 + \sigma^2 \sum_{i=1}^n c_i^2. \tag{2.24}$$

We now show that there does not exist $T_* = \sum_{i=1}^n c_i^* X_i$ such that $R_{T_*}(P) \le R_T(P)$ for any $P \in \mathcal{P}$ and $T \in \Im_1$. If there is such a T_*, then $(c_1^*, ..., c_n^*)$ is a minimum of the function of $(c_1, ..., c_n)$ on the right-hand side of (2.24). Then $c_1^*, ..., c_n^*$ must be the same and equal to $\mu^2/(\sigma^2 + n\mu^2)$, which depends on P. Hence T_* is not a statistic. This shows that there is no $\Im_1$-optimal rule.

Consider now a subclass $\Im_2 \subset \Im_1$ with c_i's satisfying $\sum_{i=1}^n c_i = 1$. From (2.24), $R_T(P) = \sigma^2 \sum_{i=1}^n c_i^2$ if $T \in \Im_2$. Minimizing $\sigma^2 \sum_{i=1}^n c_i^2$ subject to $\sum_{i=1}^n c_i = 1$ leads to an optimal solution of $c_i = n^{-1}$ for all i. Thus, the sample mean $\bar{X}$ is $\Im_2$-optimal.

There may not be any optimal rule if we consider a small class of decision rules. For example, if $\Im_3$ contains all the rules in $\Im_2$ except $\bar{X}$, then one can show that there is no $\Im_3$-optimal rule. ∎

Example 2.23. Assume that the sample X has the binomial distribution $Bi(\theta, n)$ with an unknown $\theta \in (0, 1)$ and a fixed integer $n > 1$. Consider the hypothesis testing problem described in Example 2.20 with $H_0 : \theta \in (0, \theta_0]$ versus $H_1 : \theta \in (\theta_0, 1)$, where $\theta_0 \in (0, 1)$ is a fixed value. Suppose that we are only interested in the following class of nonrandomized decision rules: $\Im = \{T_j : j = 0, 1, ..., n-1\}$, where $T_j(X) = I_{\{j+1,...,n\}}(X)$. From Example 2.20, the risk function for T_j under the 0-1 loss is

$$R_{T_j}(\theta) = P(X > j) I_{(0,\theta_0]}(\theta) + P(X \le j) I_{(\theta_0,1)}(\theta).$$

For any integers k and j, $0 \le k < j \le n-1$,

$$R_{T_j}(\theta) - R_{T_k}(\theta) = \begin{cases} -P(k < X \le j) < 0 & 0 < \theta \le \theta_0 \\ P(k < X \le j) > 0 & \theta_0 < \theta < 1. \end{cases}$$

Hence, neither T_j nor T_k is better than the other. This shows that every T_j is $\Im$-admissible and, thus, there is no $\Im$-optimal rule. ∎

In view of the fact that an optimal rule often does not exist, statisticians adopt the following two approaches to choose a decision rule. The first approach is to define a class $\Im$ of decision rules that have some desirable properties (statistical and/or nonstatistical) and then try to find the best rule in $\Im$. In Example 2.22, for instance, any estimator T in $\Im_2$ has the property that T is linear in X and $E[T(X)] = \mu$. In a general estimation problem, we can use the following concept.

Definition 2.8 (Unbiasedness). In an estimation problem, the *bias* of an estimator $T(X)$ of a real-valued parameter ϑ of the unknown population is defined to be $b_T(P) = E[T(X)] - \vartheta$ (which is denoted by $b_T(\theta)$ when P is in a parametric family indexed by θ). An estimator $T(X)$ is said to be *unbiased* for ϑ if and only if $b_T(P) = 0$ for any $P \in \mathcal{P}$. ∎

Thus, $\Im_2$ in Example 2.22 is the class of unbiased estimators linear in X. In Chapter 3, we discuss how to find a $\Im$-optimal estimator when $\Im$ is the class of unbiased estimators or unbiased estimators linear in X.

Another class of decision rules can be defined after we introduce the concept of *invariance*.

Definition 2.9 Let X be a sample from $P \in \mathcal{P}$.
(i) A class $\mathcal{G}$ of one-to-one transformations of X is called a *group* if and only if $g_i \in \mathcal{G}$ implies $g_1 \circ g_2 \in \mathcal{G}$ and $g_i^{-1} \in \mathcal{G}$.
(ii) We say that $\mathcal{P}$ is *invariant* under $\mathcal{G}$ if and only if $\bar{g}(P_X) = P_{g(X)}$ is a one-to-one transformation from $\mathcal{P}$ onto $\mathcal{P}$ for each $g \in \mathcal{G}$.
(iii) A decision problem is said to be *invariant* if and only if $\mathcal{P}$ is invariant under $\mathcal{G}$ and the loss $L(P, a)$ is invariant in the sense that, for every $g \in \mathcal{G}$ and every $a \in \mathbb{A}$, there exists a unique $g(a) \in \mathbb{A}$ such that $L(P_X, a) = L\left(P_{g(X)}, g(a)\right)$. (Note that $g(X)$ and $g(a)$ are different functions in general.)
(iv) A decision rule $T(x)$ is said to be *invariant* if and only if, for every $g \in \mathcal{G}$ and every $x \in \mathcal{X}$, $T(g(x)) = g(T(x))$. ∎

Invariance means that our decision is not affected by one-to-one transformations of data.

In a problem where the distribution of X is in a location-scale family

$\mathcal{P}$ on $\mathcal{R}^k$, we often consider location-scale transformations of data X of the form $g(X) = AX + c$, where $c \in \mathcal{C} \subset \mathcal{R}^k$ and $A \in \mathcal{T}$, a class of invertible $k \times k$ matrices. Assume that if $A_i \in \mathcal{T}$, $i = 1, 2$, then $A_i^{-1} \in \mathcal{T}$ and $A_1 A_2 \in \mathcal{T}$, and that if $c_i \in \mathcal{C}$, $i = 1, 2$, then $-c_i \in \mathcal{C}$ and $Ac_1 + c_2 \in \mathcal{C}$ for any $A \in \mathcal{T}$. Then the collection of all transformations is a group. A special case is given in the following example.

Example 2.24. Let X have i.i.d. components from a population in a location family $\mathcal{P} = \{P_\mu : \mu \in \mathcal{R}\}$. Consider the location transformation $g_c(X) = X + cJ_k$, where $c \in \mathcal{R}$ and J_k is the k-vector whose components are all equal to 1. The group of transformation is $\mathcal{G} = \{g_c : c \in \mathcal{R}\}$, which is a location-scale transformation group with $\mathcal{T} = \{I_k\}$ and $\mathcal{C} = \{cJ_k : c \in \mathcal{R}\}$. $\mathcal{P}$ is invariant under $\mathcal{G}$ with $\bar{g}_c(P_\mu) = P_{\mu+c}$. For estimating μ under the loss $L(\mu, a) = L(\mu - a)$, where $L(\cdot)$ is a nonnegative Borel function, the decision problem is invariant with $g_c(a) = a + c$. A decision rule T is invariant if and only if $T(x + cJ_k) = T(x) + c$ for every $x \in \mathcal{R}^k$ and $c \in \mathcal{R}$. An example of an invariant decision rule is $T(x) = l^\tau x$ for some $l \in \mathcal{R}^k$ with $l^\tau J_k = 1$. Note that $T(x) = l^\tau x$ with $l^\tau J_k = 1$ is in the class $\Im_2$ in Example 2.22. ∎

In §4.2 and §6.3, we discuss the problem of finding a $\Im$-optimal rule when $\Im$ is a class of invariant decision rules.

The second approach to finding a good decision rule is to consider some characteristic R_T of $R_T(P)$, for a given decision rule T, and then minimize R_T over $T \in \Im$. The following are two popular ways to carry out this idea. The first one is to consider an average of $R_T(P)$ over $P \in \mathcal{P}$:

$$r_T(\Pi) = \int_{\mathcal{P}} R_T(P) d\Pi(P),$$

where Π is a known probability measure on $(\mathcal{P}, \mathcal{F}_\mathcal{P})$ with an appropriate σ-field $\mathcal{F}_\mathcal{P}$. $r_T(\Pi)$ is called the *Bayes risk* of T w.r.t. Π. If $T_* \in \Im$ and $r_{T_*}(\Pi) \leq r_T(\Pi)$ for any $T \in \Im$, then T_* is called a $\Im$-*Bayes rule* (or Bayes rule when $\Im$ contains all possible rules) w.r.t. Π. The second method is to consider the worst situation, i.e., $\sup_{P \in \mathcal{P}} R_T(P)$. If $T_* \in \Im$ and $\sup_{P \in \mathcal{P}} R_{T_*}(P) \leq \sup_{P \in \mathcal{P}} R_T(P)$ for any $T \in \Im$, then T_* is called a $\Im$-*minimax* rule (or minimax rule when $\Im$ contains all possible rules). Bayes and minimax rules are discussed in Chapter 4.

Example 2.25. We usually try to find a Bayes rule or a minimax rule in a parametric problem where $P = P_\theta$ for a $\theta \in \mathcal{R}^k$. Consider the special case of $k = 1$ and $L(\theta, a) = (\theta - a)^2$, the squared error loss. Note that

$$r_T(\Pi) = \int_{\mathcal{R}} E[\theta - T(X)]^2 d\Pi(\theta),$$

which is equivalent to $E[\boldsymbol{\theta} - T(X)]^2$, where $\boldsymbol{\theta}$ is a random variable having the distribution Π and, given $\boldsymbol{\theta} = \theta$, the conditional distribution of X is P_θ. Then, the problem can be viewed as a prediction problem for $\boldsymbol{\theta}$ using functions of X. Using the result in Example 1.22, the best predictor is $E(\boldsymbol{\theta}|X)$, which is the $\Im$-Bayes rule w.r.t. Π with $\Im$ being the class of rules $T(X)$ satisfying $E[T(X)]^2 < \infty$ for any θ.

As a more specific example, let $X = (X_1, ..., X_n)$ with i.i.d. components having the $N(\mu, \sigma^2)$ distribution with an unknown $\mu = \theta \in \mathcal{R}$ and a known σ^2, and let Π be the $N(\mu_0, \sigma_0^2)$ distribution with known μ_0 and σ_0^2. Then the conditional distribution of $\boldsymbol{\theta}$ given $X = x$ is $N(\mu_*(x), c^2)$ with

$$\mu_*(x) = \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 + \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\bar{x} \qquad \text{and} \qquad c^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2} \tag{2.25}$$

(exercise). The Bayes rule w.r.t. Π is $E(\boldsymbol{\theta}|X) = \mu_*(X)$.

In this special case we can show that the sample mean $\bar{X}$ is $\Im$-minimax with $\Im$ being the collection of all decision rules. For any decision rule T,

$$\begin{aligned}
\sup_{\theta\in\mathcal{R}} R_T(\theta) &\geq \int_{\mathcal{R}} R_T(\theta) d\Pi(\theta) \\
&\geq \int_{\mathcal{R}} R_{\mu_*}(\theta) d\Pi(\theta) \\
&= E\left\{[\boldsymbol{\theta} - \mu_*(X)]^2\right\} \\
&= E\left\{E\{[\boldsymbol{\theta} - \mu_*(X)]^2|X\}\right\} \\
&= E(c^2) \\
&= c^2,
\end{aligned}$$

where $\mu_*(X)$ is the Bayes rule given in (2.25) and c^2 is also given in (2.25). Since this result is true for any $\sigma_0^2 > 0$ and $c^2 \to \sigma^2/n$ as $\sigma_0^2 \to \infty$,

$$\sup_{\theta\in\mathcal{R}} R_T(\theta) \geq \frac{\sigma^2}{n} = \sup_{\theta\in\mathcal{R}} R_{\bar{X}}(\theta),$$

where the equality holds because the risk of $\bar{X}$ under the squared error loss is, by (2.20), σ^2/n and independent of $\theta = \mu$. Thus, $\bar{X}$ is minimax.

A minimax rule in a general case may be difficult to obtain. It can be seen that if both μ and σ^2 are unknown in the previous discussion, then

$$\sup_{\theta\in\mathcal{R}\times(0,\infty)} R_{\bar{X}}(\theta) = \infty, \tag{2.26}$$

where $\theta = (\mu, \sigma^2)$. Hence $\bar{X}$ cannot be minimax unless (2.26) holds with $\bar{X}$ replaced by any decision rule T, in which case minimaxity becomes meaningless. ∎

2.4 Statistical Inference

The loss function plays a crucial role in statistical decision theory. Loss functions can be obtained from a utility analysis (Berger, 1985), but in many problems they have to be determined subjectively. In *statistical inference*, we make an inference about the unknown population based on the sample X and *inference procedures* without using any loss function, although any inference procedure can be cast in decision-theoretic terms as a decision rule.

There are three main types of inference procedures: *point estimators*, *hypothesis tests*, and *confidence sets*.

2.4.1 Point estimators

The problem of estimating an unknown parameter related to the unknown population is introduced in Example 2.19 and the discussion after Example 2.19 as a special statistical decision problem. In statistical inference, however, estimators of parameters are derived based on some principle (such as the unbiasedness, invariance, sufficiency, substitution principle, likelihood principle, Bayesian principle, etc.), not based on a loss or risk function. Since confidence sets are sometimes also called *interval estimators* or *set estimators*, estimators of parameters are called point estimators.

In Chapters 3 through 5, we consider how to derive a "good" point estimator based on some principle. Here we focus on how to assess performance of point estimators.

Let $\vartheta \in \tilde{\Theta} \subset \mathcal{R}$ be a parameter to be estimated, which is a function of the unknown population P or θ if P is in a parametric family. An estimator is a statistic with range $\tilde{\Theta}$. First, one has to realize that any estimator $T(X)$ of ϑ is subject to an estimation error $T(x) - \vartheta$ when we observe $X = x$. This is not just because $T(X)$ is random. In some problems $T(x)$ never equals ϑ. A trivial example is when $T(X)$ has a continuous c.d.f. so that $P(T(X) = \vartheta) = 0$. As a nontrivial example, let $X_1, ..., X_n$ be i.i.d. binary random variables (also called Bernoulli variables) with $P(X_i = 1) = p$ and $P(X_i = 0) = 1 - p$. The sample mean $\bar{X}$ is shown to be a good estimator of $\vartheta = p$ in later chapters, but $\bar{x}$ never equals ϑ if ϑ is not one of j/n, $j = 0, 1, ..., n$. Thus, we cannot assess the performance of $T(X)$ by the values of $T(x)$ with particular x's and it is also not worthwhile to do so.

The bias $b_T(P)$ and unbiasedness of a point estimator $T(X)$ is defined in Definition 2.8. Unbiasedness of $T(X)$ means that the mean of $T(X)$ is equal to ϑ. An unbiased estimator $T(X)$ can be viewed as an estimator without "systematic" error, since, on the average, it does not overestimate (i.e., $b_T(P) > 0$) or underestimate (i.e., $b_T(P) < 0$). However, an unbiased

estimator $T(X)$ may have large positive and negative errors $T(x)-\vartheta$, $x \in \mathfrak{X}$, although these errors cancel each other in the calculation of the bias, which is the average $\int [T(x) - \vartheta] dP_X(x)$.

Hence, for an unbiased estimator $T(X)$, it is desired that the values of $T(x)$ be highly concentrated around ϑ. The variance of $T(X)$ is commonly used as a measure of the dispersion of $T(X)$. The *mean squared error* (mse) of $T(X)$ as an estimator of ϑ is defined to be

$$\text{mse}_T(P) = E[T(X) - \vartheta]^2 = [b_T(P)]^2 + \text{Var}(T(X)), \tag{2.27}$$

which is denoted by $\text{mse}_T(\theta)$ if P is in a parametric family. $\text{mse}_T(P)$ is equal to the variance $\text{Var}(T(X))$ if and only if $T(X)$ is unbiased. Note that the mse is simply the risk of T in statistical decision theory under the squared error loss.

In addition to the variance and the mse, the following are other measures of dispersion that are often used in point estimation problems. The first one is the *mean absolute error* of an estimator $T(X)$ defined to be $E|T(X)-\vartheta|$. The second one is the probability of falling outside a stated distance of ϑ, i.e., $P(|T(X) - \vartheta| \geq \epsilon)$ with a fixed $\epsilon > 0$. Again, these two measures of dispersion are risk functions in statistical decision theory with loss functions $|\vartheta - a|$ and $I_{(\epsilon,\infty)}(|\vartheta - a|)$, respectively.

For the bias, variance, mse, and mean absolute error, we have implicitly assumed that certain moments of $T(X)$ exist. On the other hand, the dispersion measure $P(|T(X)-\vartheta| \geq \epsilon)$ depends on the choice of ϵ. It is possible that some estimators are good in terms of one measure of dispersion, but not in terms of other measures of dispersion. The mse, which is a function of bias and variance according to (2.27), is mathematically easy to handle and, hence, is used the most often in the literature. In this book, we use the mse to assess and compare point estimators unless otherwise stated.

Examples 2.19 and 2.22 provide some examples of estimators and their biases, variances, and mse's. The following are two more examples.

