1 Probability theory

1.1 Probability spaces and random elements

1.1.1 σ-fields and measures

- σ-fields
  - σ-field (Definition 1.1 in Section 1.1.1)
  - $\sigma(\mathcal{C})$: the smallest σ-field containing $\mathcal{C}$
  - Borel σ-field
- Measure related definitions
  - Measurable space
  - Measure (Definition 1.2 in Section 1.1.1)
  - Example. Counting measure
- Uniqueness
  - **Theorem 10.3** (Billingsley 1986) Suppose that μ_1 and μ_2 are measures on $\sigma(\mathcal{P})$, where $\mathcal{P}$ is a π-system, and suppose they are σ-finite on $\mathcal{P}$. If μ_1 and μ_2 agree on $\mathcal{P}$, then they agree on $\sigma(\mathcal{P})$.
  - **Definition.** Suppose that $\mathcal{C}$ is a collection of some subsets of Ω. $\mathcal{C}$ is π-system if it is closed under finite intersections.
  - **Definition.** Suppose that $\mathcal{C}$ is a collection of some subsets of Ω. A measure μ is σ-finite on $\mathcal{C}$ if there exists $\{A_k\}$: a sequence of sets in $\mathcal{C}$ such that

    $$\Omega = \cup_k A_k \text{ and } \mu(A_k) < \infty \text{ for all } k.$$

  - Example. Lebesgue measure
  - Product measure
- Properties (Proposition 1.1 in Section 1.1.1)
  - Monotonicity
  - Subadditivity
  - Continuity

1.1.2 Measurable functions and distributions

- Definition of a measurable function (Definition 1.3 in Section 1.1.2).
- Examples of measurable functions
  - Indicator functions
  - Operations applied to Borel functions that give Borel functions (Proposition 1.4 in Section 1.1.2): arithmetic, sup, inf, liminf, limsup.
  - Suppose that $f_1, \ldots, f_k$ are measurable from $(\Omega, \mathcal{F})$ to $(R, \mathcal{B})$, let $f = (f_1, \ldots, f_k)$, then f is measurable from $(\Omega, \mathcal{F})$ to $(R, \mathcal{B}(R^k))$.
  - Simple functions
  - Continuous functions (Proposition 1.4 (v))
  - Composition of measurable functions (Proposition 1.4 (iv))
- Approximation property. Suppose that f is measurable from $(\Omega, \mathcal{F})$ to $(\overline{R}, \overline{\mathcal{B}})$, where $\overline{R} = R \cup \{\infty, -\infty\}$ and $\overline{\mathcal{B}} = \sigma(\mathcal{B} \cup \{\{\infty\}, \{-\infty\}\})$.
  - Suppose that $f \geq 0$. Then there exists $\{f_n\}$: a sequence of real-valued simple functions such that $0 \leq f_n \leq f_{n+1} < \infty$ and $\lim_{n\to\infty} f_n = f$.
  - Let
$$f^+(w) = \begin{cases} f(w) & \text{if } f(w) > 0; \\ 0 & \text{otherwise,} \end{cases}$$
and
$$f^-(w) = \begin{cases} -f(w) & \text{if } f(w) < 0; \\ 0 & \text{otherwise.} \end{cases}$$
Then $f = f^+ - f^-$ and f^+ and f^- are measurable from $(\Omega, \mathcal{F})$ to $(\overline{R}, \overline{\mathcal{B}})$.
- σ-field induced by a function. Suppose that f is a function from Ω to Λ. Suppose that $\mathcal{G}$ is a σ-field on Λ. Then $f^{-1}(\mathcal{G}) = \{f^{-1}(A) : A \in \mathcal{G}\}$ is called the σ-field induced by f. When it is clear what $\mathcal{G}$ is, $f^{-1}(\mathcal{G})$ is often denoted by $\sigma(f)$.
  - $\sigma(f)$ is the smallest σ-field that makes f measurable.

