Contents

1 Useful Analysis results

[Power set] 2^Ω power set := all subsets of Ω, trivial σ-field

[Cartesian product] $A \times B$ Cartesian product of two sets $A, B := \{(a, b) : a \in A, b \in B\}$, two spaces $\mathcal{F}_1 \times \mathcal{F}_2 := \{A_i \times B_i : A_i \in \mathcal{F}_1, B_i \in \mathcal{F}_2\}$

[Sets under mappings] For any function $f : \Omega \to \Lambda$ and any index set I
(a) $f\left(\cup_{\alpha\in I} A_\alpha\right) = \cup_{\alpha\in I} f(A_\alpha)$ where $A_\alpha \subset \Omega$
(b) $f^{-1}\left(\cup_{\alpha\in I} B_\alpha\right) = \cup_{\alpha\in I} f^{-1}(B_\alpha)$ where $B_\alpha \subset \Lambda$
(c) $f^{-1}(B^c) = (f^{-1}(B))^c$ where $B \subset \Lambda$

[Checking convexity] If f is twice-differentiable, then convexity of f is implied by positive semi-definiteness of Hessian $[\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0]$. checked (1) Non-negative eigenvalues $[\det(\mathbf{A} - \lambda\mathbf{I}) = 0]$, (2) Leading principal minors are positive $[\det(\triangle_k) > 0]$

[Matrix operations] $c^T c = c_1^2 + \cdots + c_k^2$, cc^T is $k \times k$ matrix with (i, j)th element as $c_i c_j$

[Max function] $\max(a, b) = \frac{a+b-|a-b|}{2}$

[Set operation] $\cup_i (C \cap A_i) = C \cap (\cup_i A_i) \in \mathcal{F}_C$
$f(\omega) = \int I_{(0,f(\omega))}(t) dm(t)$

2 Probability Theory

2.1 Measure Space $(\Omega, \mathcal{F}, \nu)$ and Measurable functions

2.1.1 measure spaces

Ω := sample space, $(\Omega, \mathcal{F})$:= measurable space, $A_i \in \mathcal{F}$:= $\mathcal{F}$-measurable, $(\Omega, \mathcal{F}, \nu)$:= measure space
if $\nu(\Omega) = 1$:= probability space, $\nu(A) = P(A)$:= probability of event A

2.1.2 σ-field

A collection $\mathcal{F}$ of subsets of a set Ω is called a σ-field if
(1) $\emptyset \in \mathcal{F}$ (2) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$ (3) $A_i \in \mathcal{F} \Rightarrow \cup_{i=1}^\infty A_i \in \mathcal{F}$

2.1.3 smallest σ-field

Given collection $\mathcal{C}$, there exists a smallest σ-field $\mathcal{F} := \sigma(\mathcal{C})$ s.t.
(1) $\mathcal{C} \subset \mathcal{F}$ (2) if $\mathcal{E}$ is a σ-field containing $\mathcal{C}$, then $\mathcal{F} \subset \mathcal{E}$

2.1.4 Borel σ-field

$\mathcal{B} := \sigma(\mathcal{O})$ or $\mathcal{B}^d = \sigma(\mathcal{O}^d)$ where $\Omega = \mathcal{R}$, $\mathcal{O}$ is the collection of all open sets

2.1.5 measure

A (positive) measure ν on a measurable space $(\Omega, \mathcal{F})$ is a non-negative function $\nu : \mathcal{F} \to \mathcal{R}$ s.t.
(1. non-negativity) $0 \leq \nu(A) \leq \infty \ \forall\ A \in \mathcal{F}$
(2. empty is zero) $\nu(\emptyset) = 0$
(3. σ-additivity) $\sum_{i=1}^\infty \nu(A_i)$ $\nu(\cup_{i=1}^\infty) A_i$ if $A_i \in \mathcal{F}$ are disjoint

2.1.6 common measure

(counting measure $\nu(A)$) number of elements in $A\ \forall\ A \in \Omega$, can be ∞
(Lebesgue measure $m([a,b])$) $(b-a)I(a<b)$ default for countable sets in $(\mathcal{R},\mathcal{B})$, is σ-finite as $\exists A_i = [-i,i]$

2.1.7 measure properties

(1. Monotonicity) $A \subset B \Rightarrow \nu(A) \leq \nu(B)$
(2. Sub-additivity) any sequence of potentially non-disjoint set A_n, $\nu(\cup_{n=1}^{\infty} A_n) \leq \sum_{n=1}^{\infty} \nu(A_n)$
(3. Continuity of Increasing sequences) $\lim_{n\to\infty} A_n := \cup_{n=1}^{\infty} A_n$ and $\nu(\lim_{n\to\infty} A_n) = \lim_{n\to\infty} \nu(A_n)$
(4. Continuity of Decreasing sequences) $\lim_{n\to\infty} A_n := \cap_{n=1}^{\infty} A_n$, and if $\nu(A_1) < \infty$ then $\nu(\lim_{n\to\infty} A_n) = \lim_{n\to\infty} \nu(A_n)$

2.1.8 σ-finite

measure is σ-finite if $\exists$ a seq of measurable sets A_i s.t. $\cup_{i=1}^{\infty} A_i = \Omega$ and $\nu(A_i) < \infty\ \forall\ i$

2.1.9 product

σ-field $\prod_{i=1}^{d} \mathcal{F}_i]$ $(\Omega_1 \times \Omega_2 \times \cdots \Omega_d, A_1 \times \cdots \times A_d : A_i \in \mathcal{F}_i)$

2.1.10 product measure

Suppose $(\Omega_i, \mathcal{F}_i, \nu_i)$ are measure spaces and ν_i are all σ-finite. There exists a unique σ-field measure on product σ-field s.t. $\forall A_i \in \mathcal{F}_i$

$$\nu_i \times \cdots \times \nu_d(A_1 \times \cdots \times A_d) = \prod_{i=1}^{d} \nu_i(A_i)$$

2.1.11 measurable function

Let $(\Omega,\mathcal{F}),(\Lambda,\mathcal{G})$ be measurable spaces and $f : \Omega \to \Lambda$ be a function. The function f is measurable function from $(\Omega,\mathcal{F})$ to $(\Lambda,\mathcal{G})$ if

$$f^{-1}(A) \in \mathcal{F} \text{ for all } A \in \mathcal{G}$$

Note $f^{-1}(\mathcal{G}) = \sigma(f) = \{f^{-1}(A) : A \in \mathcal{G}\}$ is a sub-σ-field to $\mathcal{F}$
(Richness) addition, multiplication, division, composition, sup, limit preserve measurability

2.1.12 simple function

with a finite k, $c_i \geq 0, A_i \in \mathcal{F}$

$$f(\omega) := \sum_{i=1}^{k} c_i I_{A_i}(\omega)$$

2.1.13 approximation by simple function

Any non-negative Borel function f can be approximated by s sequence of increasing simple functions φ_n and $\lim_n \varphi_n(x) = f(x)$ for every $x \in \Omega$

$$\varphi_n = \sum_{i=0}^{n2^n-1} \frac{i}{2^n} I\left(\frac{i}{2^n} \leq f \leq \frac{i+1}{2^n}\right) + nI(f \geq n)\ \forall\ n \in \mathcal{N}$$

2.1.14 measurable function in probability

(random element) denoted as X, (random variable) if X is real valued, (random vector) if X is vector-valued

2.1.15 a.e., a.s., w.p

A is null set if $\nu(A) = 0$ (almost everywhere) statement holds ν-a.e. if statement holds for all $\omega \in A^c$ (almost surely) if ν is a probability (with probability 1) alternatively, we say the statement holds w.p. 1

2.2 Integration and Expectation

[Integral for simple function] ∞ is allowed and $c \times \infty = \infty$ if $c > 0$ else 0

$$\int f d\nu := \sum_{i=1}^{k} c_i \nu(A_i)$$

2.2.1 Integral for non-negative Borel functions

$\mathcal{S}_f :=$ collection of all non-negative simple functions s.t. $g \le f$ if $g \in \mathcal{S}_f$

$$\int f d\nu := \sup\left\{\int f d\nu : g \in \mathcal{S}_f\right\} = \lim_n \int \varphi_n d\nu$$

if $0 \le \varphi_0 \le \varphi_1 \le \cdots \le f$ and $\lim_n \varphi_n = f$

2.2.2 Integral for arbitrary Borel functions

$f = f_+ - f_-$, $f_+ = \max\{f(x), 0\}$, $f_- = \max\{-f(x), 0\}$

$$\int f d\nu := \int f_+ d\nu - \int f_- d\nu$$

when both f_+, f_- are finite, f is integrable

2.2.3 integral over subset, and notation

I_A is measurable, so is the product $I_A f$. If it exist, then $\int_A f d\nu = \int I_A f d\nu$

$$\int f d\nu = \int_\Omega f d\nu = \int f(x) d\nu(x) = \int f(x)\nu(dx)$$

2.2.4 Expectation

for a probability measure P

$$EX = E(X) = \int X dP$$

2.2.5 Expectation properties

(Linearity) $E(aX + bY) = aEX + bEY$
(Absolute finite) EX is finite if and only if $E|X|$ is finite
(Order) if $X \ge 0$ a.s. then $EX \ge 0$, if $X \le Y$ a.s. then $EX \le EY$, if $X = Y$ a.s. then $EX = EY$
(Absolute order) $|EX| \le E|X|$
(Deduce $X = 0$) If $X \ge 0$ a.s. and $EX = 0$ then $X = 0$ a.s.

