# Advanced Theory of Statistics

Lectured by Wang Miao LaTeXed by Chengxin Gong

2022 年 10 月 13 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 $\sigma$-field (or $\sigma$-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 $\sigma$-field on $\Omega$.

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

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

Definition 2: Let $(\Omega, \mathscr{F})$ be a measurable space. A set function $\nu$ 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 $\nu$ 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 $\infty/\infty$ is not defined.

Example 3 (Important examples of measures): (a) Let $x \in \Omega$ be a fixed point and $\delta_x(A) = \begin{cases} c & x \in A \\ 0 & x \notin A \end{cases}$. This is called a point mass at $x$. (b) Let $\mathscr{F}$ = all subsets of $\Omega$ and $\nu(A)$ = the number of elements in $A \in \mathscr{F}$ ($\nu(A) = \infty$ if $A$ contains infinitely many elements). Then $\nu$ 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 $\lim_{n\to\infty} AA_n = \cup_{i=1}^{\infty} A_i$ (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 $\infty$. $\Gamma_i, i \in \mathscr{I}$, 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 $\sigma$-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 ($\sigma$-finite): A measure $\nu$ on $(\Omega, \mathscr{F})$ is said to be $\sigma$-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 $\sigma$-finite. The Lebesgue measure on $\mathscr{F}$ is $\sigma$-finite.

Proposition 3 (Product measure theorem): Let $(\Omega_i, \mathscr{F}_i, \nu_i), i = 1, \cdots, k$, be measure spaces with $\sigma$-finite measures. There exists a unique $\sigma$-finite measure on $\sigma$-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 $\Omega$ to $\Lambda$. $f$ 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 $\phi$ w.r.t.$\nu$ is defined as $\int \phi d\nu = \sum_{i=1}^k a_i \nu(A_i)$. (b) Let $f$ be a nonnegative Borel function and let $\mathscr{S}_f$ 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. $\nu$ 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 $\Omega$. Suppose that $\partial f(\omega, \theta)/\partial\theta$ exists a.e. for $\theta \in (a, b) \subset \mathbb{R}$ and that $|\partial f(\omega, \theta)/\partial\theta| \leq g(\omega)$ a.e., where $g$ is an integrable function on $\Omega$. Then for each $\theta \in (a, b)$, $\partial f(\omega, \theta)/\partial\theta$ is integrable and, by Theorem 1(ii), $\frac{d}{d\theta} \int f(\omega, \theta) d\nu = \int \frac{\partial f(\omega,\theta)}{\partial\theta} d\nu$.

Theorem 2 (Change of variables): Let $f$ be measurable from $(\Omega, \mathscr{F}, \nu)$ 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 $\nu_i$ be a $\sigma$-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. $\nu_2$ and defines a Borel function on $\Omega_2$ whose integral w.r.t. $\nu_2$ exists, and $\int_{\Omega \times \Omega} f(\omega_1, \omega_2) d\nu_1 \times \nu_2 = \int_{\Omega_2} [\int_{\Omega_1} f(\omega_1, \omega_2) d\nu_1] d\nu_2$.

