

# Advanced Theory of Statistics

Lectured by Wang Miao

L<sup>A</sup>T<sub>E</sub>Xed by Chengxin Gong

2022 年 12 月 8 日

## 目录

<b>1</b>	<b>Probability Theory</b>	<b>3</b>
1.1	Measure space, measurable function, and integration . . . . .	3
1.2	Integration theory and Radon-Nikodym derivative . . . . .	4
1.3	Densities, moments, inequalities, and generating functions . . . . .	5
1.4	Conditional expectation and independence . . . . .	6
1.5	Convergence modes and relationships . . . . .	7
1.6	Uniform integrability and weak convergence . . . . .	8
1.7	Convergence of transformations and law of large numbers . . . . .	9
1.8	The central limit theorem . . . . .	10
<b>2</b>	<b>Fundamentals of Statistics</b>	<b>10</b>
2.1	Models, data, statistics, and sampling distributions . . . . .	10
2.2	Sufficiency and minimal sufficiency . . . . .	11
2.3	Completeness . . . . .	12
2.4	Statistical decision . . . . .	12
2.5	Statistical inference . . . . .	14
<b>3</b>	<b>Unbiased Estimation</b>	<b>16</b>
3.1	UMVUE: functions of sufficient and complete statistics . . . . .	16
3.2	Characteristic of UMVUE and Fisher information bound . . . . .	18
3.3	U- and V-statistics . . . . .	19
3.4	Construction of unbiased or approximately unbiased estimators and method of moments . . . . .	20
<b>4</b>	<b>Estimation in Parametric Models</b>	<b>21</b>
4.1	Bayesian approach . . . . .	21
4.2	Bayes rule and computation . . . . .	22
4.3	Minimaxity and admissibility . . . . .	23
4.4	Simultaneous estimation and shrinkage estimators . . . . .	24
4.5	Likelihood and maximum likelihood estimator (MLE) . . . . .	26

## 目录

4.6	Asymptotically efficient estimation . . . . .	26
4.7	MLE in generalized linear models (GLM) and quasi-MLE . . . . .	27
4.8	Other asymptotically efficient estimators and pseudo MLE . . . . .	29
<b>5</b>	<b>Estimation in Non-Parametric Models</b>	<b>29</b>
5.1	Empirical c.d.f. and empirical likelihoods . . . . .	29
5.2	Profile likelihoods, GEE, and GMM . . . . .	30
<b>6</b>	<b>Hypothesis Tests</b>	<b>32</b>
6.1	Neyman-Pearson lemma and monotone likelihood ratio . . . . .	32
6.2	UMP tests and unbiased tests . . . . .	33
6.3	Likelihood ratio and asymptotic tests . . . . .	35
6.4	Asymptotic chi-square tests . . . . .	36
<b>7</b>	<b>Confidence Sets</b>	<b>37</b>
7.1	Pivotal quantities and confidence sets . . . . .	37
7.2	Inverting acceptance regions of tests, UMA and UMAU confidence sets . . . . .	39
7.3	Lengths and expected lengths of confidence intervals . . . . .	40
7.4	Asymptotic confidence sets . . . . .	41
7.5	Variance estimation, replication, jackknife, and bootstrap . . . . .	42
7.6	Bootstrap confidence intervals . . . . .	43

# 1 Probability Theory

## 1.1 Measure space, measurable function, and integration

**Definition 1:** A collection of subsets of  $\Omega, \mathcal{F}$ , is a  $\sigma$ -field (or  $\sigma$ -algebra) if (i) The empty set  $\emptyset \in \mathcal{F}$ ; (ii) If  $A \in \mathcal{F}$ , then the complement  $A^c \in \mathcal{F}$ ; (iii) If  $A_i \in \mathcal{F}, i = 1, 2, \dots$ , then their union  $\cup A_i \in \mathcal{F}$ .  $(\Omega, \mathcal{F})$  is a measurable space if  $\mathcal{F}$  is a  $\sigma$ -field on  $\Omega$ .

**Example 1:**  $\mathcal{C}$  = a collection of subsets of interest.  $\sigma(\mathcal{C})$  = the smallest  $\sigma$ -field containing  $\mathcal{C}$  (the  $\sigma$ -field generated by  $\mathcal{C}$ ).  $\sigma(\mathcal{C}) = \mathcal{C}$  if  $\mathcal{C}$  itself is a  $\sigma$ -field.  $\sigma(\{A\}) = \{\emptyset, A, A^c, \Omega\}$ .

**Example 2 (Borel  $\sigma$ -field):**  $\mathbb{R}^k$ : the  $k$ -dimensional Euclidean space ( $\mathbb{R}^1 = \mathbb{R}$  is the real line).  $\mathcal{O}$  = all open sets,  $\mathcal{C}$  = all closed sets.  $\mathcal{B}^k = \sigma(\mathcal{O}) = \sigma(\mathcal{C})$ : the Borel  $\sigma$ -field on  $\mathbb{R}^k$ .  $C \in \mathcal{B}^k, \mathcal{B}_C = \{C \cap B : B \in \mathcal{B}^k\}$  is the Borel  $\sigma$ -field on  $C$ .

**Definition 2:** Let  $(\Omega, \mathcal{F})$  be a measurable space. A set function  $\nu$  defined on  $\mathcal{F}$  is a measure if (i)  $0 \leq \nu(A) \leq \infty$  for any  $A \in \mathcal{F}$ ; (ii)  $\nu(\emptyset) = 0$ ; (iii) If  $A_i \in \mathcal{F}, i = 1, 2, \dots$ , and  $A_i$ 's are disjoint, i.e.  $A_i \cap A_j = \emptyset$  for any  $i \neq j$ , then  $\nu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \nu(A_i)$ .  $(\Omega, \mathcal{F}, \nu)$  is a measure if  $\nu$  is a measure on  $\mathcal{F}$  in  $(\Omega, \mathcal{F})$ .

**Convention 1:** For any  $x \in \mathbb{R}$ ,  $\infty + x = \infty$ ,  $x\infty = \infty$  if  $x > 0$ ,  $x\infty = -\infty$  if  $x < 0$ .  $0\infty = 0$ ,  $\infty + \infty = \infty$ ,  $\infty^a = \infty$  for any  $a > 0$ .  $\infty - \infty$  or  $\infty/\infty$  is not defined.

**Example 3 (Important examples of measures):** (a) Let  $x \in \Omega$  be a fixed point and  $\delta_x(A) = \begin{cases} c & x \in A \\ 0 & x \notin A \end{cases}$ . This is called a point mass at  $x$ . (b) Let  $\mathcal{F} =$  all subsets of  $\Omega$  and  $\nu(A) =$  the number of elements in  $A \in \mathcal{F}$  ( $\nu(A) = \infty$  if  $A$  contains infinitely many elements). Then  $\nu$  is a measure on  $\mathcal{F}$  and is called the counting measure. (c) There is a unique measure  $m$  on  $(\mathbb{R}, \mathcal{B})$ , that satisfies  $m([a, b]) = b - a$  for every finite interval  $[a, b]$ ,  $-\infty < a \leq b < \infty$ . This is called the Lebesgue measure.

**Proposition 1 (Properties of measures):** Let  $(\Omega, \mathcal{F}, \nu)$  be a measure space. (1) Monotonicity: If  $A \subset B$ , then  $\nu(A) \leq \nu(B)$ . (2) Subadditivity: For any sequence  $A_1, A_2, \dots$ ,  $\nu(\cup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \nu(A_i)$ . (3) Continuity: If  $A_1 \subset A_2 \subset A_3 \subset \dots$  (or  $A_1 \supset A_2 \supset A_3 \supset \dots$  and  $\nu(A_1) < \infty$ ), then  $\nu(\lim_{n \rightarrow \infty} A_n) = \lim_{n \rightarrow \infty} \nu(A_n)$  where  $\lim_{n \rightarrow \infty} A_n = \cup_{i=1}^{\infty} A_i$  (or  $= \cap_{i=1}^{\infty} A_i$ ).

**Definition 3:** Let  $P$  be a probability measure on  $(\mathbb{R}, \mathcal{B})$ . The cumulative distribution function (c.d.f.) of  $P$  is defined to be  $F(x) = P((-\infty, x])$ ,  $x \in \mathbb{R}$ .

**Proposition 2 (Properties of c.d.f.'s):** (i) Let  $F$  be a c.d.f. on  $\mathbb{R}$ . (a)  $F(-\infty) = \lim_{x \rightarrow -\infty} F(x) = 0$ ; (b)  $F(\infty) = \lim_{x \rightarrow \infty} F(x) = 1$ ; (c)  $F$  is nondecreasing, i.e.  $F(x) \leq F(y)$  if  $x \leq y$ ; (d)  $F$  is right continuous, i.e.  $\lim_{y \rightarrow x+0} F(y) = F(x)$ . (ii) Suppose a real-valued function  $F$  on  $\mathbb{R}$  satisfies (a)-(d) in part (i). Then  $F$  is the c.d.f. of a unique probability measure on  $(\mathbb{R}, \mathcal{B})$ .

**Definition 4 (Product space):**  $\mathcal{I} = \{1, \dots, k\}$ ,  $k$  is finite or  $\infty$ .  $\Gamma_i, i \in \mathcal{I}$ , are some sets.  $\prod_{i \in \mathcal{I}} \Gamma_i = \Gamma_1 \times \dots \times \Gamma_k = \{(a_1, \dots, a_k) : a_i \in \Gamma_i, i \in \mathcal{I}\}$ . Let  $(\Omega_i, \mathcal{F}_i), i \in \mathcal{I}$  be measurable spaces.  $\sigma(\prod_{i \in \mathcal{I}} \mathcal{F}_i)$  is called the product  $\sigma$ -field on the product space  $\prod_{i \in \mathcal{I}} \Omega_i$ .  $(\prod_{i \in \mathcal{I}} \Omega_i, \sigma(\prod_{i \in \mathcal{I}} \mathcal{F}_i))$  is denoted by  $\prod_{i \in \mathcal{I}} (\Omega_i, \mathcal{F}_i)$ .

**Definition 5 ( $\sigma$ -finite):** A measure  $\nu$  on  $(\Omega, \mathcal{F})$  is said to be  $\sigma$ -finite iff there exists a sequence  $\{A_1, A_2, \dots\}$  such that  $\cup A_i = \Omega$  and  $\nu(A_i) < \infty$  for all  $i$ . Any finite measure is clearly  $\sigma$ -finite. The Lebesgue measure on  $\mathcal{F}$  is  $\sigma$ -finite.

**Proposition 3 (Product measure theorem):** Let  $(\Omega_i, \mathcal{F}_i, \nu_i), i = 1, \dots, k$ , be measure spaces with  $\sigma$ -finite measures. There exists a unique  $\sigma$ -finite measure on  $\sigma$ -field  $\sigma(\mathcal{F}_1 \times \dots \times \mathcal{F}_k)$ , called the product measure and denoted by  $\nu_1 \times \dots \times \nu_k$ , such that  $\nu_1 \times \dots \times \nu_k(A_1 \times \dots \times A_k) = \nu_1(A_1) \dots \nu_k(A_k)$  for all  $A_i \in \mathcal{F}_i, i = 1, \dots, k$ .

**Definition 6 (Measurable function):** Let  $(\Omega, \mathcal{F})$  and  $(\Lambda, \mathcal{G})$  be measurable spaces. Let  $f$  be a function from  $\Omega$  to  $\Lambda$ .  $f$  is called a measurable function from  $(\Omega, \mathcal{F})$  to  $(\Lambda, \mathcal{G})$  iff  $f^{-1}(\mathcal{G}) \subset \mathcal{F}$ .

**Definition 7 (Integration):** (a) The integral of a nonnegative simple function  $\phi$  w.r.t.  $\nu$  is defined as  $\int \phi d\nu = \sum_{i=1}^k a_i \nu(A_i)$ . (b) Let  $f$  be a nonnegative Borel function and let  $\mathcal{S}_f$  be the collection of all nonnegative simple functions satisfying  $\phi(\omega) \leq f(\omega)$  for any  $\omega \in \Omega$ . The integral of  $f$  w.r.t.  $\nu$  is defined as  $\int f d\nu = \sup\{\int \phi d\nu : \phi \in \mathcal{S}_f\}$  (Hence, for any Borel function  $f \geq 0$ , there exists a sequence of simple functions  $\phi_1, \phi_2, \dots$  such that  $0 \leq \phi_i \leq f$  for all  $i$  and  $\lim_{n \rightarrow \infty} \int \phi_n d\nu = \int f d\nu$ ). (c) Let  $f$  be a Borel function,  $f_+(\omega) = \max\{f(\omega), 0\}$  be the positive part of  $f$ , and  $f_-(\omega) = \max\{-f(\omega), 0\}$  be the negative part of  $f$ . We say that  $\int f d\nu$  exists if and only if at least one of  $\int f_+ d\nu$  and  $\int f_- d\nu$  is finite, in which case  $\int f d\nu = \int f_+ d\nu - \int f_- d\nu$ . (d) When both  $\int f_+ d\nu$  and  $\int f_- d\nu$  are finite, we say that  $f$  is integrable. Let  $A$  be a measurable set and  $I_A$  be its indicator function. The integral of  $f$  over  $A$  is defined as  $\int_A f d\nu = \int I_A f d\nu$ .

**Example 4 (Extended set):** For convenience, we define the integral of a measurable  $f$  from  $(\Omega, \mathcal{F}, \nu)$  to  $(\bar{\mathbb{R}}, \bar{\mathcal{B}})$ , where  $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}, \bar{\mathcal{B}} = \sigma(\mathcal{B} \cup \{\infty, -\infty\})$ . Let  $A_+ = \{f = \infty\}$  and  $A_- = \{f = -\infty\}$ . If  $\nu(A_+) = 0$ , we define  $\int f_+ d\nu$  to be  $\int I_{A_+} f_+ d\nu$ ; otherwise  $\int f_+ d\nu = \infty$ .  $\int f_- d\nu$  is similarly defined. If at least one of  $\int f_+ d\nu$  and  $\int f_- d\nu$  is finite, then  $\int f d\nu = \int f_+ d\nu - \int f_- d\nu$  is well defined.

## 1.2 Integration theory and Radon-Nikodym derivative

**Proposition 1:**  $(\Omega, \mathcal{F}, \nu)$  be a measure space and  $f$  and  $g$  be Borel functions. (i) If  $f \leq g$  a.e., then  $\int f d\nu \leq \int g d\nu$ , provided that the integrals exist. (ii) If  $f \geq 0$  a.e. and  $\int f d\nu = 0$ , then  $f = 0$  a.e.

**Theorem 1:** Let  $f_1, f_2, \dots$  be a sequence of Borel functions on  $(\Omega, \mathcal{F}, \nu)$ . (i) Fatou's lemma: If  $f_n \geq 0$ , then  $\int \liminf_n f_n d\nu \leq \liminf_n \int f_n d\nu$ . (ii) Dominated convergence theorem: If  $\lim_{n \rightarrow \infty} f_n = f$  a.e. and  $|f_n| \leq g$  a.e. for integrable  $g$ , then  $\int \lim_{n \rightarrow \infty} f_n d\nu = \lim_{n \rightarrow \infty} \int f_n d\nu$ . (iii) Monotone convergence theorem: If  $0 \leq f_1 \leq f_2 \leq \dots$  and  $\lim_{n \rightarrow \infty} f_n = f$  a.e., then  $\int \lim_{n \rightarrow \infty} f_n d\nu = \lim_{n \rightarrow \infty} \int f_n d\nu$ .

**Example 1 (Interchange of differentiation and integration):** Let  $(\Omega, \mathcal{F}, \nu)$  be a measure space and, for any fixed  $\theta \in \mathbb{R}$ , let  $f(\omega, \theta)$  be a Borel function on  $\Omega$ . Suppose that  $\partial f(\omega, \theta)/\partial \theta$  exists a.e. for  $\theta \in (a, b) \subset \mathbb{R}$  and that  $|\partial f(\omega, \theta)/\partial \theta| \leq g(\omega)$  a.e., where  $g$  is an integrable function on  $\Omega$ . Then for each  $\theta \in (a, b)$ ,  $\partial f(\omega, \theta)/\partial \theta$  is integrable and, by Theorem 1(ii),  $\frac{d}{d\theta} \int f(\omega, \theta) d\nu = \int \frac{\partial f(\omega, \theta)}{\partial \theta} d\nu$ .

**Theorem 2 (Change of variables):** Let  $f$  be measurable from  $(\Omega, \mathcal{F}, \nu)$  to  $(\Lambda, \mathcal{G})$  and  $g$  be Borel on  $(\Lambda, \mathcal{G})$ . Then  $\int_\Omega g \circ f d\nu = \int_\Lambda g d(\nu \circ f^{-1})$ , i.e., if either integral exists, then so does the other, and the two are the same.

**Theorem 3 (Fubini's theorem):** Let  $\nu_i$  be a  $\sigma$ -finite measure on  $(\Omega_i, \mathcal{F}_i), i = 1, 2$ , and  $f$  be a Borel function on  $\prod_{i=1}^2 (\Omega_i, \mathcal{F}_i)$  with  $f \geq 0$  or  $\int |f| d\nu_1 \times \nu_2 < \infty$ . Then  $g(\omega_2) = \int_{\Omega_1} f(\omega_1, \omega_2) d\nu_1$  exists a.e.  $\nu_2$  and defines a Borel function on  $\Omega_2$  whose integral w.r.t.  $\nu_2$  exists, and  $\int_{\Omega \times \Omega} f(\omega_1, \omega_2) d\nu_1 \times \nu_2 = \int_{\Omega_2} [\int_{\Omega_1} f(\omega_1, \omega_2) d\nu_1] d\nu_2$ .

## PROBABILITY THEORY

**Definition 1 (Absolutely continuous):** Let  $\lambda$  and  $\nu$  be two measures on a measurable space  $(\Omega, \mathcal{F}, \nu)$ . We say  $\lambda$  is absolutely continuous w.r.t.  $\nu$  and write  $\lambda \ll \nu$  iff  $\nu(A) = 0$  implies  $\lambda(A) = 0$ .

**Theorem 4 (Radon-Nikodym theorem):** Let  $\nu$  and  $\lambda$  be two measure on  $(\Omega, \mathcal{F})$  and  $\nu$  be  $\sigma$ -finite. If  $\lambda \ll \nu$ , then there exists a nonnegative Borel function  $f$  on  $\Omega$  such that  $\lambda(A) = \int_A f d\nu, A \in \mathcal{F}$ . Furthermore,  $f$  is unique a.e.  $\nu$ , i.e. if  $\lambda(A) = \int_A g d\nu$  for any  $A \in \mathcal{F}$ , then  $f = g$  a.e.  $\nu$ .

**Example 2:** A continuous c.d.f. may not have a p.d.f. w.r.t. Lebesgue measure. A necessary and sufficient condition for a c.d.f.  $F$  having a p.d.f. w.r.t. Lebesgue measure is that  $F$  is absolute continuous in the sense that for any  $\epsilon > 0$ , there exists a  $\delta > 0$  such that for each finite collection of disjoint bounded open intervals  $(a_i, b_i)$ ,  $\sum(b_i - a_i) < \delta$  implies  $\sum[F(b_i) - F(a_i)] < \epsilon$ .

**Proposition 2 (Calculus with Radon-Nikodym derivatives):** Let  $\nu$  be a  $\sigma$ -finite measure on a measure space  $(\Omega, \mathcal{F})$ . (i) If  $\lambda$  is a measure,  $\lambda \ll \nu$ , and  $f \geq 0$ , then  $\int f d\lambda = \int f \frac{d\lambda}{d\nu} d\nu$ . (ii) If  $\lambda_i, i = 1, 2$ , are measures and  $\lambda_i \ll \nu$ , then  $\lambda_1 + \lambda_2 \ll \nu$  and  $\frac{d(\lambda_1 + \lambda_2)}{d\nu} = \frac{d\lambda_1}{d\nu} + \frac{d\lambda_2}{d\nu}$  a.e.  $\nu$ . (iii) If  $\tau$  is a measure,  $\lambda$  is a  $\sigma$ -finite measure, and  $\tau \ll \lambda \ll \nu$ , then  $\frac{d\tau}{d\nu} = \frac{d\tau}{d\lambda} \frac{d\lambda}{d\nu}$  a.e.  $\nu$ . In particular, if  $\lambda \ll \nu$  and  $\nu \ll \lambda$  (in which case  $\lambda$  and  $\nu$  are equivalent), then  $\frac{d\lambda}{d\nu} = \left(\frac{d\nu}{d\lambda}\right)^{-1}$  a.e.  $\nu$  or  $\lambda$ . (iv) Let  $(\Omega_i, \mathcal{F}_i, \nu_i)$  be a measure space and  $\nu_i$  be  $\sigma$ -finite,  $i = 1, 2$ . Let  $\lambda_i$  be a  $\sigma$ -finite measure on  $(\Omega, \mathcal{F}_i)$  and  $\lambda_i \ll \nu_i, i = 1, 2$ . Then  $\lambda_1 \times \lambda_2 \ll \nu_1 \times \nu_2$  and  $\frac{d(\lambda_1 \times \lambda_2)}{d(\nu_1 \times \nu_2)}(\omega_1, \omega_2) = \frac{d\lambda_1}{d\nu_1}(\omega_1) \frac{d\lambda_2}{d\nu_2}(\omega_2)$  a.e.  $\nu_1 \times \nu_2$ .

### 1.3 Densities, moments, inequalities, and generating functions

**Example 1:** Let  $X$  be a random variable on  $(\Omega, \mathcal{F}, P)$  whose c.d.f.  $F_X$  has a Lebesgue p.d.f.  $f_x$  and  $F_x(c) < 1$ , where  $c$  is a fixed constant. Let  $Y = \min\{X, c\}$ . Note that  $Y^{-1}((-\infty, X]) = \Omega$  if  $x \geq c$  and  $Y^{-1}((-\infty, x]) = X^{-1}((-\infty, x])$  if  $x < c$ . Hence  $Y$  is a random variable and the c.d.f. of

$$Y \text{ is } F_Y(x) = \begin{cases} 1 & x \geq c \\ F_X(x) & x < c \end{cases}. \text{ This c.d.f. is discontinuous at } c, \text{ since } F_x(c) < 1. \text{ Thus, it does}$$

not have a Lebesgue p.d.f. It is not discrete either. Does  $P_Y$ , the probability measure corresponding to  $F_y$ , have a p.d.f. w.r.t. some measure? Consider the point mass probability measure on  $(\mathbb{R}, \mathcal{B})$ :

$$\delta_c(A) = \begin{cases} 1 & c \in A \\ 0 & c \notin A \end{cases}, A \in \mathcal{B}. \text{ Then } P_Y \ll m + \delta_c, \text{ and the p.d.f. of } P_Y \text{ is } f_Y(x) = \frac{dP_Y}{d(m + \delta_c)}(x) =$$

$$\begin{cases} 0 & x > c \\ 1 - F_X(c) & x = c \\ f_X(x) & x < c \end{cases}. \text{ To show this, it suffices to show that } \int_{(-\infty, x]} f_Y(t) d(m + \delta_c) = P_Y((-\infty, x])$$

for any  $x \in \mathcal{B}$ .

**Proposition 1 (Transformation):** Let  $X$  be a random  $k$ -vector with a Lebesgue p.d.f.  $f_X$  and let  $Y = g(X)$ , where  $g$  is a Borel function from  $(\mathbb{R}^k, \mathcal{B}^k)$  to  $(\mathbb{R}^l, \mathcal{B}^l)$ . Let  $A_1, \dots, A_m$  be disjoint sets in  $\mathcal{B}^k$  such that  $\mathcal{B}^k - (A_1 \cup \dots \cup A_m)$  has Lebesgue measure 0 and  $g$  on  $A_j$  is one-to-one with a nonvanishing Jacobian, i.e., the determinant  $\text{Det}(\partial g(x)/\partial x) \neq 0$  on  $A_j, j = 1, \dots, m$ . Then  $Y$  has the following Lebesgue p.d.f.:  $f_Y(x) = \sum_{j=1}^m |\text{Det}(\partial h_j(x)/\partial x)| f_X(h_j(x))$ , where  $h_j$  is the inverse function of  $g$  on  $A_j, j = 1, \dots, m$ .

**Example 2 (F-distribution):** Let  $X_1$  and  $X_2$  be independent random variables having the chi-

## PROBABILITY THEORY

square distributions  $\chi_{n_1}^2$  and  $\chi_{n_2}^2$ , respectively. One can show that the p.d.f. of  $Y = (X_1/n_1)/(X_2/n_2)$  is the p.d.f. of the F-distribution  $F_{n_1, n_2}$ .

**Example 3 (t-distribution):** Let  $U_1$  be a random variable having the standard normal distribution  $N(0, 1)$  and  $U_2$  a random variable having the chi-square distribution  $\chi_n^2$ . One can show that if  $U_1$  and  $U_2$  are independent, then the distribution of  $T = U_1/\sqrt{U_2/n}$  is the t-distribution  $t_n$ .

**Example 4 (Noncentral chi-square distribution):** Let  $X_1, \dots, X_n$  be independent random variables and  $X_i \sim N(\mu_i, \sigma^2)$ . The distribution of  $Y = (X_1^2 + \dots + X_n^2)/\sigma^2$  is called the noncentral chi-square distribution and denoted by  $\chi_n^2(\delta)$ , where  $\delta = (\mu_1^2 + \dots + \mu_n^2)/\sigma^2$  is the noncentrality parameter. If  $Y_1, \dots, Y_k$  are independent random variables and  $Y_i$  has the noncentral independent chi-square distribution  $\chi_{n_i}^2(\delta_i)$ ,  $i = 1, \dots, k$ , then  $Y = Y_1 + \dots + Y_k$  has the noncentral chi-square distribution  $\chi_{n_1 + \dots + n_k}^2(\delta_1 + \dots + \delta_k)$ .

**Definition 1 (Moments):** If  $\mathbb{E}X^k$  is finite, where  $k$  is a positive integer,  $\mathbb{E}X^k$  is called the  $k$ -th moment of  $X$  or  $P_X$ . If  $\mathbb{E}|X|^a < \infty$  for some real number  $a$ ,  $\mathbb{E}|X|^a$  is called the  $a$ -th absolute moment of  $X$  or  $P_X$ . If  $\mu = \mathbb{E}X$ ,  $\mathbb{E}(X - \mu)^k$  is called the  $k$ -th central moment of  $X$  or  $P_X$ .  $\text{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2$  is called the variance of  $X$  or  $P_X$ . For random matrix  $M = (M_{ij})$ ,  $\mathbb{E}M = (\mathbb{E}M_{ij})$ . For random vector  $X$ ,  $\text{Var}(X) = \mathbb{E}(X - \mathbb{E}X)(X - \mathbb{E}X)^T$  is its covariance matrix, whose  $(i, j)$ -th element,  $i \neq j$ , is called the covariance of  $X_i$  and  $X_j$  and denoted by  $\text{Cov}(X_i, X_j)$ . If  $\text{Cov}(X_i, X_j) = 0$ , then  $X_i$  and  $X_j$  are said to be uncorrelated. Independence implies uncorelation, not converse. If  $X$  is random and  $c$  is fixed, then  $\mathbb{E}(c^T X) = c^T \mathbb{E}(X)$  and  $\text{Var}(c^T X) = c^T \text{Var}(X)c$ .

**Definition 2 (Moment generating and characteristic functions):** Let  $X$  be a random  $k$ -vector. (i) The moment generating function (m.g.f.) of  $X$  or  $P_X$  is defined as  $\psi_X(t) = \mathbb{E}e^{t^T X}$ ,  $t \in \mathbb{R}^k$ . (ii) The characteristic function (ch.f.) of  $X$  or  $P_X$  is defined as  $\phi_X(t) = \mathbb{E}e^{it^T X} = \mathbb{E}[\cos(t^T X)] + i\mathbb{E}[\sin(t^T X)]$ ,  $t \in \mathbb{R}^k$ .

**Proposition 2 (Properties of m.g.f. and ch.f.):** If the m.g.f. is finite in a neighborhood of  $0 \in \mathbb{R}^k$ , then (i) moments of  $X$  of any order are finite; (ii)  $\phi_X(t)$  can be obtained by replacing  $t$  in  $\psi_X(t)$  by  $it$ . If  $Y = A^T X + c$ , where  $A$  is a  $k \times m$  matrix and  $c \in \mathbb{R}^m$ , then  $\psi_Y(u) = e^{c^T u} \psi_X(Au)$  and  $\phi_Y(u) = e^{ic^T u} \phi_X(Au)$ ,  $u \in \mathbb{R}^m$ . For independent  $X_1, \dots, X_k$ ,  $\psi_{\sum_i X_i}(t) = \prod_i \psi_{X_i}(t)$  and  $\phi_{\sum_i X_i}(t) = \prod_i \phi_{X_i}(t)$ ,  $t \in \mathbb{R}^k$ . For  $X = (X_1, \dots, X_k)$  with m.g.f.  $\psi_X$  finite in a neighborhood of 0,  $\frac{\partial \psi_X(t)}{\partial t}|_{t=0} = \mathbb{E}X$ ,  $\frac{\partial^2 \psi_X(t)}{\partial t \partial t^T}|_{t=0} = \mathbb{E}(XX^T)$ . If  $\mathbb{E}|X_1^{r_1} \dots X_k^{r_k}| < \infty$  for nonnegative integers  $r_1, \dots, r_k$ , then  $\frac{\partial \phi_X(t)}{\partial t}|_{t=0} = i\mathbb{E}X$ ,  $\frac{\partial^2 \phi_X(t)}{\partial t \partial t^T}|_{t=0} = -\mathbb{E}(XX^T)$ .

**Theorem 1 (Uniqueness):** Let  $X$  and  $Y$  be random  $k$ -vectors. (i) If  $\phi_X(t) = \phi_Y(t)$  for all  $t \in \mathbb{R}^k$ , then  $P_X = P_Y$ ; (2) If  $\psi_X(t) = \psi_Y(t) < \infty$  for all  $t$  in a neighborhood of 0, then  $P_X = P_Y$ .

### 1.4 Conditional expectation and independence

**Definition 1:** Let  $X$  be an integrable random variable on  $(\Omega, \mathcal{F}, P)$ . (i) The conditional expectation of  $X$  given  $\mathcal{A}$  (a sub- $\sigma$ -field of  $\mathcal{F}$ ), denoted by  $\mathbb{E}(X|\mathcal{A})$ , is the a.s. unique random variable satisfying the following two conditions: (a)  $\mathbb{E}(X|\mathcal{A})$  is measurable from  $(\Omega, \mathcal{A})$  to  $(\mathbb{R}, \mathcal{B})$ ; (b)  $\int_A \mathbb{E}(X|\mathcal{A}) dP = \int_A X dP$  for any  $A \in \mathcal{A}$ . (ii) The conditional probability of  $B \in \mathcal{F}$  given  $\mathcal{A}$  is defined to be  $P(B|\mathcal{A}) = \mathbb{E}(I_B|\mathcal{A})$ . (iii) Let  $Y$  be measurable from  $(\Omega, \mathcal{F}, P)$  to  $(\Lambda, \mathcal{G})$ . The conditional expectation of  $X$  given  $Y$  is defined to be  $\mathbb{E}(X|Y) = \mathbb{E}[X|\sigma(Y)]$ .

**Theorem 1:** Let  $Y$  be measurable from  $(\Omega, \mathcal{F})$  to  $(\Lambda, \mathcal{G})$  and  $Z$  a function from  $(\Omega, \mathcal{F})$  to  $\mathbb{R}^k$ . Then  $Z$  is measurable from  $(\Omega, \sigma(Y))$  to  $(\mathbb{R}^k, \mathcal{B}^k)$  iff there is a measurable function  $h$  from  $(\Lambda, \mathcal{G})$  such that  $Z = h \circ Y$ .

**Example 1:** Let  $X$  be an integrable random variable on  $(\Omega, \mathcal{F}, P)$ ,  $A_1, A_2, \dots$  be disjoint events on  $(\Omega, \mathcal{F}, P)$  such that  $\cup A_i = \Omega$  and  $P(A_i) > 0$  for all  $i$ , and let  $a_1, a_2, \dots$  be distinct real numbers. Define  $Y = a_1 I_{A_1} + a_2 I_{A_2} + \dots$ . We can show that  $\mathbb{E}(X|Y) = \sum_{i=1}^{\infty} \frac{\int_{A_i} X dP}{P(A_i)} I_{A_i}$ .

**Proposition 1:** Let  $X$  be a random  $n$ -vector and  $Y$  a random  $m$ -vector. Suppose that  $(X, Y)$  has a joint p.d.f.  $f(x, y)$  w.r.t.  $\nu \times \lambda$ , where  $\nu$  and  $\lambda$  are  $\sigma$ -finite measures on  $(\mathbb{R}^n, \mathcal{B}^n)$  and  $(\mathbb{R}^m, \mathcal{B}^m)$ , respectively. Let  $g(x, y)$  be a Borel function on  $\mathbb{R}^{n+m}$  for which  $\mathbb{E}|g(X, Y)| < \infty$ . Then  $\mathbb{E}[g(X, Y)|Y] = \frac{\int g(x, Y)f(x, Y)d\nu(x)}{\int f(x, Y)d\nu(x)}$  a.s.