2.3 Convergence Theorem

Let $\{f_n\}_{n=1}^\infty$ be a sequence of Borel functions on $(\Omega, \mathcal{F}, \nu)$

2.3.1 Monotone convergence theorem

if $0 \le f_1 \le f_2 \le \cdots$ and $\lim_n f_n = f$ a.e. then

$$\int \lim_n f_n d\nu = \lim_n \int f_n d\nu$$

2.3.2 Fatou's lemma

If $f_n \ge 0$

$$\int \liminf_n f_n d\nu \le \liminf_n \int f_n d\nu$$

2.3.3 Dominated convergence theorem

If $\lim_{n\to\infty} f_n = f$ and $\exists$ integrable function g s.t. $|f_n| \le g$ a.e.

$$\int \lim_n f_n d\nu = \lim_n \int f_n d\nu$$

2.4 Change of variables

2.4.1 Interchange differentiation and Integration

(1) Suppose $\exists\ (a, b) \subset \mathcal{R}$ which $\partial f(\omega, \theta)/\partial\theta$ exists a.e. (2) There is an integrable function g on ω s.t. $|\partial f(\omega, \theta/\partial\theta| \le g(\omega)$ a.e.

$$\frac{d}{d\theta}\int f(\omega, \theta) d\nu(\omega) = \int \frac{\partial f(\omega, \theta)}{\partial\theta} d\nu(\omega)$$

2.4.2 Change of variable

Let $(\Omega, \mathcal{F}, \nu)$ be measure space, $(\Lambda, \mathcal{G})$ be measurable space, f is measurable from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$, f induce a measure $\nu \circ f^{-1}(B) := \nu(f^{-1}(B)) \in \Lambda \ \forall \ B \in \mathcal{G}$. Suppose g is Borel function on $(\Lambda, \mathcal{G})$,

$$\int_\Omega g \circ f d\nu = \int_\Lambda g d(\nu \circ f^{-1})$$

2.4.3 Change of Var Formula

$Y = g(X)$, A_i disjoint, h_j is inverse function of g on A_j.

$$f_Y(y) = \sum_{j:1 \le j \le m, y \in g(A_j)} \left| \det \left(\frac{\partial h_j(y)}{\partial y} \right) \right| f_X(h_j(y))$$

2.5 Cumulative Distribution Function

2.5.1 Law of X (or the distribution of X)

when $\nu = P$, $(\Lambda, \mathcal{G}) = (\mathcal{R}, \mathcal{B})$ and $f = X$ (random variable). Then $P \circ X^{-1}$ if often denoted by P_X or F_X (CDF). Where $F_X(c) = P(X \le c)$ By change of variable

$$Eg(X) = \int_\Omega g(X(\omega)) dP(\omega) = \int_\mathcal{R} g(x) dP_X(x) = \int_\mathcal{R} g(x) dF_X(x)$$

2.5.2 CDF properties

$F(-\infty) = \lim_{x \to -\infty} F(x) = 0; F(\infty) = \lim_{x \to \infty} F(x) = 1$
F is non-decreasing. $F(x) \le F(y)$ if $x \le y$
F is right-continuous. $\lim_{y \to x+0} F(y) = F(x)$

2.6 Fubini's Theorem

Suppose $f \ge 0$ or $\int |f| d(\nu_1 \times \nu_2) < \infty$ then

$$g(\omega_2) = \int_{\Omega_1} f(\omega_1, \omega_2) d\nu_1(\omega_1)$$

$$\int_{\Omega_1 \times \Omega_2} f d(\nu_1 \times \nu_2) = \int_{\Omega_1} \left[\int_{\Omega_2} f(\omega_1, \omega_2) d\nu_1(\omega_1) \right] d\nu_2(\omega_2)$$

2.7 Radon-Nikodym derivatives

2.7.1 Absolutely continuity

$\lambda << \nu$ iff for any $A \in \mathcal{F}$, $\nu(A) = 0 \Rightarrow \lambda(A) = 0$

2.7.2 Radon-Nikodym

$\lambda << \nu$, there exist unique f s.t.

$$\lambda(A) = \int_A f d\nu, A \in \mathcal{F}$$

2.7.3 PDF

Given probability measure P and σ-finite ν, if $P << \nu$ then $\frac{dP}{d\nu}$ is called the pdf of P w.r.t. ν

2.7.4 Lebesgue PDF

When $P << m$ (Lebesgue measure), $\frac{dP}{dm}$ is called Lebesgue PDF of P (or of F). If F has derivative f, then

$$P((-\infty, x]) = F(x) = \int_{-\infty}^x f(y) dm(y) = \int_{-\infty}^x f(y) dy$$

2.7.5 Calculus with Radon-Nikodym derivatives

If $\lambda << \nu$ and $f \geq 0$

$$\int f d\lambda = \int f \frac{d\lambda}{d\nu} d\nu$$

If $\lambda_i << \nu$, then $\lambda_1 + \lambda_2 << \nu$

$$\frac{d(\lambda_1 + \lambda_2)}{d\nu} = \frac{d\lambda_1}{d\nu} + \frac{d\lambda_2}{d\nu}$$

If λ is σ-finite, and $\tau << \lambda << \nu$, then ν-a.e.

$$\frac{d\tau}{d\nu} = \frac{d\tau}{d\lambda}\frac{d\lambda}{d\nu}$$

If $\lambda << \nu$ and $\nu << \lambda$, then with ν or λ-a.e.

$$\frac{d\lambda}{d\nu} = \left(\frac{d\nu}{d\lambda}\right)^{-1}$$

If ν_i is σ-finite, $\lambda_i << \nu_i$ then $\lambda_1 \times \lambda_2 << \nu_1 \times \nu_2$, then for $(\nu_1 \times \nu_2)$-a.e.

$$\frac{d(\lambda_1 \times \lambda_2)}{d(\nu_1 \times \nu_2)}(\omega_1, \omega_2) = \frac{d\lambda_1}{d\nu_1}(\omega_1) \cdot \frac{d\lambda_2}{d\nu_2}(\omega_2)$$

2.8 Moments

[pth absolute moment] $E[|X|^p]$ [kth moment] $E[X^k]$ [kth central moment] $E[(X-\mu)^k]$

2.8.1 Variance, Covariance

$Var(X) = E[(X - EX)(X - EX)^T]$
$Cov(X, Y) = E[(X - EX)(Y - EY)^T]$
$Corr(X, Y) = Cov(X, Y)/(\sigma_X \sigma_Y)$
$E(a^T X) = a^T EX$
$Var(a^T X) = a^T Var(X) a$

2.9 Probability Inequalities

2.9.1 Cauchy-Schewarz inequality

$$Cov(X, Y)^2 \leq Var(X) Var(Y)$$
$$(EXY)^2 \leq EX^2 EY^2$$

2.9.2 Jensen's inequality

A is a convex set in $\mathcal{R}^d$, φ is a convex function on A and $X \in A$ is a d-random vector

$$\varphi(EX) \leq E\varphi(X)$$

If φ is strictly convex and $\varphi(X)$ is not a constant, then $\varphi(EX) < E\varphi(X)$

$$(EX)^{-1} < E(X^{-1})$$
$$E(logX) < log(EX)$$
$$\int f \log\left(\frac{f}{g}\right) d\nu \geq 0$$

2.9.3 Chebyshev's inequality

X is R.V, φ is nonnegative and symmetric function ($\varphi(-x) = \varphi(x)$) and is non-decreasing on $[0, \infty)$, then for each constant $t \geq 0$

$$\varphi(t) P(|X| \geq t) \leq \int_{\{|X| \geq t\}} \varphi(X) dP \leq E\varphi(X)$$

Common results

$$P(|X - \mu| \geq t) \leq \frac{\sigma_X^2}{t^2}, P(|X| \geq t) \leq \frac{E|X|}{t}$$

2.9.4 Hölder's inequality

suppose $p, q > 0$ are Hölder's conjugate s.t. $1/p + 1/q = 1 \Rightarrow q = p/(p-1)$

$$E|XY| \leq (E|X|^p)^{1/p}(E|Y|^q)^{1/q}$$

If both $E|X|^p$ and $E|Y|^q$ are finite, equality holds if and only if $|X|^p$ and $|Y|^q$ are linearly dependent

2.9.5 Young's inequality

equality if and only if $a^p = b^q$

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}$$

2.9.6 Minkowski's inequality

$p \geq 1$

$$(E|X+Y|^p)^{1/p} \leq (E|X|^p)^{1/p} + (E|Y|^p)^{1/p}$$

2.9.7 Lyapunov's inequality

for $0 < s < t$

$$(E|X|^s)^{1/2} \leq (E|X|^t)^{1/t}$$

2.9.8 Kullback-Leibler Information

$K(f_0, f_1) = E_0 \log \frac{f_0(X)}{f_1(X)} = \int \log\left(\frac{f_0(x)}{f_1(x)}\right) f_0(x) d\nu(x) \geq 0$ with equality if and only if $f_1(\omega) = f_0(\omega)$ ν-a.e.

2.9.9 Shannon-Kolmogorov information equality

$K(f_0, f_1) \geq 0$ with equality if and only if $f_1(\omega) = f_0(\omega)$ ν-a.e.

