Advanced Theory of Statistics

Lectured by Wang Miao LaTeXed by Chengxin Gong

2022 年 11 月 16 日

目录

1 Probability Theory

1.1 Measure space, measurable function, and integration

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

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

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

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

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 ∞/∞ 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 $\mathscr{F}$ = all subsets of Ω and $\nu(A)$ = the number of elements in $A \in \mathscr{F}$ ($\nu(A) = \infty$ if A contains infinitely many elements). Then ν is a measure on $\mathscr{F}$ and is called the counting measure. (c) There is a unique measure m on $(\mathbb{R}, \mathscr{B})$, that satisfies $m([a, b]) = b - a$ for every finite interval $[a, b]$, $-\infty < a \le b < \infty$. This is called the Lebesgue measure.

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

Definition 3: Let P be a probability measure on $(\mathbb{R}, \mathscr{B})$. 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) $F(-\infty) = \lim_{x\to-\infty} F(x) = 0$; (b) $F(\infty) = \lim_{x\to\infty} F(x) = 1$; (c) F is nondecreasing, i.e. $F(x) \le F(y)$ if $x \le y$; (d) F is right continuous, i.e. $\lim_{y\to x+0} F(y) = F(x)$. (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 meausre on $(\mathbb{R}, \mathscr{B})$.

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

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

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

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

Definition 7 (Integration): (a) The integral of a nonnegative simple function ϕ w.r.t.ν 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 $\mathscr{S}_f$ 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. ν is defined as $\int f d\nu = \sup\{\int \phi d\nu : \phi \in \mathscr{S}_f\}$ (Hence, for any Borel function $f \geq 0$, there exists aa sequence of simple functions $\phi_1, \phi_2, \cdots$ such that $0 \leq \phi_i \leq f$ for all i and $\lim_{n\to\infty} \int \phi_n d\nu = \int f d\nu$). (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 $(\Omega, \mathscr{F}, \nu)$ to $(\bar{\mathbb{R}}, \bar{\mathscr{B}})$, where $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}, \bar{\mathscr{B}} = \sigma(\mathscr{B} \cup \{\infty, -\infty\})$. Let $A_+ = \{f = \infty\}$ and $A_- = \{f = -\infty\}$. If $\nu(A_+) = 0$, we define $\int f_+ d\nu$ to be $\int I_{A_+^c} f_+ d\nu$; 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: $(\Omega, \mathscr{F}, \nu)$ 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 itegrals exist. (ii) If $f \geq 0$ a.e. and $\int f d\nu = 0$, then $f = 0$ a.e.

Theorem 1: Let $f_1, f_2, \cdot$ be a sequence of Borel functions on $(\Omega, \mathscr{F}, \nu)$. (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 $\lim_{n\to\infty} f_n = f$ a.e. and $|f_n| \leq g$ a.e. for integrable g, then $\int \lim_{n\to\infty} f_n d\nu = \lim_{n\to\infty} \int f_n d\nu$. (iii) Monotone convergence theorem: If $0 \leq f_1 \leq f_2 \leq \cdots$ and $\lim_{n\to\infty} f_n = f$ a.e., then $\int \lim_{n\to\infty} f_n d\nu = \lim_{n\to\infty} \int f_n d\nu$.

Example 1 (Interchange of differentiation and integration): Let $(\Omega, \mathscr{F}, \nu)$ be a measure space and, for any fixed $\theta \in \mathbb{R}$, let $f(\omega, \theta)$ be a Borel function on Ω. 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 Ω. 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 $(\Omega, \mathscr{F}, \nu)$ to $(\Lambda, \mathscr{G})$ and g be Borel on $(\Lambda, \mathscr{G})$. 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 ν_i be a σ-finite measure on $(\Omega_i, \mathscr{F}_i), i = 1, 2$, and f be a Borel function on $\prod_{i=1}^{2}(\Omega_i, \mathscr{F}_i)$ 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. ν_2 and defines a Borel function on Ω_2 whose integral w.r.t. ν_2 exists, and $\int_{\Omega\times\Omega} f(\omega_1, \omega_2) d\nu_1 \times \nu_2 = \int_{\Omega_2} [\int_{\Omega_1} f(\omega_1, \omega_2) d\nu_1] d\nu_2$.

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

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

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 ν be a σ-finite measure on a measure space $(\Omega, \mathscr{F})$. (i) If λ is a measure, $\lambda << \nu$, 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 $\lambda_i << \nu$, then $\lambda_1 + \lambda_2 << \nu$ and $\frac{d(\lambda_1+\lambda_2)}{d\nu} = \frac{d\lambda_1}{d\nu} + \frac{d\lambda_2}{d\nu}$ a.e. ν. (iii) If τ is a measure, λ is a σ-finite measure, and $\tau << \lambda << \nu$, then $\frac{d\tau}{d\nu} = \frac{d\tau}{d\lambda}\frac{d\lambda}{d\nu}$ a.e. ν. In particular, if $\lambda << \nu$ and $\nu << \lambda$ (in which case λ and ν are equivalent), then $\frac{d\lambda}{d\nu} = (\frac{d\nu}{d\lambda})^{-1}$ a.e. ν or λ. (iv) Let $(\Omega_i, \mathscr{F}_i, \nu_i)$ be a measure space and ν_i be σ-finite, $i = 1, 2$. Let λ_i be a σ-finite measure on $(\Omega, \mathscr{F}_i)$ and $\lambda_i << \nu_i, i = 1, 2$. Then $\lambda_1 \times \lambda_2 << \nu_1 \times \nu_2$ 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 $(\Omega, \mathscr{F}, P)$ 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 is $F_Y(x) = \begin{cases} 1 & x \geq c \\ F_X(x) & x < c \end{cases}$. This c.d.f. is discontinuous at c, since $F_x(c) < 1$. 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 $(\mathbb{R}, \mathscr{B})$: $\delta_c(A) = \begin{cases} 1 & c \in A \\ 0 & c \notin A \end{cases}, A \in \mathscr{B}$. Then $P_Y << m + \delta_c$, and the p.d.f. of P_Y is $f_Y(x) = \frac{dP_Y}{d(m+\delta_c)}(x) = \begin{cases} 0 & x > c \\ 1 - F_X(c) & x = c \\ f_X(x) & x < c \end{cases}$. To show this, it suffices to show that $\int_{(-\infty,x]} f_Y(t) d(m + \delta_c) = P_Y((-\infty, x])$ for any $x \in \mathscr{B}$.

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 $(\mathbb{R}^k, \mathscr{B}^k)$ to $(\mathbb{R}^k, \mathscr{B}^l)$. Let $A_1, \cdots, A_m$ be disjoint sets in $\mathscr{B}^k$ such that $\mathscr{R}^k - (A_1 \cup \cdots \cup A_m)$ 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 $A_j, j = 1, \cdots, m$. 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 $A_j, j = 1, \cdots, m$.

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

square distributions $\chi^2_{n_1}$ and $\chi^2_{n_2}$, 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 χ^2_n. 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 $X_1, \cdots, X_n$ be independent random variables and $X_i \sim N(\mu_i, \sigma^2)$. The distribution of $Y = (X_1^2 + \cdots + X_n^2)/\sigma^2$ is called the noncentral chi-square distribution and denoted by $\chi^2_n(\delta)$, where $\delta = (\mu_1^2 + \cdots + \mu_n^2)/\sigma^2$ is the noncentrality parameter. If $Y_1, \cdots, Y_k$ are independent random variables aand Y_i has the noncentral independent chi-square distribution $\chi^2_{n_i}(\delta_i), i = 1, \cdots, k$, then $Y = Y_1 + \cdots + Y_k$ has the noncentral chi-square distribution $\chi^2_{n_1+\cdots+n_k}(\delta_1 + \cdots + \delta_k)$.

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. $\mathrm{Var}(X) = \mathbb{E}(X-\mathbb{E}X)^2$ 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, $\mathrm{Var}(X) = \mathbb{E}(X-\mathbb{E}X)(X-\mathbb{E}X)^T$ 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 $\mathrm{Cov}(X_i, X_j)$. If $\mathrm{Cov}(X_i, X_j) = 0$, then X_i and X_j aare 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 $\mathrm{Var}(c^T X) = c^T \mathrm{Var}(X)c$.

