ECON-GA 2001

First Half

Quang Vuong

New York University

Revised October 2024

I. Probability Spaces and Random Variables

$(\Omega, \mathcal{F}, P)$ Probability space

$\rightarrow \Omega$: underlying space (nonempty)

$\rightarrow \mathcal{F}$: σ-algebra or σ-field

$\rightarrow P$: Probability measure

I.1. σ-algebra. (Bierens, Sections 1.1, 1.3–1.4)

Definition: An *algebra/field* is a nonempty collection of subsets of Ω satisfying

(i) If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$ (closed under complementation)

(ii) If $A, B \in \mathcal{F}$ then $A \cup B \in \mathcal{F}$ (closed under finite unions)

Remarks:

- In the definition, (ii) can be replaced by

(ii') If $A, B \in \mathcal{F}$ then $A \cap B \in \mathcal{F}$ (closed under finite intersections)

- Examples: $\mathcal{F}_1 = \{\emptyset, \Omega\}$ (trivial σ-algebra), $\mathcal{F}_2 = \{\emptyset, A, A^c, \Omega\}$, $\mathcal{F}_3 = 2^\Omega$ (power set)

Definition: A *σ-algebra/field* is a nonempty collection of subsets of Ω satisfying

(i) If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$ (closed under complementation)

(ii) If $A_1, A_2, \ldots \in \mathcal{F}$ then $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$ (closed under countable unions)

The pair $(\Omega, \mathcal{F})$ is called a *measurable space*.

Remarks:

- In the definition, (ii) can be replaced by

(ii') If $A_1, A_2, \ldots \in \mathcal{F}$ then $\cap_{i=1}^{\infty} A_i \in \mathcal{F}$ (closed under countable intersections)

- Examples: $\mathcal{F}_1$, $\mathcal{F}_2$, $\mathcal{F}_3$ are σ-algebras

Properties of σ-algebra:

(a) $\mathcal{F}$ is a σ-algebra $\Rightarrow$ $\mathcal{F}$ is an algebra. The reverse is not necessarily true.

(b) $\mathcal{F}$ is an algebra with a finite number of elements $\Rightarrow$ $\mathcal{F}$ is a σ-algebra.

- Let $\mathcal{F}_\theta$, $\theta \in \Theta$ be a (possibly uncountable) collection of σ-algebras of Ω indexed by θ. Then

  $\rightarrow \cap_{\theta\in\Theta}\mathcal{F}_\theta$ is a σ-algebra.

  $\rightarrow \cup_{\theta\in\Theta}\mathcal{F}_\theta$ is not necessarily a σ-algebra.

DEFINITION: The smallest σ-algebra containing a collection $\mathcal{C}$ of subsets of Ω is called the *σ-algebra generated by $\mathcal{C}$*. It is denoted $\sigma(\mathcal{C})$ and satisfies $\sigma(\mathcal{C}) = \cap_{\{\mathcal{F}:\mathcal{C}\subseteq\mathcal{F}\}}\mathcal{F}$.

EXAMPLES:

- The σ-algebra generated by unions $\vee_{\theta\in\Theta}\mathcal{F}_\theta \equiv \sigma(\cup_{\theta\in\Theta}\mathcal{F}_\theta)$

- Let $\Omega = \mathbb{R}^k$ and $\mathcal{C} = \{\times_{i=1}^k (a_i, b_i); a_i < b_i \in \mathbb{R}\}$. Then

  $\mathcal{B}^k \equiv \sigma(\mathcal{C})$ is the (Euclidean) *Borel σ-algebra* and any element of $\mathcal{B}^k$ is called a *Borel set*.

REMARKS:

$\rightarrow \mathcal{B}^k = \sigma(\{\times_{i=1}^k([a_i, b_i]; a_i \leq b_i \in \mathbb{R}\}) = \sigma(\{\times_{i=1}^k(-\infty, a_i]; a_i \in \mathbb{R}\})$, etc.

$\rightarrow \mathcal{B}^k = \sigma(\{\text{Open sets of } \mathbb{R}^{\mathrm{k}}\}) = \sigma(\{\text{Closed sets of } \mathbb{R}^{\mathrm{k}}\})$

$\rightarrow \mathcal{B}^k \neq 2^{\mathbb{R}^k}$ (Vitali example)

I.2. Probability Measures. (Bierens, Sections 1.1, 1.5–1.7)

DEFINITION: A *measure* μ on $(\Omega, \mathcal{F})$ is a mapping $\mathcal{F} \to I\!R \cup \{+\infty\}$ satisfying

(i) $\mu(A) \geq 0$ for all $A \in \mathcal{F}$ (nonnegativity)

(ii) $\mu(\emptyset) = 0$

(iii) If $A_1, A_2, \ldots \in \mathcal{F}$ are disjoint, then $\mu(\cup_{i=1}^{\infty} A_i) = \Sigma_{i=1}^{\infty} \mu(A_i)$ (countable additivity).

REMARKS: $(\Omega, \mathcal{F}, \mu)$ is called a *measured space.*

$\rightarrow$ μ is *finite* if $\mu(\Omega) < \infty$

$\rightarrow$ μ is σ*-finite* if $\Omega = \cup_{i=1}^{\infty} A_i$ with $\mu(A_i) < \infty$

EXAMPLES:

- Let $\mathcal{F} = 2^{\Omega}$ and $\mu_c(A) \equiv \# A$. Then μ_c is called the *counting measure*. It is finite if $|\Omega| < \infty$.

- Let $(\Omega, \mathcal{F}) = (I\!R^k, \mathcal{B}^k)$. The *Lebesgue measure* λ on $(I\!R^k, \mathcal{B}^k)$ is

$$\lambda(B) \equiv \inf_{B \subseteq \cup_{j=1}^{\infty} \{\times_{i=1}^{k} (a_{ij}, b_{ij})\}} \sum_{j=1}^{\infty} \prod_{i=1}^{k} (b_{ij} - a_{ij})$$

The Lebesgue measure λ is σ-finite.

DEFINITION: A *probability measure* P on a *measurable space* $(\Omega, \mathcal{F})$ is a measure with (ii) replaced by (ii) $P(\Omega) = 1$. The triplet $(\Omega, \mathcal{F}, P)$ is called a *probability space.*

REMARK: A probability measure P on an algebra $\mathcal{A}$ is defined similarly by restricting $\cup_{i=1}^{\infty} A_i \in \mathcal{A}$ in (iii). It can be uniquely extended to $\sigma(\mathcal{A})$. (Carathéodory Theorem)

EXAMPLE: Let $(\Omega, \mathcal{F}) = (\{1, 2, 3, 4, 5, 6\}, 2^{\Omega})$. The *uniform probability measure* is $P_0(A) = \frac{\#A}{6}$.

PROPERTIES OF MEASURES AND PROBABILITY MEASURES:

(a) $P(\emptyset) = 0$

(b) $P(A^c) = 1 - P(A)$ (only for probability measures)

(c) $A \subseteq B$ implies $P(A) \leq P(B)$

(d) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

(e) If $A_1 \subseteq A_2 \subseteq \ldots$, then $P(A_n) \uparrow P(\cup_{i=1}^{\infty} A_n)$

(f) If $A_1 \supseteq A_2 \supseteq \ldots$, then $P(A_n) \downarrow P(\cap_{i=1}^{\infty} A_n)$ (only for finite measures)

(g) $P(\cup_{i=1}^{\infty} A_n) \leq \Sigma_{i=1}^{\infty} P(A_n)$

I.3. Random Variables/Vectors. (Bierens, Sections 1.8, 1.10)

DEFINITION: Let $(\Omega, \mathcal{F}, P)$ be a probability space and $(\tilde{\Omega}, \tilde{\mathcal{F}})$ be a measurable space. A mapping $X : \Omega \to \tilde{\Omega}$ is a random element iff $X(\cdot)$ is $\mathcal{F}/\tilde{\mathcal{F}}$*-measurable*, i.e. iff

$$X^{-1}(\tilde{B}) \in \mathcal{F} \quad \text{for every} \quad \tilde{B} \in \tilde{\mathcal{F}}$$

where $X^{-1}(\tilde{B}) \equiv \{\omega \in \Omega : X(\omega) \in \tilde{B}\}$.

EXAMPLE: When $\tilde{\Omega} = \{f(\cdot) : \mathcal{T} \to I\!R\}$ and $\mathcal{F}$ is the σ-algebra generated by the open sets relative to a metric on $\tilde{\Omega}$, then $X(\cdot)$ is a *random process* indexed by $\mathcal{T}$ and denoted $X(\omega; t)$.

REMARKS:

- The collection $\mathcal{F}_X \equiv \{X^{-1}(\tilde{B}) : \tilde{B} \in \tilde{\mathcal{F}}\}$ is a σ-algebra called the *σ-algebra generated by X*.

  $X(\cdot)$ is a random element iff $\mathcal{F}_X \subseteq \mathcal{F}$.

- Let $P_X : \tilde{\mathcal{F}} \to I\!R \cup \{+\infty\}$ defined by $P_X(\tilde{B}) \equiv P[X^{-1}(\tilde{B})]$ for every $\tilde{B} \in \tilde{\mathcal{F}}$. Then $(\tilde{\Omega}, \tilde{\mathcal{F}}, P_X)$ is a probability space, and P_X is called the *probability measure induced by X*.

Definition: When $(\tilde{\Omega}, \tilde{\mathcal{F}}) = (I\!R^k, \mathcal{B}^k)$, the random element $X : \Omega \to I\!R^k$ is called a *random variable/vector* (r.v.).

Remarks:

- The $\mathcal{F}/\mathcal{B}^k$-measurability condition $\{X^{-1}(\tilde{B}) : \tilde{B} \in \mathcal{B}^k\} \subseteq \mathcal{F}$ is equivalent to

$$\{X^{-1}((-\infty, x]) : x \in I\!R^k\} \subseteq \mathcal{F}$$

- Let $F_X : I\!R^k \to [0, 1]$ defined by $F(x) \equiv P_X[(-\infty, x]] = P\{X^{-1}[(-\infty, x]]\}$ for every $x \in I\!R^k$. $F_X(\cdot)$ is called the *distribution function of* X.