2.10 Characteristic function, moment generating function

$\forall\ t \in \mathcal{R}^d$
[Char func] $|\phi_X| \leq 1$, $\phi_{-X} = \overline{\phi_X(t)}$

$$\phi_X(t) = E\left[exp(\sqrt{-1}t^T X\right] = E\left[\cos(t^T X) + \sqrt{-1}\sin(t^T X)\right]$$

[MGF] $\psi_{-X}(t) = \psi_X(-t)$

$$\psi_X(t) = E\left[exp(t^T X)\right]$$

if ψ is finite in neighborhood of $\mathbf{0} \in \mathcal{R}^d$, then moments of X of any order are finite, and $\phi_X(t) = \psi_X(\sqrt{-1}t)$

2.11 Condition on information

2.11.1 Conditional expectation

[Conditional Expectation] $E(X|\mathcal{A})$ is random variable satisfying
(1) $E(X|\mathcal{A})$ is measurable from $(\Omega, \mathcal{A})$ to $(\mathcal{R}, \mathcal{B})$
(2) $\int_C E(X|\mathcal{A})dP = \int_C XdP$ for any $C \in \mathcal{A}$. Such $E(X|\mathcal{A})$ exists and is unique
[Conditional probability] $P(B|\mathcal{A}) = E(I_B|\mathcal{A})$
[Conditional Expectation given Y] $E(X|Y) := E[X|\sigma(Y)]$

2.11.2 Conditional Expectation given y

Let Y be measurable from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$ and Z a function from $(\Omega, \mathcal{F})$ to $\mathcal{R}^k$. If Z is Borel on $(\Omega, \sigma(Y))$, then there is a Borel function h on $(\Lambda, \mathcal{G})$ such that $Z = h \circ Y$. We can denote h as $E(X|Y = y)$

2.11.3 Simple function Y, disjoint A_i

A_i disjoint and $\cup A_i = \Omega$, $P(A_i) > 0$, $Y = \sum_{i \geq 1} c_i I_{A_i}$

$$E(X|Y) = \sum_{i=1}^{\infty} \frac{\int_{A_i} XdP}{P(A_i)} I_{A_i}$$

2.11.4 Tower property

If $\mathcal{H} \subset \mathcal{G}$ is a $\sigma-$field, so $\mathcal{H} \subset \mathcal{G}\mathcal{F}$, then

$$E(X|\mathcal{H}) = E\{E(X|\mathcal{G})|\mathcal{H}\} \text{ a.s.}$$

Let $\mathcal{H} = \{\emptyset, \Omega\}$, then $E(X) = E(E(X|\mathcal{G}))$

2.11.5 Independence

[Independent events] $P(\cap_{i\geq 1} A_i) = \prod_{i\geq 1} A_i$
[Independent collections] $\mathcal{C}_i \subset \mathcal{F}$ are independent iff events $\{A_i \in \mathcal{C}_i : i \in \mathcal{I}\}$ are independent
[Independent random variables] R.V are independent iff $\sigma(X_i)$ $\forall$ i are independent

2.11.6 Checking independence

$P(X_1 \leq a_1, \cdots, X_n \leq a_n) = P(X_1 \leq a_1) \cdots P(X_n \leq a_n)$
If $(X_1, \cdots, X_n)$ has joint pdf f w.r.t product measure $\nu_1 \times \cdots \times \nu_n$ iff $f(x_1, \cdots, x_n) = f_1(x_1) \cdots f_n(x_n)$

2.11.7 Properties

$E(XY) = E(X)E(Y)$,
$E(X|Y) = E(X)$ $P-$a.s.,
$E(g(X,Y)|Y = y) = E(g(X,y))$ P_Y-a.s.,
Any Borel functions are also independent

2.11.8 Tuple independence

If (X, Y_1) and Y_2 are independent
$E(X|(Y_1, Y_2)) = E(X|Y_1)$ a.s.
$P(A|Y_1, Y_2) = P(A|Y_1)$ a.s. for any $A \in \sigma(X)$

2.12 Conditional distribution

2.12.1 Conditional distribution

Suppose X is a random $n-$vector on probability space $(\Omega, \mathcal{F}, P)$, and Y is measurable from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$. Then there exists a function $P_{X|Y}(B|y)$ on $\mathcal{B}^n \times \Lambda$ s.t.
(1) $P_{X|Y}(\cdot|y)$ is a probability measure on $(\mathcal{R}^n, \mathcal{B}^n)$ for any fixed $y \in \Lambda$
(2) $P_{X|Y}(B|y) = P[X \in B|Y = y]$ a.s. P_Y for any fixed $B \in \mathcal{B}^n$

2.12.2 Conditional PDF

If (X, Y) have PDF $f(x, y)$ w.r.t $\nu \times \lambda$ on $\mathcal{B}^n, \mathcal{B}^m$ and both σ-finite. Let $f_Y(y) = \int f(x, y) d\nu(x)$ be marginal PDF of Y w.r.t λ and $A = \{y \in \mathcal{R}^m : f_Y(y) > 0\}$. Then
(a) For any fixed $y \in A$ the PDF of $P_{X|Y=y}$ w.r.t ν is given by

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}$$

(b) Furthermore, if $g(x, y)$ is a Borel function on $\mathcal{R}^{n+m}$ and $Eg(X, Y) < \infty$, then

$$E[g(X,Y)|Y] = \int g(x, Y) f_{X|Y}(x|Y) d\nu(x) \text{ a.s.}$$

2.12.3 Joint distribution

Let $(\Lambda, \mathcal{G}, P_0)$ be probability space. Suppose Q is a function from $\mathcal{B}^n \times \Lambda$ to $\mathcal{R}$ and satisfies
(1) $Q(\cdot, y)$ is a probability measure on $(\mathcal{R}^n, \mathcal{B}^n)$ for any $y \in \Lambda$
(2) $Q(B, \cdot)$ is $\mathcal{G}-$measurable for any $B \in \mathcal{B}^n$.
Then there is a unique probability measure P on $(\mathcal{R}^n \times \Lambda, \sigma(\mathcal{B}^n) \times \mathcal{G})$ s.t. for $B \in \mathcal{B}^n$ and $C \subset \mathcal{G}$

$$P(B \times C) = \int_C Q(B, y) dP_0(y)$$

2.13 Convergence

2.13.1 Almost sure convergence

$X_1, X_2, \cdots$ converges almost surely to rvs X: $X_n \to^{\text{a.s.}} X$ if

$$P\left(\lim_{n\to\infty} X_n = X\right) = 1$$

Can be shown by showing $\forall \epsilon > 0, \sum_{i=1}^{\infty} P(|X_n - X| > \epsilon) < \infty$ via Borel-cantelli lemmas.

2.13.2 Infinity often

Let $\{A_n\}_{n=1}^{\infty}$ be an infinite sequencs of events. We say $\{A_n\}$ happens infinitelty often if ω belongs to following set

$$\{A_n \ i.o.\} = \cap_{n\geq 1} \cup_{j\geq n} A_j := \limsup_{n\to\infty} A_n$$

2.13.3 Borel-Cantelli lemmas

[First Borel-Cantelli] For a sequence of events $\{A_n\}_{n=1}^{\infty}$, if $\sum_{n=1}^{\infty} P(A_n) < \infty$, then $P(A_n \ i.o.) = 0$
[Second Borel-Cantelli] For a sequences of pairwisely independent events $\{A_n\}_{n=1}^{\infty}$, if $\sum_{n=1}^{\infty} P(A_n) = \infty$, then $P(A_n \ i.o.) = 1$

2.13.4 a.s. and First BC

Let X and $X_1, X_2, \cdots$ defined on a common probability space. For a constant $\epsilon > 0$, define the sequences of events $\{A_n(\epsilon)\}_{n=1}^{\infty}$ to be

$$A_n(\epsilon) = \{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \epsilon\}$$

If $\sum_{n=1}^{\infty} P(\{A_n(\epsilon\}) < \infty$ for all $\epsilon > 0$, then $X_n \to^{\text{a.s.}} X$

2.13.5 Convergence in L^p

A sequence of $\{X_n\}_{n=1}^{\infty}$ of rvs converges to a random variable X in the L^p sense for some $p > 0$ if $E|X|^p < \infty$ and $E|X_n|^p < \infty$ and

$$\lim_{n\to\infty} E|X_n - X|^p = 0$$

2.13.6 Convergence in probability

A sequence $\{X_n\}_{n=1}^{\infty}$ of rvs converges to a rando variable X in probability if for all $\epsilon > 0$

$$\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0$$

denoted by $X_n \to^P X$. Can also be showned by $E(X_n) = X$, $\lim_{n\to\infty} Var(X_n) = 0$

2.13.7 Convergence in distribution (weak convergence)

A sequence $\{X_n\}_{n=1}^{\infty}$ of rvs converges to a random variable X in distribution/in law/weakly if

$$\lim_{n\to} F_n(x) = F(x)$$

for every $x \in \mathcal{R}$ at which F is continuous, where F_n, F are CDF of X_n, X respectively. Denoted by $X_n \to^D X$ or $F_n \Rightarrow F$

2.13.8 Relations between Convergence Modes

$L^p \Rightarrow L^q \Rightarrow P$, $a.s. \Rightarrow P$, $P \Rightarrow D$.
If $X_n \to_D C$, then $X_n \to_P C$. If $X_n \to_P X$, there exists sub-sequence s.t. $X_{n_j} \to_{\text{a.s.}} X$.

2.13.9 Continuous mapping

Let $\{X_n\}_{n=1}^{\infty}$ be seq of random $k-$vectors and X is random $k-$vector in the same probability space. Let $g : \mathcal{R}^k \to \mathcal{R}$ be continuous. Then If $X_n \to^* X$, then $g(X_n) \to^* g(X)$, where * is either a.s., P or D.

2.13.10 Convengence properties

1. Unique in limit: $X = Y$ if $X_n \to X$ and Y when a.s., P, L^p. If $F_n \Rightarrow F$ and G, then $F(t) = G(t)\ \forall\ t$
2. Concatenation: $(X_n, Y_n) \to (X, Y)$ when P or a.s., $(X_n, Y_n) \to_D (X, c)$ only for constant.
3. Linearity: $(aX_n + bY_n) \to aX + bY$ when a.s., P, L^p NOT for distribution.
4. Cramér-Wold device: for k-random vectors, $X_n \to_D X \Leftrightarrow c^T X_n \to_D c^T X$ for every $c \in \mathcal{R}^k$

2.13.11 Lévy continuity

$\{X_n\}$ converges in distribution to X iff corresponding characteristic functions $\{\phi_n\}$ converges pointwise to ϕ_X

2.13.12 Scheffés theorem

Let $\{f_n\}$ be seq of pdfs on $\mathcal{R}^k$ wrt measure ν. Suppose $\lim_{n\to\infty} f_n(x) = f(x)$ a.e. ν and $f(x)$ is pdf wrt ν. Then $\lim_{n\to\infty} \int |f)n(x) - f(x)|d\nu = 0$ and $P_{f_n} \Rightarrow P_f$. Useful for checking convergence in distribution via pdfs.