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 $X_1, \cdots, X_k$, $\psi_{\sum_i X_i}(t) = \prod_i \psi_{X_i}(t)$ and $\phi_{\sum_i X_k}(t) = \prod_i \phi_{X_i}(t), t \in \mathbb{R}^k$. For $X = (X_1, \cdots, X_k)$ with m.g.f. ψ_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 $\mathbb{E}|X_1^{r_1} \cdots X_k^{r_k}| < \infty$ for nonnegative integers $r_1, \cdots, r_k$, 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 $(\Omega, \mathscr{F}, P)$. (i) The conditional expectation of X given $\mathscr{A}$ (a sub-σ-field of $\mathscr{F}$), denoted by $\mathbb{E}(X|\mathscr{A})$, is the a.s. unique random variable satisfying the following two conditions: (a) $\mathbb{E}(X|\mathscr{A})$ is a measurable from $(\Omega, \mathscr{A})$ to $(\mathbb{R}, \mathscr{B})$; (b) $\int_A \mathbb{E}(X|\mathscr{A})dP = \int_A XdP$ for any $A \in \mathscr{A}$. (ii) The conditional probability of $B \in \mathscr{F}$ given $\mathscr{A}$ is defined to be $P(B|\mathscr{A}) = \mathbb{E}(I_B|\mathscr{A})$. (iii) Let Y be measurable from $(\Omega, \mathscr{F}, P)$ to $(\Lambda, \mathscr{G})$. The conditionala 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 $(\Omega, \mathscr{F})$ to $(\Lambda, \mathscr{G})$ and Z a function from $(\Omega, \mathscr{F})$ to $\mathbb{R}^k$. Then Z is measurable from $(\Omega, \sigma(Y))$ to $(\mathbb{R}^k, \mathscr{B}^k)$ iff there is a measurable function h from $(\Lambda, \mathscr{G})$ such that $Z = h \circ Y$.

Example 1: Let X be an integrable random variable on $(\Omega, \mathscr{F}, P)$, $A_1, A_2, \cdots$ be disjoint events on $(\Omega, \mathscr{F}, P)$ such that $\cup A_i = \Omega$ and $P(A_i) > 0$ for all i, and let $a_1, a_2, \cdots$ be distinct real numbers. Define $Y = a_1 I_{A_1} + a_2 I_{A_2} + \cdots$. 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 ν and λ are σ-finite measures on $(\mathbb{R}^n, \mathscr{B}^n)$ and $(\mathbb{R}^m, \mathscr{B}^m)$, 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 marginl p.d.f. of Y w.r.t. λ.

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

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

Theorem 2: Let $\mathscr{C}_i, i \in \mathscr{I}$ be independent collections of events. If each $\mathscr{C}_i$ is a π-system, then $\sigma(\mathscr{C}_i), i \in \mathscr{I}$ 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 $X, X_1, X_2, \cdots$ 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 $X_n \to_{\text{a.s.}} X$ iff $\lim_{n\to\infty} X_n = X$ a.s. (ii) We say that $\{X_n\}$ converges to X in probability and write $X_n \to_p X$ iff for every fixed $\epsilon > 0$, $\lim_{n\to\infty} P(||X_n - X|| > \epsilon) = 0$. (iii) We say that $\{X_n\}$ converges to X in L_r (or in rth moment) with a fixed $r > 0$ and write $X_n \to_{L_r} X$ iff $\lim_{n\to\infty} \mathbb{E}||X_n - X||_r^r = 0$. (iv)

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

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

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

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 $A_1, A_2, \cdots$ re pairwise independent aaand $\sum_{n=1}^{\infty} P(A_n) = \infty$, then $P(\limsup_n A_n) = 1$.

Definition 2: Let $X_1, X_2, \cdots$ be random vectors and $Y_1, Y_2, \cdots$ 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 $X_n/Y_n \to_{\text{a.s.}} 0$. (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 $X_n/Y_n \to_p 0$.

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

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 $X, X_1, X_2, \cdots$ are positive random variables. Then $X_n \to_{\text{a.s.}} X$ iff for every $\epsilon > 0$, $\lim_{n\to\infty} P(\sup_{k\geq n} \frac{X_k}{X} > 1 + \epsilon) = 0$, and $\lim_{n\to\infty} P(\sup_{k\geq n} \frac{X}{X_k} > 1 + \epsilon) = 0$. (ii) Suppose $X, X_1, X_2, \cdots$ 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 $X_n \to_{\text{a.s.}} X$.

1.6 Uniform integrability and weak convergence

Definition 1 (Tightness): A sequence $\{P_n\}$ of probability measure on $(\mathbb{R}^k, \mathscr{B}^k)$ 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 $(\mathbb{R}^k, \mathscr{B}^k)$. (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 $(\mathbb{R}^k, \mathscr{B}^k)$ such that $P_{n_j} \to_w P$ as $j \to \infty$. (ii) If $\{P_n\}$ is tight and if each subsequence that converges weakly at all converges to the same probability measure P, then $P_n \to_w P$.

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

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

Example 1: Let $X_1, \cdots, X_n$ be independent random variables having a common c.d.f. and $T_n = X_1 + \cdots + X_n, n = 1, 2, \cdots$. 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 $|t| \to 0$, where $\mu = \mathbb{E}X_1$. Then, the ch.f. of T_n/n is $\phi_{T_n/n}(t) = [\phi_{X_1}(\frac{t}{n})]^n = [1 + \frac{\sqrt{-1}\mu t}{n} + o(\frac{t}{n})]^n \to e^{\sqrt{-1}\mu t}$ for any $t \in \mathbb{R}$ as $n \to \infty$. $e^{\sqrt{-1}\mu t}$ is the ch.f. of the point mass probability measure at μ. Thus $T_n/n \to_d \mu$ and $T_n/n \to_p \mu$.

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

1.7 Convergence of transformations and law of large numbers

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

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

Theorem 3: Let $X_1, X_2, \cdots$ and $Y = (Y_1, \cdots, Y_k)$ be random k-vectors satisfying $a_n(X_n - c) \to_d Y$, where $c \in \mathbb{R}^k$ and $\{a_n\}$ is a sequence of positive numbers with $\lim_{n\to\infty} a_n = \infty$. Let g be a function from $\mathbb{R}^k \to \mathbb{R}$. (i) If g is differentiable at c, then $a_n[g(X_n) - g(c)] \to_d [\nabla g(c)^T]Y$, 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, $1 \le j \le m-1$, vanishing at c, but with the mth-order partial derivatives not all vanishing at c. Then $a_n^m[g(X_n) - g(c)] \to_d \frac{1}{m!} \sum_{i_1=1}^k \cdots \sum_{i_m=1}^k \frac{\partial^m g}{\partial x_{i_1} \cdots \partial x_{i_m}}|_{x=c} Y_{i_1} \cdots Y_{i_m}$.

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

Theorem 5: Let $X_1, X_2, \cdots$ 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 $\frac{1}{n}\sum_{i=1}^n X_i - a_n \to_p 0$ is that $nP(|X_1| > n) \to 0$, in which case we may take $a_n = \mathbb{E}(X_1 1_{\{|X_1| \le n\}})$. (ii) The SLLN. A necessary and sufficient condition for the existence of a constant c for which $\frac{1}{n}\sum_{i=1}^n X_i \to_{\text{a.s.}} c$ is that $\mathbb{E}|X_1| < \infty$, in which case $c = \mathbb{E}X_1$ and $\frac{1}{n}\sum_{i=1}^n c_i(X_i - \mathbb{E}X_1) \to_{\text{a.s.}} 0$ for any bounded sequence of real numbers $\{c_i\}$.

Theorem 6: Let $X_1, X_2, \cdots$ 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 $\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}X_i) \to_{\text{a.s.}} 0$. (ii) The WLLN. If there is a constant $p \in [1, 2]$ such that $\lim_{n\to\infty} \frac{1}{n^p}\sum_{i=1}^n \mathbb{E}|X_i|^p = 0$, then $\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}X_i) \to_p 0$.

1.8 The central limit theorem

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

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

Theorem 3 (Berry-Esséen bound): For i.i.d. $\{X_n\}$ and $W_n = \sqrt{n}(\bar{X} - \mu)/\sigma$, $\sup_t |F_{W_n}(t) - \phi(t)| \leq \frac{33}{4}\frac{\mathbb{E}|X_1-\mu|^3}{\sigma^3\sqrt{n}}, n = 1, 2, \cdots$. Thus, the convergence speed of F_{W_n} to ϕ 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_θ on $(\Omega, \mathscr{F})$ 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_θ is a known probability measure when θ is known. The set Θ is called the parameter space and d is called its dimension. $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ 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 $\mathscr{P}$ is dominated by ν (a σ-finite measure) if $P << \nu$ for all $P \in \mathscr{P}$, in which case $\mathscr{P}$ can be identified by the family of densities $\{\frac{dP}{d\nu} : P \in \mathscr{P}\}$ or $\{\frac{dP_\theta}{d\nu} : \theta \in \Theta\}$.