Properties of Distribution Functions:

(a) A distribution function F_X on $I\!R^k$ corresponds to a unique probability measure P_X satisfying

$$P_X[(-\infty, x]] = F_X(x) \text{ for every } x \in I\!R^k.$$

(b) A distribution function $F_X(\cdot)$ is monotone nondecreasing, right-continuous with left-limits (*cadlag*), at most countable discontinuities, $\lim_{x\downarrow-\infty} F_X(x) = 0$ and $\lim_{x\uparrow+\infty} F_X(x) = 1$.

DEFINITION: (i) The events $A_1, A_2, \ldots$ in $\mathcal{F}$ of a probability space $(\Omega, \mathcal{F}, P)$ are (mutually) *independent* iff for every finite collection $A_{j_1}, \ldots, A_{j_n}$ and every n

$$P(\cap_{i=1}^{n} A_{j_i}) = \prod_{i=1}^{n} P(A_{j_i})$$

(ii) The σ-algebras $\mathcal{F}_1, \mathcal{F}_2, \ldots$ included in $\mathcal{F}$ on a probability space $(\Omega, \mathcal{F}, P)$ are (mutually) *independent* iff every sequence $A_1, A_2, \ldots$ with $A_j \in \mathcal{F}_j$ is independent.

(iii) The r.v. $X_1, X_2, \ldots$ on a probability space $(\Omega, \mathcal{F}, P)$ are (mutually) *independent* iff the sequence $\mathcal{F}_{X_1}, \mathcal{F}_{X_2}, \ldots$ is independent.

PROPERTY OF INDEPENDENCE:

- The r.v. $X_1, X_2, \ldots$ on a probability space $(\Omega, \mathcal{F}, P)$ are (mutually) independent iff

$$F_X(x) = \prod_{i=1}^{n} F_{X_{j_i}}(x_{j_i})$$

for every $x = (x_{j_1}, \ldots, x_{j_n})$, $(j_1, \ldots, j_n)$ and n, where $X = (X_{j_1}, \ldots, X_{j_n})$.

$F_X(\cdot)$ and $F_{X_j}(\cdot)$ are called the *joint and marginal distribution functions*, respectively.

II. Measurable Functions and Lebesgue Integration

II.1. Measurable Functions. (Bierens, Sections 2.2, 2.4)

DEFINITION: A $\mathcal{F}/\mathcal{B}^k$-*measurable function* $g(\cdot)$ is a mapping from a measurable space $(\Omega, \mathcal{F})$ to $(\mathbb{R}^k, \mathcal{B}^k)$ that satisfies $g^{-1}(B) \in \mathcal{F}$ for every $B \in \mathcal{B}^k$ where $g^{-1}(B) \equiv \{\omega : g(\omega) \in B\}$.

EXAMPLES:

- Same as a r.v. except that the latter is defined on $(\Omega, \mathcal{F}, P)$.

- When $(\Omega, \mathcal{F}) = (\mathbb{R}^\ell, \mathcal{B}^\ell)$, then $g(\cdot)$ is simply called a *measurable function.*

  $\rightarrow$ Continuous functions from $\mathbb{R}^\ell$ to $\mathbb{R}^k$ are measurable functions. The reverse is not true.

PROPERTIES:

(a) $g(\cdot) = [g_1(\cdot), \ldots, g_k(\cdot)]'$ is $\mathcal{F}/\mathcal{B}^k$-measurable iff $g_i(\cdot)$ is $\mathcal{F}/\mathcal{B}$-measurable for every $i = 1, \ldots, k$.

(b) Let $X(\cdot)$ and $Y(\cdot)$ be two r.v. on $(\Omega, \mathcal{F}, P)$. Then $\mathcal{F}_Y \subseteq \mathcal{F}_X$ iff there exists a measurable function $g(\cdot) : \mathbb{R}^{\dim X} \rightarrow \mathbb{R}^{\dim Y}$ such that $Y(\omega) = g[X(\omega)]$ for all $\omega \in \Omega$.

DEFINITION: A *simple function* $g(\cdot)$ is a $\mathcal{F}/\mathcal{B}$-measurable function from $(\Omega, \mathcal{F})$ to $(I\!R, \mathcal{B})$ of the form

$$g(\omega) = \sum_{j=1}^{m} b_j \, I\!I(\omega \in A_j)$$

where $m < \infty$, $b_1, \ldots, b_m \in I\!R$ and $\{A_j\}_{j=1}^m$ a partition of Ω with $A_j \in \mathcal{F}$ for $j = 1, \ldots, m$.

PROPERTIES:

(a) A function $g(\cdot) : (\Omega, \mathcal{F}) \to (I\!R, \mathcal{B})$ is $\mathcal{F}/\mathcal{B}$-measurable iff it is the (poinwise) limit of a sequence of simple functions $\{g_n(\cdot)\}_{j=1}^{\infty}$, i.e. $\lim_{j\to\infty} g_n(\omega) = g(\omega)$ for every $\omega \in \Omega$.

(b) If $f(\cdot)$ and $g(\cdot)$ are $\mathcal{F}/\mathcal{B}$-measurable, then $f(\cdot) + g(\cdot)$, $f(\cdot) - g(\cdot)$, $f(\cdot) \times g(\cdot)$ and $f(\cdot)/g(\cdot)$ (with $g(\cdot) \neq 0$) are $\mathcal{F}/\mathcal{B}$-measurable,

(c) Let $\{g_n(\cdot)\}_{n=1}^{\infty}$ be a sequence of $\mathcal{F}/\mathcal{B}$-measurable functions. Provided the RHS are well-defined,

(c.1) $\underline{f}(\cdot) \equiv \inf_{n\geq 1} g_n(\cdot)$ and $\overline{f}(\cdot) \equiv \sup_{n\geq 1} g_n(\cdot)$ are $\mathcal{F}/\mathcal{B}$-measurable,

(c.2) $\underline{h}(\cdot) \equiv \liminf_{n\to\infty} g_n(\cdot)$ and $\overline{h}(\cdot) \equiv \limsup_{n\to\infty} g_n(\cdot)$ are $\mathcal{F}/\mathcal{B}$-measurable,

(c.3) $g(\cdot) \equiv \lim_{n\to\infty} g_n(\cdot)$ is $\mathcal{F}/\mathcal{B}$-measurable.

II.2. Lebesgue Integration. (Bierens, Sections 2.2, 2.4, 4.3–4.4)

DEFINITION: Let $(\Omega, \mathcal{F}, \mu)$ be a measured space.

(i) The *integral* of a simple function $g(\cdot)$ w.r.t. μ is

$$\int_\Omega g(\omega) d\mu(\omega) \equiv \sum_{j=1}^{m} b_j \mu(A_j)$$

(ii) The *integral* of a (μ-a.e.) nonnegative $\mathcal{F}/\mathcal{B}$-measurable function $g(\cdot)$ w.r.t. μ is

$$\int_\Omega g(\omega) d\mu(\omega) \equiv \sup_{0 \leq g_S(\cdot) \leq g(\cdot)} \int_\Omega g_S(\omega) d\mu(\omega)$$

where $g_S(\cdot)$ is a simple function.

(iii) The *integral* of a $\mathcal{F}/\mathcal{B}$-measurable function $g(\cdot)$ w.r.t. μ is

$$\int_\Omega g(\omega) d\mu(\omega) \equiv \int_\Omega g_+(\omega) d\mu(\omega) - \int_\Omega g_-(\omega) d\mu(\omega)$$

where $g_+(\cdot) \equiv \max\{g(\cdot), 0\}$ and $g_-(\cdot) \equiv \max\{-g(\cdot), 0\}$ provided $\int_\Omega g_+(\omega) d\mu(\omega) < \infty$ or $\int_\Omega g_-(\omega) d\mu(\omega) < \infty$. If both are finite, then $g(\cdot)$ is *μ-integrable.*

Remarks:

- $g(\cdot)$ is μ-integrable iff $|g(\cdot)|$ is μ-integrable, in which case $\left|\int_\Omega g(\omega)d\mu(\omega)\right| \leq \int_\Omega |g(\omega)|d\mu(\omega)$.
- If $g(\cdot)$ is μ-integrable, then $g(\cdot) < \infty$ μ-a.e. (allowing $g(\cdot)$ to have $\pm\infty$ values)
- Let $\int_A g(\omega)d\mu(\omega) \equiv \int_\Omega \mathbb{I}(\omega \in A)g(\omega)d\mu(\omega)$ for $A \in \mathcal{F}$. If $\mu(A) = 0$, then $\int_A g(\omega)d\mu(\omega) = 0$.
- When $(\Omega, \mathcal{F}) = (\mathbb{R}, \lambda)$, then $\int_A g(x)d\lambda(x)$ is called the *Lebesgue integral of $g(\cdot)$ on A.*
  - $\rightarrow$ If a bounded function $g(\cdot)$ is Riemann integrable, then $g(\cdot)$ is Lebesgue integrable and $\int_a^b g(x)dx = \int_{[a,b]} g(x)d\lambda(x) < \infty$. Counterexample: $g(\cdot)$ is the Dirichlet function.
  - $\rightarrow$ A bounded function $g(\cdot) : [a, b] \rightarrow \mathbb{R}$ is Riemann integrable iff its set of discontinuities is of λ-measure zero. (Lebesgue Theorem)

Properties of Lebesgue Integration: Let $g(\cdot)$ and $h(\cdot)$ be μ-integrable. Then

(a) If $g(\cdot) = 0$ μ-a.e., then $\int_\Omega g(\omega)d\mu(\omega) = 0$.

(b) $\int_\Omega [\alpha g(\omega) + \beta h(\omega)]d\mu(\omega) = \alpha\int_\Omega g(\omega)d\mu(\omega) + \beta\int_\Omega h(\omega)d\mu(\omega)$ for $\alpha, \beta \in \mathbb{R}$

(c) If $g(\cdot) \leq h(\cdot)$ μ-a.e., then $\int_\Omega g(\omega)d\mu(\omega) \leq \int_\Omega h(\omega)d\mu(\omega)$

Limits and Lebesgue Integration:

- Let $g(\cdot)$ be μ-integrable, and $A_1, A_2, \ldots \in \mathcal{F}$.

