Lecture 1. Random Variables & Distributions Families

[1] *School of Data Science, The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen)*

1. Random Variable

- A probability space is a triple $(\Omega, \mathcal{F}, \mathsf{P})$, where
  - (i) Ω is a nonempty set of elements to be called "points" and denoted generically by ω.
  - (ii) $\mathcal{F}$ is a nonempty collection of subsets of Ω closed under complement and countable unions, and $\emptyset \in \mathcal{F}$. $\mathcal{F}$ is called a σ-field or a Borel field (B.F.).
  - (iii) P (or more commonly seen as $\mathbb{P}$), which is called a probability measure, is a numerically valued set function with domain $\mathcal{F}$, satisfying the following axioms:
    - (a) $\forall E \in \mathcal{F}$, $\mathsf{P}(E) \geq 0$.
    - (b) If $\{E_j\}_{j\geq 1}$ is a countable collection of disjoint sets in $\mathcal{F}$, then

$$\mathsf{P}\left(\bigcup_j E_j\right) = \sum_j \mathsf{P}(E_j).$$

    - (c) $\mathsf{P}(\Omega) = 1$.
- Let's denote $\mathbb{R} = (-\infty, \infty)$ the real line, $\mathcal{B}$ the Euclidean Borel field on $\mathbb{R}$. Then a random variable is defined as

Definition 1.1. A real-valued random variable is a function X whose domain is a set E in $\mathcal{F}$ and whose range is contained in $\mathbb{R}$ such that for each $B \in \mathcal{B}$, we have

$$\{\omega : X(\omega) \in B\} \in E \cap \mathcal{F} \subset \mathcal{F}.$$

Here $E \cap \mathcal{F}$ is the trace of $\mathcal{F}$ on E, i.e., $E \cap \mathcal{F} = \{E \cap E_j : \forall E_j \in \mathcal{F}\}$.

- The (cumulative) distribution function of a random variable X is a function $F : \mathbb{R} \mapsto [0,1]$ such that $F(x) \triangleq \mathsf{P}(X \leq x)$,

(i) A real-value function is a distribution function iff is non-decreasing, right continuous, and $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} F(x) = 1$.

(ii) We may define the inverse of a distribution function F as F^{-1}, whom sometimes also been called as the quantile function,

$$F^{-1}(y) = \inf\{x : F(x) \geq y\}, \quad \text{for } y \in [0,1].$$

Theorem 1.2 (♣ **Inverse Transform Method**)**.** *Let X be a random variable with a known cumulative distribution function (CDF) $F(x)$. If U is uniformly distributed on $[0,1]$, then the random variable $Y \triangleq F^{-1}(U)$ follows the same probability distribution as X.*

Therefore, in order to generate $X_1, X_2, \cdots, X_n$ i.i.d follow the distribution F, we only need to generate $U_1, U_2, \cdots, U_n$ i.i.d follow Uniform$[0,1]$, and simply let $X_i = F^{-1}(U_i)$, $1 \leq i \leq n$.

- The characteristic function of a random variable X is a function $\varphi_X : \mathbb{R} \mapsto \mathbb{C}$, such that $\varphi_X(t) \triangleq \mathsf{E}e^{itX} = \int_{-\infty}^{\infty} e^{itx} dF(x)$, where i here is the imaginary unit satisfying $i^2 = -1$. Every random variable (or distribution function) and a unique characteristic function are one-to-one matched.

  (i) $|\varphi_X(t)| \leq \varphi_X(0) = 1$ and $\varphi_X(-t) = \overline{\varphi_X(t)}$.

  (ii) $\varphi_X(t)$ is uniformly continuous on $\mathbb{R}$.

  (iii) If random variables $X_1, \cdots, X_n$ are mutually independent, and $\eta = \sum_{i=1}^n X_i$, then

  $$\varphi_\eta(t) = \prod_{i=1}^n \varphi_{X_i}(t).$$

  (iv) If $\mathbb{E}|X|^n$ exists and $\varphi_X(t)$ is n-th differentiable, then for $k \leq n$,

  $$\varphi_X^{(k)}(0) = i^k \mathbb{E}X^k.$$

Theorem 1.3 (Bochner-Khinchine Theorem)**.** *A function $\varphi(t)$ is a characteristic function iff $\varphi(t)$ is continuous, $\varphi(0) = 1$, and non-negative definite, that is, for arbitrary real numbers $t_1, \cdots, t_n$ and complex numbers $\lambda_1, \cdots, \lambda_n$, we have*

$$\sum_{i=1}^n \sum_{j=1}^n \varphi(t_i - t_j)\lambda_i \overline{\lambda_j} \geq 0.$$

Theorem 1.4 (♣ **Inverse Formula**)**.** *If $F(x)$ is the distribution function corresponding to the characteristic function $\varphi(t)$, then for two continuous points x_1, x_2 of $F(x)$, we have*

$$F(x_2) - F(x_1) = \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{itx_1} - e^{itx_2}}{it} \varphi(t) dt.$$

Theorem 1.5 (Inverse Fourier Transformation)**.** *If $F(x)$ is the distribution function corresponding to the characteristic function $\varphi(t)$, and $\varphi(t)$ is absolutely integrable, i.e.,*

$$\int_{-\infty}^{\infty} |\varphi(t)| dt < +\infty$$

then the density function, i.e., the derivative of the distribution function, $f(x) = F'(x)$ exists, it's continuous and

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{itx} \varphi(t) dt.$$

2. Some probabilistic identities and inequalities

- ♣ **Event probability**: For arbitrary event $E \in \mathcal{F}$,

$$P(E) = \mathbb{E}(\mathbb{1}\{E\}).$$

- Inclusion-Exclusion formula: For events $E_1, \cdots, E_n$ in a probability space,

$$\mathbb{P}\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{k=1}^{n} \left((-1)^{k-1} \sum_{I \subset \{1,\cdots,n\}, |I|=k} \mathbb{P}(E_I) \right),$$

where the last summation runs over all subsets I of $\{1, \cdots, n\}$ with cardinality k and $E_I = \cap_{i\in I} E_i$.