Definition 3 (Exponential families): A parametric family $\{P_\theta : \theta :\in \Theta\}$ dominated by a σ-finite measure ν on $(\Omega, \mathscr{F})$ is called on 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 $\mathscr{P}$ be a natural exponential family. (i) Let $T = (Y, U)$ and $\eta = (\theta, \phi)$, Y and θ 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 σ-finite measure depending on ϕ. 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 σ-finite measure depending on u. (ii) If η_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 neighbbrhood 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 $(\mathbb{R}^k, \mathscr{B}^k)$, $\mathscr{V} \subset \mathbb{R}^k$, and $\mathscr{M}_k$ be a collection of $k \times k$ symmetric positive definite matrices. The family $\{P_{(\mu,\Sigma)} : \mu \in \mathscr{V}, \Sigma \in \mathscr{M}_k\}$ is called a location-scale family (on $\mathbb{R}^k$), where $P_{(\mu,\Sigma)}(B) = P(\Sigma^{-1/2}(B - \mu)), B \in \mathscr{B}^k$. The parameters μ 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 $X_1, \cdots, X_n$ 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 $X = (X_1, \cdots, X_n)$ with i.i.d. random components. Let $X_{(i)}$ be the ith smallest value of $X_1, \cdots, X_n$. The statistics $X_{(1)}, \cdots, X_{(n)}$ are called the order statistics.

2.2 Sufficiency and minimal sufficiency

Definition 1 (Sufficiency): Let X be a sample from an unknown population $P \in \mathscr{P}$, where $\mathscr{P}$ is a family of populations. A statistic $T(X)$ is said to be sufficient for $P \in \mathscr{P}$ iff conditional distribution of X given T is known.

Theorem 1 (The factorization theorem): Suppose that X is a sample from $P \in \mathscr{P}$ and $\mathscr{P}$ is a family of probability measures on $(\mathbb{R}^n, \mathscr{B}^n)$ dominated by a σ-finite measure ν. Then $T(X)$ is sufficient for $P \in \mathscr{P}$ 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 $\mathscr{P}$ is dominated by a σ-finite measure, then $\mathscr{P}$ 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 $P_i \in \mathscr{P}$.

Convention 1: If a statement holds except for outcomes in an event A satisfying $P(A) = 0$ for all $P \in \mathscr{P}$, then we say that the statement holds a.s. $\mathscr{P}$.

Definition 2 (Minimal sufficiency): Let T be a sufficient statistic for $P \in \mathscr{P}$. T is called a minimal sufficient statistic iff, for any other statistic S sufficient for $P \in \mathscr{P}$, there is a measurable function ψ such that $T = \psi(S)$ a.s. $\mathscr{P}$.

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

Theorem 4: Let $\mathscr{P}$ be a family of distributions on $\mathbb{R}^k$. (i) Suppose that $\mathscr{P}_0 \subset \mathscr{P}$ and a.s. $\mathscr{P}_0$ implies a.s. $\mathscr{P}$. If T is sufficient for $P \in \mathscr{P}$ and minimal sufficient for $P \in \mathscr{P}_0$, then T is minimal sufficient for $P \in \mathscr{P}$. (ii) Suppose that $\mathscr{P}$ contains p.d.f.'s $f_0, f_1, f_2, \cdots$ w.r.t. a σ-finite ν. 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 $f_\infty(x) > 0, i = 0, 1, 2, \cdots$. Then $T(x) = (T_0, T_1, T_2, \cdots)$ is minimal sufficient for $P \in \mathscr{P}$. 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 $T(x) = (T_1, T_2, \cdots)$ is minimal sufficient for $P \in \mathscr{P}$. (iii) Suppose that $\mathscr{P}$ contains p.d.f.'s f_p w.r.t. a σ-finite measure and that there exists a sufficient statistic $T(x)$ such that, for any possible values x and y of X, $f_p(x) = f_p(y)\phi(x, y)$ for all P implies $T(x) = T(y)$, where ϕ is a measurable function. Then $T(x)$ is minimal sufficient for $P \in \mathscr{P}$.

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 θ and, hence, a further simplification of T may still be needed.

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

Remark 2: If T is complete (or boundedly complete) and $S = \psi(T)$ for a measurable ψ, 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 $X_1, \cdots, X_n$ 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 $X_1, \cdots, X_n$ 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 $X = (X_1, \cdots, X_n)$ 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 η.

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

Theorem 1 (Basu's theorem): Let V and T be two statistics of X from a population $P \in \mathscr{P}$. If V is ancillary and T is boundedly complete and sufficient for $P \in \mathscr{P}$, then V and T are independent w.r.t. any $P \in \mathscr{P}$.

Example 3: $X_1, \cdots, X_n$ 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 θ. 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 $P \in \mathscr{P}$. Decision: an action we take after observing X. $\mathscr{A}$: the set of allowable actions. $(\mathscr{A}, \mathscr{F}_\mathscr{A})$: the action space. $\mathscr{X}$: the range of X. Decision rule: a measurable function T from $(\mathscr{X}, \mathscr{F}_\mathscr{X})$ to $(\mathscr{A}, \mathscr{F}_\mathscr{A})$. If $X = x$ is observed, then we take the action $T(x) \in \mathscr{A}$.

Definition 1 (Loss function): $L(P, a)$: a function from $\mathscr{P} \times \mathscr{A}$ 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 $R_T(P) := \mathbb{E}[L(P, T(X))] = \int_\mathscr{X} L(P, T(X))dP_X(x)$.

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 $P \in \mathscr{P}$ 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 $P \in \mathscr{P}$. Optimal rule: If T^* is as good as any other rule in $\mathscr{E}$, a claass of allowable decision rules, then T^* is $\mathscr{E}$-optimal.

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

Example 1: $X = (X_1, \cdots, X_n)$ is a vector of i.i.d. measurements for a parameter $\theta \in \mathbb{R}$. We want to estimate θ. Action space: $(\mathscr{A}, \mathscr{F}_{\mathscr{A}}) = (\mathbb{R}, \mathscr{B})$. A common loss function in this problem is the squared error loss $L(P, a) = (\theta - a)^2, a \in \mathscr{A}$. Let $T(X) = \bar{X}$, the sample mean. The loss for $\bar{X}$ is $(\bar{X} - \theta)^2$. If the population has mean μ 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 $\mathscr{P}$ be a family of distributions, $\mathscr{P}_0 \subset \mathscr{P}$, $\mathscr{P}_1 = \{P \in \mathscr{P} : P \notin \mathscr{P}_0\}$. A hypothesis testing problem can be formulated as that of deciding which of the following two statements is true: $H_0 : P \in \mathscr{P}_0$ versus $H_1 : P \in \mathscr{P}_1$. 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., $\mathscr{A} = \{0, 1\}$, 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 msut have the form $I_C(X)$, where $C \in \mathscr{F}_{\mathscr{X}}$ 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 $R_T(P) = \begin{cases} P(T(X) = 1) = P(X \in C) & P \in \mathscr{P}_0 \\ P(T(X) = 0) = P(X \notin C) & P \in \mathscr{P}_1 \end{cases}$.

Definition 6 (Admissibility): Let $\mathscr{E}$ be a class of decision rules. A decision rule $T \in \mathscr{E}$ is called $\mathscr{E}$-admissible iff there does not exist any $S \in \mathscr{E}$ 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 $P \in \mathscr{P}$ and let δ_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 δ_0 if $R_{\delta_0}(P) < \infty$ for any $P \in \mathscr{P}$.

Theorem 1: Suppose that $\mathscr{A}$ is a convex subset of $\mathbb{R}^k$ and that for any $P \in \mathscr{P}$, $L(P, a)$ is a convex function of a. (i) Let δ be a randomized rule satisfying $\int_{\mathscr{A}} ||a|| d\delta(x, a) < \infty$ for any $x \in \mathscr{X}$ and let $T_1(x) = \int_{\mathscr{A}} a d\delta(x, a)$. 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 $x \in \mathscr{X}$ and $P \in \mathscr{P}$. (ii) Rao-Blackwell theorem. Let T be a sufficient statistic for $P \in \mathscr{P}$, $T_0 \in \mathbb{R}^k$ 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 $P \in \mathscr{P}$. 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 θ of the unknown population is defined to be $b_T(P) = \mathbb{E}[T(X)] - \theta$. An estimator $T(X)$ is unbiased for θ iff $b_T(P) = 0$ for any $P \in \mathscr{P}$.

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

Approach 2: Consider some characteristic R_T of $R_T(P)$, for a given decision rule T, and then minimize R_T over $T \in \mathscr{E}$. 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 $H_0 : P \in \mathscr{P}_0$ versus $H_1 : P \in \mathscr{P}_1$, 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 $\mathscr{X}$ to $\{0, 1\}$.