  If $\lim_{n\to\infty} \mu(A_n) = 0$, then $\lim_{n\to\infty} \int_{A_n} g(\omega)d\mu(\omega) = 0$.

- *Monotone Convergence Theorem*: Let $\{g_n(\cdot)\}_{n=1}^{\infty}$ be a nondecreasing sequence of nonnegative $\mathcal{F}/\mathcal{B}$-measurable functions converging (pointwise) to a function $g(\cdot)$ μ-a.e. Then

$$\lim_{n\to\infty} \int_\Omega g_n(\omega)d\mu(\omega) = \int_\Omega g(\omega)d\mu(\omega)$$

- *Lebesgue Dominated Convergence Theorem*: Let $\{g_n(\cdot)\}_{n=1}^{\infty}$ be a sequence of $\mathcal{F}/\mathcal{B}$-measurable functions converging (pointwise) to a function $g(\cdot)$ satisfying $|g_n(\cdot)| \leq h(\cdot)$ μ-a.e. for every n where $h(\cdot)$ is μ-integrable. Then, the previous conclusion holds.

Remark: Let $Q(\theta) \equiv \int_\Omega g(\omega;\theta)d\mu(\omega)$ where $g(\cdot;\theta)$ is μ-integrable for every $\theta \in \Theta \subset \mathbb{R}^p$. Then

(i) $Q(\cdot)$ is continuous and

(ii) $Q'(\theta) = \int_\Omega g'(\omega;\theta)d\mu(\omega)$, where $g'(\omega;\cdot)$ is the derivative of $g(\omega;\cdot)$.

See e.g. Theorem 16.8 in Billinglsey (1995) for conditions.

CHANGE-OF-VARIABLE: Let $g(\cdot)$ and $h(\cdot)$ be $\mathcal{B}^k/\mathcal{B}$ and $\mathcal{B}^k/\mathcal{B}^k$-measurable. If $h(\cdot)$ is one-to-one from $A\in\mathcal{B}^k \to B\in\mathcal{B}^k$, and continuously differentiable with Jacobian $J(x)\equiv\partial h/\partial x'$, then

$$\int_B g(y)\ d\lambda(y) = \int_A g[h(x)] \cdot |J(x)|\ d\lambda(x).$$

PRODUCT MEASURE AND FUBINI'S THEOREM. (Billingsley, Section 18)

Let $(\Omega, \mathcal{F}, \mu)$ and $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mu})$ be two measure spaces with σ-finite measures μ and $\tilde{\mu}$.

DEFINITION: The *product σ-algebra* on $\Omega \times \tilde{\Omega}$ is $\mathcal{F} \times \tilde{\mathcal{F}} \equiv \sigma(\{A \times \tilde{A} : A \in \mathcal{F}, \tilde{A} \in \tilde{\mathcal{F}}\})$.
The *product measure* ν on $\mathcal{F} \times \tilde{\mathcal{F}}$ is the unique σ-finite measure satisfying

$$\nu(A \times \tilde{A}) = \mu(A) \cdot \tilde{\mu}(\tilde{A}) \quad \text{for every} \quad A \in \mathcal{F}, \tilde{A} \in \tilde{\mathcal{F}}.$$

The triplet $(\Omega \times \tilde{\Omega}, \mathcal{F} \times \tilde{\mathcal{F}}, \nu)$ is called the *product measure space.*

PROPERTY: (Fubini's Theorem) Let $g(\cdot) : \Omega \times \tilde{\Omega} \to I\!R$ be $(\mathcal{F} \times \tilde{\mathcal{F}})/\mathcal{B}$-measurable. Then

$$\int_{\Omega\times\tilde{\Omega}} g(\varpi)d\nu(\varpi) = \int_\Omega \left[\int_{\tilde{\Omega}} g(\omega, \tilde{\omega})d\tilde{\mu}(\tilde{\omega})\right] d\mu(\omega) = \int_{\tilde{\Omega}} \left[\int_\Omega g(\omega, \tilde{\omega})d\mu(\omega)\right] d\tilde{\mu}(\tilde{\omega})$$

where $\varpi = (\omega, \tilde{\omega})$, provided one of the integrals is finite.

III. Expectations and Conditional Expectations

III.1. Expectations. (Bierens, Sections 2.5–2.8, 5.1)

DEFINITION: Let $(\Omega, \mathcal{F}, P)$ be a probability space, $X(\cdot)$ a r.v. in $I\!R^{\dim X}$, and $g(\cdot) : I\!R^{\dim X} \to I\!R^k$ measurable. Then $\mathrm{E}[g(X)] \equiv \int_\Omega g[X(\omega)]dP(\omega) \equiv \left[\int_\Omega g_1[X(\omega)]dP(\omega), \ldots, \int_\Omega g_k[X(\omega)]dP(\omega)\right]'$.

EXAMPLES:

- When $\dim X = 1$, the *raw and central p-th moment* of X are $\mathrm{E}[X^p]$ and $\mathrm{E}[X - \mathrm{E}(X)]^p$.
  $\to \mathrm{E}[X]$ and $\mathrm{E}[X - E(X)]^2$ are the *mean* and *variance* of X.
- The *variance-covariance matrix* of X is $\mathrm{E}[X - E(X)][X - E(X)]' = \mathrm{E}[XX'] - \mathrm{E}[X]\mathrm{E}[X]'$.

PROPERTIES:

(a) $\mathrm{E}[g(X)] = \int_{R^{\dim X}} g(x)dP_X(x) = \int_{R^{\dim Y}} ydP_Y(y)$ where $Y \equiv g(X)$, provided one of the integrals is finite.

(b) The r.v. X and Y are independent iff $\mathrm{E}[g(X)h(Y)] = \mathrm{E}[g(X)] \cdot \mathrm{E}[h(Y)]$ for all measurable functions $g(\cdot)$ and $h(\cdot)$ on $I\!R^{\dim X}$ and $I\!R^{\dim Y}$.

Some Inequalities:

- Markov inequality $\Pr[|X| \geq \epsilon] \leq \mathrm{E}[|X|^p]/\epsilon^p$ for every $p > 0$. ($\Rightarrow$ Chebyshev inequality)
- Jensen inequality $\mathrm{E}[g(X)] \leq g(\mathrm{E}[X])$ for $g(\cdot)$ concave on $\mathcal{S}_X$. ($\Rightarrow$ Liapounov inequality)
- Holder inequality $\mathrm{E}(|XY|) \leq (\mathrm{E}[|X|^p])^{1/p}(\mathrm{E}[|Y|^q])^{1/q}$ for $1/p+1/q=1$. ($\Rightarrow$ Cauchy-Schwarz in.)
- Triangle inequality $(\mathrm{E}[|X+Y|^p])^{1/p} \leq (\mathrm{E}[|X|^p])^{1/p} + (\mathrm{E}[|Y|^p])^{1/p}$ for $p \geq 1$. ($\Rightarrow$ Minkowski in.)

Definition: The *moment generating function* of X is $M_X(t) \equiv \mathrm{E}[\exp(t'X)]$ for $t \in \mathcal{D}_X \subseteq \mathbb{R}^{\dim X}$. The *characteristic function* of X is $\phi_X(t) \equiv \mathrm{E}[\exp(it'X)]$ for $t \in \mathbb{R}^{\dim X}$.

Properties:

(a) If $(-t_o, t_o) \subseteq \mathcal{D}_X \subseteq \mathbb{R}$ with $t_o > 0$, then $M^{(p)}(0) = \mathrm{E}[X^p] < \infty$ for all $p \in \mathbb{N}$,

$M_X(t) = \sum_{p=0}^{\infty} \frac{\mathrm{E}[X^p]}{p!} t^p$ and $P_X(\cdot)$ is uniquely determined by $M_X(\cdot)$ (and by the moments of X).

(b) $P_X(\cdot)$ is uniquely determined by $\phi_X(\cdot)$.

When $\dim X = 1$, $\phi_X(\cdot)$ is uniformly continuous. Moreover, $\phi^{(p)}(0) = i^p \mathrm{E}[X^p]$ if $\mathrm{E}[X^p] < \infty$.

(c) $\phi_X(t) = M_X(it)$ and $M_X(t) = \phi_X(-it)$ whenever $M_X(\cdot)$ is well-defined.

III.2. Conditional Expectations. (Bierens, Chapter 3)

$\rightarrow$ $\mathrm{E}[Y|X]$ where $Y(\cdot)$ is a r.v. and $X(\cdot)$ is a random element on $(\Omega, \mathcal{F}, P)$

EXAMPLE: $(\Omega, \mathcal{F}, P) = (\{1,2,3,4,5,6\}, 2^{\Omega}, P_0)$ where P_0 is the uniform probability measure. Let $Y(\omega) = \omega$ and $X(\omega) = \mathbb{1}[\omega \in \{2,4,6\}]$. Then $\mathrm{E}[Y|X] = 3 + X$.

DEFINITION: Let Y be a random variable on $(\Omega, \mathcal{F}, P)$ with $\mathrm{E}[Y] < \infty$. Let $\mathcal{F}_0 \subseteq \mathcal{F}$ be a σ-algebra. The *conditional expectation of* Y *given* $\mathcal{F}_0$ is a mapping $\mathrm{E}[Y|\mathcal{F}_0](\cdot) : \Omega \rightarrow \mathbb{R}$ satisfying

(i) $\mathrm{E}[Y|\mathcal{F}_0](\cdot)$ is $\mathcal{F}_0/\mathcal{B}$-measurable,

(ii) $\int_A \mathrm{E}[Y|\mathcal{F}_0](\omega) dP(\omega) = \int_A Y(\omega) dP(\omega)$, i.e., $\mathrm{E}\{\mathbb{1}_A \, \mathrm{E}[Y|\mathcal{F}_0]\} = \mathrm{E}[\mathbb{1}_A \, Y]$ for every $A \in \mathcal{F}_0$.

REMARKS:

- $\mathrm{E}[Y|\mathcal{F}_0](\cdot)$ is uniquely defined P-a.s.