$*$ For a special case where $E \in \mathcal{F}$, and we have a sequence of events in the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, $\{B_i\}_{i\geq 1}$, which is a partition of Ω, i.e., $\cup_{i\geq 1} B_i = \Omega$ and $B_i \cap B_j = \emptyset$ for $\forall i \neq j$, then

$$\mathbb{P}(E) = \sum_{i\geq 1} \mathbb{P}(E \cap B_i).$$

- ♣ **Bayes' formula**: For a partition of Ω, denoted as $\{B_i\}_{i\geq 1}$, and an event $E \in \mathcal{F}$ for which $\mathbb{P}(E) > 0$, we have

$$\mathbb{P}(B_i|E) = \frac{\mathbb{P}(E|B_i)\mathbb{P}(B_i)}{\mathbb{P}(E)} = \frac{\mathbb{P}(E|B_i)\mathbb{P}(B_i)}{\sum_{j\geq 1}\mathbb{P}(E|B_j)\mathbb{P}(B_j)}$$

- ♣ **Chebyshev's inequality**: Suppose $\varphi : \mathbb{R} \mapsto \mathbb{R}$ is a non-negative function, i.e., $\varphi \geq 0$, let $A \in \mathcal{B}$ and let $i_A = \inf\{\varphi(y) : y \in A\}$. Then

$$i_A \cdot \mathbb{P}(X \in A) \leq \mathbb{E}\big(\varphi(X) \cdot \mathbb{1}(X \in A)\big) \leq \mathbb{E}(\varphi(X)).$$

One special case is, when $\mathbb{E}X^2 < \infty$, then for $\forall\ \epsilon > 0$, we have

$$P(|X - \mathbb{E}X| > \epsilon) \leq \frac{\mathrm{Var}(X)}{\epsilon^2}$$

- ♣ **Cauchy's inequality**: If $\mathbb{E}X^2 < \infty$, $\mathbb{E}Y^2 < \infty$, then

$$\big(\mathbb{E}|XY|\big)^2 \leq (\mathbb{E}X^2)\cdot(\mathbb{E}Y^2).$$

- Hölder's inequality: If $X \in L^p$, i.e., $\mathbb{E}|X|^p < \infty$, $Y \in L^q$, where $p, q \geq 1$ and $1/p + 1/q = 1$, then

$$\|XY\|_1 \leq \|X\|_p \|Y\|_q.$$

$*$ Here $\|\xi\|_r = \big(\mathbb{E}|\xi|^r\big)^{1/r}$ for $r \geq 1$. And for a special case, where $X \in L^q$ for some $q \geq p \geq 1$, then

$$\|X\|_p \leq \|X\|_q.$$

- Minkowski's inequality: If $X, Y \in L^p$ for some $p \geq 1$, then

$$\|X + Y\|_p \leq \|X\|_p + \|Y\|_p.$$

- ♣ **Jensen's inequality**: Suppose φ is convex, that is,

$$\lambda\varphi(x) + (1-\lambda)\varphi(y) \geq \varphi(\lambda x + (1-\lambda)y)$$

for all $\lambda \in (0, 1)$ and $x, y \in \mathbb{R}$. Then

$$\mathbb{E}\big(\varphi(X)\big) \geq \varphi\big(\mathbb{E}X\big)$$

provided both expectations exists, i.e., $\mathbb{E}|X|$ and $\mathbb{E}|\varphi(X)| < \infty$.

- Function of random variables:

Theorem 2.1. *Let $X_1, X_2, \cdots, X_k$ be random variables with joint pdf. $f(x_1, x_2, \cdots, x_k)$, let*

$$\begin{cases} y_1 = g_1(x_1, x_2, \cdots, x_k) \\ y_2 = g_2(x_1, x_2, \cdots, x_k) \\ \quad\vdots \\ y_k = g_k(x_1, x_2, \cdots, x_k) \end{cases}$$

and define the support $\mathcal{X} = \{x = (x_1, x_2, \cdots, x_k)^T : f(x) > 0\}$. Suppose there exists a partition, $A_0, A_1, \ldots, A_m$, of $\mathcal{X}$ such that $\mathbb{P}(X \in A_0) = 0$ and on each $A_i, 1 \leq i \leq m$, there exist functions $g_1^{(i)}(x), \ldots, g_k^{(i)}(x)$, s.t.,

(i) For fixed $1 \leq i \leq m$, we have $g_j^{(i)}(x) = g_j(x)$, for $x \in A_i$, $1 \leq j \leq k$.

(ii) For $1 \leq i \leq m$, $\{g_j^{(i)}, 1 \leq j \leq k\}$ has a unique inverse function $\{(g_j^{(i)})^{-1}, 1 \leq j \leq k\}$, s.t.,

$$\begin{cases} x_1 = (g_1^{(i)})^{-1}(y_1, y_2, \cdots, y_k) \\ x_2 = (g_2^{(i)})^{-1}(y_1, y_2, \cdots, y_k) \\ \qquad \vdots \\ x_k = (g_k^{(i)})^{-1}(y_1, y_2, \cdots, y_k) \end{cases}$$

