

Advanced Theory of Statistics

Lectured by Wang Miao

L^AT_EXed by Chengxin Gong

2022 年 10 月 10 日

目录

1	Probability Theory	2
1.1	Measure space, measurable function, and integration	2
1.2	Integration theory and Radon-Nikodym derivative	3
1.3	Densities, moments, inequalities, and generating functions	4
1.4	Conditional expectation and independence	5
1.5	Conditional distribution, Markov chains, and martingales	6
1.6	Convergence modes and relationships	6
1.7	Uniform integrability and weak convergence	7
1.8	Convergence of transformations and law of large numbers	8
1.9	The law of large numbers and central limit theorem	8
2	Fundamentals of Statistics	9
2.1	Models, data, statistics, and sampling distributions	9
2.2	Sufficiency and minimal sufficiency	9
2.3	Completeness	10
2.4	Statistical decision and inference	11

1 Probability Theory

1.1 Measure space, measurable function, and integration

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

Example 1: \mathcal{C} = a collection of subsets of interest. $\sigma(\mathcal{C})$ = the smallest σ -field containing \mathcal{C} (the σ -field generated by \mathcal{C}). $\sigma(\mathcal{C}) = \mathcal{C}$ if \mathcal{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). \mathcal{O} = all open sets, \mathcal{C} = all closed sets. $\mathcal{B}^k = \sigma(\mathcal{O}) = \sigma(\mathcal{C})$: the Borel σ -field on \mathbb{R}^k . $C \in \mathcal{B}^k, \mathcal{B}_C = \{C \cap B : B \in \mathcal{B}^k\}$ is the Borel σ -field on C .

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

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

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

Example 3 (σ -finite): A measure ν on (Ω, \mathcal{F}) is said to be σ -finite iff there exists a sequence $\{A_1, A_2, \dots\}$ such that $\cup A_i = \Omega$ and $\nu(A_i) < \infty$ for all i . Any finite measure is clearly σ -finite. The Lebesgue measure on \mathcal{F} is σ -finite.

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

Definition 3 (Measurable function): Let (Ω, \mathcal{F}) and (Λ, \mathcal{G}) be measurable spaces. Let f be a function from Ω to Λ . f is called a measurable function from (Ω, \mathcal{F}) to (Λ, \mathcal{G}) iff $f^{-1}(\mathcal{G}) \subset \mathcal{F}$.

Definition 4 (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 \mathcal{S}_f be the collection of all nonnegative simple functions satisfying $\phi(\omega) \leq f(\omega)$ for any $\omega \in \Omega$. The integral of f w.r.t. ν is defined as $\int f d\nu = \sup\{\int \phi d\nu : \phi \in \mathcal{S}_f\}$ (Hence, for any Borel function $f \geq 0$, there exists a sequence of simple functions ϕ_1, ϕ_2, \dots such that $0 \leq \phi_i \leq f$ for all i and $\lim_{n \rightarrow \infty} \int \phi_n d\nu = \int f d\nu$). (c) Let f be a Borel function, $f_+(\omega) = \max\{f(\omega), 0\}$ be the positive part of f , and $f_-(\omega) = \max\{-f(\omega), 0\}$ be the negative part of f . We say that $\int f d\nu$ exists if and only if at least one of $\int f_+ d\nu$ and $\int f_- d\nu$ is finite, in which case $\int f d\nu = \int f_+ d\nu - \int f_- d\nu$. (d) When both $\int f_+ d\nu$ and $\int f_- d\nu$

$\int f_- d\nu$ are finite, we say that f is integrable. Let A be a measurable set and I_A be its indicator function. The integral of f over A is defined as $\int_A f d\nu = \int I_A f d\nu$.

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

1.2 Integration theory and Radon-Nikodym derivative

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

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

Example 1 (Interchange of differentiation and integration): Let $(\Omega, \mathcal{F}, \nu)$ be a measure space and, for any fixed $\theta \in \mathbb{R}$, let $f(\omega, \theta)$ be a Borel function on 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 $(\Omega, \mathcal{F}, \nu)$ to (Λ, \mathcal{G}) and g be Borel on (Λ, \mathcal{G}) . Then $\int_{\Omega} g \circ f d\nu = \int_{\Lambda} g d(\nu \circ f^{-1})$, i.e., if either integral exists, then so does the other, and the two are the same.