Example 2.26. Consider the life-time testing problem in Example 2.2. Let $X_1, ..., X_n$ be i.i.d. from an unknown c.d.f. F. Suppose that the parameter of interest is $\vartheta = 1 - F(t)$ for a fixed $t > 0$. If F is not in a parametric family, then a *nonparametric* estimator of $F(t)$ is the *empirical* c.d.f.

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,t]}(X_i), \qquad t \in \mathcal{R}. \tag{2.28}$$

Since $I_{(-\infty,t]}(X_1), ..., I_{(-\infty,t]}(X_n)$ are i.i.d. binary random variables with $P(I_{(-\infty,t]}(X_i) = 1) = F(t)$, the random variable $nF_n(t)$ has the binomial distribution $Bi(F(t), n)$. Consequently, $F_n(t)$ is an unbiased estimator of

$F(t)$ and $\mathrm{Var}(F_n(t)) = \mathrm{mse}_{F_n(t)}(P) = F(t)[1-F(t)]/n$. Since any linear combination of unbiased estimators is unbiased for the same linear combination of the parameters (by the linearity of expectations), an unbiased estimator of ϑ is $U(X) = 1 - F_n(t)$, which has the same variance and mse as $F_n(t)$.

The estimator $U(X) = 1 - F_n(t)$ can be improved in terms of the mse if there is further information about F. Suppose that F is the c.d.f. of the exponential distribution $E(0,\theta)$ with an unknown $\theta > 0$. Then $\vartheta = e^{-t/\theta}$. From §2.2.2, the sample mean $\bar{X}$ is sufficient for $\theta > 0$. Since the squared error loss is strictly convex, an application of Theorem 2.5(ii) (Rao-Blackwell theorem) shows that the estimator $T(X) = E[1-F_n(t)|\bar{X}]$, which is also unbiased, is better than $U(X)$ in terms of the mse. Figure 2.1 shows graphs of the mse's of $U(X)$ and $T(X)$, as functions of θ, in the special case of $n = 10$, $t = 2$, and $F(x) = (1-e^{-x/\theta})I_{(0,\infty)}(x)$. ∎

Example 2.27. Consider the sample survey problem in Example 2.3 with a constant selection probability $p(\boldsymbol{s})$ and univariate y_i. Let $\vartheta = Y = \sum_{i=1}^N y_i$, the population total. We now show that the estimator $\hat{Y} = \frac{N}{n}\sum_{i\in\boldsymbol{s}} y_i$ is an unbiased estimator of Y. Let $a_i = 1$ if $i \in \boldsymbol{s}$ and $a_i = 0$ otherwise. Thus, $\hat{Y} = \frac{N}{n}\sum_{i=1}^N a_i y_i$. Since $p(\boldsymbol{s})$ is constant, $E(a_i) = P(a_i = 1) = n/N$ and

$$E(\hat{Y}) = E\left(\frac{N}{n}\sum_{i=1}^N a_i y_i\right) = \frac{N}{n}\sum_{i=1}^N y_i E(a_i) = \sum_{i=1}^N y_i = Y.$$

Note that

$$\mathrm{Var}(a_i) = E(a_i) - [E(a_i)]^2 = \frac{n}{N}\left(1-\frac{n}{N}\right)$$

and for $i \neq j$,

$$\mathrm{Cov}(a_i, a_j) = P(a_i = 1, a_j = 1) - E(a_i)E(a_j) = \frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2}.$$

Hence, the variance or the mse of $\hat{Y}$ is

$$\begin{aligned}
\mathrm{Var}(\hat{Y}) &= \frac{N^2}{n^2}\mathrm{Var}\left(\sum_{i=1}^N a_i y_i\right) \\
&= \frac{N^2}{n^2}\left[\sum_{i=1}^N y_i^2 \mathrm{Var}(a_i) + 2\sum_{1\le i<j\le N} y_i y_j \mathrm{Cov}(a_i, a_j)\right] \\
&= \frac{N}{n}\left(1-\frac{n}{N}\right)\left(\sum_{i=1}^N y_i^2 - \frac{2}{N-1}\sum_{1\le i<j\le N} y_i y_j\right) \\
&= \frac{N^2}{n(N-1)}\left(1-\frac{n}{N}\right)\sum_{i=1}^N \left(y_i - \frac{Y}{N}\right)^2. \quad ∎
\end{aligned}$$

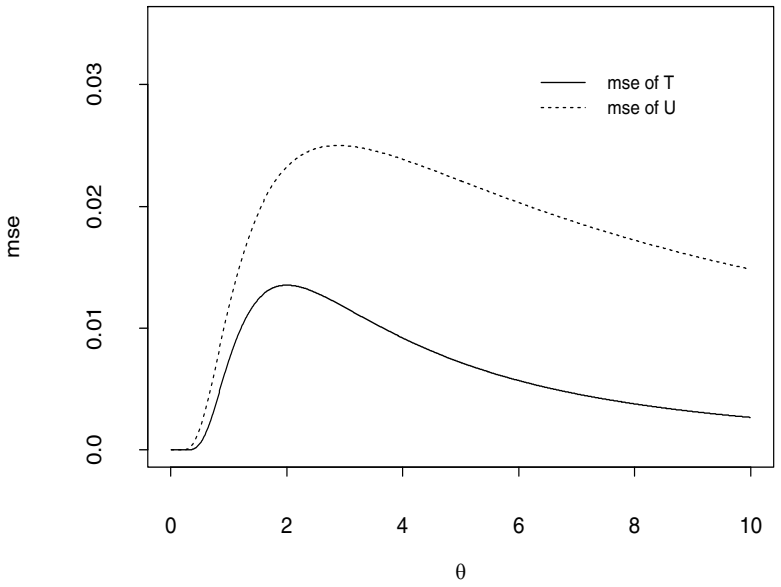

Figure 2.1: mse's of $U(X)$ and $T(X)$ in Example 2.26

2.4.2 Hypothesis tests

The basic elements of a hypothesis testing problem are described in Example 2.20. In statistical inference, tests for a hypothesis are derived based on some principles similar to those given in an estimation problem. Chapter 6 is devoted to deriving tests for various types of hypotheses. Several key ideas are discussed here.

To test the hypotheses H_0 versus H_1 given in (2.21), there are only two types of statistical errors we may commit: rejecting H_0 when H_0 is true (called the *type I error*) and accepting H_0 when H_0 is wrong (called the *type II error*). In statistical inference, a test T, which is a statistic from $\mathcal{X}$ to $\{0, 1\}$, is assessed by the probabilities of making two types of errors:

$$\alpha_T(P) = P(T(X) = 1) \qquad P \in \mathcal{P}_0 \tag{2.29}$$

and

$$1 - \alpha_T(P) = P(T(X) = 0) \qquad P \in \mathcal{P}_1, \tag{2.30}$$

which are denoted by $\alpha_T(\theta)$ and $1 - \alpha_T(\theta)$ if P is in a parametric family indexed by θ. Note that these are risks of T under the 0-1 loss in statistical decision theory. However, an optimal decision rule (test) does not exist even for a very simple problem with a very simple class of tests (Example 2.23).

That is, error probabilities in (2.29) and (2.30) cannot be minimized simultaneously. Furthermore, these two error probabilities cannot be bounded simultaneously by a fixed $\alpha \in (0,1)$ when we have a sample of a fixed size.

Therefore, a common approach to finding an "optimal" test is to assign a small bound α to one of the error probabilities, say $\alpha_T(P)$, $P \in \mathcal{P}_0$, and then to attempt to minimize the other error probability $1-\alpha_T(P)$, $P \in \mathcal{P}_1$, subject to

$$\sup_{P\in\mathcal{P}_0} \alpha_T(P) \le \alpha. \tag{2.31}$$

The bound α is called the *level of significance*. The left-hand side of (2.31) is called the *size* of the test T. Note that the level of significance should be positive, otherwise no test satisfies (2.31) except the silly test $T(X) \equiv 0$ a.s. $\mathcal{P}$.

Example 2.28. Let $X_1, ..., X_n$ be i.i.d. from the $N(\mu, \sigma^2)$ distribution with an unknown $\mu \in \mathcal{R}$ and a known σ^2. Consider the hypotheses

$$H_0 : \mu \le \mu_0 \qquad \text{versus} \qquad H_1 : \mu > \mu_0,$$

where μ_0 is a fixed constant. Since the sample mean $\bar{X}$ is sufficient for $\mu \in \mathcal{R}$, it is reasonable to consider the following class of tests: $T_c(X) = I_{(c,\infty)}(\bar{X})$, i.e., H_0 is rejected (accepted) if $\bar{X} > c$ ($\bar{X} \le c$), where $c \in \mathcal{R}$ is a fixed constant. Let Φ be the c.d.f. of $N(0,1)$. Then, by the property of the normal distributions,

$$\alpha_{T_c}(\mu) = P(T_c(X) = 1) = 1 - \Phi\left(\frac{\sqrt{n}(c-\mu)}{\sigma}\right). \tag{2.32}$$

Figure 2.2 provides an example of a graph of two types of error probabilities, with $\mu_0 = 0$. Since $\Phi(t)$ is an increasing function of t,

$$\sup_{P\in\mathcal{P}_0} \alpha_{T_c}(\mu) = 1 - \Phi\left(\frac{\sqrt{n}(c-\mu_0)}{\sigma}\right).$$

In fact, it is also true that

$$\sup_{P\in\mathcal{P}_1} [1 - \alpha_{T_c}(\mu)] = \Phi\left(\frac{\sqrt{n}(c-\mu_0)}{\sigma}\right).$$

If we would like to use an α as the level of significance, then the most effective way is to choose a c_α (a test $T_{c_\alpha}(X)$) such that

$$\alpha = \sup_{P\in\mathcal{P}_0} \alpha_{T_{c_\alpha}}(\mu),$$

in which case c_α must satisfy

$$1 - \Phi\left(\frac{\sqrt{n}(c_\alpha-\mu_0)}{\sigma}\right) = \alpha,$$

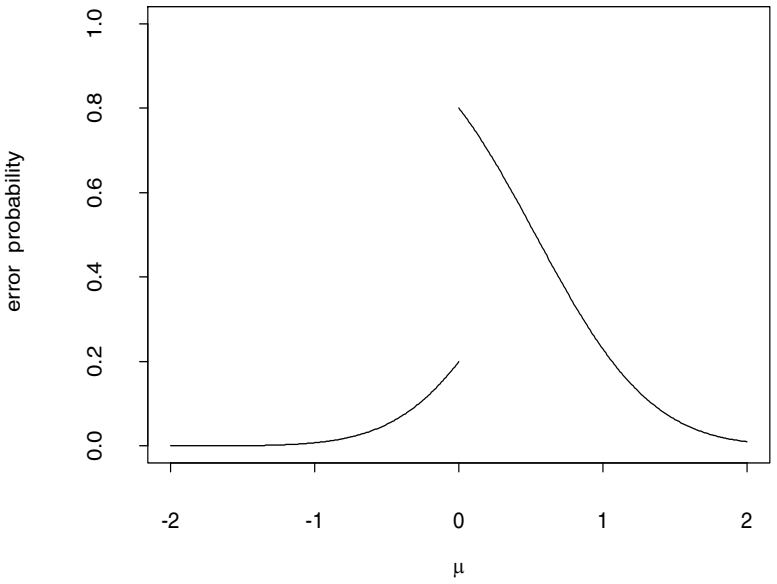

Figure 2.2: Error probabilities in Example 2.28

i.e., $c_\alpha = \sigma z_{1-\alpha}/\sqrt{n} + \mu_0$, where $z_a = \Phi^{-1}(a)$. In Chapter 6, it is shown that for any test $T(X)$ satisfying (2.31),

$$1 - \alpha_T(\mu) \geq 1 - \alpha_{T_{c_\alpha}}(\mu), \qquad \mu > \mu_0. \quad \blacksquare$$

The choice of a level of significance α is usually somewhat subjective. In most applications there is no precise limit to the size of T that can be tolerated. Standard values, such as 0.10, 0.05, or 0.01, are often used for convenience.

For most tests satisfying (2.31), a small α leads to a "small" rejection region. It is good practice to determine not only whether H_0 is rejected or accepted for a given α and a chosen test T_α, but also the smallest possible level of significance at which H_0 would be rejected for the computed $T_\alpha(x)$, i.e., $\hat{\alpha} = \inf\{\alpha \in (0,1) : T_\alpha(x) = 1\}$. Such an $\hat{\alpha}$, which depends on x and the chosen test and is a statistic, is called the *p-value* for the test T_α.

Example 2.29. Consider the problem in Example 2.28. Let us calculate the p-value for T_{c_α}. Note that

$$\alpha = 1 - \Phi\left(\frac{\sqrt{n}(c_\alpha - \mu_0)}{\sigma}\right) > 1 - \Phi\left(\frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma}\right)$$

if and only if $\bar{x} > c_\alpha$ (or $T_{c_\alpha}(x) = 1$). Hence

$$1 - \Phi\left(\frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma}\right) = \inf\{\alpha \in (0,1) : T_{c_\alpha}(x) = 1\} = \hat{\alpha}(x)$$

is the p-value for T_{c_α}. It turns out that $T_{c_\alpha}(x) = I_{(0,\alpha)}(\hat{\alpha}(x))$. ∎

With the additional information provided by p-values, using p-values is typically more appropriate than using fixed-level tests in a scientific problem. However, a fixed level of significance is unavoidable when acceptance or rejection of H_0 implies an imminent concrete decision. For more discussions about p-values, see Lehmann (1986) and Weerahandi (1995).

In Example 2.28, the equality in (2.31) can always be achieved by a suitable choice of c. This is, however, not true in general. In Example 2.23, for instance, it is possible to find an α such that

$$\sup_{0<\theta\leq\theta_0} P(T_j(X) = 1) \neq \alpha$$

for all T_j's. In such cases, we may consider *randomized tests*, which are introduced next.

Recall that a randomized decision rule is a probability measure $\delta(x, \cdot)$ on the action space for any fixed x. Since the action space contains only two points, 0 and 1, for a hypothesis testing problem, any randomized test $\delta(X, A)$ is equivalent to a statistic $T(X) \in [0,1]$ with $T(x) = \delta(x, \{1\})$ and $1 - T(x) = \delta(x, \{0\})$. A nonrandomized test is obviously a special case where $T(x)$ does not take any value in $(0,1)$.

For any randomized test $T(X)$, we define the type I error probability to be $\alpha_T(P) = E[T(X)]$, $P \in \mathcal{P}_0$, and the type II error probability to be $1 - \alpha_T(P) = E[1 - T(X)]$, $P \in \mathcal{P}_1$. For a class of randomized tests, we would like to minimize $1 - \alpha_T(P)$ subject to (2.31).

Example 2.30. Consider Example 2.23 and the following class of randomized tests:

$$T_{j,q}(X) = \begin{cases} 1 & X > j \\ q & X = j \\ 0 & X < j, \end{cases}$$

where $j = 0, 1, ..., n-1$ and $q \in [0,1]$. Then

$$\alpha_{T_{j,q}}(\theta) = P(X > j) + qP(X = j) \qquad 0 < \theta \leq \theta_0$$

and

$$1 - \alpha_{T_{j,q}}(\theta) = P(X < j) + (1-q)P(X = j) \qquad \theta_0 < \theta < 1.$$

It can be shown that for any $\alpha \in (0,1)$, there exist an integer j and $q \in (0,1)$ such that the size of $T_{j,q}$ is α (exercise). ∎

2.4.3 Confidence sets

Let ϑ be a k-vector of unknown parameters related to the unknown population $P \in \mathcal{P}$ and $C(X) \in \mathcal{B}^k_{\tilde{\Theta}}$ depending only on the sample X, where $\tilde{\Theta} \in \mathcal{B}^k$ is the range of ϑ. If

$$\inf_{P \in \mathcal{P}} P(\vartheta \in C(X)) \geq 1 - \alpha, \tag{2.33}$$

where α is a fixed constant in $(0, 1)$, then $C(X)$ is called a *confidence set* for ϑ with *level of significance* $1 - \alpha$. The left-hand side of (2.33) is called the *confidence coefficient* of $C(X)$, which is the highest possible level of significance for $C(X)$. A confidence set is a random element that covers the unknown ϑ with certain probability. If (2.33) holds, then the *coverage probability* of $C(X)$ is at least $1-\alpha$, although $C(x)$ either covers or does not cover ϑ whence we observe $X = x$. The concepts of level of significance and confidence coefficient are very similar to the level of significance and size in hypothesis testing. In fact, it is shown in Chapter 7 that some confidence sets are closely related to hypothesis tests.

Consider a real-valued ϑ. If $C(X) = [\underline{\vartheta}(X), \overline{\vartheta}(X)]$ for a pair of real-valued statistics $\underline{\vartheta}$ and $\overline{\vartheta}$, then $C(X)$ is called a *confidence interval* for ϑ. If $C(X) = (-\infty, \overline{\vartheta}(X)]$ (or $[\underline{\vartheta}(X), \infty)$), then $\overline{\vartheta}$ (or $\underline{\vartheta}$) is called an upper (or a lower) *confidence bound* for ϑ.

