STAT 560/561 Lecture Notes

Statistical Inference

University of British Columbia

Jiajun Zhang
Department of Statistics
September 2026

Contents

1 Review of Probability

1.1 Probability Measures

We denote $(\Omega, \mathcal{F}, \mathbb{P})$ as the **probability space**, where

- Ω is the **sample space**, a set containing all possible outcome of a certain event;
- $\mathcal{F}$ is the **σ-algebra** on Ω. In particular, it satisfies the following:
  - $\Omega \in \mathcal{F}$;
  - If $A \in \mathcal{F}$, then $A^C = \Omega \setminus A \in \mathcal{F}$;
  - If $\{A_n\}_{n=1}^\infty \subset \mathcal{F}$, then $\bigcup_{n=1}^\infty A_n \in \mathcal{F}$.

  We denote $\sigma(\Omega)$ as the σ-algebra generated by Ω, which is the smallest σ-algebra that contains Ω. Later, when the sample space is $\mathbb{R}$, we will denote $\mathfrak{B}(\mathbb{R})$ as the **Borel σ-algebra** on $\mathbb{R}$, which is generated by the collection of open sets in $\mathbb{R}$.

  In the event of a coin-tossing, if we denote H to be the head, and T to be the tail, then the sample space is $\Omega = \{H, T\}$, and the σ-algebra on Ω is $\mathcal{F} := \{\varnothing, \{H, T\}, \{H\}, \{T\}\}$.
- $\mathbb{P}$ is the **probability measure** on Ω, which is a mapping $\mathbb{P} : \mathcal{F} \to [0, 1]$ such that it satisfies *Kolmogorov's axioms*:
  - $\mathbb{P}(\Omega) = 1$;
  - $\mathbb{P}(A) \geq 0, \forall A \in \mathcal{F}$;
  - If $A_i \cap A_j = \varnothing, \forall i \neq j$, then $\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$.

  Based on Kolmogorov's axioms more properties of probability measure follow:
  - $\mathbb{P}(A^C) = \mathbb{P}(\Omega \setminus A) = 1 - \mathbb{P}(A)$;
  - If $A \cap B = \varnothing$, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$;
  - If $A \subset B$, then $\mathbb{P}(A) \subset \mathbb{P}(B)$;
  - **The principle of inclusion and exclusion:** Let $A_1, A_2, \cdots, A_n \in \mathcal{F}$, then

  $$\mathbb{P}\left(\bigcup_{k=1}^{n} A_k\right) = \sum_{k=1}^{n} \mathbb{P}(A_k) - \sum_{k_1 < k_2} \mathbb{P}(A_{k_1} \cap A_{k_2}) + \cdots + (-1)^{n+1} \sum_{k_1 < \cdots < k_n} \mathbb{P}(A_{k_1} \cap \cdots \cap A_{k_n}).$$

  In the special case of $n = 2, 3$, we have the following: $\forall A, B, C \in \Omega$,

  $$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B);$$

  $$\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) + \mathbb{P}(A \cap B \cap C).$$

Theorem 1 (**Continuity of $\mathbb{P}$**). Let $\{A_n\}_{n=1}^\infty$ be a sequence of non-decreasing events in $\mathcal{F}$, i.e $A_n \subset A_{n+1}$, then

$$\mathbb{P}(\lim_{n\to\infty} A_n) = \lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right).$$

Proof. Let $B_1 = A_1, B_2 = A_2 \setminus A_1, \cdots, B_n = A_n \setminus \bigcup_{i=1}^{n-1} A_i$. Then $B_i \cap B_j = \varnothing, \forall i \neq j$, and $A_n = \bigcup_{i=1}^{n} B_i$. Then by countable additivity of $\mathbb{P}$, we have

$$\mathbb{P}(A_n) = \mathbb{P}\left(\bigcup_{i=1}^{n} B_i\right) = \sum_{i=1}^{n} \mathbb{P}(B_i).$$

By taking $n \to \infty$, we have

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}(B_i) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} B_n\right) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \mathbb{P}(\lim_{n\to\infty} A_n).$$

□

Theorem 2 (**Continuity of $\mathbb{P}$, cont'd**)**.** Let $\{A_n\}_{n=1}^{\infty}$ be a sequence of non-increasing events in $\mathcal{F}$, i.e $A_{n+1} \subset A_n$, then

$$\mathbb{P}(\lim_{n\to\infty} A_n) = \lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n\right).$$

Proof. Consider $\{A_n^C\}_{n=1}^{\infty}$ which is non-decreasing. □

We next review conditional probability.

Definition 1. Suppose $A, H \in \Omega$, then we define $\mathbb{P}(A|H)$ to be the probability of A *given that* H occurs, known as the **conditional probability** of A given H.