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

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

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

Proposition 2 (Calculus with Radon-Nikodym derivatives): Let $\nu$ be a $\sigma$-finite measure on a measure space $(\Omega, \mathscr{F})$. (i) If $\lambda$ 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. $\nu$. (iii) If $\tau$ is a measure, $\lambda$ is a $\sigma$-finite measure, and $\tau << \lambda << \nu$, then $\frac{d\tau}{d\nu} = \frac{d\tau}{d\lambda}\frac{d\lambda}{d\nu}$ a.e. $\nu$. In particular, if $\lambda << \nu$ and $\nu << \lambda$ (in which case $\lambda$ and $\nu$ are equivalent), then $\frac{d\lambda}{d\nu} = (\frac{d\nu}{d\lambda})^{-1}$ a.e. $\nu$ or $\lambda$. (iv) Let $(\Omega_i, \mathscr{F}_i, \nu_i)$ be a measure space and $\nu_i$ be $\sigma$-finite, $i = 1, 2$. Let $\lambda_i$ be a $\sigma$-finite measure on $(\Omega, \mathscr{F}_i)$ 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 $\chi^2_n$. 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. $\psi_X$ finite in a neighborhood of 0, $\frac{\partial \psi_X(t)}{\partial t}|_{t=0} = \mathbb{E}X$, $\frac{\partial^2 \psi_X(t)}{\partial t \partial t^T}|_{t=0} = \mathbb{E}(XX^T)$. If $\mathbb{E}|X_1^{r_1} \cdots X_k^{r_k}| < \infty$ 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-$\sigma$-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) $\int_A \mathbb{E}(X|\mathscr{A}) dP = \int_A X dP$ for any $\mathscr{A} \in \mathscr{A}$. (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 thaat $(X, Y)$ has a joint p.d.f. $f(x, y)$ w.r.t. $\nu \times \lambda$, where $\nu$ and $\lambda$ are $\sigma$-finite measures on $(\mathbb{R}^n, \mathscr{B}^n)$ 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. $\lambda$.

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-$\sigma$-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-$\sigma$-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 $\pi$-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 $r$th 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 ais 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 eixsts 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) $\limsup_n P_{X_n}(C) \leq P_X(C)$ 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 $\phi_X, \phi_{X_1}, \phi_{X_2}$ 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 $\mu$. 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. $\nu$. Suppose that $\lim_{n\to\infty} f_n(x) = f(x)$ a.e. and $f(x)$ is a p.d.f. w.r.t. $\nu$. 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 $j$, $1 \leq j \leq m-1$, vanishing at $c$, but with the $m$th-order partial derivatives not all vanishing at $c$. Then $a_n^m[g(X_n) - g(c)] \to_d \frac{1}{m!} \sum_{i_1=1}^k \cdots \sum_{i_m=1}^k \frac{\partial^m g}{\partial x_{i_1} \cdots \partial x_{i_m}}|_{x=c} Y_{i_1} \cdots Y_{i_m}$.

Theorem 4 (The $\delta$-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 $a_n = \mathbb{E}(X_1 1_{\{|X_1| \leq n\}})$. (ii) The SLLN. A necessary and sufficient condition for the existence of a constant $c$ for which $\frac{1}{n}\sum_{i=1}^n X_i \to_{\text{a.s.}} c$ 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 $\sum_{i=1}^i nfty \frac{\mathbb{E}|X_i|^p}{i^p} < \infty$, 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 law of large numbers and 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 $0 < \sigma_n^2 = \mathrm{var}(\sum_{j=1}^{k_n} X_{nj}) < \infty, n = 1, 2, \cdots$. 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 $\Sigma = \mathrm{var}(X_1)$, $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - \mathbb{E}X_1) \to_d \mathcal{N}_k(0, \Sigma)$.

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 $\phi$ is of the order $n^{-1/2}$.

# 2 Fundamentals of Statistics

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

Definition 1: A set of probability measures $P_\theta$ on $(\Omega, \mathscr{F})$ 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 $P_\theta$ is a known probability measure when $\theta$ is known. The set $\Theta$ 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 $\nu$ (a $\sigma$-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 $\sigma$-finite measure $\nu$ 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 $\theta$ have the same dimension. Then, $Y$ has the p.d.f. $f_\eta(y) = \exp\{\theta^T y - \zeta(\eta)\}$. In particular, $T$ has a p.d.f. in a natural exponential family. Furthermore, the conditional distribution of $Y$ given $U = u$ has the p.d.f. $f_{\theta,u}(y) = \exp\{\theta^T y - \zeta_u(\theta)\}$ w.r.t. a $\sigma$-finite measure depending on $\phi$. Furthermore, the conditional distribution of $Y$ given $U = u$ has the p.d.f. $f_{\theta,u}(y) = \exp(\theta^T y - \zeta_u(\theta))$ w.r.t. a $\sigma$-finite measure depending on $u$. (ii) If $\eta_0$ is an interior point of the natural parameter space, then the m.g.f. of $P_{\eta_0} \circ T^{-1}$ is finite in a 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 $\mu$ and $\Sigma^{1/2}$ are called the location and scale parameters, respectively.