- Example 1. $\Omega = \{1, 2, 3, 4\}$. $Y(1) = 4$, $Y(2) = 5$, $Y(3) = Y(4) = 6$. Take $\mathcal{F}$ to be the smallest σ-field on Ω such that Y is measurable from $(\Omega, \mathcal{F})$ to $(R, \mathcal{B})$. Then $\mathcal{F} = \sigma(\{\{1\}, \{2\}, \{3, 4\}\})$. $\mathcal{F}$ is denoted by $\sigma(Y)$.

- Suppose that f is measurable from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$. Suppose that $\mathcal{G}$ contains all singletons in Λ. Then the value of f is determined if, for each event $A \in \mathcal{F}$, whether A occurs or not is determined. (See Example 1)

- **Lemma.** (Theorem A.42 in "Theory of Statistics" by Schervish (1995); Modified version of Lemma 1.2 (Theorem 1.6 in 1st Ed) in Section 1.4.1) Suppose that Y is measurable from $(\Omega, \mathcal{F})$ to $(\Lambda_Y, \mathcal{G}_Y)$ and Z is measurable from $(\Omega, \mathcal{F})$ to $(\Lambda_Z, \mathcal{G}_Z)$. Suppose that $\mathcal{G}_Z$ contains all singletons in Λ_Z. Let T be the range of Y and $T \cap \mathcal{G}_Y$ be the σ-field on T defined by $\{T \cap A : A \in \mathcal{G}_Y\}$. Then Z is measurable from $(\Omega, \sigma(Y))$ to $(\Lambda_Z, \mathcal{G}_Z)$ if and only if $Z = h \circ Y$ for some h that is measurable from $(T, T \cap \mathcal{G}_Y)$ to $(\Lambda_Z, \mathcal{G}_Z)$. (See Example 1)

- Proof of Lemma (the "only if" direction).

  Suppose that Z is measurable from $(\Omega, \sigma(Y))$ to $(\Lambda_Z, \mathcal{G}_Z)$. The existence of h can be established by noting that

  (*) for w_1, $w_2 \in \Omega$, $Y(w_1) = Y(w_2)$ implies that $Z(w_1) = Z(w_2)$.

  To see that (*) holds, suppose that $Y(w_1) = Y(w_2) = a$. Since Z is measurable (wrt $\sigma(Y)$), there exists $A \in \mathcal{G}_Y$ such that $Y^{-1}(A) = Z^{-1}(\{Z(w_1)\})$. Since $w_1 \in Z^{-1}(\{Z(w_1)\})$, we have $w_1 \in Y^{-1}(A)$ and $a \in A$, which gives $w_2 \in Y^{-1}(A)$ and $w_2 \in Z^{-1}(\{Z(w_1)\})$, so $Z(w_2) = Z(w_1)$.

  (*) implies that there exists a function h so that $Z(w) = h(Y(w))$ for $w \in \Omega$, where the domain of h is T: the range of Y. To prove the measurablity of h wrt $T \cap \mathcal{G}_Y$, for $B \in \mathcal{G}_Z$, let A be an event in $\mathcal{G}_Y$ such that $Y^{-1}(A) = Z^{-1}(B)$ (the existence of A is garanteed by the measurablity of Z wrt $\sigma(Y)$), then h is measurable if $h^{-1}(B) = A \cap T$, where T is the range of Y. Below is the proof for $h^{-1}(B) = A \cap T$.

  - $h^{-1}(B) \subset A \cap T$. Suppose that $a \in h^{-1}(B)$, then $h(a) \in B$ and $a = Y(w)$ for some $w \in \Omega$, so $Z(w) = h(Y(w)) \in B$ and $w \in Z^{-1}(B) = Y^{-1}(A)$, which gives $a = Y(w) \in A$.

- $A \cap T \subset h^{-1}(B)$. Suppose that $a \in A \cap T$. Then $a = Y(w) \in A$ for some $w \in \Omega$, so $w \in Y^{-1}(A) = Z^{-1}(B)$ and $h(a) = Z(w) \in B$, which gives $a \in h^{-1}(B)$.