- The *conditional expectation of* Y *given* X is $\mathrm{E}[Y|X](\cdot) \equiv \mathrm{E}[Y|\mathcal{F}_X](\cdot)$.

$\rightarrow$ When $X \in \mathbb{R}^k$, then $\mathrm{E}[Y|X](\cdot) = g[X(\cdot)]$ for some $\mathcal{B}^k/\mathcal{B}$-measurable function $g(\cdot)$.

$\rightarrow$ $\mathrm{E}[Y|X=x] \equiv g(x) = \mathrm{E}[Y|X](\omega)$ for all $\omega \in X^{-1}(x)$.

PROPERTIES OF CONDITIONAL EXPECTATIONS:

(a) If $Y = c$, then $\mathrm{E}[Y|\mathcal{F}_0] = c$.

(b) $\mathrm{E}[\alpha X + \beta Y|\mathcal{F}_0] = \alpha \mathrm{E}[X|\mathcal{F}_0] + \beta \mathrm{E}[Y|\mathcal{F}_0]$.

(c) If $X(\cdot) \leq Y(\cdot)$, then $\mathrm{E}[X|\mathcal{F}_0] \leq \mathrm{E}[Y|\mathcal{F}_0]$.

(d) If $\mathcal{F}_0 = \{\emptyset, \Omega\}$, then $\mathrm{E}[Y|\mathcal{F}_0] = \mathrm{E}[Y]$.

(e) If $\mathcal{F}_Y \perp \mathcal{F}_0$, then $\mathrm{E}[Y|\mathcal{F}_0] = \mathrm{E}[Y]$.

(f) If $\mathcal{F}_X \subseteq \mathcal{F}_0$, then $\mathrm{E}[X|\mathcal{F}_0] = X$ and $\mathrm{E}[XY|\mathcal{F}_0] = X\mathrm{E}[Y|\mathcal{F}_0]$.

(g) *Law of Iterated Expectations:* If $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}$, then

$$\mathrm{E}\{\mathrm{E}[Y|\mathcal{F}_0]|\mathcal{F}_1\} = \mathrm{E}\{\mathrm{E}[Y|\mathcal{F}_1]|\mathcal{F}_0\} = \mathrm{E}[Y|\mathcal{F}_0].$$

(h) The r.v.s X and Y are *conditionally independent* given $\mathcal{F}_0$ iff $\mathrm{E}[g(X)h(Y)|\mathcal{F}_0] =$ $\mathrm{E}[g(X)|\mathcal{F}_0] \cdot \mathrm{E}[h(Y)|\mathcal{F}_0]$ for all measurable functions $g(\cdot)$ and $h(\cdot)$ on $\mathbb{R}^{\dim X}$ and $\mathbb{R}^{\dim Y}$.

(i) *Best MSE:* If $\mathrm{E}[Y^2] < \infty$, then $\mathrm{E}[Y|X] = \arg\min_{h \in \mathcal{H}} MSE(h) \equiv \mathrm{E}[Y - h(X)]^2$ where $\mathcal{H}$ is the set of squared-integrable (measurable) functions of X.

Limits and Conditional Expectations:

- *Monotone Convergence Theorem*: Let $\{Y_n(\cdot)\}_{n=1}^{\infty}$ be a nondecreasing sequence of nonnegative r.v.s with $\mathrm{E}[\sup_{n\geq 1} Y_n] < \infty$. Then $\lim_{n\to\infty} \mathrm{E}[Y_n|\mathcal{F}_0] = \mathrm{E}[\lim_{n\to\infty} Y_n|\mathcal{F}_0]$.
- *Dominated Convergence Theorem*: Let $\{Y_n(\cdot)\}_{n=1}^{\infty}$ be a sequence of r.v.s converging to $Y(\cdot)$ P-a.s. and satisfying $|Y_n(\cdot)| \leq M(\cdot)$ with $\mathrm{E}[M|\mathcal{F}_0] < \infty$. Then, $\lim_{n\to\infty} \mathrm{E}[Y_n|\mathcal{F}_0] = \mathrm{E}[Y|\mathcal{F}_0]$.
- Let $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \ldots \subseteq \mathcal{F}$. Then $\lim_{n\to\infty} \mathrm{E}[Y|\mathcal{F}_n] = \mathrm{E}[Y|\mathcal{F}_\infty]$ where $\mathcal{F}_\infty \equiv \vee_{n=1}^{\infty}\mathcal{F}_n = \sigma(\cup_{n=1}^{\infty}\mathcal{F}_n)$.

  $\rightarrow \lim_{n\to\infty} \mathrm{E}[Y_t|Y_{t-1},\ldots,Y_{t-n}] = \mathrm{E}[Y_t|Y_{t-1},Y_{t-2},\ldots]$

Definition: Let Y be a r.v. on $(\Omega,\mathcal{F},P)$. The *conditional probability measure of* $Y \in I\!R^\ell$ *given* $\mathcal{F}_0 \subseteq \mathcal{F}$ is the mapping $P_{Y|\mathcal{F}_0}(\cdot,\cdot) : (B,\omega) \in \mathcal{B}^\ell \times \Omega \rightarrow P_{Y|\mathcal{F}_0}(B,\omega) \equiv \mathrm{E}[1\!\!I(Y \in B)|\mathcal{F}_0](\omega)$.

Remarks:

- For each $\omega \in \Omega$, $(I\!R, \mathcal{B}^\ell, P_{Y|\mathcal{F}_0}(\cdot,\omega))$ is a probability space.
- The *cond. prob. measure of* $Y \in I\!R^\ell$ *given* $X \in I\!R^k$ is $P_{Y|X}(\cdot|x) \equiv P_{Y|\mathcal{F}_X}(\cdot,\omega)$ for $\omega \in X^{-1}(x)$.

  $\rightarrow \int_A P_{Y|X}(B|x)dP_X(x) = P[X \in A, Y \in B]$ for every $(A,B) \in \mathcal{B}^k \times \mathcal{B}^\ell$.

IV. Densities and Distributions

IV.1. Densities. (Bierens, Sections 1.9)

DEFINITION: Let X be r.v. from $(\omega, \mathcal{F}, P)$ to $(\mathbb{R}^k, \mathcal{B}^k, P_X)$. Then P_X is *absolutely continuous w.r.t. a measure* μ on $(\mathbb{R}^k, \mathcal{B}^k)$, i.e. $P_X \ll \mu$, iff $\mu(B) = 0 \Rightarrow P_X(B) = 0$ for every $B \in \mathcal{B}^k$.

EXAMPLES: $k = 1$

- When $\mu(B) = \mu_c(B) \equiv \#(B \cap \mathbb{N})$, then P_X (or X) is *discrete* and $p_X(i) \equiv P[X = i]$.

- When $\mu(B) = \lambda(B)$, then P_X (or X) is *(absolutely) continuous.*

RADON-NIKODYM THEOREM: If $P_X \ll \mu$ with σ-finite μ, then there exists a nonnegative $\mathcal{B}^k/\mathcal{B}$-measurable function $f_X(\cdot)$ satisfying $P_X(B) = \int_B f_X(x) d\mu(x)$ for every $B \in \mathcal{B}^k$.

$\rightarrow$ The *(probability) density* $f_X(\cdot)$ is μ-a.e. unique.

$\rightarrow$ $\int_B g(x) dP_X(x) = \int_B g(x) f_X(x) d\mu(x)$ and $F_X(\cdot) = \int_{(-\infty, \cdot]} f_X(x) d\mu(x)$.

$\rightarrow$ When $\mu = \mu_c$, then $\int_B g(x) d\mu_c(x) = \sum_{i \in B} g(i) p_X(i)$ and $F_X(x) = \sum_{i \leq x} p_X(i)$.

$\rightarrow$ When $\mu = \lambda$ and $F_X(\cdot)$ is continuously differentiable, then $f_X(\cdot) = F'_X(\cdot)$.

IV.2. Some Discrete and Continuous R.V.s. (Bierens, Sections 4.1, 4.5–4.6)

SOME UNIVARIATE DISCRETE R.V.S:

$\rightarrow$ *Binomial Distribution*: $X \sim \mathcal{B}(n, p)$ where $n \in \mathbb{N}$ and $0 < p < 1$.

$p_X(i) = \binom{n}{p}\, p^i(1-p)^{n-i}$ for $i = 0, 1, \ldots, n$.

$\mathrm{E}[X] = np$, $\mathrm{Var}[X] = np(1-p)$, $\phi_X(t) = [1 - p + p\exp(it)]^n$.

$\rightarrow$ *Poisson Distribution*: $X \sim \mathcal{P}(\lambda)$ where $\lambda > 0$

$p_X(i) = \frac{\lambda^i}{i!}\exp(-\lambda)$ for $i = 0, 1, 2, \ldots$.

$\mathrm{E}[X] = \lambda$, $\mathrm{Var}[X] = \lambda$, $\phi_X(t) = \exp\{\lambda[\exp(it) - 1]\}$.

SOME UNIVARIATE CONTINUOUS R.V.S:

$\rightarrow$ *Normal Distribution*: $X \sim \mathcal{N}(\mu, \sigma^2)$ where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$.

$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}]$ for $x \in \mathbb{R}$.

$\mathrm{E}[X] = \mu$, $\mathrm{Var}[X] = \sigma^2$, $\phi_X(t) = \exp(i\mu t - \sigma^2 t/2)$.

Standard Normal Distribution Z: $\mu = 0$ and $\sigma^2 = 1$.

SOME UNIVARIATE CONTINUOUS R.V.S: (continued)

→ *Chi-Square Distribution*: $X \sim \chi^2_k = \sum_{j=1}^{k} Z_j^2$ where $k \in I\!N$ and Z_j are i.i.d. $\mathcal{N}(0,1)$.

$f_X(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} \exp(-x/2)$ for $x \geq 0$, where $\Gamma(\alpha) \equiv \int_0^\infty x^{\alpha-1} \exp(-x) dx$.

$\mathrm{E}[X] = k$, $\mathrm{Var}[X] = 2k$, $\phi_X(t) = \frac{1}{(1-2it)^{p/2}}$.