2.13.13 Slutsky's theorem

If $X_n \to^D X$ and $Y_n \to^D c$ for a constant c. Then $X_n + Y_n \to^D X + c$, $X_n Y_n \to^D cX$, $X_n/Y_n \to^D X/c$ if $c \neq 0$

2.13.14 Skorohod's theorem

If $X_n \to^D X$, then there are some random vectors $Y, Y_1, Y_2, \cdots$ defined on a common probability space such that $P_{Y_n} = P_{X_n}, n = 1, 2, \cdots$, $P_Y = P_X$ and $Y_n \to^{\text{a.s.}} Y$

2.13.15 δ-method

Let $X_1, X_2, \cdots, Y$ be rvs, $\{a_n\} > 0$ with $\lim_{n\to} \infty a_n = \infty$ and $a_n(X_n - c) \to^D Y$ where $c \in \mathcal{R}$. Let g be a function from $\mathcal{R}$ to $\mathcal{R}$.
If $g(c)$ differentiable at c, then

$$a_n[g(X_n) - g(c)] \to^D g'(c)Y$$

Suppose that g has continuous derivatives of order $m > 1$ in a neighbourhood of c s.t. $g^{(j)}(c) = 0$ for all $1 \leq j \leq m-1$ and $g^{(m)}(c) \neq 0$. Then

$$a_n^m[g(X_n) - g(c)] \to^D \frac{1}{m!} g^{(m)}(c)Y^m$$

If X_i, Y are k-vectors rvs and $c \in \mathcal{R}^k$

$$a_n[g(X_n) - g(c)] \to_D [\nabla g(c)]^T Y = N\left(0, g(c)^T \Sigma g(c)\right) \text{ if } Y \text{ is normal}$$

2.14 Stochastic order

2.14.1 real numbers

$\{a_n\}, \{b_n\}$ For a constant c and all n

$$a_n = O(b_n) \Leftrightarrow |a_n| \leq c|b_n|$$

$$a_n = o(b_n) \Leftrightarrow \lim_{n\to\infty} a_n/b_n = 0$$

2.14.2 rvs

$\{X_n\}, \{Y_n\}$

$$X_n = O_{\text{a.s.}}(Y_n) \Leftrightarrow P\{|X_n| = O(|Y_n|)\} = 1$$

$$X_n = o_{\text{a.s.}}(Y_n) \Leftrightarrow X_n/Y_n \to^{\text{a.s.}} 0$$

$\forall\ \epsilon > 0, \exists C_\epsilon > 0, n_\epsilon \in \mathcal{N} s.t.$

$$X_n = O_P(Y_n) \Leftrightarrow \sup_{n \geq n_\epsilon} P\left(\{\omega \in \Omega : |X_n(\omega) \geq C_\epsilon |Y_n(\omega)|\}\right) < \epsilon$$

If $X_n = O_P(1)$, $\{X_n\}$ is bounded in probability

$$X_n = o_P(Y_n) \Leftrightarrow X_n/Y_n \to^P 0$$

2.14.3 Properties

If $X_n \to_{\text{a.s.}} X$, then $\{\sup_{n\geq k}|X_n|\}_k$ is $O_p(1)$.
If $X_n \to_D X$ for a rvs, then $X_n = O_P(1)$ (tightness).
If $E|X_n| = O(a_n)$, then $X_n = O_P(a_n)$; If $E|X_n| = o(a_n)$, then $X_n = o_P(a_n)$

2.15 Law of Large Numbers (LLN)

Let $X_1, X_2, \cdots$ be independent rvs.

2.15.1 Strong LLN

If X_i are identical, let $c := EX_i$

$$E|X_i| < \infty \Leftrightarrow \frac{1}{n}\sum_{i=1}^{n} X_i \to^{\text{a.s.}} c$$

2.15.2 SLLN, non-identical

If there is a constant $p \in [1,2]$ s.t. $\sum_{i=1}^{\infty} E|X_i|^p/i^p < \infty$, then

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - EX_i) \to^{\text{a.s.}} 0$$

2.15.3 Uniform SLLN for iid samples

Suppose (1) $U(x,\theta)$ is continuous in θ for any fixed x (2) For each θ, $\mu(\theta) = EU(X,\theta)$ is finite (3) Θ is compact (4) There exists function $M(x)$ s.t. $EM(X) < \infty$ and $|U(x,\theta) \leq M(x)|$ for all x, θ. Then

$$P\left\{\lim_{n\to\infty}\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}U(X_j,\theta) - \mu(\theta)\right| = 0\right\} = 1$$

2.15.4 Weak LLN

If X_i are identical, $\{a_n\}$ exist and take $a_n = E(X_1 I_{\{|X_1|\leq n\}}) \in [-n, n]$

$$nP(|X_1| > n) \to 0 \Leftrightarrow \frac{1}{n}\sum_{i=1}^{n} X_i - a_n \to^{\mathcal{P}} 0$$

2.15.5 WLLN, non-identical

If there is a constant $p \in [1,2]$ s.t. $\lim_{n\to\infty}\frac{1}{n^p}\sum_{i=1}^{n} E|X_i|^p = 0$, then

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - EX_i) \to^{P} 0$$

2.15.6 Weak Convergency

Seq of probability measures ν_n converges weakly to ν if $\int f d\nu_n \to \int f d\nu$ for every bounded and continous real function f.
Suppose X_n's and X are k-vectors, $X_n \to_D X$ is equivalent to any of the conditions below
(1) $E[h(X_n)] \to E[h(X)]$ for every bounded continuous function h (convergence of probability measures)
(2) $\limsup_n P_{X_n}(C) \leq P_X(C)$ for any closed set $C \subset \mathcal{R}^k$
(3) $\liminf_n P_{X_n}(O) \geq P_X(O)$ for any open set $O \subset \mathcal{R}^k$

2.15.7 Central Limit Theorem, Classical iid

Let $\{X_n\}_{n=1}^{\infty}$ be seq of iid random $k-$vectors. Suppose $\Sigma = VarX_1 < \infty$, then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - EX_i) \to^{D} N(0,\Sigma)$$

2.15.8 Lindeberg's CLT for non-identical

For each n, let $\{X_{nj}, j = 1, \cdots, k_n\}$ be set of independent rvs. Suppose
(1) $k_n \to \infty$ as $n \to \infty$
(2) $0 < \sigma_n^2 = Var\left(\sum_{j=1}^{k_n} X_{nj}\right) < \infty, n = 1, 2, \cdots$. [Lindeberg's condition]
(3) If for any $\epsilon > 0$, $\frac{1}{\sigma_n^2}\sum_{j=1}^{k_n} E\left\{(X_{nj} - EX_{nj})^2 I_{\{|X_{nj}-EX_{nj}|>\epsilon\sigma_n\}}\right\} \to 0$. Then

$$\frac{1}{\sigma_n}\sum_{j=1}^{k_n}(X_{nj} - EX_{nj}) \to^D N(0,1)$$

2.15.9 Checking Lindeberg's condition

Following 2 implied Linderberg's condition
[Lyapunov condition]

$$\frac{1}{\sigma_n^{2+\delta}}\sum_{j=1}^{k_n} E|X_{nj} - EX_{nj}|^{2+\delta} \to 0 \text{ for some } \delta > 0$$

[Uniform boundedness] If $|X_{nj}| \le M$ for all n and j and $\sigma_n^2 = \sum_{j=1}^{k_n} Var(X_{nj}) \to \infty$
[Feller's condition] In general, Lindeberg's condition is not necessary for convergence result. However, if Feller's condition is met then it is sufficient and necessary.

$$\lim_{n\to\infty}\max_{j\le k_n}\frac{Var(X_{nj})}{\sigma_n^2} = 0$$

2.15.10 Berry-Esseen Theorem

There exist a universal constant C such that following holds. Suppose $Y_1, Y_2, \cdots, Y_n$ are iid rvs with $E(Y_i) = 0, E(Y_i^2) = \sigma^2 > 0, E(|Y_i|^3) = \rho < \infty$. Let F_n be CDF of $\frac{\sum_{i=1}^n Y_i}{\sigma\sqrt{n}}$, Φ be CDF of $N(0,1)$, then

$$\sup_{y\subset\mathcal{R}}|F_n(y) - \Phi(y)| \le \frac{C\rho}{\sigma^3\sqrt{n}}$$

---

3 Statistical Estimation

3.1 Basics terms

[Estimator] Estimate θ, is a function of data (is a statistics)

[Statistical models] A statistical model is often express as

$$P \in \mathcal{P} = \{Q : Q \text{ satisfies some condition}\}$$

[Population and Sample] A population is a probability space $(\Omega, \mathcal{F}, P)$, we also refer to P as the population.
A sample is a random element defined on the probability space. The data set is a realization of the sample.
A population P is known iff $P(A)$ is known value for every event $A \in \mathcal{F}$.

[Parametric family] A set of probability measures P_θ on $(\Omega, \mathcal{F})$ indexed by parameter $\theta \in \Theta$ is parametric family iff $\Theta \subset \mathcal{R}^d$.
Θ is the parameter space, d is its dimensions.