- Random variables/vectors and induced measures.

  - Definition. X is a random vector on a probability space $(\Omega, \mathcal{F}, P)$ means X is measurable from $(\Omega, \mathcal{F})$ to $(R^k, \mathcal{B}^k)$, where $\mathcal{B}^k$ denotes the Borel σ-field on R^k. When $k = 1$, X is called a random variable.

  - Definition. Suppose that f is measurable from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$. and ν is a measure on $(\Omega, \mathcal{F})$. Then the measure on $(\Lambda, \mathcal{G})$ induced by f, denoted by $\nu \circ f^{-1}$, is defined by

$$\nu \circ f^{-1}(A) = \nu(f^{-1}(A)) \text{ for } A \in \mathcal{G}.$$

  - Suppose X is a random varible on a probability space $(\Omega, \mathcal{F}, P)$. Then the induced measure $P \circ X^{-1}$ is called the distribution of X, which is often characterized by its cumulative distribution function (c.d.f.).

1.2 Integration and differentiation

1.2.1 Integration

- Definition of integration. (Definition 1.4 in Section 1.2.1)

  - Integrable functions.

  - Integration over a set.

  - Example 2. $\Omega = \{1, 2, 3, 4, 5, 6\}$. $X(\omega) = \omega$. ν: counting measure on $(\Omega, 2^\Omega)$. $P(A) = \nu(A)/\nu(\Omega)$. Find $\int X dP$.

- Basic properties of integration

  - Linearity (Proposition 1.5 in Section 1.2.1)

  - Monotonicity (Proposition 1.6(i) in Section 1.2.1)

  - If $f \geq 0$ ν-a.e. and $\int f d\nu = 0$, then $f = 0$ ν-a.e. (Proposition 1.6(ii) in Section 1.2.1).

  - $\nu(A) = 0$ implies that $\int_A f d\nu = 0$.

- Limits of integrals (Theorem 1.1 and Example 1.8 in Section 1.2.1)
  - Fatou's lemma
  - Dominated convergence theorem
  - Monotone convergence theorem
  - Interchange of differentiation and integration
- Change of variable (Theorem 1.2 in Section 1.2.1)
- Fubini's theorem (Theorem 1.3 in Section 1.2.1)
- Suppose that f is Riemann integrable on a finite interval I with endpoints a and b, where $a < b$. Let λ be the Lebesgue measure on (R, B). Then $\int_I f d\lambda = \int_a^b f(x)dx$.
- Suppose that Ω is a countable set and ν is a measure on $(\Omega, 2^\Omega)$. Then for a nonnegative f that is measurable from $(\Omega, 2^\Omega)$ to $(R, \mathcal{B})$,

$$\int f d\nu = \sum_{\omega \in \Omega} f(\omega)\nu(\{\omega\}).$$

1.2.2 Radon-Nikodym derivative

- Absolute continuity (Equation (1.19) in Section 1.2.2)
- Radon-Nikodym Theorem (Theorem 1.4 in Section 1.2.2). Note. Measures are assumed to be σ-finite.

  Example 3. Suppose that $\Omega = \{1, 2, 3\}$ and P and ν are measures on $(\Omega, 2^\Omega)$ so that $P(\{k\}) = k/6$ and $\nu(\{k\}) = 1$ for $k \in \Omega$. Show that P is absolute continuous with respect to ν and find $dP/d\nu$.

  Example 4. Suppose that F is the c.d.f. of a random variable X and F is continuously differentiable. Let λ be the Lebesgue measure on $(R, \mathcal{B})$. Then $F' = dP \circ X^{-1}/d\lambda$.

- Suppose that X is a random variable and ν is a measure on $(R, \mathcal{B})$. If $P \circ X^{-1}$ is absolute continuous with respect to ν, then $dP \circ X^{-1}/d\nu$ is called the p.d.f. of X with respect to ν.