A confidence set (or interval) is also called a set (or an interval) estimator of ϑ, although it is very different from a point estimator (discussed in §2.4.1).

Example 2.31. Consider Example 2.28. Suppose that a confidence interval for $\vartheta = \mu$ is needed. Again, we only need to consider $\underline{\vartheta}(\bar{X})$ and $\overline{\vartheta}(\bar{X})$, since the sample mean $\bar{X}$ is sufficient. Consider confidence intervals of the form $[\bar{X} - c, \bar{X} + c]$, where $c \in (0, \infty)$ is fixed. Note that

$$P\left(\mu \in [\bar{X} - c, \bar{X} + c]\right) = P\left(|\bar{X} - \mu| \leq c\right) = 1 - 2\Phi\left(-\sqrt{n}c/\sigma\right),$$

which is independent of μ. Hence, the confidence coefficient of $[\bar{X} - c, \bar{X} + c]$ is $1 - 2\Phi\left(-\sqrt{n}c/\sigma\right)$, which is an increasing function of c and converges to 1 as $c \to \infty$ or 0 as $c \to 0$. Thus, confidence coefficients are positive but less than 1 except for silly confidence intervals $[\bar{X}, \bar{X}]$ and $(-\infty, \infty)$. We can choose a confidence interval with an arbitrarily large confidence coefficient, but the chosen confidence interval may be so wide that it is practically useless.

If σ^2 is also unknown, then $[\bar{X} - c, \bar{X} + c]$ has confidence coefficient 0 and, therefore, is not a good inference procedure. In such a case a different confidence interval for μ with positive confidence coefficient can be derived (Exercise 97 in §2.6). ∎

This example tells us that a reasonable approach is to choose a level of significance $1-\alpha \in (0,1)$ (just like the level of significance in hypothesis testing) and a confidence interval or set satisfying (2.33). In Example 2.31, when σ^2 is known and c is chosen to be $\sigma z_{1-\alpha/2}/\sqrt{n}$, where $z_a = \Phi^{-1}(a)$, the confidence coefficient of the confidence interval $[\bar{X}-c, \bar{X}+c]$ is *exactly* $1-\alpha$ for any fixed $\alpha \in (0,1)$. This is desirable since, for all confidence intervals satisfying (2.33), the one with the shortest interval length is preferred.

For a general confidence interval $[\underline{\vartheta}(X), \overline{\vartheta}(X)]$, its length is $\overline{\vartheta}(X)-\underline{\vartheta}(X)$, which may be random. We may consider the expected (or average) length $E[\overline{\vartheta}(X)-\underline{\vartheta}(X)]$. The confidence coefficient and expected length are a pair of good measures of performance of confidence intervals. Like the two types of error probabilities of a test in hypothesis testing, however, we cannot maximize the confidence coefficient and minimize the length (or expected length) simultaneously. A common approach is to minimize the length (or expected length) subject to (2.33).

For an unbounded confidence interval, its length is ∞. Hence we have to define some other measures of performance. For an upper (or a lower) confidence bound, we may consider the distance $\overline{\vartheta}(X)-\vartheta$ (or $\vartheta-\underline{\vartheta}(X)$) or its expectation.

To conclude this section, we discuss an example of a confidence set for a two-dimensional parameter. General discussions about how to construct and assess confidence sets are given in Chapter 7.

Example 2.32. Let $X_1, ..., X_n$ be i.i.d. from the $N(\mu, \sigma^2)$ distribution with both $\mu \in \mathcal{R}$ and $\sigma^2 > 0$ unknown. Let $\theta = (\mu, \sigma^2)$ and $\alpha \in (0,1)$ be given. Let $\bar{X}$ be the sample mean and S^2 be the sample variance. Since $(\bar{X}, S^2)$ is sufficient (Example 2.15), we focus on $C(X)$ that is a function of $(\bar{X}, S^2)$. From Example 2.18, $\bar{X}$ and S^2 are independent and $(n-1)S^2/\sigma^2$ has the chi-square distribution χ^2_{n-1}. Since $\sqrt{n}(\bar{X}-\mu)/\sigma$ has the $N(0,1)$ distribution (Exercise 43 in §1.6),

$$P\left(-\tilde{c}_\alpha \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le \tilde{c}_\alpha\right) = \sqrt{1-\alpha},$$

where $\tilde{c}_\alpha = \Phi^{-1}\left(\frac{1+\sqrt{1-\alpha}}{2}\right)$ (verify). Since the chi-square distribution χ^2_{n-1} is a known distribution, we can always find two constants $c_{1\alpha}$ and $c_{2\alpha}$ such that

$$P\left(c_{1\alpha} \le \frac{(n-1)S^2}{\sigma^2} \le c_{2\alpha}\right) = \sqrt{1-\alpha}.$$

Then

$$P\left(-\tilde{c}_\alpha \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le \tilde{c}_\alpha, c_{1\alpha} \le \frac{(n-1)S^2}{\sigma^2} \le c_{2\alpha}\right) = 1-\alpha,$$

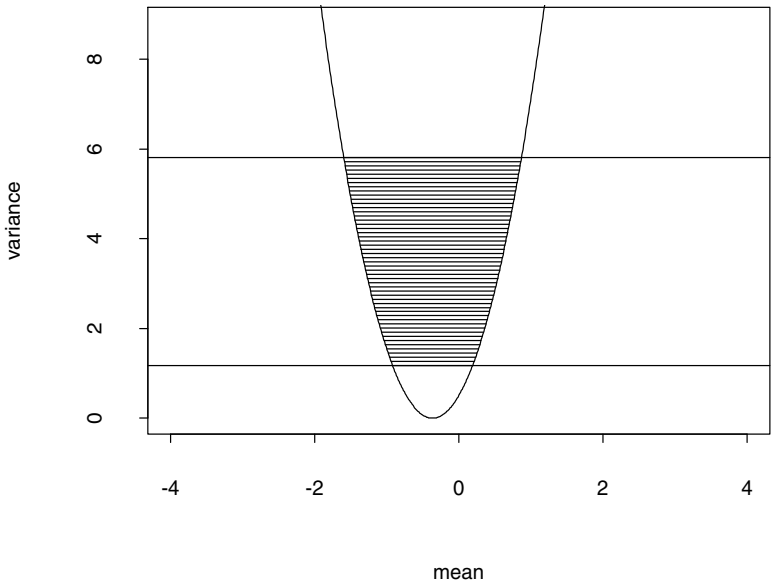

Figure 2.3: A confidence set for θ in Example 2.32

or

$$P\left(\frac{n(\bar{X}-\mu)^2}{\tilde{c}_\alpha^2} \le \sigma^2, \frac{(n-1)S^2}{c_{2\alpha}} \le \sigma^2 \le \frac{(n-1)S^2}{c_{1\alpha}}\right) = 1-\alpha. \quad (2.34)$$

The left-hand side of (2.34) defines a set in the range of $\theta = (\mu, \sigma^2)$ bounded by two straight lines, $\sigma^2 = (n-1)S^2/c_{i\alpha}$, $i = 1, 2$, and a curve $\sigma^2 = n(\bar{X}-\mu)^2/\tilde{c}_\alpha^2$ (see the shadowed part of Figure 2.3). This set is a confidence set for θ with confidence coefficient $1-\alpha$, since (2.34) holds for any θ. ∎

2.5 Asymptotic Criteria and Inference

We have seen that in statistical decision theory and inference, a key to the success of finding a good decision rule or inference procedure is being able to find some moments and/or distributions of various statistics. Although many examples are presented (including those in the exercises in §2.6), there are more cases in which we are not able to find exactly the moments or distributions of given statistics, especially when the problem is not parametric (see, e.g., the discussions in Example 2.8).

In practice, the sample size n is often large, which allows us to approximate the moments and distributions of statistics that are impossible

to derive, using the asymptotic tools discussed in §1.5. In an asymptotic analysis, we consider a sample $X = (X_1, ..., X_n)$ not for fixed n, but as a member of a sequence corresponding to $n = n_0, n_0 + 1, ...$, and obtain the limit of the distribution of an appropriately normalized statistic or variable $T_n(X)$ as $n \to \infty$. The limiting distribution and its moments are used as approximations to the distribution and moments of $T_n(X)$ in the situation with a large but actually finite n. This leads to some asymptotic statistical procedures and asymptotic criteria for assessing their performances, which are introduced in this section.

The asymptotic approach is not only applied to the situation where no exact method is available, but also used to provide an inference procedure simpler (e.g., in terms of computation) than that produced by the exact approach (the approach considering a fixed n). Some examples are given in later chapters.

In addition to providing more theoretical results and/or simpler inference procedures, the asymptotic approach requires less stringent mathematical assumptions than does the exact approach. The mathematical precision of the optimality results obtained in statistical decision theory, for example, tends to obscure the fact that these results are approximations in view of the approximate nature of the assumed models and loss functions. As the sample size increases, the statistical properties become less dependent on the loss functions and models. However, a major weakness of the asymptotic approach is that typically no good estimates for the precision of the approximations are available and, therefore, we cannot determine whether a particular n in a problem is large enough to safely apply the asymptotic results. To overcome this difficulty, asymptotic results are frequently used in combination with some numerical/empirical studies for selected values of n to examine the *finite sample* performance of asymptotic procedures.

2.5.1 Consistency

A reasonable point estimator is expected to perform better, at least on the average, if more information about the unknown population is available. With a fixed model assumption and sampling plan, more data (larger sample size n) provide more information about the unknown population. Thus, it is distasteful to use a point estimator T_n which, if sampling were to continue indefinitely, could possibly have a nonzero estimation error, although the estimation error of T_n for a fixed n may never equal 0 (see the discussion in §2.4.1).

Definition 2.10 (Consistency of point estimators). Let $X = (X_1, ..., X_n)$ be a sample from $P \in \mathcal{P}$ and $T_n(X)$ be a point estimator of ϑ for every n. (i) $T_n(X)$ is called *consistent* for ϑ if and only if $T_n(X) \to_p \vartheta$ w.r.t. any

$P \in \mathcal{P}$.
(ii) Let $\{a_n\}$ be a sequence of positive constants diverging to ∞. $T_n(X)$ is called a_n-*consistent* for ϑ if and only if $a_n[T_n(X) - \vartheta] = O_p(1)$ w.r.t. any $P \in \mathcal{P}$.
(iii) $T_n(X)$ is called *strongly consistent* for ϑ if and only if $T_n(X) \to_{a.s.} \vartheta$ w.r.t. any $P \in \mathcal{P}$.
(iv) $T_n(X)$ is called L_r-*consistent* for ϑ if and only if $T_n(X) \to_{L_r} \vartheta$ w.r.t. any $P \in \mathcal{P}$ for some fixed $r > 0$. ∎

Consistency is actually a concept relating to a sequence of estimators, $\{T_n, n = n_0, n_0 + 1, ...\}$, but we usually just say "consistency of T_n" for simplicity. Each of the four types of consistency in Definition 2.10 describes the convergence of $T_n(X)$ to ϑ in some sense, as $n \to \infty$. In statistics, consistency according to Definition 2.10(i), which is sometimes called *weak consistency* since it is implied by any of the other three types of consistency, is the most useful concept of convergence of T_n to ϑ. L_2-consistency is also called *consistency in mse*, which is the most useful type of L_r-consistency.

Example 2.33. Let $X_1, ..., X_n$ be i.i.d. from $P \in \mathcal{P}$. If $\vartheta = \mu$, which is the mean of P and is assumed to be finite, then by the SLLN (Theorem 1.13), the sample mean $\bar{X}$ is strongly consistent for μ and, therefore, is also consistent for μ. If we further assume that the variance of P is finite, then by (2.20), $\bar{X}$ is consistent in mse and is $\sqrt{n}$-consistent. With the finite variance assumption, the sample variance S^2 is strongly consistent for the variance of P, according to the SLLN.

Consider estimators of the form $T_n = \sum_{i=1}^n c_{ni} X_i$, where $\{c_{ni}\}$ is a double array of constants. If P has a finite variance, then by (2.24), T_n is consistent in mse if and only if $\sum_{i=1}^n c_{ni} \to 1$ and $\sum_{i=1}^n c_{ni}^2 \to 0$. If we only assume the existence of the mean of P, then T_n with $c_{ni} = c_i/n$ satisfying $n^{-1} \sum_{i=1}^n c_i \to 1$ and $\sup_i |c_i| < \infty$ is strongly consistent (Theorem 1.13(ii)). ∎

One or a combination of the law of large numbers, the CLT, Slutsky's theorem (Theorem 1.11), and the continuous mapping theorem (Theorems 1.10 and 1.12) are typically applied to establish consistency of point estimators. In particular, Theorem 1.10 implies that if T_n is (strongly) consistent for ϑ and g is a continuous function of ϑ, then $g(T_n)$ is (strongly) consistent for $g(\vartheta)$. For example, in Example 2.33 the point estimator $\bar{X}^2$ is strongly consistent for μ^2. To show that $\bar{X}^2$ is $\sqrt{n}$-consistent under the assumption that P has a finite variance σ^2, we can use the identity

$$\sqrt{n}(\bar{X}^2 - \mu^2) = \sqrt{n}(\bar{X} - \mu)(\bar{X} + \mu)$$

and the fact that $\bar{X}$ is $\sqrt{n}$-consistent for μ and $\bar{X} + \mu = O_p(1)$. (Note that

$\bar{X}^2$ may not be consistent in mse since we do not assume that P has a finite fourth moment.) Alternatively, we can use the fact that $\sqrt{n}(\bar{X}^2 - \mu^2) \to_d N(0, 4\mu^2\sigma^2)$ (by the CLT and Theorem 1.12) to show the $\sqrt{n}$-consistency of $\bar{X}^2$.

The following example shows another way to establish consistency of some point estimators.

Example 2.34. Let $X_1, ..., X_n$ be i.i.d. from an unknown P with a continuous c.d.f. F satisfying $F(\theta) = 1$ for some $\theta \in \mathcal{R}$ and $F(x) < 1$ for any $x < \theta$. Consider the largest order statistic $X_{(n)}$. For any $\epsilon > 0$, $F(\theta - \epsilon) < 1$ and

$$P(|X_{(n)} - \theta| \geq \epsilon) = P(X_{(n)} \leq \theta - \epsilon) = [F(\theta - \epsilon)]^n,$$

which imply (according to Theorem 1.8(v)) $X_{(n)} \to_{a.s.} \theta$, i.e., $X_{(n)}$ is strongly consistent for θ. If we assume that $F^{(i)}(\theta-)$, the ith-order left-hand derivative of F at θ, exists and vanishes for any $i \leq m$ and that $F^{(m+1)}(\theta-)$ exists and is nonzero, where m is a nonnegative integer, then

$$1 - F(X_{(n)}) = \frac{(-1)^m F^{(m+1)}(\theta-)}{(m+1)!}(\theta - X_{(n)})^{m+1} + o\left(|\theta - X_{(n)}|^{m+1}\right) \quad \text{a.s.}$$

This result and the fact that $P\left(n[1 - F(X_{(n)})] \geq s\right) = (1 - s/n)^n$ imply that $(\theta - X_{(n)})^{m+1} = O_p(n^{-1})$, i.e., $X_{(n)}$ is $n^{(m+1)^{-1}}$-consistent. If $m = 0$, then $X_{(n)}$ is n-consistent, which is the most common situation. If $m = 1$, then $X_{(n)}$ is $\sqrt{n}$-consistent. The limiting distribution of $n^{(m+1)^{-1}}(X_{(n)} - \theta)$ can be derived as follows. Let

$$h_n(\theta) = \left[\frac{(-1)^m(m+1)!}{nF^{(m+1)}(\theta-)}\right]^{(m+1)^{-1}}.$$

For $t \leq 0$, by Slutsky's theorem,

$$\begin{aligned}
\lim_{n\to\infty} P\left(\frac{X_{(n)} - \theta}{h_n(\theta)} \leq t\right) &= \lim_{n\to\infty} P\left(\left[\frac{\theta - X_{(n)}}{h_n(\theta)}\right]^{m+1} \geq (-t)^{m+1}\right) \\
&= \lim_{n\to\infty} P\left(n[1 - F(X_{(n)})] \geq (-t)^{m+1}\right) \\
&= \lim_{n\to\infty} \left[1 - (-t)^{m+1}/n\right]^n \\
&= e^{-(-t)^{m+1}}. \quad \blacksquare
\end{aligned}$$

It can be seen from the previous examples that there are many consistent estimators. Like the admissibility in statistical decision theory, consistency is a very essential requirement in the sense that any inconsistent estimators should not be used, but a consistent estimator is not necessarily good. Thus, consistency should be used together with one or a few more criteria.