[Parametric models] Parametric model refers to assumption that the population P is in a parametric family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$

[Identifiable] Parametric family is identifiable iff $\forall\ \theta_1, \theta_2 \in \Theta$ and $\theta_1 \ne \theta_2 \Rightarrow P_{\theta_1} \ne P_{\theta_2}$

[Dominated by] If $P << \nu$ (σ-finite measure) $\forall\ P \in \mathcal{P}$, then $\mathcal{P}$ is dominated by ν. Also, then $\mathcal{P}$ can be identified by the family of densities $\{\frac{dP}{d\nu} : P \in \mathcal{P}\}$

3.2 Statistics

A statistics $T(X)$ is a measurable function of sample X
[Sample mean] $\bar{X} = \frac{1}{n}\sum_i X_i$, $E(\bar{X}) = \mu$, $Var(\bar{X}) = \sigma^2/n$. If $P \sim N(\mu, \sigma^2), \bar{X} \sim N(\mu, \sigma^2/n)$, If $P \sim E(0, \theta), n\bar{X} \sim \Gamma(n, \theta)$.
[Sample Variance] $S^2 = \frac{1}{n-1}\sum_i (X_i - \bar{X})^2$
[Ordered Statistics] $X_(k)$ which is the kth smallest value of $X_1, \cdots, X_n$.

$$X_{(n)} = [F(x)]^n, f_{X_{(n)}} = nf(x)[F(x)]^{n-1}$$

$$X_{(1)} = 1 - [1 - F(x)]^n, f_{X_{(1)}} = nf(x)[1 - F(x)]^{n-1}$$

[Min, Max] $X_{(1)}, X_{(n)}$
[Empirical variance] $\frac{1}{n}\sum_i (X_i - \bar{X})^2$
[Empirical distribution] $P_n(A) = \frac{1}{n}\#\{i : X_i \in A\}$, $\forall\ A \in B$

3.3 Exponential families

Parametric family $\{P_\theta : \theta \in \Theta\}$ dominated by σ-finite measure ν on (Ω, ε is called exponential family iff for $\omega \in \Omega$

$$f_\theta(\omega) = \frac{dP_\theta}{d\nu}(\omega) = \exp\left\{[\eta(\theta)]^T T(\omega) - \xi(\theta)\right\} h(\omega)$$

where T is a random p-vector, η is a function from Ω to $\mathcal{R}^p$, h is non-negative Borel function on (ω, ε) and

$$\xi(\theta) = \log\left\{\int_\Omega \exp\left\{[\eta(\theta)]^T T(\omega)\right\} h(\omega) d\nu(\omega)\right\}$$

3.3.1 Canonical form and Natural Exp Families

Since exp fam representation is not unique, consider $\eta = \eta(\theta)$

$$f_\eta(\omega) = \exp\left\{\eta^T T(\omega) - \mathcal{C}(\eta)\right\} h(\omega)$$
$$\mathcal{C}(\eta) = \log\left\{\int_\Omega \exp\left\{\eta^T T(\omega)\right\} h(\omega) d\nu(\omega)\right\}$$

η is called natural parameter and natural parameter space $\Xi = \{\eta(\theta) : \theta \in \Theta\} \subset \mathcal{R}^p$. Full rank if Ξ contains open set in $\mathcal{R}^p$

3.3.2 Joint Exp Fam

Suppose $X_i \sim f_i$ independently with f_i Exp Fam, then joint distribution $X_1, \cdots, X_n$ is also Exp Fam.

3.3.3 Showing non Exp Fam

For an exp fam P_θ, there is nonzero measure λ s.t. $\frac{dP_\theta}{d\lambda}(\omega) > 0$ λ-a.e. and for all θ.
Consider $f = \frac{dP_\theta}{d\lambda} I_{(t,\infty)}(x)$, $\int f d\lambda = 0, f \geq 0 \Rightarrow f = 0$. Since $\frac{dP_\theta}{d\lambda} > 0$ (assume), then $I_{(t,\infty)}(x) = 0 \Rightarrow v([t, \infty)) = 0$. Since t is arbitary, consider $v(\mathcal{R}) = 0$ (contradiction)

3.3.4 Separate statistics T

Let $T = (Y, U)$ and $\eta = (\nu, \varphi)$ where Y and ν have same dimension. Then Y has PDF

$$f_\eta(y) = \exp\left\{\nu^T y - \mathcal{C}(\eta)\right\}$$

w.r.t σ-finite measure depending on φ. If T has a PDF in NEF, the conditional distirbution of Y given $U = u$ has PDF (w.r.t σ-finite measure depending on u)

$$f_{\nu,u}(y) = \exp\left\{\nu^T y - \mathcal{C}_u(\nu)\right\}$$

which is in a NEF indexed by ν

3.3.5 MGF of NEFs

If η_0 is an interior point on natural parameter space, then MGF $\phi_{\eta_0}(t)$ of T (with $P = P_{\eta_0}$ is finite in neighborhood of $t = 0$ and is given by

$$\psi_{\eta_0}(t) = \exp\{\mathcal{C}(\eta_0 + t) - \mathcal{C}(\eta_0)\}$$

Let $A(\theta) = \mathcal{C}(\eta_0(\theta))$, $\frac{dA(\theta)}{d\theta} = \frac{d\mathcal{C}(\eta_0(\theta))}{d\eta_0(\theta)} \cdot \frac{d\eta_0(\theta)}{d\theta}$

$$E_{\eta_0} T = \frac{d\psi_{\eta_0}}{dt}|_{t=0} = \frac{d\mathcal{C}}{d\eta_0} = \frac{A'(\theta)}{\eta_0'(\theta)}$$

$$E_{\eta_0} T^2 = \mathcal{C}''(\eta_0) + \mathcal{C}'(\eta_0)^2$$

$$Var(T) = \mathcal{C}''(\eta_0) = \frac{A''(\theta)}{[\eta_0(\theta)]^2} - \frac{\eta_0(\theta)'' A'(\theta)}{[\eta_0(\theta)']^3}$$

3.3.6 Differential identities of NEFs

For a Borel function g, let Ξ_g be set of values of η such that

$$\int |g(\omega)| \exp\left\{\eta^T T(\omega) - \mathcal{C}(\eta)\right\} h(\omega) d\nu(\omega) < \infty$$

Define G on Ξ_g by

$$G(\eta) := \int g(\omega) \exp\left\{\eta^T T(\omega) - \mathcal{C}(\eta)\right\} h(\omega) d\nu(\omega)$$

Then for η in interior of Ξ_g
(1) G is continuous and has continuous derivatives of all orders.
(2) These derivatives can be computed by differentiation under the integral sign.

$$\frac{dG(\eta)}{d\eta} = E_\eta \left[g(\omega) \left(T(\omega) - \frac{\partial}{\partial \eta} \xi(\eta) \right) \right]$$

3.4 Data Reduction

3.4.1 Sufficiency

Let X be a sample from an unknown population $P \in \mathcal{P}$. Statistics $T(X)$ is sufficient for $P \in \mathcal{P}$ iff $P_X(x|Y)$ is known and does not depend on P. If $\mathcal{P}$ is parametric family, we can also say $T(X)$ is sufficient for θ. Suppose T is sufficient for $\mathcal{P}_0$, $\mathcal{P}_0 \subset \mathcal{P} \subset \mathcal{P}_1$. Then $T(X)$ is sufficient for $\mathcal{P}_0$ but not ncessarily $\mathcal{P}_1$.

$$P(X = x|T = t) \text{ does not depend on } \theta$$

3.4.2 Factorization theorem

$T(X)$ is sufficient for $P \in \mathcal{P}$ iff there are non-negative Borel functions
(1) $h(x)$ does not depend on P
(2) $g_P(t)$ which depends on P
s.t.

$$\frac{dP}{d\nu}(x) = g_P(T(x))h(x)$$

3.4.3 Minimal sufficiency

Let T be a sufficient statistics for $P \in \mathcal{P}$. T is called minimal sufficient statistics iff for any other statistics S sufficient for $P \in \mathcal{P}$, there is a measurable function ψ s.t. $T = \psi(S)$ $\mathcal{P}$-a.s.

3.4.4 Min Suff - Method 1

[Theorem A] Suppose $\mathcal{P}_0 \subset \mathcal{P}$ and $\mathcal{P}_0$-a.s. implies $\mathcal{P}$-a.s. If T is sufficient for $P \in \mathcal{P}$ and minimal sufficient for $P \in \mathcal{P}_0$, then T is minimal sufficient for $P \in \mathcal{P}$
[Theorem B] Suppose $\mathcal{P}$ contains PDFs $f_0, f_1, \cdots$ w.r.t a σ-finite measure.
(1) Define $f_\infty(x) = \sum_{i=0}^\infty c_i f_i(x)$, $T_i(x) = f_i(x)/f_\infty(x)$, then $T(X) = (T_0(X), T_1(X), \cdots)$ is minimal sufficient for $\mathcal{P}$. Where $c_i > 0, \sum_{i=0}^\infty c_i = 1, f_\infty(x) > 0$.
(2) If $\{x : f_i(x) > 0\} \subset \{x : f_0(x) > 0\}$ for all i, then $T(X) = (f_1(x)/f_0(x), f_2(x)/f_0(x), \cdots$ is minimal sufficient for $\mathcal{P}$

3.4.5 Min Suff - Method 2

[Theorem C] Suppose $\mathcal{P}$ contains PDFs f_P w.r.t. σ-finite measure ν. If
(a) $T(X)$ is a sufficient statistics, and
(b) There is a measurable function ϕ s.t. for any possible values x, y of X, or $x, y \in \{x : h(x) > 0\}$ for NEF.

$$f_P(x) = f_P(y)\phi(x, y) \forall P \in \mathcal{P} \Rightarrow T(x) = T(y)$$

Then $T(X)$ is minimal sufficient for $\mathcal{P}$

3.4.6 Special min suff result for NEF

If there exists $\Theta_0 = \{\theta_0, \theta_1, \cdots, \theta_p\} \subset \Theta$ s.t. vectors $\eta_i = \eta(\theta_i) - \eta(\theta_0), i \in [1, p]$ are linearly independent in $\mathcal{R}^p$, then T is also minimal sufficient. Check $det([\eta_1, \cdots, \eta_p])$ is non-zero OR $\Xi = \{\eta(\theta) : \theta \in \Theta\}$ contains $(p+1)$ points that do not lie on the same hyperplane OR Ξ is full rank.