Example 5. A standard normal random variable has a Lebesgue p.d.f. f of the form

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \qquad x \in R.$$

Example 6. Suppose that Z is a standard normal random variable and $X = ZI_{[1,\infty)}(Z)$. Let μ be the probability measure on $(R, \mathcal{B})$ such that $\mu(A) = I_A(0)$ for $A \in \mathcal{B}$ and λ be the Lebesgue measure on $(R, \mathcal{B})$. Show that X has a p.d.f. with respect to $\mu + \lambda$.

- Integration using Radon-Nikodym derivative (Proposition 1.7(i) in Section 1.2.2)

Example 7. Suppose that X is a standard normal random variable on a probability space $(\Omega, \mathcal{F}, P)$. Find $\int X dP$.

1.3 Distributions and their characteristics

- Ways of characterizing a distribution: p.d.f., c.d.f, characteristic function and moment generating function.
- Find the p.d.f. of a transformed random variable: Proposition 1.8 in Section 1.3.1.

Example 8. Suppose that X is a random variable with Lebesgue p.d.f. f_X and $f_X(x) = 0$ for $x \leq 0$. Let $Y = X^2$ and

$$g(y) = \frac{f_X(\sqrt{y})}{2\sqrt{y}} I_{(0,\infty)}(y).$$

Then g is a Lebesgue p.d.f. of Y.

Proof. Let λ be the Lebesgue measures on $(R, \mathcal{B}(R))$ and let $\lambda^+(A) = \int_A I_{(0,\infty)}(x) d\lambda(x)$ for $A \in \mathcal{B}(R)$. Let $h(y) = \sqrt{y} I_{(0,\infty)}(y)$ for $y \in (-\infty, \infty)$. Then for $0 < b < \infty$,

$$\begin{aligned} \int_{(-\infty,b]} g(y) d\lambda(y) &= \int I_{(\sqrt{0},\sqrt{b}]}(\sqrt{y}) \frac{f_X(\sqrt{y})}{2\sqrt{y}} d\lambda^+(y) \\ &= \int I_{(0,\sqrt{b}]}(x) \frac{f_X(x)}{2x} d\lambda^+ \circ h^{-1}(x). \qquad (1) \end{aligned}$$

Note that

$$\frac{d\lambda^+ \circ h^{-1}}{d\lambda}(x) = 2xI_{(0,\infty)}(x) \ (\lambda\text{-a.e.}) \tag{2}$$

since for $0 < b < \infty$,

$$\lambda^+ \circ h^{-1}((-\infty, b]) = \lambda^+((0, b^2]) = b^2 = \int_{(-\infty,b]} 2xI_{(0,\infty)}(x)d\lambda(x),$$

and for $b \leq 0$,

$$\lambda^+ \circ h^{-1}((-\infty, b]) = 0 = \int_{(-\infty,b]} 2xI_{(0,\infty)}(x)d\lambda(x).$$

From (1) and (2), we have

$$\begin{aligned}
\int_{(0,b]} g(y)d\lambda(y) &= \int I_{(0,\sqrt{b}]}(x)f_X(x)I_{(0,\infty)}(x)d\lambda(x) \\
&= \int I_{(-\infty,\sqrt{b}]}(x)f_X(x)d\lambda(x) \\
&= P(X \in (-\infty, \sqrt{b}]) = P(Y \in (-\infty, b])
\end{aligned}$$

for $0 < b < \infty$. Moreover, $\int_{(-\infty,b]} g(y)d\lambda(y) = 0 = P(Y \in (-\infty, b])$ for $b \leq 0$, so

$$\int_A g(y)d\lambda(y) = P(Y \in A)$$

for every $A \in \mathcal{B}(R)$ and g is the Lebesgue density of Y.

1.4 Conditional expectations

1.4.1 Conditional expectations

- Definitions of $E(X|\mathcal{A})$, $P(B|\mathcal{A})$ and $E(X|Y)$ (Definition 1.6 in Section 1.4.1).
  - Existence and uniqueness.