We now discuss a situation in which finding a consistent estimator is crucial. Suppose that an estimator T_n of ϑ satisfies

$$c_n[T_n(X) - \vartheta] \to_d \sigma Y, \tag{2.35}$$

where Y is a random variable with a known distribution, $\sigma > 0$ is an unknown parameter, and $\{c_n\}$ is a sequence of constants; for example, in Example 2.33, $\sqrt{n}(\bar{X} - \mu) \to_d N(0, \sigma^2)$; in Example 2.34, (2.35) holds with $c_n = n^{(m+1)^{-1}}$ and $\sigma = [(-1)^m(m+1)!/F^{(m+1)}(\theta-)]^{(m+1)^{-1}}$. If a consistent estimator $\hat{\sigma}_n$ of σ can be found, then, by Slutsky's theorem,

$$c_n[T_n(X) - \vartheta]/\hat{\sigma}_n \to_d Y$$

and, thus, we may approximate the distribution of $c_n[T_n(X) - \vartheta]/\hat{\sigma}_n$ by the known distribution of Y.

2.5.2 Asymptotic bias, variance, and mse

Unbiasedness as a criterion for point estimators is discussed in §2.3.2 and §2.4.1. In some cases, however, there is no unbiased estimator (Exercise 84 in §2.6). Furthermore, having a "slight" bias in some cases may not be a bad idea (see Exercise 63 in §2.6). Let $T_n(X)$ be a point estimator of ϑ for every n. If ET_n exists for every n and $\lim_{n\to\infty} E(T_n - \vartheta) = 0$ for any $P \in \mathcal{P}$, then T_n is said to be *approximately unbiased*.

There are many reasonable point estimators whose expectations are not well defined. For example, consider i.i.d. $(X_1, Y_1), ..., (X_n, Y_n)$ from a bivariate normal distribution with $\mu_x = EX_1$ and $\mu_y = EY_1 \neq 0$. Let $\vartheta = \mu_x/\mu_y$ and $T_n = \bar{X}/\bar{Y}$, the ratio of two sample means. Then ET_n is not defined for any n. It is then desirable to define a concept of *asymptotic bias* for point estimators whose expectations are not well defined.

Definition 2.11. (i) Let $\xi, \xi_1, \xi_2, ...$ be random variables and $\{a_n\}$ be a sequence of positive numbers satisfying $a_n \to \infty$ or $a_n \to a > 0$. If $a_n\xi_n \to_d \xi$ and $E|\xi| < \infty$, then $E\xi/a_n$ is called an *asymptotic expectation* of ξ_n.
(ii) Let T_n be a point estimator of ϑ for every n. An asymptotic expectation of $T_n - \vartheta$, if it exists, is called an asymptotic bias of T_n and denoted by $\tilde{b}_{T_n}(P)$ (or $\tilde{b}_{T_n}(\theta)$ if P is in a parametric family). If $\lim_{n\to\infty} \tilde{b}_{T_n}(P) = 0$ for any $P \in \mathcal{P}$, then T_n is said to be *asymptotically unbiased*. ∎

Like the consistency, the asymptotic expectation (or bias) is a concept relating to sequences $\{\xi_n\}$ and $\{E\xi/a_n\}$ (or $\{T_n\}$ and $\{\tilde{b}_{T_n}(P)\}$). Note that the exact bias $b_{T_n}(P)$ is not necessarily the same as $\tilde{b}_{T_n}(P)$ when both of them exist (Exercise 115 in §2.6). The following result shows that the asymptotic expectation defined in Definition 2.11 is essentially unique.

Proposition 2.3. Let $\{\xi_n\}$ be a sequence of random variables. Suppose that both $E\xi/a_n$ and $E\eta/b_n$ are asymptotic expectations of ξ_n defined according to Definition 2.11(i). Then, one of the following three must hold: (a) $E\xi = E\eta = 0$; (b) $E\xi \neq 0$, $E\eta = 0$, and $b_n/a_n \to 0$; or $E\xi = 0$, $E\eta \neq 0$, and $a_n/b_n \to 0$; (c) $E\xi \neq 0$, $E\eta \neq 0$, and $(E\xi/a_n)/(E\eta/b_n) \to 1$.

Proof. According to Definition 2.11(i), $a_n\xi_n \to_d \xi$ and $b_n\xi_n \to_d \eta$.

(i) If both ξ and η have nondegenerate c.d.f.'s, then the result follows from Exercise 129 of §1.6.

(ii) Suppose that ξ has a nondegenerate c.d.f. but η is a constant. If $\eta \neq 0$, then by Theorem 1.11(iii), $a_n/b_n \to \xi/\eta$, which is impossible since ξ has a nondegenerate c.d.f. If $\eta = 0$, then by Theorem 1.11(ii), $b_n/a_n \to 0$.

(iii) Suppose that both ξ and η are constants. If $\xi = \eta = 0$, the result follows. If $\xi \neq 0$ and $\eta = 0$, then $b_n/a_n \to 0$. If $\xi \neq 0$ and $\eta \neq 0$, then $b_n/a_n \to \eta/\xi$. ∎

If T_n is a consistent estimator of ϑ, then $T_n = \vartheta + o_p(1)$ and, by Definition 2.11(ii), T_n is asymptotically unbiased, although T_n may not be approximately unbiased; in fact, $g(T_n)$ is asymptotically unbiased for $g(\vartheta)$ for any continuous function g. For the example of $T_n = \bar{X}/\bar{Y}$, $T_n \to_{a.s.} \mu_x/\mu_y$ by the SLLN and Theorem 1.10. Hence T_n is asymptotically unbiased, although ET_n may not be defined. In Example 2.34, $X_{(n)}$ has the asymptotic bias $\tilde{b}_{X_{(n)}}(P) = h_n(\theta)EY$, which is of order $n^{-(m+1)^{-1}}$.

When $a_n(T_n - \vartheta) \to_d Y$ with $EY = 0$ (e.g., $T_n = \bar{X}^2$ and $\vartheta = \mu^2$ in Example 2.33), a more precise order of the asymptotic bias of T_n may be obtained (for comparing different estimators in terms of their asymptotic biases). Suppose that there is a sequence of random variables $\{\eta_n\}$ such that

$$a_n\eta_n \to_d Y \qquad \text{and} \qquad a_n^2(T_n - \vartheta - \eta_n) \to_d W, \tag{2.36}$$

where Y and W are random variables with finite means, $EY = 0$ and $EW \neq 0$. Then we may define a_n^{-2} to be the order of $\tilde{b}_{T_n}(P)$ or define EW/a_n^2 to be the a_n^{-2} order asymptotic bias of T_n. However, η_n in (2.36) may not be unique. Some regularity conditions have to be imposed so that the order of asymptotic bias of T_n can be uniquely defined. In the following we focus on the case where $X_1, ..., X_n$ are i.i.d. random k-vectors. Suppose that T_n has the following expansion:

$$T_n - \vartheta = \frac{1}{n}\sum_{i=1}^{n}\phi(X_i) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\psi(X_i, X_j) + o_p\left(\frac{1}{n}\right), \tag{2.37}$$

where ϕ and ψ are functions that may depend on P, $E\phi(X_1) = 0$, $E[\phi(X_1)]^2 < \infty$, $\psi(x, y) = \psi(y, x)$, $E\psi(x, X_1) = 0$ for all x, $E[\psi(X_i, X_j)]^2 < \infty$, $i \leq j$, and $E\psi(X_1, X_1) \neq 0$. From the result for V-statistics in §3.5.3 (Theorem

3.16 and Exercise 113 in §3.6),

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\psi(X_i,X_j)\to_d W,$$

where W is a random variable with $EW = E\psi(X_1,X_1)$. Hence (2.36) holds with $a_n = \sqrt{n}$ and $\eta_n = n^{-1}\sum_{i=1}^{n}\phi(X_i)$. Consequently, we can define $E\psi(X_1,X_1)/n$ to be the n^{-1} order asymptotic bias of T_n. Examples of estimators that have expansion (2.37) are provided in §3.5.3 and §5.2.1. In the following we consider the special case of functions of sample means.

Let $X_1,...,X_n$ be i.i.d. random k-vectors with finite $\Sigma = \mathrm{Var}(X_1)$, $\bar{X} = n^{-1}\sum_{i=1}^{n}X_i$, and $T_n = g(\bar{X})$, where g is a function on $\mathcal{R}^k$ that is second-order differentiable at $\mu = EX_1 \in \mathcal{R}^k$. Consider T_n as an estimator of $\vartheta = g(\mu)$. Using Taylor's expansion, we obtain expansion (2.37) with $\phi(x) = [\nabla g(\mu)]^\tau(x-\mu)$ and $\psi(x,y) = (x-\mu)^\tau\nabla^2 g(\mu)(y-\mu)/2$, where ∇g is the k-vector of partial derivatives of g and $\nabla^2 g$ is the $k\times k$ matrix of second-order partial derivatives of g. By the CLT and Theorem 1.10(iii),

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\psi(X_i,X_j) = \frac{n}{2}(\bar{X}-\mu)^\tau\nabla^2 g(\mu)(\bar{X}-\mu)\to_d \frac{Z_\Sigma^\tau\nabla^2 g(\mu)Z_\Sigma}{2},$$

where $Z_\Sigma = N_k(0,\Sigma)$. Thus,

$$\frac{E[Z_\Sigma^\tau\nabla^2 g(\mu)Z_\Sigma]}{2n} = \frac{\mathrm{tr}\left(\nabla^2 g(\mu)\Sigma\right)}{2n} \tag{2.38}$$

is the n^{-1} order asymptotic bias of $T_n = g(\bar{X})$, where $\mathrm{tr}(A)$ denotes the trace of the matrix A. Note that the quantity in (2.38) is the same as the leading term in the exact bias of $T_n = g(\bar{X})$ obtained under a much more stringent condition on the derivatives of g (Lehmann, 1983, Theorem 2.5.1).

Example 2.35. Let $X_1,...,X_n$ be i.i.d. binary random variables with $P(X_i = 1) = p$, where $p \in (0,1)$ is unknown. Consider first the estimation of $\vartheta = p(1-p)$. Since $\mathrm{Var}(\bar{X}) = p(1-p)/n$, the n^{-1} order asymptotic bias of $T_n = \bar{X}(1-\bar{X})$ according to (2.38) with $g(x) = x(1-x)$ is $-p(1-p)/n$. On the other hand, a direct computation shows $E[\bar{X}(1-\bar{X})] = E\bar{X} - E\bar{X}^2 = p - (E\bar{X})^2 - \mathrm{Var}(\bar{X}) = p(1-p) - p(1-p)/n$. Hence, the exact bias of T_n is the same as the n^{-1} order asymptotic bias.

Consider next the estimation of $\vartheta = p^{-1}$. In this case, there is no unbiased estimator of p^{-1} (Exercise 84 in §2.6). Let $T_n = \bar{X}^{-1}$. Then, an n^{-1} order asymptotic bias of T_n according to (2.38) with $g(x) = x^{-1}$ is $(1-p)/(p^2 n)$. On the other hand, $ET_n = \infty$ for every n. ∎

Like the bias, the mse of an estimator T_n of ϑ, $\mathrm{mse}_{T_n}(P) = E(T_n-\vartheta)^2$, is not well defined if the second moment of T_n does not exist. We now

define a version of *asymptotic mean squared error* (amse) and a measure of assessing different point estimators of a common parameter.

Definition 2.12. Let T_n be an estimator of ϑ for every n and $\{a_n\}$ be a sequence of positive numbers satisfying $a_n \to \infty$ or $a_n \to a > 0$. Assume that $a_n(T_n - \vartheta) \to_d Y$ with $0 < EY^2 < \infty$.
(i) The asymptotic mean squared error of T_n, denoted by $\text{amse}_{T_n}(P)$ or $\text{amse}_{T_n}(\theta)$ if P is in a parametric family indexed by θ, is defined to be the asymptotic expectation of $(T_n - \vartheta)^2$, i.e., $\text{amse}_{T_n}(P) = EY^2/a_n^2$. The asymptotic variance of T_n is defined to be $\sigma^2_{T_n}(P) = \text{Var}(Y)/a_n^2$.
(ii) Let T_n' be another estimator of ϑ. The *asymptotic relative efficiency* of T_n' w.t.r. T_n is defined to be $e_{T_n', T_n}(P) = \text{amse}_{T_n}(P)/\text{amse}_{T_n'}(P)$.
(iii) T_n is said to be *asymptotically more efficient* than T_n' if and only if $\limsup_n e_{T_n', T_n}(P) \le 1$ for any P and < 1 for some P. ∎

The amse and asymptotic variance are the same if and only if $EY = 0$. By Proposition 2.3, the amse or the asymptotic variance of T_n is essentially unique and, therefore, the concept of asymptotic relative efficiency in Definition 2.12(ii)-(iii) is well defined.

In Example 2.33, $\text{amse}_{\bar{X}^2}(P) = \sigma^2_{\bar{X}^2}(P) = 4\mu^2\sigma^2/n$. In Example 2.34, $\sigma^2_{X_{(n)}}(P) = [h_n(\theta)]^2\text{Var}(Y)$ and $\text{amse}_{X_{(n)}}(P) = [h_n(\theta)]^2 EY^2$.

When both $\text{mse}_{T_n}(P)$ and $\text{mse}_{T_n'}(P)$ exist, one may compare T_n and T_n' by evaluating the relative efficiency $\text{mse}_{T_n}(P)/\text{mse}_{T_n'}(P)$. However, this comparison may be different from the one using the asymptotic relative efficiency in Definition 2.12(ii), since the mse and amse of an estimator may be different (Exercise 115 in §2.6). The following result shows that when the exact mse of T_n exists, it is no smaller than the amse of T_n. It also provides a condition under which the exact mse and the amse are the same.

Proposition 2.4. Let T_n be an estimator of ϑ for every n and $\{a_n\}$ be a sequence of positive numbers satisfying $a_n \to \infty$ or $a_n \to a > 0$. Suppose that $a_n(T_n - \vartheta) \to_d Y$ with $0 < EY^2 < \infty$. Then
(i) $EY^2 \le \liminf_n E[a_n^2(T_n - \vartheta)^2]$ and
(ii) $EY^2 = \lim_{n\to\infty} E[a_n^2(T_n - \vartheta)^2]$ if and only if $\{a_n^2(T_n - \vartheta)^2\}$ is uniformly integrable.
Proof. (i) By Theorem 1.10(iii),

$$\min\{a_n^2(T_n - \vartheta)^2, t\} \to_d \min\{Y^2, t\}$$

for any $t > 0$. Since $\min\{a_n^2(T_n - \vartheta)^2, t\}$ is bounded by t,

$$\lim_{n\to\infty} E(\min\{a_n^2(T_n - \vartheta)^2, t\}) = E(\min\{Y^2, t\})$$

(Theorem 1.8(viii)). Then

$$\begin{aligned} EY^2 &= \lim_{t\to\infty} E(\min\{Y^2, t\}) \\ &= \lim_{t\to\infty} \lim_{n\to\infty} E(\min\{a_n^2(T_n-\vartheta)^2, t\}) \\ &= \liminf_{t,n} E(\min\{a_n^2(T_n-\vartheta)^2, t\}) \\ &\le \liminf_{n} E[a_n^2(T_n-\vartheta)^2], \end{aligned}$$

where the third equality follows from the fact that $E(\min\{a_n^2(T_n-\vartheta)^2, t\})$ is nondecreasing in t for any fixed n.
(ii) The result follows from Theorem 1.8(viii). ∎

Example 2.36. Let $X_1, ..., X_n$ be i.i.d. from the Poisson distribution $P(\theta)$ with an unknown $\theta > 0$. Consider the estimation of $\vartheta = P(X_i = 0) = e^{-\theta}$. Let $T_{1n} = F_n(0)$, where F_n is the empirical c.d.f. defined in (2.28). Then T_{1n} is unbiased and has $\mathrm{mse}_{T_{1n}}(\theta) = e^{-\theta}(1-e^{-\theta})/n$. Also, $\sqrt{n}(T_{1n}-\vartheta) \to_d N(0, e^{-\theta}(1-e^{-\theta}))$ by the CLT. Thus, in this case $\mathrm{amse}_{T_{1n}}(\theta) = \mathrm{mse}_{T_{1n}}(\theta)$.

Next, consider $T_{2n} = e^{-\bar{X}}$. Note that $ET_{2n} = e^{n\theta(e^{-1/n}-1)}$. Hence $nb_{T_{2n}}(\theta) \to \theta e^{-\theta}/2$. Using Theorem 1.12 and the CLT, we can show that $\sqrt{n}(T_{2n}-\vartheta) \to_d N(0, e^{-2\theta}\theta)$. By Definition 2.12(i), $\mathrm{amse}_{T_{2n}}(\theta) = e^{-2\theta}\theta/n$. Thus, the asymptotic relative efficiency of T_{1n} w.r.t. T_{2n} is

$$e_{T_{1n},T_{2n}}(\theta) = \theta/(e^\theta - 1),$$

which is always less than 1. This shows that T_{2n} is asymptotically more efficient than T_{1n}. ∎

The result for T_{2n} in Example 2.36 is a special case (with $U_n = \bar{X}$) of the following general result.