The conditional probability of A given H is given by

$$\mathbb{P}(A|H) = \frac{\mathbb{P}(A \cap H)}{\mathbb{P}(H)}.$$

We say A, H are **independent** if $\mathbb{P}(A \cap H) = \mathbb{P}(A) \cap \mathbb{P}(H)$. Using the above formula, we have that $\mathbb{P}(A|H) = \mathbb{P}(A)$, i.e, the occurrence of H does not influence the occurrence of A.

Note that conditional probability on H, $\mathbb{P}(\cdot|H)$ is also a probability measure on $(H, \sigma(H))$, i.e, it satisfies Kolmogorov's axioms:

- $\mathbb{P}(H|H) = 1$;
- $\mathbb{P}(B|H) \geq 0, \forall B \in \sigma(H)$;
- $\mathbb{P}(B|H) = 1 - \mathbb{P}(B^C|H)$, where $B^C = H \setminus B$;
- If $H_1, H_2, \cdots$ are disjoint, then $\mathbb{P}\left(\bigcup_{n=1}^{\infty} H_n \middle| B\right) = \sum_{n=1}^{\infty} \mathbb{P}(H_n|B)$.

A sequence $\{A_n\}_{n=1}^{\infty}$ of events in $\mathcal{F}$ is called a **partition** of Ω if $A_i \cap A_j = \varnothing, i \neq j$ and $\bigcup_{n=1}^{\infty} A_n = \Omega$.

Theorem 3 (**Law of Total Probability**)**.** Let $\{A_n\}_{n=1}^{\infty}$ be a partition of Ω, then $\forall B \in \mathcal{F}$,

$$\mathbb{P}(B) = \sum_{n=1}^{\infty} \mathbb{P}(B|A_n)\mathbb{P}(A_n).$$

Proof. We know that $B = B \cap \Omega = B \cap \bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} B \cap A_n$, thus

$$\begin{aligned}\mathbb{P}(B) &= \mathbb{P}\left(\bigcup_{n=1}^{\infty} B \cap A_n\right) \\ &= \sum_{n=1}^{\infty} \mathbb{P}(B \cap A_n), \quad \textit{since } B \cap A_n\textit{'s are disjoint} \\ &= \sum_{n=1}^{\infty} \mathbb{P}(B|A_n)\mathbb{P}(A_n).\end{aligned}$$

□

Theorem 4 (**Bayes Theorem**)**.** Let $\{A_n\}_{n=1}^{\infty}$ be a partition of Ω, with $\mathbb{P}(A_n) > 0, \forall n$. Then $\forall B \in \mathcal{F}, \mathbb{P}(B) > 0$, and $\forall k \in \mathbb{N}$,

$$\mathbb{P}(A_k|B) = \frac{\mathbb{P}(A_k)P(B|A_k)}{\sum_{n=1}^{\infty} P(A_n)\mathbb{P}(B|A_n)}.$$

Proof. We know that $\mathbb{P}(A_k|B) = \frac{\mathbb{P}(A_k \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_k)\mathbb{P}(A_k)}{\mathbb{P}(B)}$, and use law of total probability to represent $\mathbb{P}(B)$. □

1.2 Random Variable and Distributions

Definition 2 (**Random variable**)**.** A random variable X on $(\Omega, \mathcal{F})$ is a mapping (measurable function) $X : \Omega \to \mathbb{R}$, such that for any Borel set $B \in \mathfrak{B}(\mathbb{R})$,

$$X^{-1}(B) := \{w \in \Omega : X(\omega) \in B\} \in \mathcal{F}.$$

Some analysis show that, the Borel σ-algebra is also generated by the intervals of the form $(a, b]$, hence if $\forall x \in \mathbb{R}$,

$$X^{-1}((-\infty, x]) = \{\omega \in \Omega : X(\omega) \in (-\infty, x]\} \in \mathcal{F},$$

then X is a random variable.

Example 1 (Coin Tossing)**.** Consider tossing a fair coin two times, denote H = head face up and T = tail face up, then we know that $\Omega = \{HH, HT, TH, TT\}$ and $\mathcal{F}$ is the σ-algebra generated by Ω. We define $X(\omega)$ = Number of H's in ω, then clearly $X(HH) = 2, X(HT) = X(TH) = 1, X(TT) = 0$, and further we have

$$X^{-1}((-\infty, x]) = \begin{cases} \varnothing : x < 0 \\ \{TT\} : 0 \le x < 1 \\ \{TT, HT, TH\} : 1 \le x < 2 \\ \Omega : 2 \le x \end{cases} \in \mathcal{F}$$

and thus X is a random variable.