Example 9. Suppose that $\Omega = \{1, 2, 3, 4\}$. P is a measure on $(\Omega, 2^\Omega)$ such that $P(\{k\}) = 1/4$ for $k \in \Omega$. Suppose that $X(k) = k$ for $k \in \Omega$ and $Y(1) = 4$, $Y(2) = 5$, $Y(3) = Y(4) = 6$. Find $E(X|\sigma(Y))$.

- Some facts following from the definition.

  - Suppose that X is measurable from $(\Omega, \mathcal{A}_0)$ to $(R, \mathcal{B})$, where $\mathcal{A}_0$ is a sub-σ-field of $\mathcal{A}$. Then $E(X|\mathcal{A}) = X$.
  - If $\mathcal{A} = \{\emptyset, \Omega\}$, then $E(X|\mathcal{A}) = E(X)$.

- Properties of conditional expecations (Proposition 1.10 or Proposition 1.12 in the first edition).

  - Suppose that X and Y are random vectors on $(\Omega, \mathcal{F}, P)$ and X and Y take values in R^m and R^n respectively. X and Y are independent if and only if $P((X, Y) \in A \times B)$ for all $A \in \mathcal{B}(R^m)$, $B \in \mathcal{B}(R^n)$.

- $E(X|\mathcal{A})$ is the "best guess" of X given the knowledge of occurrences of events in $\mathcal{A}$ in the following sense

$$\int (X - E(X|\mathcal{A}))^2 dP \leq \int (X - Y)^2 dP \tag{3}$$

  for all Y: measurable from $(\Omega, \mathcal{A})$ to $(R, \mathcal{B})$.

1.4.2 Independence

- Definition of independence (Definition 1.7)

- Conditional expectations and independence.

  Fact 1 *Suppose that X is a random variable on $(\Omega, \mathcal{F}, P)$ with $E|X| < \infty$, and $\mathcal{A}_1$ and $\mathcal{A}_2$ are sub-σ-fields of $\mathcal{F}$. If $\sigma(\sigma(X) \cup \mathcal{A}_1)$ and $\mathcal{A}_2$ are independent, then*

$$E(X|\sigma(\mathcal{A}_1 \cup \mathcal{A}_2)) = E(X|\mathcal{A}_1) \text{ a.s.}$$

  The proof of Fact 1 is based on the result that

$$\int_{A_1 \cap A_2} E(X|\mathcal{A}_1) dP = \int_{A_1 \cap A_2} X dP \text{ for } A_1 \in \mathcal{A}_1 \text{ and } A_2 \in \mathcal{A}_2,$$

  which can be established from the following fact:

  Fact 2 *Suppose that X is a nonnegative random variable on $(\Omega, \mathcal{F}, P)$ and $\mathcal{A}_2$ is a sub-σ-field of $\mathcal{F}$. If $\mathcal{A}_2$ is independent of $\sigma(X)$, then for $A_2 \in \mathcal{A}_2$,*

$$E(X I_{A_2}) = P(A_2) E(X).$$

The proof of Fact 2 is based on Proposition 1.10 (vii).

Special cases of Fact 1:

- Proposition 1.11 in Section 1.4.2 (Proposition 1.14 in the first edition). Note: $\sigma((Y_1, Y_2)) = \sigma(\sigma(Y_1) \cup \sigma(Y_2))$.
- Suppose that X is a random variable on $(\Omega, \mathcal{F}, P)$ with $E|X| < \infty$ and Y is a measurable function from $(\Omega, \mathcal{F})$ to a measurable space. Suppose that $\sigma(X)$ and $\sigma(Y)$ are independent. Then

$$E(X|Y) = E(X) \text{ a.s.}$$

1.4.3 Conditional distributions

- Definition of a random measure on $(R^n, \mathcal{B}^n)$. Suppose that $(\Omega, \mathcal{F}, P)$ is a probability space and μ is a function on $\mathcal{B}^n \times \Omega$ satisfying (i) and (ii):

  (i) For every $\omega \in \Omega$, $\mu(\cdot, \omega)$ is a measure on $(R^n, \mathcal{B}^n)$.

