MEM6804 Modeling and Simulation for Logistics & Supply Chain 物流与供应链建模与仿真

Theory Analysis

Lecture 2: Elements of Probability and Statistics

SHEN Haihui 沈海辉

Sino-US Global Logistics Institute
Shanghai Jiao Tong University

shenhaihui.github.io/teaching/mem6804f
shenhaihui@sjtu.edu.cn

Spring 2021 (full-time)

上海交通大学 SHANGHAI JIAO TONG UNIVERSITY | 董浩云航运与物流研究院 CY TUNG Institute of Maritime and Logistics
中美物流研究院（工程系统管理研究院） Sino-US Global Logistics Institute (Institute of Industrial & System Engineering)

Contents

Probability Space

A **probability space** is a triplet $(\Omega, \mathcal{F}, \mathbb{P})$:

- Ω, sample space: A set of *all* possible outcomes.
  - A set of *some* outcomes, as a subset of Ω, is called an **event**.

- $\mathcal{F}$, σ-algebra (or σ-field): A set of events, i.e., a set of some subsets of Ω, such that:
  1. $\Omega \in \mathcal{F}$;
  2. Closed under complementation: If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$;
  3. Closed under countable unions:† If $A_i \in \mathcal{F}$, $i = 1, 2, \ldots$, is a **countable** sequence of sets, then $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$.

- $\mathbb{P} : \mathcal{F} \to [0, 1]$, probability function (or probability measure): A function that assigns probabilities to events, such that:
  1. $\mathbb{P}(A) \in [0, 1]$ for any $A \in \mathcal{F}$;
  2. $\mathbb{P}(\Omega) = 1$;
  3. Countably additive: If $A_i \in \mathcal{F}$, $i = 1, 2, \ldots$, is a **countable** sequence of **disjoint** sets, then $\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

† It implies that $\mathcal{F}$ is also closed under countable intersections.

Probability Space

- Example 1: Flip a fair coin.
  - $\Omega = \{$H (head), T (tail)$\}$;
  - $\mathcal{F} = \{\emptyset, \{\text{H}\}, \{\text{T}\}, \Omega\}$;
  - $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\{\text{H}\}) = 1/2$, $\mathbb{P}(\{\text{T}\}) = 1/2$, and $\mathbb{P}(\Omega) = 1$.
- Example 2: Draw a ball out of 3 balls (red, green, blue).
  - $\Omega = \{$R (red), G (green), B (blue)$\}$;
  - $\mathcal{F} = \{\emptyset, \{\text{R}\}, \{\text{G}\}, \{\text{B}\}, \{\text{R,G}\}, \{\text{R,B}\}, \{\text{G,B}\}, \Omega\}$;
  - $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\{\text{R}\}) = \mathbb{P}(\{\text{G}\}) = \mathbb{P}(\{\text{B}\}) = 1/3$, $\mathbb{P}(\{\text{R,G}\}) = \mathbb{P}(\{\text{R,B}\}) = \mathbb{P}(\{\text{G,B}\}) = 2/3$, and $\mathbb{P}(\Omega) = 1$;
  - $\mathcal{F}_1 = \{\emptyset, \{\text{R}\}, \{\text{G,B}\}, \Omega\}$, $\mathcal{F}_2 = \{\emptyset, \{\text{G}\}, \{\text{R,B}\}, \Omega\}$...
- Example 3: Randomly "draw" a number in $[0, 1]$
  - $\Omega = [0, 1]$;
  - $\mathcal{F}_1 = \{\emptyset, [0, a), [a, 1], \Omega\}$, $\mathcal{F}_2 = \{\emptyset, (0, a), \{0\} \cup [a, 1], \Omega\}$...
  - A more practical and interesting $\mathcal{F}$ is the one that contains all intervals (no matter open or closed) on $[0, 1]$.

- **Independence** of Events: Two events A and B in $\mathcal{F}$ are called statistically independent events when

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B).$$

- **Conditional Probability**: If A and B are events in $\mathcal{F}$ and $\mathbb{P}(B) > 0$, then the conditional probability of A given B, denoted as $\mathbb{P}(A|B)$, is

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

- Bayes' Rule:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\,\mathbb{P}(A)}{\mathbb{P}(B)}.$$

- Events A and B are independent $\iff \mathbb{P}(A|B) = \mathbb{P}(A)$.

Probability Space

- For more than two events:
  - **Mutual independence** (or collective independence) intuitively means that each event is independent of any combination of other events;
  - **Pairwise independence** means any two events in the collection are independent of each other.

- Sets $A_1, \ldots, A_n$ are (mutually) independent if for any $I \subset \{1, \ldots, n\}$ we have $\mathbb{P}(\cap_{i \in I} A_i) = \prod_{i \in I} \mathbb{P}(A_i)$.

- **Warning**: Only having $\mathbb{P}(\cap_{i=1}^n A_i) = \prod_{i=1}^n \mathbb{P}(A_i)$ is not sufficient!

- Sets $A_1, \ldots, A_n$ are pairwise independent if for any $i \neq j$ we have $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\,\mathbb{P}(A_j)$.

- Clearly, mutual independence implies pairwise independence, but not vice versa!

Consider a sequence of sets $\{A_n : n \geq 1\}$.

(The First) Borel-Cantelli Lemma

If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(A_n \text{ i.o.}) = 0$, where "i.o." denotes "infinitely often".

The Secon Borel-Cantelli Lemma

If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$ and $\{A_n\}$ are independent,[†] then $\mathbb{P}(A_n \text{ i.o.}) = 1$.

- Remark: For event A, if $\mathbb{P}(A) = 1$, then we say A happens **almost surely** (a.s.).

[†] The assumption of independence can be weakened to pairwise independence, with more difficult proof.

- A **random variable** (RV) is a function from a sample space Ω into the set of real numbers $\mathbb{R}$.
- Formally, given the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a RV X is a function $X : \Omega \to \mathbb{R}$, such that for any $a \in \mathbb{R}$,

$$\{\omega \in \Omega : X(\omega) \leq a\} \in \mathcal{F}.$$

- For a particular element $\omega \in \Omega$, $X(\omega)$ is called a *realization* of X.
  - Usually, we will simply denote $X(\omega)$ as x when ω is not explicitly shown.
  - A popular convention is to denote the RVs by upper-case letters (e.g., X and Y) and their realizations by lower-case letters (e.g., x and y).