We mainly focus on two types of random variables: **discrete random variables** and **continuous random variables**. For discrete random variables X, it takes on countably many values (for example, $\mathbb{N}$);

Definition 3 (**CDF**)**.** For a random variable X, the **cumulative distribution function (CDF)** of X, denoted as $F_X(\cdot)$, is defined by $F_X(x) = \mathbb{P}(X \leq x)$, $x \in \mathbb{R}$.

Based on the definition, we know the following about CDF:

- $F_X(\cdot)$ is non-negative and non-decreasing;
- $F_X(-\infty) = 0, F_X(\infty) = 1$;
- $F_X(\cdot)$ is right continuous.

Definition 4 (**pmf**)**.** For a discrete random variable X, the **probability mass function (pmf)** of X, denote as $f_X(\cdot)$, is defined by $f_X(X = x) = \mathbb{P}(X = x)$;

Definition 5 (**continuous random variable, pdf**)**.** A random variable X is said to be continuous if there exists a non-negative function $f_X(\cdot)$, such that $\int_{\mathcal{S}} f_X(x)dx = 1$ and $\mathbb{P}(X \in B) = \int_B f_X(x)dx$, where $\mathcal{S}$ is the support of X, $B \in \mathfrak{B}(\mathbb{R})$. Then $f_X(\cdot)$ is the **probability density function (pdf)** of X, and

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt,$$

when $F_X(\cdot)$ is differentiable at x, we have $F'_X(x) = f_X(x)$.

Example 2 (**examples of random variables**)**.** Here are some examples of commonly used random variables:

- **Discrete random variables:**
  - Bernoulli random variable $X \sim \text{Bern}(p)$: Where X has binary outcomes: $X = \{0, 1\}$, $\mathbb{P}(X = 1) = p, \mathbb{P}(X = 0) = 1 - p$. A single coin toss can be modeled by a Bernoulli random variable;
  - Binomial random variable $X \sim \text{Bino}(n, p)$: Where X can be viewed as a sum of n independent Bernoulli random variables $Y_i \sim \text{Ber}(p)$. Then we have

    $$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \cdots, n.$$

    If we toss a coin n times and record the number of heads, then it would follow a Binomial distribution.
  - Poisson random variable $X \sim \text{Pois}(\lambda)$, where X can be viewed as a "modified Binomial" distribution but n is large and p is small. The probability mass function is given by

    $$\mathbb{P}(X = k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}, \quad k = 0, 1, \cdots$$

- **Continuous random variables:**

– Normal random variable $X \sim N(\mu, \sigma^2)$: One of the most commonly used continuous random variables. Its probability density function is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right);$$

– Exponential random variable $X \sim \text{Exp}(\lambda)$: Its probability density function is given by

$$f_X(x) = \frac{1}{\lambda} e^{-\lambda x}.$$

Theorem 5 (Transformation of random variable). If X is a random variable, and $g(\cdot)$ is a Borel measurable function (i.e, $\forall x \in \mathbb{R}, g^{-1}((-\infty, x]) \in \mathfrak{B}(\mathbb{R})$), then $g(X)$ is again a random variable.

We may then study the transformations of random variables. For notation simplicity, we only focus on continuous random variables. Let X be a continuous random variable and $g(\cdot)$ be a measurable function, then the CDF of $g(X)$ can be obtained by

$$F_{g(X)}(x) = \mathbb{P}(g(X) \leq x) := \mathbb{P}(X \leq g^{-1}(x)) := \int_B f_X(t)dt,$$

where $B := \{t \in \mathbb{R}, t \leq g^{-1}(x)\}$. If differentiable, the cdf can be obtained by differentiating the CDF and apply the fundamental theorem of calculus.

1.3 Moments

Definition 6. Let X be a random variable, we define the **expectation** of X, denote as $\mathbb{E}[X]$, or μ_X, as follow:

- If X is a discrete random variable taking on countably many values from I, then

$$\mathbb{E}[X] = \sum_{i \in I} i \cdot \mathbb{P}(X = i).$$

- If X is a continuous random variable, denote its support by S, the pdf by $f_X(x)$, then

$$\mathbb{E}[X] = \int_S x f_X(x) dx.$$

Expectation satisfies *linearity*. That is, if $X_1, \cdots, X_n$ are random variables and $a_1, \cdots, a_n \in \mathbb{R}$, then

$$\mathbb{E}\left[\sum_{i=1}^n a_i X_i\right] = \sum_{i=1}^n a_i \mathbb{E}[X_i].$$

For a measurable function $g(\cdot)$, if $\mathbb{E}[g(X)]$ exists, then:

- If X is a discrete random variable taking on countably many values from I, then

$$\mathbb{E}[g(X)] = \sum_{i \in I} g(i) \cdot \mathbb{P}(X = i);$$

- If X is a continuous random variable with support S and pdf $f_X(x)$, then

$$\mathbb{E}[f(X)] = \int_S g(x) \cdot f_X(x) dx.$$

Definition 7. Let X be a random variable, the **kth central moment** of X, is given by $\mathbb{E}[(X - \mathbb{E}[X])^k]$. When $k = 2$, $\mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$ is also denoted as the **variance** of X, denote as $\mathbb{V}\mathbf{ar}(X)$.