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

Example 1: Let $X_1, \cdots, X_n$ be i.i.d. from the $\mathcal{N}(\mu, \sigma^2)$ distribution with an unknown $\mu \in \mathbb{R}$ and a known σ^2. Consider the hypotheses $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$, where μ_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, $\sup_{P \in \mathscr{P}_0} \alpha_{T_c}(\mu) = 1 - \phi(\frac{\sqrt{n}(c-\mu_0)}{\sigma})$. In fact, it is also true for $\sup_{P \in \mathscr{P}_1}[1 - \alpha_{T_c}(\mu)] = \phi(\frac{\sqrt{n}(c-\mu_0)}{\sigma})$. If we woudl like to use an α as the level of significance, then the most effective way is to choose a c_α such that $\alpha = \sup_{P \in \mathscr{P}_0} \alpha_{T_{c_\alpha}}(\mu)$, in which case c_α 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 $\sup_{P \in \mathscr{P}_0} \alpha_T(P) \leq \alpha$, $1 - \alpha_T(\mu) \geq 1 - \alpha_{T_{c_\alpha}}(\mu), \mu > \mu_0$.

Definition 4 (Significance tests): A common approach of finding an "optimal" test is to assign a small bound α to the type I error rate $\alpha_T(P), P \in \mathscr{P}_0$, and then to attempt to minimize the type II error rate $1 - \alpha_T(P), P \in \mathscr{P}_1$, subject to $\sup_{P \in \mathscr{P}_0} \alpha_T(P) \leq \alpha$. The bound α 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 a and a chosen test T_α, but also the smallest possible level of significance at which H_0 would be rejected for the computed $T_\alpha(x)$, i.e., $\hat{\alpha} = \inf\{\alpha \in (0, 1) : T_\alpha(x) = 1\}$. Such an $\hat{\alpha}$, which depends on x and the chosen test and is a statistic, is called the p-value for the test T_α.

Example 2: Let us calculate the p-value for T_{c_α} 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_α}. It turns out that $T_{c_\alpha}(x) = I_{(0,\alpha)}(\hat{\alpha}(X))$.

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

Example 3: Let $X_1, \cdots, X_n$ 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, $P(-\widetilde{c}_\alpha \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq \widetilde{c}_\alpha) = \sqrt{1-\alpha}$, where $\widetilde{c}_\alpha = \Phi^{-1}(\frac{1+\sqrt{1-\alpha}}{2})$. Since the χ^2 distribution distribution χ^2_{n-1} is a known distribution, we can always find two constants $c_{1\alpha}$ and $c_{2\alpha}$ such that $P(c_{1\alpha} \leq \frac{(n-1)S^2}{\sigma^2} \leq c_{2\alpha}) = \sqrt{1-\alpha}$. Then $P(-\widetilde{c}_\alpha \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq \widetilde{c}_\alpha, c_{1\alpha} \leq \frac{(n-1)S^2}{\sigma^2} \leq c_{2\alpha}) = 1-\alpha$. 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 $\sigma^2 = n(\bar{X}-\mu)^2/\widetilde{c}^2_\alpha$. This set is a confidence set for θ 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 $\alpha_T(P) = \mathbb{E}[T(X)], P \in \mathscr{P}_0$, and the type II error probability to be $1-\alpha_T(P) = \mathbb{E}[1-T(X)], P \in \mathscr{P}_1$. For a class of randomized tests, we would like to minimize $1-\alpha_T(P)$ subject to $\sup_{P\in\mathscr{P}_0} \alpha_T(P) = \alpha$.

Definition 8 (Consistency of point estimators): Let $X = (X_1, \cdots, X_n)$ be a sample from $P \in \mathscr{P}$, $T_n(X)$ be an estimator of θ for every n, and $\{a_n\}$ be a sequence of positive constants, $a_n \to \infty$. (i) $T_n(x)$ is consistent for θ iff $T_n(x) \to_p \theta$ w.r.t. any P. (ii) $T_n(x)$ is a_n-consistent for θ iff $a_n[T_n(X)-\theta] = O_p(1)$ w.r.t. any P. (iii) $T_n(x)$ is strongly consistent for θ iff $T_n(x) \to_{\text{a.s.}} \theta$ w.r.t. any P. (iv) $T_n(X)$ is L_r-consistent for θ iff $T_n(x) \to_{L_r} \theta$ 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 $\xi, \xi_1, \xi_2, \cdots$ be random variables and $\{a_n\}$ be a sequence of positive numbers satisfying $a_n \to \infty$ or $a_n \to a > 0$. If $a_n\xi_n \to_d \xi$ and $\mathbb{E}|\xi| < \infty$, then $\mathbb{E}\xi/a_n$ is called an asymptotic expectation of ξ_n. (ii) For a point estimator T_n of θ, an asymptotic expectation of $T_n - \theta$, if it exists,

is called an asymptotic bias of T_n and denoted by $\widetilde{b}_{T_n}(P)$. If $\lim_{n\to\infty}\widetilde{b}_{T_n}(P)=0$ 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 ξ_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 $b_n/a_n\to 0$; (c) $\mathbb{E}\xi\neq 0, \mathbb{E}\eta\neq 0$, and $(\mathbb{E}\xi/a_n)/(\mathbb{E}\eta/b_n)\to 1$.

Example 4 (Functions of sample means): We consider the case where $X_1,\cdots,X_n$ are i.i.d. random k-vectors with finite $\Sigma=\mathrm{Var}(X_1), T_n=g(\bar{X})$, 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, $2^{-1}n(\bar{X}-\mu)\nabla^2 g(\mu)(\bar{X}-\mu)\to_d 2^{-1}Z_\Sigma^T\nabla^2 g(\mu)Z_\Sigma$, where $Z_\Sigma=\mathcal{N}_k(0,\Sigma)$. Thus, $\frac{\mathbb{E}[Z_\Sigma^T\nabla^2 g(\mu)Z_\Sigma]}{2n}=\frac{\mathrm{tr}(\nabla^2 g(\mu)\Sigma)}{2n}$ 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 θ for every n and $\{a_n\}$ be a sequence of positive numbers satisfying $a_n\to\infty$ or $a_n\to a>0$. Assume that $a_n(T_n-\theta)\to_d Y$ with $0<\mathbb{E}Y^2<\infty$. (i) The asymptotic mean squared error of T_n, denoted by $\mathrm{amse}_{T_n}(P)$, is defined as the asymptotic expectation of $(T_n-\theta)^2$, $\mathrm{amse}_{T_n}(P)=\mathbb{E}Y^2/a_n^2$. The asymptotic variance of T_n is defined as $\sigma^2_{T_n}(P)=\mathrm{Var}(Y)/a_n^2$. (ii) Let T_n' be another estimator of θ. The asymptotic relative efficiency of T_n' w.r.t. T_n is defined as $e_{T_n',T_n}=\mathrm{amse}_{T_n}(P)/\mathrm{amse}_{T_n'}(P)$. (iii) T_n is said to be asymptotically more efficient than T_n' iff $\limsup_n e_{T_n',T_n}(P)\le 1$ for any P and <1 for some P.

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

Example 5: Let $X_1,\cdots,X_n$ 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 $\mathrm{mse}_{T_{1n}}(\theta)=e^{-\theta}(1-e^{-\theta})/n$. Also, $\sqrt{n}(T_{1n}-\theta)\to_d\mathcal{N}(0,e^{-\theta}(1-e^{-\theta}))$ by the CLT. Thus, in the case $\mathrm{amse}_{T_{1n}}(\theta)=\mathrm{mse}_{T_{1n}}(\theta)$. Consider $T_{2n}=e^{-\bar{X}}$. Note that $\mathbb{E}T_{2n}=e^{n\theta(e^{-1/n}-1)}$, hence $nb_{T_{2n}}(\theta)\to\theta e^{-\theta}/2$. Using the CLT, we can show that $\sqrt{n}(T_{2n}-\theta)\to_d\mathcal{N}(0,e^{-2\theta}\theta)$. Then $\mathrm{amse}_{T_{2n}}(\theta)=e^{-2\theta}\theta/n$. Thus, the asymptotic relative efficiency of T_{1n} w.r.t. T_{2n} is $e_{T_{1n},T_{2n}}=\theta/(e^\theta-1)<1$. 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 ϑ, then ϑ is called an estimable parameter.

Definition 2 (UMVUE): An unbiased estimator $T(X)$ of θ is called uniformly minimum variance unbiased estimator (UMVUE) iff $\mathrm{Var}(T(X))\le\mathrm{Var}(U(X))$ for any $P\in\mathscr{P}$ aand any other unbiased estimator $U(X)$ of θ.

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

statistic $T(X)$ for $P \in \mathscr{P}$. If θ is estimable, i.e., there is a unique unbiased estimator of θ, then there is a unique UMVUE of θ 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 θ. If $\mathbb{E}[h(T)] = \theta$ for all P, what should h be?