**Definition 2 (Conditional p.d.f.):** Let  $(X, Y)$  be a random vector with a joint p.d.f.  $f(x, y)$  w.r.t.  $\nu \times \lambda$ . The conditional p.d.f. of  $X$  given  $Y = y$  is defined to be  $f_{X|Y}(x|y)/f_Y(y)$  where  $f_Y(y) = \int f(x, y)d\nu(x)$  is the marginal p.d.f. of  $Y$  w.r.t.  $\lambda$ .

**Proposition 2:** Let  $X, Y, X_1, X_2, \dots$  be integrable random variables on  $(\Omega, \mathcal{F}, P)$  and  $\mathcal{A}$  be a sub- $\sigma$ -field of  $\mathcal{F}$ . (i) If  $X = c$  a.s.,  $c \in \mathbb{R}$ , then  $\mathbb{E}(X|\mathcal{A}) = c$  a.s. (ii) If  $X \leq Y$  a.s., then  $\mathbb{E}(X|\mathcal{A}) \leq \mathbb{E}(Y|\mathcal{A})$  a.s. (iii) If  $a, b \in \mathbb{R}$ , then  $\mathbb{E}(aX + bY|\mathcal{A}) = a\mathbb{E}(X|\mathcal{A}) + b\mathbb{E}(Y|\mathcal{A})$  a.s. (iv)  $\mathbb{E}[\mathbb{E}(X|\mathcal{A})] = \mathbb{E}X$ . (v)  $\mathbb{E}[\mathbb{E}(X|\mathcal{A})|\mathcal{A}_0] = \mathbb{E}(X|\mathcal{A}_0) = \mathbb{E}[\mathbb{E}(X|\mathcal{A}_0)|\mathcal{A}]$  a.s., where  $\mathcal{A}_0$  is a sub- $\sigma$ -field of  $\mathcal{A}$ . (vi) If  $\sigma(Y) \subset \mathcal{A}$  and  $\mathbb{E}|XY| < \infty$ , then  $\mathbb{E}(XY|\mathcal{A}) = Y\mathbb{E}(X|\mathcal{A})$  a.s. (vii) If  $X$  and  $Y$  are independent and  $\mathbb{E}|g(X, Y)| < \infty$  for a Borel function  $g$ , then  $\mathbb{E}[g(X, Y)|Y = y] = \mathbb{E}[g(X, y)]$  a.s.  $P_Y$ . (viii) If  $\mathbb{E}X^2 < \infty$ , then  $[\mathbb{E}(X|\mathcal{A})]^2 \leq \mathbb{E}(X^2|\mathcal{A})$  a.s. (ix) Fatou's lemma: If  $X_n \geq 0$  for any  $n$ , then  $\mathbb{E}(\liminf_n X_n|\mathcal{A}) \leq \liminf_n \mathbb{E}(X_n|\mathcal{A})$  a.s. (x) Dominated convergence theorem: If  $|X_n| \leq Y$  for any  $n$  and  $X_n \rightarrow_{\text{a.s.}} X$ , then  $\mathbb{E}(X_n|\mathcal{A}) \rightarrow_{\text{a.s.}} \mathbb{E}(X|\mathcal{A})$ .

**Definition 3 (Independence):** Let  $(\Omega, \mathcal{F}, P)$  be a probability space. (i) Let  $\mathcal{C}$  be a collection of subsets in  $\mathcal{F}$ . Events in  $\mathcal{C}$  are said to be independent iff for any positive integer  $n$  and distinct events  $A_1, \dots, A_n \in \mathcal{C}$ ,  $P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1)P(A_2) \dots P(A_n)$ . (ii) Collections  $\mathcal{C}_i \subset \mathcal{F}, i \in \mathcal{I}$  are said to be independent iff events in any collection of the form  $\{A_i \in \mathcal{C}_i : i \in \mathcal{J}\}$  are independent. (iii) Random elements  $X_i, i \in \mathcal{I}$ , are said to be independent iff  $\sigma(X_i), i \in \mathcal{I}$  are independent.

**Theorem 2:** Let  $\mathcal{C}_i, i \in \mathcal{I}$  be independent collections of events. If each  $\mathcal{C}_i$  is a  $\pi$ -system, then  $\sigma(\mathcal{C}_i), i \in \mathcal{I}$  are independent.

**Proposition 2:** Let  $X$  be a random variable with  $\mathbb{E}|X| < \infty$  and let  $Y_i$  be random  $k_i$  vectors,  $i = 1, 2$ . Suppose that  $(X, Y_1)$  and  $Y_2$  are independent. Then  $\mathbb{E}[X|(Y_1, Y_2)] = \mathbb{E}(X|Y_1)$  a.s.

**Definition 4 (Conditional independence):** Let  $X, Y, Z$  be random vectors. We say that given  $Z$ ,  $X$  and  $Y$  are conditionally independent iff  $P(A|X, Z) = P(A|Z)$  a.s. for any  $A \in \sigma(Y)$ .

## 1.5 Convergence modes and relationships

**Definition 1 (Convergence modes):** Let  $X, X_1, X_2, \dots$  be a random  $k$ -vectors defined on a probability space. (i) We say that the sequence  $\{X_n\}$  converges to  $X$  almost surely and write  $X_n \rightarrow_{\text{a.s.}} X$  iff  $\lim_{n \rightarrow \infty} X_n = X$  a.s. (ii) We say that  $\{X_n\}$  converges to  $X$  in probability and write  $X_n \rightarrow_p X$  iff for every fixed  $\epsilon > 0$ ,  $\lim_{n \rightarrow \infty} P(\|X_n - X\| > \epsilon) = 0$ . (iii) We say that  $\{X_n\}$  converges to  $X$  in  $L_r$  (or in  $r$ th moment) with a fixed  $r > 0$  and write  $X_n \rightarrow_{L_r} X$  iff  $\lim_{n \rightarrow \infty} \mathbb{E}\|X_n - X\|_r^r = 0$ . (iv)



## PROBABILITY THEORY

Let  $F, F_n, n = 1, 2, \dots$  be c.d.f.'s on  $\mathbb{R}^k$  and  $P, P_n, n = 1, 2, \dots$  be their corresponding probability measures. We say that  $\{F_n\}$  converges to  $F$  weakly (or  $\{P_n\}$  converges to  $P$  weakly) and write  $F_n \rightarrow_w F$  (or  $P_n \rightarrow_w P$ ) iff, for each continuity point  $x$  of  $F$ ,  $\lim_{n \rightarrow \infty} F_n(x) = F(x)$ . We say that  $\{X_n\}$  converges to  $X$  in distribution (or in law) and write  $X_n \rightarrow_d X$  iff  $F_{X_n} \rightarrow_w F_X$ .

**Proposition 1:** If  $F_n \rightarrow_w F$  and  $F$  is continuous on  $\mathbb{R}^k$ , then  $\lim_{n \rightarrow \infty} \sup_{x \in \mathbb{R}^k} |F_n(x) - F(x)| = 0$ .

**Theorem 1:** For random  $k$ -vectors  $X, X_1, X_2, \dots$  on a probability space,  $X_n \rightarrow_{a.s.} X$  iff for every  $\epsilon > 0$ ,  $\lim_{n \rightarrow \infty} P(\cup_{m=n}^{\infty} \{\|X_m - X\| > \epsilon\}) = 0$ .

**Theorem 2 (Borel-Cantelli lemma):** Let  $A_n$  be a sequence of events in a probability space and  $\limsup_n A_n = \cap_{n=1}^{\infty} \cup_{m=n}^{\infty} A_m$ . (i) If  $\sum_{n=1}^{\infty} P(A_n) < \infty$ , then  $P(\liminf_n A_n) = 0$ . (ii) If  $A_1, A_2, \dots$  are pairwise independent and  $\sum_{n=1}^{\infty} P(A_n) = \infty$ , then  $P(\limsup_n A_n) = 1$ .

**Definition 2:** Let  $X_1, X_2, \dots$  be random vectors and  $Y_1, Y_2, \dots$  be random variables defined on a common probability space. (i)  $X_n = O(Y_n)$  a.s. iff  $P(\|X_n\| = O(|Y_n|)) = 1$ . (ii)  $X_n = o(Y_n)$  a.s. iff  $X_n/Y_n \rightarrow_{a.s.} 0$ . (iii)  $X_n = O_p(Y_n)$  iff, for any  $\epsilon > 0$ , there is a constant  $C_\epsilon > 0$  such that  $\sup_n P(\|X_n\| \geq C_\epsilon |Y_n|) < \epsilon$ . (iv)  $X_n = o_p(Y_n)$  iff  $X_n/Y_n \rightarrow_p 0$ .

**Theorem 3:** (i) If  $X_n \rightarrow_{a.s.} X$ , then  $X_n \rightarrow_p X$ . (The converse is not true). (ii) If  $X_n \rightarrow_{L_r} X$  for an  $r > 0$ , then  $X_n \rightarrow_p X$ . (The converse is not true). (iii) If  $X_n \rightarrow_p X$ , then  $X_n \rightarrow_d X$ . (The converse is not true). (iv) (Skorohod's theorem). If  $X_n \rightarrow_d X$ , then there are random vectors  $Y, Y_1, Y_2, \dots$  defined on a common probability space such that  $P_Y = P_X, P_{Y_n} = P_{X_n}, n = 1, 2, \dots$  and  $Y_n \rightarrow_{a.s.} Y$ . (v) If, for every  $\epsilon > 0$ ,  $\sum_{n=1}^{\infty} P(\|X_n - X\| \geq \epsilon) < \infty$ , then  $X_n \rightarrow_{a.s.} X$ . (vi) If  $X_n \rightarrow_p X$ , then there is a subsequence such that  $X_{n_j} \rightarrow_{a.s.} X$  as  $j \rightarrow \infty$ . (vii) If  $X_n \rightarrow_d X$  and  $P(X = c) = 1$ , where  $c \in \mathbb{R}^k$  is a constant vector, then  $X_n \rightarrow_p c$ . (viii) Suppose that  $X_n \rightarrow_d X$ . Then for any  $r > 0$ ,  $\lim_{n \rightarrow \infty} \mathbb{E}\|X_n\|_r^r = \mathbb{E}\|X\|_r^r < \infty$  if  $\{\|X_n\|_r^r\}$  is uniformly integrable in the sense that  $\lim_{t \rightarrow \infty} \sup_n \mathbb{E}(\|X_n\|_r^r I_{\{\|X_n\|_r > t\}}) = 0$ .

**Proposition 2 (Sufficient conditions for uniform integrability):**  $\sup_n \mathbb{E}\|X_n\|_r^{r+\delta} < \infty$  for a  $\delta > 0$ .

**Proposition 3 (Properties of the quotient random variables):** (i) Suppose  $X, X_1, X_2, \dots$  are positive random variables. Then  $X_n \rightarrow_{a.s.} X$  iff for every  $\epsilon > 0$ ,  $\lim_{n \rightarrow \infty} P(\sup_{k \geq n} \frac{X_k}{X} > 1 + \epsilon) = 0$ , and  $\lim_{n \rightarrow \infty} P(\sup_{k \geq n} \frac{X}{X_k} > 1 + \epsilon) = 0$ . (ii) Suppose  $X, X_1, X_2, \dots$  are positive random variables. If  $\sum_{n=1}^{\infty} P(X_n/X > 1 + \epsilon) < \infty$  and  $\sum_{n=1}^{\infty} P(X/X_n > 1 + \epsilon) < \infty$ , then  $X_n \rightarrow_{a.s.} X$ .

### 1.6 Uniform integrability and weak convergence

**Definition 1 (Tightness):** A sequence  $\{P_n\}$  of probability measure on  $(\mathbb{R}^k, \mathcal{B}^k)$  is tight if for every  $\epsilon > 0$ , there is a compact set  $C \subset \mathbb{R}^k$  such that  $\inf_n P_n(C) > 1 - \epsilon$ . If  $\{X_n\}$  is a sequence of random  $k$ -vectors, then the tightness of  $\{P_{X_n}\}$  is the same as the boundedness of  $\{\|X_n\|\}$  in probability.

**Proposition 1:** Let  $\{P_n\}$  be a sequence of probability measures on  $(\mathbb{R}^k, \mathcal{B}^k)$ . (i) Tightness of  $\{P_n\}$  is a necessary and sufficient condition that for every subsequence  $\{P_n\}$  there exists a further subsequence  $\{P_{n_j}\} \subset \{P_n\}$  and a probability measure  $P$  on  $(\mathbb{R}^k, \mathcal{B}^k)$  such that  $P_{n_j} \rightarrow_w P$  as  $j \rightarrow \infty$ . (ii) If  $\{P_n\}$  is tight and if each subsequence that converges weakly at all converges to the same probability measure  $P$ , then  $P_n \rightarrow_w P$ .

**Theorem 1 (Useful sufficient and necessary conditions for convergence in distribution):** Let  $X, X_1, X_2, \dots$  be random  $k$ -vectors. (i)  $X_n \rightarrow_d X$  is equivalent to any one of the following conditions:



(a)  $\mathbb{E}[h(X_n)] \rightarrow \mathbb{E}[h(X)]$  for every bounded continuous function  $h$ ; (b)  $\limsup_n P_{X_n}(C) \leq P_X(C)$  for any closed set  $C \subset \mathbb{R}^k$ ; (c)  $\liminf_n P_{X_n}(O) \geq P_X(O)$  for any open set  $O \subset \mathbb{R}^k$ . (ii) Lévy-Cramér continuity theorem. Let  $\phi_X, \phi_{X_1}, \phi_{X_2}, \dots$  be the ch.f.'s of  $X, X_1, X_2, \dots$ , respectively.  $X_n \rightarrow_d X$  iff  $\lim_{n \rightarrow \infty} \phi_{X_n}(t) = \phi_X(t)$  for all  $t \in \mathbb{R}^k$ . (iii) Cramér-Wold device.  $X_n \rightarrow_d X$  iff  $c^T X_n \rightarrow_d c^T X$  for every  $c \in \mathbb{R}^k$ .

**Example 1:** Let  $X_1, \dots, X_n$  be independent random variables having a common c.d.f. and  $T_n = X_1 + \dots + X_n, n = 1, 2, \dots$ . Suppose that  $\mathbb{E}|X_1| < \infty$ . It follows from a result in calculus that the ch.f. of  $X_1$  satisfies  $\phi_{X_1}(t) = \phi_{X_1}(0) + \sqrt{-1}\mu t + o(|t|)$  as  $|t| \rightarrow 0$ , where  $\mu = \mathbb{E}X_1$ . Then, the ch.f. of  $T_n/n$  is  $\phi_{T_n/n}(t) = [\phi_{X_1}(\frac{t}{n})]^n = [1 + \frac{\sqrt{-1}\mu t}{n} + o(\frac{t}{n})]^n \rightarrow e^{\sqrt{-1}\mu t}$  for any  $t \in \mathbb{R}$  as  $n \rightarrow \infty$ .  $e^{\sqrt{-1}\mu t}$  is the ch.f. of the point mass probability measure at  $\mu$ . Thus  $T_n/n \rightarrow_d \mu$  and  $T_n/n \rightarrow_p \mu$ .

**Proposition 2 (Scheffé's theorem):** Let  $\{f_n\}$  be a sequence of p.d.f.'s on  $\mathbb{R}^k$  w.r.t.  $\nu$ . Suppose that  $\lim_{n \rightarrow \infty} f_n(x) = f(x)$  a.e. and  $f(x)$  is a p.d.f. w.r.t.  $\nu$ . Then  $\lim_{n \rightarrow \infty} \int |f_n(x) - f(x)| d\nu = 0$ .

## 1.7 Convergence of transformations and law of large numbers

**Theorem 1 (Continuous mapping theorem):** Let  $X, X_1, X_2, \dots$  be random  $k$ -vectors defined on a probability space and  $g$  be a measure function from  $(\mathbb{R}^k, \mathcal{B}^k)$  to  $(\mathbb{R}^l, \mathcal{B}^l)$ . Suppose that  $g$  is continuous a.s.  $P_X$ . Then (i)  $X_n \rightarrow_{a.s.} X$  implies  $g(X_n) \rightarrow_{a.s.} g(X)$ ; (ii)  $X_n \rightarrow_p X$  implies  $g(X_n) \rightarrow_p g(X)$ ; (iii)  $X_n \rightarrow_d X$  implies  $g(X_n) \rightarrow_d g(X)$ .

**Theorem 2 (Slutsky's theorem):** Let  $X, X_1, X_2, \dots, Y_1, Y_2, \dots$  be random variables on a probability space. Suppose that  $X_n \rightarrow_d X$  and  $Y_n \rightarrow_p c$ , where  $c$  is a constant. Then (i)  $X_n + Y_n \rightarrow_d X + c$ ; (ii)  $Y_n X_n \rightarrow_d cX$ ; (iii)  $X_n/Y_n \rightarrow_d X/c$  if  $c \neq 0$ .

**Theorem 3:** Let  $X_1, X_2, \dots$  and  $Y = (Y_1, \dots, Y_k)$  be random  $k$ -vectors satisfying  $a_n(X_n - c) \rightarrow_d Y$ , where  $c \in \mathbb{R}^k$  and  $\{a_n\}$  is a sequence of positive numbers with  $\lim_{n \rightarrow \infty} a_n = \infty$ . Let  $g$  be a function from  $\mathbb{R}^k \rightarrow \mathbb{R}$ . (i) If  $g$  is differentiable at  $c$ , then  $a_n[g(X_n) - g(c)] \rightarrow_d [\nabla g(c)^T]Y$ , where  $\nabla g(x)$  denotes the  $k$ -vector of partial derivatives of  $g$  at  $x$ . (ii) Suppose that  $g$  has continuous partial derivatives of order  $m > 1$  in a neighborhood of  $c$ , with all the partial derivatives of order  $j, 1 \leq j \leq m-1$ , vanishing at  $c$ , but with the  $m$ th-order partial derivatives not all vanishing at  $c$ . Then  $a_n^m[g(X_n) - g(c)] \rightarrow_d \frac{1}{m!} \sum_{i_1=1}^k \dots \sum_{i_m=1}^k \frac{\partial^m g}{\partial x_{i_1} \dots \partial x_{i_m}}|_{x=c} Y_{i_1} \dots Y_{i_m}$ .

**Theorem 4 (The  $\delta$ -method):** If  $Y$  has the  $\mathcal{N}_k(0, \Sigma)$  distribution, then  $a_n[g(X_n) - g(c)] \rightarrow_d \mathcal{N}(0, [\nabla g(c)]^T \Sigma \nabla g(c))$ .

**Theorem 5:** Let  $X_1, X_2, \dots$  be i.i.d. random variables. (i) The WLLN. A necessary and sufficient condition for the existence of a sequence of real numbers  $\{a_n\}$  for which  $\frac{1}{n} \sum_{i=1}^n X_i - a_n \rightarrow_p 0$  is that  $nP(|X_1| > n) \rightarrow 0$ , in which case we may take  $a_n = \mathbb{E}(X_1 1_{\{|X_1| \leq n\}})$ . (ii) The SLLN. A necessary and sufficient condition for the existence of a constant  $c$  for which  $\frac{1}{n} \sum_{i=1}^n X_i \rightarrow_{a.s.} c$  is that  $\mathbb{E}|X_1| < \infty$ , in which case  $c = \mathbb{E}X_1$  and  $\frac{1}{n} \sum_{i=1}^n c_i(X_i - \mathbb{E}X_1) \rightarrow_{a.s.} 0$  for any bounded sequence of real numbers  $\{c_i\}$ .

**Theorem 6:** Let  $X_1, X_2, \dots$  be independent random variables with finite expectations. (i) The SLLN. If there is a constant  $p \in [1, 2]$  such that  $\sum_{i=1}^{\infty} \frac{\mathbb{E}|X_i|^p}{i^p} < \infty$ , then  $\frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}X_i) \rightarrow_{a.s.} 0$ . (ii) The WLLN. If there is a constant  $p \in [1, 2]$  such that  $\lim_{n \rightarrow \infty} \frac{1}{n^p} \sum_{i=1}^n \mathbb{E}|X_i|^p = 0$ , then  $\frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}X_i) \rightarrow_p 0$ .

## 1.8 The central limit theorem

**Theorem 1 (Lindeberg's CLT):** Let  $\{X_{nj}, j = 1, \dots, k_n\}$  be independent random variables with  $k_n \rightarrow \infty$  as  $n \rightarrow \infty$  and  $0 < \sigma_n^2 = \text{var}(\sum_{j=1}^{k_n} X_{nj}) < \infty, n = 1, 2, \dots$ . If  $\frac{1}{\sigma_n^2} \sum_{j=1}^{k_n} \mathbb{E}[(X_{nj} - \mathbb{E}X_{nj})^2 I_{\{|X_{nj} - \mathbb{E}X_{nj}| > \epsilon \sigma_n\}}] \rightarrow 0$  for any  $\epsilon > 0$ , then  $\frac{1}{\sigma_n} \sum_{j=1}^{k_n} (X_{nj} - \mathbb{E}X_{nj}) \rightarrow_d \mathcal{N}(0, 1)$ .

**Theorem 2 (Multivariate CLT):** For i.i.d. random  $k$ -vectors  $X_1, \dots, X_n$  with a finite  $\Sigma = \text{var}(X_1)$ ,  $\frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - \mathbb{E}X_1) \rightarrow_d \mathcal{N}_k(0, \Sigma)$ .

**Theorem 3 (Berry-Esséen bound):** For i.i.d.  $\{X_n\}$  and  $W_n = \sqrt{n}(\bar{X} - \mu)/\sigma$ ,  $\sup_t |F_{W_n}(t) - \phi(t)| \leq \frac{33}{4} \frac{\mathbb{E}|X_1 - \mu|^3}{\sigma^3 \sqrt{n}}, n = 1, 2, \dots$ . Thus, the convergence speed of  $F_{W_n}$  to  $\phi$  is of the order  $n^{-1/2}$ .

## 2 Fundamentals of Statistics

### 2.1 Models, data, statistics, and sampling distributions

**Definition 1:** A set of probability measures  $P_\theta$  on  $(\Omega, \mathcal{F})$  indexed by a parameter  $\theta \in \Theta$  is said to be a parametric family or follow a parametric model iff  $\Theta \subset \mathbb{R}^d$  for some fixed positive integer  $d$  and each  $P_\theta$  is a known probability measure when  $\theta$  is known. The set  $\Theta$  is called the parameter space and  $d$  is called its dimension.  $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$  is identifiable iff  $\theta_1 \neq \theta_2$  and  $\theta_i \in \Theta$  imply  $P_{\theta_1} \neq P_{\theta_2}$ , which may be achieved through reparameterization.

**Definition 2 (Dominated family):** A family of populations  $\mathcal{P}$  is dominated by  $\nu$  (a  $\sigma$ -finite measure) if  $P \ll \nu$  for all  $P \in \mathcal{P}$ , in which case  $\mathcal{P}$  can be identified by the family of densities  $\{\frac{dP}{d\nu} : P \in \mathcal{P}\}$  or  $\{\frac{dP_\theta}{d\nu} : \theta \in \Theta\}$ .

**Definition 3 (Exponential families):** A parametric family  $\{P_\theta : \theta \in \Theta\}$  dominated by a  $\sigma$ -finite measure  $\nu$  on  $(\Omega, \mathcal{F})$  is called an exponential family iff  $\frac{dP_\theta}{d\nu}(\omega) = \exp\{[\eta(\theta)]^T T(\omega) - \xi(\theta)\} h(\omega), \omega \in \Omega$  where  $\xi(\theta) = \log\{\int_\Omega \exp\{[\eta(\theta)]^T T(\omega)\} h(\omega) d\nu(\omega)\}$ . In an exponential family, consider the parameter  $\eta = \eta(\theta)$  and  $f_\eta(\omega) = \exp\{\eta^T T(\omega) - \zeta(\eta)\} h(\omega), \omega \in \Omega$ . This is called the canonical form for the family, and  $\Xi = \{\eta : \zeta(\eta) \text{ is defined}\}$  is called the natural parameter space. An exponential family in canonical form is 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.

**Theorem 1:** Let  $\mathcal{P}$  be a natural exponential family. (i) Let  $T = (Y, U)$  and  $\eta = (\theta, \phi)$ ,  $Y$  and  $\theta$  have the same dimension. Then,  $Y$  has the p.d.f.  $f_\eta(y) = \exp\{\theta^T y - \zeta(\eta)\}$ . 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.  $f_{\theta, u}(y) = \exp\{\theta^T y - \zeta_u(\theta)\}$  w.r.t. a  $\sigma$ -finite measure depending on  $\phi$ . Furthermore, the conditional distribution of  $Y$  given  $U = u$  has the p.d.f.  $f_{\theta, u}(y) = \exp(\theta^T y - \zeta_u(\theta))$  w.r.t. a  $\sigma$ -finite measure depending on  $u$ . (ii) If  $\eta_0$  is an interior point of the natural parameter space, then the m.g.f. of  $P_{\eta_0} \circ T^{-1}$  is finite in a neighborhood of 0 and is given by  $\psi_{\eta_0}(t) = \exp\{\zeta(\eta_0 + t) - \zeta(\eta_0)\}$ .

**Definition 4 (Location-scale families):** Let  $P$  be a known probability measure on  $(\mathbb{R}^k, \mathcal{B}^k)$ ,  $\mathcal{V} \subset \mathbb{R}^k$ , and  $\mathcal{M}_k$  be a collection of  $k \times k$  symmetric positive definite matrices. The family  $\{P_{(\mu, \Sigma)} : \mu \in \mathcal{V}, \Sigma \in \mathcal{M}_k\}$  is called a location-scale family (on  $\mathbb{R}^k$ ), where  $P_{(\mu, \Sigma)}(B) = P(\Sigma^{-1/2}(B - \mu)), B \in \mathcal{B}^k$ . The parameters  $\mu$  and  $\Sigma^{1/2}$  are called the location and scale parameters, respectively.

**Definition 5 (Statistics and their sampling distributions):** Our data set is a realization of a sample

(random vector)  $X$  from an unknown population  $P$ . Statistic  $T(X)$ : A measurable function  $T$  of  $X$ ;  $T(X)$  is a known value whenever  $X$  is known. A nontrivial statistic  $T(X)$  is usually simpler than  $X$ . Finding the form of the distribution of  $T$  is one of the major problems in statistical inference and decision theory.

**Example 1:** Let  $X_1, \dots, X_n$  be i.i.d. random variables having a common distribution  $P$ . The sample mean and sample variance  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ ,  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$  are two commonly used statistics.

**Example 2 (Order statistics):** Let  $X = (X_1, \dots, X_n)$  with i.i.d. random components. Let  $X_{(i)}$  be the  $i$ th smallest value of  $X_1, \dots, X_n$ . The statistics  $X_{(1)}, \dots, X_{(n)}$  are called the order statistics.

## 2.2 Sufficiency and minimal sufficiency

**Definition 1 (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  $P \in \mathcal{P}$  iff conditional distribution of  $X$  given  $T$  is known.

**Theorem 1 (The factorization theorem):** Suppose that  $X$  is a sample from  $P \in \mathcal{P}$  and  $\mathcal{P}$  is a family of probability measures on  $(\mathbb{R}^n, \mathcal{B}^n)$  dominated by a  $\sigma$ -finite measure  $\nu$ . Then  $T(X)$  is sufficient for  $P \in \mathcal{P}$  iff there are nonnegative Borel functions  $h$  and  $g_p$  on the range of  $T$  such that  $\frac{dP}{d\nu}(x) = g_p(T(x))h(x)$ .

**Theorem 2:** If a family  $\mathcal{P}$  is dominated by a  $\sigma$ -finite measure, then  $\mathcal{P}$  is dominated by a probability measure  $Q = \sum_{i=1}^{\infty} c_i P_i$ , where  $c_i$ 's are nonnegative constants with  $\sum_{i=1}^{\infty} c_i = 1$  and  $P_i \in \mathcal{P}$ .

**Convention 1:** 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 (Minimal sufficiency):** Let  $T$  be a sufficient statistic for  $P \in \mathcal{P}$ .  $T$  is called a minimal sufficient statistic iff, for any other statistic  $S$  sufficient for  $P \in \mathcal{P}$ , there is a measurable function  $\psi$  such that  $T = \psi(S)$  a.s.  $\mathcal{P}$ .

**Theorem 3 (Existence and uniqueness):** Minimal sufficient statistics exist when  $\mathcal{P}$  contains distributions on  $\mathbb{R}^k$  dominated by a  $\sigma$ -finite measure. If both  $T$  and  $S$  are minimal sufficient statistics, then by definition there is one-to-one measurable function  $\psi$  such that  $T = \psi(S)$  a.s.  $\mathcal{P}$ .

**Theorem 4:** Let  $\mathcal{P}$  be a family of distributions on  $\mathbb{R}^k$ . (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  $f_0, f_1, f_2, \dots$  w.r.t. a  $\sigma$ -finite  $\nu$ . 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  $f_{\infty}(x) > 0$ ,  $i = 0, 1, 2, \dots$ . Then  $T(x) = (T_0, T_1, T_2, \dots)$  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  $f_{\infty}(x)$  for  $f_0(x)$ , in which case  $T(x) = (T_1, T_2, \dots)$  is minimal sufficient for  $P \in \mathcal{P}$ . (iii) Suppose that  $\mathcal{P}$  contains p.d.f.'s  $f_p$  w.r.t. a  $\sigma$ -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  $\phi$  is a measurable function. Then  $T(x)$  is minimal sufficient for  $P \in \mathcal{P}$ .

### 2.3 Completeness

**Definition 1 (Ancillary statistics):** A statistic  $V(x)$  is ancillary iff its distribution does not depend on any unknown quantity. A statistic  $V(X)$  is first-order ancillary iff  $\mathbb{E}[V(X)]$  does not depend on any unknown quantity.

**Remark 1:** If  $V(x)$  is a non-trivial ancillary statistic, then  $\sigma(V)$  does not contain any information about the unknown population  $P$ . If  $T(x)$  is a statistic and  $V(T(x))$  is a non-trivial ancillary statistic, it indicates that the reduced data set by  $T$  contains a non-trivial part that does not contain any information about  $\theta$  and, hence, a further simplification of  $T$  may still be needed.

**Definition 2 (Completeness):** A statistic  $T(x)$  is complete (or boundedly complete) for  $P \in \mathcal{P}$  iff, for any Borel  $f$  (or bounded Borel  $f$ ),  $\mathbb{E}[f(T)] = 0$  for all  $P \in \mathcal{P}$  implies  $f = 0$  a.s.  $\mathcal{P}$ .

**Remark 2:** If  $T$  is complete (or boundedly complete) and  $S = \psi(T)$  for a measurable  $\psi$ , then  $S$  is complete (or boundedly complete). A complete and sufficient statistic should be minimal sufficient. But a minimal sufficient statistic may be not complete.

**Proposition 1:** If  $P$  is in an exponential family of full rank with p.d.f.'s given by  $f_\eta(x) = \exp\{\eta^T T(x) - \zeta(\eta)\}h(x)$ , then  $T(x)$  is complete and sufficient for  $\eta \in \Xi$ .

**Example 1:** Suppose that  $X_1, \dots, X_n$  are i.i.d. random variables having the  $\mathcal{N}(\mu, \sigma^2)$  distribution,  $\mu \in \mathbb{R}$ ,  $\sigma > 0$ . The joint p.d.f. of  $X_1, \dots, X_n$  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  $\eta = (\eta_1, \eta_2) = (\frac{\mu}{\sigma^2}, \frac{1}{2\sigma^2})$ . Hence, the family of distributions for  $X = (X_1, \dots, X_n)$  is a natural exponential family of full rank ( $\Xi = \mathbb{R} \times (0, \infty)$ ). Thus  $T(X) = (T_1, T_2)$  is complete and sufficient for  $\eta$ .

**Example 2:**  $T(x) = (X_{(1)}, \dots, X_{(n)})$  of i.i.d. random variables  $X_1, \dots, X_n$  is sufficient for  $P \in \mathcal{P}$ , where  $\mathcal{P}$  is the family of distributions on  $\mathbb{R}$  having Lebesgue p.d.f.'s. We can show that  $T(x)$  is also complete for  $P \in \mathcal{P}$ .

**Theorem 1 (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  $P \in \mathcal{P}$ .

**Example 3:**  $X_1, \dots, X_n$  is a random sample from uniform( $\theta, \theta + 1$ ),  $\theta \in \mathbb{R}$ , and  $T = (X_{(1)}, X_{(n)})$  is the minimal sufficient statistic for  $\theta$ . We can show that  $T$  is not complete.

**Theorem 2:** Suppose that  $S$  is a minimal sufficient statistic and  $T$  is a complete and sufficient statistic. Then  $T$  must be minimal sufficient and  $S$  must be complete.

### 2.4 Statistical decision

**Convention 1 (Basic elements):**  $X$ : a sample from a population  $P \in \mathcal{P}$ . Decision: an action we take after observing  $X$ .  $\mathcal{A}$ : the set of allowable actions.  $(\mathcal{A}, \mathcal{F}_{\mathcal{A}})$ : the action space.  $\mathcal{X}$ : the range of  $X$ . Decision rule: a measurable function  $T$  from  $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$  to  $(\mathcal{A}, \mathcal{F}_{\mathcal{A}})$ . If  $X = x$  is observed, then we take the action  $T(x) \in \mathcal{A}$ .

**Definition 1 (Loss function):**  $L(P, a)$ : a function from  $\mathcal{P} \times \mathcal{A}$  to  $[0, \infty)$ .  $L(P, a)$  is Borel for each  $P$ . If  $X = x$  is observed and our decision rule is  $T$ , then our loss is  $L(P, T(x))$ .