Given $a, b \in \mathbb{R}$, we have that $\mathbb{V}\mathbf{ar}(aX + b) = a^2 \mathbb{V}\mathbf{ar}(X)$. We next discuss some inequalities related to moments.

Theorem 6 (**Markov's inequality**)**.** Let X be a non-negative random variable and $\mathbb{E}|X| < \infty$, then $\forall t > 0$,

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

Proof.

$$\mathbb{E}[X] := \int_{-\infty}^{t} x f_X(x) dx + \int_{t}^{\infty} x f_X(x) dx \geq \int_{t}^{\infty} t f_X(x) dx = t\mathbb{P}(X \geq t).$$

□

In general, if a measurable function $g(\cdot)$ is monotonically increasing and $g(t) \neq 0$, then

$$\mathbb{P}(X \geq t) = \mathbb{P}(g(X) \geq g(t)) \leq \frac{\mathbb{E}[g(X)]}{g(t)}.$$

Theorem 7 (**Chebyshev's inequality**)**.** Let X be a random variable, such that $\mu = \mathbb{E}[X], \mathbb{E}[X^2] < \infty$, then $\forall \epsilon > 0$,

$$\mathbb{P}(|X - \mu| \geq \epsilon) \leq \frac{\mathbb{V}\mathbf{ar}(X)}{\epsilon^2}.$$

Proof. Since $g(x) = x^2$ is monotonically increasing on $\mathbb{R}^+$, then

$$\mathbb{P}(|X - \mu| \geq \epsilon) = \mathbb{P}((X - \mu)^2 \geq \epsilon^2).$$

Then one can apply Markov's inequality and use the fact that $\mathbb{V}\mathbf{ar}(X) = \mathbb{E}[(X - \mu)^2]$. □

Definition 8 (**MGF**)**.** Let X be a random variable, the **moment generating function (mgf)** of X, denote as $M_X(t), t \in \mathbb{R}$ (if exists), is defined by $M_X(t) = \mathbb{E}[e^{tX}]$.

The reason why it is called moment generating function is because the moments of X ($\mathbb{E}[X^k]$), can be obtained by differentiating $M_X(t)$ with respect to t. Note that

$$\frac{d}{dt} M_X(t)\bigg|_{t=0} := \mathbb{E}[Xe^{tX}]\bigg|_{t=0} = \mathbb{E}[X],$$

$$\frac{d^2}{dt^2} M_X(t)\bigg|_{t=0} := \mathbb{E}[X^2 e^{tX}]\bigg|_{t=0} = \mathbb{E}[X^2].$$

Theorem 8 (**Chernoff's bound**)**.** Let X be a random variable, then for any $a \in \mathbb{R}$,

$$\mathbb{P}(X \geq a) \leq \inf_{t>0} e^{-at} M_X(t).$$

Proof. Since $g(x) = e^{xt}$ is increasing for $t > 0$, hence

$$\mathbb{P}(X \geq a) = \mathbb{P}(e^{tX} \geq e^{ta}) \leq \frac{\mathbb{E}[e^{tX}]}{e^{ta}} = e^{-ta} \cdot M_X(t), \forall t > 0.$$

□

We may also derive the Chernoff's bound for the other tail: Suppose $t < 0$, then $g(x) = e^{tx}$ is decreasing, hence

$$\mathbb{P}(X \leq a) = \mathbb{P}(e^{tX} \geq e^{ta}) \leq \frac{\mathbb{E}[e^{tX}]}{e^{ta}} = e^{-at} \cdot M_X(t), t < 0,$$

by taking the infinmum, we have

$$\mathbb{P}(X \leq a) \leq \inf_{t<0} e^{-at} M_X(t).$$

Note that MGF uniquely determines a random variable, however not all random variables have MGF. For example, the Cauchy random variable $X \sim \text{Cauchy}(0,1)$, with pdf

$$f_X(x) = \frac{1}{\pi(1+x^2)}, \quad x \in \mathbb{R}.$$

Then the MGF does not exist:

$$M_X(t) = \mathbb{E}[e^{tX}] = \int_{\mathbb{R}} \frac{e^{tx}}{\pi(1+x^2)} dx = \infty.$$