→ *Student-t Distribution*: $X \sim t_k = \frac{Z}{\sqrt{\chi^2_k/k}}$ where $k \in I\!N$ and $Z \sim \mathcal{N}(0,1) \perp \chi^2_k$.

$f_X(x) = \frac{1}{\sqrt{k}B(\frac{1}{2},\frac{k}{2})} \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}}$ for $x \in I\!R$, where $B(\alpha,\beta) \equiv \frac{\Gamma(\alpha)\cdot\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.

$\mathrm{E}[X] = 0$ for $k \geq 2$, $\mathrm{Var}[X] = \frac{k}{k-2}$ for $k \geq 3$, $\phi_X(t) = \frac{2\cdot\Gamma[(k+1)/2]}{\sqrt{k\pi}\cdot\Gamma(k/2)} \int_0^\infty \frac{\cos(tx)}{(1+x^2/k)^{(k+1)/2}} dx$.

Cauchy Distribution: $k = 1$. $\mathrm{E}[X^p]$ does not exist or is infinite for any $p \in I\!N$.

→ *F Distribution*: $X \sim F(m,n) = \frac{\chi^2_m/m}{\chi^2_n/n}$ where $m,n \in I\!N$ and $\chi^2_m \perp \chi^2_n$.

$f_X(x) = \frac{m^{m/2}n^{n/2}}{B(m/2,n/2)} \frac{x^{m/2-1}}{(n+mx)^{(m+n)/2}}$ for $x \in [0,+\infty]$.

$\mathrm{E}[X] = \frac{n}{n-2}$ for $n \geq 3$, $\mathrm{Var}[X] = \frac{2n^2(m+n-4)}{m(n-2)^2(n-4)}$ for $n \geq 5$.

$\mathrm{E}[X] = \infty$ for $n = 1,2$, $\mathrm{Var}[X]$ does not exist or is infinite for $n = 1,2,3,4$.

IV.3. The Multivariate Normal Distribution. (Bierens, Sections 5.2–5.3, 5.5)

DEFINITION: The r.v. X with values in $\mathbb{R}^k$ has a *multivariate normal distribution* $\mathcal{N}(\mu, \Sigma)$ where $\mu \in \mathbb{R}^k$ and Σ is positive definite matrix iff its density w.r.t. Lebesgue measure is

$$f_X(x) = \frac{1}{(\sqrt{2\pi})^n \sqrt{\det(\Sigma)}} \exp\left[-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right] \quad \text{for } x \in \mathbb{R}^k.$$

PROPERTIES:

(a) $\mathrm{E}[X] = \mu$, $\mathrm{Var}[X] = \Sigma$, $\phi_X(t) = \exp(it'\mu - t'\Sigma t/2)$.

(b) If $Y = A + BX$ where $X \sim \mathcal{N}(\mu, \Sigma)$, then $Y \sim\sim \mathcal{N}(A + B\mu, B\Sigma B')$.

(c) If $X \sim \mathcal{N}(\mu, \Sigma)$, then its components are independent iff Σ is diagonal.

(d) If $Y \in \mathbb{R}$ and $X \in \mathbb{R}^k$ are (jointly) normal distributed with $\mathrm{E}[Y]{=}\mu_Y$, $\mathrm{E}[X]{=}\mu_X$, $\mathrm{Var}[Y]{=}\sigma_Y^2$, $\mathrm{Var}[X]{=}\Sigma_{XX}$ nonsingular, and $\mathrm{E}[XY]{=}\Sigma_{XY}$, then $P_{Y|X}(\cdot|x)$ is normal with $\mathrm{E}[Y|X] = \mu_Y + (X{-}\mu_X)'\Sigma_{XX}^{-1}\Sigma_{XY}$ and $\mathrm{Var}[Y|X] = \sigma_Y^2 - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$ with $\Sigma_{YX}'{=}\Sigma_{XY}$.

(e) If $X \sim \mathcal{N}(0, \Sigma)$, then $X'\Sigma^{-1}X \sim \chi_k^2$.

V. Modes of Convergence (Bierens, Sections 6.2–6.3, 6.5–6.7)

DEFINITION: Let X and $\{X_n\}_{n=1}^{\infty}$ be random variables on $(\Omega, \mathcal{F}, P)$.

(i) X_n *converges to* X *almost surely*, i.e. $X_n \xrightarrow{a.s.} X$, iff $P[\lim_{n\to\infty} X_n = X] = 1$,

(ii) X_n *converges to* X *in probability*, i.e. $X_n \xrightarrow{p} X$, iff $\forall \epsilon > 0 \ \lim_{n\to\infty} P[|X_n - X| \leq \epsilon] = 1$,

(iii) X_n *converges to* X *in distribution*, i.e. $X_n \xrightarrow{d} X$, iff $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$

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

REMARKS:

- Alternative definitions:

  (i) $X_n \xrightarrow{a.s.} X$ iff $\forall \epsilon > 0 \ \lim_{n\to\infty} P[\sup_{m\geq n} |X_m - X| \leq \epsilon] = 1$,

  (iii) $X_n \xrightarrow{d} X$ iff $\lim_{n\to\infty} \mathrm{E}[g(X_n)] = \mathrm{E}[g(X)]$ for all bounded continuous $g(\cdot) : \mathbb{R} \to \mathbb{R}$.

- The multivariate case: $X \in \mathbb{R}^k$ and $X_n \in \mathbb{R}^k$

  (i)-(ii) $X_n \xrightarrow{a.s./p} X$ iff $X_n \xrightarrow{a.s./p} X$ componentwise,

  (iii) $X_n \xrightarrow{d} X$ iff $\lambda' X_n \xrightarrow{d} \lambda' X$ for all $\lambda \in \mathbb{R}^k$ iff $\lim_{n\to\infty} \phi_{X_n}(t) = \phi_X(t)$ for all $t \in \mathbb{R}^k$.

Properties:

(a) $X_n \xrightarrow{a.s.} X \Rightarrow X_n \xrightarrow{p} X \Rightarrow X_n \xrightarrow{d} X$.

The reverse is not true except: $X_n \xrightarrow{p} c \Leftrightarrow X_n \xrightarrow{d} c$ where c is a constant.

(b) *Continuous Mapping (Slutzky) Theorem*: Let $g(\cdot) : \mathbb{R}^k \to \mathbb{R}^\ell$ be continuous (P_X-a.s.).

If $X_n \xrightarrow{a.s./p/d} X$, then $g(X_n) \xrightarrow{a.s./p/d} g(X)$.

Examples: Let $X = (X_1', X_2')'$ and $X_n = (X_{1n}', X_{2n}')'$. If $X_n \xrightarrow{a.s./p/d} X$, then

$X_{1n}+X_{2n} \xrightarrow{a.s./p/d} X_1+X_2$, $X_{1n}X_{2n} \xrightarrow{a.s./p/d} X_1X_2$, $X_{1n}/X_{2n} \xrightarrow{a.s./p/d} X_1/X_2$ when $P[X_2=0]=0$.

Basic LLN and CLT: Let X_i, $i = 1, 2, \ldots$ be i.i.d. and $\overline{X}_n \equiv \frac{1}{n}\sum_{i=1}^{n} X_i$.

- Strong LLN (Kolmogorov): If $\mathrm{E}[X_i] \equiv \mu < \infty$, then $\overline{X}_n \xrightarrow{a.s.} \mu$.

- CLT: If $\mathrm{E}[X_i] \equiv \mu < \infty$ and $\mathrm{Var}[X_i] \equiv \Sigma < \infty$ nonsingular, then $\sqrt{n}\,\Sigma^{-1/2}(\overline{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, I)$.

Remarks: Extensions to non i.i.d X_i. See Bierens, Sections 7.2-7.3, 7.5. Let $\mu_i \equiv \mathrm{E}[X_i]$.

- SLLNs: $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_i) \xrightarrow{a.s.} 0 \Rightarrow$ WLLNs: $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_i) \xrightarrow{p} 0 \left(\Leftrightarrow \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_i) \xrightarrow{d} 0\right)$.

- CLTs: When $\dim X_i = 1$, $\frac{\sum_{i=1}^{n}(X_i - \mathrm{E}[X_i])}{\sqrt{\mathrm{Var}(\sum_{i=1}^{n} X_i)}} \xrightarrow{d} \mathcal{N}(0, 1)$.

The linear Regression Model (Bierens, Section 5.7)

- Data: $(Y_i, X_i) \in I\!R \times I\!R^k, i = 1, 2, \ldots$ i.i.d. on $(\Omega, \mathcal{F}, P_0)$.

- Model: $\mathrm{E}[Y|X] = X'\beta_0$ or $Y_i = X_i'\beta_0 + \epsilon_i$ where $\mathrm{E}[\epsilon_i|X_i] = 0$.

- *OLS estimator*: $\hat{\theta}_{OLS} = \arg\min_\beta (1/n) \Sigma_{i=1}^n (Y_i - X_i'\beta)^2$. Provided $\mathbf{X}$ is *full-column rank*

$$\hat{\beta}_{OLS} = \left(\frac{1}{n}\sum_{i=1}^n X_i X_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n X_i Y_i\right) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

- Properties of $\hat{\beta}_{OLS}$: Assume $\mathrm{E}[XX'] \equiv \Sigma_{XX}$ is p.d., and $0 < \sigma_\epsilon^2(X) \equiv \mathrm{E}[\epsilon^2|X] < \infty$

(a) Finite sample property: $\mathrm{E}[\hat{\beta}_{OLS}] = \beta_0$, i.e. $\hat{\beta}_{OLS}$ *unbiased estimator* of β_0, etc.

(b) Asymptotic properties: $\hat{\beta}_{OLS} \xrightarrow{a.s.} \beta_0$ and $\sqrt{n}(\hat{\beta}_{OLS} - \beta_0) \xrightarrow{d} \mathcal{N}\left(0, \Sigma_{XX}^{-1}\mathrm{E}[\sigma_\epsilon^2(X)XX']\Sigma_{XX}^{-1}\right)$.

(c) Approximation: $\hat{\beta}_{OLS} \approx \mathcal{N}(\beta_0, (\mathbf{X}'\mathbf{X})^{-1}\left[\Sigma_{i=1}^n e_i^2 X_i X_i'\right](\mathbf{X}'\mathbf{X})^{-1})$ where $e_i \equiv Y_i - X_i'\hat{\beta}_{OLS}$.