**Definition 2 (Risk):** The averaged loss  $R_T(P) := \mathbb{E}[L(P, T(X))] = \int_{\mathcal{X}} L(P, T(X)) dP_X(x)$ .

**Definition 3 (Comparisons):** For decision rules  $T_1$  and  $T_2$ ,  $T_1$  is as good as  $T_2$  iff  $R_{T_1}(P) \leq R_{T_2}(P)$  for any  $P \in \mathcal{P}$  and is better than  $T_2$  if, in addition,  $R_{T_1}(P) < R_{T_2}(P)$  for some  $P$ .  $T_1$  and  $T_2$  are equivalent iff  $R_{T_1}(P) = R_{T_2}(P)$  for all  $P \in \mathcal{P}$ . Optimal rule: If  $T^*$  is as good as any other rule in  $\mathcal{E}$ , a class of allowable decision rules, then  $T^*$  is  $\mathcal{E}$ -optimal.

**Definition 4 (Randomized decision rules):** A function  $\delta$  on  $\mathcal{X} \times \mathcal{F}_{\mathcal{A}}$ ; for every  $A \in \mathcal{F}_{\mathcal{A}}$ ,  $\delta(\cdot, A)$  is a Borel function and, for every  $x \in \mathcal{X}$ ,  $\delta(x, \cdot)$  is a probability measure on  $(\mathcal{A}, \mathcal{F}_{\mathcal{A}})$ . If  $X = x$  is observed, we have a distribution of actions:  $\delta(x, \cdot)$ . A nonrandomized rule  $T$  is a special randomized decision rule with  $\delta(x, \{a\}) = I_{\{a\}}(T(x))$ ,  $a \in \mathcal{A}$ ,  $x \in \mathcal{X}$ . The loss function for a randomized rule  $\delta$  is defined as  $L(P, \delta, x) = \int_{\mathcal{A}} L(P, a) d\delta(x, a)$ , which reduces to the same loss function when  $\delta$  is nonrandomized. The risk of a randomized  $\delta$  is then  $R_{\delta}(P) = \mathbb{E}[L(P, \delta, X)] = \int_{\mathcal{X}} \int_{\mathcal{A}} L(P, a) d\delta(x, a) dP_X(x)$ .

**Example 1:**  $X = (X_1, \dots, X_n)$  is a vector of i.i.d. measurements for a parameter  $\theta \in \mathbb{R}$ . We want to estimate  $\theta$ . Action space:  $(\mathcal{A}, \mathcal{F}_{\mathcal{A}}) = (\mathbb{R}, \mathcal{B})$ . A common loss function in this problem is the squared error loss  $L(P, a) = (\theta - a)^2$ ,  $a \in \mathcal{A}$ . Let  $T(X) = \bar{X}$ , the sample mean. The loss for  $\bar{X}$  is  $(\bar{X} - \theta)^2$ . If the population has mean  $\mu$  and variance  $\sigma^2 < \infty$ , then  $R_{\bar{X}}(P) = (\mu - \theta)^2 + \frac{\sigma^2}{n}$ . This problem is a special case of a general problem called estimation. In an estimation problem, a decision rule  $T$  is called an estimator.

**Example 2:** Let  $\mathcal{P}$  be a family of distributions,  $\mathcal{P}_0 \subset \mathcal{P}$ ,  $\mathcal{P}_1 = \{P \in \mathcal{P} : P \notin \mathcal{P}_0\}$ . A hypothesis testing problem can be formulated as that of deciding which of the following two statements is true:  $H_0 : P \in \mathcal{P}_0$  versus  $H_1 : P \in \mathcal{P}_1$ .  $H_0$  is called the null hypothesis and  $H_1$  is the alternative hypothesis. The action space for this problem contains only two elements, i.e.,  $\mathcal{A} = \{0, 1\}$ , where 0 is accepting  $H_0$  and 1 is rejecting  $H_0$ . This problem is a special case of a general problem called hypothesis testing. A decision rule is called a test, which must have the form  $I_C(X)$ , where  $C \in \mathcal{F}_{\mathcal{X}}$  is called the rejection or critical region.

**Definition 5 (0-1 loss):**  $L(P, a) = 0$  if a correct decision is made and 1 if an incorrect decision is made, which leads to the risk  $R_T(P) = \begin{cases} P(T(X) = 1) = P(X \in C) & P \in \mathcal{P}_0 \\ P(T(X) = 0) = P(X \notin C) & P \in \mathcal{P}_1 \end{cases}$ .

**Definition 6 (Admissibility):** Let  $\mathcal{E}$  be a class of decision rules. A decision rule  $T \in \mathcal{E}$  is called  $\mathcal{E}$ -admissible iff there does not exist any  $S \in \mathcal{E}$  that is better than  $T$  (in terms of the risk).

**Remark 1:** 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.

**Proposition 1:** Let  $T(X)$  be a sufficient statistic for  $P \in \mathcal{P}$  and let  $\delta_0$  be a decision rule. Then  $\delta_1(t, A) = \mathbb{E}[\delta_0(X, A) | T = t]$ , which is a randomized decision rule depending only on  $T$ , is equivalent to  $\delta_0$  if  $R_{\delta_0}(P) < \infty$  for any  $P \in \mathcal{P}$ .

**Theorem 1:** Suppose that  $\mathcal{A}$  is a convex subset of  $\mathbb{R}^k$  and that for any  $P \in \mathcal{P}$ ,  $L(P, a)$  is a convex function of  $a$ . (i) Let  $\delta$  be a randomized rule satisfying  $\int_{\mathcal{A}} \|a\| d\delta(x, a) < \infty$  for any  $x \in \mathcal{X}$  and let  $T_1(x) = \int_{\mathcal{A}} a d\delta(x, a)$ . Then  $L(P, T_1(x)) \leq L(P, \delta, x)$  (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  $P \in \mathcal{P}$ ,  $T_0 \in \mathbb{R}^k$  be a nonrandomized rule satisfying  $\mathbb{E}\|T_0\| < \infty$ , and  $T_1 = \mathbb{E}[T_0(X) | T]$ . Then  $R_{T_1}(P) \leq R_{T_0}(P)$  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.

**Definition 7 (Unbiasedness):** In an estimation problem, the bias of an estimator  $T(X)$  of a parameter  $\theta$  of the unknown population is defined to be  $b_T(P) = \mathbb{E}[T(X)] - \theta$ . An estimator  $T(X)$  is unbiased for  $\theta$  iff  $b_T(P) = 0$  for any  $P \in \mathcal{P}$ .

**Approach 1:** Define a class  $\mathcal{E}$  of decision rules that have some desirable properties and then try to find the best rule in  $\mathcal{E}$ .

**Approach 2:** Consider some characteristic  $R_T$  of  $R_T(P)$ , for a given decision rule  $T$ , and then minimize  $R_T$  over  $T \in \mathcal{E}$ . Methods include the Bayes rule and the minimax rule.

## 2.5 Statistical inference

**Definition 1 (Three components in statistical inference):** Point estimators, hypothesis tests, confidence sets.

**Definition 2 (Point estimators):** Let  $T(X)$  be an estimator of  $\theta \in \mathbb{R}$ . Bias:  $b_T(P) = \mathbb{E}[T(X)] - \theta$ . Mean squared error (mse):  $\text{mse}_T(P) = \mathbb{E}[T(X) - \theta]^2 = [b_T(P)]^2 + \text{Var}(T(X))$ . Bias and mse are two common criteria for the performance of point estimators, i.e., instead of considering risk functions, we use bias and mse to evaluate point estimators.

**Definition 3 (Hypothesis tests):** To test the hypotheses  $H_0 : P \in \mathcal{P}_0$  versus  $H_1 : P \in \mathcal{P}_1$ , there are two types of 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). A test  $T$ : a statistic from  $\mathcal{X}$  to  $\{0, 1\}$ .

**Theorem 1 (Probabilities of making two types of errors):** Type I error rate:  $\alpha_T(P) = P(T(X) = 1), P \in \mathcal{P}_0$ . Type II error rate:  $1 - \alpha_T(P) = P(T(X) = 0), P \in \mathcal{P}_1$ .  $\alpha_T(P)$  is also called the power function of  $T$ . Power function is  $\alpha_T(\theta)$  if  $P$  is in a parametric family indexed by  $\theta$ .

**Example 1:** Let  $X_1, \dots, X_n$  be i.i.d. from the  $\mathcal{N}(\mu, \sigma^2)$  distribution with an unknown  $\mu \in \mathbb{R}$  and a known  $\sigma^2$ . Consider the hypotheses  $H_0 : \mu \leq \mu_0$  versus  $H_1 : \mu > \mu_0$ , where  $\mu_0$  is a fixed constant. Since the sample mean  $\bar{X}$  is sufficient for  $\mu \in \mathbb{R}$ , it is reasonable to consider the following class of tests:  $T_c(X) = I_{(c, \infty)}(\bar{X})$ . By the property of the normal distributions,  $\alpha_{T_c}(\mu) = P(T_c(X) = 1) = 1 - \phi(\frac{\sqrt{n}(c-\mu)}{\sigma})$ . Since  $\phi(t)$  is an increasing function of  $t$ ,  $\sup_{P \in \mathcal{P}_0} \alpha_{T_c}(\mu) = 1 - \phi(\frac{\sqrt{n}(c-\mu_0)}{\sigma})$ . In fact, it is also true for  $\sup_{P \in \mathcal{P}_1} [1 - \alpha_{T_c}(\mu)] = \phi(\frac{\sqrt{n}(c-\mu_0)}{\sigma})$ . If we would like to use an  $\alpha$  as the level of significance, then the most effective way is to choose a  $c_\alpha$  such that  $\alpha = \sup_{P \in \mathcal{P}_0} \alpha_{T_{c_\alpha}}(\mu)$ , in which case  $c_\alpha$  must satisfy  $1 - \phi(\frac{\sqrt{n}(c_\alpha-\mu_0)}{\sigma}) = \alpha$ , i.e.,  $c_\alpha = \sigma z_{1-\alpha} / \sqrt{n} + \mu_0$ , where  $z_a = \Phi^{-1}(a)$ . It can be shown that for any test  $T(X)$  satisfying  $\sup_{P \in \mathcal{P}_0} \alpha_T(P) \leq \alpha$ ,  $1 - \alpha_T(\mu) \geq 1 - \alpha_{T_{c_\alpha}}(\mu), \mu > \mu_0$ .

**Definition 4 (Significance tests):** A common approach of finding an “optimal” test is to assign a small bound  $\alpha$  to the type I error rate  $\alpha_T(P), P \in \mathcal{P}_0$ , and then to attempt to minimize the type II error rate  $1 - \alpha_T(P), P \in \mathcal{P}_1$ , subject to  $\sup_{P \in \mathcal{P}_0} \alpha_T(P) \leq \alpha$ . The bound  $\alpha$  is called the level of significance. The left-hand side is called the size of the test  $T$ . The level of significance should be positive, otherwise no test satisfies.

**Definition 5 (p-value):** It is good practice to determine not only whether  $H_0$  is rejected for a given  $\alpha$  and a chosen test  $T_\alpha$ , 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  $T_\alpha$ .



**Example 2:** Let us calculate the  $p$ -value for  $T_{c_\alpha}$  in Example 1. Note that  $\alpha = 1 - \phi(\frac{\sqrt{n}(c_\alpha - \mu_0)}{\sigma}) > 1 - \Phi(\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma})$  if and only if  $\bar{X} > c_\alpha$  (or  $T_{c_\alpha}(x) = 1$ ). Hence,  $1 - \phi(\frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma}) = \inf\{\alpha \in (0, 1) : T_{c_\alpha}(x) = 1\} = \hat{\alpha}(X)$  is the  $p$ -value for  $T_{c_\alpha}$ . It turns out that  $T_{c_\alpha}(x) = I_{(0, \alpha)}(\hat{\alpha}(X))$ .

**Definition 6 (Confidence sets)**  $\theta$ : a  $k$ -vector of unknown parameters related to the unknown  $P \in \mathcal{P}$ . If a Borel set  $C(X)$  (in the range of  $\theta$ ) depending only on the sample  $X$  such that  $\inf_{P \in \mathcal{P}} P(\theta \in C(X)) \geq 1 - \alpha$ , where  $\alpha$  is a fixed constant in  $(0, 1)$ , then  $C(X)$  is called a confidence set for  $\theta$  with level of significance  $1 - \alpha$ . The left-hand side 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  $\theta$  with certain probability.

**Example 3:** Let  $X_1, \dots, X_n$  be i.i.d. from the  $\mathcal{N}(\mu, \sigma^2)$  distribution with both  $\mu \in \mathbb{R}$  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, we focus on  $C(X)$  that is a function of  $(\bar{X}, S^2)$ . Since  $\sqrt{n}(\bar{X} - \mu)/\sigma$  has the  $\mathcal{N}(0, 1)$  distribution,  $P(-\tilde{c}_\alpha \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq \tilde{c}_\alpha) = \sqrt{1 - \alpha}$ , where  $\tilde{c}_\alpha = \Phi^{-1}(\frac{1 + \sqrt{1 - \alpha}}{2})$ . Since the  $\chi^2$  distribution  $\chi_{n-1}^2$  is a known distribution, we can always find two constants  $c_{1\alpha}$  and  $c_{2\alpha}$  such that  $P(c_{1\alpha} \leq \frac{(n-1)S^2}{\sigma^2} \leq c_{2\alpha}) = \sqrt{1 - \alpha}$ . Then  $P(-\tilde{c}_\alpha \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq \tilde{c}_\alpha, c_{1\alpha} \leq \frac{(n-1)S^2}{\sigma^2} \leq c_{2\alpha}) = 1 - \alpha$ . The LHS 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  $\sigma^2 = n(\bar{X} - \mu)^2/\tilde{c}_\alpha^2$ . This set is a confidence set for  $\theta$  with confidence coefficient  $1 - \alpha$ .

**Definition 7 (Randomized tests):** 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  $\alpha_T(P) = \mathbb{E}[T(X)], P \in \mathcal{P}_0$ , and the type II error probability to be  $1 - \alpha_T(P) = \mathbb{E}[1 - T(X)], P \in \mathcal{P}_1$ . For a class of randomized tests, we would like to minimize  $1 - \alpha_T(P)$  subject to  $\sup_{P \in \mathcal{P}_0} \alpha_T(P) = \alpha$ .

**Definition 8 (Consistency of point estimators):** Let  $X = (X_1, \dots, X_n)$  be a sample from  $P \in \mathcal{P}$ ,  $T_n(X)$  be an estimator of  $\theta$  for every  $n$ , and  $\{a_n\}$  be a sequence of positive constants,  $a_n \rightarrow \infty$ . (i)  $T_n(x)$  is consistent for  $\theta$  iff  $T_n(x) \rightarrow_p \theta$  w.r.t. any  $P$ . (ii)  $T_n(x)$  is  $a_n$ -consistent for  $\theta$  iff  $a_n[T_n(X) - \theta] = O_p(1)$  w.r.t. any  $P$ . (iii)  $T_n(x)$  is strongly consistent for  $\theta$  iff  $T_n(x) \rightarrow_{a.s.} \theta$  w.r.t. any  $P$ . (iv)  $T_n(X)$  is  $L_r$ -consistent for  $\theta$  iff  $T_n(x) \rightarrow_{L_r} \theta$  w.r.t. for any  $P$  for some fixed  $r > 0$ ; if  $r = 2$ ,  $L_2$ -consistency is called consistency in mse.

**Remark 1 (Consistency is an essential requirement):** Like the admissibility, consistency is an essential requirement: any inconsistent estimators should not be used, but there are many consistent estimators and some may not be good. Thus, consistency should be used together with other criteria.

**Remark 2 (Approximate and asymptotic bias):** Unbiasedness is a criterion for point estimator. In some cases, however, there is no unbiased estimator. Furthermore, having a “slight” bias in some cases may not be a bad idea.

**Definition 9:** (i) Let  $\xi, \xi_1, \xi_2, \dots$  be random variables and  $\{a_n\}$  be a sequence of positive numbers satisfying  $a_n \rightarrow \infty$  or  $a_n \rightarrow a > 0$ . If  $a_n \xi_n \rightarrow_d \xi$  and  $\mathbb{E}|\xi| < \infty$ , then  $\mathbb{E}\xi/a_n$  is called an asymptotic expectation of  $\xi_n$ . (ii) For a point estimator  $T_n$  of  $\theta$ , an asymptotic expectation of  $T_n - \theta$ , if it exists,



is called an asymptotic bias of  $T_n$  and denoted by  $\tilde{b}_{T_n}(P)$ . If  $\lim_{n \rightarrow \infty} \tilde{b}_{T_n}(P) = 0$  for any  $P$ , then  $T_n$  is asymptotically unbiased.

**Proposition 1 (Asymptotic expectation is essentially unique):** For a sequence of random variables  $\{\xi_n\}$ , suppose both  $\mathbb{E}\xi/a_n$  and  $\mathbb{E}\eta/b_n$  are asymptotic expectations of  $\xi_n$ . Then, one of the following three must hold: (a)  $\mathbb{E}\xi = \mathbb{E}\eta = 0$ ; (b)  $\mathbb{E}\xi \neq 0, \mathbb{E}\eta = 0$ , and  $b_n/a_n \rightarrow 0$ ; (c)  $\mathbb{E}\xi \neq 0, \mathbb{E}\eta \neq 0$ , and  $(\mathbb{E}\xi/a_n)/(\mathbb{E}\eta/b_n) \rightarrow 1$ .

**Example 4 (Functions of sample means):** We consider the case where  $X_1, \dots, X_n$  are i.i.d. random  $k$ -vectors with finite  $\Sigma = \text{Var}(X_1)$ ,  $T_n = g(\bar{X})$ , where  $g$  is a function on  $\mathbb{R}^k$  that is second-order differentiable at  $\mu = \mathbb{E}X_1$ . Consider  $T_n$  as an estimator of  $\theta = g(\mu)$ . By Taylor's expansion,  $T_n - \theta = [\nabla g(\mu)]^T(\bar{X} - \mu) + 2^{-1}(\bar{X} - \mu)^T \nabla^2 g(\mu)(\bar{X} - \mu) + o_p(n^{-1})$ . By the CLT,  $2^{-1}n(\bar{X} - \mu) \nabla^2 g(\mu)(\bar{X} - \mu) \rightarrow_d 2^{-1}Z_\Sigma^T \nabla^2 g(\mu) Z_\Sigma$ , where  $Z_\Sigma = \mathcal{N}_k(0, \Sigma)$ . Thus,  $\frac{\mathbb{E}[Z_\Sigma^T \nabla^2 g(\mu) Z_\Sigma]}{2n} = \frac{\text{tr}(\nabla^2 g(\mu) \Sigma)}{2n}$  is the  $n^{-1}$  order asymptotic bias of  $T_n = g(\bar{X})$ .

**Definition 10 (Asymptotic variance and amse):** Let  $T_n$  be an estimator of  $\theta$  for every  $n$  and  $\{a_n\}$  be a sequence of positive numbers satisfying  $a_n \rightarrow \infty$  or  $a_n \rightarrow a > 0$ . Assume that  $a_n(T_n - \theta) \rightarrow_d Y$  with  $0 < \mathbb{E}Y^2 < \infty$ . (i) The asymptotic mean squared error of  $T_n$ , denoted by  $\text{amse}_{T_n}(P)$ , is defined as the asymptotic expectation of  $(T_n - \theta)^2$ ,  $\text{amse}_{T_n}(P) = \mathbb{E}Y^2/a_n^2$ . The asymptotic variance of  $T_n$  is defined as  $\sigma_{T_n}^2(P) = \text{Var}(Y)/a_n^2$ . (ii) Let  $T'_n$  be another estimator of  $\theta$ . The asymptotic relative efficiency of  $T'_n$  w.r.t.  $T_n$  is defined as  $e_{T'_n, T_n} = \text{amse}_{T_n}(P)/\text{amse}_{T'_n}(P)$ . (iii)  $T_n$  is said to be asymptotically more efficient than  $T'_n$  iff  $\limsup_n e_{T'_n, T_n}(P) \leq 1$  for any  $P$  and  $< 1$  for some  $P$ .

**Proposition 2:** Let  $T_n$  be an estimator of  $\theta$  for every  $n$  and  $\{a_n\}$  be a sequence of positive numbers satisfying  $a_n \rightarrow \infty$  or  $a_n \rightarrow a > 0$ . If  $a_n(T_n - \theta) \rightarrow_d Y$  with  $0 < \mathbb{E}Y^2 < \infty$ , then (i)  $\mathbb{E}Y^2 \leq \liminf_n \mathbb{E}[a_n^2(T_n - \theta)^2]$  and (ii)  $\mathbb{E}Y^2 = \lim_{n \rightarrow \infty} \mathbb{E}[a_n^2(T_n - \theta)^2]$  if and only if  $\{a_n^2(T_n - \theta)^2\}$  is uniformly integrable.

**Example 5:** Let  $X_1, \dots, X_n$  be i.i.d. from the Poisson distribution  $P(\theta)$  with an unknown  $\theta > 0$ . Consider the estimation of  $\theta = P(X_i = 0) = e^{-\theta}$ . Let  $T_{1n} = F_n(0)$ , where  $F_n$  is the empirical c.d.f. Then  $T_{1n}$  is unbiased and has  $\text{mse}_{T_{1n}}(\theta) = e^{-\theta}(1 - e^{-\theta})/n$ . Also,  $\sqrt{n}(T_{1n} - \theta) \rightarrow_d \mathcal{N}(0, e^{-\theta}(1 - e^{-\theta}))$  by the CLT. Thus, in the case  $\text{amse}_{T_{1n}}(\theta) = \text{mse}_{T_{1n}}(\theta)$ . Consider  $T_{2n} = e^{-\bar{X}}$ . Note that  $\mathbb{E}T_{2n} = e^{n\theta(e^{-1/n} - 1)}$ , hence  $nb_{T_{2n}}(\theta) \rightarrow \theta e^{-\theta}/2$ . Using the CLT, we can show that  $\sqrt{n}(T_{2n} - \theta) \rightarrow_d \mathcal{N}(0, e^{-2\theta}\theta)$ . Then  $\text{amse}_{T_{2n}}(\theta) = e^{-2\theta}\theta/n$ . Thus, the asymptotic relative efficiency of  $T_{1n}$  w.r.t.  $T_{2n}$  is  $e_{T_{1n}, T_{2n}} = \theta/(e^\theta - 1) < 1$ . This shows that  $T_{2n}$  is asymptotically more efficient than  $T_{1n}$ .

### 3 Unbiased Estimation

#### 3.1 UMVUE: functions of sufficient and complete statistics

**Definition 1 (Estimable):** If there exists an unbiased estimator of  $\vartheta$ , then  $\vartheta$  is called an estimable parameter.

**Definition 2 (UMVUE):** An unbiased estimator  $T(X)$  of  $\theta$  is called uniformly minimum variance unbiased estimator (UMVUE) iff  $\text{Var}(T(X)) \leq \text{Var}(U(X))$  for any  $P \in \mathcal{P}$  and any other unbiased estimator  $U(X)$  of  $\theta$ .

**Theorem 1 (Lehmann-Scheffé theorem):** Suppose that there exists a sufficient and complete

statistic  $T(X)$  for  $P \in \mathcal{P}$ . If  $\theta$  is estimable, i.e., there is a unique unbiased estimator of  $\theta$ , then there is a unique UMVUE of  $\theta$  that is of the form  $h(T)$  with a Borel function  $h$ .

**The first method (Directly solving for  $h$ ):** Need the distribution of  $T$ . Try some function  $h$  to see if  $\mathbb{E}[h(T)]$  is related to  $\theta$ . If  $\mathbb{E}[h(T)] = \theta$  for all  $P$ , what should  $h$  be?

**Example 1:** Let  $X_1, \dots, X_n$  be i.i.d. from the uniform distribution on  $(0, \theta)$ ,  $\theta > 0$ . Consider  $\vartheta = \theta$ . Since the sufficient and complete statistic  $X_{(n)}$  has the Lebesgue p.d.f.  $n\theta^{-n}x^{n-1}1_{(0,\theta)}(x)$ ,  $\mathbb{E}X_{(n)} = n\theta^{-n} \int_0^\theta x^n dx = \frac{n}{n+1}\theta$ . An unbiased estimator of  $\theta$  is  $(n+1)X_{(n)}/n$ , which is the UMVUE. Consider now  $\vartheta = g(\theta)$ , where  $g$  is a differentiable function on  $(0, \theta)$ . An unbiased estimator  $h(X_{(n)})$  of  $\vartheta$  must satisfy  $\theta^n g(\theta) = n \int_0^\theta h(x)x^{n-1}dx$  for all  $\theta > 0$ . Hence, the UMVUE of  $\vartheta$  is  $h(X_{(n)}) = g(X_{(n)}) + n^{-1}X_{(n)}g'(X_{(n)})$ .

**The second method (When a sufficient and complete statistic is available):** Find an unbiased estimator of  $\theta$ , say  $U(X)$ . Conditioning on a sufficient and complete statistic  $T(X)$ :  $\mathbb{E}[U(X)|T]$  is the UMVUE of  $\theta$ . We need to derive an explicit form of  $\mathbb{E}[U(X)|T]$ .

**Example 2:** Let  $X_1, \dots, X_n$  be i.i.d. from the exponential distribution  $\text{Exp}(0, \theta)$ .  $F_\theta(x) = (1 - e^{-x/\theta})1_{(0,\theta)}(x)$ . Consider the estimation of  $\vartheta = 1 - F_\theta(t)$ .  $\bar{X}$  is sufficient and complete for  $\theta > 0$ .  $1_{(t,\infty)}(X_1)$  is unbiased for  $\vartheta$ ,  $\mathbb{E}[1_{(t,\infty)}(X_1)] = P(X_1 > t) = \vartheta$ . Hence  $T(X) = \mathbb{E}[1_{(t,\infty)}(X_1)|\bar{X}] = P(X_1 > t|\bar{X})$  is the UMVUE of  $\vartheta$ . By Basu's theorem,  $X_1/\bar{X}$  and  $\bar{X}$  are independent. Thus,  $P(X_1 > t|\bar{X} = \bar{x}) = P(X_1/\bar{X} > t/\bar{x}|\bar{X} = \bar{x}) = P(X_1/\bar{X} > t/\bar{x})$ . To compute this unconditional probability, we need the distribution of  $X_1/\sum_{i=1}^n X_i = X_1/(X_1 + \sum_{i=2}^n X_i)$ . Using the transformation technique and the fact that  $\sum_{i=2}^n X_i$  is independent of  $X_1$  and has a gamma distribution, we obtain that  $X_1/\sum_{i=1}^n X_i$  has the Lebesgue p.d.f.  $(n-1)(1-x)^{n-2}1_{(0,1)}(x)$ . Hence  $P(X_1 > t|\bar{X} = \bar{x}) = (n-1) \int_{t/(n\bar{x})}^1 (1-x)^{n-2}dx = (1 - \frac{t}{n\bar{x}})^{n-1}$  and the UMVUE of  $\vartheta$  is  $T(X) = (1 - \frac{t}{n\bar{X}})^{n-1}$ .

**Example 3:** Let  $X_1, \dots, X_n$  be i.i.d. from an unknown population  $P$  in a nonparametric family  $\mathcal{P}$ . In many cases the vector of order statistics,  $T = (X_{(1)}, \dots, X_{(n)})$ , is sufficient and complete for  $P \in \mathcal{P}$ . Note that an estimator  $\phi(X_1, \dots, X_n)$  is a function of  $T$  iff the function  $\phi$  is symmetric in its  $n$  arguments. Hence, if  $T$  is sufficient and complete, then a symmetric unbiased estimator of any estimable  $\vartheta$  is the UMVUE. Specific examples:  $\bar{X}$  is the UMVUE of  $\vartheta = \mathbb{E}X_1$ ,  $S^2$  is the UMVUE of  $\text{Var}(X_1)$ ,  $n^{-1} \sum_{i=1}^n X_i^2 - S^2$  is the UMVUE of  $(\mathbb{E}X_1)^2$ ,  $F_n(t)$  is the UMVUE of  $P(X_1 \leq t)$  for any fixed  $t$ . The previous conclusions are not true if  $T$  is not sufficient and complete for  $P \in \mathcal{P}$ .

**Remark 1 (Nonexistence of any UMVUE):** If  $n > 2$  and  $\mathcal{P}$  contains all symmetric distributions having Lebesgue p.d.f.'s and finite means, then there is no UMVUE for  $\mu = \mathbb{E}X_1$ .

**Example 4 (Survey samples from a finite population):** Let  $\mathcal{P} = \{1, \dots, N\}$  be a finite population of interest. For each  $i \in \mathcal{P}$ , let  $y_i$  be a value of interest associated with unit  $i$ . Let  $s = \{i_1, \dots, i_n\}$  be a subset of distinct elements of  $\mathcal{P}$ , which is a sample selected with selection probability  $p(s)$ , where  $p$  is known. The value  $y_i$  is observed if and only if  $i \in s$ . If  $p(s)$  is constant, the sampling plan is called the simple random sampling without replacement. Consider the estimation of  $Y = \sum_{i=1}^N y_i$ , the population total as the parameter of interest. Let  $X = (X_i, i \in s)$  be the vector such that  $P(X_1 = y_{i_1}, \dots, X_n = y_{i_n}) = p(s)/n!$ . Let  $\mathcal{Y}$  be the range of  $y_i$ ,  $\theta = (y_1, \dots, y_N)$  and  $\Theta = \prod_{i=1}^N \mathcal{Y}$ . Under simple random sampling without replacement, the population under consideration is a parametric family indexed by  $\theta \in \Theta$ .

**Theorem 2 (Watson-Royall theorem):** (i) If  $p(s) > 0$  for all  $s$ , then the vector of order statistics  $X_{(1)} \leq \dots \leq X_{(n)}$  is complete for  $\theta \in \Theta$ . (ii) Under simple random sampling without replacement, the vector of order statistics is sufficient for  $\theta \in \Theta$ . (iii) Under simple random sampling without replacement, for any estimable function of  $\theta$ , its unique UMVUE is the unbiased estimator  $g(X_1, \dots, X_n)$ , where  $g$  is symmetric in its  $n$  arguments.

### 3.2 Characteristic of UMVUE and Fisher information bound

**Remark 1:** When a complete and sufficient statistic is not available, it is usually very difficult to derive a UMVUE. In some cases, the following result can be applied, if we have enough knowledge about unbiased estimators of 0.

**Theorem 1:** Let  $\mathcal{U}$  be the set of all unbiased estimators of 0 with finite variances and  $T$  be an unbiased estimator of  $\theta$  with  $\mathbb{E}(T^2) < \infty$ . (i) A necessary and sufficient condition for  $T(X)$  to be a UMVUE of  $\theta$  is that  $\mathbb{E}[T(X)U(X)] = 0$  for any  $U \in \mathcal{U}$  and any  $P \in \mathcal{P}$ . (ii) Suppose that  $T = h(\tilde{T})$ , where  $\tilde{T}$  is a sufficient statistic for  $P \in \mathcal{P}$  and  $h$  is a Borel function. Let  $\mathcal{U}_{\tilde{T}}$  be the subset of  $\mathcal{U}$  consisting of Borel functions of  $\tilde{T}$ . Then a necessary and sufficient condition for  $T$  to be a UMVUE of  $\theta$  is that  $\mathbb{E}[T(X)U(X)] = 0$  for any  $U \in \mathcal{U}_{\tilde{T}}$  and any  $P \in \mathcal{P}$ . The theorem can be used to find a UMVUE, check whether a particular estimator is a UMVUE and show the nonexistence of any UMVUE.

**Theorem 2:** (i) If  $T_j$  is a UMVUE of  $\theta_j, j = 1, \dots, k$ , then  $\sum_{j=1}^k c_j T_j$  is a UMVUE of  $\theta = \sum_{j=1}^k c_j \theta_j$  for any constants  $c_1, \dots, c_k$ . (ii) If  $T_1$  and  $T_2$  are two UMVUE's of  $\theta$ , then  $T_1 = T_2$  a.s.  $P$  for any  $P \in \mathcal{P}$ .

**Example 1:** Let  $X_1, \dots, X_n$  be i.i.d. from the uniform distribution on the interval  $(0, \theta)$ . We have shown that  $(1+n^{-1})X_{(n)}$  is the UMVUE for  $\theta$  when the parameter space is  $\Theta = (0, \infty)$ . Suppose now that  $\Theta = [1, \infty)$ . Then  $X_{(n)}$  is not complete, although it is still sufficient for  $\theta$ . We now illustrate how to use Theorem 1 to find a UMVUE of  $\theta$ . Let  $U(X_{(n)})$  be an unbiased estimator of 0. Since  $X_{(n)}$  has the Lebesgue p.d.f  $n\theta^{-n}x^{n-1}1_{(0,\theta)}(x)$ ,  $0 = \int_0^1 U(x)x^{n-1}dx + \int_1^\theta U(x)x^{n-1}dx$  for all  $\theta \geq 1$ . This implies that  $U(x) = 0$  a.e. Lebesgue measure on  $[1, \infty)$  and  $\int_0^1 U(x)x^{n-1}dx = 0$ . Consider  $T = h(X_{(n)})$ . To have  $\mathbb{E}(TU) = 0$ , we must have  $\int_0^1 h(x)U(x)x^{n-1}dx = 0$ . Thus, we may consider the

following function:  $h(x) = \begin{cases} c & 0 \leq x \leq 1 \\ bx & x > 1 \end{cases}$ , where  $c$  and  $b$  are some constants. Since  $\mathbb{E}[h(X_{(n)})] = \theta$ ,

we obtain that  $\theta = cP(X_{(n)} \leq 1) + b\mathbb{E}[X_{(n)}1_{(1,\infty)}(X_{(n)})] = c\theta^{-n} + \frac{bn}{n+1}(\theta - \theta^{-n})$ . Thus,  $c = 1$  and

$b = (n+1)/n$ . The UMVUE of  $\theta$  is then  $h(X_{(n)}) = \begin{cases} 1 & 0 \leq X_{(n)} \leq 1 \\ (1+n^{-1})X_{(n)} & X_{(n)} > 1 \end{cases}$ .