and a well defined Jacobian matrix

$$|J_i|_+ = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_k} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_k}{\partial y_1} & \frac{\partial x_k}{\partial y_2} & \cdots & \frac{\partial x_k}{\partial y_k} \end{vmatrix}_+ = \begin{vmatrix} \frac{\partial (g_1^{(i)})^{-1}}{\partial y_1} & \frac{\partial (g_1^{(i)})^{-1}}{\partial y_2} & \cdots & \frac{\partial (g_1^{(i)})^{-1}}{\partial y_k} \\ \frac{\partial (g_2^{(i)})^{-1}}{\partial y_1} & \frac{\partial (g_2^{(i)})^{-1}}{\partial y_2} & \cdots & \frac{\partial (g_2^{(i)})^{-1}}{\partial y_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial (g_k^{(i)})^{-1}}{\partial y_1} & \frac{\partial (g_k^{(i)})^{-1}}{\partial y_2} & \cdots & \frac{\partial (g_k^{(i)})^{-1}}{\partial y_k} \end{vmatrix}_+$$

defined for $x = (x_1, x_2, \cdots, x_k)^T \in A_i$, and $y = (y_1, y_2, \cdots, y_k)^T \in \mathcal{Y}_i \triangleq \{y : y_j = g_j(x), 1 \leq j \leq k, x \in A_i\}$.

Then

$$f_Y(y) = \sum_{i=1}^{m} f_X((g_1^{(i)})^{-1}(y), (g_2^{(i)})^{-1}(y), \cdots, (g_k^{(i)})^{-1}(y)) \cdot \mid J_i \mid_+ \cdot \mathbb{1}(y \in \mathcal{Y}_i).$$

Example 2.2. (♣ **Representation Result**) Assume $Z_1, Z_2, \cdots, Z_{n+1}$ are i.i.d random variables with standard exponential distribution. Let

$$Y = (Y_1, \cdots, Y_n) = \left(\frac{Z_1}{\sum_{i=1}^{n+1} Z_i}, \frac{Z_1 + Z_2}{\sum_{i=1}^{n+1} Z_i}, \cdots, \frac{\sum_{i=1}^{n} Z_i}{\sum_{i=1}^{n+1} Z_i} \right),$$

please give the joint distribution of $(Y_1, \cdots, Y_n)$.

3. Modes of Convergence

Denote $\{X_n\}_{n\geq 1}$, $\{Y_n\}_{n\geq 1}$ as two sequence of random variables, and denote X as a random variable, that are all defined in $(\Omega, \mathcal{F}, \mathbb{P})$. Denote c as a constant,

- ***Definition*** 3.1 (♣ **Converge in distribution**) $\{X_n\}_{n\geq 1}$ is said to converge to X in distribution, written as $X_n \overset{d}{\to} X$, if

  $$F_{X_n}(x) = \mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x) = F_X(x),$$

  for all continuity points x of $F_X(x)$.

  Theorem 3.2 (Helly-Bray Theorem)**.** *$X_n \overset{d}{\to} X$, iff*

  $$\mathbb{E}\big(f(X_n)\big) \to \mathbb{E}\big(f(X)\big), \quad \text{for all bounded, continuous functions.}$$

  Theorem 3.3 (Portmanteau Theorem)**.** *The followings are equivalent*

  (i) $X_n \overset{d}{\to} X$.

(ii) $\mathbb{E}\big(f(X_n)\big) \to \mathbb{E}\big(f(X)\big)$ *for all bounded, continuous function f.*

(iii) $\mathbb{E}\big(f(X_n)\big) \to \mathbb{E}\big(f(X)\big)$ *for all bounded, uniformly continuous function f.*

(iv) $\limsup_n \mathbb{P}(X_n \in F) \leq \mathbb{P}(X \in F)$ *for all closed set $F \in \mathcal{B}$.*

(v) $\liminf_n \mathbb{P}(X_n \in E) \geq \mathbb{P}(X \in E)$ *for all open set $E \in \mathcal{B}$.*

(vi) $\mathbb{P}(X_n \in A) \to \mathbb{P}(X \in A)$ *for all X-continuity sets A. Where a set in $\mathcal{B}$ is called a X-continuity set if $\mathbb{P}(X \in \partial A) = 0$, and $\partial A = \bar{A}/A^o$.*

Theorem 3.4 (♣ **Lévy's continuity Theorem**)**.** *Denote the corresponding characteristic functions of $\{X_n\}_{n\geq 1}$ and X as $\{\varphi_{X_n}(t)\}_{n\geq 1}$ and $\varphi_X(t)$ accordingly, then $X_n \overset{d}{\to} X$ iff $\varphi_{X_n}(t) \to \varphi_X(t)$ for $\forall t \in \mathbb{R}$.*

- ***Definition*** 3.5 (♣ **Converge in probability**) $\{X_n\}_{n\geq 1}$ is said to converge to X in probability, written as $X_n \overset{P}{\to} X$, if

$$\mathbb{P}\left(|X_n - X| > \epsilon\right) \to 0 \quad \text{for } \forall \epsilon > 0.$$

Theorem 3.6 (♣ **Continuous Mapping Theorem**)**.** *If function $g : \mathbb{R} \mapsto \mathbb{R}$ is continuous, then $X_n \overset{d}{\to} X$ implies $g(X_n) \overset{d}{\to} g(X)$, and $X_n \overset{P}{\to} X$ implies $g(X_n) \overset{P}{\to} g(X)$.*

Theorem 3.7 (♣ **Slutsky's Theorem**)**.** *If $X_n \overset{d}{\to} X$ and $Y_n \overset{P}{\to} c$, then*

(i) $X_n + Y_n \overset{d}{\to} X + c$.

(ii) $X_n Y_n \overset{d}{\to} cX$.

(iii) $X_n / Y_n \overset{d}{\to} X/c$, *provided that* $c \neq 0$.