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

- Example 1': Let $X(\mathsf{H}) = 0$, $X(\mathsf{T}) = 1$.
- Example 2':
  - Under $(\Omega, \mathcal{F}, \mathbb{P})$, let $X(\mathsf{R}) = 0$, $X(\mathsf{G}) = 1$, and $X(\mathsf{B}) = 2$.
  - Under $(\Omega, \mathcal{F}_1, \mathbb{P})$, let $X(\mathsf{R}) = 0$, $X(\mathsf{G}) = 1$, and $X(\mathsf{B}) = 1$.
- Example 3':
  - Under $(\Omega, \mathcal{F}_1, \mathbb{P})$, let $X(\omega) := \begin{cases} 0, & \text{if } \omega \in [0, a), \\ 1, & \text{if } \omega \in [a, 1]. \end{cases}$
  - Under $(\Omega, \mathcal{F}, \mathbb{P})$, let $X(\omega) = \omega$ for $\omega \in [0, 1]$.

- The **cumulative distribution function** (CDF) of a RV X, denoted by $F : \mathbb{R} \to [0, 1]$, is defined by

$$F(x) \coloneqq \mathbb{P}(X \leq x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\}), \ \forall x \in \mathbb{R},$$

and the following is satisfied:

  - $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to+\infty} F(x) = 1$;
  - $F(x)$ is nondecreasing in x;
  - $F(x)$ is right-continuous, that is, for any $x_0 \in \mathbb{R}$,

$$\lim_{x \downarrow x_0} F(x) = F(x_0).$$

- A RV X is said to be **discrete** if the set of its possible values is countable.
- The **probability mass function** (pmf) of a discrete RV X is given by

$$p(x) := \mathbb{P}(X = x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x\}), \ \forall x \in \mathbb{R},$$

  and the following is satisfied:
  - $p(x) \geq 0$ for all $x \in \mathbb{R}$;
  - $\sum_{x\in\mathbb{R}} p(x) = 1$.
- It is easy to see that $F(x) = \sum_{y\in(-\infty,\, x]} p(y)$.

- A RV X is said to be **continuous** if there exists a **probability density function** (pdf) $f(x)$ such that

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(t)\mathrm{d}t, \ \forall x \in \mathbb{R},$$

  and the following is satisfied:

  - $f(x) \geq 0$ for all $x \in \mathbb{R}$;
  - $\int_{-\infty}^{+\infty} f(t)\mathrm{d}t = 1$.

- Observe that $\frac{\mathrm{d}}{\mathrm{d}x}F(x) = f(x)$.

- The **joint** CDF of RVs X and Y, denoted by $F : \mathbb{R}\times\mathbb{R} \to [0,1]$, is defined by

$$\begin{aligned} F(x,y) &:= \mathbb{P}(X \le x, Y \le y) \\ &= \mathbb{P}(\{\omega : X(\omega) \le x\} \cap \{\omega : Y(\omega) \le y\}), \ \forall x, y \in \mathbb{R}. \end{aligned}$$

- For discrete RVs X and Y, the **joint** pmf is given by

$$\begin{aligned} p(x,y) &:= \mathbb{P}(X = x, X = y) \\ &= \mathbb{P}(\{\omega : X(\omega) = x\} \cap \{\omega : Y(\omega) = y\}), \ \forall x, y \in \mathbb{R}. \end{aligned}$$

- For continuous RVs X and Y, the **joint** pdf is $f(x,y)$ such that

$$F(x,y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(t,u) \mathrm{d}t\mathrm{d}u, \ \forall x, y \in \mathbb{R}.$$

- Observe that $\frac{\partial^2 F(x,y)}{\partial x \partial y} = f(x,y)$.

- Given the random vector $(X, Y)^\intercal$, the distribution of X or Y is called the **marginal distribution**.
  - The marginal CDF of X is $F_X(x) = F(x, +\infty)$.

- If $(X, Y)^\intercal$ is discrete, the marginal pmf of X is

$$p_X(x) = \sum_{y \in \mathbb{R}} p(x, y).$$

- If $(X, Y)^\intercal$ is continuous, the marginal pdf of X is

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y) \mathrm{d}y.$$

- For Y, its marginal CDF, and pmf or pdf, can be determined similarly.

- If $(X, Y)^\mathsf{T}$ is discrete, for any y such that $\mathbb{P}(Y = y) = p_Y(y) > 0$, the **conditional** pmf of X given that $Y = y$ is defined as

$$p(x|y) := \mathbb{P}(X = x|Y = y) = \frac{p(x,y)}{p_Y(y)}.$$

- If $(X, Y)^\mathsf{T}$ is continuous, for any y such that $f_Y(y) > 0$, the **conditional** pdf of X given that $Y = y$ is defined as

$$f(x|y) := \frac{f(x,y)}{f_Y(y)}.$$

Intuitively, $f(x|y)$ can be understood as follows (although it is not the most rigorous approach):

1. Note that

$$\begin{aligned} F(x|Y=y) &= \lim_{\Delta\to 0} F(x|Y \text{ between } y \text{ and } y+\Delta) \\ &= \lim_{\Delta\to 0} \frac{\mathbb{P}(X\le x, Y \text{ between } y \text{ and } y+\Delta)}{\mathbb{P}(Y \text{ between } y \text{ and } y+\Delta)} \\ &= \frac{\lim_{\Delta\to 0}[F(x,y+\Delta)-F(x,y)]/\Delta}{\lim_{\Delta\to 0}[F_Y(y+\Delta)-F_Y(y)]/\Delta} \\ &= \frac{\frac{\partial}{\partial y}F(x,y)}{\frac{\mathrm{d}}{\mathrm{d}y}F_Y(y)} = \frac{\frac{\partial}{\partial y}\int_{-\infty}^{y}\int_{-\infty}^{x} f(t,u)\mathrm{d}t\mathrm{d}u}{f_Y(y)} \\ &= \frac{\int_{-\infty}^{x} f(t,y)\mathrm{d}t}{f_Y(y)}. \end{aligned}$$

2. Then, $f(x|y) = \frac{\partial}{\partial x}F(x|Y=y) = \frac{\frac{\partial}{\partial x}\int_{-\infty}^{x} f(t,y)\mathrm{d}t}{f_Y(y)} = \frac{f(x,y)}{f_Y(y)}$.

- Two RVs X and Y are said to be statistically **independent**, which can be denoted as $X \perp Y$, when, for any $x, y \in \mathbb{R}$,

$$F(x,y) = F_X(x)F_Y(y), \text{ or,}$$
$$p(x,y) = p_X(x)p_Y(y), \text{ or,}$$
$$f(x,y) = f_X(x)f_Y(y).$$

- X and Y are independent $\Longleftrightarrow$
  - $p(x|y) \equiv p_X(x)$ or $f(x|y) \equiv f_X(x)$ regardless of the value y;
  - $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\,\mathbb{P}(X \in B)$ for any $A, B \subset \mathbb{R}$.

