機率導論: 下半學期筆記

陳 捷　B06303009　經濟三

2021-Apr-15

1 Covariance and Correlation

Definition 1 (Covariance)**.** *Assume that X and Y are two random variables. Then the covariance of X and Y,*

$$\operatorname{Cov}(x, y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]\,.$$

We say that X and Y are uncorrelated if $\operatorname{Cov}(X, Y) = 0$.

Remark 1. *If X and Y are independent, then* $\operatorname{Cov}(X, Y) = 0$. *The converse is not true.*

Example.

$$\Pr\{X = 1\} = \Pr\{X = 0\} = \Pr\{X = -1\} = \frac{1}{3} \qquad Y = \begin{cases} 0 & \text{if } X \neq 0 \\ 1 & \text{if } X = 0 \end{cases}$$

We have that X and Y are not independent but $\operatorname{Cov}(X, Y) = 0$.

1.1 Properties of Covariance

1. $\operatorname{Cov}(X, Y) = \operatorname{Cov}(Y, X)$
2. $\operatorname{Cov}(aX, Y) = a \operatorname{Cov}(X, Y)$
3. Covariance decomposition:

$$\begin{aligned} \operatorname{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) &= \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{m} X_i Y_i\right] - \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{E}[X_i]\,\mathbb{E}[Y_j] \\ &= \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{m} (X_i - \mathbb{E}[X_i])\,(Y_i - \mathbb{E}[Y_j])\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbb{E}[(X_i - \mathbb{E}[X_i])\,(Y_j - \mathbb{E}[Y_j])] = \sum_{i=1}^{n}\sum_{j=1}^{m} \operatorname{Cov}(X_i, Y_j) \end{aligned}$$

4. Variance decomposition:

$$\begin{aligned} \operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \mathbb{E}\left[\left(\sum_{i=1}^{n} (X_i - \mathbb{E}[X_i])\right)^2\right] &= \mathbb{E}\left[\sum_{i=1}^{n} (X_i - \mathbb{E}[X_i]) \cdot \sum_{j=1}^{n} (X_j - \mathbb{E}[X_j])\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \operatorname{Cov}(X_i, X_j) \\ &= \sum_{i=1}^{n} \operatorname{Cov}(X_i, X_i) + \sum_{i \neq j} \operatorname{Cov}(X_i, X_j) \\ &= \sum_{i=1}^{n} \operatorname{Var}(X_i) + 2\sum_{i<j} \operatorname{Cov}(X_i, X_j) \end{aligned}$$

1.2 Correlation

Definition 2 (Correlation)**.**

$$\begin{aligned}\rho(X,Y) &= \textit{correlation coefficient of } X \textit{ and } Y \\ &= \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)}\sqrt{\operatorname{Var}(X)}} \\ &= \operatorname{Cov}\left(\frac{X}{\sqrt{\operatorname{Var}(X)}}, \frac{Y}{\sqrt{\operatorname{Var}(Y)}}\right), \qquad \operatorname{Var}\left(\frac{X}{\sqrt{\operatorname{Var}(X)}}\right) = \frac{\operatorname{Var}(X)}{\operatorname{Var}(X)} = 1.\end{aligned}$$

Remark 2. *We have the following properties of correlation:*

1. *Notice that we have $-1 \le \rho(X,Y) \le 1$ (Cauchy-Schwarz inequality).*

$$\begin{aligned}|\,\rho(X,Y)\,| &= \left|\operatorname{Cov}\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right)\right| \\ &= \left|\mathbb{E}\left[\left(\frac{X}{\sigma_X} - \frac{\mathbb{E}[X]}{\sigma_X}\right)\cdot\left(\frac{Y}{\sigma_Y} - \frac{\mathbb{E}[Y]}{\sigma_Y}\right)\right]\right| \\ &\le \left|\mathbb{E}\left[\left(\frac{X}{\sigma_X} - \frac{\mathbb{E}[X]}{\sigma_X}\right)^2\right]\right|^{1/2} \left|\mathbb{E}\left[\left(\frac{Y}{\sigma_Y} - \frac{\mathbb{E}[Y]}{\sigma_Y}\right)^2\right]\right|^{1/2} = 1.\end{aligned}$$

2. *$\rho(X,Y)$ means the linearity between X and Y. If $\rho(X,Y) = 0$, then X and Y are uncorrelated. If $\rho(X,Y) = 1$, i.e., $\exists a > 0$ s.t.*

$$\left(\frac{X}{\sigma_X} - \frac{\mathbb{E}[X]}{\sigma_X}\right) = a\left(\frac{Y}{\sigma_Y} - \frac{\mathbb{E}[Y]}{\sigma_Y}\right) \implies X = \underbrace{a\frac{\sigma_X}{\sigma_Y}}_{\tilde{a}} Y + b,$$

X and Y are linearly correlated. If $\rho(X,Y) = -1$, i.e., $\exists \tilde{a} < 0$ s.t. $X = \tilde{a}Y + \tilde{b}$, X and Y are linearly correlated.