Theorem 2.6. Let g be a function on $\mathcal{R}^k$ that is differentiable at $\theta \in \mathcal{R}^k$ and let U_n be a k-vector of statistics satisfying $a_n(U_n - \theta) \to_d Y$ for a random k-vector Y with $0 < E\|Y\|^2 < \infty$ and a sequence of positive numbers $\{a_n\}$ satisfying $a_n \to \infty$. Let $T_n = g(U_n)$ be an estimator of $\vartheta = g(\theta)$. Then, the amse and asymptotic variance of T_n are, respectively, $E\{[\nabla g(\theta)]^\tau Y\}^2/a_n^2$ and $[\nabla g(\theta)]^\tau \mathrm{Var}(Y)\nabla g(\theta)/a_n^2$. ∎

2.5.3 Asymptotic inference

Statistical inference based on asymptotic criteria and approximations is called *asymptotic statistical inference* or simply *asymptotic inference*. We have previously considered asymptotic estimation. We now focus on asymptotic hypothesis tests and confidence sets.

Definition 2.13. Let $X = (X_1, ..., X_n)$ be a sample from $P \in \mathcal{P}$ and $T_n(X)$ be a test for $H_0 : P \in \mathcal{P}_0$ versus $H_1 : P \in \mathcal{P}_1$.
(i) If $\limsup_n \alpha_{T_n}(P) \leq \alpha$ for any $P \in \mathcal{P}_0$, then α is an *asymptotic significance level* of T_n.
(ii) If $\lim_{n\to\infty} \sup_{P\in\mathcal{P}_0} \alpha_{T_n}(P)$ exists, then it is called the *limiting size* of T_n.
(iii) T_n is called *consistent* if and only if the type II error probability converges to 0, i.e., $\lim_{n\to\infty}[1 - \alpha_{T_n}(P)] = 0$, for any $P \in \mathcal{P}_1$.
(iv) T_n is called *Chernoff-consistent* if and only if T_n is consistent *and* the type I error probability converges to 0, i.e., $\lim_{n\to\infty} \alpha_{T_n}(P) = 0$, for any $P \in \mathcal{P}_0$. T_n is called *strongly Chernoff-consistent* if and only if T_n is consistent and the limiting size of T_n is 0. ∎

Obviously if T_n has size (or significance level) α for all n, then its limiting size (or asymptotic significance level) is α. If the limiting size of T_n is $\alpha \in (0, 1)$, then for any $\epsilon > 0$, T_n has size $\alpha + \epsilon$ for all $n \geq n_0$, where n_0 is independent of P. Hence T_n has level of significance $\alpha + \epsilon$ for any $n \geq n_0$. However, if $\mathcal{P}_0$ is not a parametric family, it is likely that the limiting size of T_n is 1 (see, e.g., Example 2.37). This is the reason why we consider the weaker requirement in Definition 2.13(i). If T_n has asymptotic significance level α, then for any $\epsilon > 0$, $\alpha_{T_n}(P) < \alpha + \epsilon$ for all $n \geq n_0(P)$ but $n_0(P)$ depends on $P \in \mathcal{P}_0$; and there is no guarantee that T_n has significance level $\alpha + \epsilon$ for any n.

The consistency in Definition 2.13(iii) only requires that the type II error probability converge to 0. We may define uniform consistency to be $\lim_{n\to\infty} \sup_{P\in\mathcal{P}_1}[1 - \alpha_{T_n}(P)] = 0$, but it is not satisfied in most problems. If $\alpha \in (0, 1)$ is a pre-assigned level of significance for the problem, then a consistent test T_n having asymptotic significance level α is called *asymptotically correct*, and a consistent test having limiting size α is called *strongly* asymptotically correct.

The Chernoff-consistency (or strong Chernoff-consistency) in Definition 2.13(iv) requires that both types of error probabilities converge to 0. Mathematically, Chernoff-consistency (or strong Chernoff-consistency) is better than asymptotic correctness (or strongly asymptotic correctness). After all, both types of error probabilities should decrease to 0 if sampling can be continued indefinitely. However, if α is chosen to be small enough so that error probabilities smaller than α can be practically treated as 0, then the asymptotic correctness (or strongly asymptotic correctness) is enough, and is probably preferred, since requiring an unnecessarily small type I error probability usually results in an unnecessary increase in the type II error probability, as the following example illustrates.

Example 2.37. Consider the testing problem $H_0 : \mu \leq \mu_0$ versus $H_1 :$

$\mu > \mu_0$ based on i.i.d. $X_1, ..., X_n$ with $EX_1 = \mu \in \mathcal{R}$. If each X_i has the $N(\mu, \sigma^2)$ distribution with a known σ^2, then the test T_{c_α} given in Example 2.28 with $c_\alpha = \sigma z_{1-\alpha}/\sqrt{n} + \mu_0$ and $\alpha \in (0, 1)$ has size α (and, therefore, limiting size α). It also follows from (2.32) that for any $\mu > \mu_0$,

$$1 - \alpha_{T_{c_\alpha}}(\mu) = \Phi\left(z_{1-\alpha} + \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma}\right) \to 0 \tag{2.39}$$

as $n \to \infty$. This shows that T_{c_α} is consistent and, hence, is strongly asymptotically correct. Note that the convergence in (2.39) is not uniform in $\mu > \mu_0$, but is uniform in $\mu > \mu_1$ for any fixed $\mu_1 > \mu_0$.

Since the size of T_{c_α} is α for all n, T_{c_α} is not Chernoff-consistent. A strongly Chernoff-consistent test can be obtained as follows. Let

$$\alpha_n = 1 - \Phi(\sqrt{n}a_n), \tag{2.40}$$

where a_n's are positive numbers satisfying $a_n \to 0$ and $\sqrt{n}a_n \to \infty$. Let T_n be T_{c_α} with $\alpha = \alpha_n$ for each n. Then, T_n has size α_n. Since $\alpha_n \to 0$, The limiting size of T_n is 0. On the other hand, (2.39) still holds with α replaced by α_n. This follows from the fact that

$$z_{1-\alpha_n} + \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma} = \sqrt{n}\left(a_n + \frac{\mu_0 - \mu}{\sigma}\right) \to -\infty$$

for any $\mu > \mu_0$. Hence T_n is strongly Chernoff-consistent. However, if $\alpha_n < \alpha$, then, from the left-hand side of (2.39), $1 - \alpha_{T_{c_\alpha}}(\mu) < 1 - \alpha_{T_n}(\mu)$ for any $\mu > \mu_0$.

We now consider the case where the population P is not in a parametric family. We still assume that $\sigma^2 = \mathrm{Var}(X_i)$ is known. Using the CLT, we can show that for $\mu > \mu_0$,

$$\lim_{n\to\infty}[1 - \alpha_{T_{c_\alpha}}(\mu)] = \lim_{n\to\infty} \Phi\left(z_{1-\alpha} + \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma}\right) = 0,$$

i.e., T_{c_α} is still consistent. For $\mu \le \mu_0$,

$$\lim_{n\to\infty} \alpha_{T_{c_\alpha}}(\mu) = 1 - \lim_{n\to\infty} \Phi\left(z_{1-\alpha} + \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma}\right),$$

which equals α if $\mu = \mu_0$ and 0 if $\mu < \mu_0$. Thus, the asymptotic significance level of T_{c_α} is α. Combining these two results, we know that T_{c_α} is asymptotically correct. However, if $\mathcal{P}$ contains all possible populations on $\mathcal{R}$ with finite second moments, then one can show that the limiting size of T_{c_α} is 1 (exercise). For α_n defined by (2.40), we can show that $T_n = T_{c_\alpha}$ with $\alpha = \alpha_n$ is Chernoff-consistent (exercise). But T_n is not strongly Chernoff-consistent if $\mathcal{P}$ contains all possible populations on $\mathcal{R}$ with finite second moments. ∎

Definition 2.14. Let $X = (X_1, ..., X_n)$ be a sample from $P \in \mathcal{P}$, ϑ be a k-vector of parameters related to P, and $C(X)$ be a confidence set for ϑ.
(i) If $\liminf_n P(\vartheta \in C(X)) \geq 1 - \alpha$ for any $P \in \mathcal{P}$, then $1 - \alpha$ is an *asymptotic significance level* of $C(X)$.
(ii) If $\lim_{n\to\infty} \inf_{P\in\mathcal{P}} P(\vartheta \in C(X))$ exists, then it is called the *limiting confidence coefficient* of $C(X)$. ∎

Note that the asymptotic significance level and limiting confidence coefficient of a confidence set are very similar to the asymptotic significance level and limiting size of a test, respectively. Some conclusions are also similar. For example, in a parametric problem one can often find a confidence set having limiting confidence coefficient $1 - \alpha \in (0, 1)$, which implies that for any $\epsilon > 0$, the confidence coefficient of $C(X)$ is $1 - \alpha - \epsilon$ for all $n \geq n_0$, where n_0 is independent of P; in a nonparametric problem the limiting confidence coefficient of $C(X)$ might be 0, whereas $C(X)$ may have asymptotic significance level $1 - \alpha \in (0, 1)$, but for any fixed n, the confidence coefficient of $C(X)$ might be 0.

The confidence interval in Example 2.31 with $c = \sigma z_{1-\alpha/2}/\sqrt{n}$ and the confidence set in Example 2.32 have confidence coefficient $1 - \alpha$ for any n and, therefore, have limiting confidence coefficient $1 - \alpha$. If we drop the normality assumption and assume $EX_i^4 < \infty$, then these confidence sets have asymptotic significance level $1-\alpha$; their limiting confidence coefficients may be 0 (exercise).

2.6 Exercises

1. Consider Example 2.3. Suppose that $p(\boldsymbol{s})$ is constant. Show that X_i and X_j, $i \neq j$, are not uncorrelated and, hence, $X_1, ..., X_n$ are not independent. Furthermore, when y_i's are either 0 or 1, show that $Z = \sum_{i=1}^n X_i$ has a hypergeometric distribution and compute the mean of Z.

2. Consider Example 2.3. Suppose that we do not require that the elements in $\boldsymbol{s}$ be distinct, i.e., we consider sampling with replacement. Define a probability measure p and a sample $(X_1, ..., X_n)$ such that (2.3) holds. If $p(\boldsymbol{s})$ is constant, are $X_1, ..., X_n$ independent? If $p(\boldsymbol{s})$ is constant and y_i's are either 0 or 1, what are the distribution and mean of $Z = \sum_{i=1}^n X_i$?

3. Show that $\{P_\theta : \theta \in \Theta\}$ is an exponential family and find its canonical form and natural parameter space, when
   (a) P_θ is the Poisson distribution $P(\theta)$, $\theta \in \Theta = (0, \infty)$;
   (b) P_θ is the negative binomial distribution $NB(\theta, r)$ with a fixed r,

$\theta \in \Theta = (0, 1)$;
(c) P_θ is the exponential distribution $E(a, \theta)$ with a fixed a, $\theta \in \Theta = (0, \infty)$;
(d) P_θ is the gamma distribution $\Gamma(\alpha, \gamma)$, $\theta = (\alpha, \gamma) \in \Theta = (0, \infty) \times (0, \infty)$;
(e) P_θ is the beta distribution $B(\alpha, \beta)$, $\theta = (\alpha, \beta) \in \Theta = (0, 1) \times (0, 1)$;
(f) P_θ is the Weibull distribution $W(\alpha, \theta)$ with a fixed $\alpha > 0$, $\theta \in \Theta = (0, \infty)$.

4. Show that the family of exponential distributions $E(a, \theta)$ with two unknown parameters a and θ is not an exponential family.

5. Show that the family of negative binomial distributions $NB(p, r)$ with two unknown parameters p and r is not an exponential family.

6. Show that the family of Cauchy distributions $C(\mu, \sigma)$ with two unknown parameters μ and σ is not an exponential family.

7. Show that the family of Weibull distributions $W(\alpha, \theta)$ with two unknown parameters α and θ is not an exponential family.

8. Is the family of log-normal distributions $LN(\mu, \sigma^2)$ with two unknown parameters μ and σ^2 an exponential family?

9. Show that the family of double exponential distributions $DE(\mu, \theta)$ with two unknown parameters μ and θ is not an exponential family, but the family of double exponential distributions $DE(\mu, \theta)$ with a fixed μ and an unknown parameter θ is an exponential family.

10. Show that the k-dimensional normal family discussed in Example 2.4 is an exponential family. Identify the functions T, η, ξ, and h.

11. Obtain the variance-covariance matrix for $(X_1, ..., X_k)$ in Example 2.7, using (a) Theorem 2.1(ii) and (b) direct computation.

12. Show that the m.g.f. of the gamma distribution $\Gamma(\alpha, \gamma)$ is $(1 - \gamma t)^{-\alpha}$, $t < \gamma^{-1}$, using Theorem 2.1(ii).

13. A discrete random variable X with

$$P(X = x) = \gamma(x)\theta^x / c(\theta), \quad x = 0, 1, 2, ...,$$

where $\gamma(x) \geq 0$, $\theta > 0$, and $c(\theta) = \sum_{x=0}^{\infty} \gamma(x)\theta^x$, is called a random variable with a *power series* distribution.
(a) Show that $\{\gamma(x)\theta^x / c(\theta) : \theta > 0\}$ is an exponential family.
(b) Suppose that $X_1, ..., X_n$ are i.i.d. with a power series distribution $\gamma(x)\theta^x / c(\theta)$. Show that $\sum_{i=1}^{n} X_i$ has the power series distribution $\gamma_n(x)\theta^x / [c(\theta)]^n$, where $\gamma_n(x)$ is the coefficient of θ^x in the power series expansion of $[c(\theta)]^n$.

14. Let X be a random variable with a p.d.f. f_θ in an exponential family $\{P_\theta : \theta \in \Theta\}$ and let A be a Borel set. Show that the distribution of X truncated on A (i.e., the conditional distribution of X given $X \in A$) has a p.d.f. $f_\theta I_A / P_\theta(A)$ that is in an exponential family.

15. Let $\{P_{(\mu,\Sigma)} : \mu \in \mathcal{R}^k, \Sigma \in \mathcal{M}_k\}$ be a location-scale family on $\mathcal{R}^k$. Suppose that $P_{(0,I_k)}$ has a Lebesgue p.d.f. that is always positive and that the mean and variance-covariance matrix of $P_{(0,I_k)}$ are 0 and I_k, respectively. Show that the mean and variance-covariance matrix of $P_{(\mu,\Sigma)}$ are μ and Σ, respectively.

16. Show that if the distribution of a positive random variable X is in a scale family, then the distribution of $\log X$ is in a location family.

17. Let X be a random variable having the gamma distribution $\Gamma(\alpha, \gamma)$ with a known α and an unknown $\gamma > 0$ and let $Y = \sigma \log X$.
(a) Show that if $\sigma > 0$ is unknown, then the distribution of Y is in a location-scale family.
(b) Show that if $\sigma > 0$ is known, then the distribution of Y is in an exponential family.

18. Let $X_1, ..., X_n$ be i.i.d. random variables having a finite $E|X_1|^4$ and let $\bar{X}$ and S^2 be the sample mean and variance defined by (2.1) and (2.2). Express $E(\bar{X}^3)$, $\mathrm{Cov}(\bar{X}, S^2)$, and $\mathrm{Var}(S^2)$ in terms of $\mu_k = EX_1^k$, $k = 1, 2, 3, 4$. Find a condition under which $\bar{X}$ and S^2 are uncorrelated.

19. Let $X_1, ..., X_n$ be i.i.d. random variables having the gamma distribution $\Gamma(\alpha, \gamma_x)$ and $Y_1, ..., Y_n$ be i.i.d. random variables having the gamma distribution $\Gamma(\alpha, \gamma_y)$, where $\alpha > 0$, $\gamma_x > 0$, and $\gamma_y > 0$. Assume that X_i's and Y_i's are independent. Derive the distribution of the statistic $\bar{X}/\bar{Y}$, where $\bar{X}$ and $\bar{Y}$ are the sample means based on X_i's and Y_i's, respectively.

20. Let $X_1, ..., X_n$ be i.i.d. random variables having the exponential distribution $E(a, \theta)$, $a \in \mathcal{R}$, and $\theta > 0$. Show that the smallest order statistic, $X_{(1)}$, has the exponential distribution $E(a, \theta/n)$ and that $2\sum_{i=1}^n (X_i - X_{(1)})/\theta$ has the chi-square distribution χ^2_{2n-2}.

21. Let $(X_1, Y_1), ..., (X_n, Y_n)$ be i.i.d. random 2-vectors. Suppose that X_1 has the Cauchy distribution $C(0, 1)$ and given $X_1 = x$, Y_1 has the Cauchy distribution $C(\beta x, 1)$, where $\beta \in \mathcal{R}$. Let $\bar{X}$ and $\bar{Y}$ be the sample means based on X_i's and Y_i's, respectively. Obtain the marginal distributions of $\bar{Y}$, $\bar{Y} - \beta\bar{X}$, and $\bar{Y}/\bar{X}$.