上海交通大学 SHANGHAI JIAO TONG UNIVERSITY

- For more than two RVs $X_1, \dots, X_n$, the joint CDF, joint pmf or pdf, and the marginal pmf or pdf, are defined analogically.
- RVs $X_1, \dots, X_n$ are (mutually) independent if

$$F(x_1, \dots, x_n) \equiv F_{X_1}(x_1) \times \cdots \times F_{X_n}(x_n), \text{ or,}$$
$$p(x_1, \dots, x_n) \equiv p_{X_1}(x_1) \times \cdots \times p_{X_n}(x_n), \text{ or,}$$
$$f(x_1, \dots, x_n) \equiv f_{X_1}(x_1) \times \cdots \times f_{X_n}(x_n).$$

- RVs $X_1, \dots, X_n$ are pairwise independent if for any $i \neq j$, $X_i \perp X_j$.

- The **expectation**, or **expected value**, or **mean**, of a RV X is defined as
$$\mathbb{E}[X] := \int_\Omega X(\omega) \mathrm{d}\, \mathbb{P}(\omega),$$
provided that $\int_\Omega |X(\omega)| \mathrm{d}\, \mathbb{P}(\omega) < \infty$ or $X \geq 0$ a.s., where the integral is the Lebesgue integral, rather than the Riemann integral.
- For function $h : \mathbb{R} \to \mathbb{R}$, $\mathbb{E}[h(X)] = \int_\Omega h(X(\omega)) \mathrm{d}\, \mathbb{P}(\omega)$.
- If X is a discrete RV:
    - $\mathbb{E}[X] = \sum_{x \in \mathbb{R}} x p(x)$;
    - $\mathbb{E}[h(X)] = \sum_{x \in \mathbb{R}} h(x) p(x)$.
- If X is a continuous RV:
    - $\mathbb{E}[X] = \int_{-\infty}^{+\infty} x f(x) \mathrm{d}x$;
    - $\mathbb{E}[h(X)] = \int_{-\infty}^{+\infty} h(x) f(x) \mathrm{d}x$.

Expectations

- For integer n, $\mathbb{E}[X^n]$ is called the nth **moment** of X, and $\mathbb{E}[(X - \mathbb{E}[X])^n]$ is called the nth **central moment** of X.
- Some special moments:
  - Mean (1st moment): $\mu := \mathbb{E}[X]$.
  - **Variance** (2nd central moment): $\sigma^2 := \operatorname{Var}(X) := \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.
- Linear association:
  - **Covariance**: $\operatorname{Cov}(X, Y) := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$.
  - **Correlation**: $\rho(X, Y) := \frac{\operatorname{Cov}(X, Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}}$.
- In general, $X \perp Y \underset{\not\Longleftarrow}{\Longrightarrow} \rho(X, Y) = 0 \iff \operatorname{Cov}(X, Y) = 0$.
- If $(X, Y)^\intercal$ follows a bivariate normal distribution,† then $X \perp Y \iff \rho(X, Y) = 0$.

† **CAUTION:** It means MORE than that X and Y both follow a normal distribution! More details latter.

Expectations

- The conditional expectation of X given $Y = y$ is

$$\mathbb{E}[X|y] := \begin{cases} \sum_{x\in\mathbb{R}} xp(x|y), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{+\infty} xf(x|y)\mathrm{d}x, & \text{if } X \text{ is continuous.} \end{cases}$$

- The conditional variance of X given $Y = y$ is

$$\mathrm{Var}(X|y) := \mathbb{E}[(X - \mathbb{E}[X])^2|y] = \mathbb{E}[X^2|y] - (\mathbb{E}[X|y])^2.$$

- If $X \not\perp Y$, then $\mathbb{E}[X|y]$ and $\mathrm{Var}(X|y)$ are functions of y.
- If $X \not\perp Y$, then $\mathbb{E}[X|Y]$ and $\mathrm{Var}(X|Y)$ are also RVs, whose value depends on the value of Y.
- If $X \perp Y$, then $\mathbb{E}[X|y] = \mathbb{E}[X|Y] = \mathbb{E}[X]$, and $\mathrm{Var}(X|y) = \mathrm{Var}(X|Y) = \mathrm{Var}(X)$.

Expectations

- $\mathbb{E}[aX + bY] = a\,\mathbb{E}[X] + b\,\mathbb{E}[Y]$.
- $\text{Var}(aX + bY) = a^2\,\text{Var}(X) + 2ab\,\text{Cov}(X, Y) + b^2\,\text{Var}(Y)$.
- $\text{Cov}(aX + bY, cW + dV) = ac\,\text{Cov}(X, W) + ad\,\text{Cov}(X, V) + bc\,\text{Cov}(Y, W) + bd\,\text{Cov}(Y, V)$.
- $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$.
- $\text{Var}(X) = \mathbb{E}[\text{Var}(X|Y)] + \text{Var}(\mathbb{E}[X|Y])$.
- If $X \perp Y$, then $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$.

- $X \sim \text{Bernoulli}(p)$ or $\text{Ber}(p)$, if

$$X = \begin{cases} 1, & \text{with probability } p, \\ 0, & \text{with probability } 1-p, \end{cases} \quad p \in [0,1].$$

  - $\mathbb{E}[X] = p$, $\text{Var}(X) = p(1-p)$.
  - The value $X = 1$ is often termed a "success" and p is referred to as the success probability.

- $Y \sim \text{binomial}(n, p)$ or $\text{B}(n, p)$: The number of successes among n (mutually) **independent and identically distributed** (iid) $\text{Ber}(p)$ trials.
  - $Y = \sum_{i=1}^{n} X_i$, where $X_i \sim \text{Ber}(p)$ are iid.
  - $p(y) = \mathbb{P}(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y = 0, 1, \ldots, n.$
  - $\mathbb{E}[Y] = np$, $\text{Var}(Y) = np(1-p)$.

- If $Y_1 \sim \text{B}(n_1, p)$ and $Y_2 \sim \text{B}(n_2, p)$ are independent, then $Y_1 + Y_2 \sim \text{B}(n_1 + n_2, p)$.

- $Y \sim$ negative binomial(r,p) or $\mathrm{NB}(r,p)$: The number of iid $\mathrm{Ber}(p)$ trials to obtain r successes.
  - $p(y) = \mathbb{P}(Y=y) = \binom{y-1}{r-1} p^r (1-p)^{y-r}, \quad y = r, r+1, \ldots$.
  - $\mathbb{E}[Y] = r + r(1-p)/p$, $\mathrm{Var}(Y) = r(1-p)/p^2$.
  - When $r=1$, it becomes the geometric distribution.