(iv) *If* $X_n \overset{P}{\to} X$ *and* $Y_n \overset{P}{\to} Y$, *then* $X_n + Y_n \overset{P}{\to} X + Y$, $X_n Y_n \overset{P}{\to} XY$.

(v) *If* $X_n \overset{L^P}{\to} X$ *and* $Y_n \overset{L^p}{\to} Y$, *then* $X_n + Y_n \overset{L^P}{\to} X + Y$.

- ***Definition*** 3.8 (♣ **Converge in L^p**) $\{X_n\}_{n\geq 1}$ is said to converge to X in L^p norm, written as $X_n \overset{L^p}{\to} X$, if

$$\lim_{n\to\infty} \mathbb{E}|X_n - X|^p = 0 \quad \text{for some } p > 1.$$

- For almost surely convergence, we first define

Definition 3.9. For a sequence of events $\{A_n\}_{n\geq 1}$, we define

$$\begin{aligned}\{\omega : \omega \in A_n, i.o.\} &\triangleq \limsup A_n = \lim_{m\to\infty} \cup_{n=m}^{\infty} A_n \\ &= \{\omega \text{ that are in infinitely many } A_n\},\end{aligned}$$

where the "i.o." represents for infinitely often.

Definition 3.10 (♣ **Almost Surely Convergence**). $\{X_n\}_{n\geq 1}$ is said to converge to X almost surely, written as $X_n \overset{a.s.}{\to} X$, if

$$\mathbb{P}\big(|X_n - X| > \epsilon, i.o.\big) = \mathbb{P}\left(\lim_{m\to\infty} \cup_{n=m}^{\infty} \{|X_n - X| > \epsilon\}\right) = 0, \quad \text{for } \forall \epsilon > 0.$$

Theorem 3.11 (Borel-Cantelli Lemma)**.** *If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, then*

$$\mathbb{P}(A_n, i.o.) = 0.$$

Theorem 3.12 (The second Borel-Cantelli Lemma)**.** *If the events $\{A_n\}_{n\geq 1}$ are mutually independent, and $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$, then $\mathbb{P}(A_n, i.o.) = 1$.*

- ♣ **Relations between different mode of convergence**:

  Definition 3.13 (Uniformly integrable). A sequence of random variables $\{X_n\}_{n\geq 1}$ is said to be uniformly integrable if for $\forall \epsilon > 0$, there exists a $K > 0$, s.t.,
$$\sup_{n\geq 1} \mathbb{E}\Big(|X| \cdot \mathbb{1}(|X| \geq K)\Big) \leq \epsilon.$$

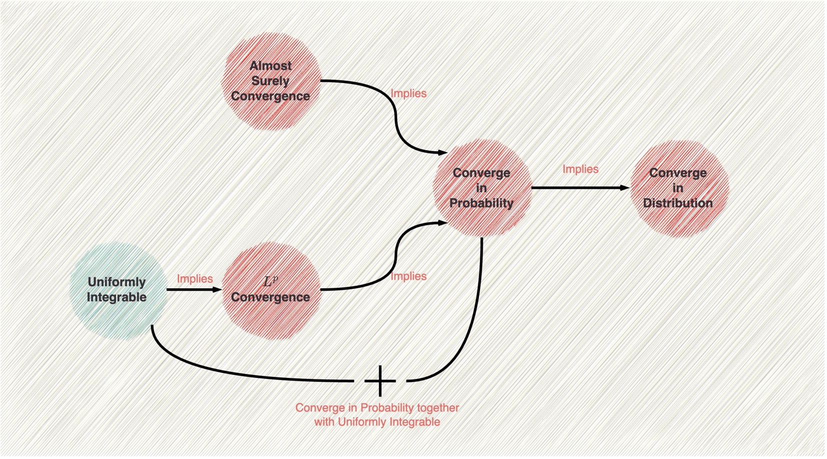

Figure 1: Relations between different mode of convergence

Theorem 3.14. *Relations between different mode of convergence:*

(i) $X_n \overset{L^p}{\to} X$ *implies* $X_n \overset{P}{\to} X$.

(ii) $X_n \overset{a.s.}{\to} X$ *implies* $X_n \overset{P}{\to} X$.

(iii) $X_n \overset{P}{\to} X$ *implies* $X_n \overset{d}{\to} X$.

(iv) *$X_n \xrightarrow{d} c$ implies $X_n \xrightarrow{P} c$, where c is a constant.*

(v) *If $X_n \xrightarrow{P} X$, and $\{|X_n|^p\}_{n\geq 1}$ is uniformly integrable, then $X_n \xrightarrow{L^p} X$.*

4. Law of Large numbers and Central Limit Theorems

Theorem 4.1 (♣ **Weak Law of Large Numbers**)**.** *Let $X_1, X_2, \cdots, X_n$ be mutually independent and identically distributed with $\mathbb{E}|X_i| < \infty$. Let $\mathbb{E}X_i = \mu$ and $S_n = X_1 + \cdots + X_n$. Then $S_n/n \xrightarrow{P} \mu$ as $n \to \infty$.*

Theorem 4.2 (♣ **Strong Law of Large Numbers**)**.** *Let $X_1, X_2, \cdots, X_n$ be mutually independent and identically distributed with $\mathbb{E}|X_i| < \infty$. Let $\mathbb{E}X_i = \mu$ and $S_n = X_1 + \cdots + X_n$. Then $S_n/n \to \mu$ a.s. as $n \to \infty$.*

Theorem 4.3 (♣ **Central Limit Theorem (Linderberg-Lévy)**)**.** *Let $X_1, X_2, \cdots, X_n$ be mutually independent and identically distributed with $\mathbb{E}|X_i|^2 < \infty$. Let $\mathbb{E}X_i = \mu$, $\operatorname{Var} X_i = \sigma^2$ and $S_n = X_1 + \cdots + X_n$. Then*

$$(S_n - n\mu)/(\sigma n^{1/2}) \xrightarrow{d} N(0,1).$$

Theorem 4.4 (Central Limit Theorem (Linderberg-Feller))**.** *Consider a triangular array, where for each n, let $X_{n,1}, X_{n,2}, \cdots, X_{n,n}$ be mutually independent random variables with $\mathbb{E}X_{n,m} = 0$ for $m = 1, \cdots, n$. Suppose*

(i) $\sum_{m=1}^{n} \mathbb{E}X_{n,m}^2 \to \sigma^2 > 0$.