  (ii) For every $B \in \mathcal{B}^n$, let $X(\omega) = \mu(B, \omega)$. Then X is measurable from $(\Omega, \mathcal{F})$ to $(\overline{R}, \overline{\mathcal{B}})$.

  Then μ is a random measure on $(R^n, \mathcal{B}^n)$ with respect to the probability space $(\Omega, \mathcal{F}, P)$. Here $\overline{R} = R \cup \{\infty\} \cup \{-\infty\}$ and $\overline{\mathcal{B}} = \sigma(\mathcal{B} \cup \{\{\infty\}, \{-\infty\}\})$. If for every $\omega \in \Omega$, $\mu(\cdot, \omega)$ is a probability measure on $(R^n, \mathcal{B}^n)$, then μ is a random probability measure on $(R^n, \mathcal{B}^n)$.

- Existence of conditional distributions (Theorem 1.7 (i); Theorem 1.7 in the first edition). Suppose that X and Y are random vectors on $(\Omega, \mathcal{F}, P)$ and take values in R^n and R^m respectively. Let $P_Y = P \circ Y^{-1}$ be the distribution of Y. Then there exists a random probability measure μ on $(R^n, \mathcal{B}^n)$ with respect to the probability space $(R^m, \mathcal{B}^m, P_Y)$ such that

$$P((X, Y) \in B \times C) = \int_C \mu(B, y) dP_Y(y) \text{ for all } B \in \mathcal{B}^n, C \in \mathcal{B}^m. \quad (4)$$

  For y in the range of Y, $\mu(\cdot, y)$ is called a version of the conditional distribution of X given $Y = y$, denoted by $P_{X|Y}(\cdot|y)$ or $P_{X|Y=y}$.

- Conditional expectation = expectation with respect to conditional distribution. Suppose that g is measurable from $(R^n \times R^m, \mathcal{B}^{n+m})$ to $(\overline{R}, \overline{\mathcal{B}})$. Suppose that g is nonnegative. Let $h(y) = \int g(x,y) dP_{X|Y}(x|y)$. Then $E(g(X,Y)|Y) = h(Y)$. Note that if $E(|g(X,Y)|) < \infty$, then h can be defined P_Y-a.e and we still have $E(g(X,Y)|Y) = h(Y)$.

- (4) can be used to construct the joint distribution of X and Y.

- Conditional p.d.f.s. Suppose that X and Y are random vectors on $(\Omega, \mathcal{F}, P)$ and take values in R^n and R^m respectively. Let P_Y denote the distribution of Y. Suppose that for every y in the range of Y, μ_y is a measure on $(R^n, \mathcal{B}^n)$ and h is a function such that $\mu(\cdot, y)$ can serve as a version of the conditional distribution of $P_{X|Y=y}$, where $\mu(B, y) = \int_B h(x,y) d\mu_y(x)$. That is, for each y, $h(x,y) \geq 0$,

  $$\int_{R^n} h(x,y) d\mu_y(x) = 1 \text{ for every } y,$$

  for each $B \in \mathcal{B}^n$, let

  $$h_1(y) = \int_B h(x,y) d\mu_y(x),$$

  then h_1 is measurable from (R^m, B^m) to $(\overline{R}, \overline{\mathcal{B}})$, and

  $$P((X,Y) \in B \times C) = \int_C \int_B h(x,y) d\mu_y(x) dP_Y(y) \text{ for all } B \in \mathcal{B}^n, C \in \mathcal{B}^m.$$

  Then $h(\cdot, y)$ is a p.d.f. for the conditional distribution $P_{X|Y}(\cdot|y)$ with respect to the measure μ_y. Such an $h(\cdot, y)$ is called a conditional p.d.f. of X given $Y = y$ respect to the measure μ_y and is denoted by $f_{X|Y}(\cdot|y)$ or $f_{X|Y=y}$.