- $Y \sim$ geometric(p) or $\mathrm{Geo}(p)$: The number of iid $\mathrm{Ber}(p)$ trials to obtain the first success.
  - $p(y) = \mathbb{P}(Y=y) = p(1-p)^{y-1}, \quad y = 1, 2, \ldots$.
  - $\mathbb{E}[Y] = 1/p$, $\mathrm{Var}(Y) = (1-p)/p^2$.
  - **Memoryless Property**: For integers $s > t$,

$$\mathbb{P}(Y>s|Y>t) = \frac{\mathbb{P}(Y>s, Y>t)}{\mathbb{P}(Y>t)} = \frac{\mathbb{P}(Y>s)}{\mathbb{P}(Y>t)} = \frac{(1-p)^s}{(1-p)^t} = (1-p)^{s-t}$$
$$= \mathbb{P}(X > s-t).$$

- If $Y_1 \sim \mathrm{NB}(r_1, p)$ and $Y_2 \sim \mathrm{NB}(r_2, p)$ are independent, then $Y_1 + Y_2 \sim \mathrm{NB}(r_1 + r_2, p)$.

- Poisson distribution is often used to model the number of occurrence in a given time interval.
- One of the basic assumptions is that, *for very small time intervals, the probability of an occurrence is proportional to the length of the time interval.*[†]
- $X \sim \text{Poisson}(\lambda)$ or $\text{Pois}(\lambda)$, with $\lambda > 0$, if

$$p(x) = \mathbb{P}(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, \ldots .$$

  - It can be verified that $\sum_{x=0}^{\infty} p(x) = 1$.
  - $\mathbb{E}[X] = \lambda$, $\text{Var}(X) = \lambda$.
- If $X_1 \sim \text{Pois}(\lambda_1)$ and $X_2 \sim \text{Pois}(\lambda_2)$ are independent,
  - $X_1 + X_2 \sim \text{Pois}(\lambda_1 + \lambda_2)$;
  - Given $X_1 + X_2 = n$, $X_1 \sim \text{B}(n, \lambda_1/(\lambda_1 + \lambda_2))$.

[†] See more detailed discussion in Lec 3.

- $X \sim \mathrm{Uniform}(a,b)$ or $\mathrm{Unif}(a,b)$ with $a < b$, if its pdf is given by

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{if } x \in [a,b], \\ 0, & \text{otherwise.} \end{cases}$$

  - $\mathbb{E}[X] = \frac{b+a}{2}$, $\mathrm{Var}(X) = \frac{(b-a)^2}{12}$.

- $X \sim \mathrm{exponential}(\lambda)$ or $\mathrm{Exp}(\lambda)$, with $\lambda > 0$, if its pdf is given by

$$f(x) = \lambda e^{-\lambda x}, \quad x \in [0, \infty).$$

  - λ is called the rate parameter.
  - $F(x) = 1 - e^{-\lambda x}$, $\mathbb{P}(X > x) = 1 - F(x) = e^{-\lambda x}$.
  - $\mathbb{E}[X] = 1/\lambda$, $\mathrm{Var}(X) = 1/\lambda^2$.
  - **Memoryless Property**: For $s > t \geq 0$,

$$\mathbb{P}(X > s | X > t) = \frac{\mathbb{P}(X > s, X > t)}{\mathbb{P}(X > t)} = \frac{\mathbb{P}(X > s)}{\mathbb{P}(X > t)} = \frac{e^{-\lambda s}}{e^{-\lambda t}} = e^{-\lambda(s-t)}$$
$$= \mathbb{P}(X > s - t).$$

- If $X_1 \sim \mathrm{Exp}(\lambda_1)$ and $X_2 \sim \mathrm{Exp}(\lambda_2)$ are independent, then $\min\{X_1, X_2\} \sim \mathrm{Exp}(\lambda_1 + \lambda_2)$.

- If $X \sim \mathrm{Exp}(\lambda)$, then for $\alpha > 0$, $Y := X^{1/\alpha} \sim \mathrm{Weibull}(\alpha, \beta)$ in shape & scale parametrization with $\beta = (1/\lambda)^{1/\alpha}$, whose pdf is
$$f(y) = \alpha\beta^{-\alpha} y^{\alpha-1} e^{-(y/\beta)^\alpha}, \quad y \in (0, \infty).$$

- $\mathrm{Erlang}(k, \lambda)$ or $\mathrm{Erl}(k, \lambda)$, with k being a positive integer, is a generalized version of $\mathrm{Exp}(\lambda)$, whose pdf is
$$f(x) = \frac{\lambda^k}{(k-1)!} x^{k-1} e^{-\lambda x}, \quad x \in [0, \infty).$$
  - $\mathbb{E}[X] = k/\lambda$, $\mathrm{Var}(X) = k/\lambda^2$.
  - $k = 1 \implies \mathrm{Exp}(\lambda)$.

- If $X_1 \sim \mathrm{Erl}(k_1, \lambda)$ and $X_2 \sim \mathrm{Erl}(k_2, \lambda)$ are independent, then $X_1 + X_2 \sim \mathrm{Erl}(k_1 + k_2, \lambda)$.

- If $X \sim \mathrm{Erl}(k, \lambda)$, then $cX \sim \mathrm{Erl}(k, \lambda/c)$ for $c > 0$.

- $X \sim \text{Gamma}(\alpha, \lambda)$ in shape & rate parametrization with $\alpha, \lambda > 0$, if its pdf is given by

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \quad x \in [0, \infty).$$

  - $\mathbb{E}[X] = \alpha/\lambda$, $\text{Var}(X) = \alpha/\lambda^2$.

- $\Gamma(\alpha) := \int_0^\infty t^{\alpha-1} e^{-t} \mathrm{d}t$ is known as the gamma function.
  - $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$; $\Gamma(n) = (n-1)!$, for integer $n > 0$.

- If $X_1 \sim \text{Gamma}(\alpha_1, \lambda)$ and $X_2 \sim \text{Gamma}(\alpha_2, \lambda)$ are independent, then $X_1 + X_2 \sim \text{Gamma}(\alpha_1 + \alpha_2, \lambda)$.

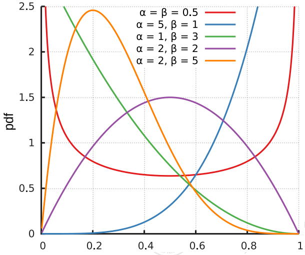
- If $X \sim \text{Gamma}(\alpha, \lambda)$, then $cX \sim \text{Gamma}(\alpha, \lambda/c)$ for $c > 0$.