(ii) *For all $\epsilon > 0$, $\lim_{n\to 0} \sum_{m=1}^{n} \mathbb{E}\big[X_{n,m}^2 \cdot \mathbb{1}(|X_{n,m}| > \epsilon)\big] = 0$*

Define $S_n = X_{n,1} + \cdots + X_{n,n}$. Then

$$S_n/\sigma \xrightarrow{d} N(0,1).$$

5. Delta Method

Conside an estimation problem, where we are fortunate to found an estimator $\hat{\theta}_n$ based on our observated data for the parameter of interests θ, and we are blessed to have $\sqrt{n}(\hat{\theta}_n - \theta)/\sigma \xrightarrow{d} N(0,1)$. Then for some continuous function g,

Theorem 5.1 (♣ **Delta Method**)**.** *Suppose $\sqrt{n}(\hat{\theta}_n - \theta)/\sigma \xrightarrow{d} N(0,1)$, and g is a continuous function,*

1. *(First order delta method) If g' exists and is continuous, with $g'(\theta) \neq 0$, then*
$$\sqrt{n}\left(\frac{g(\hat{\theta}) - g(\theta)}{|g'(\theta)|\sigma}\right) \xrightarrow{d} N(0,1).$$

2. *(Second order delta method) If g'' exists and is continuous, with $g'(\theta) = 0$ and $g''(\theta) \neq 0$, then*
$$n\left(\frac{g(\hat{\theta}) - g(\theta)}{\frac{1}{2}g''(\theta)\sigma^2}\right) \xrightarrow{d} \chi_1^2.$$

Definition 5.2 (♣ Random Sample). Random sample could mean differently under different contexts. Here, we specify that, whenever we say a random sample (or simple random sample), we mean a sample with its elements being i.i.d random variables.

Example 5.3 (♣ **Self-normalization using Delta's method**). Let $X_1, \cdots, X_n$ be a random sample from $N(\theta, \theta)$, $\theta > 0$. Apparently, $\bar{X} = (\sum_{i=1}^n X_i)/n$ is a "good" estimator of θ, please give an interval estimator for θ based on $\bar{X}$.

6. Exponential Family

6.1. Definition and some examples of exponential family

Definition 6.1 (♣ **Exponential family**). A family of pdfs or pmfs is called an exponential family if it can be expressed as

$$f(x|\theta) = h(x)c(\theta)\exp\left(\sum_{i=1}^{k}\omega_i(\theta)T_i(x)\right)$$

Where $h(x) \geq 0$ and $T_1(x), \cdots, T_k(x)$ are real valued functions of the observation $x = (x_1, \cdots, x_n)$ which does not depend on θ. And, $c(\theta) \geq 0$ and $\omega_1(\theta), \cdots, \omega_k(\theta)$ are real valued functions of the parameter $\theta = (\theta_1, \cdots, \theta_m)$ which does not depend on x. Besides,

$$\Theta \triangleq \left\{\theta : c(\theta) \geq 0,\ \omega_i(\theta) \text{ being well defined for } 1 \leq i \leq k\right\}$$

is the parameter space of this exponential family, and this exponential family is called to have dimension k (i.e., full rank) if Θ contains a open set in $\mathbb{R}^k$.

Remark 6.2. Let $\omega_i = \omega_i(\theta)$, if there exist some function $b(\cdot)$ such that $c(\theta) \equiv b(\omega)$, which infers that the density equals to

$$f(x|\omega) = h(x)b(\omega)\exp\left(\sum_{i=1}^{k}\omega_i T_i(x)\right) \tag{6.1}$$

then we call this reparameterized density the **canonical form** of the exponential family.

Remark 6.3. If Θ does not contain a open set in $\mathbb{R}^k$. Instead, Θ contains a open set in $\mathbb{R}^s$ for some $s < k$, then we say this exponential family has dimension s and this is a **curved exponential family**.

Example 6.4. (♣ **Normal exponential family**). Say we have a random sample $X_1, \cdots, X_n$ i.i.d. came from an normal distribution $\mathcal{N}(\mu, \sigma^2)$ where $\theta = (\mu, \sigma^2)$ are unknown. Then the sample forms an exponential family.

Example 6.5. (♣ **Poisson exponential family**). $X_1, \cdots, X_n$ are i.i.d. from Poisson distribution $Poisson(\lambda)$ where λ is unknown. Then the sample forms an exponential family.

Example 6.6. (♣ **Gamma exponential family**). Let $X_1, \ldots, X_n$ be a random sample from a $Gamma(\alpha, \beta)$ population with unknown $\theta = (\alpha, \beta)$. Then the sample forms an exponential family.

Example 6.7. **t_θ, Student t distribution with degree of freedom θ**. Let $X_1, \ldots, X_n$ be a random sample from a t_θ with unknown $\theta \in \mathbb{R}^+$. Then the sample does not form an exponential family.

Example 6.8. (♣ **Curved exponential family**). Let $X_1, \ldots, X_n$ be a random sample from $N(\theta, \theta^3)$ with unknown $\theta \in \mathbb{R}^+$. Then the sample forms an curved exponential family.