$\rightarrow (\hat{\beta}_j - \beta_{j0})/\hat{\sigma}_j \approx \mathcal{N}(0,1)$ where $\hat{\sigma}_j^2 \equiv \left[(\mathbf{X}'\mathbf{X})^{-1}\left[\Sigma_{i=1}^n e_i^2 X_i X_i'\right](\mathbf{X}'\mathbf{X})^{-1}\right]_{jj}$ for $j = 1, \ldots, k$.

$\hat{\sigma}_j$ is called White (1980) *Heteroscedasticity Robust Standard Error*.

(d) Homoscedastic case $\sigma_\epsilon^2(\cdot) = \sigma_{\epsilon 0}^2$: $\hat{\beta}_{OLS} \approx \mathcal{N}(\beta_0, \tilde{\sigma}_\epsilon^2(\mathbf{X}'\mathbf{X})^{-1})$ where $\tilde{\sigma}_\epsilon^2 \equiv \frac{1}{n-k}\Sigma_{i=1}^n e_i^2$.

$\rightarrow$ When $\epsilon_i \sim \mathcal{N}(0, \sigma_{\epsilon 0}^2)$, then $(\hat{\beta}_j - \beta_{j0})/\tilde{\sigma}_j \sim t_{n-k}$ where $\tilde{\sigma}_j^2 \equiv \tilde{\sigma}_\epsilon^2[(\mathbf{X}'\mathbf{X})^{-1}]_{jj}$.

VI. M-Estimation (Bierens, Sections 6.4, 6.9)

DEFINITION: An *extremum or M-estimator* $\hat{\theta}_n \in \Theta \subseteq I\!R^k$ is a solution of $\max_{\theta\in\Theta} Q_n(\theta)$ or $\min_{\theta\in\Theta} \tilde{Q}_n(\theta)$, where $Q_n(\theta) = -\tilde{Q}_n(\theta)$ is a (measurable) function of the data.

EXAMPLES:

- OLS estimator $\hat{\beta}_{OLS} \equiv \arg\min_{\beta\in R^k} \tilde{Q}_n(\beta) = \frac{1}{n}\Sigma_{i=1}^n (Y_i - X_i'\beta)^2$.
- NLLS estimator $\hat{\theta}_{NLLS} \equiv \arg\min_{\theta\in R^k} \tilde{Q}_n(\beta) = \frac{1}{n}\Sigma_{i=1}^n [Y_i - m(X_i;\theta)]^2$.
- IV estimator $\hat{\beta}_{IV} \equiv \arg\min_{\beta\in R^k} \tilde{Q}_n(\beta) = \|\frac{1}{n}\Sigma_{i=1}^n Z_i(Y_i - X_i'\beta)\|_{A_n}$.
- GMM estimator $\hat{\theta}_{GMM} \equiv \arg\min_{\theta\in\Theta} \tilde{Q}_n(\beta) = \|\frac{1}{n}\Sigma_{i=1}^n h(Y_i, X_i, Z_i;\theta)\|_{A_n}$.
- ML estimator $\hat{\theta}_{ML} \equiv \arg\max_{\theta\in\Theta} Q_n(\beta) = \frac{1}{n}\Sigma_{i=1}^n \log f(Y_i|X_i;\theta)$.

Master Consistency Theorem: For a nonstochastic function $Q_\infty(\cdot)$ on Θ and $\theta_* \in \Theta$. If

(i) *Separation*: $\sup_{\theta\in\Theta\cap\mathcal{N}^c} Q_\infty(\theta) < Q_\infty(\theta_*)$, for any neighborhood $\mathcal{N}$ of θ_*,

(ii) *Uniform convergence*: $\sup_{\theta\in\Theta} |Q_n(\theta) - Q_\infty(\theta)| \xrightarrow{a.s./p} 0$,

then $\hat{\theta}_n \xrightarrow{a.s./p} \theta_*$ and θ_* is called the *pseudo-true parameter value*.

Proof of Strong Consistency: Consider only the ω for which $\lim_{n\to\infty}\sup_{\theta\in\Theta}|Q_n(\theta)-Q_\infty(\theta)|=0$. Thus for $\epsilon>0$ and n sufficiently large, $Q_\infty(\theta)-\epsilon/3<Q_n(\theta)<Q_\infty(\theta)+\epsilon/3$ for all $\theta\in\Theta$. Hence, (a) $Q_\infty(\theta_*)-\epsilon/3<Q_n(\theta_*)$ and (b) $Q_n(\hat{\theta}_n)<Q_\infty(\hat{\theta}_n)+\epsilon/3$.

Because $Q_n(\hat{\theta}_n)\geq Q_n(\theta)$ for all $\theta\in\Theta$, then $Q_n(\hat{\theta}_n)>Q_\infty(\theta_*)-\epsilon/3$ by (a). Hence, by (b)

$$Q_\infty(\hat{\theta}_n)>Q_n(\hat{\theta}_n)-\epsilon/3>Q_\infty(\theta_*)-2\epsilon/3>Q_\infty(\theta_*)-\epsilon.$$

Let $\mathcal{N}$ be an arbitrary neighborhood of θ_* and set $\epsilon=Q_\infty(\theta_*)-\sup_{\theta\in\Theta\cap\mathcal{N}^c}Q_\infty(\theta)$. Thus, $Q_\infty(\hat{\theta}_n)>\sup_{\theta\in\Theta\cap\mathcal{N}^c}Q_\infty(\theta)\Rightarrow\hat{\theta}_n\in\Theta\cap\mathcal{N}$.

REMARKS:

- Condition (i) $\Rightarrow\theta_*$ is the unique maximizer of $Q_\infty(\cdot)$ on Θ. See Lemma below for the reverse.

- Condition (ii) is often proved using a uniform SLLN. See below.

- $\theta_*\neq\theta_0$. If condition (i) is satisfied by θ_0, then $\theta_*=\theta_0$.

- Neither $Q_n(\cdot)$ nor $Q_\infty(\cdot)$ need to be continuous on Θ.

- Generalization to $\Theta\subseteq\mathcal{G}$ with a metric d on $\mathcal{G}$.

LEMMA: Suppose (i) Θ is compact, (ii) $Q_\infty(\cdot)$ is continuous on Θ, and (iii) *Identification*: θ_* is the unique maximizer of $Q_\infty(\cdot)$ on Θ. Then condition (i) in the basic consistency theorem holds.

Proof of Lemma: By (i) $\Theta \cap \mathcal{N}^c$ is compact for any neighborhood $\mathcal{N}$ of θ_*. Thus, by (ii) $\sup_{\theta\in\Theta\cap\mathcal{N}^c} Q_\infty(\theta) = Q_\infty(\overline{\theta})$ for some $\overline{\theta} \in \Theta$. By (iii) $Q_\infty(\overline{\theta}) < Q_\infty(\theta_*)$.

EXAMPLE: $\tilde{Q}_\infty(\beta) \equiv \mathrm{E}[Y - X'\beta]^2$ and $\beta_* \equiv \arg\min_{\beta\in B} \tilde{Q}_\infty(\beta)$. If $\mathrm{E}[Y^2] < \infty$ and $\mathrm{E}[XX'] < \infty$ nonsingular, then $\tilde{Q}_\infty(\cdot)$ is continuous on B and $\beta_* = \arg\min_{\beta\in B} \mathrm{E}[\mathrm{E}(Y|X) - X'\beta]^2$.

$\rightarrow \beta_* = (\mathrm{E}[XX'])^{-1}\mathrm{E}[XY]$ if the latter $\in B \Rightarrow$ If $\mathrm{E}[Y|X] = X'\beta_0$ with $\beta_0 \in B$, then $\beta_* = \beta_0$.

BASIC UNIFORM SLLN: Let $X_1, X_2, \ldots$ be i.i.d. r.v.s in $\mathbb{R}^p$ and $g(\cdot;\cdot) : \mathbb{R}^p \times \Theta \rightarrow \mathbb{R}$ be $\mathbb{R}^p/\mathbb{R}$-measurable and continuous on compact $\Theta \subset \mathbb{R}^k$. If $\sup_{\theta\in\Theta} |g(\cdot;\theta)| \leq M(\cdot)$ with $\mathrm{E}[M(X_i)] < \infty$, then $\sup_{\theta\in\Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \{g(X_i;\theta) - \mathrm{E}[g(X_i;\theta)]\} \right| \xrightarrow{a.s.} 0$ and $\mathrm{E}[g(X_i;\cdot)]$ is continuous on Θ.

REMARKS:

- Extensions to non i.i.d X_i. See Bierens, Section 7.4.

- Extensions to *empirical processes*: $\sup_{g\in\mathcal{G}} \left| \frac{1}{n} \Sigma_{i=1}^{n} \{g(X_i) - \mathrm{E}[g(X_i)]\} \right| \xrightarrow{a.s.} 0$.

EXAMPLE (CONTINUED): $\tilde{Q}_n(\beta) \equiv \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'\beta)^2$ with B compact. We have $\tilde{g}(y, x; \beta) \equiv (y - x'\beta)^2 \leq 2[y^2 + (x'\beta)^2] \leq 2(y^2 + \|x\|^2\|\beta\|^2) \leq M(y, x) \equiv 2(y^2 + M_o^2\|x\|^2)$. If $\mathrm{E}[Y_i^2] < \infty$ and $\mathrm{E}[X_i X_i'] < \infty$, then $\mathrm{E}[M(Y_i, X_i)] < \infty$. Hence, $\sup_{\beta \in B} |\tilde{Q}_n(\beta) - \tilde{Q}_\infty(\beta)| \xrightarrow{a.s./p} 0$.