To avoid this issue, we sometime use the **characteristic function** of a random variable X defined by

$$\varphi_X(t) = \mathbb{E}[e^{iXt}], \quad t \in \mathbb{R}.$$

Characteristic function for a random variable X always exists, and it uniquely determines X. Again take $X \sim \text{Cauchy}(0,1)$, then for $t > 0$,

$$\varphi_X(t) = \frac{1}{\pi} \int_{\mathbb{R}} \frac{e^{itx}}{1+x^2} dx = \frac{1}{\pi} \lim_{R\to\infty} \int_{\gamma} \frac{e^{itz}}{1+z^2} dz - \int_0^{\pi} \frac{i\theta R e^{i\theta} e^{itRe^{i\theta}}}{1+R^2 e^{2i\theta}} d\theta = e^{-t},$$

where the first integral can be easily computed by residue theorem (γ is the upper circle with radius R).

Theorem 9 (**Jensen's inequality**)**.** Let X be a random variable, and $\mathbb{E}|X| < \infty$.

- If g is a convex function (i.e $g''(x) > 0$), then $g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)]$;
- If g is a concave function (i.e $g''(x) < 0$), then $g(\mathbb{E}[X]) \geq \mathbb{E}[g(X)]$.

Proof. Let $y = ax + b$ be the tangent line to $g(x)$ at $x_0 = \mathbb{E}[X]$. If g is convex, then the graph of $g(x)$ is below y, i.e., $g(x) \leq y$, hence

$$\mathbb{E}[g(X)] \leq \mathbb{E}[aX + b] = y(\mathbb{E}[X]) = g(\mathbb{E}[X]).$$

The argument for g is concave is similar. □

1.4 Multiple Random Variables

Definition 9 (Joint distribution). Let $X_1, \cdots, X_n$ be n random variables.

- $(X_1, \cdots, X_n)$ is said to be discrete if they take on countably many values. Then the cumulative distribution function of $(X_1, \cdots, X_n)$ is given by

$$F(x_1, \cdots, x_n) = \sum_{(a_1,\cdots,a_n)\in S} \mathbb{P}(X_1 = a_1, \cdots, X_n = a_n),$$

  where $B := \{(a_1, \cdots, a_n) : x_1 \leq a_1, \cdots, x_n \leq a_n\}$.

- $(X_1, \cdots, X_n)$ is said to be continuous if there exists a non-negative function $f(x_1, \cdots, x_n)$, such that $\int_S f(x_1, \cdots, x_n) d\boldsymbol{x} = 1$ and

$$\mathbb{P}(X_1 \leq x_1, \cdots, X_n \leq x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(t_1, \cdots, t_n) dt_n \cdots dt_1.$$

Similar to the definition we had before, random variables X, Y are **independent** if $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \in B)$ for any $A, B \in \mathfrak{B}(\mathbb{R})$. Under this case, if f_X is the density for X and f_Y is the density for Y, denote $f_{X,Y}$ as the joint density of X, Y, then $f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y), \forall x, y \in \mathbb{R}$.

If X, Y are not independent, we may define the **covariance** of X, Y, denote as $\mathbb{C}\mathbf{ov}(X, Y)$, as

$$\mathbb{C}\mathbf{ov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y],$$

where

$$\mathbb{E}[XY] = \iint_{\mathbb{R}^2} xy f_{X,Y}(x, y) dy dx.$$

Some properties of covariance follow:

- If X, Y are independent, then $\mathbb{C}\mathbf{ov}(X, Y) = 0$;
- $\mathbb{C}\mathbf{ov}(aX, Y) = \mathbb{C}\mathbf{ov}(X, aY) = a\mathbb{C}\mathbf{ov}(X, Y)$;
- $\mathbb{C}\mathbf{ov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i,j} \mathbb{C}\mathbf{ov}(X_i, Y_j)$;
- $\mathbb{V}\mathbf{ar}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbb{V}\mathbf{ar}(X_i) + 2\sum_{i\neq j} \mathbb{C}\mathbf{ov}(X_i, X_j)$.

The **correlation** of two random variables X, Y, denote as $\rho(x, y)$, is defined by

$$\rho(X, Y) = \frac{\mathbb{C}\mathbf{ov}(X, Y)}{\sqrt{\mathbb{V}\mathbf{ar}(X) \cdot \mathbb{V}\mathbf{ar}(Y)}},$$

it can be shown that $\rho(X, Y) \in [-1, 1]$.

FUN FACT: EXPECTATION IS AN INNER PRODUCT.

Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, define $L^2 := L^2(\Omega, \mathcal{F}, \mathbb{P})$ as the space of square-integrable random variables with

$$\mathbb{E}[X^2] = \int_\Omega x^2 f_X(x) dx,$$

we claim that L^2 defines a Hilbert space (a complete inner product space), with inner product defined by $\langle X, Y \rangle := \mathbb{E}[XY]$ and $\langle X, Y \rangle = 0$ if and only if X, Y are independent (or called orthogonal). The norm equipped on L^2 is simply defined by $||X||_{L^2} := \mathbb{E}[X^2]^{1/2}$.

By viewing expectation as a L^p space, we can then derive inequalities with respect to $\mathbb{E}|X|^p$ using L^p theory. Below, we assume X, Y come from the same probability space.

Theorem 10 (**Cauchy-Bunyakovsky-Schwarz inequality**)**.** Let X, Y be random variables, then

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

Theorem 11 (**Hölder's inequality**)**.** Let X, Y be random variables, $p, q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$. Then

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

Theorem 12 (**Minkouski's inequality**)**.** Let X, Y be random variables, then

$$\mathbb{E}|X+Y|^r \leq C_r(\mathbb{E}|X|^r + \mathbb{E}|Y|^r), \quad C_r = \begin{cases} 1 & r \in [0,1] \\ 2^{r-1} & r > 1 \end{cases}.$$

I will omit the proof for those three inequalities above.

For random variables $X_1, \cdots, X_n$ with joint distribution $f_{X_1, \cdots, X_n}$, it is also possible to find the distribution of $\boldsymbol{g}(\boldsymbol{X}) = (g_1(X_1), g_2(X_2), \cdots, g_n(X_n))$. This can be done by change of variable formula described below:

Theorem 13 (**Change of variable formula**)**.** Let $\boldsymbol{X} = \begin{pmatrix} X_1 & X_2 & \cdots & X_n \end{pmatrix}^\top$ be an n-dimensional random vector with joint PDF $f_{\boldsymbol{X}}(\boldsymbol{x})$. Let $G \in C^2(\mathbb{R}^n) : \mathbb{R}^n \to \mathbb{R}^n$ be invertible and let $H = G^{-1}$. Suppose a random vector $\boldsymbol{Y} = \begin{pmatrix} Y_1 & Y_2 & \cdots & Y_n \end{pmatrix}^\top$ is given by $\boldsymbol{Y} = G(\boldsymbol{X})$, hence $\boldsymbol{X} = H(\boldsymbol{Y})$, and

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} H_1(Y_1, Y_2, \cdots, Y_n) \\ H_2(Y_1, Y_2, \cdots, Y_n) \\ \vdots \\ H_n(Y_1, Y_2, \cdots, Y_n) \end{pmatrix},$$

then the PDF of $\boldsymbol{Y}$ is given by

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}(H(\boldsymbol{y})) \cdot |J|$$

where J is the Jacobian of H defined by $J = \left(\frac{\partial H_i}{\partial y_j}\right)_{i,j} \in M_n(\mathbb{R})$.

Definition 10 (**Conditional distribution**)**.** Let X, Y be random variables, $f_{X,Y}(\cdot,\cdot)$ be the joint density of X, Y. The **conditional distribution** of X given $Y = y$, denote as $f_{X|Y=y}(x|y)$ is defined by

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

where $f_Y(y)$ is the **marginal distribution** of Y, which is defined by

$$f_Y(y) = \sum_{x \in I} \mathbb{P}(X = x, Y = y), \quad \text{if } X, Y \text{ are discrete;}$$

and

$$f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x,y)dx, \quad \text{if } X, Y \text{ are continuous.}$$

1.5 Convergence of Random Variables

We will study the convergence of sequence of random variables $\{X_n\}_{n=1}^{\infty}$. We introduce 4 different kinds of convergence:

Definition 11. Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables, X be another random variable.

- **Weak convergence:** We say $X_n \to X$ weakly (or convergence in distribution, convergence in law), if $\lim_{n\to\infty} \mathbb{P}(X_n \leq x) = \mathbb{P}(X \leq x)$ for every x such that $x \mapsto \mathbb{P}(X \leq x)$ is continuous.

- **Convergence in probability:** We say $X_n \to X$ in probability ($X_n \xrightarrow{P} X$), if $\forall \epsilon > 0$,

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

- **Convergence in L^p:** We say $X_n \to X$ in L^p ($X_n \xrightarrow{L^p} X$), if $\lim_{n\to\infty} \mathbb{E}|X_n - X|^k = 0$.

- **Convergence almost surely:** We say $X_n \to X$ almost surely ($X_n \xrightarrow{a.s} X$), if

$$\mathbb{P}\left(\lim_{n\to\infty} |X_n - X| = 0\right) = 1.$$

Theorem 14. If $X_n \xrightarrow{L^p} X$, then $X_n \xrightarrow{P} X$.