Example 1: Let $X_1, \cdots, X_n$ 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 θ 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 ϑ must satisfy $\theta^n g(\theta) = n\int_0^\theta h(x)x^{n-1}dx$ for all $\theta > 0$. Hence, the UMVUE of ϑ 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 θ, say $U(X)$. Conditioning on a sufficient and complete statistic $T(X)$: $\mathbb{E}[U(X)|T]$ is the UMVUE of θ. We need to derive an explicit form of $\mathbb{E}[U(X)|T]$.

Example 2: Let $X_1, \cdots, X_n$ 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 ϑ, $\mathbb{E}[1_{(t,\theta)}(X_1)] = P(X_1 > t) = \vartheta$. Hence $T(X) = \mathbb{E}[1_{(t,\infty)}(X_1)|\bar{X}] = P(X_1 > t|\bar{X})$ is the UMVUE of ϑ. 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 ϑ is $T(X) = (1 - \frac{t}{n\bar{X}})^{n-1}$.

Example 3: Let $X_1, \cdots, X_n$ be i.i.d. from an unknown population P in a nonparametric family $\mathscr{P}$. In many cases the vector of order statistics, $T = (X_{(1)}, \cdots, X_{(n)})$, is sufficient and complete for $P \in \mathscr{P}$. Note that an estimator $\phi(X_1, \cdots, X_n)$ is a function of T iff the function ϕ is symmetric in its n arguments. Hence, if T is sufficient and complete, then a symmetric unbiased estimator of any estimable ϑ 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 $P(X_1 \le t)$ for any fixed t. The previous conclusions are not true if T is not sufficient and complete for $P \in \mathscr{P}$.

Remark 1 (Nonexistence of any UMVUE): If $n > 2$ and $\mathscr{P}$ 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 $\mathscr{P} = \{1, \cdots, N\}$ be a finite population of interest. For each $i \in \mathscr{P}$, let y_i be a value of interest associated with unit i. Let $s = \{i_1, \cdots, i_n\}$ be a subset of distinct elements of $\mathscr{P}$, 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 $P(X_1 = y_{i_1}, \cdots, X_n = y_{i_n}) = p(s)/n!$. Let $\mathscr{Y}$ be the range of y_i, $\theta = (y_1, \cdots, y_N)$ and $\Theta = \prod_{i=1}^N \mathscr{Y}$. 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 $X_{(1)} \leq \cdots \leq X_{(n)}$ 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 θ, its unique UMVUE is the unbiased estimator $g(X_1, \cdots, X_n)$, 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 $\mathscr{U}$ be the set of all unbiased estimators of 0 with finite variances and T be an unbiased estimator of θ with $\mathbb{E}(T^2) < \infty$. (i) A necessary and sufficient condition for $T(X)$ to be a UMVUE of θ is that $\mathbb{E}[T(X)U(X)] = 0$ for any $U \in \mathscr{U}$ and any $P \in \mathscr{P}$. (ii) Suppose that $T = h(\widetilde{T})$, where $\widetilde{T}$ is a sufficient statistic for $P \in \mathscr{P}$ and h is a Borel function. Let $\mathscr{U}_{\widetilde{T}}$ be the subset of $\mathscr{U}$ consisting of Borel functions of $\widetilde{T}$. Then a necessary and sufficient condition for T to be a UMVUE of θ is that $\mathbb{E}[T(X)U(X)] = 0$ for any $U \in \mathscr{U}_{\widetilde{T}}$ and any $P \in \mathscr{P}$. 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 $\theta_j, j = 1, \cdots, k$, 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 $c_1, \cdots, c_k$. (ii) If T_1 and T_2 are two UMVUE's of θ, then $T_1 = T_2$ a.s. P for any $P \in \mathscr{P}$.

Example 1: Let $X_1, \cdots, X_n$ 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 θ 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 θ. We now illustrate how to use Theorem 1 to find a UMVUE of θ. 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 θ 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 $X = (X_1, \cdots, X_n)$ be a sample from $P \in \mathscr{P} = \{P_\theta : \theta \in \Theta\}$, where Θ 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 θ; P_θ has a p.d.f. f_θ w.r.t. a measure ν for all $\theta \in \Theta$; and f_θ is differentiable as a function of θ 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 $\mathrm{Var}(T(X)) \geq [\frac{\partial}{\partial\theta}g(\theta)]^T[I(\theta)]^{-1}\frac{\partial}{\partial\theta}g(\theta)$, 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 θ contained in (X, Y) is $I_x(\theta)+I_Y(\theta)$. (ii) Suppose that X has the p.d.f. f_θ that is twice differentiable in θ 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_θ 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 ψ is differentiable, then the Fisher information that X contains about η 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 σ-finite measure is $f_\theta(x) = \exp\{[\eta(\theta)]^T T(X) - \xi(\theta)\}c(x)$, where Θ 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 η, then the variance-covariance matrix $\mathrm{Var}(T) = I(\eta)$. (iii) If $I(\theta)$ is the Fisher information matrix for the parameter $\vartheta = \mathbb{E}[T(X)]$, then $\mathrm{Var}(T) = [I(\vartheta)]^{-1}$.

3.3 U- and V-statistics

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

Example 1: Consider the estimation of μ^m, where $\mu = \mathbb{E}X_1$ and m is an integer > 0. Using $h(x_1,\cdots,x_m) = x_1,\cdots x_m$, we obtain the following U-statistic for μ^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 $h(x_1,x_2) = (x_1-x_2)^2/2$: $U_n = \frac{2}{n(n-1)}\sum_{1\le i<j\le n}\frac{(X_i-X_j)^2}{2} = \frac{1}{n-1}(\sum_{i=1}^n X_i^2 - n\bar{X}^2) = S^2$, which is the sample variance.

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

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

Example 2: Consider $h(x_1,x_2) = x_1x_2$, the U-statistic unbiased for $\mu^2, \mu = \mathbb{E}X_1$. Note that $h_1(x_1) = \mu x_1, \widetilde{h}_1(x_1) = \mu(x_1-\mu)$. $\zeta_1 = \mathbb{E}[\widetilde{h}_1(X_1)]^2 = \mu^2\mathrm{Var}(X_1) = \mu^2\sigma^2, \widetilde{h}(x_1,x_2) = x_1x_2 - \mu^2$, and $\zeta_2 = \mathrm{Var}(X_1X_2) = (\mu^2+\sigma^2)^2 - \mu^4$. Thus for $U_n = (C_n^2)^{-1}\sum_{1\le i<j\le n}X_iX_j$, $\mathrm{Var}(U_n) = (C_n^2)^{-1}(C_2^1C_{n-2}^1\zeta_1 + C_2^2C_{n-2}^0\zeta_2) = \frac{2}{n(n-1)}[2(n-2)\mu^2\sigma^2 + (\mu^2+\sigma^2)^2 - \mu^4] = \frac{4\mu^2\sigma^2}{n} + \frac{2\sigma^4}{n(n-1)}$.

Remark 1 (Asymptotic distributions of U-statistics): For nonparametric $\mathscr{P}$, 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 $X_1, \cdots, X_n$. The projection of T_n on k_n random elements $Y_1, \cdots, Y_{k_n}$ is defined to be $\check{T}_n = \mathbb{E}(T_n) + \sum_{i=1}^{k_n}[\mathbb{E}(T_n|Y_i) - \mathbb{E}(T_n)]$.

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

Example 3: For a U-statistic U_n, one can show that $\check{U}_n = \mathbb{E}(U_n) + \frac{m}{n}\sum_{i=1}^n \widetilde{h}_1(X_i)$, where $\check{U}_n$ is the projection of U_n on $X_1, \cdots, X_n$ and $\widetilde{h}_1(x) = h_1(x) - \mathbb{E}[h(X_1, \cdots, X_m)], h_1(x) = \mathbb{E}[h(x, X_2, \cdots, X_m)]$. Hence, if $\zeta_1 = \text{Var}(\widetilde{h}_1(X_i)) > 0, \text{Var}(\check{U}_n) = m^2\zeta_1/n$ and $\mathbb{E}(U_n - \check{U}_n)^2 = O(n^{-2})$. If $\zeta_1 = 0$ but $\zeta_2 > 0$, then we can show that $\mathbb{E}(U_n - \check{U}_n)^2 = O(n^{-3})$. 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 $\mathbb{E}[h(X_1, \cdots, X_m)]^2 < \infty$. (i) If $\zeta_1 > 0$, then $\sqrt{n}[U_n - \mathbb{E}(U_n)] \to_d \mathcal{N}(0, m^2\zeta_1)$. (ii) If $\zeta_1 = 0$ but $\zeta_2 > 0$, then $n[U_n - \mathbb{E}(U_n)] \to_d \frac{m(m-1)}{2}\sum_{j=1}^\infty \lambda_j(\chi^2_{1j} - 1)$, where χ^2_{1j}'s are i.i.d. random variables having the chi-square distribution χ^2_1 and λ_j's are some constants (which may depend on P) satisfying $\sum_{j=1}^\infty \lambda_j^2 = \zeta_2$.