- Finding conditional p.d.f.s using joint p.d.f.s

  - Joint p.d.f is with respect to a product measure (Proposition 1.9 in Section 1.4.1 or Proposition 1.11 in the first edition). Suppose that (X,Y) has a p.d.f. $f_{X,Y}$ with respect to a product measure $\mu \times \nu$. Let $f_Y(y) = \int f_{X,Y}(x,y) d\mu(x)$, then f_Y is the p.d.f of Y with respect to ν and $f_{X,Y}(\cdot, y)/f_Y(y)$ is the conditional p.d.f. of X given $Y = y$ with respect to μ.

– Joint p.d.f. is with respect to the distribution of another pair of random vectors.

Fact 3 *Suppose that (X, Y)'s distribution has a p.d.f. $f_{X,Y}$ with respect to (X_0, Y_0)'s distribution. Let*

$$f_Y(y) = \int f_{X,Y}(x, y) dP_{X_0|Y_0=y}(x),$$

then f_Y is a p.d.f. of Y with respect to the distribution of Y_0 and $f_{X,Y}(\cdot, y)/f_Y(y)$ is the conditional p.d.f. of X given $Y = y$ with respect to $P_{X_0|Y_0=y}$.

Example 10. Suppose that $X_1, \ldots, X_n$ are IID and $X_i \sim N(\mu, 1)$, and $Y_1, \ldots, Y_n$ are IID and $Y_i \sim N(0, 1)$. Let $X = (X_1, \ldots, X_n)$, $\bar{X} = \sum_{i=1}^n X_i/n$, $Y = (Y_1, \ldots, Y_n)$, and $\bar{Y} = \sum_{i=1}^n Y_i/n$. Show that the distribution of $(Y, \bar{Y})$ has a PDF with respect to $P_{Y,\bar{Y}}$.

1.5 Asymptotic theory

1.5.1 Convergence modes and stochastic orders

- Convergence modes (Definition 1.8)
  - Almost everywhere convergence.
  - Convergence in probability.
  - L_r convergence.
  - Convergence in distribution.
- Relation among different convergence modes. (Theorem 1.8)
  - Almost surely convergence or L_r convergence implies convergence in probability, which implies convergence in distribution.
  - Convergence in distribution to a constant implies convergence in probability.
  - Skorohod's theorem. X_n converges to X in distribution implies that there exist $\{Y_n\}$ and Y such that
    1. X_n and Y_n have the same distribution,
    2. Y and X have the same distribution, and

3. Y_n converges to Y almost surely.

- X_n converges to X in probability if and only if every subsequence of $\{X_n\}$ has a subsequence that converges to X almost surely.
- Borel Cantelli lemmas. (Lemma 1.5) Let $\limsup_n A_n = \cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k$.
  - First Borel Cantelli lemma. If $\sum_n P(A_n) < \infty$, then $P(\limsup_n A_n) = 0$. Special case: Theorem 1.8 (v)
  - Second Borel Cantelli lemma. If A_n's are independent and $\sum_n P(A_n) = \infty$, then $P(\limsup_n A_n) = 1$.
- Remark. The event $\limsup_n A_n$ occurs means A_n occurs infinitely often, so we also denote $\limsup_n A_n$ as A_n i.o. (infinitely often).
  - Fact. Suppose that $P(|X_n - X| > \varepsilon \text{i.o.}) = 0$ for every $\varepsilon > 0$, then $X_n \to X$ a.s. as $n \to \infty$.
- Stochastic orders
  - Suppose that $\{a_n\}$ and $\{b_n\}$ are sequences in R and $b_n \neq 0$. Then $a_n = o(b_n)$ means $a_n/b_n \to 0$ and $a_n = O(b_n)$ means $\{a_n/b_n\}$ is a bounded sequence.
  - $X_n = O_p(Y_n)$ (Definition 1.9 (iii)).
  - $X_n = o_p(Y_n)$ (Definition 1.9 (iv)).
  - $X_n = O_p(1)$ means that for every $\varepsilon > 0$, there exists $C > 0$ such that $\sup_n P(|X_n| \geq C) < \varepsilon$, where $|\cdot|$ denotes the Euclidean norm.
  - $X_n = o_p(1)$ means that $|X_n|$ converges to 0 in probability.