- Important special cases of $\text{Gamma}(\alpha, \lambda)$:
  - α is an integer $\Longrightarrow \text{Erl}(\alpha, \lambda)$; $\alpha = 1 \Longrightarrow \text{Exp}(\lambda)$;
  - $\alpha = p/2$, where p is an integer, and $\lambda = 1/2 \Longrightarrow$ chi-square distribution with p degrees of freedom, denoted as χ^2_p.

- Beta distribution is a very flexible distribution that in a finite interval.
- $X \sim \text{Beta}(\alpha, \beta)$ with $\alpha, \beta > 0$, if its pdf is given by

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \ \ x \in [0,1].$$

  - $\mathbb{E}[X] = \alpha/(\alpha+\beta)$, $\text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.
- $B(\alpha,\beta) := \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\mathrm{d}t$ is known as the beta function.
  - $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.
- The $\text{Beta}(\alpha,\beta)$ pdf is quite flexible
  - $\alpha = 1, \beta = 1 \implies \text{Unif}(0,1)$
  - $\alpha > 1, \beta = 1 \implies$ strictly increasing
  - $\alpha = 1, \beta > 1 \implies$ strictly decreasing
  - $\alpha < 1, \beta < 1 \implies$ U-shaped
  - $\alpha > 1, \beta > 1 \implies$ unimodal



- $X \sim$ Student's t distribution with p degrees of freedom, denoted as t_p, where p is an integer, if its pdf is given by

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

  - $\mathbb{E}[X] = 0$ if $p > 1$;
  - $\mathrm{Var}(X) = p/(p-2)$ if $p > 2$.

- t_1 is also known as the standard Cauchy distribution, or $\mathrm{Cauchy}(0, 1)$, whose pdf is simply

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

- The **normal distribution** (sometimes called the Gaussian distribution) plays a **central role** in a large body of statistics.

- $X \sim$ normal distribution with mean μ and variance σ^2, denoted as $\mathcal{N}(\mu, \sigma^2)$, with $\sigma > 0$, if its pdf is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$$

  - $\mathbb{E}[X] = \mu$, $\mathrm{Var}(X) = \sigma^2$.

- If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Z := (X - \mu)/\sigma \sim \mathcal{N}(0, 1)$.
  - Z is also known as the **standard normal** RV.
  - We often use $\Phi(z)$ and $\phi(z)$ to denote the CDF and pdf of Z.
  - $\mathbb{P}(X \le x) = \Phi((x - \mu)/\sigma)$.

- If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $a + bX \sim \mathcal{N}(a + b\mu, b^2\sigma^2)$ for $b > 0$.

- If $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are independent, then $X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

- If $Z \sim \mathcal{N}(0,1)$, then $Z^2 \sim \chi_1^2$.

Proof. Let $Y := Z^2$. For $y \in [0, \infty)$,

$$\mathbb{P}(Y \le y) = \mathbb{P}(Z^2 \le y) = \mathbb{P}(-\sqrt{y} \le Z \le \sqrt{y}) = \int_{-\sqrt{y}}^{\sqrt{y}} \phi(t)\mathrm{d}t =: F(y).$$

Then,

$$\begin{aligned} f(y) &= \frac{\mathrm{d}}{\mathrm{d}y}F(y) = \phi(\sqrt{y})\frac{\mathrm{d}}{\mathrm{d}y}\sqrt{y} - \phi(-\sqrt{y})\frac{\mathrm{d}}{\mathrm{d}y}(-\sqrt{y}) \\ &= 2\phi(\sqrt{y})\frac{\mathrm{d}}{\mathrm{d}y}\sqrt{y} = \frac{1}{\sqrt{2\pi}}e^{-\frac{y}{2}}y^{-\frac{1}{2}}. \end{aligned}$$

If $Y \sim \chi_1^2$, i.e., $Y \sim \mathrm{Gamma}(1/2, 1/2)$, it means its pdf is

$$f(y) = \frac{1}{\sqrt{2}\Gamma(\frac{1}{2})}y^{-\frac{1}{2}}e^{-\frac{y}{2}}.$$

The proof is completed by showing that $\Gamma(\frac{1}{2}) = \int_0^\infty t^{-\frac{1}{2}}e^{-t}\mathrm{d}t = \sqrt{\pi}$, which can be seen if we convert to polar coordinates. $\blacksquare$

- If $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_p$ are independent, then $\frac{Z}{\sqrt{V/p}} \sim t_p$.

Proof. Since $V \sim \chi^2_p$, by definition, its pdf is

$$f_V(v) = \frac{(\frac{1}{2})^{\frac{p}{2}}}{\Gamma(\frac{p}{2})} v^{\frac{p}{2}-1} e^{-\frac{1}{2}v}, \quad v \in [0,\infty).$$

Let $Y := \sqrt{V/p}$. For $y \in [0,\infty)$,

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}\mathbb{P}(Y \le y) = \frac{\mathrm{d}}{\mathrm{d}y}\mathbb{P}(V \le py^2) = \frac{\mathrm{d}}{\mathrm{d}y}\int_0^{py^2} f_V(v)\mathrm{d}v = 2py f_V(py^2).$$

Let $T := \frac{Z}{\sqrt{V/p}} = \frac{Z}{Y}$. For $t \in \mathbb{R}$,

$$\mathbb{P}(T \le t) = \mathbb{P}\left(\frac{Z}{Y} \le t\right) = \mathbb{P}(Z \le tY) = \int_0^\infty \mathbb{P}(Z \le ty) f_Y(y)\mathrm{d}y. \quad \text{(Why?)}$$

Then,

$$f_T(t) = \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{P}(T \le t) = \int_0^\infty \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{P}(Z \le ty) f_Y(y)\mathrm{d}y.$$

Proof. (Cont'd) Note that $\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{P}(Z \le ty) = \frac{\mathrm{d}}{\mathrm{d}t}\int_{-\infty}^{ty}\phi(z)\mathrm{d}z = y\phi(ty)$. So,

$$\begin{aligned} f_T(t) &= \int_0^\infty y\phi(ty)f_Y(y)\mathrm{d}y = \int_0^\infty y\phi(ty)2pyf_V(py^2)\mathrm{d}y \\ &= \int_0^\infty 2py^2 \cdot \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2y^2}{2}} \cdot \frac{(\frac{1}{2})^{\frac{p}{2}}}{\Gamma(\frac{p}{2})}(py^2)^{\frac{p}{2}-1}e^{-\frac{1}{2}py^2}\mathrm{d}y \\ &= \frac{1}{\Gamma(\frac{p}{2})}\frac{1}{(p\pi)^{1/2}}2^{\frac{1-p}{2}}p^{\frac{p+1}{2}}\int_0^\infty y^p e^{-\frac{1}{2}(t^2+p)y^2}\mathrm{d}y. \end{aligned}$$