22. Let $X_i = (Y_i, Z_i)$, $i = 1, ..., n$, be i.i.d. random 2-vectors. The sample correlation coefficient is defined to be

$$T(X) = \frac{1}{(n-1)\sqrt{S_Y^2 S_Z^2}} \sum_{i=1}^{n} (Y_i - \bar{Y})(Z_i - \bar{Z}),$$

where $\bar{Y} = n^{-1}\sum_{i=1}^n Y_i$, $\bar{Z} = n^{-1}\sum_{i=1}^n Z_i$, $S_Y^2 = (n-1)^{-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$, and $S_Z^2 = (n-1)^{-1}\sum_{i=1}^n (Z_i - \bar{Z})^2$.
(a) Assume that $E|Y_i|^4 < \infty$ and $E|Z_i|^4 < \infty$. Show that

$$\sqrt{n}[T(X) - \rho] \to_d N(0, c^2),$$

where ρ is the correlation coefficient between Y_1 and Z_1 and c is a constant depending on some unknown parameters.
(b) Assume that Y_i and Z_i are independently distributed as $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively. Show that T has the Lebesgue p.d.f.

$$f(t) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{n-2}{2}\right)}(1-t^2)^{(n-4)/2} I_{(-1,1)}(t).$$

(c) Assume the conditions in (b). Obtain the result in (a) using Scheffé's theorem (Proposition 1.18).

23. Let $X_1, ..., X_n$ be i.i.d. random variables with $EX_1^4 < \infty$, $T = (Y, Z)$, and $T_1 = Y/\sqrt{Z}$, where $Y = n^{-1}\sum_{i=1}^n |X_i|$ and $Z = n^{-1}\sum_{i=1}^n X_i^2$.
(a) Show that $\sqrt{n}(T - \theta) \to_d N_2(0, \Sigma)$ and $\sqrt{n}(T_1 - \vartheta) \to_d N(0, c^2)$. Identify θ, Σ, ϑ, and c^2 in terms of moments of X_1.
(b) Repeat (a) when X_1 has the normal distribution $N(0, \sigma^2)$.
(c) Repeat (a) when X_1 has the double exponential distribution $D(0, \sigma)$.

24. Prove the claims in Example 2.9 for the distributions related to order statistics.

25. Show that if T is a sufficient statistic and $T = \psi(S)$, where ψ is measurable and S is another statistic, then S is sufficient.

26. In the proof of Lemma 2.1, show that $C_0 \in \mathcal{C}$. Also, prove Lemma 2.1 when $\mathcal{P}$ is dominated by a σ-finite measure.

27. Let $X_1, ..., X_n$ be i.i.d. random variables from $P_\theta \in \{P_\theta : \theta \in \Theta\}$. In the following cases, find a sufficient statistic for $\theta \in \Theta$ that has the same dimension as θ.
(a) P_θ is the Poisson distribution $P(\theta)$, $\theta \in (0, \infty)$.
(b) P_θ is the negative binomial distribution $NB(\theta, r)$ with a known r, $\theta \in (0, 1)$.

(c) P_θ is the exponential distribution $E(0,\theta)$, $\theta \in (0,\infty)$.
(d) P_θ is the gamma distribution $\Gamma(\alpha,\gamma)$, $\theta = (\alpha,\gamma) \in (0,\infty)\times(0,\infty)$.
(e) P_θ is the beta distribution $B(\alpha,\beta)$, $\theta = (\alpha,\beta) \in (0,1)\times(0,1)$.
(f) P_θ is the log-normal distribution $LN(\mu,\sigma^2)$, $\theta = (\mu,\sigma^2) \in \mathcal{R}\times(0,\infty)$.
(g) P_θ is the Weibull distribution $W(\alpha,\theta)$ with a known $\alpha > 0$, $\theta \in (0,\infty)$.

28. Let $X_1,...,X_n$ be i.i.d. random variables from $P_{(a,\theta)}$, where $(a,\theta) \in \mathcal{R}^2$ is a parameter. Find a two-dimensional sufficient statistic for (a,θ) in the following cases.
(a) $P_{(a,\theta)}$ is the exponential distribution $E(a,\theta)$, $a \in \mathcal{R}$, $\theta \in (0,\infty)$.
(b) $P_{(a,\theta)}$ is the Pareto distribution $Pa(a,\theta)$, $a \in (0,\infty)$, $\theta \in (0,\infty)$.

29. In Example 2.11, show that $X_{(1)}$ (or $X_{(n)}$) is sufficient for a (or b) if we consider a subfamily $\{f_{(a,b)} : a < b\}$ with a fixed b (or a).

30. Let X and Y be two random variables such that Y has the binomial distribution $Bi(\pi,N)$ and, given $Y = y$, X has the binomial distribution $Bi(p,y)$.
(a) Suppose that $p \in (0,1)$ and $\pi \in (0,1)$ are unknown and N is known. Show that (X,Y) is minimal sufficient for (p,π).
(b) Suppose that π and N are known and $p \in (0,1)$ is unknown. Show whether X is sufficient for p and whether Y is sufficient for p.

31. Let $X_1,...,X_n$ be i.i.d. random variables having a distribution $P \in \mathcal{P}$, where $\mathcal{P}$ is the family of distributions on $\mathcal{R}$ having continuous c.d.f.'s. Let $T = (X_{(1)},...,X_{(n)})$ be the vector of order statistics. Show that, given T, the conditional distribution of $X = (X_1,...,X_n)$ is a discrete distribution putting probability $1/n!$ on each of the $n!$ points $(X_{i_1},...,X_{i_n}) \in \mathcal{R}^n$, where $\{i_1,...,i_n\}$ is a permutation of $\{1,...,n\}$; hence, T is sufficient for $P \in \mathcal{P}$.

32. In Example 2.13 and Example 2.14, show that T is minimal sufficient for θ by using Theorem 2.3(iii).

33. A coin has probability p of coming up heads and $1-p$ of coming up tails, where $p \in (0,1)$. The first stage of an experiment consists of tossing this coin a known total of M times and recording X, the number of heads. In the second stage, the coin is tossed until a total of $X+1$ tails have come up. The number Y of heads observed in the second stage along the way to getting the $X+1$ tails is then recorded. This experiment is repeated independently a total of n times and the two counts (X_i,Y_i) for the ith experiment are recorded, $i = 1,...,n$. Obtain a statistic that is minimal sufficient for p and derive its distribution.

34. Let $X_1, ..., X_n$ be i.i.d. random variables having the Lebesgue p.d.f.

$$f_\theta(x) = \exp\left\{-\left(\tfrac{x-\mu}{\sigma}\right)^4 - \xi(\theta)\right\},$$

where $\theta = (\mu, \sigma) \in \Theta = \mathcal{R} \times (0, \infty)$. Show that $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is an exponential family, where P_θ is the joint distribution of $X_1, ..., X_n$, and that the statistic $T = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i^3, \sum_{i=1}^n X_i^4\right)$ is minimal sufficient for $\theta \in \Theta$.

35. Let $X_1, ..., X_n$ be i.i.d. random variables having the Lebesgue p.d.f.

$$f_\theta(x) = (2\theta)^{-1}\left[I_{(0,\theta)}(x) + I_{(2\theta,3\theta)}(x)\right].$$

Find a minimal sufficient statistic for $\theta \in (0, \infty)$.

36. Let $X_1, ..., X_n$ be i.i.d. random variables having the Cauchy distribution $C(\mu, \sigma)$ with unknown $\mu \in \mathcal{R}$ and $\sigma > 0$. Show that the vector of order statistics is minimal sufficient for (μ, σ).

37. Let $X_1, ..., X_n$ be i.i.d. random variables having the double exponential distribution $DE(\mu, \theta)$ with unknown $\mu \in \mathcal{R}$ and $\theta > 0$. Show that the vector of order statistics is minimal sufficient for (μ, θ).

38. Let $X_1, ..., X_n$ be i.i.d. random variables having the Weibull distribution $W(\alpha, \theta)$ with unknown $\alpha > 0$ and $\theta > 0$. Show that the vector of order statistics is minimal sufficient for (α, θ).

39. Let $X_1, ..., X_n$ be i.i.d. random variables having the beta distribution $B(\beta, \beta)$ with an unknown $\beta > 0$. Find a minimal sufficient statistic for β.

40. Let $X_1, ..., X_n$ be i.i.d. random variables having a population P in a parametric family indexed by (θ, j), where $\theta > 0$, $j = 1, 2$, and $n \geq 2$. When $j = 1$, P is the $N(0, \theta^2)$ distribution. When $j = 2$, P is the double exponential distribution $DE(0, \theta)$. Show that $T = (\sum_{i=1}^n X_i^2, \sum_{i=1}^n |X_i|)$ is minimal sufficient for (θ, j).

41. Let $X_1, ..., X_n$ be i.i.d. random variables having a population P in a parametric family indexed by (θ, j), where $\theta \in (0, 1)$, $j = 1, 2$, and $n \geq 2$. When $j = 1$, P is the Poisson distribution $P(\theta)$. When $j = 2$, P is the binomial distribution $Bi(\theta, 1)$.
(a) Show that $T = \sum_{i=1}^n X_i$ is not sufficient for (θ, j).
(b) Find a two-dimensional minimal sufficient statistic for (θ, j).

42. Let X be a sample from $P \in \mathcal{P} = \{f_{\theta,j} : \theta \in \Theta, j = 1, ..., k\}$, where $f_{\theta,j}$'s are p.d.f.'s w.r.t. a common σ-finite measure and Θ is a set of parameters. Assume that $\{x : f_{\theta,j}(x) > 0\} \subset \{x : f_{\theta,k}(x) > 0\}$ for all

θ and $j = 1, ..., k-1$. Suppose that for each fixed j, $T = T(X)$ is a statistic sufficient for θ.
(a) Obtain a k-dimensional statistic that is sufficient for (θ, j).
(b) Derive a sufficient condition under which T is minimal sufficient for (θ, j).

43. A box has an unknown odd number of balls labeled consecutively as $-\theta, -(\theta-1), ..., -2, -1, 0, 1, 2, ..., (\theta-1), \theta$, where θ is an unknown nonnegative integer. A simple random sample $X_1, ..., X_n$ is taken without replacement, where X_i is the label on the ith ball selected and $n < 2\theta + 1$.
(a) Find a statistic that is minimal sufficient for θ and derive its distribution.
(b) Show that the minimal sufficient statistic in (a) is also complete.

44. Let $X_1, ..., X_n$ be i.i.d. random variables having the Lebesgue p.d.f. $\theta^{-1} e^{-(x-\theta)/\theta} I_{(\theta,\infty)}(x)$, where $\theta > 0$ is an unknown parameter.
(a) Find a statistic that is minimal sufficient for θ.
(b) Show whether the minimal sufficient statistic in (a) is complete.

45. Let $X_1, ..., X_n$ $(n \geq 2)$ be i.i.d. random variables having the normal distribution $N(\theta, 2)$ when $\theta = 0$ and the normal distribution $N(\theta, 1)$ when $\theta \in \mathcal{R}$ and $\theta \neq 0$. Show that the sample mean $\bar{X}$ is a complete statistic for θ but it is not a sufficient statistic for θ.

46. Let X be a random variable with a distribution P_θ in $\{P_\theta : \theta \in \Theta\}$, f_θ be the p.d.f. of P_θ w.r.t. a measure ν, A be an event, and $\mathcal{P}_A = \{f_\theta I_A / P_\theta(A) : \theta \in \Theta\}$.
(a) Show that if $T(X)$ is sufficient for $P_\theta \in \mathcal{P}$, then it is sufficient for $P_\theta \in \mathcal{P}_A$.
(b) Show that if T is sufficient and complete for $P_\theta \in \mathcal{P}$, then it is complete for $P_\theta \in \mathcal{P}_A$.

47. Show that $(X_{(1)}, X_{(n)})$ in Example 2.13 is not complete.

48. Let T be a complete (or boundedly complete) and sufficient statistic. Suppose that there is a minimal sufficient statistic S. Show that T is minimal sufficient and S is complete (or boundedly complete).

49. Let T and S be two statistics such that $S = \psi(T)$ for a measurable ψ. Show that
(a) if T is complete, then S is complete;
(b) if T is complete and sufficient and ψ is one-to-one, then S is complete and sufficient;
(c) the results in (a) and (b) still hold if the completeness is replaced by the bounded completeness.

50. Find complete and sufficient statistics for the families in Exercises 27 and 28.

51. Show that $(X_{(1)}, X_{(n)})$ in Example 2.11 is complete.

52. Let $(X_1, Y_1), ..., (X_n, Y_n)$ be i.i.d. random 2-vectors having the following Lebesgue p.d.f.

$$f_\theta(x, y) = (2\pi\gamma^2)^{-1} I_{(0,\gamma)}\left(\sqrt{(x-a)^2 + (y-b)^2}\right), \quad (x, y) \in \mathcal{R}^2,$$

where $\theta = (a, b, \gamma) \in \mathcal{R}^2 \times (0, \infty)$.
(a) If $a = 0$ and $b = 0$, find a complete and sufficient statistic for γ.
(b) If all parameters are unknown, show that the convex hull of the sample points is a sufficient statistic for θ.

53. Let X be a discrete random variable with p.d.f.

$$f_\theta(x) = \begin{cases} \theta & x = 0 \\ (1-\theta)^2\theta^{x-1} & x = 1, 2, ... \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta \in (0, 1)$. Show that X is boundedly complete, but not complete.

54. Show that the sufficient statistic T in Example 2.10 is also complete without using Proposition 2.1.

55. Let $Y_1, ..., Y_n$ be i.i.d. random variables having the Lebesgue p.d.f. $\lambda x^{\lambda-1} I_{(0,1)}(x)$ with an unknown $\lambda > 0$ and let $Z_1, ..., Z_n$ be i.i.d. discrete random variables having the power series distribution given in Exercise 13 with an unknown $\theta > 0$. Assume that Y_i's and Z_j's are independent. Let $X_i = Y_i + Z_i$, $i = 1, ..., n$. Find a complete and sufficient statistic for the unknown parameter (θ, λ) based on the sample $X = (X_1, ..., X_n)$.

56. Suppose that $(X_1, Y_1), ..., (X_n, Y_n)$ are i.i.d. random 2-vectors and X_i and Y_i are independently distributed as $N(\mu, \sigma_X^2)$ and $N(\mu, \sigma_Y^2)$, respectively, with $\theta = (\mu, \sigma_X^2, \sigma_Y^2) \in \mathcal{R} \times (0, \infty) \times (0, \infty)$. Let $\bar{X}$ and S_X^2 be the sample mean and variance given by (2.1) and (2.2) for X_i's and $\bar{Y}$ and S_Y^2 be the sample mean and variance for Y_i's. Show that $T = (\bar{X}, \bar{Y}, S_X^2, S_Y^2)$ is minimal sufficient for θ but T is not boundedly complete.

57. Let $X_1, ..., X_n$ be i.i.d. from the $N(\theta, \theta^2)$ distribution, where $\theta > 0$ is a parameter. Find a minimal sufficient statistic for θ and show whether it is complete.

58. Suppose that $(X_1, Y_1), ..., (X_n, Y_n)$ are i.i.d. random 2-vectors having the normal distribution with $EX_1 = EY_1 = 0$, $\mathrm{Var}(X_1) = \mathrm{Var}(Y_1) = 1$, and $\mathrm{Cov}(X_1, Y_1) = \theta \in (-1, 1)$.
(a) Find a minimal sufficient statistic for θ.
(b) Show whether the minimal sufficient statistic in (a) is complete or not.
(c) Prove that $T_1 = \sum_{i=1}^n X_i^2$ and $T_2 = \sum_{i=1}^n Y_i^2$ are both ancillary but (T_1, T_2) is not ancillary.

59. Let $X_1, ..., X_n$ be i.i.d. random variables having the exponential distribution $E(a, \theta)$.
(a) Show that $\sum_{i=1}^n (X_i - X_{(1)})$ and $X_{(1)}$ are independent for any (a, θ).
(b) Show that $Z_i = (X_{(n)} - X_{(i)})/(X_{(n)} - X_{(n-1)})$, $i = 1, ..., n-2$, are independent of $(X_{(1)}, \sum_{i=1}^n (X_i - X_{(1)}))$.

60. Let $X_1, ..., X_n$ be i.i.d. random variables having the gamma distribution $\Gamma(\alpha, \gamma)$. Show that $\sum_{i=1}^n X_i$ and $\sum_{i=1}^n [\log X_i - \log X_{(1)}]$ are independent for any (α, γ).

61. Let $X_1, ..., X_n$ be i.i.d. random variables having the uniform distribution on the interval (a, b), where $-\infty < a < b < \infty$. Show that $(X_{(i)} - X_{(1)})/(X_{(n)} - X_{(1)})$, $i = 2, ..., n-1$, are independent of $(X_{(1)}, X_{(n)})$ for any a and b.

62. Consider Example 2.19. Assume that $n > 2$.
(a) Show that $\bar{X}$ is better than T_1 if $P = N(\theta, \sigma^2)$, $\theta \in \mathcal{R}$, $\sigma > 0$.
(b) Show that T_1 is better than $\bar{X}$ if P is the uniform distribution on the interval $(\theta - \frac{1}{2}, \theta + \frac{1}{2})$, $\theta \in \mathcal{R}$.
(c) Find a family $\mathcal{P}$ for which neither $\bar{X}$ nor T_1 is better than the other.