Theorem 3 (Fubini's theorem): Let ν_i be a σ -finite measure on $(\Omega_i, \mathcal{F}_i)$, $i = 1, 2$, and f be a Borel function on $\prod_{i=1}^2 (\Omega_i, \mathcal{F}_i)$ with $f \geq 0$ or $\int |f| d\nu_1 \times \nu_2 < \infty$. Then $g(\omega_2) = \int_{\Omega_1} f(\omega_1, \omega_2) d\nu_1$ exists a.e. 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 $(\Omega, \mathcal{F}, \nu)$. We say λ is absolutely continuous w.r.t. ν and write $\lambda \ll \nu$ iff $\nu(A) = 0$ implies $\lambda(A) = 0$.

Theorem 4 (Radon-Nikodym theorem): Let ν and λ be two measure on (Ω, \mathcal{F}) and ν be σ -finite. If $\lambda \ll \nu$, then there exists a nonnegative Borel function f on Ω such that $\lambda(A) = \int_A f d\nu$, $A \in \mathcal{F}$. Furthermore, f is unique a.e. ν , i.e. if $\lambda(A) = \int_A g d\nu$ for any $A \in \mathcal{F}$, 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 (Ω, \mathcal{F}) . (i) If λ is a measure, $\lambda \ll \nu$, and $f \geq 0$, then $\int f d\lambda = \int f \frac{d\lambda}{d\nu} d\nu$. (ii) If $\lambda_i, i = 1, 2$, are measures and $\lambda_i \ll \nu$, then $\lambda_1 + \lambda_2 \ll \nu$ and $\frac{d(\lambda_1 + \lambda_2)}{d\nu} = \frac{d\lambda_1}{d\nu} + \frac{d\lambda_2}{d\nu}$ a.e. ν . (iii) If τ is a measure, λ is a σ -finite measure, and $\tau \ll \lambda \ll \nu$, then $\frac{d\tau}{d\nu} = \frac{d\tau}{d\lambda} \frac{d\lambda}{d\nu}$ a.e. ν . In particular, if

$\lambda \ll \nu$ and $\nu \ll \lambda$ (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 $(\Omega_i, \mathcal{F}_i, \nu_i)$ be a measure space and ν_i be σ -finite, $i = 1, 2$. Let λ_i be a σ -finite measure on (Ω, \mathcal{F}_i) and $\lambda_i \ll \nu_i, i = 1, 2$. Then $\lambda_1 \times \lambda_2 \ll \nu_1 \times \nu_2$ and $\frac{d(\lambda_1 \times \lambda_2)}{d(\nu_1 \times \nu_2)}(\omega_1, \omega_2) = \frac{d\lambda_1}{d\nu_1}(\omega_1) \frac{d\lambda_2}{d\nu_2}(\omega_2)$ a.e. $\nu_1 \times \nu_2$.

1.3 Densities, moments, inequalities, and generating functions

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

$\delta_c(A) = \begin{cases} 1 & c \in A \\ 0 & c \notin A \end{cases}, A \in \mathcal{B}$. Then $P_Y \ll m + \delta_c$, 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 $x \in \mathcal{B}$.

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

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

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

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

Definition 1 (Moments): If $\mathbb{E}X^k$ is finite, where k is a positive integer, $\mathbb{E}X^k$ is called the k -th moment of X or P_x . If $\mathbb{E}|X|^a < \infty$ for some real number a , $\mathbb{E}|X|^a$ is called the a -th absolute moment of X or P_X . If $\mu = \mathbb{E}X$, $\mathbb{E}(X - \mu)^k$ is called the k -th central moment of X or P_X . $\text{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2$

is called the variance of X or P_X . For random matrix $M = (M_{ij})$, $\mathbb{E}M = (\mathbb{E}M_{ij})$. For random vector X , $\text{Var}(X) = \mathbb{E}(X - \mathbb{E}X)(X - \mathbb{E}X)^T$ is its covariance matrix, whose (i, j) -th element, $i \neq j$, is called the covariance of X_i and X_j and denoted by $\text{Cov}(X_i, X_j)$. If $\text{Cov}(X_i, X_j) = 0$, then X_i and X_j are said to be uncorrelated. Independence implies uncorelation, not converse. If X is random and c is fixed, then $\mathbb{E}(c^T X) = c^T \mathbb{E}(X)$ and $\text{Var}(c^T X) = c^T \text{Var}(X)c$.