Proof. The proof is straightforward, we have

$$\mathbb{P}\{|X_n - X| > \epsilon\} = \mathbb{P}\{|X_n - X|^p > \epsilon^p\} \leq \frac{\mathbb{E}|X_n - X|^p}{\epsilon^p} \to 0.$$

□

Theorem 15. If $X_n \xrightarrow{P} X$, then there exists a subsequence n_k of $\mathbb{N}$ such that $X_{n_k} \xrightarrow{a.s} X$.

Proof. Assume $X_n \xrightarrow{P} X$, then $\forall k \geq 1$, $\exists n_k$ such that $\mathbb{P}(\{X_{n_k} - X| > 1/k\}) \leq 1/k^2$, denote $A_k := \{|X_{n_k} - X| > 1/k\}$, then by Borel-Cantelli Lemma, we have

$$\mathbb{P}\left(\bigcap_{l=1}^{\infty}\bigcup_{k=l}^{\infty} A_k\right) = \lim_{l\to\infty} \mathbb{P}\left(\bigcup_{k=l}^{\infty} A_k\right) \leq \lim_{l\to\infty} \sum_{k=l}^{\infty} \mathbb{P}(A_k) = 0,$$

meaning that for almost everywhere, $\exists l$, such that $\forall k \geq l : |X_{n_k} - X| \leq 1/k$, which means that for almost everywhere, $\lim_{k\to\infty} |X_{n_k} - X| = 0$, thus $X_{n_k} \xrightarrow{a.s} X$. □

The figure below shows the relations between different modes of convergence. The arrows in red means direct implication without any further condition, while arrows in orange will hold if extra conditions are given. The general structure is that, convergence in probability implies convergence almost surely along a subsequence; Convergence in probability with uniform integrability would imply convergence in L^p; Convergence almost surely when dominated convergence theorem applies will imply convergence in L^p.

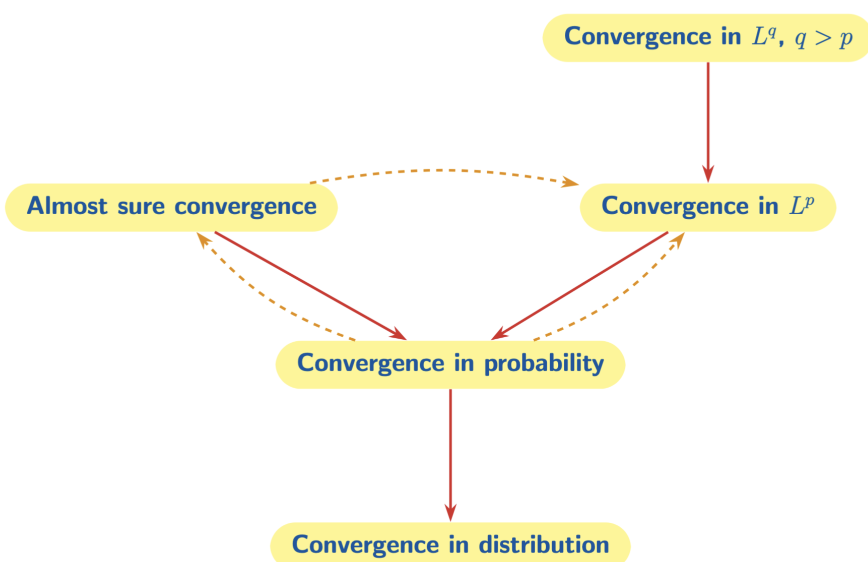

Figure 1: Different modes of convergences and their relations

Theorem 16 (**Continuous mapping theorem**)**.** Let $\{X_n\} \subset \mathbb{R}^k$ be a sequence of random vectors, $X \subset \mathbb{R}^k$ be another random vector. Suppose $g : \mathbb{R}^k \to \mathbb{R}^m$ is continuous almost everywhere on the support of $\{X_n\}, X$, then

$$X_n \to X \implies g(X_n) \to g(X),$$

where the convergence could be convergence in distribution, in probability, or almost everywhere.

Theorem 17 (**Slutsky's theorem**)**.** Let $\{X_n\}, \{Y_n\}, X, Y$ be random vectors, $c \in \mathbb{R}$. Assume $X_n \xrightarrow{d} X, Y_n \xrightarrow{P} c$, then

$$X_n + Y_n \xrightarrow{d} X + c; \quad Y_n X_n \xrightarrow{d} cX; \quad \frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}, \quad c, Y_n \neq 0.$$

Theorem 18 (**Lévy's continuity theorem**)**.** The characteristic function of X is defined as

$$\mathbb{1}_X(t) := \mathbb{E}[\exp(itX)].$$

Then $X_n \xrightarrow{d} X$ iff $\mathbb{1}_{X_n(t)} \overset{\text{pointwise}}{\rightarrow} \mathbb{1}_{X(t)}$ for all $t \in \mathbb{R}$.