6.2. Some properties of exponential family

In this subsection, we consider the canonical form of the exponential family $f(x|\omega)$ defined in (6.1) if no other explanation provided.

Theorem 6.9 (♣ **Differentiation of exponential family density**). *Denote the parameter space of a canonical expoential family $f(x|\omega)$ as Θ, then for any integrable function $g(x)$, i.e.,*

$$\int g(x) \cdot h(x) b(\omega) \exp\left(\sum_{i=1}^{k} \omega_i T_i(x) \right) dx < \infty, \tag{6.2}$$

and for any ω_0 in the interior of Θ, i.e., there exists some $\epsilon > 0$, such that

$$\Big\{ \omega : \|\omega - \omega_0\|_2 < \epsilon \Big\} \subset \Theta,$$

we have the integral (6.2) is continuous and has derivatives of all orders with respect to ω_0, and this can be obtained by differentiating under the integral sign.

Remark 6.10. For a special case where $g(x) \equiv 1$, we differentiate the identity

$$\int h(x) b(\omega) \exp\left(\sum_{i=1}^{k} \omega_i T_i(x) \right) dx = 1$$

with respect to ω_i gives

$$\mathbb{E} T_i(X) = -\frac{1}{b(\omega)} \frac{\partial b(\omega)}{\partial \omega_i} = -\frac{\partial \log b(\omega)}{\partial \omega_i}, \quad \text{for } i = 1, \cdots, k. \tag{6.3}$$

Similarly, we differentiate the identity (6.3) with respect to ω_j gives

$$\mathrm{Cov}(T_i(X), T_j(X)) = -\frac{\partial^2 \log b(\omega)}{\partial \omega_i \partial \omega_j}.$$

Theorem 6.11 (Stein's identity)**.** *If X is a random variable distributed with density*

$$f(x|\omega) = h(x)b(\omega)\exp\left(\sum_{i=1}^{k}\omega_i T_i(x)\right).$$

For any differentiable function g, if the support of X is $(-\infty, \infty)$, $f(x|\omega)$ satisfy that $\lim_{x\to\infty} f(x|\omega) = \lim_{x\to-\infty} f(x|\omega) = 0$, then

$$E\left\{\left[\frac{h'(X)}{h(X)} + \sum_{i=1}^{k}\omega_i T_i'(X)\right]g(X)\right\} = -Eg'(X)$$

if it's provided that $\mathbb{E}|g'(X)| < \infty$ and $\mathbb{E}\left|\frac{f'(X|\omega)}{f(X|\omega)}g(X)\right| < \infty$.

Example 6.12 (♣ **Stein's identity for normal distribution**). If $X \sim N(\mu, \sigma^2)$, then Stein's identity implies that for suitable function $g(\cdot)$ satisfy the condition in Theorem 6.11, we have

$$\mathbb{E}\Big[g(X)(X-\mu)\Big] = \sigma^2\mathbb{E}g'(X).$$

This immediately shows that $\mathbb{E}X = \mu$ (if we take $g(x) = 1$) and $\mathbb{E}X^2 = \sigma^2 + \mu^2$ (if we take $g(x) = x$). Higher-order moments are equally easy to calculate.

6.3. Score function, Fisher Information and The Second Bartlett's Identities

Recall that the score function (Fisher score function) is defined as the partial derivative of the log-likelihood, it measures the sensitivity of log-likelihood $\log f(x|\theta)$ to its parameter θ,

Definition 6.13 (♣ **Score function**). The score function of a likelihood function $L(\theta|x)$ is defined as

$$s(\theta|x) = \frac{\partial \log L(\theta|x)}{\partial\theta}.$$

Notice that the defnition does not ensure the existence of score function. Only when the partial derivative of log-likelihood exists, we call it the score function. Similarly, when the second moment of score function exists, we call it the Fisher Information.

Definition 6.14 (♣ **Fisher Information**). The Fisher Information of a joint probability distribution function $f(x|\theta)$ is defined as

$$I(\theta) = \mathbb{E}\left[\frac{\partial \log f(x|\theta)}{\partial\theta}\right]^2 = \mathbb{E}\left[s(\theta|x)\right]^2.$$

- ♣ **Unbiasedness of the score function.**

  Now, assume the joint probability distribution function $f(x|\theta)$ forms an expoential family. Then for arbitray θ in the interior of our parameter space Θ, by taking integrable function $g(x) \equiv 1$, we have

  $$\begin{aligned}\mathbb{E}\big[s(\theta|x)\big] &= \int \frac{\partial \log f(x|\theta)}{\partial\theta} f(x|\theta)dx = \int \frac{1}{f(x|\theta)}\frac{\partial f(x|\theta)}{\partial\theta} f(x|\theta)dx \\ &= \int \frac{\partial f(x|\theta)}{\partial\theta} dx = \frac{\partial}{\partial\theta}\int f(x|\theta)dx = \frac{\partial}{\partial\theta} 1 = 0. \qquad (6.4)\end{aligned}$$

  according to Theorem 6.9. Equation (6.4) means that score function is an unbiased estimator of 0, which further infers that we can construct an **Estimating equation** like

  $$\hat{\theta} \text{ satisfies } \frac{1}{n}\sum_{i=1}^{n} s(\hat{\theta}|x_i) = 0.$$

- ♣ **The second Bartlett's Identities.**

  Meanwhile, notice that

  $$\frac{\partial^2 \log f(x|\theta)}{(\partial\theta)^2} = \frac{\frac{\partial^2}{(\partial\theta)^2} f(x|\theta)}{f(x|\theta)} - \left(\frac{\frac{\partial}{\partial\theta} f(x|\theta)}{f(x|\theta)}\right)^2 = \frac{\frac{\partial^2}{(\partial\theta)^2} f(x|\theta)}{f(x|\theta)} - \left(\frac{\partial}{\partial\theta}\log f(x|\theta)\right)^2.$$