Master Asymptotic Normality Theorem: Let $\hat{\theta}_n \xrightarrow{a.s./p} \theta_*$. If

(i) $\theta_* \in \mathrm{int}(\Theta)$, and $Q_n(\cdot)$ is twice continuously differentiable in a neighborhood $\mathcal{N}$ of θ_*,

(ii) $\sup_{\theta \in \mathcal{N}} \|\partial^2 Q_n(\theta)/\partial\theta\partial\theta' - H(\theta)\| \xrightarrow{a.s./p} 0$ where $H(\cdot)$ is continuous and nonsingular at θ_*,

(iii) $\sqrt{n}\; \partial Q_n(\theta_*)/\partial\theta \xrightarrow{d} \mathcal{N}(0, \Sigma)$ for Σ nonsingular,

then $\sqrt{n}(\hat{\theta}_n - \theta_*) \xrightarrow{d} \mathcal{N}(0, H^{-1}\Sigma H^{-1})$ where $H \equiv H(\theta_*)$.

Proof of Asymptotic Normality: By (i) $\partial Q_n(\hat{\theta}_n)/\partial\theta = 0$ where $\hat{\theta}_n \in \mathcal{N}$ with probability one or approaching one. Taking a second-order Taylor expansion at θ_* on the neighborhood $\mathcal{N}$ gives

$$0 = \sqrt{n}\frac{\partial Q_n(\theta_*)}{\partial\theta} + \frac{\partial^2 Q_n(\overline{\theta}_n)}{\partial\theta\partial\theta'}\sqrt{n}(\hat{\theta}_n - \theta_*) \qquad \text{where } \overline{\theta}_n \in [\theta_*, \hat{\theta}_n] \text{ (componentwise)}$$

By (ii) $\partial^2 Q_n(\overline{\theta}_n)/\partial\theta\partial\theta' \xrightarrow{a.s./p} H = H(\theta_*)$ since $\overline{\theta}_n \xrightarrow{a.s./p} \theta_*$. By (iii) and the continuous mapping theorem, the result follows since H is nonsingular.

REMARK: Condition (ii) by USLLN while condition (iii) from CLT at θ_*. See NLLS and MLE.

EXAMPLE (CONTINUED): $\frac{\partial^2 \tilde{Q}_n(\beta)}{\partial\beta\partial\beta'} = \frac{2}{n}\sum_{i=1}^{n} X_i X_i' \xrightarrow{a.s./p} 2\Sigma_{XX}$, $\sqrt{n}\frac{\partial \tilde{Q}_n(\beta_*)}{\partial\beta} = \frac{-2}{\sqrt{n}}\sum_{i=1}^{n} X_i(Y_i - X_i'\beta_*) \xrightarrow{d}$ $\mathcal{N}(0, 4\mathrm{E}[\sigma^2_{\epsilon_*}(X)XX'])$, where $\sigma_{\epsilon_*}(X) \equiv \mathrm{E}[\epsilon^2_*|X]$ and $\epsilon_* \equiv Y - X'\beta_*$. Thus, $\sqrt{n}(\hat{\beta}_{OLS} - \beta_*) \xrightarrow{d} \mathcal{N}(0, \Sigma^{-1}_{XX}\mathrm{E}[\sigma^2_{\epsilon_*}(X)XX']\Sigma^{-1}_{XX})$ same result as before with β_* instead of β_0.

$\rightarrow$ Approximation: $\hat{\beta}_{OLS} \approx \mathcal{N}(\beta_*, (\mathbf{X}'\mathbf{X})^{-1}\left[\sum_{i=1}^{n} e_i^2 X_i X_i'\right](\mathbf{X}'\mathbf{X})^{-1})$ where $e_i \equiv Y_i - X_i'\hat{\beta}_{OLS}$.

THEOREM (DELTA METHOD): Let $\sqrt{n}(X_n - c) \xrightarrow{d} \mathcal{N}_p(0, \Omega)$ and $g(\cdot) : I\!R^p \rightarrow I\!R^m$ be continuously differentiable on a neighborhood of $c \in I\!R^p$. Then $\sqrt{n}[g(X_n) - g(c)] \xrightarrow{d} \mathcal{N}_m\Big(0, \frac{\partial g(c)}{\partial c'}\Omega\frac{\partial g(c)'}{\partial c}\Big)$.

REMARKS:

- Different from continuous mapping theorem.

- $n[g(X_n) - g(c)]'\left[\frac{\partial g(X_n)}{\partial c'}\hat{\Omega}\frac{\partial g(X_n)'}{\partial c}\right]^{-1}[g(X_n) - g(c)] \xrightarrow{d} \chi^2_p$ where $\hat{\Omega} \xrightarrow{a.s./p} \Omega$ if $\frac{\partial g(c)}{\partial c'}\Omega\frac{\partial g(c)'}{\partial c}$ nonsingular.

$\rightarrow$ Application to testing $H_0 : g(\theta_*) = 0$ vs. $H_A : g(\theta_*) \neq 0$. If $\sqrt{n}(\hat{\theta}_n - \theta_*) \xrightarrow{d} \mathcal{N}_k(0, \Omega)$,
then $n\ g(\hat{\theta}_n)'\left[\frac{\partial g(\hat{\theta}_n)}{\partial\partial'}\hat{\Omega}\frac{\partial g(\hat{\theta}_n)'}{\partial\theta}\right]^{-1} g(\hat{\theta}_n) \xrightarrow{d} \chi^2_p$ under H_0 and $\xrightarrow{a.s./p} +\infty$ under H_1
where $\hat{\Omega} \equiv \hat{H}^{-1}\hat{\Sigma}\hat{H}^{-1} \xrightarrow{a.s./p} \Omega = H^{-1}\Sigma H^{-1}$ and $\hat{H} = \frac{\partial^2 Q_n(\hat{\theta}_n)}{\partial\theta\partial\theta'}$. See below for $\hat{\Sigma}$.

VII. NLLS and ML Estimation (Bierens, Sections 6.4, 8.1-8.2, 8.4)

VII.1. Non-Linear Least Squares: $\hat{\theta}_{NLLS} \equiv \arg\min_{\theta\in\Theta} \tilde{Q}_n(\theta) = \frac{1}{n}\Sigma_{i=1}^n [Y_i - m(X_i;\theta)]^2$

THEOREM (NLLS): Let $(Y_1, X_1), (Y_2, X_2), \ldots$ be i.i.d. r.v.s in $\mathbb{R}\times\mathbb{R}^p$ and $m(\cdot;\cdot) : \mathbb{R}^p\times\Theta \to \mathbb{R}$ be measurable on $\mathbb{R}^p$ and continuous on $\Theta \subset \mathbb{R}^k$. Suppose that $\mathrm{E}[Y^2] < \infty$ and θ_* uniquely solves $\min_{\theta\in\Theta} Q_\infty = \mathrm{E}[Y - m(X;\theta)]^2$.

(i) If Θ is compact and $\sup_{\theta\in\Theta} |m(\cdot;\theta)| \le M(\cdot)$ with $\mathrm{E}[M^2(X)] < \infty$, then $\hat{\theta}_{NLLS} \xrightarrow{a.s./p} \theta_*$.

(ii) If in addition $\theta_* \in \mathrm{int}(\Theta)$ with $m(\cdot,\cdot)$ twice cont. differentiable on a neighborhood $\mathcal{N}$ of θ_*,

$\sup_{\theta\in\mathcal{N}} \|\frac{\partial m(\cdot;\theta)}{\partial\theta}\| \le M_1(\cdot)$ with $\mathrm{E}[M_1^2(X)] < \infty$, $\sup_{\theta\in\mathcal{N}} \|\frac{\partial^2 m(\cdot;\theta)}{\partial\theta\partial\theta'}\| \le M_2(\cdot)$ with $\mathrm{E}[M_2^2(X)] < \infty$,

then $\sqrt{n}(\hat{\theta}_{NLLS} - \theta_*) \xrightarrow{d} \mathcal{N}(0, H^{-1}\Sigma H^{-1})$ provided $H = 2\mathrm{E}\left[\epsilon_* \frac{\partial^2 m(X;\theta_*)}{\partial\theta\partial\theta'} - \frac{\partial m(X;\theta_*)}{\partial\theta}\frac{\partial m(X;\theta_*)}{\partial\theta'}\right]$

is nonsingular and $\Sigma = 4\mathrm{E}\left[\sigma^2_{\epsilon_*}(X)\frac{\partial m(X;\theta_*)}{\partial\theta}\frac{\partial m(X;\theta_*)}{\partial\theta'}\right]$ with $\sigma^2_{\epsilon_*}(X) = \mathrm{E}[\epsilon_*^2|X]$ and $\epsilon_* = Y - m(X;\theta_*)$.

REMARKS: Let $e_i \equiv Y_i - m(X_i;\hat{\theta}_{NLLS})$. Then $\hat{\Sigma} = \frac{4}{n}\sum_{i=1}^n e_i^2 \frac{\partial m(X_i;\hat{\theta}_{NLLS})}{\partial\theta}\frac{\partial m(X_i;\hat{\theta}_{NLLS})}{\partial\theta'} \xrightarrow{a.s./p} \Sigma$

and $\hat{H} = \frac{2}{n}\sum_{i=1}^n \left[e_i \frac{\partial^2 m(X;\hat{\theta}_{NLLS})}{\partial\theta\partial\theta'} - \frac{\partial m(X_i;\hat{\theta}_{NLLS})}{\partial\theta}\frac{\partial m(X_i;\hat{\theta}_{NLLS})}{\partial\theta'}\right] \xrightarrow{a.s./p} H$.

- If $\mathrm{E}[Y|X] = m(X;\theta_0)$ for $\theta_0 \in \Theta$, then $\theta_* = \theta_0$. If, in addition $\sigma^2_{\epsilon_*}(X) = \sigma_0^2$, then

$H = -2\mathrm{E}\left[\frac{\partial m(X;\theta_0)}{\partial\theta}\frac{\partial m(X;\theta_0)}{\partial\theta'}\right]$, $\Sigma = 4\sigma_0^2\mathrm{E}\left[\frac{\partial m(X;\theta_0)}{\partial\theta}\frac{\partial m(X;\theta_0)}{\partial\theta'}\right]$, $H^{-1}\Sigma H^{-1} = \sigma_0^2\left(\mathrm{E}\left[\frac{\partial m(X;\theta_0)}{\partial\theta}\frac{\partial m(X;\theta_0)}{\partial\theta'}\right]\right)^{-1}$

VII.2. Maximum Likelihood: $\hat{\theta}_{ML} \equiv \arg\max_{\theta\in\Theta} Q_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} \log f(Y_i|X_i;\theta)$

THEOREM (ML): Let $(Y_1, X_1), (Y_2, X_2), \ldots$ be i.i.d. r.v.s in $\mathbb{R}^m \times \mathbb{R}^p$. Let the *model* be a family $\{f(\cdot|\cdot;\theta) : \theta \in \Theta\}$ of conditional densities w.r.t. a measure μ on $\mathbb{R}^m$ and continuous on $\Theta \subset \mathbb{R}^k$. Suppose that θ_* uniquely solves $\max_{\theta\in\Theta} Q_\infty = \mathrm{E}[\log f(Y|X;\theta)]$.