**Theorem 3 (Cramér-Rao lower bound):** Let  $X = (X_1, \dots, X_n)$  be a sample from  $P \in \mathcal{P} = \{P_\theta : \theta \in \Theta\}$ , where  $\Theta$  is an open set in  $\mathbb{R}^k$ . Suppose that  $T(X)$  is an estimator with  $\mathbb{E}[T(X)] = g(\theta)$  being a differentiable function of  $\theta$ ;  $P_\theta$  has a p.d.f.  $f_\theta$  w.r.t. a measure  $\nu$  for all  $\theta \in \Theta$ ; and  $f_\theta$  is differentiable as a function of  $\theta$  and satisfies  $\frac{\partial}{\partial \theta} \int h(x)f_\theta(x)d\nu = \int h(x)\frac{\partial}{\partial \theta} f_\theta(x)d\nu, \theta \in \Theta$  for  $h(x) \equiv 1$  and  $h(x) = T(x)$ . Then  $\text{Var}(T(X)) \geq [\frac{\partial}{\partial \theta} g(\theta)]^T [I(\theta)]^{-1} \frac{\partial}{\partial \theta} g(\theta)$ , where  $I(\theta) = \mathbb{E}\{\frac{\partial}{\partial \theta} \log f_\theta(X) [\frac{\partial}{\partial \theta} \log f_\theta(X)]^T\}$  is assumed to be positive definite for any  $\theta \in \Theta$  and is called the Fisher information matrix.

**Proposition 1:** (i) If  $X$  and  $Y$  are independent with the Fisher information matrices  $I_X(\theta)$  and  $I_Y(\theta)$ , respectively, then the Fisher information about  $\theta$  contained in  $(X, Y)$  is  $I_X(\theta) + I_Y(\theta)$ . (ii) Suppose that  $X$  has the p.d.f.  $f_\theta$  that is twice differentiable in  $\theta$  and  $\frac{\partial}{\partial \theta} \int h(x) f_\theta(x) d\nu = \int h(x) \frac{\partial}{\partial \theta} f_\theta(x) d\nu$  holds with  $h(x) \equiv 1$  and  $f_\theta$  replaced by  $\partial f_\theta / \partial \theta$ . Then  $I(\theta) = -\mathbb{E}[\frac{\partial^2}{\partial \theta \partial \theta^T} \log f_\theta(X)]$ .

**Remark 2:** If  $\theta = \psi(\eta)$  and  $\psi$  is differentiable, then the Fisher information that  $X$  contains about  $\eta$  is  $\frac{\partial}{\partial \eta} \psi(\eta) I(\psi(\eta)) [\frac{\partial}{\partial \eta} \psi(\eta)]^T$ . However, the Cramér-Rao lower bound is not affected by any one-to-one reparameterization.

**Proposition 2:** Suppose that the distribution of  $X$  is from an exponential family  $\{f_\theta : \theta \in \Theta\}$ , i.e., the p.d.f. of  $X$  w.r.t. a  $\sigma$ -finite measure is  $f_\theta(x) = \exp\{\eta(\theta)^T T(X) - \xi(\theta)\} c(x)$ , where  $\Theta$  is an open subset of  $\mathbb{R}^k$ . (i) The regularity condition  $\frac{\partial}{\partial \theta} \int h(x) f_\theta(x) d\nu = \int h(x) \frac{\partial}{\partial \theta} f_\theta(x) d\nu$  is satisfied for any  $h$  with  $\mathbb{E}|h(X)| < \infty$  and  $I(\theta) = -\mathbb{E}[\frac{\partial^2}{\partial \theta \partial \theta^T} \log f_\theta(X)]$ . (ii) If  $I(\eta)$  is the Fisher information matrix for the natural parameter  $\eta$ , then the variance-covariance matrix  $\text{Var}(T) = I(\eta)$ . (iii) If  $I(\theta)$  is the Fisher information matrix for the parameter  $\vartheta = \mathbb{E}[T(X)]$ , then  $\text{Var}(T) = [I(\vartheta)]^{-1}$ .

### 3.3 U- and V-statistics

**Definition 1 (U-statistics):** Let  $X_1, \dots, X_n$  be i.i.d. from an unknown population  $P$  in a non-parametric family  $\mathcal{P}$ . If the vector of order statistic is sufficient and complete for  $P \in \mathcal{P}$ , then a symmetric unbiased estimator of an estimable  $\theta$  is the UMVUE of  $\theta$ . In many problems, parameters to be estimated are of the form  $\theta = \mathbb{E}[h(X_1, \dots, X_m)]$  with a positive integer  $m$  and a Borel function  $h$  that is symmetric and satisfies  $\mathbb{E}|h(X_1, \dots, X_m)| < \infty$  for any  $P \in \mathcal{P}$ . An effective way of obtaining an unbiased estimator of  $\theta$  is to use  $U_n = (C_n^m)^{-1} \sum_c h(X_{i_1}, \dots, X_{i_m})$ , where  $\sum_c$  denotes the summation over the  $C_n^m$  combinations of  $m$  distinct elements  $\{i_1, \dots, i_m\}$  from  $\{1, \dots, n\}$ . The statistic is called a U-statistic with kernel  $h$  of order  $m$ .

**Example 1:** Consider the estimation of  $\mu^m$ , where  $\mu = \mathbb{E}X_1$  and  $m$  is an integer  $> 0$ . Using  $h(x_1, \dots, x_m) = x_1 \cdots x_m$ , we obtain the following U-statistic for  $\mu^m$ :  $U_n = (C_n^m)^{-1} \sum_c X_{i_1} \cdots X_{i_m}$ . Consider next the estimation of  $\sigma^2 = \mathbb{E}[(X_1 - X_2)^2/2]$ , we obtain the following U-statistic with kernel  $h(x_1, x_2) = (x_1 - x_2)^2/2$ :  $U_n = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \frac{(X_i - X_j)^2}{2} = \frac{1}{n-1} (\sum_{i=1}^n X_i^2 - n\bar{X}^2) = S^2$ , which is the sample variance.

**Theorem 1 (Hoeffding's theorem):** For a U-statistic  $U_n$  with  $\mathbb{E}[h(X_1, \dots, X_m)]^2 < \infty$ ,  $\text{Var}(U_n) = (C_n^m)^{-1} \sum_{k=1}^m C_m^k C_{n-m}^{m-k} \zeta_k$ , where  $\zeta_k = \text{Var}(h_k(X_1, \dots, X_k))$ ,  $h_k(x_1, \dots, x_k) = \mathbb{E}[h(X_1, \dots, X_m) | X_1 = x_1, \dots, X_k = x_k] = \mathbb{E}[h(x_1, \dots, x_k, X_{k+1}, \dots, X_m)]$ ,  $\tilde{h}_k = h_k - \mathbb{E}[h(X_1, \dots, X_m)]$ .

**Proposition 1:** (i)  $\frac{m^2}{n} \zeta_1 \leq \text{Var}(U_n) \leq \frac{m}{n} \zeta_m$ ; (ii)  $(n+1)\text{Var}(U_{n+1}) \leq n\text{Var}(U_n)$  for any  $n > m$ ; (iii) For any fixed  $m$  and  $k = 1, \dots, m$ , if  $\zeta_j = 0$  for  $j < k$  and  $\zeta_k > 0$ , then  $\text{Var}(U_n) = \frac{k!(C_m^k)^2 \zeta_k}{n^k} + O(\frac{1}{n^{k+1}})$ .

**Example 2:** Consider  $h(x_1, x_2) = x_1 x_2$ , the U-statistic unbiased for  $\mu^2$ ,  $\mu = \mathbb{E}X_1$ . Note that  $h_1(x_1) = \mu x_1$ ,  $\tilde{h}_1(x_1) = \mu(x_1 - \mu)$ .  $\zeta_1 = \mathbb{E}[\tilde{h}_1(X_1)]^2 = \mu^2 \text{Var}(X_1) = \mu^2 \sigma^2$ ,  $\tilde{h}(x_1, x_2) = x_1 x_2 - \mu^2$ , and  $\zeta_2 = \text{Var}(X_1 X_2) = (\mu^2 + \sigma^2)^2 - \mu^4$ . Thus for  $U_n = (C_n^2)^{-1} \sum_{1 \leq i < j \leq n} X_i X_j$ ,  $\text{Var}(U_n) = (C_n^2)^{-1} (C_2^1 C_{n-2}^1 \zeta_1 + C_2^2 C_{n-2}^0 \zeta_2) = \frac{2}{n(n-1)} [2(n-2)\mu^2 \sigma^2 + (\mu^2 + \sigma^2)^2 - \mu^4] = \frac{4\mu^2 \sigma^2}{n} + \frac{2\sigma^4}{n(n-1)}$ .

**Remark 1 (Asymptotic distributions of U-statistics):** For nonparametric  $\mathcal{P}$ , the exact distribution of  $U_n$  is hard to derive. We study the method of projection, which is particularly effective for studying asymptotic distributions of U-statistics.

**Definition 2:** Let  $T_n$  be a given statistic based on  $X_1, \dots, X_n$ . The projection of  $T_n$  on  $k_n$  random elements  $Y_1, \dots, Y_{k_n}$  is defined to be  $\tilde{T}_n = \mathbb{E}(T_n) + \sum_{i=1}^{k_n} [\mathbb{E}(T_n|Y_i) - \mathbb{E}(T_n)]$ .

**Theorem 2:** Let  $T_n$  be a symmetric statistics with  $\text{Var}(T_n) < \infty$  for every  $n$  and  $\tilde{T}_n$  be the projection of  $T_n$  on  $X_1, \dots, X_n$ . Then  $\mathbb{E}(T_n) = \mathbb{E}(\tilde{T}_n)$  and  $\mathbb{E}(T_n - \tilde{T}_n)^2 = \text{Var}(T_n) - \text{Var}(\tilde{T}_n)$ .

**Example 3:** For a U-statistic  $U_n$ , one can show that  $\tilde{U}_n = \mathbb{E}(U_n) + \frac{m}{n} \sum_{i=1}^n \tilde{h}_1(X_i)$ , where  $\tilde{U}_n$  is the projection of  $U_n$  on  $X_1, \dots, X_n$  and  $\tilde{h}_1(x) = h_1(x) - \mathbb{E}[h(X_1, \dots, X_m)]$ ,  $h_1(x) = \mathbb{E}[h(x, X_2, \dots, X_m)]$ . Hence, if  $\zeta_1 = \text{Var}(\tilde{h}_1(X_i)) > 0$ ,  $\text{Var}(\tilde{U}_n) = m^2 \zeta_1 / n$  and  $\mathbb{E}(U_n - \tilde{U}_n)^2 = O(n^{-2})$ . If  $\zeta_1 = 0$  but  $\zeta_2 > 0$ , then we can show that  $\mathbb{E}(U_n - \tilde{U}_n)^2 = O(n^{-3})$ . One may derive results for the cases where  $\zeta_2 = 0$ , but the case of either  $\zeta_1 > 0$  or  $\zeta_2 > 0$  is the most interesting case in applications.

**Theorem 3:** Let  $U_n$  be a U-statistic with  $\mathbb{E}[h(X_1, \dots, X_m)]^2 < \infty$ . (i) If  $\zeta_1 > 0$ , then  $\sqrt{n}[U_n - \mathbb{E}(U_n)] \rightarrow_d \mathcal{N}(0, m^2 \zeta_1)$ . (ii) If  $\zeta_1 = 0$  but  $\zeta_2 > 0$ , then  $n[U_n - \mathbb{E}(U_n)] \rightarrow_d \frac{m(m-1)}{2} \sum_{j=1}^{\infty} \lambda_j (\chi_{1j}^2 - 1)$ , where  $\chi_{1j}^2$ 's are i.i.d. random variables having the chi-square distribution  $\chi_1^2$  and  $\lambda_j$ 's are some constants (which may depend on  $P$ ) satisfying  $\sum_{j=1}^{\infty} \lambda_j^2 = \zeta_2$ .

**Proposition 2:**  $\mathbb{E}[\frac{m(m-1)}{2} \sum_{j=1}^{\infty} \lambda_j (\chi_{1j}^2 - 1)]^2 = \frac{m^2(m-1)^2}{2} \zeta_2$ .

**Definition 3 (V-statistics):** Let  $X_1, \dots, X_n$  be i.i.d. from  $P$ . For every U-statistic  $U_n$  as an estimator  $\theta = \mathbb{E}[h(X_1, \dots, X_m)]$ , there is a closely related V-statistic defined by  $V_n = \frac{1}{n^m} \sum_{i_1=1}^n \dots \sum_{i_m=1}^n h(X_{i_1}, \dots, X_{i_m})$ . As an estimator of  $\theta$ ,  $V_n$  is biased; but the bias is small asymptotically. For a fixed  $n$ ,  $V_n$  may be better than  $U_n$  in terms of the mse.

**Proposition 3:** (i) Assume that  $\mathbb{E}|h(X_{i_1}, \dots, X_{i_m})| < \infty$  for all  $1 \leq i_1 \leq \dots \leq i_m \leq m$ . Then the bias of  $V_n$  satisfies  $b_{V_n}(P) = O(n^{-1})$ . (ii) Assume that  $\mathbb{E}[h(X_{i_1}, \dots, X_{i_m})]^2 < \infty$  for all  $1 \leq i_1 \leq \dots \leq i_m \leq m$ . Then the variance of  $V_n$  satisfies  $\text{Var}(V_n) = \text{Var}(U_n) + O(n^{-2})$ .

**Theorem 4:** Let  $V_n$  be a V-statistic with  $\mathbb{E}[h(X_{i_1}, \dots, X_{i_m})]^2 < \infty$  for all  $1 \leq i_1 \leq \dots \leq i_m \leq m$ . (i) If  $\zeta_1 = \text{Var}(h_1(X_1)) > 0$ , then  $\sqrt{n}(V_n - \theta) \rightarrow_d \mathcal{N}(0, m^2 \zeta_1)$ . (ii) If  $\zeta_1 = 0$  but  $\zeta_2 = \text{Var}(h_2(X_1, X_2)) > 0$ , then  $n(V_n - \theta) \rightarrow_d \frac{m(m-1)}{2} \sum_{j=1}^{\infty} \lambda_j \chi_{1j}^2$ .

### 3.4 Construction of unbiased or approximately unbiased estimators and method of moments

**Definition 1 (Survey samples from a finite population):** Let  $\mathcal{P} = \{1, \dots, N\}$  be a finite population of interest. For each  $i \in \mathcal{P}$ , let  $y_i$  be a value of interest associated with unit  $i$ . Let  $s = \{i_1, \dots, i_n\}$  be a subset of distinct elements of  $\mathcal{P}$ , which is a sample selected with selection probability  $p(s)$ , where  $p$  is known. The value  $y_i$  is observed iff  $i \in s$ .  $Y = \sum_{j=1}^N y_j$  is the unknown population total of interest. Define  $\pi_i$  = probability that  $i \in s, i = 1, \dots, N$ .

**Theorem 1:** (i) (Horvitz-Thompson). If  $\pi_i > 0$  for  $i = 1, \dots, N$  and  $\pi_i$  is known when  $i \in s$ , then  $\hat{Y}_{ht} = \sum_{i \in s} y_i / \pi_i$  is an unbiased estimator of the population total  $Y$ . (ii) Define  $\pi_{ij}$  = probability that  $i \in s$  and  $j \in s, i = 1, \dots, N, j = 1, \dots, N$ . Then  $\text{Var}(\hat{Y}_{ht}) = \sum_{i=1}^N \sum_{j=i+1}^N (\pi_i \pi_j - \pi_{ij}) (\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j})^2$ .

**Remark 1 (Deriving asymptotically unbiased estimators):** An exactly unbiased estimator may not exist, or is hard to obtain. We often derive asymptotically unbiased estimators. Functions of sample means are popular estimators.

**Remark 2 (Functions of unbiased estimators):** If the parameter to be estimated is  $\vartheta = g(\theta)$  with a vector-valued parameter  $\theta$  and  $U_n$  is a vector of unbiased estimators of components of  $\theta$ ,

then  $T_n = g(U_n)$  is often asymptotically unbiased for  $\vartheta$ . Note that  $\mathbb{E}(T_n) = \mathbb{E}g(U_n)$  may not exist. Assume that  $g$  is differentiable and  $c_n(U_n - \theta) \rightarrow_d Y$ . Then  $\text{amse}_{T_n}(P) = \mathbb{E}\{[\nabla g(\theta)]^T Y\}^2 / c_n^2$ . Hence,  $T_n$  has a good performance in terms of amse if  $U_n$  is optimal in terms of mse.

**Definition 2 (Method of moments):** Consider a parametric problem where  $X_1, \dots, X_n$  are i.i.d. random variables from  $P_\theta, \theta \in \Theta \subset \mathbb{R}^k$ , and  $\mathbb{E}|X_1|^k < \infty$ . Let  $\mu_j = \mathbb{E}X_1^j$  be the  $j$ th moment of  $P$  and let  $\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^n X_i^j$  be the  $j$ th sample moment, which is an unbiased estimator of  $\mu_j, j = 1, \dots, k$ . Typically,  $\mu_j = h_j(\theta), j = 1, \dots, k$ , for some functions  $h_j$  on  $\mathbb{R}^k$ . By substituting  $\mu_j$ 's on the left-hand side by the sample moments  $\hat{\mu}_j$ , we obtain a moment estimator  $\hat{\theta}$ , i.e.  $\hat{\theta}$  satisfies  $\hat{\mu}_j = h_j(\hat{\theta}), j = 1, \dots, k$ . This method of deriving estimators is called the method of moments.

**Example 1:** Let  $X_1, \dots, X_n$  be i.i.d. from a population  $P_\theta$  indexed by the parameter  $\theta = (\mu, \sigma^2)$ , where  $\mu = \mathbb{E}X_1 \in \mathbb{R}$  and  $\sigma^2 = \text{Var}(X_1) \in (0, \infty)$ . Since  $\mathbb{E}X_1 = \mu$  and  $\mathbb{E}X_1^2 = \sigma^2 + \mu^2$ , setting  $\hat{\mu}_1 = \mu$  and  $\hat{\mu}_2 = \sigma^2 + \mu^2$  we obtain the moment estimator  $\hat{\theta} = (\bar{X}, \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2)$ .

## 4 Estimation in Parametric Models

### 4.1 Bayesian approach

**Definition 1 (Bayesian approach):**  $X$  is from a population in a parametric family  $\mathcal{P} = P_\theta : \theta \in \Theta$ , where  $\theta \in \mathbb{R}^k$  for a fixed integer  $k \geq 1$ .  $\theta$  is viewed as a realization of a random vector  $\theta \in \Theta$  whose prior distribution is  $\Pi$ . Prior distribution: past experience, past data, or a statistician's belief (subjective). Sample  $X \in \mathcal{X}$ : from  $P_\theta = P_{x|\theta}$ , the conditional distribution of  $X$  given  $\theta$ . Posterior distribution: updated prior distribution using observed  $X = x$ .

**Theorem 1 (Bayes formula):** Assume  $\mathcal{P} = \{P_{x|\theta} : \theta \in \Theta\}$  is dominated by a  $\sigma$ -finite measure  $\nu$  and  $f_\theta(x) = dP_{x|\theta}/d\nu$  is a Borel function on  $(\mathcal{X} \times \Theta, \sigma(\mathcal{B}_\mathcal{X} \times \mathcal{B}_\Theta))$ . Let  $\Pi$  be a prior distribution on  $\Theta$ . Suppose that  $m(x) = \int_\Theta f_\theta(x) d\Pi > 0$ . (i) The posterior distribution  $P_{\theta|x} \ll \Pi$  and  $dP_{\theta|x}/d\Pi = f_\theta(x)/m(x)$ . (ii) If  $\Pi \ll \lambda$  and  $d\Pi/d\lambda = \pi(\theta)$  for a  $\sigma$ -finite measure  $\lambda$ , then  $dP_{\theta|x}/d\lambda = f_\theta(x)\pi(\theta)/m(x)$ .

**Definition 2 (Bayes action):** Let  $\mathcal{A}$  be an action space in a decision problem and  $L(\theta, a) \geq 0$  be a loss function. For any  $x \in \mathcal{X}$ , a Bayes action w.r.t.  $\Pi$  is any  $\delta(x) \in \mathcal{A}$  such that  $\mathbb{E}[L(\theta, \delta(x))|X = x] = \min_{a \in \mathcal{A}} \mathbb{E}[L(\theta, a)|X = x]$  where the expectation is w.r.t. the posterior distribution  $P_{\theta|x}$ .

**Definition 3 (Conjugate prior):** An interesting phenomenon is that the prior and the posterior are in the same parametric family of distributions. Such a prior is called a conjugate prior.

**Definition 4 (Generalized Bayes action):** The minimization in Definition 4.1 is the same as the minimizing  $\int_\Theta L(\theta, \delta(x)) f_\theta(x) d\Pi = \min_{a \in \mathcal{A}} \int_\Theta L(\theta, a) f_\theta(x) d\Pi$ . This is still defined even if  $\Pi$  is not a probability measure but a  $\sigma$ -finite measure on  $\Theta$ , in which case  $m(x)$  may not be finite. If  $\Pi(\Theta) \neq 1$ ,  $\Pi$  is called an improper prior.  $\delta(x)$  is called a generalized Bayes action.

**Definition 5 (Hyperparameters and empirical Bayes):** A Bayes action depends on the chosen prior with a vector  $\xi$  of parameters called hyperparameters. If the hyperparameters  $\xi$  is unknown, one way to solve the problem is to estimate  $\xi$  using some historical data; the resulting Bayes action is called an empirical Bayes action. If there is no historical data, we may estimate  $\xi$  using data  $x$  and the resulting Bayes action is also called an empirical Bayes action. The simplest empirical Bayes method is to



estimate  $\xi$  by viewing  $x$  as a “sample” from the marginal distribution  $P_{x|\xi}(A) = \int_{\Theta} P_{x|\theta}(A) d\Pi_{\theta|\xi}$ ,  $A \in \mathcal{B}_X$ , where  $\Pi_{\theta|\xi}$  is a prior depending on  $\xi$  or from the marginal p.d.f.  $m(x) = \int_{\Theta} f_{\theta}(x) d\Pi$ , if  $P_{x|\theta}$  has a p.d.f.  $f_{\theta}$ . The method of moments can be applied to estimate  $\xi$ .

**Example 1:** Let  $X = (X_1, \dots, X_n)$  and  $X_i$ 's be i.i.d. with an unknown mean  $\mu \in \mathbb{R}$  and a known variance  $\sigma^2$ . Assume the prior  $\Pi_{\mu|\xi}$  has mean  $\mu_0$  and variance  $\sigma_0^2$ ,  $\xi = (\mu_0, \sigma_0^2)$ . To obtain a moment estimate of  $\xi$ , we need to calculate  $\int_{\mathbb{R}^n} x_1 m(x) dx$  and  $\int_{\mathbb{R}^n} x_1^2 m(x) dx$ ,  $x = (x_1, \dots, x_n)$ . These two integrals can be obtained without knowing  $m(x)$ . Note that  $\int_{\mathbb{R}^n} x_1 m(x) dx = \int_{\Theta} \int_{\mathbb{R}^n} x_1 f_{\mu}(x) dx d\Pi_{\mu|\xi} = \int_{\mathbb{R}} \mu d\Pi_{\mu|\xi} = \mu_0$  and  $\int_{\mathbb{R}^n} x_1^2 m(x) dx = \int_{\Theta} \int_{\mathbb{R}^n} x_1^2 f_{\mu}(x) dx d\Pi_{\mu|\xi} = \sigma^2 + \int_{\mathbb{R}} \mu^2 d\Pi_{\mu|\xi} = \sigma^2 + \mu_0^2 + \sigma_0^2$ . Thus, by viewing  $x_1, \dots, x_n$  as a sample from  $m(x)$ , we obtain the moment estimates  $\hat{\mu}_0 = \bar{x}$  and  $\hat{\sigma}_0^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 - \sigma^2$ , where  $\bar{x}$  is the sample mean of  $x_i$ 's.

**Definition 6 (Hierarchical Bayes):** Instead of estimating hyperparameters, in the hierarchical Bayes approach we put a prior on hyperparameters. Let  $\Pi_{\theta|\xi}$  be a prior with a hyperparameter vector  $\xi$  and let  $\Lambda$  be a prior on  $\Xi$ , the range of  $\xi$ . Then the “marginal” prior for  $\theta$  is defined by  $\Pi(B) = \int_{\Xi} \Pi_{\theta|\xi}(B) d\Lambda(\xi)$ ,  $B \in \mathcal{B}_{\Theta}$ . If the second-stage prior  $\Lambda$  also depends on some unknown hyperparameters, then one can go on to consider a third-stage prior. In most applications, however, two-stage priors are sufficient, since misspecifying a second-stage prior is much less serious than misspecifying a first-stage prior.

**Example 2:** If  $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$  with a known  $\sigma^2$ , the prior  $\pi(\mu|\xi)$  is the p.d.f. of  $\mathcal{N}(\xi, \sigma_0^2)$  with a known  $\sigma_0^2$ , and the prior of  $\xi$  is  $\mathcal{N}(\mu_0, \tau^2)$  with a known  $\mu_0$  and  $\tau^2$ , then the marginal prior p.d.f. of  $\mu$  is  $\mathcal{N}(\mu_0, \sigma_0^2 + \tau^2)$ .

## 4.2 Bayes rule and computation

**Theorem 1 (Admissibility of Bayes rule)** In a decision problem, let  $\delta(x)$  be a Bayes rule w.r.t. a prior  $\Pi$ . (i) If  $\delta(X)$  is a unique Bayes rule, then  $\delta(X)$  is admissible. (ii) If  $\Theta$  is countable set, the Bayes risk  $r_{\delta}(\Pi) < \infty$ , and  $\Pi$  gives positive probability to each  $\theta \in \Theta$ , then  $\delta(X)$  is admissible. (iii) Let  $\mathcal{E}$  be the class of decision rules having continuous risk functions. If  $\delta(X) \in \mathcal{E}$ ,  $r_{\delta}(\Pi) < \infty$ , and  $\Pi$  gives positive probability to any open subset of  $\Theta$ , then  $\delta(X)$  is  $\mathcal{E}$ -admissible.

**Theorem 2:** Suppose that  $\Theta$  is an open set of  $\mathbb{R}^k$ . In a decision problem, let  $\mathcal{E}$  be the class of decision rules having continuous risk functions. A decision rule  $T \in \mathcal{E}$  is  $\mathcal{E}$ -admissible if there exists a sequence  $\{\Pi_j\}$  of priors such that (a) the generalized Bayes risks  $r_T(\Pi_j)$  are finite for all  $j$ ; (2) for any  $\theta_0 \in \Theta$  and  $\eta > 0$ ,  $\lim_{j \rightarrow \infty} \frac{r_T(\Pi_j) - r_j^*(\Pi_j)}{\Pi_j(O_{\theta_0, \eta})} = 0$ , where  $r_j^*(\Pi_j) = \inf_{T \in \mathcal{E}} r_T(\Pi_j)$  and  $O_{\theta_0, \eta} = \{\theta \in \Theta : \|\theta - \theta_0\| < \eta\}$  with  $\Pi_j(O_{\theta_0, \eta}) < \infty$  for all  $j$ .

**Proposition 1 (Bayes estimators are biased):** If  $\delta(X)$  is a Bayes estimator of  $\vartheta = g(\theta)$  under the squared error loss, then  $\delta(X)$  is not unbiased except in the trivial case where  $r_{\delta}(\Pi) = 0$ .

**Theorem 3:** Suppose that  $X$  has a p.d.f.  $f_{\theta}(x)$  w.r.t. a  $\sigma$ -finite measure  $\nu$ . Suppose that  $\theta = (\theta_1, \theta_2)$ ,  $\theta_j \in \Theta_j$ , and that the prior has a p.d.f.  $\pi(\theta) = \pi_{\theta_1|\theta_2}(\theta_1) \pi_{\theta_2}(\theta_2)$  where  $\pi_{\theta_2}(\theta_2)$  is a p.d.f. w.r.t. a  $\sigma$ -finite measure  $\nu_2$  on  $\Theta_2$  and for any given  $\theta_2$ ,  $\pi_{\theta_1|\theta_2}(\theta_1)$  is a p.d.f. w.r.t. a  $\sigma$ -finite measure  $\nu_1$  on  $\Theta_1$ . Suppose further that if  $\theta_2$  is given, the Bayes estimator of  $h(\theta_1) = g(\theta_1, \theta_2)$  under the squared error loss is  $\delta(X, \theta_2)$ . Then the Bayse estimator of  $g(\theta_1, \theta_2)$  under the squared error loss is  $\delta(X)$  with  $\delta(x) = \int_{\Theta_2} \delta(x, \theta_2) p_{\theta_2|x}(\theta_2) d\nu_2$  where  $p_{\theta_2|x}(\theta_2)$  is the posterior p.d.f. of  $\theta_2$  given  $X = x$ .



**Remark 1:** Often, Bayes actions or estimators have to be computed numerically. Typically we need to compute  $\mathbb{E}_p(g) = \int_{\Theta} g(\theta)p(\theta)d\nu$  with some function  $g$ , where  $p(\theta)$  is a p.d.f. w.r.t. a  $\sigma$ -finite measure  $\nu$  on  $(\Theta, \mathcal{B}_{\Theta})$  and  $\Theta \subset \mathbb{R}^k$ . There are many numerical methods for computing integrals  $\mathbb{E}_p(g)$ .

**Definition 1 (The simple Monte Carlo method):** Generate i.i.d.  $\theta^{(1)}, \dots, \theta^{(m)}$  from a p.d.f.  $h(\theta) > 0$  w.r.t.  $\nu$ . By the SLLN, as  $m \rightarrow \infty$ ,  $\hat{\mathbb{E}}_p(g) = \frac{1}{m} \sum_{j=1}^m \frac{g(\theta^{(j)})p(\theta^{(j)})}{h(\theta^{(j)})} \rightarrow_{\text{a.s.}} \int_{\Theta} \frac{g(\theta)p(\theta)}{h(\theta)} h(\theta)d\nu = \mathbb{E}_p(g)$ .

**Remark 2:** The simple Monte Carlo method may not work well because (i) the convergence of  $\hat{\mathbb{E}}_p(g)$  is very slow when  $k$  (the dimension of  $\Theta$ ) is large; (ii) generating a random vector from some  $k$ -dimensional distribution may be difficult, if not impossible.

**Remark 3 (More sophisticated MCMC methods):** Different from the simple Monte Carlo in two aspects: (i) generating random vectors can be done using distributions whose dimensions are much lower than  $k$ ; (ii)  $\theta^{(1)}, \dots, \theta^{(m)}$  are not independent, but form a homogeneous Markov chain.

**Definition 2 (Gibbs sampler):** Let  $y = (y_1, y_2, \dots, y_d)$ .  $y_j$ 's may be vectors with different dimensions. At step  $t = 1, 2, \dots$ , given  $y^{(t-1)}$ , generate  $y_1^{(t)}$  from  $P(y_2^{(t-1)}, \dots, y_d^{(t-1)} | y_1^{(t-1)})$ ,  $\dots$ ,  $y_j^{(t)}$  from  $P(y_1^{(t)}, \dots, y_{j-1}^{(t)}, y_{j+1}^{(t-1)}, \dots, y_k^{(t-1)} | y_j^{(t-1)})$ ,  $\dots$ ,  $y_k^{(t)}$  from  $P(y_1^{(t)}, \dots, y_{k-1}^{(t)} | y_k^{(t-1)})$ .

### 4.3 Minimality and admissibility

**Definition 1 (Minimax estimator):** An estimator  $\delta$  is minimax if  $\sup_{\theta} R_{\delta}(\theta) = \inf_T \sup_{\theta} R_T(\theta)$ .

**Remark 1:** A minimax estimator can be very conservative and unsatisfactory. It tries to do as well as possible in the worst case. A unique minimax estimator is admissible, since any estimator better than a minimax estimator is also minimax.

**Theorem 1 (Minimality of a Bayes estimator):** Let  $\Pi$  be a proper prior on  $\Theta$  and  $\delta$  be a Bayes estimator of  $\theta$  w.r.t.  $\Pi$ . Suppose  $\delta$  has constant risk on  $\Theta_{\Pi}$ . If  $\Pi(\Theta_{\Pi}) = 1$ , then  $\delta$  is minimax. If, in addition,  $\delta$  is the unique Bayes estimator w.r.t.  $\Pi$ , then it is the unique minimax estimator.

**Theorem 2:** Let  $\Pi_j, j = 1, 2, \dots$  be a sequence of priors and  $r_j$  be the Bayes risk of a Bayes estimator of  $\theta$  w.r.t.  $\Pi_j$ . Let  $T$  be a constant risk estimator of  $\theta$ . If  $\liminf_j r_j \geq R_T$ , then  $T$  is minimax.