Definition 5 (Statistics and their sampling distributions): Our data set is a realization of a sample (random vector) $X$ from an unknown population $P$. Statistic $T(X)$: A measurable function T of $X$; $T(X)$ is a known value whenever $X$ is known. A nontrivial statistic $T(X)$ is usually simpler than $X$. Finding the form of the distribution of $T$ is one of the major problems in statistical inference and decision theory.

Example 1: Let $X_1, \cdots, X_n$ 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 $i$th 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 $\sigma$-finite measure $\nu$. 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 $\sigma$-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 $\psi$ 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 $\sigma$-finite measure. If both $T$ and $S$ are minimal sufficient statistics, then by definition there is one-to-one measurable function $\psi$ such that $T = \psi(S)$ a.s. $\mathscr{P}$.

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 $\sigma$-finite $\nu$. Let $f_\infty(x) = \sum_{i=0}^\infty c_i f_i(x)$, where $c_i > 0$ for all $i$ and $\sum_{i=0}^\infty c_i = 1$, and let $T_i(x) = f_i(x)/f_\infty(x)$ when $f_\infty(x) > 0, i = 0, 1, 2, \cdots$. 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 $\sigma$-finite measure and that there exists a sufficient statistic $T(x)$ such that, for any possible values $x$ and $y$ of $X$, $f_p(x) = f_p(y)\phi(x, y)$ for all $P$ implies $T(x) = T(y)$, where $\phi$ is a measurable function. Then $T(x)$ is minimal sufficient for $P \in \mathscr{P}$.

### 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 $\theta$ and, hence, a further simplification of $T$ may still be needed.

Definition 2 (Completeness): A statistic $T(x)$ is complete (or boundedly complete) for $P \in \mathscr{P}$ 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 $\psi$, then $S$ is complete (or boundedly complete). A complete and sufficient statistic should be minimal sufficient. But a minimal sufficient statistic may be not complete.

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

Example 1: Suppose that $X_1, \cdots, X_n$ 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 $\eta$.

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 $\theta$. We can show that $T$ is not complete.

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

### 2.4 Statistical decision

Convention 1 (Basic elements): $X$: a sample from a population $P \in \mathscr{P}$. 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 averaaged 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 $R_{T_1}(P) \le R_{T_2}(P)$ 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 $\delta$ 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 $\delta$ 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 $\delta$ is nonrandomized. The risk of a randomized $\delta$ 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 $\theta$. 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 $\mu$ and variance $\sigma^2 < \infty$, then $R_{\bar{X}}(P) = (\mu - \theta)^2 + \frac{\sigma^2}{n}$. This problem is a special case of a general problem called estimation. In an estimation problem, a decision rule $T$ is called an estimator.

Example 2: Let $\mathscr{P}$ 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 $\delta_0$ be a decision rule. Then $\delta_1(t, A) = \mathbb{E}[\delta_0(X, A) | T = t]$, which is a randomized decision rule depending only on $T$, is equivalent to $\delta_0$ if $R_{\delta_0}(P) < \infty$ for any $P \in \mathscr{P}$.

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 $\delta$ 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 $L(P, T_1(x)) \le L(P, \delta, x)$ (or $L(P, T_1(x)) < L(P, \delta, x)$) if $L$ is strictly convex in $a$ for any $x \in \mathscr{X}$ 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 $R_{T_1}(P) \le R_{T_0}(P)$ 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 $\theta$ of the unknown population is defined to be $b_T(P) = \mathbb{E}[T(X)] - \theta$. An estimator $T(X)$ is unbiased for $\theta$ iff $b_T(P) = 0$ for any $P \in \mathscr{P}$.