63. Let $X_1, ..., X_n$ be i.i.d. from the $N(\mu, \sigma^2)$ distribution, where $\mu \in \mathcal{R}$ and $\sigma > 0$. Consider the estimation of σ^2 with the squared error loss. Show that $\frac{n-1}{n} S^2$ is better than S^2, the sample variance. Can you find an estimator of the form cS^2 with a nonrandom c such that it is better than $\frac{n-1}{n} S^2$?

64. Let $X_1, ..., X_n$ be i.i.d. binary random variables with $P(X_i = 1) = \theta \in (0, 1)$. Consider estimating θ with the squared error loss. Calculate the risks of the following estimators:
(a) the nonrandomized estimators $\bar{X}$ (the sample mean) and

$$T_0(X) = \begin{cases} 0 & \text{if more than half of } X_i\text{'s are 0} \\ 1 & \text{if more than half of } X_i\text{'s are 1} \\ \frac{1}{2} & \text{if exactly half of } X_i\text{'s are 0;} \end{cases}$$

(b) the randomized estimators

$$T_1(X) = \begin{cases} \bar{X} & \text{with probability } \frac{1}{2} \\ T_0 & \text{with probability } \frac{1}{2} \end{cases}$$

and

$$T_2(X) = \begin{cases} \bar{X} & \text{with probability } \bar{X} \\ \frac{1}{2} & \text{with probability } 1 - \bar{X}. \end{cases}$$

65. Let $X_1, ..., X_n$ be i.i.d. random variables having the exponential distribution $E(0, \theta)$, $\theta \in (0, \infty)$. Consider estimating θ with the squared error loss. Calculate the risks of the sample mean $\bar{X}$ and $cX_{(1)}$, where c is a positive constant. Is $\bar{X}$ better than $cX_{(1)}$ for some c?

66. Consider the estimation of an unknown parameter $\theta \geq 0$ under the squared error loss. Show that if T and U are two estimators such that $T \leq U$ and $R_T(P) < R_U(P)$, then $R_{T_+}(P) < R_{U_+}(P)$, where $R_T(P)$ is the risk of an estimator T and T_+ denotes the positive part of T.

67. Let $X_1, ..., X_n$ be i.i.d. random variables having the exponential distribution $E(0, \theta)$, $\theta \in (0, \infty)$. Consider the hypotheses

$$H_0 : \theta \leq \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0,$$

where $\theta_0 > 0$ is a fixed constant. Obtain the risk function (in terms of θ) of the test rule $T_c(X) = I_{(c,\infty)}(\bar{X})$, under the 0-1 loss.

68. Let $X_1, ..., X_n$ be i.i.d. random variables having the Cauchy distribution $C(\mu, \sigma)$ with unknown $\mu \in \mathcal{R}$ and $\sigma > 0$. Consider the hypotheses

$$H_0 : \mu \leq \mu_0 \quad \text{versus} \quad H_1 : \mu > \mu_0,$$

where μ_0 is a fixed constant. Obtain the risk function of the test rule $T_c(X) = I_{(c,\infty)}(\bar{X})$, under the 0-1 loss.

69. Let $X_1, ..., X_n$ be i.i.d. binary random variables with $P(X_i = 1) = \theta$, where $\theta \in (0, 1)$ is unknown and n is an even integer. Consider the problem of testing $H_0 : \theta \leq 0.5$ versus $H_1 : \theta > 0.5$ with action space $\{0, 1\}$ (0 means H_0 is accepted and 1 means H_1 is accepted). Let the loss function be $L(\theta, a) = 0$ if H_j is true and $a = j$, $j = 0, 1$; $L(\theta, 0) = C_0$ when $\theta > 0.5$; and $L(\theta, 1) = C_1$ when $\theta \leq 0.5$, where $C_0 > C_1 > 0$ are some constants. Calculate the risk function of the following randomized test (decision rule):

$$T = \begin{cases} 0 & \text{if more than half of } X_i\text{'s are 0} \\ 1 & \text{if more than half of } X_i\text{'s are 1} \\ \frac{1}{2} & \text{if exactly half of } X_i\text{'s are 0.} \end{cases}$$

70. Consider Example 2.21. Suppose that our decision rule, based on a sample $X = (X_1, ..., X_n)$ with i.i.d. components from the $N(\theta, 1)$ distribution with an unknown $\theta > 0$, is

$$T(X) = \begin{cases} a_1 & b_1 < \bar{X} \\ a_2 & b_0 < \bar{X} \le b_1 \\ a_3 & \bar{X} \le b_0. \end{cases}$$

Express the risk of T in terms of θ.

71. Consider an estimation problem with $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ (a parametric family), $\mathbb{A} = \Theta$, and the squared error loss. If $\theta_0 \in \Theta$ satisfies that $P_\theta \ll P_{\theta_0}$ for any $\theta \in \Theta$, show that the estimator $T \equiv \theta_0$ is admissible.

72. Let $\Im$ be a class of decision rules. A subclass $\Im_0 \subset \Im$ is called $\Im$-*complete* if and only if, for any $T \in \Im$ and $T \notin \Im_0$, there is a $T_0 \in \Im_0$ that is better than T, and $\Im_0$ is called $\Im$-*minimal complete* if and only if $\Im_0$ is $\Im$-complete and no proper subclass of $\Im_0$ is $\Im$-complete. Show that if a $\Im$-minimal complete class exists, then it is exactly the class of $\Im$-admissible rules.

73. Let $X_1, ..., X_n$ be i.i.d. random variables having a distribution $P \in \mathcal{P}$. Assume that $EX_1^2 < \infty$. Consider estimating $\mu = EX_1$ under the squared error loss.
(a) Show that any estimator of the form $a\bar{X} + b$ is inadmissible, where $\bar{X}$ is the sample mean, a and b are constants, and $a > 1$.
(b) Show that any estimator of the form $\bar{X} + b$ is inadmissible, where $b \neq 0$ is a constant.

74. Consider an estimation problem with $\vartheta \in [c, d] \subset \mathcal{R}$, where c and d are known. Suppose that the action space is $\mathbb{A} \supset [c, d]$ and the loss function is $L(|\vartheta - a|)$, where $L(\cdot)$ is an increasing function on $[0, \infty)$. Show that any decision rule T with $P(T(X) \notin [c, d]) > 0$ for some $P \in \mathcal{P}$ is inadmissible.

75. Suppose that the action space is $(\Omega, \mathcal{B}_\Omega^k)$, where $\Omega \in \mathcal{B}^k$. Let X be a sample from $P \in \mathcal{P}$, $\delta_0(X)$ be a nonrandomized rule, and T be a sufficient statistic for $P \in \mathcal{P}$. Show that if $E[I_A(\delta_0(X))|T]$ is a nonrandomized rule, i.e., $E[I_A(\delta_0(X))|T] = I_A(h(T))$ for any $A \in \mathcal{B}_\Omega^k$, where h is a Borel function, then $\delta_0(X) = h(T(X))$ a.s. P.

76. Let T, δ_0, and δ_1 be as given in the statement of Proposition 2.2. Show that

$$\int_{\mathbb{A}} L(P, a) d\delta_1(X, a) = E\left[\int_{\mathbb{A}} L(P, a) d\delta_0(X, a) \middle| T\right] \quad \text{a.s. } P.$$

77. Prove Theorem 2.5.

78. In Exercise 64, use Theorem 2.5 to find decision rules that are better than T_j, $j = 0, 1, 2$.

79. In Exercise 65, use Theorem 2.5 to find a decision rule better than $cX_{(1)}$.

80. Consider Example 2.22.
(a) Show that there is no optimal rule if $\Im$ contains all possible estimators. (Hint: consider constant estimators.)
(b) Find a $\Im_2$-optimal rule if $X_1, ..., X_n$ are independent random variables having a common mean μ and $\mathrm{Var}(X_i) = \sigma^2/a_i$ with known a_i, $i = 1, ..., n$.
(c) Find a $\Im_2$-optimal rule if $X_1, ..., X_n$ are identically distributed but are correlated with a common correlation coefficient ρ.

81. Let $X_{ij} = \mu + a_i + \epsilon_{ij}$, $i = 1, ..., m$, $j = 1, ..., n$, where a_i's and ϵ_{ij}'s are independent random variables, a_i is $N(0, \sigma_a^2)$, ϵ_{ij} is $N(0, \sigma_e^2)$, and μ, σ_a^2, and σ_e^2 are unknown parameters. Define $\bar{X}_i = n^{-1}\sum_{j=1}^n X_{ij}$, $\bar{X} = m^{-1}\sum_{i=1}^m \bar{X}_i$, $\mathrm{MSA} = n(m-1)^{-1}\sum_{i=1}^m (\bar{X}_i - \bar{X})^2$, and $\mathrm{MSE} = m^{-1}(n-1)^{-1}\sum_{i=1}^m \sum_{j=1}^n (X_{ij} - \bar{X}_i)^2$. Assume that $m(n-1) > 4$. Consider the following class of estimators of $\theta = \sigma_a^2/\sigma_e^2$:

$$\left\{ \hat{\theta}(\delta) = \frac{1}{n}\left[(1-\delta)\frac{\mathrm{MSA}}{\mathrm{MSE}} - 1 \right] : \delta \in \mathcal{R} \right\}.$$

(a) Show that MSA and MSE are independent.
(b) Obtain a $\delta \in \mathcal{R}$ such that $\hat{\theta}(\delta)$ is unbiased for θ.
(c) Show that the risk of $\hat{\theta}(\delta)$ under the squared error loss is a function of (δ, θ).
(d) Show that there is a constant δ^* such that for any fixed θ, the risk of $\hat{\theta}(\delta)$ is strictly decreasing in δ for $\delta < \delta^*$ and strictly increasing for $\delta > \delta^*$.
(e) Show that the unbiased estimator of θ derived in (b) is inadmissible.

82. Let $T_0(X)$ be an unbiased estimator of ϑ in an estimation problem. Show that any unbiased estimator of ϑ is of the form $T(X) = T_0(X) - U(X)$, where $U(X)$ is an "unbiased estimator" of 0.

83. Let X be a discrete random variable with

$$P(X = -1) = p, \quad P(X = k) = (1-p)^2 p^k, \quad k = 0, 1, 2, ...,$$

where $p \in (0, 1)$ is unknown.
(a) Show that $U(X)$ is an unbiased estimator of 0 if and only if $U(k) =$

ak for all $k = -1, 0, 1, 2, ...$ and some a.
(b) Show that $T_0(X) = I_{\{0\}}(X)$ is unbiased for $\vartheta = (1-p)^2$ and that, under the squared error loss, T_0 is a $\mathfrak{S}$-optimal rule, where $\mathfrak{S}$ is the class of all unbiased estimators of ϑ.
(c) Show that $T_0(X) = I_{\{-1\}}(X)$ is unbiased for $\vartheta = p$ and that, under the squared error loss, there is no $\mathfrak{S}$-optimal rule, where $\mathfrak{S}$ is the class of all unbiased estimators of ϑ.

84. (Nonexistence of an unbiased estimator). Let X be a random variable having the binomial distribution $Bi(p, n)$ with an unknown $p \in (0, 1)$ and a known n. Consider the problem of estimating $\vartheta = p^{-1}$. Show that there is no unbiased estimator of ϑ.

85. Let $X_1, ..., X_n$ be i.i.d. random variables having the normal distribution $N(\theta, 1)$, where $\theta = 0$ or 1. Consider the estimation of θ.
(a) Let $\mathfrak{S}$ be the class of nonrandomized rules (estimators), i.e., estimators that take values 0 and 1 only. Show that there does not exist any unbiased estimator of θ in $\mathfrak{S}$.
(b) Find an estimator in $\mathfrak{S}$ that is approximately unbiased.

86. Let $X_1, ..., X_n$ be i.i.d. from the Poisson distribution $P(\theta)$ with an unknown $\theta > 0$. Find the bias and mse of $T_n = (1 - a/n)^{n\bar{X}}$ as an estimator of $\vartheta = e^{-a\theta}$, where $a \neq 0$ is a known constant.

87. Let $X_1, ..., X_n$ be i.i.d. ($n \geq 3$) from $N(\mu, \sigma^2)$, where $\mu > 0$ and $\sigma > 0$ are unknown parameters. Let $T_1 = \bar{X}/S$ be an estimator of μ/σ and $T_2 = \bar{X}^2$ be an estimator of μ^2, where $\bar{X}$ and S^2 are the sample mean and variance, respectively. Calculate the mse's of T_1 and T_2.

88. Consider a location family $\{P_\mu : \mu \in \mathcal{R}^k\}$ on $\mathcal{R}^k$, where $P_\mu = P_{(\mu, I_k)}$ is given in (2.10). Let $l_0 \in \mathcal{R}^k$ be a fixed vector and $L(P, a) = L(\|\mu - a\|)$, where $a \in \mathbb{A} = \mathcal{R}^k$ and $L(\cdot)$ is a nonnegative Borel function on $[0, \infty)$. Show that the family is invariant and the decision problem is invariant under the transformation $g(X) = X + cl_0$, $c \in \mathcal{R}$. Find an invariant decision rule.

89. Let $X_1, ..., X_n$ be i.i.d. from the $N(\mu, \sigma^2)$ distribution with unknown $\mu \in \mathcal{R}$ and $\sigma^2 > 0$. Consider the scale transformation aX, $a \in (0, \infty)$.
(a) For estimating σ^2 under the loss function $L(P, a) = (1 - a/\sigma^2)^2$, show that the problem is invariant and that the sample variance S^2 is invariant.
(b) For testing $H_0 : \mu \leq 0$ versus $H_1 : \mu > 0$ under the loss

$$L(P, 0) = \frac{\mu}{\sigma} I_{(0,\infty)}(\mu) \quad \text{and} \quad L(P, 1) = \frac{|\mu|}{\sigma} I_{(-\infty,0]}(\mu),$$

show that the problem is invariant and any test that is a function of $\bar{X}/\sqrt{S^2/n}$ is invariant.

90. Let $X_1, ..., X_n$ be i.i.d. random variables having the c.d.f. $F(x-\theta)$, where F is symmetric about 0 and $\theta \in \mathcal{R}$ is unknown.
(a) Show that the c.d.f. of $\sum_{i=1}^n w_i X_{(i)} - \theta$ is symmetric about 0, where $X_{(i)}$ is the ith order statistic and w_i's are constants satisfying $w_i = w_{n-i+1}$ and $\sum_{i=1}^n w_i = 1$.
(b) Show that $\sum_{i=1}^n w_i X_{(i)}$ in (a) is unbiased for θ if the mean of F exists.
(c) Show that $\sum_{i=1}^n w_i X_{(i)}$ is location invariant when $\sum_{i=1}^n w_i = 1$.

91. In Example 2.25, show that the conditional distribution of $\boldsymbol{\theta}$ given $X = x$ is $N(\mu_*(x), c^2)$ with $\mu_*(x)$ and c^2 given by (2.25).

92. A *median* of a random variable Y (or its distribution) is any value m such that $P(Y \le m) \ge \frac{1}{2}$ and $P(Y \ge m) \ge \frac{1}{2}$.
(a) Show that the set of medians is a closed interval $[m_0, m_1]$.
(b) Suppose that $E|Y| < \infty$. If c is not a median of Y, show that $E|Y-c| \ge E|Y-m|$ for any median m of Y.
(c) Let X be a sample from P_θ, where $\theta \in \Theta \subset \mathcal{R}$. Consider the estimation of θ under the absolute error loss function $|a-\theta|$. Let Π be a given distribution on Θ with finite mean. Find the $\Im$-Bayes rule w.r.t. Π, where $\Im$ is the class of all rules.

93. (Classification). Let X be a sample having a p.d.f. $f_j(x)$ w.r.t. a σ-finite measure ν, where j is unknown and $j \in \{1, ..., J\}$ with a known integer $J \ge 2$. Consider a decision problem in which the action space $\mathbb{A} = \{1, ..., J\}$ and the loss function is

$$L(j,a) = \begin{cases} 0 & \text{if } a = j \\ 1 & \text{if } a \ne j. \end{cases}$$

(a) Let $\Im$ be the class of all nonrandomized decision rules. Obtain the risk of a $\delta \in \Im$.
(b) Let Π be a probability measure on $\{1, ..., J\}$ with $\Pi(\{j\}) = \pi_j$, $j = 1, ..., J$. Obtain the Bayes risk of $\delta \in \Im$ w.r.t. Π.
(c) Obtain a $\Im$-Bayes rule w.r.t. Π in (b).
(d) Assume that $J = 2$, $\pi_1 = \pi_2 = 0.5$, and $f_j(x) = \phi(x - \mu_j)$, where $\phi(x)$ is the p.d.f. of the standard normal distribution and μ_j, $j = 1, 2$, are known constants. Obtain the Bayes rule in (c) and compute the Bayes risk.
(e) Obtain the risk and the Bayes risk (w.r.t. Π in (b)) of a randomized decision rule.
(f) Obtain a Bayes rule w.r.t. Π.
(g) Obtain a minimax rule.