**Example 1:** Let  $X_1, \dots, X_n$  be i.i.d. components having the  $\mathcal{N}(\mu, \sigma^2)$  distribution with an known  $\mu = \theta \in \mathbb{R}$  and a known  $\sigma^2$ . If the prior is  $\mathcal{N}(\mu_0, \sigma_0^2)$ , then the posterior of  $\theta$  given  $X = x$  is  $\mathcal{N}(\mu_*(x), c^2)$  with  $\mu_*(x) = \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0 + \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \bar{X}$  and  $c^2 = \frac{\sigma_0^2}{\sigma^2} n\sigma_0^2 + \sigma^2$ . We now show that  $\bar{X}$  is minimax under the squared error loss. For any decision rule  $T$ ,  $\sup_{\theta \in \mathbb{R}} R_T(\theta) \geq \int_{\mathbb{R}} R_T(\theta) d\Pi(\theta) \geq \int_{\mathbb{R}} R_{\mu_*}(\theta) d\Pi(\theta) = \mathbb{E}\{[\theta - \mu_*(X)]^2\} = \mathbb{E}\{\mathbb{E}\{[\theta - \mu_*(X)]^2 | X\}\} = \mathbb{E}(c^2) = c^2$ . Since this result is true for any  $\sigma_0^2 > 0$  and  $c^2 \rightarrow \sigma^2/n$  as  $\sigma_0^2 \rightarrow \infty$ ,  $\sup_{\theta \in \mathbb{R}} R_T(\theta) \geq \frac{\sigma^2}{n} = \sup_{\theta \in \mathbb{R}} R_{\bar{X}}(\theta)$  where the equality holds because the risk of  $\bar{X}$  under the squared error loss is  $\sigma^2/n$  and independent of  $\theta = \mu$ . Thus,  $\bar{X}$  is minimax.

**Theorem 3:** Let  $\Theta_0$  be a subset of  $\Theta$  and  $T$  be a minimax estimator of  $\theta$  when  $\Theta_0$  is the parameter space. Then  $T$  is minimax estimator if  $\sup_{\theta \in \Theta} R_T(\theta) = \sup_{\theta \in \Theta_0} R_T(\theta)$ .

**Theorem 4 (Admissibility in one-parameter exponential families):** Suppose that  $X$  has the p.d.f.  $c(\theta)e^{\theta T(x)}$  w.r.t. a  $\sigma$ -finite measure  $\nu$ , where  $T(x)$  is real-valued and  $\theta \in (\theta_-, \theta_+) \subset \mathbb{R}$ . Consider the estimation of  $\theta = \mathbb{E}[T(X)]$  under the squared error loss. Let  $\lambda \geq 0$  and  $\gamma$  be known constants

and let  $T_{\lambda,\gamma}(X) = (T + \gamma\lambda)/(1 + \lambda)$ . Then a sufficient condition for the admissibility of  $T_{\lambda,\gamma}$  is that  $\int_{\theta_0}^{\theta_+} \frac{e^{-\gamma\lambda\theta}}{[c(\theta)]^\lambda} d\theta = \int_{\theta_-}^{\theta_0} \frac{e^{-\gamma\lambda\theta}}{[c(\theta)]^\lambda} d\theta = \infty$ , where  $\theta_0 \in (\theta_-, \theta_+)$ .

**Theorem 5:** Assume that  $X$  has the p.d.f. as described in Theorem 4 with  $\theta_- = -\infty$  and  $\theta_+ = \infty$ . (i) As an estimator of  $\theta = \mathbb{E}(T)$ ,  $T(X)$  is admissible under the squared error loss and the loss  $(a - \theta)^2/\text{Var}(T)$ . (ii)  $Y$  is the unique minimax estimator of  $\theta$  under the loss  $(a - \theta)^2/\text{Var}(T)$ .

**Example 2:** Let  $X_1, \dots, X_n$  be i.i.d. from  $\mathcal{N}(0, \sigma^2)$  with an unknown  $\sigma^2 > 0$  and let  $Y = \sum_{i=1}^n X_i^2$ . Consider the estimation of  $\sigma^2$ . The risk of  $Y/(n+2)$  is a constant under the loss  $(a - \sigma^2)^2/\sigma^4$ . We now apply Theorem 4 to show that  $Y/(n+2)$  is admissible. Note that the joint p.d.f. of  $X_i$ 's is of the form  $c(\theta)e^{\theta T(x)}$  with  $\theta = -n/(4\sigma^2)$ ,  $c(\theta) = (-2\theta/n)^{n/2}$ ,  $T(X) = 2Y/n$ ,  $\theta_- = -\infty$  and  $\theta_+ = 0$ . By Theorem 4,  $T_{\lambda,\gamma} = (T + \gamma\lambda)/(1 + \lambda)$  is admissible under the squared error loss if, for some  $c > 0$ ,  $\int_{-\infty}^{-c} e^{-\gamma\lambda\theta} \left(\frac{-2\theta}{n}\right)^{-n\lambda/2} d\theta = \int_0^c e^{\gamma\lambda\theta} \theta^{-n\lambda/2} d\theta = \infty$ . This means  $T_{\lambda,\gamma}$  is admissible if  $\gamma = 0$  and  $\lambda = 2/n$ , or if  $\gamma > 0$  and  $\lambda \geq 2/n$ . In particular,  $2Y/(n+2)$  is admissible for estimating  $\mathbb{E}(T) = 2\mathbb{E}(Y)/n = 2\sigma^2$ , under the squared error loss. It is easy to see that  $Y/(n+2)$  is then an admissible estimator of  $\sigma^2$  under the squared error loss and the loss  $(a - \sigma^2)^2/\sigma^4$ . Hence  $Y/(n+2)$  is minimax under the loss  $(a - \sigma^2)^2/\sigma^4$ .

#### 4.4 Simultaneous estimation and shrinkage estimators

**Definition 1 (Simultaneous estimation):** Estimation of a  $p$ -vector  $\vartheta$  of parameters (functions of  $\theta$ ) under the decision theory approach.

**Remark 1 (Difference from estimating  $\vartheta$  component-by-component):** A single loss function  $L(\vartheta, a)$ , instead of  $p$  loss functions.

**Definition 2 (Squared error loss):** A natural generalization of the squared error loss is  $L(\theta, a) = \|a - \theta\|^2 = \sum_{i=1}^p (a_i - \theta_i)^2$ .

**Definition 3 (James-Stein estimator):** We start with the simple case where  $X$  is from  $\mathcal{N}_p(\theta, I_p)$  with an unknown  $\theta \in \mathbb{R}^p$ . James and Stein proposed the following class of estimators of  $\theta$  having smaller risks than  $X$  when the squared error loss is used and  $p \geq 3$ :  $\delta_c = X - \frac{p-2}{\|X-c\|^2}(X-c)$ , where  $c \in \mathbb{R}^p$  is fixed and the choice of  $c$  is discussed later.

**Definition 4 (Extended James-Stein estimators):** For the purpose of generalizing the results to more complicated situations, we consider the following extension of the James-Stein estimator:  $\delta_{c,r} = X - \frac{r(p-2)}{\|X-c\|^2}(X-c)$ , where  $c \in \mathbb{R}^p$  and  $r \in \mathbb{R}$  are known.

**Motivation 1 (Shrink the observation toward a given point  $c$ ):** Suppose it were thought a priori likely, though not certain, that  $\theta = c$ . Then we might first test a hypothesis  $H_0 : \theta = c$  and estimate  $\theta$  by  $c$  if  $H_0$  is accepted and by  $X$  otherwise. The best rejection region has the form  $\|X - c\|^2 > t$  for some constant  $t > 0$  so that we might estimate  $\theta$  by  $I_{(t,\infty)}(\|X - c\|^2)X + [1 - I_{(t,\infty)}(\|X - c\|^2)]c$ .  $\delta_{c,r}$  is a smoothed version of this estimator, since, for some function  $\psi$ ,  $\delta_{c,r} = \psi(\|X - c\|^2)X + [1 - \psi(\|X - c\|^2)]c$ . Any estimator having this form is called a shrinkage estimator.

**Motivation 2 (Empirical Bayes estimator):** A Bayes estimator of  $\theta$  is of the form  $\delta = (1 - B)X + Bc$ , where  $c$  is the prior mean of  $\theta$  and  $B$  involves prior variances.  $1 - B$  is “estimated” by  $\psi(\|X - c\|^2)$ .  $\delta_{c,r}$  can be viewed as an empirical Bayes estimator.

**Theorem 1 (Risks of shrinkage estimators):** Suppose that  $X$  is from  $\mathcal{N}_p(\theta, I_p)$  with  $p \geq 3$ . Then,

under the squared error loss, the risks of the following shrinkage estimators of  $\theta$ ,  $\delta_{c,r} = X - \frac{r(p-2)}{\|X-c\|^2}(X-c)$ , where  $c \in \mathbb{R}^p$  and  $r \in \mathbb{R}$  are known, are given by  $R_{\delta_{c,r}}(\theta) = p - (2r - r^2)(p-2)^2\mathbb{E}(\|X-c\|^{-2})$ .

**Remark 2:** The risk of  $\delta_{c,r}$  is smaller than  $p$ , the risk of  $X$  for every value of  $\theta$  when  $p \geq 3$  and  $0 < r < 2$ .  $\delta = \delta_{c,1}$  is better than any  $\delta_{c,r}$  with  $r \neq 1$ .

**Remark 3 (The improvement):** To see that  $\delta_c$  may have a substantial improvement over  $X$  in terms of risks, consider the special case where  $\theta = c$ . Since  $\|X-c\|^2$  has the chi-square distribution  $\chi_p^2$  when  $\theta = c$ ,  $\mathbb{E}\|X-c\|^{-2} = (p-2)^{-1}$  and  $R_{\delta_{c,1}}(\theta) = p - (2r - r^2)(p-1)^2\mathbb{E}(\|X-c\|^{-2}) = 2$ . The ratio  $R_X(\theta)/R_{\delta_c}(\theta)$  equals  $p/2$  when  $\theta = c$  and can be substantially larger than 1 near  $\theta = c$  when  $p$  is large.

**Remark 4 (Minimaxity and admissibility of  $\delta_c$ ):** Since  $X$  is minimax,  $\delta_{c,r}$  is minimax provided that  $p \geq 3$  and  $0 < r < 2$ . Unfortunately, the James-Stein estimator  $\delta_c$  with any  $c$  is also inadmissible. It is dominated by  $\delta_c^+ = X - \min\{1, \frac{p-2}{\|X-c\|^2}\}(X-c)$ . This estimator, however, is still inadmissible. Although neither the James-Stein estimator  $\delta_c$  nor  $\delta_c^+$  is admissible, it is found that no substantial improvements over  $\delta_c^+$  are possible.

**Definition 5 (Extension of Theorem 1 to  $\text{Var}(X) = \sigma^2 D$ ):** Consider the case where  $\text{Var}(X) = \sigma^2 D$  with an unknown  $\sigma^2 > 0$  and a known positive definite matrix  $D$ . If  $\sigma^2$  is known, then an extended James-Stein estimator is  $\tilde{\delta}_{c,r} = X - \frac{(p-2)r\sigma^2}{\|D^{-1}(X-c)\|^2}D^{-1}(X-c)$ . Under the squared error loss, the risk of  $\tilde{\delta}_{c,r}$  is  $\sigma^2[\text{tr}(D) - (2r - r^2)(p-2)^2\mathbb{E}(\|D^{-1}(X-c)\|^{-2})]$ . When  $\sigma^2$  is unknown, we assume that there exists a statistic  $S_0^2$  such that  $S_0^2$  is independent of  $X$  and  $S_0^2/\sigma^2$  has the chi-square distribution  $\chi_m^2$ . Replacing  $r\sigma^2$  in  $\tilde{\delta}_{c,r}$  by  $\hat{\sigma}^2 = tS_0^2$  with a constant  $t > 0$  leads to the following extended James-Stein estimator:  $\tilde{\delta}_c = X - \frac{(p-2)\hat{\sigma}^2}{\|D^{-1}(X-c)\|^2}D^{-1}(X-c)$ . From the risk formula for  $\tilde{\delta}_{c,r}$  and the independence of  $\hat{\sigma}^2$  and  $X$ , the risk of  $\tilde{\delta}_c$  is  $R_{\tilde{\delta}_c}(\theta) = \sigma^2\{\text{tr}(D) - [2tm - t^2m(m+2)](p-2)^2\sigma^2\kappa(\theta)\}$ , where  $\theta = (\theta, \sigma^2)$  and  $\kappa(\theta) = \mathbb{E}(\|D^{-1}(X-c)\|^{-2})$ . Replacing  $t$  by  $1/(m+2)$  leads to  $R_{\tilde{\delta}_c}(\theta) = \sigma^2[\text{tr}(D) - m(m+2)^{-1}(p-2)^2\sigma^2\mathbb{E}(\|D^{-1}(X-c)\|^{-2})]$ , which is smaller than  $\sigma^2\text{tr}(D)$  (the risk of  $X$ ) for any fixed  $\theta, p \geq 3$ .

**Example 1:** Consider the general linear model  $X = Z\beta + \epsilon$  with  $\epsilon \sim \mathcal{N}_p(0, \sigma^2)$ ,  $p \geq 3$ , and a full rank  $Z$ . Consider the estimation of  $\theta = \beta$  under the squared error loss. The LSE  $\hat{\beta}$  is from  $\mathcal{N}(\beta, \sigma^2 D)$  with a known matrix  $D = (Z^T Z)^{-1}$ ,  $S_0^2 = \text{SSR}$  is independent of  $\hat{\beta}$ ,  $S_0^2/\sigma^2$  has the chi-square distribution  $\chi_{n-p}^2$ . Hence, from the previous discussion, the risk of the shrinkage estimator  $\hat{\beta} - \frac{(p-2)\hat{\sigma}^2}{\|Z^T Z(\hat{\beta}-c)\|^2}Z^T Z(\hat{\beta}-c)$  is smaller than that of  $\hat{\beta}$  for any  $\beta$  and  $\sigma^2$ , where  $c \in \mathbb{R}^p$  is fixed and  $\hat{\sigma}^2 = \text{SSR}/(n-p+2)$ .

**Definition 6 (Other shrinkage estimators):** From the previous discussion, the James-Stein estimators improve  $X$  substantially when we shrink the observations toward a vector  $c$  that is near  $\theta = \mathbb{E}X$ . One may consider shrinking the observations toward the mean of the observations rather than a given point; that is, one may obtain a shrinkage estimator by replacing  $c$  in  $\delta_{c,r}$  by  $\bar{X}J_p$ , where  $\bar{X} = p^{-1}\sum_{i=1}^p X_i$  and  $J_p$  is the  $p$ -vectors of ones. However, we have to replace the factor  $p-2$  in  $\delta_{c,r}$  by  $p-3$ . This leads to shrinkage estimators  $X - \frac{p-3}{\|X-\bar{X}J_p\|^2}(X-\bar{X}J_p)$  and  $X - \frac{(p-3)\hat{\sigma}^2}{\|D^{-1}(X-\bar{X}J_p)\|^2}D^{-1}(X-\bar{X}J_p)$ . These estimators are better than  $X$  (and, hence, are minimax) when  $p \geq 4$ , under the squared error loss.

**Remark 5:** The idea of shrinkage has been used in problems with high dimensions, e.g. LASSO.

## 4.5 Likelihood and maximum likelihood estimator (MLE)

**Definition 1:** Let  $X \in \mathcal{X}$  be a sample with a p.d.f.  $f_\theta$  w.r.t. a  $\sigma$ -finite measure  $\nu$ , where  $\theta \in \Theta \subset \mathbb{R}^k$ . (i) For each  $x \in \mathcal{X}$ ,  $f_\theta(x)$  considered as a function of  $\theta$  is called the likelihood function and denoted by  $l(\theta)$ . (ii) Let  $\bar{\Theta}$  be the closure of  $\Theta$ . A  $\hat{\theta} \in \bar{\Theta}$  satisfying  $l(\hat{\theta}) = \max_{\theta \in \bar{\Theta}} l(\theta)$  is called a maximum likelihood estimate (MLE) of  $\theta$ . If  $\hat{\theta}$  is a Borel function of  $X$  a.e.  $\nu$ , then  $\hat{\theta}$  is called a maximum likelihood estimator *MLE* of  $\theta$ . (iii) Let  $g$  be a Borel function from  $\Theta$  to  $\mathbb{R}^p, p \leq k$ . If  $\hat{\theta}$  is an MLE of  $\theta$ , then  $\hat{\vartheta} = g(\hat{\theta})$  is defined to be an MLE of  $\vartheta = g(\theta)$ .

**Remark 1 (Finding an MLE):** Since  $\log x$  is a strictly increasing function,  $\hat{\theta}$  is an MLE if and only if it maximizes the log-likelihood function  $\log l(\theta)$ . If  $l(\theta)$  is differentiable on  $\Theta^\circ$ , then possible candidates for MLE's are the values of  $\theta \in \Theta^\circ$  satisfying  $\frac{\partial \log l(\theta)}{\partial \theta} = 0$ , which is called the likelihood equation or log-likelihood equation.

**Example 1:** Let  $X_1, \dots, X_n$  be i.i.d. binary random variables with  $P(X_1 = 1) = p \in \Theta = (0, 1)$ . When  $(X_1, \dots, X_n) = (x_1, \dots, x_n)$  is observed, the likelihood function is  $l(p) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} = p^{n\bar{x}} (1-p)^{n(1-\bar{x})}$ , where  $\bar{x} = n^{-1} \sum_{i=1}^n x_i$ . Note that  $\bar{\Theta} = [0, 1]$  and  $\Theta^\circ = \Theta$ . The likelihood equation is  $\frac{n\bar{x}}{p} - \frac{n(1-\bar{x})}{1-p} = 0$ . If  $0 < \bar{x} < 1$ , then this equation has a unique solution  $\bar{x}$ . The second-order derivative of  $\log l(p)$  is  $-\frac{n\bar{x}}{p^2} - \frac{n(1-\bar{x})}{(1-p)^2}$ , which is always negative. Also, when  $p$  tends to 0 or 1 (the boundary of  $\Theta$ ),  $l(p) \rightarrow 0$ . Thus,  $\bar{x}$  is the unique MLE of  $p$ .

**Definition 2 (The Newton-Raphson method):** In applications, MLE's typically do not have analytic forms and some numerical methods have to be used to compute MLE's. A commonly used numerical method is the Newton-Raphson iteration method, which repeatedly computes  $\hat{\theta}^{(t+1)} = \hat{\theta}^{(t)} - [\frac{\partial^2 \log l(\theta)}{\partial \theta \partial \theta^T} |_{\theta=\hat{\theta}^{(t)}}]^{-1} \frac{\partial \log l(\theta)}{\partial \theta} |_{\theta=\hat{\theta}^{(t)}}$ ,  $t = 0, 1, \dots$ , where  $\hat{\theta}^{(0)}$  is an initial value and  $\partial^2 \log l(\theta) / \partial \theta \partial \theta^T$  is assumed of full rank for every  $\theta \in \Theta$ .

**Definition 3 (The Fisher-scoring method):** If, at each iteration, we replace  $[\frac{\partial^2 \log l(\theta)}{\partial \theta \partial \theta^T} |_{\theta=\hat{\theta}^{(t)}}]^{-1}$  by  $[E(\frac{\partial^2 \log l(\theta)}{\partial \theta \partial \theta^T}) |_{\theta=\hat{\theta}^{(t)}}]^{-1}$ , where the expectation is taken under  $P_\theta$ , then the method is known as the Fisher-scoring method.

## 4.6 Asymptotically efficient estimation

**Definition 1 (Asymptotic comparison):** Let  $\{\hat{\theta}_n\}$  be a sequence of estimators of  $\theta$  based on a sequence of samples  $\{X = (X_1, \dots, X_n), n = 1, 2, \dots\}$ . Suppose that as  $n \rightarrow \infty$ ,  $\hat{\theta}_n$  is asymptotically normal (AN) in the sense that  $[V_n(\theta)]^{-1/2}(\hat{\theta}_n - \theta) \rightarrow_d \mathcal{N}_k(0, I_k)$ , where, for each  $n$ ,  $V_n(\theta)$  is a  $k \times k$  positive definite matrix depending on  $\theta$ . If  $\theta$  is one-dimensional, then  $V_n(\theta)$  is the asymptotic variance as well as the amse of  $\hat{\theta}_n$ . When  $k > 1$ ,  $V_n(\theta)$  is called the asymptotic covariance matrix of  $\hat{\theta}_n$  and can be used as a measure of asymptotic performance of estimators. If  $\hat{\theta}_{j_n}$  is AN with asymptotic covariance matrix  $V_{j_n}(\theta)$ ,  $j = 1, 2$ , and  $V_{1n}(\theta) \leq V_{2n}(\theta)$  for all  $\theta \in \Theta$ , then  $\hat{\theta}_{1n}$  is said to be asymptotically more efficient than  $\hat{\theta}_{2n}$ .

**Theorem 1:** Let  $X_1, \dots, X_n$  be i.i.d. from a p.d.f.  $f_\theta$  w.r.t. a  $\sigma$ -finite measure  $\nu$  on  $(\mathbb{R}, \mathcal{B})$ , where  $\theta \in \Theta$  and  $\Theta$  is an open set in  $\mathbb{R}^k$ . Suppose that for every  $x$  in the range of  $X_1$ ,  $f_\theta(x)$  is twice continuously differentiable in  $\theta$  and satisfies  $\frac{\partial}{\partial \theta} \int \psi_\theta(x) d\nu = \int \frac{\partial}{\partial \theta} \psi_\theta(x) d\nu$  for  $\psi_\theta(x) = f_\theta(x)$  and  $= \partial f_\theta(x) / \partial \theta$ ; the Fisher information matrix  $I_1(\theta) = E\{\frac{\partial}{\partial \theta} \log f_\theta(X_1) [\frac{\partial}{\partial \theta} \log f_\theta(X_1)]^T\}$  is positive

definite; and for any given  $\theta \in \Theta$ , there exists a positive number  $c_\theta$  and a positive function  $h_\theta$  such that  $\mathbb{E}[h_\theta(X_1)] < \infty$  and  $\sup_{\gamma: \|\gamma - \theta\| < c_\theta} \|\frac{\partial^2 \log f_\gamma(x)}{\partial \gamma \partial \gamma^T}\| \leq h_\theta(x)$  for all  $x$  in the range of  $X_1$ , where  $\|A\| = \sqrt{\text{tr}(A^T A)}$  for any matrix  $A$ . If  $\hat{\theta}_n$  is an estimator of  $\theta$  and is AN with  $V_n(\theta) = V(\theta)/n$ , then there is a  $\Theta_0 \subset \Theta$  with Lebesgue measure 0 such that the information inequality  $V_n(\theta) \geq [I_n(\theta)]^{-1}$  holds if  $\theta \notin \Theta_0$ .

**Definition 2 (Asymptotic efficiency):** Assume that the Fisher information matrix  $I_n(\theta)$  is well defined and positive definite for every  $n$ . A sequence of estimators  $\{\hat{\theta}_n\}$  that is AN is said to be asymptotically efficient or asymptotically optimal if and only if  $V_n(\theta) = [I_n(\theta)]^{-1}$ .

**Remark 1 (Estimating a function of  $\theta$ ):** Suppose that we are interested in estimating  $\vartheta = g(\theta)$ , where  $g$  is a differentiable function from  $\Theta$  to  $\mathbb{R}^p$ ,  $1 \leq p \leq k$ . If  $\hat{\theta}_n$  is AN, then  $\hat{\vartheta}_n = g(\hat{\theta}_n)$  is asymptotically distributed as  $\mathcal{N}_p(\vartheta, [\nabla g(\theta)]^T V_n(\theta) \nabla g(\theta))$ . Thus, the information inequality becomes  $[\nabla g(\theta)]^T V_n(\theta) \nabla g(\theta) \geq [I_n(\vartheta)]^{-1}$ , where  $I_n(\vartheta)$  is the Fisher information matrix about  $\vartheta$  contained in  $X$ . If  $p = k$  and  $g$  is one-to-one, then  $[I_n(\vartheta)]^{-1} = [\nabla g(\theta)]^T [I_n(\theta)]^{-1} \nabla g(\theta)$  and, therefore,  $\hat{\theta}_n$  is asymptotically efficient if and only if  $\hat{\theta}_n$  is asymptotically efficient.

**Theorem 2:** Assume the conditions of Theorem 1. (i) Asymptotic existence and consistency. There is a sequence of estimators  $\{\hat{\theta}_n\}$  such that  $P(s_n(\hat{\theta}_n) = 0) \rightarrow 1$  and  $\hat{\theta}_n \rightarrow_p \theta$ , where  $s_n(\gamma) = \frac{\partial \log l(\gamma)}{\partial \gamma}$ . (ii) Asymptotic efficiency. Any consistent sequence  $\tilde{\theta}_n$  of RLE (root of the likelihood equation)'s is asymptotically normal and asymptotically efficient.

**Theorem 3:** Assume the conditions of Theorem 1. Let  $\pi(\gamma)$  be a prior p.d.f w.r.t. the Lebesgue measure on  $\Theta$  and  $p_n(\gamma)$  be the posterior p.d.f., given  $X_1, \dots, X_n$ ,  $n = 1, 2, \dots$ . Assume that there exists an  $n_0$  such that  $p_{n_0}(\gamma)$  is continuous and positive for all  $\gamma \in \Theta$ ,  $\int p_{n_0}(\gamma) d\gamma = 1$  and  $\int \|\gamma\| p_{n_0}(\gamma) d\gamma < \infty$ . Suppose further that, for any  $\epsilon > 0$ , there exists a  $\delta > 0$  such that  $\lim_{n \rightarrow \infty} P(\sup_{\|\gamma - \theta\| \geq \epsilon} \frac{\log l(\gamma) - \log l(\theta)}{n} > -\delta) = 0$ ,  $\lim_{n \rightarrow \infty} P(\sup_{\|\gamma - \theta\| \leq \delta} \frac{\|\nabla s_n(\gamma) - \nabla s_n(\theta)\|}{n} \geq \epsilon) = 0$ , where  $l(\gamma)$  is the likelihood function and  $s_n(\gamma)$  is the score function. (i) Let  $p_n^*(\gamma)$  be the posterior p.d.f of  $\sqrt{n}(\gamma - T_n)$ , where  $T_n = \theta + [I_n(\theta)]^{-1} s_n(\theta)$  and  $\theta$  is the true parameter value, and let  $\psi(\gamma)$  be the p.d.f. of  $\mathcal{N}_k(0, [I_1(\theta)]^{-1})$ . Then  $\int (1 + \|\gamma\|) |p_n^*(\gamma) - \psi(\gamma)| d\gamma \rightarrow_p 0$ . (ii) The Bayes estimator of  $\theta$  under the squared error loss is asymptotically efficient.

**Proposition 1:** The posterior p.d.f. is approximately normal with mean  $\theta + [I_n(\theta)]^{-1} s_n(\theta)$  and covariance matrix  $[I_n(\theta)]^{-1}$ .

**Remark 2:** The results hold regardless of the prior being used, indicating that the effect of the prior declines as  $n \rightarrow \infty$ .

## 4.7 MLE in generalized linear models (GLM) and quasi-MLE

**Definition 1 (The structure of a GLM):** The sample  $X = (X_1, \dots, X_n)$  has independent  $X_i$ 's and  $X_i$  has the p.d.f.  $\exp\{\frac{\eta_i x_i - \zeta(\eta_i)}{\phi_i}\} h(x_i, \phi_i)$ ,  $i = 1, \dots, n$ , w.r.t. a  $\sigma$ -finite measure  $\nu$ , where  $\eta_i$  and  $\phi_i$  are unknown,  $\phi_i > 0$ ,  $\eta_i \in \Xi = \{\eta : 0 < \int h(x, \phi) e^{\eta x / \phi} d\nu(x) < \infty\} \subset \mathbb{R}$  for all  $i$ ,  $\zeta$  and  $h$  are known functions, and  $\zeta''(\eta) > 0$  is assumed for all  $\eta \in \Xi$ . Note that the p.d.f. belongs to an exponential family if  $\phi_i$  is known. As a consequence,  $\mathbb{E}(X_i) = \zeta'(\eta_i)$  and  $\text{Var}(X_i) = \phi_i \zeta''(\eta_i)$ ,  $i = 1, \dots, n$ . Define  $\mu(\eta) = \zeta'(\eta)$ . It is assumed that  $\eta_i$  is related to  $Z_i$ , the  $i$ th value of a  $p$ -value of covariates, through  $\mu(\eta_i) = \beta^T Z_i$ ,  $i = 1, \dots, n$ , where  $\beta$  is a  $p$ -vector of unknown parameters and  $g$ , called a link

function, is a known one-to-one, third-order continuously differentiable function on  $\{\mu(\eta) : \eta \in \Xi^\circ\}$ . If  $\mu = g^{-1}$ , then  $\eta_i = \beta^T Z_i$  and  $g$  is called the canonical or natural link function. If  $g$  is not canonical, we assume that  $\frac{d}{d\eta}(g \circ \mu)(\eta) \neq 0$  for all  $\eta$ . In a GLM, the parameter of interest is  $\beta$ . We assume the range of  $\beta$  is  $B = \{\beta : (g \circ \mu)^{-1}(\beta^T z) \in \Xi^\circ \text{ for all } z \in \mathcal{Z}\}$ , where  $\mathcal{Z}$  is the range of  $Z_i$ 's.  $\phi_i$ 's are called dispersion parameters and are considered to be nuisance parameters.

**Proposition 1 (MLE in GLM):** An MLE of  $\beta$  in a GLM is considered under assumption  $\phi_i = \phi/t_i, i = 1, \dots, n$ , with an unknown  $\phi > 0$  and known positive  $t_i$ 's. Let  $\theta = (\beta, \phi)$  and  $\psi = (g \circ \mu)^{-1}$ .  $\log l(\theta) = \sum_{i=1}^n [\log h(x_i, \frac{\phi}{t_i}) + \frac{\psi(\beta^T Z_i)x_i - \zeta(\psi(\beta^T Z_i))}{\phi/t_i}]$ ,  $\frac{\partial \log l(\theta)}{\partial \beta} = \frac{1}{\phi} \sum_{i=1}^n \{[x_i - \mu(\psi(\beta^T Z_i))] \psi'(\beta^T Z_i) t_i Z_i\} = 0$ ,  $\frac{\partial \log l(\theta)}{\partial \phi} = \sum_{i=1}^n \{ \frac{\partial \log h(x_i, \phi/t_i)}{\partial \phi} - \frac{t_i [\psi(\beta^T Z_i)x_i - \zeta(\psi(\beta^T Z_i))]}{\phi^2} \} = 0$ . From the first likelihood equation, an MLE of  $\beta$ , if it exists, can be obtained without estimating  $\phi$ . The second likelihood equation, however, is usually difficult to solve. Some other estimators of  $\phi$  are suggested by various researchers. Suppose there is a solution  $\hat{\beta}$  to the likelihood equation.  $\text{Var}(\frac{\partial \log l(\theta)}{\partial \beta}) = \frac{M_n(\beta)}{\phi}$ ,  $\frac{\partial^2 \log l(\theta)}{\partial \beta \partial \beta^T} = \frac{R_n(\beta) - M_n(\beta)}{\phi}$ , where  $M_n(\beta) = \sum_{i=1}^n [\psi'(\beta^T Z_i)]^2 \zeta''(\psi(\beta^T Z_i)) t_i Z_i Z_i^T$ ,  $R_n(\beta) = \sum_{i=1}^n [x_i - \mu(\psi(\beta^T Z_i))] \psi''(\beta^T Z_i) t_i Z_i Z_i^T$ . Consider first the simple case of canonical  $g$ ,  $\psi'' = 0$  and  $R_n = 0$ . If  $M_n(\beta)$  is positive definite for all  $\beta$ , then  $-\log l(\theta)$  is strictly convex in  $\beta$  for any fixed  $\phi$  and, therefore,  $\hat{\beta}$  is the unique MLE of  $\beta$ . For noncanonical  $g$ ,  $R_n(\beta) \neq 0$  and  $\hat{\beta}$  is not necessarily an MLE. If  $R_n(\beta)$  is dominated by  $M_n(\beta)$ , i.e.,  $[M_n(\beta)]^{-1/2} R_n(\beta) [M_n(\beta)]^{-1/2} \rightarrow 0$  in some sense, then  $-\log l(\theta)$  is convex and  $\hat{\beta}$  is an MLE for large  $n$ . In a GLM, an MLE  $\hat{\beta}$  usually does not have an analytic form and a numerical method such as the Newton-Raphson has to be applied.

**Example 1:** Consider the GLM with  $\zeta(\eta) = \eta^2/2, \eta \in \mathbb{R}$ . If  $g$  is the canonical link, then the model is the same as a linear model with independent  $\epsilon_i$ 's distributed as  $\mathcal{N}(0, \phi_i)$ . Suppose now that  $g$  is noncanonical but  $\phi_i \equiv \phi$ . Then the model reduces to the one with independent  $X_i$ 's and  $X_i = \mathcal{N}(g^{-1}(\beta^T Z_i), \phi), i = 1, \dots, n$ . This type of model is called a nonlinear regression model (with normal errors) and an MLE of  $\beta$  under this model is also called a nonlinear LSE, since maximizing the log-likelihood is equivalent to minimizing the sum of squares  $\sum_{i=1}^n [X_i - g^{-1}(\beta^T Z_i)]^2$ . Under certain conditions the matrix  $R_n(\beta)$  is dominated by  $M_n(\beta)$  and an MLE of  $\beta$  exists.