Definition 2 (Moment generating and characteristic functions): Let X be a random k -vector.

- (i) The moment generating function (m.g.f.) of X or P_X is defined as $\psi_X(t) = \mathbb{E}e^{t^T X}$, $t \in \mathbb{R}^k$.
- (ii) The characteristic function (ch.f.) of X or P_X is defined as $\phi_X(t) = \mathbb{E}e^{it^T X} = \mathbb{E}[\cos(t^T X)] + i\mathbb{E}[\sin(t^T X)]$, $t \in \mathbb{R}^k$.

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

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

1.4 Conditional expectation and independence

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

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

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

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

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

Proposition 2: Let X, Y, X_1, X_2, \dots be integrable random variables on (Ω, \mathcal{F}, P) and \mathcal{A} be

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

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

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

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

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

1.5 Conditional distribution, Markov chains, and martingales

1.6 Convergence modes and relationships

Definition 1 (Convergence modes): Let X, X_1, X_2, \dots be a random k -vectors defined on a probability space. (i) We say that the sequence $\{X_n\}$ converges to X almost surely and write $X_n \rightarrow_{\text{a.s.}} X$ iff $\lim_{n \rightarrow \infty} X_n = X$ a.s. (ii) We say that $\{X_n\}$ converges to X in probability and write $X_n \rightarrow_p X$ iff for every fixed $\epsilon > 0$, $\lim_{n \rightarrow \infty} P(\|X_n - X\| > \epsilon) = 0$. (iii) We say that $\{X_n\}$ converges to X in L_r (or in r th moment) with a fixed $r > 0$ and write $X_n \rightarrow_{L_r} X$ iff $\lim_{n \rightarrow \infty} \mathbb{E}\|X_n - X\|_r^r = 0$. (iv) Let $F, F_n, n = 1, 2, \dots$ be c.d.f.'s on \mathbb{R}^k and $P, P_n, n = 1, 2, \dots$ be their corresponding probability measures. We say that $\{F_n\}$ converges to F weakly (or $\{P_n\}$ converges to P weakly) and write $F_n \rightarrow_w F$ (or $P_n \rightarrow_w P$) iff, for each continuity point x of F , $\lim_{n \rightarrow \infty} F_n(x) = F(x)$. We say that $\{X_n\}$ converges to X in distribution (or in law) and write $X_n \rightarrow_d X$ iff $F_{X_n} \rightarrow_w F_X$.

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

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

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

Definition 2: Let X_1, X_2, \dots be random vectors and Y_1, Y_2, \dots be random variables defined on a common probability space. (i) $X_n = O(Y_n)$ a.s. iff $P(\|X_n\| = O(|Y_n|)) = 1$. (ii) $X_n = o(Y_n)$

a.s. iff $X_n/Y_n \rightarrow_{\text{a.s.}} 0$. (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 $X_n/Y_n \rightarrow_p 0$.

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

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

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

1.7 Uniform integrability and weak convergence

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

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

Theorem 1 (Useful sufficient and necessary conditions for convergence in distribution): Let X, X_1, X_2, \dots be random k -vectors. (i) $X_n \rightarrow_d X$ is equivalent to any one of the following conditions: (a) $\mathbb{E}[h(X_n)] \rightarrow \mathbb{E}[h(X)]$ for every bounded continuous function h ; (b) $\limsup_n P_{X_n}(C) \leq P_X(C)$ for any closed set $C \subset \mathbb{R}^k$; (c) $\liminf_n P_{X_n}(O) \geq P_X(O)$ for any open set $O \subset \mathbb{R}^k$. (ii) Lévy-Cramér continuity theorem. Let $\phi_X, \phi_{X_1}, \phi_{X_2}, \dots$ be the ch.f.'s of X, X_1, X_2, \dots , respectively. $X_n \rightarrow_d X$ iff $\lim_{n \rightarrow \infty} \phi_{X_n}(t) = \phi_X(t)$ for all $t \in \mathbb{R}^k$. (iii) Cramér-Wold device. $X_n \rightarrow_d X$ iff $c^T X_n \rightarrow_d c^T X$ for every $c \in \mathbb{R}^k$.