Definition 12 (**Stochastic o notations**)**.** .

- A sequence of random vectors $\{X_n\}$ is **bounded in probability** (or uniformly tight), if $\forall \epsilon > 0$, $\exists M_\epsilon$, such that $\sup_n \mathbb{P}(||X_n|| \geq M_\epsilon) \leq \epsilon$. In this case we say $X_n = O_P(1)$.
- If a sequence of random vectors $\{X_n\}$ satisfies $X_n \xrightarrow{P} 0$, then we say $X_n = o_P(1)$.

Theorem 19 (**Central limit theorem**)**.** Let $\{X_i\}_{i=1}^\infty \overset{i.i.d}{\sim} f$ and assume $\mathbb{E}[X^2] < \infty$, then

$$\sqrt{n}\left(\overline{X}_n - \mathbb{E}[X]\right) \xrightarrow{d} N(0, \mathbb{V}\mathbf{ar}(X)).$$

Proof. WLOG assume $\mathbb{E}X = 0, \mathbb{V}\mathbf{ar}(X) = 1$ then

$$\begin{aligned}
\mathbb{1}_{\sqrt{n}\overline{X}_n}(t) &= \mathbb{E}\left[\exp\left(\sqrt{n}it\overline{X}_n\right)\right] \\
&= \mathbb{E}\left[\exp\left(\frac{it(X_1 + \cdots + X_n)}{\sqrt{n}}\right)\right] \\
&= \left(\mathbb{E}\left[\exp\left(\frac{itX}{\sqrt{n}}\right)\right]\right)^n \\
&= \left[\mathbb{1}_X\left(\frac{t}{\sqrt{n}}\right)\right]^n.
\end{aligned}$$

Then a Taylor expansion around 0 will yield

$$\begin{aligned}
\mathbb{1}_X\left(\frac{t}{\sqrt{n}}\right) &= \mathbb{1}_X(0) + \mathbb{1}_X'(0) \cdot \frac{t}{\sqrt{n}} + \mathbb{1}_X'' \cdot \frac{t^2}{2n} + o\left(\frac{1}{n}\right) \\
&= 1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)
\end{aligned}$$

This is because

$$\begin{aligned}
\mathbb{1}_X'(t)\bigg|_{t=0} &= \frac{d}{dt}\bigg|_{t=0} \int_B f(x) \cdot \exp(itX)dx \\
&= \int_B \frac{d}{dt}\bigg|_{t=0} f(x) \cdot \exp(itX)dx \\
&= \int_B ixf(x) \cdot \exp(itX)\bigg|_{t=0} dx \\
&:= i\mathbb{E}[X] \\
&= 0.
\end{aligned}$$

A similar statement can be drawn: $\mathbb{1}''_X(0) = i^2\mathbb{E}[X^2] = -1$. Use the fact that $\left(1 + \frac{x}{n}\right)^n \sim e^x$ for large n, we have

$$\begin{aligned}
\mathbb{1}_{\sqrt{n}\overline{X}_n}(t) &= \left[\mathbb{1}_X\left(\frac{t}{\sqrt{n}}\right)\right]^n \\
&= \left[1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right]^n \\
&= \exp\left(-\frac{t^2}{2}\right).
\end{aligned}$$

By the uniqueness of the characteristic function, we have finished the proof.

$\square$

Theorem 20 (**Delta method**)**.** Suppose T_n is a estimator (a $\mathfrak{B}(\mathbb{R})$-measurable function of data independent of unknown parameter θ) of θ. Let $\phi : \mathbb{R}^k \to \mathbb{R}^m$ be a map defined on a subset of $\mathbb{R}^k$ and differentiable at θ, and T_n takes values inside the domain of ϕ. If $r_n(T_n - \theta) \xrightarrow{d} T$ as $n \to \infty$, then

$$r_n(\phi(T_n) - \phi(\theta)) \xrightarrow{d} \phi'_\theta \cdot T; \quad r_n(\phi(T_n) - \phi(\theta)) - \phi'_\theta(r_n(T_n - \theta)) \xrightarrow{P} 0.$$

Here ϕ'_θ is the Jacobian of ϕ evaluated at θ.

2 Basics of Statistical inference

3 Bibliography

References

Axler, Sheldon (2020). *Measure, Integration and Real Analysis*.
McDonald, Daniel J. and Ben Bloem-Reddy (2026). *Lecture Notes for STAT 560 (Statistical Inference I)*.
Wasserman, Larry (2003). *All of Statistics*.
Zhang, Jiajun (2024). *Lecture Notes for MATH 356 (Honors Probability)*.