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

Definition 3 (V-statistics): Let $X_1, \cdots, X_n$ be i.i.d. from P. For every U-statistic U_n as an estimator $\theta = \mathbb{E}[h(X_1, \cdots, X_m)]$, there is a closely related V-statistic defined by $V_n = \frac{1}{n^m}\sum_{i_1=1}^n \cdots \sum_{i_m=1}^n h(X_{i_1}, \cdots, X_{i_m})$. As an estimator of θ, 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 $\mathbb{E}|h(X_{i_1}, \cdots, h_{i_m})| < \infty$ for all $1 \le i_1 \le \cdots \le i_m \le m$. Then the bias of V_n satisfies $b_{V_n}(P) = O(n^{-1})$. (ii) Assume that $\mathbb{E}[h(X_{i_1}, \cdots, X_{i_m})]^2 < \infty$ for all $1 \le i_1 \le \cdots \le i_m \le m$. 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 $\mathbb{E}[h(X_{i_1}, \cdots, X_{i_m})]^2 < \infty$ for all $1 \le i_1 \le \cdots \le i_m \le m$. (i) If $\zeta_1 = \text{Var}(h_1(X_1)) > 0$, then $\sqrt{n}(V_n - \theta) \to_d \mathcal{N}(0, m^2\zeta_1)$. (ii) If $\zeta_1 = 0$ but $\zeta_2 = \text{Var}(h_2(X_1, X_2)) > 0$, then $n(V_n - \theta) \to_d \frac{m(m-1)}{2}\sum_{j=1}^\infty \lambda_j\chi^2_{1j}$.

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

Definition 1 (Survey samples from a finite population): Let $\mathscr{P} = \{1, \cdots, N\}$ be a finite population of interest. For each $i \in \mathscr{P}$, let y_i be a value of interest associated with unit i. Let $s = \{i_1, \cdots, i_n\}$ be a subset of distinct elements of $\mathscr{P}$, 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 $i \in s, i = 1, \cdots, N$.

Theorem 1: (i) (Horvitz-Thompson). If $\pi_i > 0$ for $i = 1, \cdots, N$ and π_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 $j \in s, i = 1, \cdots, N, j = 1, \cdots, N$. 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 θ and U_n is a vector of unbiased estimators of components of θ,

then $T_n = g(U_n)$ is often asymptotically unbiased for ϑ. Note that $\mathbb{E}(T_n) = \mathbb{E}g(U_n)$ may not exists. Assume that g is differentiable and $c_n(U_n - \theta) \to_d Y$. 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 $X_1, \cdots, X_n$ 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 jth moment of P and let $\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^n X_i^j$ be the jth sample moment, which is an unbiased estimator of $\mu_j, j = 1, \cdots, k$. Typically, $\mu_j = h_j(\theta), j = 1, \cdots, k$, for some functions h_j on $\mathbb{R}^k$. By substituting μ_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 $\hat{\mu}_j = h_j(\hat{\theta}), j = 1, \cdots, k$. This method of deriving estimators is called the method of moments.

Example 1: Let $X_1, \cdots, X_n$ be i.i.d. from a population P_θ 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 $\mathscr{P} = P_\theta : \theta \in \Theta$, where $\theta \subset \mathbb{R}^k$ for a fixed integer $k \geq 1$. θ is viewed as a realization of a random vector $\theta \in \Theta$ whose prior distribution is Π. Prior distribution: past experience, past data, or a statistician's belief (subjective). Sample $X \in \mathscr{X}$: from $P_\theta = P_{x|\theta}$, the conditional distribution of X given θ. Posterior distribution: updated prior distribution using observed $X = x$.

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

Definition 2 (Bayes action): Let $\mathscr{A}$ be an action space in a decision problem and $L(\theta, a) \geq 0$ be a loss function. For any $x \in \mathscr{X}$, a Bayes action w.r.t. Π is any $\delta(x) \in \mathscr{A}$ such that $\mathbb{E}[L(\theta, \delta(x))|X = x] = \min_{a \in \mathscr{A}} \mathbb{E}[L(\theta, a)|X = x]$ 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 $\int_\Theta L(\theta, \delta(x)) f_\theta(x) d\Pi = \min_{a \in \mathscr{A}} \int_\Theta L(\theta, a) f_\theta(x) d\Pi$. This is still defined even if Π is not a probability measure but a σ-finite measure on Θ, in which case $m(x)$ may not be finite. If $\Pi(\Theta) \neq 1$, Π 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 ξ of parameters called hyperparameters. If the hyperparamters ξ is unknown, one way to solve the problem is to estimate ξ using some historical data; the resulting Bayes action is called an empirical Bayes action. If there is no historical data, we may estimate ξ using data x and the resulting Bayes action is also called an empirical Bayes action. The simplest empirical Bayes method is to

estimate ξ by viewing x as a "sample" from the marginal distribution $P_{x|\xi}(A) = \int_\Theta P_{x|\theta}(A)d\Pi_{\theta|\xi}, A \in \mathscr{B}_\mathscr{X}$, where $\Pi_{\theta|\xi}$ is a prior depending on ξ 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_θ. The method of moments can be applied to estimate ξ.

Example 1: Let $X = (X_1, \cdots, X_n)$ and X_i's be i.i.d. with an unknown mean $\mu \in \mathbb{R}$ and a known variance σ^2. Assume the prior $\Pi_{\mu|\xi}$ has mean μ_0 and variance σ_0^2, $\xi = (\mu_0, \sigma_0^2)$. To obtain a moment estimate of ξ, we need to calculate $\int_{\mathbb{R}^n} x_1 m(x)dx$ and $\int_{\mathbb{R}^n} x_1^2 m(x)dx, x = (x_1, \cdots, x_n)$. These two integrals can be obtained without knowing $m(x)$. Note that $\int_{\mathbb{R}^n} x_1 m(x)dx = \int_\Theta \int_{\mathbb{R}^n} x_1 f_\mu(x)dxd\Pi_{\mu|\xi} = \int_\mathbb{R} \mu d\Pi_{\mu|\xi} = \mu_0$ and $\int_{\mathbb{R}^n} x_1^2 m(x)dx = \int_\Theta \int_{\mathbb{R}^n} x_1^2 f_\mu(x)dxd\Pi_{\mu|\xi} = \sigma^2 + \int_\mathbb{R} \mu^2 d\Pi_{\mu|\xi} = \sigma^2 + \mu_0^2 + \sigma_0^2$. Thus, by viewing $x_1, \cdots, x_n$ 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 ξ and let Λ be a prior on Ξ, the range of ξ. Then the "marginal" prior for θ is defined by $\Pi(B) = \int_\Xi \Pi_{\theta|\xi}(B)d\Lambda(\xi), B \in \mathscr{B}_\Theta$. If the second-stage prior Λ 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 σ^2, the prior $\pi(\mu|\xi)$ is the p.d.f of $\mathcal{N}(\xi, \sigma_0^2)$ with a known σ_0^2, and the prior of ξ is $\mathcal{N}(\mu_0, \tau^2)$ with a known μ_0 and τ^2, then the marginal prior p.d.f of μ 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 Π. (i) If $\delta(X)$ is a unique Bayes rule, then $\delta(X)$ is admissible. (ii) If Θ is countable set, the Bayes risk $r_\delta(\Pi) < \infty$, and Π gives positive probability to each $\theta \in \Theta$, then $\delta(X)$ is admissible. (iii) Let $\mathscr{E}$ be the class of decision rules having continuous risk functions. If $\delta(X) \in \mathscr{E}, r_\delta(\Pi) < \infty$, and Π gives positive probability to any open subset of Θ, then $\delta(X)$ is $\mathscr{E}$-admissible.

Theorem 2: Suppose that Θ is an open set of $\mathbb{R}^k$. In a decision problem, let $\mathscr{E}$ be the class of decision rules having continuous risk functions. A decision rule $T \in \mathscr{E}$ is $\mathscr{E}$-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 $\eta > 0$, $\lim_{j\to\infty} \frac{r_T(\Pi_j) - r_j^*(\Pi_j)}{\Pi_j(O_{\theta_0,\eta})} = 0$, where $r_j^*(\Pi_j) = \inf_{T\in\mathscr{E}} r_T(\Pi_j)$ 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 σ-finite measure ν. 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 σ-finite measure ν_2 on Θ_2 and for any given θ_2, $\pi_{\theta_1|\theta_2}(\theta_1)$ is a p.d.f. w.r.t. a σ-finite measure ν_1 on Θ_1. Suppose further that if θ_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 θ_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 σ-finite measure ν on $(\Theta, \mathscr{B}_\Theta)$ 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. $\theta^{(1)}, \cdots, \theta^{(m)}$ from a p.d.f. $h(\theta) > 0$ w.r.t. ν. By the SLLN, as $m \to \infty$, $\hat{\mathbb{E}}_p(g) = \frac{1}{m}\sum_{j=1}^m \frac{g(\theta^{(j)})p(\theta^{(j)})}{h(\theta^j)} \to_{\text{a.s.}} \int_\Theta \frac{g(\theta)p(\theta)}{h(\theta)}h(\theta)d\nu = \mathbb{E}_p(g)$.