  And for arbitray θ in the interior of our parameter space Θ, by taking integrable function $g(x) \equiv 1$, Theorem 6.9 ensures that

  $$\mathbb{E}\left[\frac{\frac{\partial^2}{(\partial\theta)^2} f(x|\theta)}{f(x|\theta)}\right] = \int \frac{\partial^2}{(\partial\theta)^2} f(x|\theta)dx = \frac{\partial^2}{(\partial\theta)^2}\int f(x|\theta)dx = 0.$$

  which concludes that

  $$I(\theta) = \mathbb{E}\left[\frac{\partial \log f(x|\theta)}{\partial\theta}\right]^2 = -\mathbb{E}\left[\frac{\partial^2 \log f(x|\theta)}{\partial\theta^2}\right]. \qquad (6.5)$$

  Equation (6.5) is called the Second Bartlett's Identities.

However, not every distribution would satisfy these two magnificent properties. For instance,

Example 6.15. Assume $X \sim \text{Uniform}[0, \theta]$, the score function of X don't have the unbiasedness.

* Therefore, in practice, we would usually impose the following **Fisher Information regularity condition** for $\mathcal{F} = \{f(x|\theta), \theta \in \Theta\}$ (or C-R regularity condition) and restrict our attention to such distributions,

1. Θ is open regard to its topology (In order to make the differential well defined, obviouly if we are consider Θ as a subset of $\mathbb{R}$, then just need the boundary of Θ is a null set).
2. the distributions in $\mathcal{F}$ has same support. (Classical example, the information theorey does not working on $\{\text{Uniform}(0,\theta) : \theta > 0\}$ even though we can still define the fisher information, but it's generally off interests).
3. For $\forall x \in \mathcal{X}$ and $\theta \in \Theta$, $\frac{\partial f(x|\theta)}{\partial \theta}$ exist.
4. The following equation holds and the term on the both side of the equation are well defined.

$$\frac{\partial}{\partial \theta} \int f(x|\theta) dx = \int \frac{\partial}{\partial \theta} f(x|\theta) dx$$

5. The following expectation exist and

$$0 < I(\theta) = E_\theta \left[\frac{\partial \log f(x|\theta)}{\partial \theta} \right]^2 < \infty$$

These are called the **Fisher Information regularity condition** or $C-R$ regularity condition and a family satisfy the above condition is called the $C-R$ regular family, $I(\theta)$ is called the Fisher information of this family. Apparently, in order to have the second Bartlett's Indentities, we sometimes further impose

6. For $\forall x \in \mathcal{X}$ and $\theta \in \Theta$, $\frac{\partial^2 f(x|\theta)}{(\partial \theta)^2}$ exist, and the following equation holds while the term on the both side of the equation are well defined.

$$\frac{\partial^2}{(\partial \theta)^2} \int f(x|\theta) dx = \int \frac{\partial^2}{(\partial \theta)^2} f(x|\theta) dx.$$

7. Location-scale Family

7.1. Definition and some examples of Location-Scale family

Let U be a random variable with a fixed distribution F. If a constant a is added to U, the resulting variable

$$X = U + a \tag{7.1}$$

has distribution

$$P\left(X \leq x\right) = F\left(x - a\right). \tag{7.2}$$

The totality of distribution (7.2), for fixed F and as a varies from $-\infty$ to ∞, is said to constitute a *location family.* Analogously, a *scale family* is generated by the transformations

$$X = bU, \quad b > 0, \tag{7.3}$$

and has the form

$$P\left(X \le x\right) = F\left(x/b\right).$$

Combining these two types of transformations into

$$X = a + bU, \quad b > 0, \tag{7.4}$$

one obtains the *location-scale* family

$$P\left(X \le x\right) = F\left(\frac{x-a}{b}\right). \tag{7.5}$$

Rigorously, we have

Definition 7.1 (♣ **Location-Scale family**). For a known distribution function F, the location family, the scale family, and the location-scale family generated from F are

$$\mathcal{F}_l = \left\{\tilde{F}(x) : \tilde{F}(x) = F(x-a), \forall a \in \mathbb{R}\right\},$$

$$\mathcal{F}_s = \left\{\tilde{F}(x) : \tilde{F}(x) = F(x/b), \forall b > 0\right\},$$

$$\mathcal{F} = \left\{\tilde{F}(x) : \tilde{F}(x) = F\left(\frac{x-a}{b}\right), \forall a \in \mathbb{R}, b > 0\right\}.$$

In application of these families, F usually has a density f with respect to Lebesgue measure. The density of (7.5) is then given by

$$\frac{1}{b} f\left(\frac{x-a}{b}\right).$$

The following table exhibits several such densities.

Table 1
Example of Location-Scale families

Density	**Support**	**Name**
$\frac{1}{\sqrt{2\pi b}} e^{-(x-a)^2/2b^2}$	$-\infty < x < \infty$	Normal
$\frac{1}{2b} e^{-\lvert x-a\rvert/b}$	$-\infty < x < \infty$	Double exponential
$\frac{b}{\pi} \frac{1}{b^2+(x-a)^2}$	$-\infty < x < \infty$	Cauchy
$\frac{1}{b} \frac{e^{-(x-a)/b}}{[1+e^{-(x-a)/b}]^2}$	$-\infty < x < \infty$	Logistic
$\frac{1}{b} e^{-(x-a)/b} I_{[a,\infty)}(x)$	$-a < x < \infty$	Exponential
$\frac{1}{b} I_{[a-b/2,a+b/2]}(x)$	$-a - \frac{b}{2} < x < a + \frac{b}{2}$	Uniform

Definition* 7.2 (Generalized Location-Scale family). A family of distributions $\mathcal{F}$ is called a generalized location-scale family if for $\forall F \in \mathcal{F}$, implies that

$$F\Big(\frac{x-a}{b}\Big) \in \mathcal{F},$$

for $\forall a \in \mathbb{R}$ and $b > 0$.