94. Let $\hat{\theta}$ be an unbiased estimator of an unknown $\theta \in \mathcal{R}$.
(a) Under the squared error loss, show that the estimator $\hat{\theta} + c$ is not

minimax unless $\sup_\theta R_T(\theta) = \infty$ for any estimator T, where $c \neq 0$ is a known constant.
(b) Under the squared error loss, show that the estimator $c\hat{\theta}$ is not minimax unless $\sup_\theta R_T(\theta) = \infty$ for any estimator T, where $c \in (0, 1)$ is a known constant.
(c) Consider the loss function $L(\theta, a) = (a-\theta)^2/\theta^2$ (assuming $\theta \neq 0$). Show that $\hat{\theta}$ is not minimax unless $\sup_\theta R_T(\theta) = \infty$ for any T.

95. Let X be a binary observation with $P(X = 1) = \theta_1$ or θ_2, where $0 < \theta_1 < \theta_2 < 1$ are known values. Consider the estimation of θ with action space $\{a_1, a_2\}$ and loss function $L(\theta_i, a_j) = l_{ij}$, where $l_{21} \geq l_{12} > l_{11} = l_{22} = 0$. For a decision rule $\delta(X)$, the vector $(R_\delta(\theta_1), R_\delta(\theta_2))$ is defined to be its risk point.
(a) Show that the set of risk points of all decision rules is the convex hull of the set of risk points of all nonrandomized rules.
(b) Find a minimax rule.
(c) Let Π be a distribution on $\{\theta_1, \theta_2\}$. Obtain the class of all Bayes rules w.r.t. Π. Discuss when there is a unique Bayes rule.

96. Consider the decision problem in Example 2.23.
(a) Let Π be the uniform distribution on $(0, 1)$. Show that a $\Im$-Bayes rule w.r.t. Π is $T_{j^*}(X)$, where j^* is the largest integer in $\{0, 1, ..., n-1\}$ such that $B_{j+1,n-j+1}(\theta_0) \geq \frac{1}{2}$ and $B_{a,b}(\cdot)$ denotes the c.d.f. of the beta distribution $B(a, b)$.
(b) Derive a $\Im$-minimax rule.

97. Let $X_1, ..., X_n$ be i.i.d. from the $N(\mu, \sigma^2)$ distribution with unknown $\mu \in \mathcal{R}$ and $\sigma^2 > 0$. To test the hypotheses

$$H_0 : \mu \leq \mu_0 \qquad \text{versus} \qquad H_1 : \mu > \mu_0,$$

where μ_0 is a fixed constant, consider a test of the form $T_c(X) = I_{(c,\infty)}(T_{\mu_0})$, where $T_{\mu_0} = (\bar{X} - \mu_0)/\sqrt{S^2/n}$ and c is a fixed constant.
(a) Find the size of T_c. (Hint: T_{μ_0} has the t-distribution t_{n-1}.)
(b) If α is a given level of significance, find a c_α such that T_{c_α} has size α.
(c) Compute the p-value for T_{c_α} derived in (b).
(d) Find a c_α such that $[\bar{X} - c_\alpha\sqrt{S^2/n}, \bar{X} + c_\alpha\sqrt{S^2/n}]$ is a confidence interval for μ with confidence coefficient $1 - \alpha$. What is the expected interval length?

98. In Exercise 67, calculate the size of $T_c(X)$; find a c_α such that T_{c_α} has size α, a given level of significance; and find the p-value for T_{c_α}.

99. In Exercise 68, assume that σ is known. Calculate the size of $T_c(X)$; find a c_α such that T_{c_α} has size α, a given level of significance; and find the p-value for T_{c_α}.

100. Let $\alpha \in (0,1)$ be given and $T_{j,q}(X)$ be the test given in Example 2.30. Show that there exist integer j and $q \in (0,1)$ such that the size of $T_{j,q}$ is α.

101. Let $X_1, ..., X_n$ be i.i.d. from the exponential distribution $E(a, \theta)$ with unknown $a \in \mathcal{R}$ and $\theta > 0$. Let $\alpha \in (0,1)$ be given.
(a) Using $T_1(X) = \sum_{i=1}^n (X_i - X_{(1)})$, construct a confidence interval for θ with confidence coefficient $1-\alpha$ and find the expected interval length.
(b) Using $T_1(X)$ and $T_2(X) = X_{(1)}$, construct a confidence interval for a with confidence coefficient $1-\alpha$ and find the expected interval length.
(c) Using the method in Example 2.32, construct a confidence set for the two-dimensional parameter (a, θ) with confidence coefficient $1-\alpha$.

102. Suppose that X is a sample and a statistic $T(X)$ has a distribution in a location family $\{P_\mu : \mu \in \mathcal{R}\}$. Using $T(X)$, derive a confidence interval for μ with level of significance $1-\alpha$ and obtain the expected interval length. Show that if the c.d.f. of $T(X)$ is continuous, then we can always find a confidence interval for μ with confidence coefficient $1-\alpha$ for any $\alpha \in (0,1)$.

103. Let $X = (X_1, ..., X_n)$ be a sample from P_θ, where $\theta \in \{\theta_1, ..., \theta_k\}$ with a fixed integer k. Let $T_n(X)$ be an estimator of θ with range $\{\theta_1, ..., \theta_k\}$.
(a) Show that $T_n(X)$ is consistent if and only if $P_\theta(T_n(X) = \theta) \to 1$.
(b) Show that if $T_n(X)$ is consistent, then it is a_n-consistent for any $\{a_n\}$.

104. Let $X_1, ..., X_n$ be i.i.d. from the uniform distribution on $(\theta - \frac{1}{2}, \theta + \frac{1}{2})$, where $\theta \in \mathcal{R}$ is unknown. Show that $(X_{(1)} + X_{(n)})/2$ is strongly consistent for θ and also consistent in mse.

105. Let $X_1, ..., X_n$ be i.i.d. from a population with the Lebesgue p.d.f. $f_\theta(x) = 2^{-1}(1 + \theta x) I_{(-1,1)}(x)$, where $\theta \in (-1,1)$ is an unknown parameter. Find a consistent estimator of θ. Is your estimator $\sqrt{n}$-consistent?

106. Let $X_1, ..., X_n$ be i.i.d. observations. Suppose that T_n is an unbiased estimator of ϑ based on $X_1, ..., X_n$ such that for any n, $\mathrm{Var}(T_n) < \infty$ and $\mathrm{Var}(T_n) \le \mathrm{Var}(U_n)$ for any other unbiased estimator U_n of ϑ based on $X_1, ..., X_n$. Show that T_n is consistent in mse.

107. Consider the Bayes rule $\mu_*(X)$ in Example 2.25. Show that $\mu_*(X)$ is a strongly consistent, $\sqrt{n}$-consistent, and L_2-consistent estimator of μ. What is the order of the bias of $\mu_*(X)$ as an estimator of μ?

108. In Exercise 21, show that
(a) $\bar{Y}/\bar{X}$ is an inconsistent estimator of β;
(b) $\hat{\beta} = Z_{(m)}$ is a consistent estimator of β, where $m = n/2$ when n is even, $m = (n+1)/2$ when n is odd, and $Z_{(i)}$ is the ith smallest value of Y_i/X_i, $i = 1, ..., n$.

109. Show that the estimator T_0 of θ in Exercise 64 is inconsistent.

110. Let g_1, g_2,... be continuous functions on $(a, b) \subset \mathcal{R}$ such that $g_n(x) \to g(x)$ uniformly for x in any closed subinterval of (a, b). Let T_n be a consistent estimator of $\theta \in (a, b)$. Show that $g_n(T_n)$ is consistent for $\vartheta = g(\theta)$.

111. Let $X_1, ..., X_n$ be i.i.d. from P with unknown mean $\mu \in \mathcal{R}$ and variance $\sigma^2 > 0$, and let $g(\mu) = 0$ if $\mu \neq 0$ and $g(0) = 1$. Find a consistent estimator of $\vartheta = g(\mu)$.

112. Establish results for the smallest order statistic $X_{(1)}$ (based on i.i.d. random variables $X_1, ..., X_n$) similar to those in Example 2.34.

113. (Consistency for finite population). In Example 2.27, show that $\hat{Y} \to_p Y$ as $n \to N$ for any fixed N and population. Is $\hat{Y}$ still consistent if sampling is with replacement?

114. Assume that $X_i = \theta t_i + e_i$, $i = 1, ..., n$, where $\theta \in \Theta$ is an unknown parameter, Θ is a closed subset of $\mathcal{R}$, e_i's are i.i.d. on the interval $[-\tau, \tau]$ with some unknown $\tau > 0$ and $Ee_i = 0$, and t_i's are fixed constants. Let

$$T_n = S_n(\tilde{\theta}_n) = \min_{\gamma \in \Theta} S_n(\gamma),$$

where

$$S_n(\gamma) = 2 \max_{i \le n} |X_i - \gamma t_i| / \sqrt{1 + \gamma^2}.$$

(a) Assume that $\sup_i |t_i| < \infty$ and $\sup_i t_i - \inf_i t_i > 2\tau$. Show that the sequence $\{\tilde{\theta}_n, n = 1, 2, ...\}$ is bounded a.s.
(b) Let $\theta_n \in \Theta$, $n = 1, 2, ...$ If $\theta_n \to \theta$, show that

$$S_n(\theta_n) - S_n(\theta) = O(|\theta_n - \theta|) \quad \text{a.s.}$$

(c) Under the conditions in (a), show that T_n is a strongly consistent estimator of $\vartheta = \min_{\gamma \in \Theta} S(\gamma)$, where $S(\gamma) = \lim_{n \to \infty} S_n(\gamma)$ a.s.

115. Let $X_1, ..., X_n$ be i.i.d. random variables with $EX_1^2 < \infty$ and $\bar{X}$ be the sample mean. Consider the estimation of $\mu = EX_1$.
(a) Let $T_n = \bar{X} + \xi_n/\sqrt{n}$, where ξ_n is a random variable satisfying $\xi_n = 0$ with probability $1 - n^{-1}$ and $\xi_n = n^{3/2}$ with probability n^{-1}.

Show that $b_{T_n}(P) \neq \tilde{b}_{T_n}(P)$ for any P.
(b) Let $T_n = \bar{X} + \eta_n/\sqrt{n}$, where η_n is a random variable that is independent of $X_1, ..., X_n$ and equals 0 with probability $1 - 2n^{-1}$ and $\pm\sqrt{n}$ with probability n^{-1}. Show that $\text{amse}_{T_n}(P) = \text{amse}_{\bar{X}}(P) = \text{mse}_{\bar{X}}(P)$ and $\text{mse}_{T_n}(P) > \text{amse}_{T_n}(P)$ for any P.

116. Let $X_1, ..., X_n$ be i.i.d. random variables with finite $\theta = EX_1$ and $\text{Var}(X_1) = \theta$, where $\theta > 0$ is unknown. Consider the estimation of $\vartheta = \sqrt{\theta}$. Let $T_{1n} = \sqrt{\bar{X}}$ and $T_{2n} = \bar{X}/S$, where $\bar{X}$ and S^2 are the sample mean and sample variance.
(a) Obtain the n^{-1} order asymptotic biases of T_{1n} and T_{2n} according to (2.38).
(b) Obtain the asymptotic relative efficiency of T_{1n} w.r.t. T_{2n}.

117. Let $X_1, ..., X_n$ be i.i.d. according to $N(\mu, 1)$ with an unknown $\mu \in \mathcal{R}$. Let $\vartheta = P(X_1 \leq c)$ for a fixed constant c. Consider the following estimators of ϑ: $T_{1n} = F_n(c)$, where F_n is the empirical c.d.f. defined in (2.28), and $T_{2n} = \Phi(c - \bar{X})$, where Φ is the c.d.f. of $N(0, 1)$.
(a) Find the n^{-1} order asymptotic bias of T_{2n} according to (2.38).
(b) Find the asymptotic relative efficiency of T_{1n} w.r.t. T_{2n}.

118. Let $X_1, ..., X_n$ be i.i.d. from the $N(0, \sigma^2)$ distribution with an unknown $\sigma > 0$. Consider the estimation of $\vartheta = \sigma$. Find the asymptotic relative efficiency of $\sqrt{\pi/2}\sum_{i=1}^n |X_i|/n$ w.r.t. $(\sum_{i=1}^n X_i^2/n)^{1/2}$.

119. Let $X_1, ..., X_n$ be i.i.d. from P with $EX_1^4 < \infty$ and unknown mean $\mu \in \mathcal{R}$ and variance $\sigma^2 > 0$. Consider the estimation of $\vartheta = \mu^2$ and the following three estimators: $T_{1n} = \bar{X}^2$, $T_{2n} = \bar{X}^2 - S^2/n$, $T_{3n} = \max\{0, T_{2n}\}$, where $\bar{X}$ and S^2 are the sample mean and variance. Show that the amse's of T_{jn}, $j = 1, 2, 3$, are the same when $\mu \neq 0$ but may be different when $\mu = 0$. Which estimator is the best in terms of the asymptotic relative efficiency when $\mu = 0$?

120. Prove Theorem 2.6.

121. Let $X_1, ..., X_n$ be i.i.d. with $EX_i = \mu$, $\text{Var}(X_i) = 1$, and $EX_i^4 < \infty$. Let $T_{1n} = n^{-1}\sum_{i=1}^n X_i^2 - 1$ and $T_{2n} = \bar{X}^2 - n^{-1}$ be estimators of $\vartheta = \mu^2$.
(a) Find the asymptotic relative efficiency of T_{1n} w.r.t. T_{2n}.
(b) Show that $e_{T_{1n},T_{2n}}(P) \leq 1$ if the c.d.f. of $X_i - \mu$ is symmetric about 0 and $\mu \neq 0$.
(c) Find a distribution P for which $e_{T_{1n},T_{2n}}(P) > 1$.

122. Let $X_1, ..., X_n$ be i.i.d. binary random variables with unknown $p = P(X_i = 1) \in (0, 1)$. Consider the estimation of p. Let a and b be two positive constants. Find the asymptotic relative efficiency of the estimator $(a + n\bar{X})/(a + b + n)$ w.r.t. $\bar{X}$.

123. Let $X_1, ..., X_n$ be i.i.d. from $N(\mu, \sigma^2)$ with an unknown $\mu \in \mathcal{R}$ and a known σ^2. Let $T_1 = \bar{X}$ be the sample mean and $T_2 = \mu_*(X)$ be the Bayes estimator given in (2.25). Assume that $EX_1^4 < \infty$.
(a) Calculate the exact mse of both estimators. Can you conclude that one estimator is better than the other in terms of the mse?
(b) Find the asymptotic relative efficiency of T_1 w.r.t. T_2.

124. In Example 2.37, show that
(a) the limiting size of T_{c_α} is 1 if $\mathcal{P}$ contains all possible populations on $\mathcal{R}$ with finite second moments;
(b) $T_n = T_{c_\alpha}$ with $\alpha = \alpha_n$ (given by (2.40)) is Chernoff-consistent;
(c) T_n in (b) is not strongly Chernoff-consistent if $\mathcal{P}$ contains all possible populations on $\mathcal{R}$ with finite second moments.

125. Let $X_1, ..., X_n$ be i.i.d. with unknown mean $\mu \in \mathcal{R}$ and variance $\sigma^2 > 0$. For testing $H_0 : \mu \le \mu_0$ versus $H_1 : \mu > \mu_0$, consider the test T_{c_α} obtained in Exercise 97(b).
(a) Show that T_{c_α} has asymptotic significance level α and is consistent.
(b) Find a test that is Chernoff-consistent.

126. Consider the test T_j in Example 2.23. For each n, find a $j = j_n$ such that T_{j_n} has asymptotic significance level $\alpha \in (0, 1)$.

127. Show that the test T_{c_α} in Exercise 98 is consistent, but T_{c_α} in Exercise 99 is not consistent.

128. In Example 2.31, suppose that we drop the normality assumption but assume that $\mu = EX_i$ and $\sigma^2 = \mathrm{Var}(X_i)$ are finite.
(a) Show that when σ^2 is known, the asymptotic significance level of the confidence interval $[\bar{X} - c_\alpha, \bar{X} + c_\alpha]$ is $1 - \alpha$, where $c_\alpha = \sigma z_{1-\alpha/2}/\sqrt{n}$ and $z_a = \Phi^{-1}(a)$.
(b) Show that when σ^2 is known, the limiting confidence coefficient of the interval in (a) might be 0 if $\mathcal{P}$ contains all possible populations on $\mathcal{R}$.
(c) Show that the confidence interval in Exercise 97(d) has asymptotic significance level $1 - \alpha$.

129. Let $X_1, ..., X_n$ be i.i.d. with unknown mean $\mu \in \mathcal{R}$ and variance $\sigma^2 > 0$. Assume that $EX_1^4 < \infty$. Using the sample variance S^2, construct a confidence interval for σ^2 that has asymptotic significance level $1 - \alpha$.

130. Consider the sample correlation coefficient T defined in Exercise 22. Construct a confidence interval for ρ that has asymptotic significance level $1 - \alpha$, assuming that (Y_i, Z_i) is normally distributed. (Hint: show that the asymptotic variance of T is $(1 - \rho^2)^2$.)

http://www.springer.com/978-0-387-95382-3

Mathematical Statistics
Shao, J.
2003, XVI, 591 p., Hardcover
ISBN: 978-0-387-95382-3