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 Θ) 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) $\theta^{(1)}, \cdots, \theta^{(m)}$ are not independent, but form a homogeneous Markov chain.

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

4.3 Minimaxity and admissibility

Definition 1 (Minimax estimator): An estimator δ 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 (Minimaxity of a Bayes estimator): Let Π be a proper prior on Θ and δ be a Bayes estimator of θ w.r.t. Π. Suppose δ has constant risk on Θ_Π. If $\Pi(\Theta_\Pi) = 1$, then δ is minimax. If, in addition, δ is the unique Bayes estimator w.r.t. Π, then it is the unique minimax estimator.

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

Example 1: Let $X_1, \cdots, X_n$ be i.i.d. components having the $\mathcal{N}(\mu, \sigma^2)$ distribution with an known $\mu = \theta \in \mathbb{R}$ and a known σ^2. If the prior is $\mathcal{N}(\mu_0, \sigma_0^2)$, then the posterior of θ 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 $c^2 \to \sigma^2/n$ as $\sigma_0^2 \to \infty$, $\sup_{\theta\in\mathbb{R}} R_T(\theta) \geq \frac{\sigma^2}{n} = \sup_{\theta\in\mathbb{R}} R_{\bar{X}}(\theta)$ where the equality holds because the risk of $\bar{X}$ under the squared error loss is σ^2/n and independent of $\theta = \mu$. Thus, $\bar{X}$ is minimax.

Theorem 3: Let Θ_0 be a subset of Θ and T be a minimax estimator of θ when Θ_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 σ-finite measure ν, 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 γ 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 $(a - \theta)^2/\mathrm{Var}(T)$. (ii) Y is the unique minimax estimator of θ under the loss $(a - \theta)^2/\mathrm{Var}(T)$.

Example 2: Let $X_1, \cdots, X_n$ 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 σ^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 $c > 0$, $\int_{-\infty}^{-c} e^{-\gamma\lambda\theta}(\frac{-2\theta}{n})^{-n\lambda/2} d\theta = \int_0^c e^{\gamma\lambda\theta}\theta^{-n\lambda/2} d\theta = \infty$. 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 σ^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 ϑ of parameters (functions of θ) under the decision theory approach.

Remark 1 (Difference from estimating ϑ 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 θ 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 θ 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 θ 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 ψ, $\delta_{c,r} = \psi(||X - c||^2)X + [1 - \psi(||X - c||^2)]c$. Any estimator having this form is called a shrinkage estimator.

Motivation 2 (Empirical Bayes estimator): A Bayes estimator of θ is of the form $\delta = (1 - B)X + Bc$, where c is the prior mean of θ 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 θ, $\delta_{c,r} = X - \frac{r(p-2)}{||X-c||^2}(X-c)$, 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 θ 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 δ_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 χ^2_p 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 δ_c). Since X is minimax, $\delta_{c,r}$ is minimax provided that $p \geq 3$ and $0 < r < 2$. Unfortunately, the James-Stein estimator δ_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 δ_c nor δ_c^+ is admissible, it is found that no substantial improvements over δ_c^+ are possible.

Definition 5 (Extension of Theorem 1 to $\mathrm{Var}(X) = \sigma^2 D$): Consider the case where $\mathrm{Var}(X) = \sigma^2 D$ with an unknown $\sigma^2 > 0$ and a known positive definite matrix D. If σ^2 is known, then an extended James-Stein estimator is $\widetilde{\delta}_{c,r} = X - \frac{(p-2)r\delta^2}{||D^{-1}(X-c)||^2}D^{-1}(X-c)$. Under the squared error loss, the risk of $\widetilde{\delta}_{c,r}$ is $\sigma^2[\mathrm{tr}(D) - (2r-r^2)(p-2)^2\sigma^2\mathbb{E}(||D^{-1}(X-c)||^{-2})]$. When σ^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/σ^2 has the chi-square distribution χ^2_m. Replacing $r\sigma^2$ in $\widetilde{\delta}_{c,r}$ by $\hat{\sigma}^2 = tS_0^2$ with a constant $t > 0$ leads to the following extended James-Stein estimator: $\widetilde{\delta}_c = X - \frac{(p-2)\hat{\sigma}^2}{||D^{-1}(X-c)||^2}D^{-1}(X-c)$. From the risk formula for $\widetilde{\delta}_{c,r}$ and the independence of $\hat{\sigma}^2$ and X, the risk of $\widetilde{\delta}_c$ is $R_{\widetilde{\delta}_c}(\theta) = \sigma^2\{\mathrm{tr}(D) - [2tm - t^2m(m+2)](p-2)^2\sigma^2\kappa(\theta)\}$, where $\theta = (\theta, \sigma^2)$ and $\kappa(\theta) = \mathbb{E}(||D^{-1}(X-c)||^{-2})$. Replacing t by $1/(m+2)$ leads to $R_{\widetilde{\delta}_c}(\theta) = \sigma^2[\mathrm{tr}(D) - m(m+2)^{-1}(p-2)^2\sigma^2\mathbb{E}(||D^{-1}(X-c)||^{-2})]$, which is smaller than $\sigma^2\mathrm{tr}(D)$ (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 $D = (Z^TZ)^{-1}$, $S_0^2 = \mathrm{SSR}$ is independent of $\hat{\beta}$, S_0^2/σ^2 has the chi-sqaure distribution χ^2_{n-p}. Hence, from the previous discussion, the risk of the shrinkage estimator $\hat{\beta} - \frac{(p-2)\hat{\sigma}^2}{||Z^TZ(\hat{\beta}-c)||^2}Z^TZ(\hat{\beta}-c)$ is smaller than that of $\hat{\beta}$ for any β and σ^2, where $c \in \mathbb{R}^p$ is fixed and $\hat{\sigma}^2 = \mathrm{SSR}/(n-p+2)$

Definition 6 (Other shinkage 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 $X \in \mathscr{X}$ be a sample with a p.d.f. f_θ w.r.t. a σ-finite measure ν, where $\theta \in \Theta \subset \mathbb{R}^k$. (i) For each $x \in \mathscr{X}$, $f_\theta(x)$ considered as a function of θ is called the likelihood function and denoted by $l(\theta)$. (ii) Let $\bar{\Theta}$ be the closure of Θ. A $\hat{\theta} \in \Theta$ satisfying $l(\hat{\theta}) = \max_{\theta\in\Theta} l(\theta)$ is called a maximum likelihood estimate (MLE) of θ. If $\hat{\theta}$ is a Borel function of X a.e. ν, then $\hat{\theta}$ is called a maximum likelihood estimator MLE of θ. (iii) Let g be a Borel function from Θ to $\mathbb{R}^p, p \le k$. If $\hat{\theta}$ is an MLE of θ, 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 Θ°, tthen 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 $X_1, \cdots, X_n$ be i.i.d. binary random variables with $P(X_1 = 1) = p \in \Theta = (0, 1)$. When $(X_1, \cdots, X_n) = (x_1, \cdots, x_n)$ 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 Θ), $l(p) \to 0$. 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 $\hat{\theta}^{(t+1)} = \hat{\theta}^{(t)} - [\frac{\partial^2 \log l(\theta)}{\partial\theta\partial\theta^T}|_{\theta=\hat{\theta}^{(t)}}]^{-1} \frac{\partial \log l(\theta)}{\partial\theta}|_{\theta=\hat{\theta}^{(t)}}, t = 0, 1, \cdots$, 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 $[\{\mathbb{E}(\frac{\partial^2 \log l(\theta)}{\partial\theta\partial\theta^T})\}|_{\theta=\hat{\theta}^{(t)}}]^{-1}$, where the expectation is taken under P_θ, 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 θ based on a sequence of samples $\{X = (X_1, \cdots, X_n), n = 1, 2, \cdots\}$. Suppose that as $n \to \infty$, $\hat{\theta}_n$ is asymptotically normal (AN) in the sense that $[V_n(\theta)]^{-1/2}(\hat{\theta}_n - \theta) \to_d \mathcal{N}_k(0, I_k)$, where, for each n, $V_n(\theta)$ is a $k \times k$ positive definite matrix depending on θ. If θ 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 $\hat{\theta}_{jn}$ is AN with asymptotic covariance matrix $V_{jn}(\theta), j = 1, 2$, and $V_{1n}(\theta) \le V_{2n}(\theta)$ for all $\theta \in \Theta$, then $\hat{\theta}_{1n}$ is said to be asymptotically more efficient than $\hat{\theta}_{2n}$.