3.4.7 Completeness

[Ancillary statistics] A statistics $V(X)$ is ancillary for $\mathcal{P}$ if its distribution does not depend on population $P \in \mathcal{P}$
[First-order ancillary] if $E_P[V(X)]$ does not depend on $P \in \mathcal{P}$
[Completeness] Statistics $T(X)$ is complete for $P \in \mathcal{P}$ iff for any Borel function f, $E_P f(T) = 0$ for all $P \in \mathcal{P}$ implies $f(T) = 0$ $\mathcal{P}$-a.s. T is boundedly complete iff statements holds for bounded Borel functions f.

3.4.8 Completeness + Sufficiency ⇒ Minimal Sufficiency

Suppose X is a sample from unknown $P \in \mathcal{P}$, and suppose a minmal sufficient statistics exists. If a statistics U is sufficient and boundedly complete, then U is minimal sufficient

3.4.9 Complete sufficient statistics for NEF

If $\mathcal{P}$ is NEF of full rank then $T(X)$ is complete and sufficient for $\eta \in \Xi$

3.5 Basu's theorem

Let V and T be two statistics of X from a population $P \in \mathcal{P}$. If V is ancillary and T is boundedly complete and sufficient for $P \in \mathcal{P}$, then V and T are independent w.r.t any $P \in \mathcal{P}$

---

4 Evaluation

4.1 Decision rules

Statistical decision is action taken after observing data X, e.g. estimate θ by $\hat{\theta}$, choose between different hypothesis, make a statement about parameter range.
[Action space] $\mathcal{A}$ set of allowable actions, endowed with a σ-field $\mathcal{F}_\mathcal{A}$
[Decision rule] measurable function from $(\mathcal{X}, \mathcal{F}_\mathcal{X})$ to $(\mathcal{A}, \mathcal{F}_\mathcal{A})$. $\mathcal{X}$ is range of X and $\mathcal{F}_\mathcal{X}$ is a σ-field on $\mathcal{X}$. Choose a decision rule T, and take action $T(X) \in \mathcal{A}$ after X is observed.

4.1.1 Loss function

Function $L : \mathcal{P} \times \mathcal{A} \to [0, \infty)$ that is Borel for each fixed $P \in \mathcal{P}$

4.1.2 Risk

$R_T(P) = E_P L(P, T(X)) = \int L(P, T(X)) dP$. Average loss under population P, depending on P, T, L

4.1.3 Hypothesis tests

Let $\mathcal{P}$ be a family of distributions, $\mathcal{P}_0 \subset \mathcal{P}, \mathcal{P}_1 = \mathcal{P} \backslash \mathcal{P}_0$. Hypothesis testing decides between $H_0 : P \in \mathcal{P}_0, H_1 : P \in \mathcal{P}_1$. Action space $\mathcal{A} = \{0, 1\}$, decision rule is called a test $T : \mathcal{X} \to \{0, 1\} \Rightarrow T(X) = I_C(X)$ for some $C \subset \mathcal{X}$. C is called the region/critical region.

4.1.4 $0-1$ loss

Common loss function for hypo test, $L(P, j) = 0$ for $P \in \mathcal{P}_j$ and $= 1$ for $P \in \mathcal{P}_{1-j}, j \in \{0, 1\}$
Risk $R_T(P) = P(T(X) = 1) = P(X \in C)$ if $P \in \mathcal{P}_0$ or $P(T(X) = 0) = P(X \notin C)$ if $P \in \mathcal{P}_1$

4.1.5 Type I and II errors

Type I: H_0 is rejected when H_0 is true. Error rate: $\alpha_T(P) = P(T(X) = 1), P \in \mathcal{P}_0$
Type II: H_0 is accepted when H_0 is false. Error rate: $1 - \alpha_T(P) = P(T(X) = 1), P \in \mathcal{P}_1$

4.1.6 Power function of T

$\alpha_T(P)$, Type I and Type II error rates cannot be minimized simultaneously.

4.1.7 Significance level

Under Neyman-Pearson framework, assign pre-specified bound α (significance level of test):

$$\sup_{P \subset \mathcal{P}_0} P(T(X) = 1) \leq \alpha$$

4.1.8 size of test

α' is the size of the test

$$\sup_{P \subset \mathcal{P}_0} P(T(X) = 1) = \alpha'$$

4.2 Comparing decision rules

4.2.1 Compare decision rules

T_1 is ... as T_2 if ...: as good as if $R_{T_1}(P) \leq P_{T_2}(P)$. $\forall\ P \in \mathcal{P}$
better if $R_{T_1}(P) < R_{T_2}(P)$ for some $P \in \mathcal{P}$ (and T_2 is dominated by T_1).
equivalent if $R_{T_2}(P) = R_{T_2}(P)$ for all $P \in \mathcal{P}$

4.2.2 Optimal

Let $\mathcal{J}$ be collection of decision rules in consideration. T_* is $\mathcal{J}$-optional if T_* is as good as any other rule in $\mathcal{J}$, Optimal if T_* is as good as any other possible rule

4.2.3 Admissibility

Let $\mathcal{J}$ be a class of decision rules. A decision rule $T \in \mathcal{J}$ is called $\mathcal{J}$-admissible if no $S \in \mathcal{J}$ is better than T in terms of the risk.

4.2.4 Minimaxity

Let $\mathcal{J}$ be a class of decision rules. A decision rule $T_* \in \mathcal{J}$ is called $\mathcal{J}$-minimax if $\sup_{P \subset \mathcal{P}} R_{T_*}(P) \leq \sup_{P \subset \mathcal{P}} R_T(P)$ for any $T \in \mathcal{J}$

4.2.5 Bayes Risk and Rule

A form of averaging $R_T(P)$ over $P \in \mathcal{P}$. Bayes risk $r_T(\Pi) = \int_{\mathcal{P}} R_T(P) d\Pi(P)$, Π is known probability measure. $R_T(\Pi)$ is Bayes risk of T wrt Π. If $T_* \in \mathcal{J}$, $r_{T_*}(\Pi) \leq r_T(\Pi)$ for any $T \in \mathcal{J}$, then T_* is called $\mathcal{J}$-Bayes rule wrt Π.

4.2.6 Finding Bayes rule

Let $\tilde{\theta} \sim \pi$, $X|\tilde{\theta} \sim P_{\tilde{\theta}}$, then $r_\pi(T) = E\left[L(\tilde{\theta}, T(X)\right] = E\left[E\left[L(\tilde{\theta}, T(X)\right]|X\right]$ where E is taken jointly over $(\tilde{\theta}, X)$. Then find $T_*(x)$ that minimises the conditional risk.

4.2.7 Point estimators evaluation

.....
[Bias] $E_\theta(\hat{\theta}) - \theta$
If there exists an unbiased estimator of θ, then θ is called an estimable parameter.
.....
[Variance] $Var(\hat{\theta})$
.....
[MSE] $E(||\hat{\theta} - \theta||^2) = E||\hat{\theta} - E(\hat{\theta})||^2 + (E\hat{\theta} - \theta)^2 = Var(\hat{\theta}) + \text{Bias}^2$

4.3 Rao-Blackwell

Require convex loss $L(P,a)$ and sufficient statistics T for $P \in \mathcal{P}$. Suppose S_0 is decision rule satisfying $E_P|||S_0|| < \infty$ for all $P \in \mathcal{P}$. Let $S_1 = E[S_0(X)|T]$, then $R_{S_1}(P) \leq R_{S_0}(P)$. If $L(P,a)$ is strictly convex in a, and S_0 is not a funciton of T, then S_0 is inadmissible and dominated by S_1.

---

5 Estimators

5.1 Method of Moments

[jth moments of P_θ] $\mu_j = E_\theta X^j$
[Unbiased estimate] $\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^n X^j$
[MoM] Find the first n moments, n = number of parameters. Note $EX^2 = Var(X) + (EX)^2$

5.1.1 Properties

5.2 Maximum Likelihood

[Likelihood function] $\ell(\theta) = f_\theta(x)$
[Maximum likelihood estimate of θ] $\hat{\theta} \in \Theta$ satisfying $\ell(\hat{\theta}) = \max_{\theta\in\Theta} \ell(\theta)$
[Maximum likelihood estimator] If $\hat{\theta}$ is a Borel function X a.e. ν
[MLE est] $\max_\theta \ell(\theta)$
[Score function] $\frac{\partial}{\partial\theta}\log f_\theta(X)$
[Limitations] MLE might not exist, solvable, no PDFs, not MSE optimal. Note need to check all critical points

5.2.1 Numerical methods

Assuming Hessian matrix is full rank. Solving $\frac{\partial \log \ell(\theta)}{\partial\theta} = \mathbf{0}$

$$\hat{\theta}^{(t+1)} = \hat{\theta}^{(t)} - \left[\frac{\partial^2 \log \ell(\theta)}{\partial\theta\partial\theta^T} \mid_{\theta=\hat{\theta}^{(t)}}\right]^{-1} \frac{\partial \log \ell(\theta)}{\partial\theta} \mid_{\theta=\hat{\theta}^{(t)}}$$

5.2.2 MLE for Exp Fam

NEF: $\ell(\eta) = \exp\left[\eta^T T(x) - \mathcal{C}(\eta)\right] h(x)$

$$T(x) = \frac{\partial \mathcal{C}(\eta)}{\partial\eta}, Var(T) = \frac{\partial^2 \mathcal{C}(\eta)}{\partial\eta\partial\eta^T}$$

General: $\ell(\theta) = \exp\left[\eta(\theta)^T T(x) - \xi(\theta)\right] h(x)$, note $\xi(\theta) = \mathcal{C}(\eta(\theta))$

$$\hat{\theta} = \eta^{-1}(\hat{\eta}), \text{ or solution of } \frac{\partial\eta(\theta)}{\partial\theta} T(x) = \frac{\partial\xi(\theta)}{\partial\theta}$$

5.2.3 Consistency

Suppose (1) Θ is compact (2) $f(x|\theta)$ is continuous in θ for all x (3) There exists a function $M(x)$ s.t. $E_{\theta_0}[M(X)] < \infty$ and $|\log f(x|\theta) - \log f(x|\theta_0)| \leq M(x)$ for all x, θ (4) identifiability holds $f(x|\theta) = f(x|\theta_0)$ ν-a.e. $\Rightarrow \theta = \theta_0$. Then for any sequence of maximum likelihood-likelihood estimates $\hat{\theta}_n$ of θ

$$\hat{\theta}_n \to^{\text{a.s}} \theta_0$$

5.3 Unbiased Estimators

5.3.1 Uniformly minimum variance unbiased estimator UMVUE

$T(X)$ of θ is UMVUE $\Leftrightarrow Var(T(X) \leq Var(U(X))$ for any $P \in \mathcal{P}$ and any other unbiased estimator $U(X)$ of θ

5.3.2 Lehmann-Scheffé

Suppose there exists sufficient and complete statistic $T(X)$ for $P \in \mathcal{P}$, and θ is related to P. If θ is estimable, then there is a unique unbiased estimator of θ that is of the form $h(T)$ with a Borel function h. Furthermore, $h(T)$ is the unique UMVUE of θ.