1.5.2 Weak convergence

- Checking for convergence in distribution. Theorem 1.9.

1.5.3 Convergence of transformations

- Continuous mapping theorem (Theorem 1.10)
- Slutsky's theorem (Theorem 1.11)
- Generalized delta method (Theorem 1.12)

Example 11. Suppose that $\sqrt{n}(X_n - 2)$ converges to $N(0,1)$ in distribution. Then $n(X_n-2)^2$ converges to $\chi^2(1)$ in distribution and $\sqrt{n}(X_n^2 - 4)$ converges to $N(0,16)$ in distribution.

1.5.4 The Law of large number

- Convergence of the average of independent random variables. (Theorem 1.14)

1.5.5 The center limit theorem

- Lindeberg's CLT (Theorem 1.15). Suppose that $\{k_n\}$ is a sequence of positive integers and for each n, $X_{n,1}, \ldots, X_{n,k_n}$ are independent random variables. Let $\sigma_n = \sqrt{Var(\sum_{j=1}^{k_n} X_{n,j})}$ and suppose that $0 < \sigma_n^2 < \infty$. Then under the Lindeberg's condition, $\sigma_n^{-1} \sum_{j=1}^{k_n}(X_{n,j} - EX_{n,j})$ converges to $N(0,1)$ in distribution.

- Lindeberg's condition.

$$\lim_{n\to\infty} \sigma_n^{-2} \sum_{j=1}^{k_n} E(X_{n,j} - EX_{n,j})^2 I_{\{|X_{n,j}-EX_{n,j}|>\varepsilon\sigma_n\}} = 0 \text{ for every } \varepsilon > 0.$$

- Remarks on Lindeberg's CLT.
  - Proof of Lindeberg's CLT is based on approximating the characteristic function of $\sigma_n^{-1} \sum_{j=1}^{k_n}(X_{n,j} - EX_{n,j})$ and the following inequalities

$$\left|\prod_{k=1}^{m} b_k - \prod_{k=1}^{m} a_k\right| \le \sum_{k=1}^{m} |b_k - a_k| \text{ if } |a_k| \le 1 \text{ and } |b_k| \le 1 \text{ for all } k,$$

$$\left|e^{i\theta} - \left(1 + i\theta - \frac{\theta^2}{2}\right)\right| \le \min(\theta^2, |\theta|^3) \text{ for } \theta \in (-\infty, \infty),$$

    and

$$|E(Z)| \le E|Z|.$$

  - Lindeberg's condition is implied by the Liapounov's condition: $\sum_{j=1}^{k_n} E|X_{n,j} - EX_{n,j}|^{2+\delta} = o(\sigma_n^{2+\delta})$ for some $\delta > 0$.

– Lindeberg's condition implies Feller's condition:

$$\frac{\max_{1\le j\le k_n} Var(X_{n,j})}{\sigma_n^2} \to 0 \text{ as } n \to \infty.$$

– It seems that the condition $k_n \to \infty$ in the text is not used in the proof, but the condition follows from the Feller's condition.

– The usual CLT is implied by Lindeberg's CLT.

Example 12. Suppose that $\varepsilon_1, \ldots, \varepsilon_n, \ldots$ are IID random variables with mean 0 and variance σ^2 and $\{w_i\}_{i=1}^{\infty}$ is a sequence of real numbers. Let $S_n = \sum_{i=1}^n w_i\varepsilon_i$ and $\sigma_n = \sigma\sqrt{\sum_{i=1}^n w_i^2}$. If $\lim_{n\to\infty} \frac{\max_{1\le i\le n} |w_i|}{\sigma_n} = 0$, then S_n/σ_n converges to $N(0,1)$ in distribution.

- Multivariate CLT (Corollary 1.2).