**Example 2 (The Poisson model):** Consider the GLM with  $\zeta(\eta) = e^\eta, \eta \in \mathbb{R}, \phi_i = \phi/t_i$ . If  $\phi_i = 1$ , then  $X_i$  has the Poisson distribution with mean  $e^{\eta_i}$ . Under the canonical link  $g(t) = \log t$ ,  $M_n(\beta) = \sum_{i=1}^n e^{\beta^T Z_i} t_i Z_i Z_i^T$ , which is positive definite if  $\inf_i e^{\beta^T Z_i} > 0$  and the matrix  $(\sqrt{t_1} Z_1, \dots, \sqrt{t_n} Z_n)$  is of full rank. There is one noncanonical link that deserves attention. Suppose that we choose a link function so that  $[\psi'(t)]^2 \zeta''(\psi(t)) \equiv 1$ . Then  $M_n(\beta) = \sum_{i=1}^n t_i Z_i Z_i^T$  does not depend on  $\beta$ . It is shown that the asymptotic variance of the MLE  $\hat{\beta}$  is  $\phi [M_n(\beta)]^{-1}$ . The fact that  $M_n(\beta)$  does not depend on  $\beta$  makes the estimation of the asymptotic variance (and, thus, statistical inference) easy. Under the Poisson model,  $\zeta''(t) = e^t$  and, therefore, we need to solve the differentiable equation  $[\psi'(t)]^2 e^{\psi(t)} = 1$ . A solution is  $\psi(t) = 2 \log(t/2)$  and the link  $g(\mu) = 2\sqrt{\mu}$ .

**Theorem 1:** Consider the GLM with  $\phi_i = \phi/t_i$  and  $t_i$ 's in a fixed interval  $(t_0, t_\infty), 0 < t_0 \leq t_\infty < \infty$ . Assume that the range of unknown parameter  $\beta$  is an open subset of  $\mathbb{R}^p$ ; at the true value of  $\beta, 0 < \inf_i \phi(\beta^T Z_i) \leq \sup_i \phi(\beta^T Z_i) < \infty$ , where  $\phi(t) = [\psi'(t)]^2 \zeta''(\psi(t))$ ; as  $n \rightarrow \infty, \max_{i \leq n} Z_i^T (Z^T Z)^{-1} Z_i \rightarrow 0$  and  $\lambda_- [Z^T Z] \rightarrow \infty$ , where  $Z$  is the  $n \times p$  matrix whose  $i$ th row is the vector  $Z_i$  and  $\lambda_- [A]$  is the smallest eigenvalue of  $A$ . (i) There is a unique sequence of estimators  $\{\hat{\beta}_n\}$  such that  $P(s_n(\hat{\beta}_n) = 0) \rightarrow$



1 and  $\hat{\beta}_n \rightarrow_p \beta$ , where  $s_n(\beta) = \partial \log l(\beta, \phi) / \partial \phi$  is the score function. (ii) Let  $I_n(\beta) = \text{Var}(s_n(\beta))$ . Then  $[I_n(\beta)]^{1/2}(\hat{\beta}_n - \beta) \rightarrow_d \mathcal{N}_p(0, I_p)$ . (iii) If  $\phi$  is known or the p.d.f. indexed by  $\theta = (\beta, \phi)$  satisfies the conditions for  $f_\theta$  in Theorem 1 of section 4.6, then  $\hat{\beta}_n$  is asymptotically efficient.

**Definition 2 (Quasi-MLE):** If assumption  $\phi_i$  is arbitrary, or the distribution assumption on  $X_i$  does not hold, but  $\mathbb{E}(X_i) = \zeta'(\eta_i)$ ,  $\text{Var}(X_i) = \phi_i \zeta''(\eta_i)$ ,  $i = 1, \dots, n$  and  $g(\mu(\eta_i)) = \beta^T Z_i$ ,  $i = 1, \dots, n$  still hold, we estimate  $\beta$  by solving equation  $G_n(\beta) = \sum_{i=1}^n \{[x_i - \mu(\psi(\beta^T Z_i))] \psi'(\beta^T Z_i) t_i Z_i\} = 0$ , then the resulting estimator is called a quasi-MLE. This method is also called the method of generalized estimating equations (GEE). They are efficient if the GEE is a likelihood equation, and is robust if it is not.

**Remark 1:** The asymptotic existence and consistency of quasi-MLE can be shown using a similar argument to the proof of Theorem 2 of section 4.6.

## 4.8 Other asymptotically efficient estimators and pseudo MLE

**Definition 1 (One-Step MLE):** Let  $s_n(\gamma)$  be the score function. Let  $\hat{\theta}_n^{(0)}$  be an estimator of  $\theta$  that may not be asymptotically efficient. The one-step MLE is the first iteration in computing an RLE using the Newton-Raphson method with  $\hat{\theta}_n^{(0)}$  as the initial value,  $\hat{\theta}_n^{(1)} = \hat{\theta}_n^{(0)} - [\nabla s_n(\hat{\theta}_n^{(0)})]^{-1} s_n(\hat{\theta}_n^{(0)})$ . Without any further iteration,  $\hat{\theta}_n^{(1)}$  is asymptotically efficient under some conditions.

**Theorem 1:** Assume that the conditions in Theorem 1 of section 4.6 hold and that  $\hat{\theta}_n^{(0)}$  is  $\sqrt{n}$ -consistent for  $\theta$ . (i) The one-step MLE  $\hat{\theta}_n^{(1)}$  is asymptotically efficient. (ii) The one-step MLE obtained by replacing  $\nabla s_n(\gamma)$  with its expected value,  $-I_n(\gamma)$  (the Fisher-scoring method), is asymptotically efficient.

**Definition 2 (Pseudo MLE):** Let  $X_1, \dots, X_n$  be a random sample from a pdf in a family indexed by two parameters  $\theta$  and  $\pi$  with likelihood  $l(\theta, \pi)$ . The method of pseudo MLE may be viewed as follows. Based on the sample, an estimate  $\hat{\pi}$  of  $\pi$  is obtained using some technique other than MLE. The pseudo MLE of  $\theta$  is then obtained by maximizing the likelihood  $l(\theta, \hat{\pi})$ .

**Remark 1:**  $\pi$  is viewed as a nuisance parameter. Pseudo MLE consists of replacing  $\pi$  by an estimate and solving a reduced system of likelihood equations, which works when a higher dimensional MLE is intractable but a lower dimensional MLE is feasible. The consistency and asymptotic normality hold under fairly standard regularity conditions.

**Theorem 2 (Asymptotic existence and consistency of pseudo MLE):** Assume the conditions in Theorem 1 of section 4.6. Assume also  $\hat{\pi}$  is a consistent estimator of  $\pi_0$ . As  $n \rightarrow \infty$ , with probability tending to 1, there exists  $\hat{\theta}$  such that  $\frac{\partial \log l(\hat{\theta}, \hat{\pi})}{\partial \theta} = 0$  and  $\hat{\theta} \rightarrow_p \theta_0$  where  $\theta_0$  is the true value of  $\theta$ .

# 5 Estimation in Non-Parametric Models

## 5.1 Empirical c.d.f. and empirical likelihoods

**Definition 1 (Estimation in nonparametric models):** Data  $X = (X_1, \dots, X_n)$ , where  $X_i$ 's are random  $d$ -vectors i.i.d. from an unknown c.d.f.  $F$  in a nonparametric family. We study mainly two topics: estimation of the c.d.f.  $F$  and estimation of  $\theta = T(F)$ , where  $T$  is a functional.



**Definition 2 (Empirical c.d.f.):**  $F_n(t) = \frac{1}{n} \sum_{i=1}^n I_{(-\infty, t]}(X_i)$ ,  $t \in \mathbb{R}^d$ , where  $(-\infty, a]$  denotes the set  $(-\infty, a_1] \times \cdots \times (-\infty, a_d]$  for any  $a = (a_1, \dots, a_d) \in \mathbb{R}^d$ .  $F_n$  is the distribution putting mass  $n^{-1}$  at each  $X_i$ ,  $i = 1, \dots, n$ .

**Proposition 1 (Properties of empirical c.d.f.):** (i) For any  $t \in \mathbb{R}^d$ ,  $nF_n(t)$  has the binomial distribution  $B(F(t), n)$ ; (ii)  $F_n(t)$  is unbiased variance  $F(t)[1 - F(t)]/n$ ; (iii)  $F_n(t)$  is the UMVUE under some nonparametric models; (iv)  $F_n(t)$  is  $\sqrt{n}$ -consistent for  $F(t)$ .

**Theorem 1:** Define sup-norm distance  $\rho_\infty(G_1, G_2) = \|G_1 - G_2\|_\infty = \sup_{t \in \mathbb{R}^d} |G_1(t) - G_2(t)|$ ,  $G_j \in \mathcal{F}$ . (i) When  $d = 1$ , there exists a positive constant  $C$  (not depending on  $F$ ) such that  $P(\rho_\infty(F_n, F) > z) \leq Ce^{-2nz^2}$ ,  $z > 0$ ,  $n = 1, 2, \dots$ . (ii) When  $d \geq 2$ , for any  $\epsilon > 0$ , there exists a positive constant  $C_{\epsilon, d}$  (not depending on  $F$ ) such that  $P(\rho_\infty(F_n, F) > z) \leq C_{\epsilon, d}e^{-(2-\epsilon)nz^2}$ ,  $z > 0$ ,  $n = 1, 2, \dots$ .

**Theorem 2:** Let  $F_n$  be the empirical c.d.f. of i.i.d.  $X_1, \dots, X_n$  from a c.d.f.  $F$  on  $\mathbb{R}^d$ . (i)  $\rho_\infty(F_n, F) \rightarrow_{a.s.} 0$  as  $n \rightarrow \infty$ ; (ii)  $\mathbb{E}[\sqrt{n}\rho_\infty(F_n, F)]^s = O(1)$  for any  $s > 0$ .

**Theorem 3:** Let  $F_n$  be the empirical c.d.f. based on i.i.d. random variables  $X_1, \dots, X_n$  from a c.d.f  $F \in \mathcal{F}_1$ . (i)  $\rho_{L_p}(F_n, F) \rightarrow_{a.s.} 0$ ; (ii)  $\mathbb{E}[\sqrt{n}\rho_{L_p}(F_n, F)] = O(1)$  if  $1 < p < 2$  and  $\int \{F(t)[1 - F(t)]\}^{p/2} dt < \infty$  if  $p \geq 2$ .

**Theorem 4:** For  $X_1, \dots, X_n$  i.i.d. from  $F \in \mathcal{F}$ , the empirical c.d.f.  $F_n$  maximizes the nonparametric likelihood function  $l(G)$  over  $G \in \mathcal{F}$ .

**Definition 3 (Empirical likelihoods):** The nonparametric MLE can be extended to various situations with some modifications of  $l(G)$  and/or constraints on  $p_i$ 's. Modifications of the likelihood  $l(G)$  are called empirical likelihoods. An estimator obtained by maximizing an empirical likelihood is then called a maximum empirical likelihood estimator (MELE).

**Remark 1 (Estimation of  $F$  with auxiliary information about  $F$ ):** In some cases we have some information about  $F$ . For instance, suppose that there is a known Borel function  $u$  from  $\mathbb{R}^d$  to  $\mathbb{R}^s$  such that  $\int u(x)dF = 0$ . It is reasonable to expect that any estimate  $\hat{F}$  of  $F$  has property  $\int u(x)d\hat{F} = 0$ , which is not true for the empirical c.d.f  $F_n$ , since  $\int u(x)dF_n = \frac{1}{n} \sum_{i=1}^n u(X_i) \neq 0$  even if  $\mathbb{E}[u(X_1)] = 0$ . Using the method of empirical likelihoods, a natural solution is to put another constraint in the process of maximizing the likelihood. That is, we maximize  $l(G)$  subject to  $p_i > 0$ ,  $i = 1, \dots, n$ ,  $\sum_{i=1}^n p_i = 1$ , and  $\sum_{i=1}^n p_i u(x_i) = 0$  where  $p_i = P_G(\{x_i\})$ . Using the Lagrange multiplier method, it can be shown that an MELE of  $F$  is  $\hat{F}(t) = \sum_{i=1}^n \hat{p}_i I_{(-\infty, t]}(X_i)$ , where  $\hat{p}_i = n^{-1}[1 + \lambda_n^T u(X_i)]^{-1}$ ,  $i = 1, \dots, n$  and  $\lambda_n \in \mathbb{R}^s$  is the Lagrange multiplier satisfying  $\sum_{i=1}^n \hat{p}_i u(X_i) = \frac{1}{n} \sum_{i=1}^n \frac{u(X_i)}{1 + \lambda_n^T u(X_i)} = 0$ .

**Theorem 5:** Let  $u$  be a Borel function on  $\mathbb{R}^d$  satisfying  $\int u(x)dF = 0$  and  $\hat{F}$  be the MELE of  $F$ . Suppose that  $U = \text{Var}(u(X_1))$  is positive definite. Then, for any  $m$  fixed distinct  $t_1, \dots, t_m \in \mathbb{R}^d$ ,  $\sqrt{n}[(\hat{F}(t_1), \dots, \hat{F}(t_m)) - (F(t_1), \dots, F(t_m))] \rightarrow_d \mathcal{N}_m(0, \Sigma_u)$ , where  $\Sigma_u = \Sigma - W^T U^{-1} W$ ,  $\Sigma$  is the covariance matrix of  $\sqrt{n}[(F_n(t_1), \dots, F_n(t_m)) - (F(t_1), \dots, F(t_m))]$ ,  $W = (W(t_1), \dots, W(t_m))$ , and  $W(t_j) = \mathbb{E}[u(X_1)I_{(-\infty, t_j]}(X_1)]$ .

## 5.2 Profile likelihoods, GEE, and GMM

**Definition 1 (Profile likelihoods):** Let  $l(\theta, \xi)$  be a likelihood (or empirical likelihood), where  $\theta$  and  $\xi$  are not necessarily vector-valued. It may be difficult to maximize the likelihood  $l(\theta, \xi)$  simultaneously over  $\theta$  and  $\xi$ . For each fixed  $\theta$ , let  $\xi(\theta)$  satisfy  $l(\theta, \xi(\theta)) = \sup_\xi l(\theta, \xi)$ . The function

$l_p(\theta) = l(\theta, \xi(\theta))$  is called a profile likelihood function for  $\theta$ . Suppose that  $\hat{\theta}_p$  maximizes  $l_p(\theta)$ . Then  $\hat{\theta}_p$  is called a maximum profile likelihood estimator of  $\theta$ . Although this idea can be applied to parametric models, it is more useful in semi-parametric models, especially when  $\theta$  is a parametric component and  $\xi$  is a nonparametric component.

**Example 1 (Missing data):** Assume that  $X_1, \dots, X_n$  are i.i.d. random variables from an unknown c.d.f.  $F$  and some  $X_i$ 's are missing. Let  $\delta_i = 1$  if  $X_i$  is observed and  $\delta_i = 0$  if  $X_i$  is missing. Suppose that  $(X_i, \delta_i)$  are i.i.d. and let  $\pi(x) = P(\delta_i = 1 | X_i = x)$ . If  $X_i$  and  $\delta_i$  are independent, i.e.  $\pi(x) \equiv \pi$  does not depend on  $x$ , then the empirical c.d.f based on observed data, i.e., the c.d.f. putting mass  $r^{-1}$  to each observed  $X_i$ , where  $r$  is the number of observed  $X_i$ 's, is an unbiased and consistent estimator of  $F$ , provided that  $\pi > 0$ . On the other hand, if  $\pi(x)$  depends on  $x$  (called nonignorable missingness), then the empirical c.d.f. based on observed data is a biased and inconsistent estimator of  $F$ . In fact, the empirical c.d.f. based on observed data is an unbiased estimator of  $P(X_i \leq x | \delta_i = 1)$ , which is generally different from the unconditional probability  $F(x) = P(X_i \leq x)$ . If both  $\pi$  and  $F$  are in parametric models, then we can apply the method of maximum likelihood. For example, if  $\pi(x) = \pi_\theta(x)$  and  $F(x) = F_\vartheta(x)$  has a p.d.f.  $f_\vartheta$ , where  $\theta$  and  $\vartheta$  are vectors of unknown parameters, then a parametric likelihood of  $(\theta, \vartheta)$  is  $l(\theta, \vartheta) = \prod_{i=1}^n [\pi_\theta(x_i) f_\vartheta(x_i)]^{\delta_i} (1 - \pi)^{1-\delta_i}$ , where  $\pi = \int \pi_\theta(x) f_\vartheta(x) dx$ . computationally, it may be difficult to maximizing this likelihood, since  $\pi$  is an integral. Suppose now that  $\pi(x) = \pi_\theta(x)$  is the parametric component and  $F$  is the nonparametric component. Then an empirical likelihood can be defined as  $l(\theta, G) = \prod_{i=1}^n [\pi_\theta(x_i) p_i]^{\delta_i} (1 - \pi)^{1-\delta_i}$ ,  $p_i = P_G(\{x_i\})$  subject to  $p_i \geq 0$ ,  $\sum_{i=1}^n \delta_i p_i = 1$ ,  $\sum_{i=1}^n \delta_i p_i [\pi_\theta(x_i) - \pi] = 0$ ,  $i = 1, \dots, n$ . It can be shown that the logarithm of the profile empirical likelihood for  $(\theta, \pi)$  with a Lagrange multiplier is  $\sum_{i=1}^n \{\delta_i \log(\pi_\theta(x_i)) + (1 - \delta_i) \log(1 - \pi) - \delta_i \log(1 + \lambda[\pi_\theta(x_i) - \pi])\}$ . Under some conditions, it can be shown that the estimators  $\hat{\theta}$ ,  $\hat{\pi}$  and  $\hat{\lambda}$  obtained by maximizing this likelihood are consistent and asymptotically normal and that the empirical c.d.f. putting mass  $\hat{p}_i = r^{-1} \{1 + \hat{\lambda}[\pi_\theta(X_i) - \hat{\pi}]\}^{-1}$  to each observed  $X_i$  is consistent for  $F$ . The result can be extended when there is an observed covariate.

**Definition 2 (Generalized estimating equation (GEE)):** Assume that  $X_1, \dots, X_n$  are independent random vectors, where the dimension of  $X_i$  is  $d_i$ ,  $i = 1, \dots, n$  ( $\sup_i d_i < \infty$ ), and that we are interested in estimating  $\theta$ , a  $k$ -vector of unknown parameters related to the unknown population. Let  $\Theta \subset \mathbb{R}^k$  be the range of  $\theta$ ,  $\psi_i$  be a Borel function form  $\mathbb{R}^{d_i} \times \Theta$  to  $\mathbb{R}^k$ ,  $i = 1, \dots, n$ , and  $s_n(\gamma) = \sum_{i=1}^n \psi_i(X_i, \gamma)$ ,  $\gamma \in \Theta$ . If  $\theta$  is estimated by  $\hat{\theta} \in \Theta$  satisfying  $s_n(\hat{\theta}) = 0$ , then  $\hat{\theta}$  is called a GEE estimator. The equation  $s_n(\gamma) = 0$  is called a GEE.

**Remark 1 (Motivation):** Usually GEE's are chosen so that  $\mathbb{E}[s_n(\theta)] = \sum_{i=1}^n \mathbb{E}[\psi_i(X_i, \theta)] = 0$ , where the expectation  $\mathbb{E}$  may be replaced by an asymptotic expectation if the exact expectation does not exist.

**Proposition 1 (Consistency of GEE estimators):** Suppose that  $X_1, \dots, X_n$  are i.i.d. from  $F$  and  $\psi \equiv \psi$ , a bounded and continuous function form  $\mathbb{R}^d \times \Theta$  to  $\mathbb{R}^k$ . Let  $g(t) = \int \psi(x, t) dF(x)$ . Suppose that  $g(\theta) = 0$  and  $\partial g(t)/\partial t$  exists and is of full rank at  $t = \theta$ . Then  $\hat{\theta}_n \rightarrow_p \theta$ .

**Definition 3 (Generalized method of moments (GMM)):** Suppose that we have a set of  $m \geq k$  functions  $\psi_j(x, \theta)$ ,  $j = 1, \dots, m$  such that  $\mathbb{E}_\theta[\psi_j(X, \theta)] = 0$  for all  $j$  and  $\psi_j$ 's are not linearly independent, i.e., the  $m \times m$  matrix whose  $(j, j')$ th element is  $\mathbb{E}_\theta[\psi_j(X_i, \theta) \psi_{j'}(X_i, \theta)]$  is positive

## HYPOTHESIS TESTS

definite, which can usually be achieved by eliminating some redundant functions where  $\psi_j$ 's are linearly dependent. Let  $G_n(\theta) = (\frac{1}{n} \sum_{i=1}^n \psi_1(x_i, \theta), \dots, \frac{1}{n} \sum_{i=1}^n \psi_m(x_i, \theta))^T, \theta \in \Theta$ . if  $m = k$ , a solution to  $G_n(\theta) = 0$  is a GEE estimator. If  $m > k$ , a solution to  $G_n(\theta) = 0$  may not exist. Then we can minimize  $G_n^T(\theta)G_n(\theta)$ , using a data driven procedure.

**Definition 4 (GMM algorithm):** A GMM estimate of  $\theta$  can be obtained using the following two-step algorithm (the second step is to gain efficiency). (1) Obtain  $\hat{\theta}^{(1)}$  by minimizing  $G_n^T(\theta)G_n(\theta)/2$  over  $\theta \in \Theta$ . (2) Let  $\hat{W}$  be the inverse matrix of the  $m \times m$  matrix whose  $(j, j')$  element is equal to  $\frac{1}{n} \sum_{i=1}^n \psi_j(x_i, \hat{\theta}^{(1)})\psi_{j'}(x_i, \hat{\theta}^{(1)})$ . The GMM estimate  $\hat{\theta}$  is obtained by minimizing  $G_n^T(\theta)\hat{W}G_n(\theta)/2$  over  $\theta \in \Theta$ .

## 6 Hypothesis Tests

### 6.1 Neyman-Pearson lemma and monotone likelihood ratio

**Definition 1 (Theory of testing hypotheses):**  $X$ : a sample from a population  $P \in \mathcal{P}$ , a family of populations. Based on the observed  $X$ , we test a given hypothesis  $H_0 : P \in \mathcal{P}_0$  vs  $H_1 : P \in \mathcal{P}_1$  where  $\mathcal{P}_0$  and  $\mathcal{P}_1$  are two disjoint subsets of  $\mathcal{P}$  and  $\mathcal{P}_0 \cup \mathcal{P}_1 = \mathcal{P}$ . A test for a hypothesis is a statistic  $T(X)$  taking values in  $[0, 1]$ . When  $X = x$  is observed, we reject  $H_0$  with probability  $T(x)$ . If  $T(X) = 1$  or  $0$  a.s.  $\mathcal{P}$ , then  $T(X)$  is a nonrandomized test; otherwise  $T(X)$  is randomized. For a given test  $T(X)$ , the power function of  $T(X)$  is defined to be  $\beta_T(P) = \mathbb{E}[T(X)], P \in \mathcal{P}$ , which is the type I error probability of  $T(X)$  when  $P \in \mathcal{P}_0$  and one minus the type II error probability of  $T(X)$  when  $P \in \mathcal{P}_1$ .

**Definition 2 (Significance tests):** With a sample of a fixed size, we are not able to minimize two error probabilities simultaneously. Our approach involves maximizing the power  $\beta_T(P)$  over all  $P \in \mathcal{P}_1$  (i.e., minimizing the type II error probability) and over all tests  $T$  satisfying  $\sup_{P \in \mathcal{P}_0} \beta_T(P) \leq \alpha$ , where  $\alpha \in [0, 1]$  is a given level of significance. The left-hand side of the last expression is defined to be the size of  $T$ .

**Definition 3:** A test  $T_*$  of size  $\alpha$  is a uniformly most powerful (UMP) test if and only if  $\beta_{T_*}(P) \geq \beta_T(P)$  for all  $P \in \mathcal{P}_1$  and  $T$  of level  $\alpha$ .

**Proposition 1 (Using sufficient statistics):** If  $U(X)$  is a sufficient statistic for  $P \in \mathcal{P}$ , then for any test  $T(X)$ ,  $\mathbb{E}(T|U)$  has the same power function as  $T$  and, therefore, to find a UMP test we may consider tests that re functions of  $U$  only.

**Theorem 1 (Neyman-Pearson lemma):** Suppose that  $\mathcal{P}_0 = \{P_0\}$  and  $\mathcal{P}_1 = \{P_1\}$ . Let  $f_j$  be the p.d.f of  $P_j$  w.r.t. a  $\sigma$ -finite measure  $\nu$  (e.g.,  $\nu = P_0 + P_1$ ),  $j = 0, 1$ . (i) Existence of a UMP test. For

every  $\alpha$ , there exists a UMP test of size  $\alpha$ , which is  $T_*(X) = \begin{cases} 1 & f_1(X) > cf_0(X) \\ \gamma & f_1(X) = cf_0(X) \text{ where } \gamma \in [0, 1] \\ 0 & f_1(X) < cf_0(X) \end{cases}$

and  $c \geq 0$  are some constants chosen so that  $\mathbb{E}[T_*(X)] = \alpha$  when  $P = P_0$  ( $c = \infty$  is allowed). (ii)

Uniqueness. If  $T_{**}$  is a UMP test of size  $\alpha$ , then  $T_{**}(X) = \begin{cases} 1 & f_1(X) > cf_0(X) \\ 0 & f_1(X) < cf_0(X) \end{cases}$  a.s.  $\mathcal{P}$ .

## HYPOTHESIS TESTS

**Example 1:** Suppose that  $X$  is a sample of size 1,  $\mathcal{P}_0 = \{P_0\}$ , and  $\mathcal{P}_1 = \{P_1\}$ , where  $P_0$  is  $\mathcal{N}(0, 1)$  and  $P_1$  is the double exponential distribution  $\text{DE}(0, 2)$  with the p.d.f  $4^{-1}e^{-|x|/2}$ . Since  $P(f_1(X) = cf_0(X)) = 0$ , there is a unique nonrandomized UMP test. By theorem 1, the UMP test  $T_*(x) = 1$  if and only if  $\frac{\pi}{8}e^{x^2-|x|} > c^2$  for some  $c > 0$ , which is equivalent to  $|x| > t$  or  $|x| < 1 - t$  for some  $t > \frac{1}{2}$ . Suppose that  $\alpha < \frac{1}{3}$ . To determine  $t$ , we use  $\alpha = \mathbb{E}_0[T_*(X)] = P_0(|X| > t) + P_0(|X| < 1 - t)$ . If  $t \leq 1$ , then  $P_0(|X| > t) \geq P_0(|X| > 1) = 0.3374 > \alpha$ . Hence  $t$  should be larger than 1 and  $\alpha = P_0(|X| > t) = \Phi(-t) + 1 - \Phi(t)$ . Thus,  $t = \Phi^{-1}(1 - \alpha/2)$  and  $T_*(X) = I_{(t, \infty)}(|X|)$ . Note that it is not necessary to find out what  $c$  is.

**Theorem 2:** Suppose that there is a test  $T_*$  of size  $\alpha$  such that for every  $P_1 \in \mathcal{P}_1$ ,  $T_*$  is UMP for testing  $H_0$  versus the hypothesis  $P = P_1$ . Then  $T_*$  is the UMP for testing  $H_0$  versus  $H_1$ .

**Definition 4:** Suppose that the distribution of  $X$  is in  $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ , a parametric family indexed by a real-valued  $\theta$ , and that  $\mathcal{P}$  is dominated by a  $\sigma$ -finite measure  $\nu$ . Let  $f_\theta = dP_\theta/d\nu$ . The family  $\mathcal{P}$  is said to have monotone likelihood ration in  $Y(X)$  (a real-valued statistic) if and only if, for any  $\theta_1 < \theta_2$ ,  $f_{\theta_2}(x)/f_{\theta_1}(x)$  is a nondecreasing function of  $T(x)$  for values  $x$  at which at least one of  $f_{\theta_1}(x)$  and  $f_{\theta_2}(x)$  is positive.

**Example 2:** Let  $\theta$  be real-valued and  $\eta(\theta)$  be a nondecreasing function of  $\theta$ . Then the one-parameter exponential family with  $f_\theta(x) = \exp\{\eta(\theta)Y(x) - \xi(\theta)\}h(x)$  has monotone likelihood ration in  $Y(X)$ .

**Theorem 3:** Suppose that  $X$  has a distribution in  $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$  ( $\Theta \subset \mathbb{R}$ ) that has monotone likelihood ratio in  $Y(X)$ . Consider the problem of testing  $H_0 : \theta \leq \theta_0$  versus  $H_1 : \theta > \theta_0$ , where  $\theta_0$  is

a given constant. (i) There exists a UMP test of size  $\alpha$ , which is given by  $T_*(X) = \begin{cases} 1 & Y(X) > c \\ \gamma & Y(X) = c \\ 0 & Y(X) < c \end{cases}$

where  $c$  and  $\gamma$  are determined by  $\beta_{T_*}(\theta_0) = \alpha$ , and  $\beta_T(\theta) = \mathbb{E}[T(X)]$  is the power function of a test  $T$ . (ii)  $\beta_{T_*}(\theta)$  is strictly increasing for all  $\theta$ 's for which  $0 < \beta_{T_*}(\theta) < 1$ . (iii) For any  $\theta < \theta_0$ ,  $T_*$  minimizes  $\beta_T(\theta)$  among all tests  $T$  satisfying  $\beta_T(\theta_0) = \alpha$ . (iv) Assume that  $P_\theta(f_\theta(X) = cf_{\theta_0}(X)) = 0$  for any  $\theta > \theta_0$  and  $c \geq 0$ , where  $f_\theta$  is the p.d.f. of  $P_\theta$ . If  $T$  is a test with  $\beta_T(\theta_0) = \beta_{T_*}(\theta_0)$ , then for any  $\theta > \theta_0$ , either  $\beta_T(\theta) < \beta_{T_*}(\theta)$  or  $T = T_*$  a.s.  $P_\theta$ . (v) For any fixed  $\theta_1$ ,  $T_*$  is UMP for testing  $H_0 : \theta \leq \theta_1$  versus  $H_1 : \theta > \theta_1$ , with size  $\beta_{T_*}(\theta_1)$ .

**Theorem 4 (one-parameter exponential families):** Suppose that  $X$  has a p.d.f. in a one-parameter exponential family with  $\eta$  being a strictly monotone function of  $\theta$ . If  $\eta$  is increasing, then  $T_*$  given by Theorem 3 is UMP for testing  $H_0 : \theta \leq \theta_0$  versus  $H_1 : \theta > \theta_0$ , where  $\gamma$  and  $c$  are determined by  $\beta_{T_*}(\theta_0) = \alpha$ . If  $\eta$  is decreasing or  $H_0 : \theta \geq \theta_0$  ( $H_1 : \theta < \theta_0$ ), the result is still valid by reversing inequalities in the definition of  $T_*$ .

## 6.2 UMP tests and unbiased tests

**Theorem 1:** Suppose that the distribution of  $X$  is in a parametric family  $P$  indexed by a real-valued  $\theta$  and that  $P$  has monotone likelihood ratio in  $Y(X)$ . If  $\psi$  is a nondecreasing function of  $Y$ , then  $g(\theta) = \mathbb{E}[\psi(Y)]$  is a nondecreasing function of  $\theta$ .

**Proposition 1 (Generalized Neyman-Pearson lemma):** Let  $f_1, \dots, f_{m+1}$  be Borel functions on

## HYPOTHESIS TESTS

$\mathbb{R}^p$  integrable w.r.t. a  $\sigma$ -finite  $\nu$ . For given constants  $t_1, \dots, t_m$ , let  $\mathcal{T}$  be the class of Borel functions  $\phi$  (from  $\mathbb{R}^p \rightarrow [0, 1]$ ) satisfying  $\int \phi f_i d\nu \leq t_i, i = 1, \dots, m$ , and  $\mathcal{T}_0$  be the set of  $\phi$ 's in  $\mathcal{T}$  satisfying  $\int \phi f_i d\nu = t_i, i = 1, \dots, m$ . If there are constants  $c_1, \dots, c_m$  such that  $\phi_*(x) = \begin{cases} 1 & f_{m+1}(x) > c_1 f_1(x) + \dots + c_m f_m(x) \\ 0 & f_{m+1}(x) < c_1 f_1(x) + \dots + c_m f_m(x) \end{cases}$  is a member of  $\mathcal{T}_0$ , then  $\phi_*$  maximizes  $\int \phi f_{m+1} d\nu$  over  $\phi \in \mathcal{T}_0$ . If  $c_i \geq 0$  for all  $i$ , then  $\phi_*$  maximizes  $\int \phi f_{m+1} d\nu$  over  $\phi \in \mathcal{T}$ .

**Theorem 2:** Let  $f_1, \dots, f_m$  and  $\nu$  be given by Proposition 1. Then the set  $M = \{(\int \phi f_1 d\nu, \dots, \int \phi f_m d\nu) : \phi \text{ is from } \mathbb{R}^p \rightarrow [0, 1]\}$  is convex and closed. If  $(t_1, \dots, t_m)$  is an interior point of  $M$ , then there exist constant  $c_1, \dots, c_m$  such that the function  $\phi_*$  defined in Proposition 1 is in  $\mathcal{T}_0$ .