*7.2. *Transformation Group and Invariant Family*

For a fixed distribution function $F(\cdot)$, and a random variable $U \sim F$, we can build a location-scale family

$$\mathcal{F} = \left\{ F\left(\frac{x-a}{b}\right), \forall a \in \mathbb{R}, b > 0 \right\},$$

and for each element of $\mathcal{F}$, say $F((x-a)/b)$, we have variable $X = a + bU$ following this distribution. Now, if we define $\mathcal{X} = \{a + bU : a \in \mathbb{R}, b > 0\ \}$, then we know there is a one-to-one mapping between $\mathcal{F}$ and $\mathcal{X}$.

Define a set of functions (transformations) $\mathcal{G} = \{g : \mathcal{X} \mapsto \mathcal{X} \mid g(x) = a + bx\}$. Then for $\forall g_1(x) = a_1 + b_1 x, g_2 = a_2 + b_2 x \in \mathcal{G}$, we have

$$g_2 \circ g_1(x) = g_2(g_1(x)) = a_2 + b_2(a_1 + b_1 x) = (a_2 + b_2 a_1) + b_2 b_1 x,$$

which implies that $g_2 \circ g_1 \in \mathcal{G}$, we say the class $\mathcal{G}$ is closure under composition. Besides,

$$g_1^{-1}(y) = \frac{y - a_1}{b_1} = -\frac{a_1}{b_1} + \frac{1}{b_1} y,$$

which implies that $g_1^{-1} \in \mathcal{G}$, we say the class $\mathcal{G}$ is closure under inversion.

Definition 7.3. (Transformation Group) A class $\mathcal{G}$ of transformations is called a transformation group if it is closed under both comosition and inversion.

Apparently, for aribitrary $X = a_1 + b_1 U \in \mathcal{X}$ and $g(x) = a_2 + b_2 x \in \mathcal{G}$, we have $g(X) \in \mathcal{X}$. Thus, we say $\mathcal{X}$ is closure under the transformation group $\mathcal{G}$.

Consider a family of densities $\mathcal{F} = \{f(x|\theta) : \theta \in \Theta\}$, we come up with a statistic $T(X)$ to estimate θ and a loss function $L(\theta, T)$ to evalute this statistic and the estimation.

Definition 7.4. (Location invariant). We say the density family

$$\mathcal{F} = \{f(x|\theta) : \theta \in \Theta\}$$

and the loss function $L(\theta, T)$ are location invariant if, respectively, we have $f(x|\theta) = f(x'|\theta')$ and $L(\theta, T(x)) = L(\theta', T(x'))$ whenever $x' = x + a$, $\theta' = \theta + a$ and $T(x') = T(x + a) = T(x) + a$. If both the densities and the loss function are location invariant, the problem of estimating θ is said to be location invariant under the transformation $\{g : g(x) = x + a, a \in \mathbb{R}\}$.

Within the definition of location invariant, we made the assumption that we can found some statistics T, such that $T(x') = T(x+a) = T(x)+a$ as $x' = x+a$. More generally, for a sample (for simplicity, let assume it is a random sample) $X_1, \cdots, X_n$, we may hopping to found the statistics T such that

$$T(X_1 + a, \cdots, X_n + a) = T(X_1, \cdots, X_n) + a \tag{7.6}$$

Definition 7.5*.* (Location equivariant estimator). An estimator satisfying (7.6) will be called an location equivariant estimator.

The above argument can be parallelly extended to scale equivariant estimator.

8. *Supplement

Here we list some of the proofs of theorems listed before.

Proof. (of Theorem.6.11) Notice that,

$$\begin{aligned}
\int_{-\infty}^{0} g'(x) f(x|\omega) dx &= \int_{-\infty}^{0} g'(x) \int_{-\infty}^{x} f'(y|\omega) dy dx \\
&= \lim_{a\to\infty} \int_{-a}^{0} g'(x) \int_{-a}^{x} f'(y|\omega) dy dx \\
&= \lim_{a\to\infty} \int_{-a}^{0} f'(y|\omega) \int_{y}^{0} g'(x) dx dy \\
&= \lim_{a\to\infty} \int_{-a}^{0} f'(y|\omega) \left[g(0) - g(y)\right] dy \\
&= \lim_{a\to\infty} \left[f(0|\omega) g(0) - f(-a|\omega) g(0) - \int_{-a}^{0} f'(y|\omega) g(y) dy \right] \\
&= f(0|\omega) g(0) - \int_{-\infty}^{0} f'(x|\omega) g(x) dx
\end{aligned}$$

Similarly, using the fact that $f(x|\omega) = -\int_x^\infty f'(y)dy$, we have the other side as

$$\int_{0}^{+\infty} g'(x) f(x|\omega) dx = -f(0|\omega) g(0) - \int_{0}^{+\infty} f'(x|\omega) g(x) dx.$$

Combine the above two equations, we conclude that

$$\begin{aligned}
-\mathbb{E} g'(X) &= -\int g'(x) f(x|\omega) dx = \int f'(x|\omega) g(x) dx = \mathbb{E}\left[\frac{f'(X|\omega)}{f(X|\omega)} g(X)\right] \\
&= E\left\{ \left[\frac{h'(X)}{h(X)} + \sum_{i=1}^{k} \omega_i T_i'(X) \right] g(X) \right\}.
\end{aligned}$$

□