5.3.3 Finding UMVUE method 1

Using Lehmann-Scheffé, manipulate $E(h(T)) = \theta$ to get $\hat{\theta}$ where T is sufficient and complete. Useful when $E(h(T))$ is easy to solve.

5.3.4 Finding UMVUE method 2

Using Rao-Blackwellization. Find (1) unbiased estimator of $\theta = U(X)$, (2) sufficient and complete statistics $T(X)$, then $E(U|T)$ is the UMVUE of θ by Lehmann-Scheffé. Useful if $E(U|T)$ is easy to solve.

5.3.5 UMVUE method 3 - necessary and sufficient condition

Useful when no complete and sufficient statistics. Can use to find UMVUE, check if estimator is UMVUE, show nonexistence of UMVUE.
Let T is an unbiased estimator of *eta* with finite variance, $\mathcal{U}$ is set of all unbiased estimators of 0 with finite variances. $T(X)$ is UMVUE $\Leftrightarrow E[T(X)U(X)] = 0$ for any $U \in \mathcal{U}$ and any $P \in \mathcal{P}$.
Suppose $T = h(S)$, where S is sufficient statistics for $P \in \mathcal{P}$ and h is a Borel function. Let $\mathcal{U}_S$ be the subset of $\mathcal{U}$ consisting of Borel functions of S. $T(X)$ is UMVUE $\Leftrightarrow E[T(X)U(X)] = 0$ for any $U \in \mathcal{U}_S$ and any $P \in \mathcal{P}$

5.3.6 Using method3

(1) Find $U(x)$ via $E[U(x)] = 0$ (2) Construct $T = h(S)$ s.t. T is unbiased (3) Find T via $E[TU] = 0$

5.3.7 Corollary

If T_j is UMVUE of η_j with finite variances, then $T = \sum_{j=1}^k c_j T_j$ is UMVUE of $\eta = \sum_{j=1}^k c_j \eta_j$.
If T_1, T_2 are UMVUE of η with finite variances, then $T_1 = T_2$ a.s. $P, P \in \mathcal{P}$

5.4 Fisher information

Suppose fixed support, for any $\theta \in \Theta$, $\frac{\partial f_\theta(x)}{\partial \theta}$ exists and is finite $P_\theta-$a.s., X is a sample from $P_\theta \in \mathcal{P}$. Amount of information from X is

$$I(\theta) = E\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right)^2 = \int\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right)^2 f_\theta(X)d\nu(x) = E\left\{\frac{\partial}{\partial\theta}\log f_\theta(X)\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right]^T\right\}$$

5.4.1 Parameterization

If $\theta = \psi(\eta)$ and ψ' exists

$$\tilde{I}(\eta) = \psi'(\eta)^2 I(\psi(\eta))$$

5.4.2 Twice differentiable

Suppose f_θ is twice differentiable in θ and $\int \frac{\partial^2}{\partial\theta^2} f_\theta(x) I_{f_\theta(x)>0} d\nu = 0$, then

$$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\log f_\theta(X)\right] = -E\left[\frac{\partial^2}{\partial\theta\partial\theta^T}\log f_\theta(X)\right]$$

5.4.3 Independent samples

If regularity condition $\int \frac{\partial}{\partial\theta} f_\theta(x)d\nu = 0$ holds, then

$$I_{(X,Y)}(\theta) = I_X(\theta) + I_Y(\theta)$$

5.4.4 iid samples

If regularity condition holds

$$I_{(X_1,\cdots,X_n)}(\theta) = nI_X(X_1)(\theta)$$

5.4.5 Exp fam

For any S with $E[S(X)] < \infty$, it holds that $\frac{\partial}{\partial\theta}\int S(x)f_\theta(x)d\nu = \int S(x)\frac{\partial}{\partial\theta}f_\theta(x)d\nu$ and $I(\theta) = -E\left[\frac{\partial^2}{\partial\theta\partial\theta^T}\log f_\theta(X)\right]$
If $\underline{I}(\eta)$ is fisher information matrix for natural parameter η, then covariance matrix $Var(T) = \underline{I}(\eta)$.
Let $\psi = E[T(X)]$. Suppose $\bar{I}(\psi)$ is fisher info matrix for parameter ψ, then $Var(T) = [\bar{I}(\psi)]^{-1}$

5.4.6 Cramér-Rao Lower Bound

Suppose (1) Θ is an open set; P_θ has pdf f_θ (2) f_θ is differentiable and $0 = \frac{\partial}{\partial\theta}\int f_\theta(x)d\nu = \int \frac{\partial}{\partial\theta} f_\theta(x)d\nu, \theta \in \Theta$.
Suppose $g(\theta)$ is differentiable. $T(X)$ is unbiased estimator of $g(\theta)$ s.t. $g'(\theta) = \frac{\partial}{\partial\theta}\int T(x)f_\theta(x)d\nu = \int T(x)\frac{\partial}{\partial\theta} f_\theta(x)d\nu, \theta \in \Theta$.
Then

$$Var(T(X)) \geq \frac{g'(\theta)^2}{I(\theta)} = \left[\frac{\partial}{\partial\theta}g(\theta)\right]^T [I(\theta)]^{-1}\frac{\partial}{\partial\theta}g(\theta)$$

where $I(\theta) > 0$ for any $\theta \in \Theta$

5.4.7 CR LB for biasd estimator

$Var(T) \geq \frac{[g'(\theta)+b'(\theta)]^2}{I(\theta)}$

5.4.8 CR LB equality

CR achieve equality iff $T = \left[\frac{g'(\theta)}{I(\theta)}\right]\frac{\partial}{\partial\theta}\log f_\theta(X) + g(\theta)$ a.s. One such example is exp fam.

---

6 Asymptotics

6.1 Consistency of point estimators

$X = (X_1, \cdots, X_n)$ is sample from $P \in \mathcal{P}$ and $T_n(X)$ be estimator of θ for P.
[consistent] $\Leftrightarrow T_n(X) \to^P \theta$
[strongly consistent] $\Leftrightarrow T_n(X) \to^{\text{a.s.}} \theta$
[a_n-consistent] $\Leftrightarrow a_n(T_n(X) - \theta) = O_P(1)$, $\{a_n\} > 0$ and diverge to ∞
[L_r-consistent] $T_n(X) \to^{L^P} \theta$ for some fixed $r > 0$
A combination of LLN, CLT, Slustky's, continuous mapping, δ-method are used. If T_n is (strongly) consistent for θ and g is continuous at θ then $g(T_n)$ is (strongly) consistent for $g(\theta)$

6.1.1 Affine estimator

Consider $T_n = \sum_{i=1}^n c_{ni}X_i$
(1) If $c_{ni} = c_i/n$ satisfy (1) $\frac{1}{n}\sum_{i=1}^n c_i \to 1$ and $\sup_i |c_i| < \infty$ then T_n is strongly consistent.
(2) If population variance is finite, then T_n is consistent in mse $\Leftrightarrow \sum_{i=1}^n c_{ni} \to 1$ and $\sum_{i=1}^n c_{ni}^2 \to 0$

6.2 Asympotics bias, variance, MSE

[Approximate unbiased] Estimator $T_n(X)$ for θ is approximately unbiased if $b_{T_n}(P) \to 0$ as $n \to \infty$, $b_{T_n}(P) := ET_n(X) - \theta$

. . . . . .

When estimator's expectations or second moment are not well defined, we need asymptotic behaviours.
[Asymptotic statistics conditions] $\{a_n\} > 0$ and either (a) $a_n \to \infty$ or (b) $a_n \to a > 0$. If

$$a_n(T_n - \theta) \to^D Y$$

[Asymptotic expectation] If $a_n\xi_n \to^D \xi$, $E|\xi| < \infty$, then asymptotic expectation of ξ_n is $E\xi/a_n$
[Asymptotic bias] $\tilde{b}_{T_n} = EY/a_n$, asymptotically unbiased if $\lim_{n\to\infty}\tilde{b}_{T_n}(P) = 0$ for any $P \in \mathcal{P}$.
[Asymptotic MSE] amse is the asymptotic expectation of $(T_n - \theta)^2$ or $\text{amse}_{T_n}(P) = EY^2/a_n^2$
[Asymptotic Variance] $\sigma^2_{T_n}(P) = Var(Y)/a_n^2$
[Remark] $EY^2 \leq \liminf_{n\to\infty} E[a_n^2(T_n - v)^2]$ (amse is no greater than exact mse)

6.2.1 Asym Relative Efficiency

$e_{T_{1n},T_{2n}} = amse_{T_{2n}(P)}/amse_{T_{1n}(P)}$. Note efficiency of estimator T refers to $1/[I(\theta)MSE_T(\theta)]$

6.2.2 δ-method corollary

If $a_n \to \infty$, g is differentiable at θ, $U_n = g(T_n)$. Then amse of U_n is $[g'(\theta)^2EY^2]/a_n^2$, asym var of U_n is $[g'(\theta)^2Var(Y)]/a_n^2$

6.3 Properties of MOM

θ_n is unique if h^{-1} exists. Strongly consistent if h^{-1} is continuous via SLLN and continuous mapping. If h^{-1} is differentiable and $E|X_1|^{2k} < \infty$ then by CLT and δ-method. V_μ is $k \times k$ with $(i,j) = \mu_{i+j} - \mu_i\mu_j$

$$\sqrt{n}(\hat{\theta}_n - \theta) \to_D N(0, [\nabla g]^T V_\mu \nabla g)$$

MOM is $\sqrt{n}$-consistent, and if $k = 1$ $amse_{\hat{\theta}_n}(\theta) = g'(\mu_1)^2\sigma^2/n$, $\sigma^2 = \mu_2 - \mu_1^2$

6.4 Asym Properties of UMVUE

Typically consistent, exactly unbiased, ratio of mse over Cramér-Rao LB converges to 1 (asym they are the same).