**Definition 1 (Two-sided hypotheses):** The following hypotheses are called two-sided hypotheses:  $H_0 : \theta \leq \theta_1$  or  $\theta \geq \theta_2$  versus  $H_1 : \theta_1 < \theta < \theta_2$ ,  $H_0 : \theta_1 \leq \theta \leq \theta_2$  versus  $H_1 : \theta < \theta_1$  or  $\theta > \theta_2$ ,  $H_0 : \theta = \theta_0$  versus  $H_1 : \theta \neq \theta_0$ , where  $\theta_0, \theta_1, \theta_2$  are given constants and  $\theta_1 < \theta_2$ .

**Theorem 3 (UMP tests for two-sided hypotheses):** Suppose that  $X$  has a p.d.f in a one-parameter exponential family, i.e., the p.d.f is  $f_\theta(x) = \exp\{\eta(\theta)Y(x) - \xi(\theta)\}h(x)$  w.r.t. a  $\sigma$ -finite measure, where  $\eta$  is a strictly increasing function of  $\theta$ . (i) For testing hypotheses  $H_0 : \theta \leq \theta_1$  or  $\theta \geq \theta_2$

versus  $H_1 : \theta_1 < \theta < \theta_2$ , a UMP test of size  $\alpha$  is  $T_*(x) = \begin{cases} 1 & c_1 < Y(X) < c_2 \\ \gamma_i & Y(X) = c_i, i = 1, 2, \\ 0 & Y(X) < c_1 \text{ or } Y(X) > c_2 \end{cases}$ , where

$c_i$ 's and  $\gamma_i$ 's are determined by  $\beta_{T_*}(\theta_1) = \beta_{T_*}(\theta_2) = \alpha$ . (ii)  $T_*$  minimizes  $\beta_T(\theta)$  over all  $\theta < \theta_1, \theta > \theta_2$ , and  $T$  satisfying  $\beta_T(\theta_1) = \beta_T(\theta_2) = \alpha$ . (iii) If  $T_*$  and  $T_{**}$  are two tests satisfying  $T(x) =$

$\begin{cases} 1 & c_1 < Y(X) < c_2 \\ \gamma_i & Y(X) = c_i, i = 1, 2, \\ 0 & Y(X) < c_1 \text{ or } Y(X) > c_2 \end{cases}$  and  $\beta_{T_*}(\theta_1) = \beta_{T_{**}}(\theta_1)$  and if the region  $\{T_{**} = 1\}$  is to the right

of  $\{T_* = 1\}$ , then  $\beta_{T_*}(\theta_1) < \beta_{T_{**}}(\theta)$  for  $\theta > \theta_1$  and  $\beta_{T_*}(\theta) > \beta_{T_{**}}(\theta)$  for  $\theta < \theta_1$ . If both  $T_*$  and  $T_{**}$

satisfy  $T(x) = \begin{cases} 1 & c_1 < Y(X) < c_2 \\ \gamma_i & Y(X) = c_i, i = 1, 2, \\ 0 & Y(X) < c_1 \text{ or } Y(X) > c_2 \end{cases}$  and  $\beta_{T_*}(\theta_1) = \beta_{T_*}(\theta_2) = \alpha$ , then  $T_* = T_{**}$  a.s.  $\mathcal{P}$ .

**Example 1:** Let  $X_1, \dots, X_n$  be i.i.d. from  $\mathcal{N}(\theta, 1)$ . By Theorem 3, a UMP test for testing  $H_0 : \theta \leq \theta_1$  or  $\theta \geq \theta_2$  versus  $H_1 : \theta_1 < \theta < \theta_2$  is  $T_*(X) = I_{(c_1, c_2)}(\bar{X})$ , where  $c_i$ 's are determined by  $\Phi(\sqrt{n}(c_2 - \theta_1)) - \Phi(\sqrt{n}(c_1 - \theta_1)) = \alpha$  and  $\Phi(\sqrt{n}(c_2 - \theta_2)) - \Phi(\sqrt{n}(c_2 - \theta_2)) = \alpha$ .

**Remark 1 (Nonexistence of UMP tests):** Unfortunately, a UMP test does not exist in general for testing hypotheses  $H_0 : \theta_1 \leq \theta \leq \theta_2$  versus  $H_1 : \theta < \theta_1$  or  $\theta > \theta_2$ ,  $H_0 : \theta = \theta_0$  versus  $H_1 : \theta \neq \theta_0$ .

**Definition 2:** Let  $\alpha$  be a given level of significance. A test  $T$  for  $H_0 : P \in \mathcal{P}_0$  versus  $P \in \mathcal{P}_1$  is said to be unbiased of level  $\alpha$  if and only if  $\beta_T(P) \leq \alpha, P \in \mathcal{P}_0$  and  $\beta_T(P) \geq \alpha, P \in \mathcal{P}_1$ . A test of size  $\alpha$  is called a uniformly most powerful unbiased (UMPU) test if and only if it is UMP within the class of unbiased tests of level  $\alpha$ .

**Definition 3 (Similarity):** Consider the hypotheses  $H_0 : \theta \in \Theta_0$  vs  $H_1 : \theta \in \Theta_1$ . Let  $\alpha$  be a given level of significance and let  $\bar{\Theta}_{01}$  be the common boundary of  $\Theta_0$  and  $\Theta_1$ , i.e., the set of points  $\theta$  that are points or limit points of both  $\Theta_0$  and  $\Theta_1$ . A test  $T$  is similar on  $\bar{\Theta}_{01}$  if and only if  $\beta_T(P) = \alpha$  for

all  $\theta \in \bar{\Theta}_{01}$ .

**Remark 2:** It is more convenient to work with similarity than to work with unbiasedness for testing  $H_0 : \theta \in \Theta_0$  vs  $H_1 : \theta \in \Theta_1$ .

**Theorem 4:** Consider hypotheses  $H_0 : \theta \in \Theta_0$  vs  $H_1 : \theta \in \Theta_1$ . Suppose that, for every  $T$ ,  $\beta_T(P)$  is continuous in  $\theta$ . If  $T_*$  is uniformly most powerful among all similar tests and has size  $\alpha$ , then  $T_*$  is a UMPU test.

### 6.3 Likelihood ratio and asymptotic tests

**Definition 1:** Let  $l(\theta) = f_\theta(X)$  be the likelihood function. For testing  $H_0 : \theta \in \Theta_0$  versus  $H_1 : \theta \in \Theta_1$ , a likelihood ratio (LR) test is any test that rejects  $H_0$  if and only if  $\lambda(X) < c$ , where  $c \in [0, 1]$  and  $\lambda(X)$  is the likelihood ratio defined by  $\lambda(X) = \sup_{\theta \in \Theta_0} l(\theta) / \sup_{\theta \in \Theta} l(\theta)$ .

**Remark 1 (Optimality):** When a UMP or UMPU test exists, an LR test is often the same as this optimal test.

**Proposition 1:** Suppose that  $X$  has a p.d.f. in one-parameter exponential family:  $f_\theta(x) = \exp\{\eta(\theta)Y(x) - \xi(\theta)\}h(x)$  w.r.t. a  $\sigma$ -finite measure  $\nu$ , where  $\eta$  is a strictly increasing and differentiable function of  $\theta$ . (i) For testing  $H_0 : \theta \leq \theta_0$  versus  $H_1 : \theta > \theta_0$ , there is an LR test whose rejection region is the same as that of the UMP test  $T_*$ . (ii) For testing  $H_0 : \theta \leq \theta_1$  or  $\theta \geq \theta_2$  versus  $H_1 : \theta_1 < \theta < \theta_2$ , there is an LR test whose rejection is the same as that of the UMP test  $T_*$ . (iii) For testing the other two-sided hypotheses, there is an LR test whose rejection region is equivalent to  $Y(X) < c_1$  or  $Y(X) > c_2$  for some constants  $c_1$  and  $c_2$ .

**Example 1:** Consider the testing problem  $H_0 : \theta = \theta_0$  versus  $H_1 : \theta \neq \theta_0$  based on i.i.d.  $X_1, \dots, X_n$  from the uniform distribution  $U(0, \theta)$ . We now show that the UMP test with rejection region  $X_{(n)} > \theta_0$  or  $X_{(n)} \leq \theta_0 \alpha^{1/n}$  is an LR test. Note that  $l(\theta) = \theta^{-n} I_{(X_{(n)}, \infty)}(\theta)$ . Hence  $\lambda(X) = \begin{cases} (X_{(n)}/\theta_0)^n & X_{(n)} \leq \theta_0 \\ 0 & X_{(n)} > \theta_0 \end{cases}$  and  $\lambda(X) < c$  is equivalent to  $X_{(n)} > \theta_0$  or  $X_{(n)}/\theta_0 < c^{1/n}$ . Taking  $c = \alpha$  ensures that the LR test has size  $\alpha$ .

**Definition 2 (Asymptotic tests):** Let  $X = (X_1, \dots, X_n)$  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  $\alpha$  is an asymptotic significance level of  $T_n$ . (ii) If  $\lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_0} \alpha_{T_n}(P)$  exists, it is called the limiting size of  $T_n$ . (iii)  $T_n$  is consistent iff the type II error probability converges to 0.

**Remark 2:** If  $\mathcal{P}_0$  is not a parametric family, the limiting size of  $T_n$  may be 1. This is the reason why we consider the weaker requirement in (i).

**Definition 3:** If  $\alpha \in (0, 1)$  is pre-assigned level of significance for the problem, then a consistent test  $T_n$  having an asymptotic significance level  $\alpha$  is called asymptotically correct, and a consistent test having limiting size  $\alpha$  is called strongly asymptotically correct.

**Theorem 1 (Asymptotic distribution of likelihood ratio):** Assume the conditions of Theorem 1 in section 4.6. Suppose that  $H_0 : \theta = g(\vartheta)$ , where  $\vartheta$  is  $(k - r)$ -vector of unknown parameters and  $g$  is a continuously differentiable function from  $\mathbb{R}^{k-r}$  to  $\mathbb{R}^k$  with a full rank  $\partial g(\vartheta)/\partial \vartheta$ . Under  $H_0$ ,  $-2 \log \lambda_n \rightarrow_d \chi_r^2$ , where  $\lambda_n = \lambda(X)$  and  $\chi_r^2$  is a random variable having the chi-square distribution



## HYPOTHESIS TESTS

$\chi_r^2$ . Consequently, the LR test with rejection region  $\lambda_n < e^{-\chi_{r,\alpha}^2/2}$  has asymptotic significance level  $\alpha$ , where  $\chi_{r,\alpha}^2$  is the  $(1 - \alpha)$ th quantile of the chi-square distribution  $\chi_r^2$ .

**Theorem 2:** Assume the conditions of Theorem 1 in section 4.6. (i) Under  $H_0 : R(\theta) = 0$ , where  $R(\theta)$  is a continuously differentiable function from  $\mathbb{R}^k$  to  $\mathbb{R}^r$ ,  $\hat{\theta}$  is an MLE or RLE of  $\theta$ ,  $\tilde{\theta}$  is an MLE or RLE of  $\theta$  under  $H_0 : R(\theta) = 0$ ,  $W_n = [R(\hat{\theta})]^T \{[C(\hat{\theta})]^T [I_n(\hat{\theta})]^{-1} C(\hat{\theta})\}^{-1} R(\hat{\theta}) \rightarrow_d \chi_r^2$  and, therefore, the test rejects  $H_0$  if and only if  $W_n > \chi_{r,\alpha}^2$  has asymptotic significance level  $\alpha$ , where  $\chi_{r,\alpha}^2$  is the  $(1 - \alpha)$ th quantile of the chi-square distribution  $\chi_r^2$ . (ii) The result in (i) still holds if  $W_n$  is replaced by  $R_n = [s_n(\tilde{\theta})]^T [I_n(\tilde{\theta})]^{-1} s_n(\tilde{\theta})$ .

### 6.4 Asymptotic chi-square tests

**Definition 1 (Testing in multinomial distributions):** Consider  $n$  independent trials with  $k$  possible outcomes for each trial. Let  $p_j > 0$  be the probability that the  $j$ th outcome occurs in a given trial and  $X_j$  be the number of occurrences of the  $j$ th outcome in  $n$  trials. Then  $X = (X_1, \dots, X_k)$  has the multinomial distribution with the parameter  $p = (p_1, \dots, p_k)$ . Let  $\xi_i = (0, \dots, 0, 1, 0, \dots, 0)$ , where the single nonzero component 1 is located in the  $j$ th position if the  $i$ th trial yields the  $j$ th outcome. Then  $\xi_1, \dots, \xi_n$  are i.i.d. and  $X/n = \bar{\xi} = \sum_{i=1}^n \xi_i/n$ .  $X/n$  is an unbiased estimator of  $p$  and by the CLT,  $Z_n(p) = \sqrt{n}(\frac{X}{n} - p) = \sqrt{n}(\bar{\xi} - p) \rightarrow_d \mathcal{N}_k(0, \Sigma)$ , where  $\Sigma = \text{Var}(X/\sqrt{n})$  is a symmetric  $k \times k$  matrix whose  $i$ th diagonal element is  $p_i(1 - p_i)$  and  $(i, j)$ th off-diagonal element is  $-p_i p_j$ . We first consider the problem of testing  $H_0 : p = p_0$  versus  $H_1 : p \neq p_0$  where  $p_0 = (p_{01}, \dots, p_{0k})$  is a known vector of cell probabilities.

**Definition 2 ( $\chi^2$  tests):** For testing  $H_0 : p = p_0$  vs  $H_1 : p \neq p_0$ , a class of tests related to the asymptotic tests, a popular test is based on the following  $\chi^2$ -statistic:  $\chi^2 = \sum_{j=1}^k \frac{(X_j - np_{0j})^2}{np_{0j}} = \|D(p_0)Z_n(p_0)\|^2$ , where  $D(c)$  with  $c = (c_1, \dots, c_k)$  is the  $k \times k$  diagonal matrix whose  $j$ th diagonal element is  $c_j^{-1/2}$ . Another popular test is based on the following modified  $\chi^2$ -statistic:  $\tilde{\chi}^2 = \sum_{j=1}^k \frac{(X_j - np_{0j})^2}{X_j} = \|D(X/n)Z_n(p_0)\|^2$ .

**Theorem 1:** Let  $\phi = (\sqrt{p_1}, \dots, \sqrt{p_k})$  and  $\Lambda$  be a  $k \times k$  projection matrix. (i) If  $\Lambda\phi = a\phi$ , then  $[Z_n(p)]^T D(p)\Lambda D(p)Z_n(p) \rightarrow_d \chi_r^2$ , where  $\chi_r^2$  has the chi-square distribution  $\chi_r^2$  with  $r = \text{tr}(\Lambda) - a$ . (ii) The same result holds if  $D(p)$  in (i) is replaced by  $D(X/n)$ .

**Example 1 (Goodness of fit tests):** Let  $Y_1, \dots, Y_n$  be i.i.d. from  $F$ . Consider the problem of testing  $H_0 : F = F_0$  versus  $H_1 : F \neq F_0$ , where  $F_0$  is a known c.d.f. One way to test  $H_0 : F = F_0$  is to partition the range of  $Y_1$  into  $k$  disjoint events  $A_1, \dots, A_k$  and test  $H_0 : p = p_0$  with  $p_j = P_F(A_j)$  and  $p_{0j} = P_{F_0}(A_j)$ ,  $j = 1, \dots, k$ . Let  $X_j$  be the number of  $Y_i$ 's in  $A_j$ ,  $j = 1, \dots, k$ . Based on  $X_j$ 's, the  $\chi^2$ -tests discussed previously can be applied.

**Definition 3 (Generalized  $\chi^2$ -statistics):** The generalized  $\chi^2$ -statistics  $\chi^2$  and  $\tilde{\chi}^2$  are defined to be the previously defined  $\chi^2$ -statistics with  $p_{0j}$ 's replaced by  $p_j(\hat{\theta})$ 's, where  $\theta \in \Theta \subset \mathbb{R}^s$  ( $s < k$ ) and  $\hat{\theta}$  is an MLE of  $\theta$  under  $H_0$ .

**Theorem 2:** Under  $H_0 : p = p(\theta)$ , the generalized  $\chi^2$ -statistics converge in distribution to  $\chi_{k-s-1}^2$ .

**Definition 4 (Testing independence):** Testing independence of  $\{A_j : j = 1, \dots, c\}$  and  $\{B_i, i = 1, \dots, r\}$  is equivalent to testing hypotheses  $H_0 : p_{ij} = p_i p_j$  for all  $i, j$  versus  $H_1 : p_{ij} \neq p_i p_j$  for some  $i, j$ , where  $p_{ij} = P(A_j \cap B_i) = \mathbb{E}[X_{ij}]/n$ ,  $p_i = P(B_i)$  and  $p_j = P(A_j)$ ,  $i = 1, \dots, r$ ,  $j = 1, \dots, c$ . Under



## CONFIDENCE SETS

$H_0$ , MLE's of  $p_{i\cdot}$  and  $p_{\cdot j}$  are  $\bar{X}_{i\cdot} = n_{i\cdot}/n$  and  $\bar{X}_{\cdot j}m_j/n$ , respectively,  $i = 1, \dots, r, j = 1, \dots, c$ . The number of free parameters is  $rc-1$ . Under  $H_0$ , the number of free parameters is  $r-1+c-1 = r+c-2$ . The difference of the two is  $(r-1)(c-1)$ . Then the  $\chi^2$ -test rejects  $H_0$  when  $\chi^2 > \chi_{(r-1)(c-1), \alpha}^2$  where  $\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(X_{ij} - n\bar{X}_{i\cdot}\bar{X}_{\cdot j})^2}{n\bar{X}_{i\cdot}\bar{X}_{\cdot j}}$  and  $\chi_{(r-1)(c-1), \alpha}^2$  is the  $(1-\alpha)$ th quantile of the chi-square distribution of  $\chi_{(r-1)(c-1)}^2$ .

**Remark 1 (Construction of asymptotic tests):** A simple method of constructing asymptotic tests (for almost all problems, parametric or nonparametric) for testing  $H_0 : \theta = \theta_0$  versus  $H_1 : \theta \neq \theta_0$ , where  $\theta$  is a vector of parameters, is to use an asymptotically normally distributed estimator  $\theta$ . Let  $\hat{\theta}_n$  be an estimator of  $\theta$  based on a sample size  $n$  from  $P$ . Suppose that under  $H_0$ ,  $V_n^{-1/2}(\hat{\theta}_n - \theta) \rightarrow_d \mathcal{N}_k(0, I_k)$ , where  $V_n$  is the asymptotic covariance matrix of  $\hat{\theta}_n$ . If  $V_n$  is known when  $\theta = \theta_0$ , then we define a test with rejection region  $(\hat{\theta}_n - \theta_0)^T V_n^{-1}(\hat{\theta}_n - \theta_0) > \chi_{k, \alpha}^2$ , where  $\chi_{k, \alpha}^2$  is the  $(1-\alpha)$ th quantile of the chi-squared distribution  $\chi_k^2$ . If  $V_n$  depends on the unknown population  $P$  even if  $H_0$  is true ( $\theta = \theta_0$ ), then we have to replace  $V_n$  by an estimator  $\hat{V}_n$ . If  $\hat{V}_n$  is consistent, then the resulting test still has asymptotic significance level  $\alpha$ .

**Theorem 3:** Assume that  $V_n^{-1/2}(\hat{\theta}_n - \theta) \rightarrow_d \mathcal{N}_k(0, I_k)$ , holds for any  $P$ . Assume also that  $\lambda_+[V_n] \rightarrow 0$ , where  $\lambda_+[V_n]$  is the largest eigenvalue of  $V_n$ . (i) The test having rejection region  $(\hat{\theta}_n - \theta_0)^T V_n^{-1}(\hat{\theta}_n - \theta_0) > \chi_{k, \alpha}^2$  with a known  $V_n$  (or with  $V_n$  replaced by a consistent estimator  $\hat{V}_n$ ) is consistent. (ii) If we choose  $\alpha = \alpha_n \rightarrow 0$  as  $n \rightarrow \infty$  and  $\chi_{k, 1-\alpha_n}^2 \lambda_+[V_n] = o(1)$ , then the test in (i) is Chernoff-consistent.

## 7 Confidence Sets

### 7.1 Pivotal quantities and confidence sets

**Definition 1 (Confidence sets):**  $X$ : a sample from a population  $P \in \mathcal{P}$ .  $\theta = \theta(P)$ : a functional from  $\mathcal{P}$  to  $\Theta \subset \mathbb{R}^k$  for a fixed integer  $k$ .  $C(X)$ : a confidence set for  $\theta$ , a set in  $\mathcal{B}_\Theta$  depending only on  $X$ . The confidence coefficient of  $C(X) : \inf_{P \in \mathcal{P}} P(\theta \in C(X))$ . If the confidence coefficient of  $C(X)$  is  $\geq 1 - \alpha$  for fixed  $\alpha \in (0, 1)$ , then we say that  $C(X)$  has confidence level  $1 - \alpha$  or  $C(X)$  is a level  $1 - \alpha$  confidence set.

**Definition 2:** A known Borel function  $h$  of  $(X, \theta)$  is called a pivotal quantity if and only if the distribution of  $h(X, \theta)$  does not depend on  $P$ .

**Construction:** First, find two constants  $c_1$  and  $c_2$  such that  $P(c_1 \leq h(x, \theta) \leq c_2) \geq 1 - \alpha$ . Next, define  $C(X) = \{\theta \in \Theta : c_1 \leq h(X, \theta) \leq c_2\}$ . Then  $C(X)$  is a level  $1 - \alpha$  confidence set. The confidence coefficient of  $C(X)$  may not be  $1 - \alpha$ . If  $h(X, \theta)$  has a continuous c.d.f., then we can choose  $c_i$ 's such that the equality in the last expression holds and the confidence set  $C(X)$  has confidence coefficient  $1 - \alpha$ .

**Computation:** When  $h(X, \theta)$  and  $c_i$ 's are chosen, we need to compute the confidence set  $C(X) = \{c_1 \leq h(X, \theta) \leq c_2\}$ . This can be done by inverting  $c_1 \leq h(X, \theta) \leq c_2$ . For example, if  $\theta$  is real-valued and  $h(X, \theta)$  is monotone in  $\theta$  when  $X$  is fixed, then  $C(X) = \{\theta : \underline{\theta}(X) \leq \theta \leq \bar{\theta}(X)\}$  for some  $\underline{\theta}(X) < \bar{\theta}(X)$ , i.e.,  $C(X)$  is an interval (finite or infinite). If  $h(X, \theta)$  is not monotone, then  $C(X)$  may be a union of several intervals.

## CONFIDENCE SETS

**Example 1 (Location-scale families):** Suppose that  $X_1, \dots, X_n$  are i.i.d. with a Lebesgue p.d.f.  $\frac{1}{\sigma}f(\frac{x-\mu}{\sigma})$ , where  $\mu \in \mathbb{R}, \sigma > 0$ , and  $f$  is a known Lebesgue p.d.f. Consider first the case where  $\sigma$  is known and  $\theta = \mu$ .  $\bar{X} - \mu$  is a pivotal quantity. Let  $c_1$  and  $c_2$  be constants such that  $P(c_1 \leq \bar{X} - \mu \leq c_2) = 1 - \alpha$ , then  $C(X) = \{\mu : \bar{X} - c_2 \leq \mu \leq \bar{X} - c_1\}$ . Consider next the case where  $\mu$  is known and  $\theta = \sigma$ .  $S/\sigma$  is a pivotal quantity, where  $S^2$  is the sample variance. Let  $c_1$  and  $c_2$  be chosen such that  $P(c_1 \leq S/\sigma \leq c_2) = 1 - \alpha$ . If both  $c_i$ 's are positive, then  $C(X) = \{\sigma : S/c_2 \leq \sigma S/c_1\} = [S/c_2, S/c_1]$  is a finite interval. Similarly, if  $c_1 = 0$  or  $c_2 = \infty$ , then  $C(X) = [S/c_2, \infty)$  or  $(0, S/c_1]$ . When  $\theta = \sigma$  and  $\mu$  is also known,  $S/\sigma$  is still a pivotal quantity and, hence, confidence intervals of  $\sigma$  based on  $S$  are still valid. Finally, we consider the case where both  $\mu$  and  $\sigma$  are unknown and  $\theta = \mu$ . There are still many different pivotal quantities, but the most commonly used pivotal quantity is  $t(X) = \sqrt{n}(\bar{X} - \mu)/S$ . The distribution of  $t(X)$  does not depend on  $(\mu, \sigma)$ . When  $f$  is normal,  $t(X)$  has the  $t$ -distribution  $t_{n-1}$ . A confidence interval for  $\mu$  based on  $t(X)$  is of the form  $\{\mu : c_1 \leq \sqrt{n}(\bar{X} - \mu)/S \leq c_2\} = [\bar{X} - c_2 S/\sqrt{n}, \bar{X} - c_1 S/\sqrt{n}]$ , where  $c_i$ 's are chosen so that  $P(c_1 \leq t(X) \leq c_2) = 1 - \alpha$ .

**Example 2:** Let  $X_1, \dots, X_n$  be i.i.d. random variables from the uniform distribution  $U(0, \theta)$ . Consider the problem of finding a confidence set for  $\theta$ . Note that  $X_{(n)}/\theta$  has the Lebesgue p.d.f.  $nx^{n-1}I_{(0,1)}(x)$ . Hence  $c_i$ 's should satisfy  $c_2^n - c_1^n = 1 - \alpha$ . The resulting confidence interval for  $\theta$  is  $[c_2^{-1}X_{(n)}, c_1^{-1}X_{(n)}]$ .

**Proposition 1 (Existence of pivotal quantities in parametric problems):** Let  $T(X) = (T_1(X), \dots, T_s(X))$  and  $T_1, \dots, T_s$  be independent statistics. Suppose that each  $T_i$  has a continuous c.d.f.  $F_{T_i, \theta}$  indexed by  $\theta$ . Then  $h(X, \theta) = \prod_{i=1}^s F_{T_i, \theta}(T_i(X))$  is a pivotal quantity.

**Theorem 1:** Suppose that  $P$  is in a parametric family indexed by a real-valued  $\theta$ . Let  $T(X)$  be a real-valued statistic with c.d.f.  $F_{T, \theta}(t)$  and let  $\alpha_1$  and  $\alpha_2$  be fixed positive constants such that  $\alpha_1 + \alpha_2 = \alpha < \frac{1}{2}$ . (i) Suppose that  $F_{T, \theta}(t)$  and  $F_{T, \theta}(t-)$  are nonincreasing in  $\theta$  for each fixed  $t$ . Define  $\bar{\theta} = \sup\{\theta : F_{T, \theta}(T) \geq \alpha_1\}$  and  $\underline{\theta} = \inf\{\theta : F_{T, \theta}(T-) \leq 1 - \alpha_2\}$ . Then  $[\underline{\theta}(T), \bar{\theta}(T)]$  is a level  $1 - \alpha$  confidence interval for  $\theta$ . (ii) If  $F_{T, \theta}(t)$  and  $F_{T, \theta}(t-)$  are nondecreasing in  $\theta$  for each  $t$ , then the same result holds with  $\underline{\theta} = \inf\{\theta : F_{T, \theta}(T) \geq \alpha_1\}$  and  $\bar{\theta} = \sup\{\theta : F_{T, \theta}(T-) \leq 1 - \alpha_2\}$ . (iii) If  $F_{T, \theta}$  is a continuous c.d.f. for any  $\theta$ , then  $F_{T, \theta}(T)$  is a pivotal quantity and the confidence interval in (i) or (ii) has confidence coefficient  $1 - \alpha$ .

**Definition 3 (Asymptotic criterion):** In some problems, especially in nonparametric problems, it is difficult to find a reasonable confidence set with a given confidence coefficient or confidence level  $1 - \alpha$ . A common approach is to find a confidence set whose confidence coefficient or confidence level is nearly  $1 - \alpha$  when the sample size  $n$  is large. A confidence set  $C(X)$  for  $\theta$  has asymptotic confidence level  $1 - \alpha$  if  $\liminf_n P(\theta \in C(X)) \geq 1 - \alpha$  for any  $P \in \mathcal{P}$ . If  $\lim_{n \rightarrow \infty} P(\theta \in C(X)) = 1 - \alpha$  for any  $P \in \mathcal{P}$ , then  $C(X)$  is a  $1 - \alpha$  asymptotically correct confidence set.

**Definition 4 (Asymptotically pivotal quantities):** A known Borel function of  $(X, \theta)$ ,  $h_n(X, \theta)$ , is said to be asymptotically pivotal iff the limiting distribution of  $h_n(X, \theta)$  does not depend on  $P$ . Like a pivotal quantity in constructing confidence sets with a given confidence coefficient or confidence level, an asymptotically pivotal quantity can be used in constructing asymptotically correct confidence sets.

**Example 3 (Functions of means):** Suppose that  $X_1, \dots, X_n$  are i.i.d. random vectors having

## CONFIDENCE SETS

a c.d.f.  $F$  on  $\mathbb{R}^d$  and that the unknown parameter of interest  $\theta = g(\mu)$ , where  $\mu = \mathbb{E}(X_1)$  and  $g$  is a known differentiable function from  $\mathbb{R}^d$  to  $\mathbb{R}^k$ ,  $k \leq d$ . From the CLT,  $\hat{\theta}_n = g(\bar{X})$  satisfies  $V_n^{-1/2}(\hat{\theta}_n - \theta) \rightarrow_d \mathcal{N}_k(0, I_k)$ ,  $V_n = [\nabla g(\mu)]^T \text{Var}(X_1) \nabla g(\mu)/n$ . A consistent estimator of the asymptotic covariance matrix  $V_n$  is  $\hat{V}_n = [\nabla g(\bar{X})]^T S^2 \nabla g(\bar{X})/n$ . Thus,  $C(X) = \{\theta : \|\hat{V}_n^{-1/2}(\hat{\theta}_n - \theta)\|^2 \leq \chi_{k,\alpha}^2\}$ , is a  $1 - \alpha$  asymptotically correct confidence set for  $\theta$ .

### 7.2 Inverting acceptance regions of tests, UMA and UMAU confidence sets

**Remark 1 (Confidence sets and hypothesis tests):** Another popular method of constructing confidence sets is to use a close relationship between confidence sets and hypothesis tests. For any test  $T$ , the set  $\{x : T(x) \neq 1\}$  is called the acceptance region. This terminology is not precise when  $T$  is a randomized test.

**Theorem 1:** For each  $\theta_0 \in \Theta$ , let  $T_{\theta_0}$  be a test for  $H_0 : \theta = \theta_0$  with significance level  $\alpha$  and acceptance region  $A(\theta_0)$ . For each  $x$  in the range of  $X$ , define  $C(x) = \{\theta : x \in A(\theta)\}$ . Then  $C(X)$  is a level  $1 - \alpha$  confidence set for  $\theta$ . If  $T_{\theta_0}$  is nonrandomized and has size  $\alpha$  for every  $\theta_0$ , then  $C(X)$  has confidence coefficient  $1 - \alpha$ .

**Proposition 1:** Let  $C(X)$  be a confidence set for  $\theta$  with confidence level  $1 - \alpha$ . For any  $\theta_0 \in \Theta$ , define a region  $A(\theta_0) = \{x : \theta_0 \in C(x)\}$ . Then the test  $T(X) = 1 - I_{A(\theta_0)}(X)$  has significance level  $\alpha$  for testing  $H_0 : \theta = \theta_0$  versus some  $H_1$ .

**Example 1:** Suppose that  $X$  has the following p.d.f. in a one-parameter exponential family:  $f_\theta(x) = \exp\{\eta(\theta)Y(x) - \xi(\theta)\}h(x)$ , where  $\theta$  is real-valued and  $\eta(\theta)$  is nondecreasing in  $\theta$ . Consider  $H_0 : \theta = \theta_0$  and  $H_1 : \theta > \theta_0$ . The acceptance region of the UMP test of size  $\alpha$  is  $A(\theta_0) = \{x : Y(x) \leq c(\theta_0)\}$ . It can be shown that  $c(\theta)$  is nondecreasing in  $\theta$ . Inverting  $A(\theta)$ , we obtain  $C(X) = [\underline{\theta}(X), \infty)$  or  $(\underline{\theta}(X), \infty)$ , a one-sided confidence interval for  $\theta$  with confidence level  $1 - \alpha$ .  $\underline{\theta}(X)$  is called a lower confidence bound for  $\theta$ . If  $H_0 : \theta = \theta_0$  and  $H_1 : \theta < \theta_0$  are considered, then  $C(X) = \{\theta : Y(X) \geq c(\theta)\}$  and is of the form  $(-\infty, \bar{\theta}(X)]$  or  $(-\infty, \bar{\theta}(X))$ .  $\bar{\theta}(X)$  is called an upper confidence bound for  $\theta$ .

**Remark 2 (Confidence sets related to optimal tests):** For a confidence set obtained by inverting the acceptance regions of some UMP or UMPU tests, it is expected that the confidence set inherits some optimality property.

**Definition 1:** Let  $\theta \in \Theta$  be an unknown parameter and  $\Theta'$  be a subset of  $\Theta$  that does not contain the true parameter value  $\theta$ . A confidence set  $C(X)$  for  $\theta$  with confidence coefficient  $1 - \alpha$  is said to be  $\Theta'$ -uniformly most accurate iff for any other confidence set  $C_1(X)$  with confidence level  $1 - \alpha$ ,  $P(\theta' \in C(X)) \leq P(\theta' \in C_1(X))$  for all  $\theta' \in \Theta'$ .  $C(X)$  is UMA iff it is  $\Theta'$ -UMA with  $\Theta' = \{\theta\}^c$ .