Let $x := y^2$. Then, integration by substitution shows that

$$\int_0^\infty y^p e^{-\frac{1}{2}(t^2+p)y^2}\mathrm{d}y = \frac{1}{2}\int_0^\infty x^{\frac{p-1}{2}}e^{-\frac{1}{2}(t^2+p)x}\mathrm{d}x =: \frac{1}{2}\int_0^\infty x^{\alpha-1}e^{-\lambda x}\mathrm{d}x,$$

where $\alpha := \frac{p+1}{2}$ and $\lambda := \frac{1}{2}(t^2+p)$. Recalling the pdf of $\Gamma(\alpha,\lambda)$, it is easy to see that $\int_0^\infty x^{\alpha-1}e^{-\lambda x}\mathrm{d}x = \Gamma(\alpha)/\lambda^\alpha$. Finally,

$$\begin{aligned} f_T(t) &= \frac{1}{\Gamma(\frac{p}{2})}\frac{1}{(p\pi)^{1/2}}2^{\frac{1-p}{2}}p^{\frac{p+1}{2}} \cdot \frac{1}{2}\frac{\Gamma(\frac{p+1}{2})}{(1/2)^{(p+1)/2}(t^2+p)^{(p+1)/2}} \\ &= \frac{\Gamma(\frac{p+1}{2})}{\Gamma(\frac{p}{2})}\frac{1}{(p\pi)^{1/2}}\frac{1}{(1+t^2/p)^{(p+1)/2}}. \end{aligned}$$

■

- $\boldsymbol{X} := (X_1, \ldots, X_k)^\intercal$ is said to follow a k-variate normal distribution, if **every** linear combination of $X_1, \ldots, X_k$ follows a (univariate) normal distribution.
  - $\boldsymbol{X}$ is also called a (k dimensional) normal random vector.
  - If $k = 2$, $\boldsymbol{X} = (X_1, X_2)^\intercal$ is also said to follow a *bivariate* normal distribution.

- $\boldsymbol{X} \sim$ a k-variate normal distribution, denoted as $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if its joint pdf is given by

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{k/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}, \ \boldsymbol{x} \in \mathbb{R}^k,$$

  where $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$.
  - $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)^\intercal = \mathbb{E}[\boldsymbol{X}] = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_k])^\intercal \in \mathbb{R}^k$.
  - $\boldsymbol{\Sigma} = (\Sigma_{ij}) = \mathrm{Cov}(\boldsymbol{X}, \boldsymbol{X}) = (\mathrm{Cov}(Z_i, Z_j)) \in \mathbb{R}^{k\times k}$.
  - $\boldsymbol{\Sigma}$ is a symmetric and positive definite matrix.
  - $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, $i = 1, \ldots, k$.

- If $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is k dimensional, then
  - $\boldsymbol{Z} := \boldsymbol{A}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, where $\boldsymbol{A}$ satisfies $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^\intercal$ (Cholesky decomposition), $\mathbf{0} \in \mathbb{R}^k$, and $\boldsymbol{I} \in \mathbb{R}^{k\times k}$ denotes the identity matrix.
  - $\boldsymbol{Z} = (Z_1, \ldots, Z_k)^\intercal$, where $Z_i \sim \mathcal{N}(0, 1)$, $i = 1, \ldots, k$, **iid**.
  - $\boldsymbol{a} + \boldsymbol{B}\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{a} + \boldsymbol{B}\boldsymbol{\mu}, \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^\intercal)$.†

- Suppose $\boldsymbol{X}$ is a k dimensional random vector. Then,
  $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \iff$
  There exist $\boldsymbol{\mu} \in \mathbb{R}^k$ and $\boldsymbol{A} \in \mathbb{R}^{k\times \ell}$ such that $\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{Z}$, where $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ with $\mathbf{0} \in \mathbb{R}^\ell$ and $\boldsymbol{I} \in \mathbb{R}^{\ell\times\ell}$.
  - Such $\boldsymbol{A}$ must satisfy $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^\intercal$.

† The multivariate normal distribution will be degenerate if $\boldsymbol{B}$ does not have full row rank ($\boldsymbol{B}$ 不行满秩).

- Bivariate normal distribution: $(X_1, X_2)^\intercal \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (\mu_1, \mu_2)^\intercal$, and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathrm{Cov}(X_1, X_1) & \mathrm{Cov}(X_1, X_2) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_2, X_2) \end{bmatrix} =: \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix},$$

and the joint pdf is

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]}.$$

- To see $\rho = 0 \Longrightarrow X_1 \perp X_2$, let $\rho = 0$, and note

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2} e^{-\frac{1}{2}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]}$$

$$= \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}} \times \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x_2-\mu_2)^2}{2\sigma_2^2}} = f_{X_1}(x_1) f_{X_2}(x_2).$$

- If $(X_1, X_2)^\intercal \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $X_i \sim \mathcal{N}(\mu_i, \sigma^2)$, $i = 1, 2$, then $X_1 + X_2 \perp X_1 - X_2$.

Proof. Note that

$$\boldsymbol{Y} := \begin{bmatrix} X_1 + X_2 \\ X_1 - X_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} =: \boldsymbol{B} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}.$$

Since $\boldsymbol{B}$ has full row rank, $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{B}\boldsymbol{\mu}, \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^\intercal)$, which is non-degenerate. Hence, to prove $X_1 + X_2 \perp X_1 - X_2$, it suffices to show $\mathrm{Cov}(X_1 + X_2, X_1 - X_2) = 0$. Note that

$$\begin{aligned} \mathrm{Cov}(X_1 + X_2, X_1 - X_2) &= \mathrm{Cov}(X_1, X_1) - \mathrm{Cov}(X_2, X_2) \\ &= \sigma^2 - \sigma^2 = 0. \end{aligned}$$ ■

- There are many other relationships among various probability distributions.
  - See, for example, Song (2005);
  - Or, Leemis & McQueston (2008) and their online interactive graph http://www.math.wm.edu/~leemis/chart/UDR/UDR.html

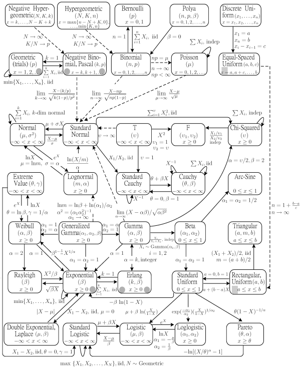


Figure: Relationships Among 35 Distributions (*from* *Song (2005)*)

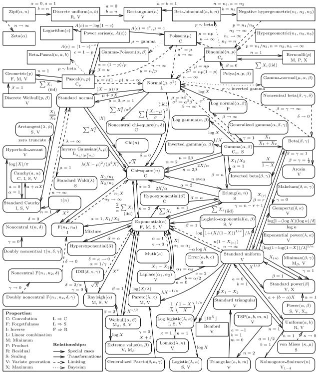


Figure: Relationships Among 76 Distributions (*from* *Leemis & McQueston (2008)*)

Markov's Inequality

Let X be a RV. If $\mathbb{P}(X \geq 0) = 1$ and $\mathbb{P}(X = 0) < 1$, then, for any $r > 0$,

$$\mathbb{P}(X \geq r) \leq \frac{\mathbb{E}[X]}{r},$$

with equality if and only if

$$X = \begin{cases} r, & \text{with probability } p, \\ 0, & \text{with probability } 1 - p. \end{cases}$$

- Markov's Inequality has many variations, which are usually called Chebyshev's Inequality.