(i) If Θ is compact and $\sup_{\theta\in\Theta} |\log f(\cdot|\cdot;\theta)| \leq M(\cdot,\cdot)$ with $\mathrm{E}[M(Y,X)] < \infty$, then $\hat{\theta}_{ML} \xrightarrow{a.s./p} \theta_*$.

(ii) If in addition $\theta_* \in \mathrm{int}(\Theta)$ with $f(\cdot|\cdot;\cdot)$ twice continuously differentiable on a neighborhood $\mathcal{N}$ of θ_*, $\sup_{\theta\in\mathcal{N}} \|\frac{\partial^2 \log f(\cdot|\cdot;\theta)}{\partial\theta\partial\theta'}\| \leq M_2(\cdot,\cdot)$, $\mathrm{E}[M_2(Y,X)] < \infty$ and $\mathrm{E}\left[\frac{\partial^2 \log f(Y|X;\theta_*)}{\partial\theta\partial\theta'}\right]$ nonsingular, then

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_*) \xrightarrow{d} \mathcal{N}(0, A^{-1}BA^{-1})$$

where $A = A(\theta_*) \equiv \mathrm{E}\left[\frac{\partial^2 \log f(Y|X;\theta_*)}{\partial\theta\partial\theta'}\right]$ and $B = B(\theta_*) \equiv \mathrm{E}\left[\frac{\partial \log f(Y|X;\theta_*)}{\partial\theta}\frac{\partial \log f(Y|X;\theta_*)}{\partial\theta'}\right]$.

REMARKS:

- If $\sup_{\theta\in\mathcal{N}} \|\frac{\partial \log f(\cdot;\theta)}{\partial\theta}\| \leq M_1(\cdot)$ with $\mathrm{E}[M_1(X)] < \infty$, then

$$\hat{A} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log f(Y_i|X_i;\hat{\theta}_{ML})}{\partial\theta\partial\theta'} \xrightarrow{a.s./p} A \text{ and } \hat{B} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log f(Y_i|X_i;\hat{\theta}_{ML})}{\partial\theta}\frac{\partial \log f(Y_i|X_i;\hat{\theta}_{ML})}{\partial\theta'} \xrightarrow{a.s./p} B.$$

-*Kullback-Leibler Divergence*: $\theta_* = \arg\min_{\theta\in\Theta} KLIC[f_{Y|X}(\cdot|\cdot); f(\cdot|\cdot;\theta)] \equiv \mathrm{E}\left[\log \frac{f_{Y|X}(Y|X)}{f(Y|X;\theta)}\right] \geq 0$.

DEFINITION: $\{f(\cdot|\cdot;\theta) : \theta \in \Theta\}$ is *correctly specified* iff $f_{Y|X}(\cdot|\cdot) = f(\cdot|\cdot;\theta_0)$ a.s. for $\theta_0 \in \Theta$. It is *one-to-one parameterized* iff $f(\cdot|\cdot;\theta_1) = f(\cdot|\cdot,\theta_2)$ a.s. $\Leftrightarrow \theta_1 = \theta_2$ for any $\theta_1, \theta_2 \in \Theta$.

LEMMA: If $\{f(\cdot|\cdot;\theta) : \theta \in \Theta\}$ is correctly specified and one-to-one parameterized, then θ_0 uniquely solves $\max_{\theta\in\Theta} \mathrm{E}[\log f(Y|X;\theta)]$.

Proof of Lemma: If $f(y|x;\theta_0) > 0$, then $\log \frac{f(y|x;\theta)}{f(y|x;\theta_0)} \leq \frac{f(y|x;\theta)}{f(y|x;\theta_0)} - 1$ with equality iff $\frac{f(y|x;\theta)}{f(y|x;\theta_0)} = 1$. Thus $\mathrm{E}\left[\log \frac{f(y|x;\theta)}{f(y|x;\theta_0)}\right] \leq \mathrm{E}\left[\frac{f(y|x;\theta)}{f(y|x;\theta_0)}\right] - 1 \leq \int_{\{f(y|x;\theta_0)>0\}} f(y|x;\theta) d\mu(y) dF_X(x) - 1 \leq 0$. Hence, $\theta_0 = \arg\max_{\theta\in\Theta} \mathrm{E}[\log f(Y|X;\theta)]$. Moreover, $\mathrm{E}\left[\log \frac{f(y|x;\theta)}{f(y|x;\theta_0)}\right] = 0 \Rightarrow \mathrm{E}\left[\log \frac{f(y|x;\theta)}{f(y|x;\theta_0)} - \frac{f(y|x;\theta)}{f(y|x;\theta_0)} + 1\right] = 0$ $\Rightarrow \log \frac{f(y|x;\theta)}{f(y|x;\theta_0)} = \frac{f(y|x;\theta)}{f(y|x;\theta_0)} - 1$ a.s. $\Rightarrow \frac{f(y|x;\theta)}{f(y|x;\theta_0)} = 1$ a.s. $\Rightarrow \theta = \theta_0$, i.e. θ_0 is the unique maximizer.

LEMMA (INFORMATION MATRIX EQUALITY): Assume that $\{f(\cdot|\cdot;\theta) : \theta \in \Theta\}$ is correctly specified. If one can switch integration and differentiation, then $A(\theta_0) + B(\theta_0) = 0$.

Proof of Lemma: Differentiating twice $\int f(y|x;\theta) d\mu(y) = 1$ w.r.t. θ gives

$$\int \frac{\partial^2 \log f(y|x;\theta)}{\partial\theta\partial\theta'} f(y|x;\theta) d\mu(y) + \int \frac{\partial \log f(y|x;\theta)}{\partial\theta} \frac{\partial \log f(y|x;\theta)}{\partial\theta'} f(y|x;\theta) d\mu(y) = 0 \text{ for every } \theta \in \Theta.$$

Letting $\theta = \theta_0$ in $A(\theta)$ and $B(\theta)$ gives the result because $f_{Y|X}(\cdot|\cdot) = f(\cdot|\cdot;\theta_0)$.

REMARKS:

- *Information Matrix Test*: White (1982) $H_0 : A(\theta_*) + B(\theta_*) = 0$ vs. $H_A : A(\theta_*) + B(\theta_*) \neq 0$.

- If $\{f(\cdot|\cdot;\theta) : \theta \in \Theta\}$ is correctly specified and one-to-one parameterized, then

$\rightarrow \sqrt{n}(\hat{\theta}_{ML} - \theta_0) \xrightarrow{d} \mathcal{N}(0, B^{-1}) = \mathcal{N}(0, -A^{-1}) \Rightarrow \hat{\theta}_{ML} \approx \mathcal{N}\left(\theta_0, -\left[\sum_{i=1}^{n} \frac{\partial^2 \log f(Y_i|X_i;\hat{\theta}_{ML})}{\partial\theta\partial\theta'}\right]^{-1}\right)$.

$\rightarrow \hat{\theta}_{ML}$ is ***asymptotically efficient*** among consistent estimators $\hat{\theta}_n$, i.e., $\lim_{n\to\infty}\mathrm{Var}[\sqrt{n}\hat{\theta}_n] \geq B^{-1}$.

$B = \mathrm{E}\left[\frac{\partial \log f(Y|X;\theta_0)}{\partial\theta}\frac{\partial \log f(Y|X;\theta_0)}{\partial\theta'}\right]$ is the *Fisher Information Matrix for one observation.*

THEOREM (CRAMER-RAO): If $\hat{\theta}_n$ is an unbiased estimator of θ_0, then $\mathrm{Var}[\sqrt{n}\hat{\theta}_n] \geq B^{-1}$.

Proof of Cramer-Rao Bound: We have $\int g_n(y_1, x_1, \ldots, y_n, x_n)\, \Pi_{i=1}^{n} f(y_i|x_i;\theta_0) d\mu(y_i) dF_X(x_i) = \theta_0$ for any $\theta_0 \in \Theta$, where $\hat{\theta}_n = g_n(Y_1, X_1, \ldots, Y_n, X_n)$. Differentiating w.r.t. θ_0 gives $\int g_n(y_1, x_1, \ldots, y_n, x_n)\left[\Sigma_{i=1}^{n} \frac{\partial \log f(y_i|x_i;\theta_0)}{\partial\theta'}\right] \Pi_{i=1}^{n} f(y_i|x_i;\theta_0) d\mu(y_i) dF_X(x_i) = I$, i.e., $\mathrm{E}[\hat{\theta}_n S_n'] = I$ where $S_n \equiv \Sigma_{i=1}^{n} \frac{\partial \log f(Y_i|X_i;\theta_0)}{\partial\theta}$. But $\mathrm{E}[S_n] = 0$ by differentiating $\int f(y_i|x_i;\theta_0) d\mu(y_i) dF_X(x_i) = 1$ w.r.t. θ_0. Hence $\mathrm{Cov}[\hat{\theta}_n, S_n] = I$ and $\mathrm{Var}[S_n] = nB$. Now, $\mathrm{Var}[\hat{\theta}_n] = \mathrm{Var}[(nB)^{-1}S_n + \hat{\theta}_n - (nB)^{-1}S_n] = (nB)^{-1} + \mathrm{Var}[\hat{\theta}_n - (nB)^{-1}S_n] \geq (nB)^{-1}$ since $\mathrm{Cov}[(nB)^{-1}S_n, \hat{\theta}_n - (nB)^{-1}S_n] = 0$.

That's All Folks !