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

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

1.8 Convergence of transformations and law of large numbers

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

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

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

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

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

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

1.9 The law of large numbers and central limit theorem

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

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

Theorem 3 (Berry-Esséen bound): For i.i.d. $\{X_n\}$ and $W_n = \sqrt{n}(\bar{X} - \mu)/\sigma$, $\sup_t |F_{W_n}(t) - \phi(t)| \leq \frac{33}{4} \frac{\mathbb{E}|X_1 - \mu|^3}{\sigma^3 \sqrt{n}}, n = 1, 2, \dots$. Thus, the convergence speed of F_{W_n} to 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 (Ω, \mathcal{F}) indexed by a parameter $\theta \in \Theta$ is said to be a parametric family or follow a parametric model iff $\Theta \subset \mathbb{R}^d$ for some fixed positive integer d and each P_θ is a known probability measure when θ is known. The set Θ is called the parameter space and d is called its dimension. $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is identifiable iff $\theta_1 \neq \theta_2$ and $\theta_i \in \Theta$ imply $P_{\theta_1} \neq P_{\theta_2}$, which may be achieved through reparameterization.

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

Definition 3 (Exponential families): A parametric family $\{P_\theta : \theta \in \Theta\}$ dominated by a σ -finite measure ν on (Ω, \mathcal{F}) is called an exponential family iff $\frac{dP_\theta}{d\nu}(\omega) = \exp\{\eta(\theta)^T T(\omega) - \xi(\theta)\} h(\omega)$, $\omega \in \Omega$ where $\xi(\theta) = \log\{\int_\Omega \exp\{\eta(\theta)^T T(\omega)\} h(\omega) d\nu(\omega)\}$. In an exponential family, consider the parameter $\eta = \eta(\theta)$ and $f_\eta(\omega) = \exp\{\eta^T T(\omega) - \zeta(\eta)\} h(\omega)$, $\omega \in \Omega$. This is called the canonical form for the family, and $\Xi = \{\eta : \zeta(\eta) \text{ is defined}\}$ is called the natural parameter space. An exponential family in canonical form is a natural exponential family. If there is an open set contained in the natural parameter space of an exponential family, then the family is said to be of full rank.

Theorem 1: Let \mathcal{P} be a natural exponential family. (i) Let $T = (Y, U)$ and $\eta = (\theta, \phi)$, Y and 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. $f_{\theta, u}(y) = \exp\{\theta^T y - \zeta_u(\theta)\}$, which is in a natural exponential family indexed by θ . (ii) If η_0 is an interior point of the natural parameter space, then the m.g.f. of $P_{\eta_0} \circ T^{-1}$ is finite in a neighborhood of 0 and is given by $\psi_{\eta_0}(t) = \exp\{\zeta(\eta_0 + t) - \zeta(\eta_0)\}$.

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

2.2 Sufficiency and minimal sufficiency

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

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

sufficient for $P \in \mathcal{P}$ iff there are nonnegative Borel functions h and g_p on the range of T such that $\frac{dP}{d\nu}(x) = g_p(T(x))h(x)$.

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

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

Theorem 2 (Existence and uniqueness): Minimal sufficient statistics exist when \mathcal{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. \mathcal{P} .

Theorem 3: Let \mathcal{P} be a family of distributions on \mathbb{R}^k . (i) Suppose that $\mathcal{P}_0 \subset \mathcal{P}$ and a.s. \mathcal{P}_0 implies a.s. \mathcal{P} . If T is sufficient for $P \in \mathcal{P}$ and minimal sufficient for $P \in \mathcal{P}_0$, then T is minimal sufficient for $P \in \mathcal{P}$. (ii) Suppose that \mathcal{P} contains p.d.f.'s f_0, f_1, f_2, \dots w.r.t. a 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 $f_\infty(x) > 0, i = 0, 1, 2, \dots$. Then $T(x) = (T_0, T_1, T_2, \dots)$ is minimal sufficient for $P \in \mathcal{P}$. Furthermore, if $\{x : f_i(x) > 0\} \subset \{x : f_0(x) > 0\}$ for all i , then we may replace $f_\infty(x)$ for $f_0(x)$, in which case $T(x) = (T_1, T_2, \dots)$ is minimal sufficient for $P \in \mathcal{P}$. (iii) Suppose that \mathcal{P} contains p.d.f.'s f_p w.r.t. a σ -finite measure and that there exists a sufficient statistic $T(x)$ such that, for any possible values x and y of X , $f_p(x) = f_p(y)\phi(x, y)$ for all P implies $T(x) = T(y)$, where ϕ is a measurable function. Then $T(x)$ is minimal sufficient for $P \in \mathcal{P}$.