**Theorem 2:** Let  $C(X)$  be a confidence set for  $\theta$  obtained by inverting the acceptance regions of nonrandomized tests  $T_{\theta_0}$  for testing  $H_0 : \theta = \theta_0$  versus  $H_1 : \theta \in \Theta_{\theta_0}$ . Suppose that for each  $\theta_0$ ,  $T_{\theta_0}$  is UMP of size  $\alpha$ . Then  $C(X)$  is  $\Theta'$ -UMA with confidence coefficient  $1 - \alpha$ , where  $\Theta' = \{\theta' : \theta \in \Theta_{\theta'}\}$ .

**Definition 2:** Let  $\theta \in \Theta$  be an unknown parameter,  $\Theta'$  be a subset of  $\Theta$  that does not contain the true parameter value  $\theta$ , and  $1 - \alpha$  be a given confidence level. (i) A level  $1 - \alpha$  confidence set  $C(X)$  is said to be  $\Theta'$ -unbiased (unbiased when  $\Theta' = \{\theta\}^c$ ) iff  $P(\theta' \in C(X)) \leq 1 - \alpha$  for all  $\theta' \in \Theta'$ . (ii) Let  $C(X)$  be a  $\Theta'$ -unbiased confidence set with confidence coefficient  $1 - \alpha$ . If  $P(\theta' \in C(X)) \leq P(\theta' \in C_1(X))$  for all  $\theta' \in \Theta'$  holds for any other  $\Theta'$ -unbiased confidence set  $C_1(X)$

## CONFIDENCE SETS

with confidence level  $1 - \alpha$ , then  $C(X)$  is  $\Theta'$ -uniformly most accurate unbiased (UMAU).  $C(X)$  is UMAU if and only if it is  $\Theta'$ -UMAU with  $\Theta' = \{\theta\}^c$ .

**Theorem 3:** Let  $C(X)$  be a confidence set for  $\theta$  obtained by inverting the acceptance regions of nonrandomized tests  $T_{\theta_0}$  for testing  $H_0 : \theta = \theta_0$  versus  $H_1 : \theta \in \Theta_{\theta_0}$ . If  $T_{\theta_0}$  is unbiased of size  $\alpha$  for each  $\theta_0$ , then  $C(X)$  is  $\Theta'$ -unbiased with confidence coefficient  $1 - \alpha$ , where  $\Theta' = \{\theta' : \theta \in \Theta_{\theta'}\}$ . If  $T_{\theta_0}$  is also UMPU for each  $\theta_0$ , then  $C(X)$  is  $\Theta'$ -UMAU.

**Definition 3:** Consider a sample  $X$  from a population in a parametric family dominated by a  $\sigma$ -finite measure. Let  $f_{\theta}(x)$  be the p.d.f. of  $X$  and  $\pi(\theta)$  be a prior p.d.f. w.r.t. a  $\sigma$ -finite measure  $\lambda$  on  $(\Theta, \mathcal{B}_{\Theta})$ . Let  $p_x(\theta) = f_{\theta}(x)\pi(\theta)/m(x)$  be the posterior p.d.f. w.r.t.  $\lambda$ , where  $x$  is the observed  $X$  and  $m(x) = \int_{\Theta} f_{\theta}(x)\pi(\theta)d\lambda$ . For any  $\alpha \in (0, 1)$ , a level  $1 - \alpha$  credible set for  $\theta$  is any  $C \in \mathcal{B}_{\Theta}$  with  $P_{\theta|x}(\theta \in C) = \int_C p_x(\theta)d\lambda \geq 1 - \alpha$ . A level  $1 - \alpha$  highest posterior density (HPD) credible set for  $\theta$  is defined to be the event  $C(x) = \{\theta : p_x(\theta) \geq c_{\alpha}\}$ , where  $c_{\alpha}$  is chosen so that  $\int_{C(x)} p_x(\theta)d\lambda \geq 1 - \alpha$ .

**Example 2:** Let  $X_1, \dots, X_n$  be i.i.d. as  $\mathcal{N}(\theta, \sigma^2)$  with an unknown  $\theta \in \mathbb{R}$  and a known  $\sigma^2$ . Let  $\pi(\theta)$  be the p.d.f. of  $\mathcal{N}(\mu_0, \sigma_0^2)$  with known  $\mu_0$  and  $\sigma_0^2$ . Then  $p_x(\theta)$  is the p.d.f. of  $\mathcal{N}(\mu_*(x), c^2)$ , where  $\mu_*(x) = \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 + \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\bar{x}$  and  $c^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}$ , and the HPD credible set is  $C(x) = \{\theta : e^{-[\theta - \mu_*(x)]^2/(2c^2)} \geq c_{\alpha}\sqrt{2\pi c}\} = \{\theta : |\theta - \mu_*(x)| \leq \sqrt{2}c[-\log(c_{\alpha}\sqrt{2\pi c})]^{1/2}\}$ . Let  $\Phi$  be the standard normal c.d.f. The quantity  $\sqrt{2}c[-\log(c_{\alpha}\sqrt{2\pi c})]^{1/2}$  must be  $cz_{1-\alpha/2}$ , where  $z_{\alpha} = \Phi^{-1}(\alpha)$ , since it is chosen so that  $P_{\theta|x}(C(x)) = 1 - \alpha$  and  $P_{\theta|x} = \mathcal{N}(\mu_*(x), c^2)$ . Therefore,  $C(x) = [\mu_*(x) - cz_{1-\alpha/2}, \mu_*(x) + cz_{1-\alpha/2}]$ .

### 7.3 Lengths and expected lengths of confidence intervals

**Remark 1 (Length criterion):** For confidence intervals of a real-valued with the same confidence coefficient, an apparent measure of their performance is the interval length. Shorter confidence intervals are preferred, since they are more informative. When confidence intervals are constructed by using pivotal quantities or by inverting acceptance regions of tests, choosing a reasonable class of confidence intervals amounts to selecting good pivotal quantities or tests.

**Theorem 1:** Let  $\theta$  be a real-valued parameter and  $T(X)$  be a real-valued statistic. (i) Let  $U(X)$  be a positive statistic. Suppose that  $(T - \theta)/U$  is a pivotal quantity having a Lebesgue p.d.f.  $f$  that is unimodal at  $x_0 \in \mathbb{R}$  in the sense that  $f(x)$  is nondecreasing for  $x \leq x_0$  and  $f(x)$  is nonincreasing for  $x \geq x_0$ . Consider the following class of confidence intervals for  $\theta$ :  $\mathcal{C} = \{[T - bU, T - aU] : a \in \mathbb{R}, b \in \mathbb{R}, \int_a^b f(x)dx = 1 - \alpha\}$ . If  $[T - b_*U, T - a_*U] \in \mathcal{C}$ ,  $f(a_*) = f(b_*) > 0$ , and  $a_* \leq x_0 \leq b_*$ , then the interval  $[T - b_*U, T - a_*U]$  has the shortest length within  $\mathcal{C}$ . (ii) Suppose that  $T > 0, \theta > 0, T/\theta$  is a pivotal quantity having a Lebesgue p.d.f.  $f$ , and that  $x^2f(x)$  is unimodal at  $x_0$ . Consider the following class of confidence intervals for  $\theta$ :  $\mathcal{C} = \{[b^{-1}T, a^{-1}T] : a > 0, b > 0, \int_a^b f(x)dx = 1 - \alpha\}$ . If  $[b_*^{-1}T, a_*^{-1}T] \in \mathcal{C}$ ,  $a_*^2f(a_*) = b_*^2f(b_*) > 0$ , and  $a_* \leq x_0 \leq b_*$ , then the interval  $[b_*^{-1}T, a_*^{-1}T]$  has the shortest length within  $\mathcal{C}$ .

**Example 1:** Let  $X_1, \dots, X_n$  be i.i.d. from  $\mathcal{N}(\mu, \sigma^2)$  with unknown  $\mu$  and  $\sigma^2$ . Confidence intervals for  $\theta = \mu$  using the pivotal quantity  $\sqrt{n}(\bar{X} - \mu)/S$  form the class  $\mathcal{C}$  in Theorem 1(i) with  $f$  being the p.d.f. of the  $t$ -distribution  $t_{n-1}$ , which is unimodal at  $x_0 = 0$ . Since  $f$  is symmetric about 0,  $f(a_*) = f(b_*)$  implies  $a_* = -b_*$ . Therefore, the equal-tail confidence interval  $[\bar{X} - t_{n-1, \alpha/2}S/\sqrt{n}, \bar{X} +$

## CONFIDENCE SETS

$t_{n-1, \alpha/2} S / \sqrt{n}$  has the shortest length within  $\mathcal{C}$ . If  $\theta = \mu$  and  $\sigma^2$  is known, then we replace  $S$  by  $\sigma$  and  $f$  by the standard normal p.d.f. The resulting confidence interval is  $[\bar{X} - \Phi^{-1}(1 - \alpha/2)\sigma/\sqrt{n}, \bar{X} + \Phi^{-1}(1 - \alpha/2)\sigma/\sqrt{n}]$ , which is the shortest interval of the form  $[\bar{X} - b, \bar{X} - a]$  with confidence coefficient  $1 - \alpha$ . Consider next confidence intervals for  $\theta = \sigma^2$  using the pivotal quantity  $(n - 1)S^2/\sigma^2$ , which form the class  $\mathcal{C}$  in Theorem 1(ii) with  $f$  being the p.d.f. of the chi-square distribution  $\chi_{n-1}^2$ . Note that  $x^2 f(x)$  is unimodal, but not symmetric. By Theorem 1(ii), the shortest-length interval within  $\mathcal{C}$  is  $[b_*^{-1}(n - 1)S^2, a_*^{-1}(n - 1)S^2]$ , where  $a_*$  and  $b_*$  are solutions of  $a_*^2 f(a_*) = b_*^2 f(b_*)$  and  $\int_{a_*}^{b_*} f(x)dx = 1 - \alpha$ . Numerical values of  $a_*$  and  $b_*$  can be obtained. Note that this interval is not equal-tail.

**Remark 2 (Expected length):** In a problem where a shortest-length confidence interval does not exist, we may have to use the expected length as the criterion in comparing confidence intervals.

**Theorem 2 (Pratt's theorem):** Let  $X$  be a sample from  $P$  and  $C(X)$  be a confidence set for  $\theta \in \mathbb{R}^k$ . Suppose that  $\text{vol}(C(x)) = \int_{C(x)} d\theta'$  is finite a.s.  $P$ . Then the expected volume of  $C(X)$  is  $\mathbb{E}[\text{vol}(C(X))] = \int_{\theta \neq \theta'} P(\theta' \in C(X)) d\theta'$ .

**Remark 3:** It follows from Theorem 2 that if  $C(X)$  is UMA (or UMAU) with confidence coefficient  $1 - \alpha$ , then it has the smallest expected volume among all confidence sets (or all unbiased confidence sets) with confidence level  $1 - \alpha$ .

### 7.4 Asymptotic confidence sets

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

**Proposition 1:** Let  $C_j(X) = \{\theta : \|\hat{V}_{jn}^{-1/2}(\hat{\theta}_{jn} - \theta)\|^2 \leq \chi_{k, \alpha}^2\}$ ,  $j = 1, 2$ , be the confidence sets based on  $\hat{\theta}_{jn}$  satisfying  $V_{jn}^{-1/2}(\hat{\theta}_{jn} - \theta) \rightarrow_d \mathcal{N}_k(0, I_k)$ , where  $\hat{V}_{jn}$  is consistent for  $V_{jn}$ ,  $j = 1, 2$ . If  $\det(V_{1n}) < \det(V_{2n})$  for sufficiently large  $n$ , then  $P(\text{vol}(C_1(X)) < \text{vol}(C_2(X))) \rightarrow 1$ .

**Example 1 (Parametric likelihoods):** Consider the case where  $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$  is a parametric family dominated by a  $\sigma$ -finite measure, where  $\Theta \subset \mathbb{R}^k$ . Consider  $\theta = (\vartheta, \varphi)$  and confidence sets for  $\vartheta$  with dimension  $r$ . Let  $l(\theta)$  be the likelihood function based on the observation  $X = x$ . The acceptance region of the LR test with  $\Theta_0 = \{\theta : \vartheta = \vartheta_0\}$  is  $A(\vartheta_0) = \{x : l(\vartheta_0, \hat{\varphi}_{\vartheta_0}) \geq e^{-c_\alpha/2} l(\hat{\theta})\}$ , where  $l(\hat{\theta}) = \sup_{\theta \in \Theta} l(\theta)$ ,  $l(\vartheta, \hat{\varphi}_\vartheta) = \sup_\varphi l(\vartheta, \varphi)$ , and  $c_\alpha$  is a constant related to the significance level  $\alpha$ . If  $c_\alpha$  is chosen to be  $\chi_{r, \alpha}^2$ , the  $(1 - \alpha)$ th quantile of the chi-square distribution  $\chi_r^2$ , then  $C(X) = \{\vartheta : l(\vartheta, \hat{\varphi}_\vartheta) \geq e^{-c_\alpha/2} l(\hat{\theta})\}$  is a  $1 - \alpha$  asymptotically correct confidence set. Note that this confidence set and the one given by  $C(X) = \{\theta : \|\hat{V}_n^{-1/2}(\hat{\theta}_n - \theta)\|^2 \leq \chi_{k, \alpha}^2\}$  are generally different. In many cases  $-l(\vartheta, \varphi)$  is a convex function of  $\vartheta$  and, therefore,  $C(X)$  based on LR tests is a bounded set in  $\mathbb{R}^k$ . When  $\Theta_0 = \{\theta : \vartheta = \vartheta_0\}$ , Wald's test has acceptance region  $A(\vartheta_0) = \{x : (\hat{\vartheta} - \vartheta_0)^T \{C^T[I_n(\hat{\theta})]^{-1}C\}^{-1}(\hat{\vartheta} - \vartheta_0) \leq \chi_{r, \alpha}^2\}$ , where  $\hat{\theta} = (\hat{\vartheta}, \hat{\varphi})$  is an MLE or RLE of  $\theta = (\vartheta, \varphi)$ ,  $I_n(\theta)$  is the Fisher information matrix based on  $X$ ,  $C^T = (I_r, 0)$ , and  $0$  is an  $r \times (k - r)$  matrix of 0's. The confidence set obtained by inverting  $A(\vartheta)$  is  $C(X) = \{\theta : \|\hat{V}_n^{-1/2}(\hat{\vartheta} - \vartheta)\|^2 \leq \chi_{k, \alpha}^2\}$  with

## CONFIDENCE SETS

$\hat{V}_n = C^T [I_n(\hat{\theta})]^{-1} C$ . When  $\Theta_0 = \{\theta : \vartheta = \vartheta_0\}$ , Rao's score test has acceptance region  $A(\vartheta_0) = \{x : [s_n(\vartheta_0, \hat{\varphi}_{\vartheta_0})]^T [I_n(\vartheta_0, \hat{\varphi}_{\vartheta_0})]^{-1} s_n(\vartheta_0, \hat{\varphi}_{\vartheta_0}) \leq \chi_{r,\alpha}^2\}$ , where  $s_n(\theta) = \partial \log l(\theta) / \partial \theta$ . The confidence set obtained by inverting  $A(\vartheta)$  is also  $1 - \alpha$  asymptotically correct.

**Remark 1 (Confidence intervals for quantiles):** Let  $X_1, \dots, X_n$  be i.i.d. from a continuous c.d.f.  $F$  on  $\mathbb{R}$  and let  $\theta = F^{-1}(p)$  be the  $p$ th quantile of  $F$ ,  $0 < p < 1$ . The general methods can be applied to obtain a confidence set for  $\theta$ , but we introduce here a method that works particularly for quantile problems.

**Theorem 1 (Refinement of Bahadur's representation):** Let  $X_1, \dots, X_n$  be i.i.d. from a continuous c.d.f.  $F$  on  $\mathbb{R}$  that is twice differentiable at  $\theta = F^{-1}(p)$ ,  $0 < p < 1$ , with  $F'(\theta) > 0$ . Let  $F_n$  be the empirical c.d.f. Let  $\{k_n\}$  be a sequence of interges satisfying  $1 \leq k_n \leq n$  and  $k_n/n = p + o((\log n)^\delta / \sqrt{n})$  for some  $\delta > 0$ . Then  $X_{(k_n)} = \theta + \frac{(k_n/n) - F_n(\theta)}{F'(\theta)} + O\left(\frac{(\log n)^{(1+\delta)/2}}{n^{3/4}}\right)$  a.s.

**Theorem 2:** Assume the conditions in Theorem 1 and  $k_n/n = p + cn^{-1/2} + o(n^{-1/2})$  with a constant  $c$ . Then  $\sqrt{n}(X_{(k_n)} - F_n^{-1}(p)) \rightarrow_{\text{a.s.}} c/F'(\theta)$ .

**Theorem 3:** Assume the conditions in Theorem 1. Let  $\{k_{1n}\}$  and  $\{k_{2n}\}$  be two sequences of integers satisfying  $1 \leq k_{1n} < k_{2n} \leq n$ ,  $k_{1n}/n = p - z_{1-\alpha/2} \sqrt{p(1-p)/n} + o(n^{-1/2})$ , and  $k_{2n}/n = p + z_{1-\alpha/2} \sqrt{p(1-p)/n} + o(n^{-1/2})$ , where  $z_a = \Phi^{-1}(a)$ . Then the confidence interval  $C(X) = [X_{(k_{1n})}, X_{(k_{2n})}]$  has the property that  $P(\theta \in C(X))$  does not depend on  $P$  and  $\lim_{n \rightarrow \infty} \inf_{P \in \mathcal{P}} P(\theta \in C(X)) = \lim_{n \rightarrow \infty} P(\theta \in C(X)) = 1 - \alpha$ . Furthermore, the length of  $C(X) = \frac{2z_{1-\alpha/2} \sqrt{p(1-p)}}{F'(\theta) \sqrt{n}} + o(\frac{1}{\sqrt{n}})$  a.s.

## 7.5 Variance estimation, replication, jackknife, and bootstrap

**Motivation:** To evaluate and compare different estimators, we need consistent estimators of variances or asymptotic variances of estimators. Traditional approach to estimate  $\text{Var}(\hat{\theta})$ : Derivation and substitution. (i) First, we derive a theoretical formula; (ii) Approximation (asymptotic theory) is usually needed; (iii) The formula may depend on unknown quantities; (iv) We then substitute unknown quantities by estimators.

**Method 1 (The  $\delta$ -method):**  $Y_1, \dots, Y_n$  are i.i.d. ( $k$ -dimensional).  $\theta = g(\mu)$ ,  $\hat{\theta} = g(\bar{Y})$ ,  $\text{Var}(\hat{\theta}) \approx [\nabla g(\mu)]^T \text{Var}(\bar{Y}) [\nabla g(\mu)]$ . An estimator of  $\text{Var}(\hat{\theta})$  is  $\hat{V}_n = [\nabla g(\bar{Y})]^T (S^2/n) \nabla g(\bar{Y})$ .

**Method 2:** Suppose we can independently obtain  $B$  copies of the data set  $X$ , say  $X^1, \dots, X^B$ . Then we can calculate  $\hat{\theta}^b = \hat{\theta}(X^b)$ ,  $b = 1, \dots, B$ . Variance of  $\hat{\theta}$  can be estimated as  $\frac{1}{B} \sum_{b=1}^B (\hat{\theta}^b - \frac{1}{B} \sum_{l=1}^B \hat{\theta}^l)^2$ . No derivation is needed.

**Definition 1 (Jackknife):** Consider pseudo replicates  $X^i = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$ ,  $i = 1, \dots, n$ . Let  $\hat{\theta}_{-i}$  be the same estimator as  $\hat{\theta}_n$  but based on  $X^i$ ,  $i = 1, \dots, n$ . Since  $\hat{\theta}_n$  and  $\hat{\theta}_{-1}, \dots, \hat{\theta}_{-n}$  estimate the same quantity, the following "sample variance" can be used as a measure of the variation of  $\hat{\theta}_n$ :  $\frac{1}{n-1} \sum_{i=1}^n (\hat{\theta}_{-i} - \bar{\theta}_n)^2$ ,  $\bar{\theta}_n = \frac{1}{n} \sum_{i=1}^n \hat{\theta}_{-i}$ . If  $\hat{\theta}_n = \bar{X}$  is the sample mean,  $\hat{\theta}_{-i} - \bar{\theta}_n = (n-1)^{-1} (\bar{X} - X_i)$  and  $\frac{1}{n-1} \sum_{i=1}^n (\hat{\theta}_{-i} - \bar{\theta}_n)^2 = \frac{S^2}{(n-1)^2}$ . Thus, the correction factor  $(n-1)^2/n$  should be multiplied, which leads to the jackknife variance estimator of  $\text{Var}(\hat{\theta}_n)$ :  $\hat{V}_J = \frac{n-1}{n} \sum_{i=1}^n (\hat{\theta}_{-i} - \bar{\theta}_n)^2$ .

**Theorem 1:** Let  $X_1, \dots, X_n$  be i.i.d. random  $d$ -vectors from  $F$  with finite  $\mu = \mathbb{E}(X_1)$  and  $\text{Var}(X_1)$ , and let  $\hat{\theta}_n = g(\bar{X})$ . Suppose that  $\nabla g$  is continuous at  $\mu$  and  $\nabla g(\mu) \neq 0$ . Then the jackknife variance estimator  $\hat{V}_J$  is strongly consistent for  $\text{Var}(\hat{\theta}_n)$ .



## CONFIDENCE SETS

**Definition 2 (Bootstrap):** Create bootstrap pseudo-replicate datasets  $X^{*1}, \dots, X^{*B}$  randomly generated from  $X$ . Let  $\hat{\theta}^{*b}$  be the same as an estimator  $\hat{\theta}$  but based on  $X^{*b}, b = 1, \dots, B$ . Is  $\frac{1}{B} \sum_{b=1}^B (\hat{\theta}^{*b} - \frac{1}{B} \sum_{l=1}^B \hat{\theta}^{*l})^2$  still a valid estimator of  $\text{Var}(\hat{\theta})$ ? In fact, the cdf  $G(t) = P(\hat{\theta} - \theta \leq t)$  can be estimated as  $\frac{1}{B} \sum_{b=1}^B I(\hat{\theta}^{*b} - \hat{\theta} \leq t) = \frac{\# \text{ of } b\text{'s such that } \hat{\theta}^{*b} - \hat{\theta} \leq t}{B}$ .

**Remark 1 (A heuristic description for the bootstrap):**  $\mathcal{P}$ : the population producing data  $X$ .  $\hat{\mathcal{P}}$ : an estimated of the population based on data  $X$ .  $X^*$ : the bootstrap data produced by  $\mathcal{P}$ . Real world:  $\mathcal{P} \Rightarrow X \Rightarrow \hat{\theta} = \hat{\theta}(X)$ . Bootstrap:  $\hat{\mathcal{P}} \Rightarrow X^* \Rightarrow \hat{\theta}^* = \hat{\theta}(X^*)$ . If  $\mathcal{P}$  is close to  $\hat{\mathcal{P}}$ , then  $\hat{G}(t)$  is close to  $G(t)$  and  $\text{Var}^*(\hat{\theta}^*)$  is close to  $\text{Var}(\hat{\theta})$ .

**Definition 3 (Parametric bootstrap):** Let  $X_1, \dots, X_n$  be i.i.d. with a c.d.f.  $F_\theta$  where  $\theta$  is an unknown parameter vector and  $F_\theta$  is known when  $\theta$  is known. Let  $\hat{\theta}$  be an estimator of  $\theta$  based on  $X = (X_1, \dots, X_n)$ . Parametric bootstrap data set  $X^* = (X_1^*, \dots, X_n^*)$  is obtained by generating i.i.d.  $X_1^*, \dots, X_n^*$  from  $F_{\hat{\theta}}$ .

**Example 1 (Location-scale problems):** Let  $F_\theta(x) = F_0(\frac{x-\mu}{\sigma})$ , where  $\mu = \mathbb{E}(X_1), \sigma^2 = \text{Var}(X_1)$  and  $F_0$  is a known cdf. Let  $\bar{X}$  be the sample mean,  $S^2$  be the sample variance, and  $T = \frac{\sqrt{n}(\bar{X}-\mu)}{S} = \frac{\sqrt{n} \sum_{i=1}^n \frac{X_i - \mu}{S}}{S}$ . The distribution of  $T$  does not depend on any parameter. Let  $\hat{\theta} = (\bar{X}, S^2)$  generate i.i.d.  $X_i^*, i = 1, \dots, n$  from  $F_{\hat{\theta}}$ . Then  $(X_i^* - \bar{X})/S \sim F_0, T^* = \sqrt{n} \sum_{i=1}^n \frac{X_i^* - \bar{X}}{S} \sim T$ .

**Definition 4 (Nonparametric bootstrap):** Without any model, we can apply the simple nonparametric bootstrap. If  $X = (X_1, \dots, X_n), X_1, \dots, X_n$  are i.i.d., then  $\mathcal{P}$  is the cdf of  $X_1$  and  $\hat{\mathcal{P}}$  is the empirical cdf based on  $X_1, \dots, X_n$ . If we generate i.i.d. bootstrap data  $X_1^*, \dots, X_n^*$  from  $\hat{\mathcal{P}}$ , then it is the same as taking a simple random sample with replacement from  $X$ .

**Proposition 1 (Property of  $\text{Var}^*(\hat{\theta}^*)$ ):** Consider the estimation of  $g(\mu)$ , where  $\mu = \mathbb{E}(X_1)$  and  $g$  is a continuously differentiable function. Our estimator is  $\hat{\theta} = g(\bar{X})$ . The bootstrap analog is  $\hat{\theta}^* = g(\bar{X}^*)$ . When  $n$  is large,  $g(\bar{X}^*) \approx g(\bar{X}) + \nabla g(\bar{X})(\bar{X}^* - \bar{X})$ , and  $\text{Var}^*(\hat{\theta}^*) = \text{Var}[g(\bar{X}^*)] \approx \nabla g(\bar{X}) \text{Var}(\bar{X}^* - \bar{X}) \nabla g(\bar{X})^T \approx \frac{n-1}{n^2} \nabla g(\bar{X}) S^2 \nabla g(\bar{X})^T$ .

## 7.6 Bootstrap confidence intervals

**Remark 1:** We want to find limits  $\underline{\theta}$  and  $\bar{\theta}$  such that  $P(\underline{\theta} \leq \theta) = P(\theta \leq \bar{\theta}) = 1 - \alpha$ . Traditional asymptotic approach is  $G(t) \rightarrow \Phi(t/\sigma), n \rightarrow \infty$  where  $\Phi$  is the standard normal cdf,  $\sigma$  is an unknown scale parameter. Let  $\hat{\sigma}$  be a consistent estimator of  $\sigma$ , then  $G(\hat{\sigma}t) \rightarrow \Phi(t)$ . Normal approximation  $100(1 - \alpha)\%$  confidence limits are  $\underline{\theta}_N = \hat{\theta} - \hat{\sigma} z_{1-\alpha}/\sqrt{n}, \bar{\theta}_N = \hat{\theta} + \hat{\sigma} z_{1-\alpha}/\sqrt{n}$ .

**Definition 1 (Hybrid bootstrap):** A bootstrap estimator of  $G(t) = P(\sqrt{n}(\hat{\theta} - \theta) \leq t)$  is  $\hat{G}(t) = P_*(\sqrt{n}(\hat{\theta}^* - \hat{\theta}) \leq t)$ .  $G^{-1}(1 - \alpha)$  can be estimated by  $\hat{G}^{-1}(1 - \alpha)$ . HB lower and upper confidence limits:  $\underline{\theta}_{\text{HB}} = \hat{\theta} - \hat{G}^{-1}(1 - \alpha)/\sqrt{n}, \bar{\theta}_{\text{HB}} = \hat{\theta} - \hat{G}^{-1}(\alpha)/\sqrt{n}$ .

**Definition 2 (Bootstrap-t):**  $G(t) = P(\sqrt{n}(\hat{\theta} - \theta) \leq t) = \Phi(t/\sigma)$ .  $\hat{\sigma}$ : a consistent estimator of  $\sigma$ . Then  $H(t) = P(\sqrt{n}(\hat{\theta} - \theta)/\hat{\sigma} \leq t) \rightarrow \Phi(t)$ . A bootstrap estimator of  $H(t)$ :  $\hat{H}(t) = P_*(\sqrt{n}(\hat{\theta}^* - \hat{\theta})/\hat{\sigma}^* \leq t)$ . BT lower and upper confidence limits:  $\underline{\theta}_{\text{BT}} = \hat{\theta} - \hat{\sigma} \hat{H}^{-1}(1 - \alpha)/\sqrt{n}, \bar{\theta}_{\text{BT}} = \hat{\theta} + \hat{\sigma} \hat{H}^{-1}(1 - \alpha)/\sqrt{n}$ .

**Definition 3 (Bootstrap percentile):** Bootstrap distribution (histogram):  $K(t) = P_*(\hat{\theta}^* \leq t) \approx \frac{1}{B} (\# \text{ of times } \hat{\theta}^{*b} \leq t)$ .  $100(1 - \alpha)\%$  BP lower confidence limit  $\underline{\theta}_{\text{BP}} = K^{-1}(\alpha) = \inf\{t : K(t) \leq \alpha\}$ , i.e.,  $\underline{\theta}_{\text{BP}} \approx \alpha B$ th ordered valued of  $\hat{\theta}^{*1}, \dots, \hat{\theta}^{*B}$ .  $100(1 - \alpha)\%$  BP upper confidence limit  $\bar{\theta}_{\text{BP}} = K^{-1}(1 - \alpha) = \inf\{t : K(t) \geq 1 - \alpha\}$ , i.e.,  $\bar{\theta}_{\text{BP}} \approx (1 - \alpha)B$ th ordered value of  $\hat{\theta}^{*1}, \dots, \hat{\theta}^{*B}$ .

## CONFIDENCE SETS

**Assumption 1:** There is a monotone transformation  $\phi$  such that  $P(\hat{\phi} - \phi \leq t) = \Psi(t)$  for all  $t$  and all  $P$  (including  $P = P_*$ ), where  $\phi = \phi(\theta)$ ,  $\hat{\phi} = \phi(\hat{\theta})$  and  $\Psi$  is a continuous, increasing cdf symmetric about 0.

**Theorem 1:** If  $\phi$  is known, then  $\underline{\theta}_E = \phi^{-1}(\hat{\phi} + z_\alpha)$ ,  $z_\alpha = \Psi^{-1}(\alpha)$  is an exact  $100(1 - \alpha)\%$  lower confidence bound, i.e.,  $P(\underline{\theta}_E \leq \theta) = 1 - \alpha$ . Under assumption 1,  $\underline{\theta}_{BP} = \underline{\theta}_E$ .

**Assumption 2:** There is a monotone transformation  $\phi$  and a constant  $z_0$  such that  $P(\hat{\phi} - \phi + z_0 \leq t) = \Psi(t)$  for all  $t$  and all  $P$  (including  $P = P_*$ ), where  $\Psi$  is a continuous, increasing cdf symmetric about 0 and  $z_0$  is a constant that may depend on  $P$  and  $n$ .

**Theorem 2:**  $\underline{\theta}_E = \phi^{-1}(\hat{\phi} + z_\alpha + z_0)$  is an exact  $100(1 - \alpha)\%$  lower confidence limit for  $\theta$ . Under assumption 2, the bootstrap bias-corrected percentile  $\underline{\theta}_{BC} := K^{-1}(\Psi(z_\alpha + 2z_0)) = K^{-1}(\Psi(z_\alpha + 2\Psi^{-1}(K(\hat{\theta})))) = \underline{\theta}_E$ .

**Assumption 3:** There is a monotone transformation  $\phi$  and constants  $z_0$  and  $a$  (acceleration constant) such that  $P(\frac{\hat{\phi} - \phi}{1 + a\phi} + z_0 \leq t) = \Psi(t)$  for all  $t$  and all  $P$  (including  $P = P_*$ ), where  $\Psi$  is a continuous, increasing cdf symmetric about 0.

**Theorem 3:**  $\underline{\theta}_E = \phi^{-1}(\hat{\phi} + \frac{(z_\alpha + z_0)(1 + a\hat{\phi})}{1 - a(z_\alpha + z_0)})$  is an exact  $1 - (1 - \alpha)\%$  lower confidence limit for  $\theta$ . Under assumption 3, if  $a$  is known, the  $BC_a$  lower confidence bound  $\underline{\theta}_{BC_a} = K^{-1}(\Psi(z_0 + \frac{z_\alpha + z_0}{1 - a(z_\alpha + z_0)})) = \underline{\theta}_E$ .

**Definition 4 (Asymptotic accuracy):** A confidence set  $C$  is first order accurate if  $P(\theta \in C) = 1 - \alpha + O(n^{-1/2})$  and second order accurate if  $P(\theta \in C) = 1 - \alpha + O(n^{-1})$ .

**Theorem 5:** For the case of  $\hat{\theta}$  is a smooth function of sample means, we have shown the following summary: (i) The BY and bootstrap  $BC_a$  one-sided confidence intervals are second order accurate. (ii) The BP, BC, HB, and NA one-sided confidence intervals are in general first order accurate. (iii) The equal-tail two-sided confidence intervals produced by all five bootstrap methods and the normal approximation are second order accurate (errors cancel each other).