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): $\mathrm{mse}_T(P) = \mathbb{E}[T(X) - \theta]^2 = [b_T(P)]^2 + \mathrm{Var}(T(X))$. 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 $\theta$.

Definition 4 (Significance tests): A common approach of finding an "optimal" test is to assign a small bound $\alpha$ to the type I error rate $\alpha_T(P), P \in \mathscr{P}_0$, 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 $\alpha$ is called the level of significance. The left-hand side is called the size of the test $T$. The level of significance should be positive, otherwise no test satisfies.

Definition 5 ($p$-value): It is good practice to determine not only whether $H_0$ is rejected for a given a and a chosen test $T_\alpha$, but also the smallest possible level of significance at which $H_0$ would be rejected for the computed $T_\alpha(x)$, i.e., $\hat{\alpha} = \inf\{\alpha \in (0, 1) : T_\alpha(x) = 1\}$. Such an $\hat{\alpha}$, which depends on $x$ and the chosen test and is a statistic, is called the $p$-value for the test $T_\alpha$.

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

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 $X = (X1, \cdots, Xn)$ be a sample from $P \in \mathscr{P}$, $T_n(X)$ be an estimator of $\theta$ for every $n$, and $\{a_n\}$ be a sequence of positive constants, $a_n \to \infty$. (i) $T_n(x)$ is consistent for $\theta$ iff $T_n(x) \to_p \theta$ w.r.t. any $P$. (ii) $T_n(x)$ is $a_n$-consistent for $\theta$ iff $a_n[T_n(X)-\theta] = O_p(1)$ w.r.t. any $P$. (iii) $T_n(x)$ is strongly consistent for $\theta$ iff $T_n(x) \to_{\text{a.s.}} \theta$ w.r.t. any $P$. (iv) $T_n(X)$ is $L_r$-consistent for $\theta$ 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 (Approximate and asymptotic bis): 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 $\xi_n$. (ii) For a point estimator $T_n$ of $\theta$, an asymptotic expectation of $T_n - \theta$, if it exists, is called an asymptotic bias of $T_n$ and denoted by $\widetilde{b}_{T_n}(P)$. 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 $\xi_n$. Then, one of the following three must hold: (a) $\mathbb{E}\xi = \mathbb{E}\eta = 0$; (b) $\mathbb{E}\xi \neq 0, \mathbb{E}\eta = 0$, and $b_n/a_n \to 0$; (c) $\mathbb{E}\xi \neq 0, \mathbb{E}\eta \neq 0$, and $(\mathbb{E}\xi/a_n)/(\mathbb{E}\eta/b_n) \to 1$.

Definition 10 (Asymptotic variance and amse): Let $T_n$ be an estimator of $\theta$ for every $n$ and $\{a_n\}$ be a sequence of positive numbers satisfying $a_n \to \infty$ 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 $\text{amse}_{T_n}(P)$, is defined as the asymptotic expectation of $(T_n-\theta)^2$, $\text{amse}_{T_n}(P) = \mathbb{E}Y^2/a_n^2$. The asymptotic variance of $T_n$ is defined as $\sigma^2_{T_n}(P) = \text{Var}(Y)/a_n^2$. (ii) Let $T_n'$ be another estimator of $\theta$. The asymptotic relative efficiency of $T_n'$ w.r.t. $T_n$ is defined as $e_{T_n',T_n} = \text{amse}_{T_n}(P)/\text{amse}_{T_n'}(P)$. (iii) $T_n$ is said to be asymptotically more efficient than $T_n'$ iff $\limsup_n e_{T_n',T_n}(P) \le 1$ for any $P$ and $< 1$ for some $P$.

Proposition 2: Let $T_n$ be an estimator of $\theta$ for every $n$ and $\{a_n\}$ be a sequence of positive numbers satisfying $a_n \to \infty$ 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.