6.5 Asymptotic properties of sample quantiles

$X_1, X_2, \cdots$ iid rvs with CDF F, $\gamma \in (0,1)$, $\hat{\theta}_n := \lfloor \gamma n \rfloor$-th order statistics. Suppose $F(\theta) = \gamma$ and $F'(\theta) > 0$ and exists.

$$\sqrt{n}(\hat{\theta}_n - \theta) \to^D N\left(0, \frac{\gamma(1-\gamma)}{[F'(\theta)]^2}\right)$$

6.6 Consistency and Asymptotic efficiency of MLEs and RLEs

6.6.1 Continuous in θ

Suppose (1) Θ is compact (2) $f(x|\theta)$ is continuous in θ for all x (3) there exists a function $M(x)$ s.t. $E_{\theta_0}|M(X)| < \infty$ and $|\log f(x|\theta) - log f(x|\theta_0)| \leq M(x)$ for all x and θ (4) identifiable $f(x|\theta) = f(x|\theta_0)$ ν-a.e. $\Rightarrow \theta = \theta_0$. Then for any sequence of MLE $\hat{\theta}_n \to_{\text{a.s.}} \theta_0$

6.6.2 Upper semi-continuous (usc)

$$\lim_{\rho \to 0} \left\{ \sup_{||\theta' - \theta|| < \rho} f(x|\theta') \right\} = f(x|\theta)$$

6.6.3 USC in θ

Suppose (1) Θ is compact with metric $d(\cdot,\cdot)$ (2) $f(x|\theta)$ is usc in θ and for all x (3) there exists a function $M(x)$ s.t. $E_{\theta_0}|M(X)| < \infty$ and $\log f(x|\theta) - \log f(x|\theta_0) \leq M(x)$ for all x and θ (4) for all $\theta \in \Theta$ and sufficiency small $\rho > 0$, $\sup_{d(\theta',\theta)<\rho} f(x|\theta')$ is measurable in x (5) identifiable $f(x|\theta) = f(x|\theta_0)$ ν-a.e. $\Rightarrow \theta = \theta_0$. Then $d(\hat{\theta}_n, \theta_0) \to_{\text{a.s.}} 0$

6.6.4 M-estimators

General method to find $\hat{\theta}_n$ maximises criterion function $S_\theta(x)$, for MLE $s_\theta(x) = \log f(x|\theta)$. $E_{\theta_0} s_\theta(X) < E_{\theta_0} s_{\theta_0}(X)$ $\forall\ \theta \neq \theta_0$.

$$\theta \mapsto S_n(\theta) = \frac{1}{n}\sum_{i=1}^n s_\theta(X_i)$$

6.6.5 Consistency of M-estimators

$S_n(\theta)$ is random function while $S(\theta)$ is fixed s.t. $\sup_{\theta \in \Theta} |S_n(\theta) - S(\theta)| \to_P 0$ and for every $\rho > 0$ $\sup_{\theta: d(\theta,\theta_0) \geq \rho} S(\theta) < S(\theta_0)$. Then any sequence of estimators $\hat{\theta}_n$ with $S_n(\hat{\theta}_n) \geq S_n(\theta_0) - o_P(1)$ converges in probability to θ_0

6.6.6 RLE: Roots of the Likelihood Equation

θ that solves $\frac{\partial}{\partial\theta} \log L_n(\theta) = 0$

6.6.7 Basic Regularity conditions

Suppose (1) Θ is open subset of $\mathcal{R}^k$ (2) $f(x|\theta)$ is twice continuously differentiable in θ for all x, and $\frac{\partial}{\partial\theta}\int f(x|\theta)d\nu = \int \frac{\partial}{\partial\theta} f(x|\theta)d\nu$, $\frac{\partial}{\partial\theta}\int \frac{\partial}{\partial\theta^T} f(x|\theta)d\nu = \int \frac{\partial^2}{\partial\theta\partial\theta^T} f(x|\theta)d\nu$. (3) $\Psi(x,\theta) = \frac{\partial^2}{\partial\theta\partial\theta^T}\log f(x|\theta)$, there exists a constant c and non-negative function H s.t. $EH(X) < \infty$ and $\sup_{||\theta - \theta_*|| < c} ||\Psi(x,\theta)|| \leq H(x)$. (4) Identifiable

6.6.8 Consistency of RLEs

Under basic regularity conditions, there exists a sequence of $\hat{\theta}_n$ s.t. $\frac{\partial}{\partial\theta}\log L_n(\hat{\theta}_n) = 0$ and $\hat{\theta}_n \to_{\text{a.s.}} \theta_*$. More useful if likelihood is concave or unique.

6.6.9 Asymptotic Normality of RLEs

Assume basic regularity conditions, and $I(\theta) = \int \frac{\partial}{\partial\theta} \log f(x|\theta) \left[\frac{\partial}{\partial\theta} \log f(x|\theta)\right]^T d\nu(x)$ is positive definite and $\theta = \theta_*$. Then any consistent sequence $\{\tilde{\theta_n}\}$ of RLE it holds

$$\sqrt{n}(\tilde{\theta_n} - \theta_*) \to_D N\left(0, \frac{1}{I(\theta_*)}\right)$$

6.6.10 NEF RLEs

Basic regularity condition (1, 2, 3, 4) holds due to proposition 3.2 and theorem 2.1, and result for (3). Only need to check condition on Fisher Info, then when n is large, there exists $\hat{\eta}_n$ s.t. $g(\hat{\eta}_n) = \hat{\mu}_n$ and $\hat{\eta}_n \to_{\text{a.s.}} \eta$

$$\sqrt{n}(\hat{\eta}_n - \eta) \to_D N\left(0, \left[\frac{\partial^2}{\partial\eta\partial\eta^T}\mathcal{C}(\eta)\right]^{-1}\right)$$

Where $g(\eta) = \frac{\partial\mathcal{C}(\eta)}{\partial\eta}$ and $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n T(X_i)$

6.6.11 Asym Covariance Matrix

$V_n(\theta)$ is $k \times k$ positive definite matrix called asym covariance matrix. $V_n(\theta)$ is usually in form of $n^{-\delta}V(\theta)$, higher δ means faster convergence.

$$[V_n(\theta)]^{-1/2}(\hat{\theta}_n - \theta) \to_D N_k(0, I_k)$$

6.6.12 Information Inequalities

$A \preccurlyeq B$ means $B - A$ is positive semi-definite. Suppose two estimators $\hat{\theta}_{1n}, \hat{\theta}_{2n}$ satisfy asym covariance matrix with $V_{1n}(\theta), V_{2n}(\theta)$. $\hat{\theta}_{1n}$ is asym more efficient thant $\hat{\theta}_{2n}$ if
(1) $V_{1n}(\theta) \preccurlyeq V_{2n}(\theta)$ for all $\theta \in \Theta$ and all large n
(2) $V_{1n}(\theta) \prec V_{2n}(\theta)$ for at least one $\theta \in \Theta$
But note $\hat{\theta}_n$ is asym unbiased but CR LB might not hold even if regularity condition is satisfied.

6.6.13 Hodges' estimator

$X_i \sim N(\theta, 1)$, $\hat{\theta}_n = \bar{X}_n$ if $\bar{X}_n \geq n^{-1/4}$ and $t\bar{X}_n$ otherwise. $V_n(\theta) = 1/n$ if $\theta \neq 0$ and t^2/n otherwise.
if $\theta \neq 0$: $\sqrt{n}(\hat{\theta} - \theta) = \sqrt{n}(\bar{X}_n - \theta) - (1-t)\sqrt{n}\bar{X}_n I_{|\bar{\theta}_n| < n^{-1/4}}$ if $\theta = 0$: $= t\sqrt{n}(\bar{X}_n - \theta) + (1-t)\sqrt{n}\bar{X}_n I_{|\bar{X}_n| \geq n^{-1/4}}$

6.6.14 Super-efficiency

Point where UMVUE failed Hodeges' estiamtor in information inequality (2). But under the basic regularity condition and if Fisher Information is positive definite at $\theta = \theta_*$, if $\hat{\theta}_n$ satisfies Asym covariance matrix, then there is a $\Theta_0 \subset \Theta$ with Lebesgue measure 0 s.t. information inequality (2) holds for any $\theta \notin \Theta_0$

6.6.15 Asym efficiency

Assume Fisher Info $I_n(\theta)$ is well-defined and positive definite for every n, seq of estimators $\{\hat{\theta}_n\}$ satisfies asym cov matrix is asym efficient or asym optimal if and only if $V_n(\theta) = [I_n(\theta)]^{-1}$.

6.6.16 One-step MLE

Often asym efficient, useful to adjust an non asym efficient estimators provided $\hat{\theta}_n^{(0)}$ is $\sqrt{n}$-consistent.

$$\hat{\theta}_n^{(1)} = \hat{\theta}_n^{(0)} - \left[\nabla s_n(\hat{\theta}_n^{(0)})\right]^{-1} s_n(\hat{\theta}_n^{(0)})$$