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 $P \in \mathcal{P}$ iff, for any Borel f (or bounded Borel f), $\mathbb{E}[f(T)] = 0$ for all $P \in \mathcal{P}$ implies $f = 0$ a.s. \mathcal{P} .

Remark 2: If T is complete (or boundedly complete) and $S = \psi(T)$ for a measurable 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, \dots, X_n are i.i.d. random variables having the $\mathcal{N}(\mu, \sigma^2)$ distribution, $\mu \in \mathbb{R}$, $\sigma > 0$. The joint p.d.f. of X_1, \dots, X_n is $(2\pi)^{-n/2} \exp\{\eta_1 T_1 + \eta_2 T_2 - n\zeta(\eta)\}$, where $T_1 = \sum_{i=1}^n X_i, T_2 = -\sum_{i=1}^n X_i^2$ and $\eta = (\eta_1, \eta_2) = (\frac{\mu}{\sigma^2}, \frac{1}{2\sigma^2})$. Hence, the family of distributions for $X = (X_1, \dots, X_n)$ is a natural exponential family of full rank ($\Xi = \mathbb{R} \times (0, \infty)$). Thus $T(X) = (T_1, T_2)$

is complete and sufficient for η .

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

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

Example 3: X_1, \dots, X_n is a random sample from $\text{uniform}(\theta, \theta + 1)$, $\theta \in \mathbb{R}$, 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 and inference

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

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

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

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

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

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

is the alternative hypothesis. The action space for this problem contains only two elements, i.e., $\mathcal{A} = \{0, 1\}$, where 0 is accepting H_0 and 1 is rejecting H_0 . This problem is a special case of a general problem called hypothesis testing. A decision rule is called a test, which must have the form $I_C(X)$, where $C \in \mathcal{F}_{\mathcal{X}}$ is called the rejection or critical region.

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

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

Remark 1: An admissible decision rule is not necessarily good. For example, in an estimation problem a silly estimator $T(X) \equiv a$ constant may be admissible.

Proposition 1: Let $T(X)$ be a sufficient statistic for $P \in \mathcal{P}$ and let 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 $P \in \mathcal{P}$.

Theorem 1: Suppose that \mathcal{A} is a convex subset of \mathbb{R}^k and that for any $P \in \mathcal{P}$, $L(P, a)$ is a convex function of a . (i) Let δ be a randomized rule satisfying $\int_{\mathcal{A}} \|a\| d\delta(x, a) < \infty$ for any $x \in \mathcal{X}$ and let $T_1(x) = \int_{\mathcal{A}} a d\delta(x, a)$. Then $L(P, T_1(x)) \leq L(P, \delta, x)$ (or $L(P, T_1(x)) < L(P, \delta, x)$) if L is strictly convex in a for any $x \in \mathcal{X}$ and $P \in \mathcal{P}$. (ii) Rao-Blackwell theorem. Let T be a sufficient statistic for $P \in \mathcal{P}$, $T_0 \in \mathbb{R}^k$ be a nonrandomized rule satisfying $\mathbb{E}\|T_0\| < \infty$, and $T_1 = \mathbb{E}[T_0(X) | T]$. Then $R_{T_1}(P) \leq R_{T_0}(P)$ for any $P \in \mathcal{P}$. If L is strictly convex in a and T_0 is not a function of T , then T_0 is inadmissible.

Definition 7 (Unbiasedness): In an estimation problem, the bias of an estimator $T(X)$ of a parameter 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 $P \in \mathcal{P}$.

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

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