Chebyshev's Inequality

Let X be a RV and $g(x)$ be a nonnegative function. Then, for any $r > 0$,

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

Chebyshev's Inequality

Let X be a RV. Then, for any $r, p > 0$,

$$\mathbb{P}(|X| \geq r) \leq \frac{\mathbb{E}[|X|^p]}{r^p},$$

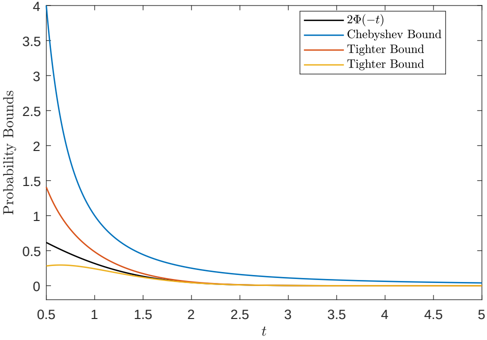
$$\mathbb{P}(|X - \mu| \geq r) \leq \frac{\sigma^2}{r^2},$$

where $\mu := \mathbb{E}[X]$, and $\sigma^2 := \mathrm{Var}(X)$.

- Chebyshev's Inequality is typically very conservative.
- If $Z \sim \mathcal{N}(0,1)$, a tighter bound is available: For any $t > 0$,

$$2\Phi(-t) = \mathbb{P}(|Z| \geq t) \leq \sqrt{\frac{2}{\pi}} \frac{1}{t} e^{-t^2/2},$$

$$2\Phi(-t) = \mathbb{P}(|Z| \geq t) \geq \sqrt{\frac{2}{\pi}} \frac{t}{1+t^2} e^{-t^2/2}.$$



- A function $g(x)$ is **convex** if

$$g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y),$$

for all x and y, and $\lambda \in (0,1)$.

- A function $g(x)$ is concave if $-g(x)$ is convex.

Jensen's Inequality

Let X be a RV. If $g(x)$ is a convex function, then

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X]),$$

with equality if and only if $g(x)$ is a linear function on some set A such that $\mathbb{P}(X \in A) = 1$.

Hölder's Inequality

Let X and Y be any two RVs, and let p and q be any two positive numbers (necessarily greater than 1) satisfying

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Then,

$$|\mathbb{E}[XY]| \leq \mathbb{E}[|XY|] \leq \{\mathbb{E}[|X|^p]\}^{1/p}\{\mathbb{E}[|Y|^q]\}^{1/q}.$$

Cauchy-Schwarz Inequality ($p = q = 2$)

Let X and Y be any two RVs, then

$$|\mathbb{E}[XY]| \le \mathbb{E}[|XY|] \le \{\mathbb{E}[|X|^2]\}^{1/2}\{\mathbb{E}[|Y|^2]\}^{1/2}.$$

Liapounov's Inequality ($Y \equiv 1$)

Let X be a RV, then for any $s > r > 1$,

$$\{\mathbb{E}[|X|^r]\}^{1/r} \le \{\mathbb{E}[|X|^s]\}^{1/s}.$$

Minkowski's Inequality

Let X and Y be any two RVs. Then, for $p \geq 1$,

$$\{\mathbb{E}[|X+Y|^p]\}^{1/p} \leq \{\mathbb{E}[|X|^p]\}^{1/p} + \{\mathbb{E}[|Y|^p]\}^{1/p}.$$

- **Remark**: The preceding Hölder's Inequality (including its special cases) and Minkowski's Inequality also apply to numerical sums where there is no explicit reference to an expectation.

Consider a sequence of RVs $\{X_n : n \geq 1\}$ and another RV X.

- **Convergence Almost Surely** (a.s.), $X_n \xrightarrow{a.s.} X$:

$$\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1.$$

- **Convergence in Probability**, $X_n \xrightarrow{p} X$:

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

- **Convergence in Distribution**, $X_n \xrightarrow{d} X$ or $X_n \Rightarrow X$:

$$\lim_{n\to\infty} F_n(x) = F(x), \text{ for any continuous point } x \text{ of } F(x),$$

where F_n and F are CDF of X_n and X, respectively.

- **Convergence in L^r Norm** ($r \in [1, \infty)$), $X_n \xrightarrow{L^r} X$:

$$\lim_{n\to\infty} \mathbb{E}(|X_n - X|^r) = 0,$$

given $\mathbb{E}[|X_n|^r] < \infty$ for any $n \geq 1$ and $\mathbb{E}[|X|^r] < \infty$.

- Simple relationships:

$$
\begin{array}{ccccc}
X_n \xrightarrow{a.s.} X & \Longrightarrow & X_n \xrightarrow{p} X & \Longrightarrow & X_n \Rightarrow X \\
 & & \Uparrow & & \\
X_n \xrightarrow{L^s} X & \overset{s>r\geq 1}{\Longrightarrow} & X_n \xrightarrow{L^r} X & \Longrightarrow & \mathbb{E}[|X_n|^r] \to \mathbb{E}[|X|^r]
\end{array}
$$

- $X_n \Rightarrow$ a constant $c \quad \Longrightarrow \quad X_n \xrightarrow{p} c$.
- $X_n \xrightarrow{L^1} X \quad \Longrightarrow \quad \mathbb{E}[X_n] \to \mathbb{E}[X]$.
- $X_n \xrightarrow{a.s.} X \quad \Longleftrightarrow \quad \sup_{j\geq n} |X_j - X| \xrightarrow{p} 0$.
- $X_n \xrightarrow{p} X \quad \Longleftrightarrow$ For every subsequence $X_n(m)$ there is a further subsequence $X_n(m_k)$ such that $X_n(m_k) \xrightarrow{a.s.} X$.