Theorem 1: Let $X_1, \cdots, X_n$ be i.i.d. from a p.d.f. f_θ w.r.t. a σ-finite measure ν on $(\mathbb{R}, \mathscr{B})$, where $\theta \in \Theta$ and Θ 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 θ 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 $I_1(\theta) = \mathbb{E}\{\frac{\partial}{\partial\theta}\log f_\theta(X_1)[\frac{\partial}{\partial\theta}\log f_\theta(X_1)]^T\}$ is positive

definite; and for any given $\theta \in \Theta$, there exists a positive number c_θ and a positive function h_θ 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 θ 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 θ): Suppose that we are interested in estimating $\vartheta = g(\theta)$, where g is a differentiable function from Θ 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 ϑ 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 $P(s_n(\hat{\theta}_n) = 0) \to 1$ and $\hat{\theta}_n \to_p \theta$, where $s_n(\gamma) = \frac{\partial \log l(\gamma)}{\partial \gamma}$. (ii) Asymptotic efficiency. Any consistent sequence $\widetilde{\theta}_n$ 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 Θ and $p_n(\gamma)$ be the posterior p.d.f., given $X_1, \cdots, X_n$, $n = 1, 2, \cdots$. 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 $\lim_{n\to\infty} P(\sup_{||\gamma-\theta||\geq\epsilon} \frac{\log l(\gamma) - \log l(\theta)}{n} > -\delta) = 0, \lim_{n\to\infty} P(\sup_{||\gamma-\theta||\leq\delta} \frac{||\nabla s_n(\gamma) - \nabla s_n(\theta)||}{n} \geq \epsilon) = 0$, 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 θ is the true parameter value, and let $\psi(\gamma)$ be the p.d.f. of $\mathcal{N}_k(0, [I_1(\theta)]^{-1})$. Then $\int (1 + ||\gamma||) |p_n^*(\gamma) - \psi(\gamma)| d\gamma \to_p 0$. (ii) The Bayes estimator of θ 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 $n \to \infty$.

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

Definition 1 (The structure of a GLM): The sample $X = (X_1, \cdots, X_n)$ has independent X_i's and X_i has the p.d.f. $\exp\{\frac{\eta_i x_i - \zeta(\eta_i)}{\phi_i}\} h(x_i, \phi_i), i = 1, \cdots, n$, w.r.t. a σ-finite measure ν, where η_i and ϕ_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, ζ and h are known functions, and $\zeta''(\eta) > 0$ is assumed for all $\eta \in \Xi^\circ$. Note that the p.d.f. belongs to an exponential family if ϕ_i is known. As a consequence, $\mathbb{E}(X_i) = \zeta'(\eta_i)$ and $\text{Var}(X_i) = \phi_i \zeta''(\eta_i), i = 1, \cdots, n$. Define $\mu(\eta) = \zeta'(\eta)$. It is assumed that η_i is related to Z_i, the ith value of a p-value of covariates, through $g(\mu(\eta_i)) = \beta^T Z_i, i = 1, \cdots, n$, where β 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 η. In a GLM, the parameter of interest is β. We assume the range of β is $B = \{\beta : (g \circ \mu)^{-1}(\beta^T z) \in \Xi^\circ \text{ for all } z \in \mathscr{Z}\}$, where $\mathscr{Z}$ is the range of Z_i's. ϕ_i's are called dispersion parameters and are considered to be nuisance parameters.

Proposition 1 (MLE in GLM): An MLE of β in a GLM is considered under assumption $\phi_i = \phi/t_i, i = 1, \cdots, n$, 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 β, if it exists, can be obtained without estimating ϕ. The second likelihood equation, however, is usually difficult to solve. Some other estimators of ϕ 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 β, then $-\log l(\theta)$ is strictly convex in β for any fixed ϕ and, therefore, $\hat{\beta}$ is the unique MLE of β. 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., $[M_n(\beta)]^{-1/2} R_n(\beta) [M_n(\beta)]^{-1/2} \to 0$ 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 ϵ_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 $X_i = \mathcal{N}(g^{-1}(\beta^T Z_i), \phi), i = 1, \cdots, n$. This type of model is called a nonlinear regression model (with normal errors) and an MLE of β 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 β 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^{η_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 $(\sqrt{t_1} Z_1, \cdots, \sqrt{t_n} Z_n)$ 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 β. 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 β 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 β 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 $n \to \infty, \max_{i \leq n} Z_i^T (Z^T Z)^{-1} Z_i \to 0$ and $\lambda_-[Z^T Z] \to \infty$, where Z is the $n \times p$ matrix whose ith 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 $P(s_n(\hat{\beta}_n) = 0) \to$

1 and $\hat{\beta}_n \to_p \beta$, 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 $[I_n(\beta)]^{1/2}(\hat{\beta}_n - \beta) \to_d \mathcal{N}_p(0, I_p)$. (iii) If ϕ is known or the p.d.f. indexed by $\theta = (\beta, \phi)$ satisifies the conditions for f_θ in Theorem 1 of section 4.6, then $\hat{\beta}_n$ is asymptotically efficient.

Definition 2 (Quasi-MLE): If assumption ϕ_i is arbitrary, or the distribution assumption on X_i does not hold, but $\mathbb{E}(X_i) = \zeta'(\eta_i), \text{Var}(X_i) = \phi_i\zeta''(\eta_i), i = 1, \cdots, n$ and $g(\mu(\eta_i)) = \beta^T Z_i, i = 1, \cdots, n$ still hold, we estimate β 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 θ 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 θ. (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 $X_1, \cdots, X_n$ be a random sample from a pdf in a family indexed by two parameters θ and π with likelihood $l(\theta, \pi)$. The method of pseudo MLE may be viewed as follows. Based on the sample, an estimate $\hat{\pi}$ of π is obtained using some technique other than MLE. The pseudo MLE of θ is then obtained by maximizing the likelihood $l(\theta, \hat{\pi})$.

Remark 1: π is viewed as a nuisance parameter. Pseudo MLE consists of replacing π 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 π_0. As $n \to \infty$, with probability tending to 1, there exists $\hat{\theta}$ such that $\frac{\partial \log l(\hat{\theta}, \hat{\pi})}{\partial\theta} = 0$ and $\hat{\theta} \to_p \theta_0$ where θ_0 is the true value of θ.

5 Estimation in Non-Parametric Models

5.1 Empirical c.d.f. and empirical likelihoods

Definition 1 (Estimation in nonparametric models): Data $X = (X_1, \cdots, X_n)$, 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 $a = (a_1, \cdots, a_d) \in \mathbb{R}^d$. F_n is the distribution putting mass n^{-1} at each $X_i, i = 1, \cdots, n$.

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 $\rho_\infty(G_1, G_2) = ||G_1 - G_2||_\infty = \sup_{t\in\mathbb{R}^d} |G_1(t) - G_2(t)|, G_j \in \mathscr{F}$. (i) When $d = 1$, there exists a positive constant C (not depending on F) such that $P(\rho_\infty(F_n, F) > z) \le Ce^{-2nz^2}, z > 0, n = 1, 2, \cdots$. (ii) When $d \ge 2$, for any $\epsilon > 0$, there exists a positive constant $C_{\epsilon,d}$ (not depending on F) such that $P(\rho_\infty(F_n, F) > z) \le C_{\epsilon,d}e^{-(2-\epsilon)nz^2}, z > 0, n = 1, 2, \cdots$.

Theorem 2: Let F_n be the empirical c.d.f. of i.i.d. $X_1, \cdots, X_n$ from a c.d.f. F on $\mathbb{R}^d$. (i) $\rho_\infty(F_n, F) \to_{\text{a.s.}} 0$ as $n \to \infty$; (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 $X_1, \cdots, X_n$ from a c.d.f $F \in \mathscr{F}_1$. (i) $\rho_{L_p}(F_n, F) \to_{\text{a.s.}} 0$; (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 $p \ge 2$.

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

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).

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 $t_1, \cdots, t_m \in \mathbb{R}^d$, $\sqrt{n}[(\hat{F}(t_1), \cdots, \hat{F}(t_m)) - (F(t_1), \cdots, F(t_m))] \to_d \mathcal{N}_m(0, \Sigma_u)$, where $\Sigma_u = \Sigma - W^TU^{-1}W$, Σ is the covariance matrix of $\sqrt{n}[(F_n(t_1), \cdots, F_n(t_m)) - (F(t_1), \cdots, F(t_m))], W = (W(t_1), \cdots, W(t_m))$, and $W(t_j) = \mathbb{E}[u(X_1)I_{(-\infty,t_j]}(X_1)]$.