- Question: If $X_n \Rightarrow X$ or $X_n \xrightarrow{p} X$ or $X_n \xrightarrow{a.s.} X$, does it imply $\mathbb{E}[X_n] \to \mathbb{E}[X]$?

Monotone Convergence Theorem (MCT)

Suppose $X_n \xrightarrow{a.s.} X$, and $0 \le X_1 \le X_2 \le \cdots$ a.s.. Then $\mathbb{E}[X_n] \to \mathbb{E}[X]$.

Fatou's Lemma

Suppose $X_n \ge Y$ a.s. for all n where $\mathbb{E}[|Y|] < \infty$. Then $\mathbb{E}[\liminf_{n\to\infty} X_n] \le \liminf_{n\to\infty} \mathbb{E}[X_n]$. In particular, if $X_n \ge 0$ a.s. for all n, then the result holds.

Dominated Convergence Theorem (DCT)

Suppose $X_n \xrightarrow{a.s.} X$, $|X_n| \leq Y$ a.s. for all n, and $\mathbb{E}[|Y|] < \infty$. Then $\mathbb{E}[X_n] \to \mathbb{E}[X]$.

- The DCT is still true if $\xrightarrow{a.s.}$ is replaced by $\xrightarrow{p}$.
- An **even more general** result:
  Suppose $X_n \xrightarrow{p} X$, $|X_n| \leq Y$ a.s. for all n, and $\mathbb{E}[|Y|^r] < \infty$ with $r \geq 1$. Then, $\mathbb{E}[|X_n|^r] < \infty$, $\mathbb{E}[|X|^r] < \infty$, and $X_n \xrightarrow{L^r} X$.

- $X = Y$ a.s., if *any one* of the following holds:
  - $X_n \xrightarrow{a.s.} X$ and $X_n \xrightarrow{a.s.} Y$;
  - $X_n \xrightarrow{p} X$ and $X_n \xrightarrow{p} Y$;
  - $X_n \xrightarrow{L^r} X$ and $X_n \xrightarrow{L^r} Y$.
- If $X_n \xrightarrow{a.s.} X$ and $Y_n \xrightarrow{a.s.} Y$, then $(X_n, Y_n)^\intercal \xrightarrow{a.s.} (X, Y)^\intercal$.
  $\Longrightarrow aX_n + bY_n \xrightarrow{a.s.} aX + bY$; $X_nY_n \xrightarrow{a.s.} XY$. (Due to CMT)
- If $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$, then $(X_n, Y_n)^\intercal \xrightarrow{p} (X, Y)^\intercal$.
  $\Longrightarrow aX_n + bY_n \xrightarrow{p} aX + bY$; $X_nY_n \xrightarrow{p} XY$. (Due to CMT)
- If $X_n \xrightarrow{L^r} X$ and $Y_n \xrightarrow{L^r} Y$, then $(X_n, Y_n)^\intercal \xrightarrow{L^r} (X, Y)^\intercal$.
  $\Longrightarrow aX_n + bY_n \xrightarrow{L^r} aX + bY$.
- None of the above are true for convergence in distribution.
- If $X_n \Rightarrow X$ and $Y_n \Rightarrow$ **constant** c, then $(X_n, Y_n)^\intercal \Rightarrow (X, c)^\intercal$.
  $\Longrightarrow aX_n + bY_n \Rightarrow aX + bc$; $X_nY_n \Rightarrow cX$. (Due to CMT; also known as Slutsky's theorem)

Continuous Mapping Theorem (CMT)

Consider a sequence of RVs $\{X_n : n \geq 1\}$ and another RV X. Suppose g is a function that has the set of discontinuity points D such that $\mathbb{P}(X \in D) = 0$. Then,

$$\begin{aligned} X_n \xrightarrow{a.s.} X \quad &\Longrightarrow \quad g(X_n) \xrightarrow{a.s.} g(X); \\ X_n \xrightarrow{p} X \quad &\Longrightarrow \quad g(X_n) \xrightarrow{p} g(X); \\ X_n \Rightarrow X \quad &\Longrightarrow \quad g(X_n) \Rightarrow g(X). \end{aligned}$$

- CMT also holds for **random vectors**.
- Caution: For convergence in L^r norm, stronger assumption of g than continuity is required to ensure $g(X_n) \xrightarrow{L^r} g(X)$.

- Let $X_1, \dots, X_n$ be a **random sample** from a distribution with mean μ and variance σ^2, i.e., $X_1, \dots, X_n$ are iid, and $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$, $i = 1, \dots, n$.
- Define

$$\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i, \text{ and } S^2 := \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}.$$

- For a **general** distribution, the following is true:
  1. $\bar{X}$ is an **unbiased** estimator of μ, i.e., $\mathbb{E}[\bar{X}] = \mu$;
  2. S^2 is an **unbiased** estimator of σ^2, i.e, $\mathbb{E}[S^2] = \sigma^2$;
  3. $\mathrm{Var}(\bar{X}) = \sigma^2/n$.
- If the distribution is $\mathcal{N}(\mu, \sigma^2)$, we further have:
  4. $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$, i.e., $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$;
  5. $\bar{X} \perp S^2$;
  6. $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$;
  7. $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$.

- For a **general** distribution, *what can we say about the distribution of* $\bar{X}$?
- $\mathrm{Var}(\bar{X}) = \sigma^2/n$ intuitively means that the randomness of $\bar{X}$ vanishes and $\bar{X}$ concentrates around μ when n gets large.
- Denote $\bar{X}$ as $\bar{X}_n$, to explicitly indicate the effect of **sample size** n.

Weak Law of Large Numbers (WLLN)

Suppose $X_1, \ldots, X_n$ are iid with mean μ and variance $\sigma^2 < \infty$.[†] Then, $\bar{X}_n \xrightarrow{p} \mu$.

Strong Law of Large Numbers (SLLN)

Suppose $X_1, \ldots, X_n$ are iid with mean μ and variance $\sigma^2 < \infty$.[†] Then, $\bar{X}_n \xrightarrow{a.s.} \mu$.

[†] Mutual independence can be weakened to pairwise independence; $\sigma^2 < \infty$ can be weakened to $\mathbb{E}[|X_i|] \leq \infty$.

- Note that for **normal** distribution, $\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$, regardless of the value of n.
- For a **general** distribution, *what can we say about the distribution of* $\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}$?
- Note that $\mathbb{E}\left[\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}\right] = 0$ and $\mathrm{Var}\left(\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}\right) = 1$, regardless of the distribution and the value of n.

Central Limit Theorem (CLT)

Suppose $X_1, \ldots, X_n$ are iid with mean μ and variance $\sigma^2 \in (0, \infty)$. Then,

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \Rightarrow \mathcal{N}(0,1).$$