Definition 3. *If $\rho(X,Y) > 0$, then X and Y are positively correlated; if $\rho(X,Y) < 0$, then X and Y are negatively correlated.*

Example. Let A and B be two event. Define χ_A and χ_B as

$$\chi_A = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise} \end{cases} \qquad \chi_B = \begin{cases} 1 & \text{if } \omega \in B \\ 0 & \text{otherwise} \end{cases}.$$

Then we have

$$\mathbb{E}[\chi_A] = \Pr\{A\}\left(= \int_A f(x)dx\right) \qquad \mathbb{E}[\chi_B] = \Pr\{B\}\left(= \int_B f(x)dx\right).$$

That is, let $(\Omega, \mathscr{F}, \Pr)$ be a probability space. A random variable $X : \Omega \to \mathbb{R}$ is Borel measurable function. Then $\mathbb{E}[X] = \int X d\Pr$. If $X = \chi_A$, $A \in \mathscr{F}$, then $\mathbb{E}[\chi_A] = \int \chi_A d\Pr = \int_A d\Pr = \Pr\{A\}$. Similarly, we have

$$\begin{aligned}\mathbb{E}[\chi_A\chi_B] &= \Pr\{A \cap B\} \\ \operatorname{Cov}(\chi_A, \chi_B) &= \mathbb{E}[\chi_A\chi_B] - \mathbb{E}[\chi_A]\,\mathbb{E}[\chi_B] \\ &= \Pr\{A\cap B\} - \Pr\{A\}\Pr\{B\} \\ &= \Pr\{B\}\left[\frac{\Pr\{A\cap B\}}{\Pr\{B\}} - \Pr\{A\}\right] && \text{(if } \Pr\{B\} \neq 0) \\ &= \Pr\{B\}\left[\Pr\{A \mid B\} - \Pr\{A\}\right]\end{aligned}$$

Observe that

1. If $\Pr\{A \mid B\} = \Pr\{A\}$, then $\operatorname{Cov}(\chi_a, \chi_B) = 0$.
2. If $\Pr\{A \mid B\} > \Pr\{A\} \iff \chi_A$ and χ_B are positively correlated.
3. If $\Pr\{A \mid B\} < \Pr\{A\} \iff \chi_A$ and χ_B are negatively correlated.

2 Conditional Density Functions and Conditional Expectation

Recall: Conditional probability $\Pr\{A \mid B\} = \Pr\{A \cap Bk\} / \Pr\{B\}$. Given an $B \in \mathscr{F}$, $\Pr\{\cdot \mid B\}$ is a probability measure on $(\Omega, \mathscr{F})$.

If X and Y are random variables, then

$$\Pr\{X \in A \mid Y \in B\} = \frac{\Pr\{X \in A, Y \in B\}}{\Pr\{Y \in B\}}.$$

Given $Y \in B$, then $\Pr\{X \in A \mid Y \in B\} = \Pr_{\{Y \in B\}}\{X \in A\}$. That is, we can define a new measure on the probability space.

Consider X and Y be two discrete random variables with value x and y. Assume that the joint probability mass function $p(x, y)$, then

$$\Pr\{X = x \mid Y = y\} = \frac{\Pr\{X = x, Y = y\}}{\Pr\{Y = y\}} = \frac{p(x, y)}{p_Y(y)}.$$

where $p_Y(y)$ is the marginal probability mass function.

Definition 4 (Conditional pmf)**.** *The conditional probability mass of X given $Y = y$ denoted by $\Pr_{XgivenY}\{x \mid y\}$.*

$$\Pr{}_{XgivenY}\{x \mid y\} = \frac{p(x, y)}{p_Y(y)} = \frac{p(x, y)}{\sum_x p(x, y)} > 0.$$

Note that $\sum_x \Pr_{XgivenY}\{x \mid y\} = 1$.

Definition 5 (Conditional distribution and expectation)**.** *Conditional distribution of X given $Y = y$ is*

$$F_{XgivenY}(xgiveny) = \Pr{}_{XgivenY}\{X \leq x \mid Y = y\}.$$

Conditional expectation of X given $Y = y$ is

$$\mathbb{E}[XgivenY]X \mid y = \sum_x \Pr{}_{XgivenY}\{x \mid y\} = \sum_x \frac{xp(x, y)}{p_Y(y)}.$$

Notice that conditional expectation is also a random variables of Y.

Example. $X \sim \text{Poisson}(\lambda_1)$, $Y \sim \text{Poisson}(\lambda_2)$ where X and Y are independent. Compute the conditional mass function of X given $X + Y = n$.

Solution: $X + Y \sim \text{Poisson}(\lambda_1 + \lambda_2)$, i.e.,

$$\Pr\{X + Y = n\} = \frac{(\lambda_1 + \lambda_2)^n e^{-(\lambda_1 + \lambda_2)}}{n!}.$$

Given $0 \le k \le n$,

$$\begin{aligned}
\Pr\{X = k \mid X + Y = n\} &= \frac{\Pr\{X = k, X + Y = n\}}{\Pr\{X + Y = n\}} \\
&= \frac{\Pr\{X = k, Y = n - k\}}{\Pr\{X + Y = n\}} \\
&= \frac{\Pr\{X = k\} \Pr\{Y = n - k\}}{\Pr\{X + Y = n\}} \\
&= \frac{\left(\frac{\lambda_1^k e^{-\lambda_1}}{k!}\right)\left(\frac{\lambda_2^{n-k} e^{-\lambda_2}}{(n-k)!}\right)}{\frac{(\lambda_1+\lambda_2)^n e^{-(\lambda_1+\lambda_2)}}{n!}} \\
&= \frac{n!}{k!(n-k)!}\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^k \left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{n-k} \sim \text{Binomial.}
\end{aligned}$$

Remark 3. *Consider the following:*

$$\Pr_X\{X \in A\} = \sum_{x \in A} \sum_y p(x,y) = \sum_{x \in A} \sum_y \frac{p(x,y)}{p_Y(y)} p_Y(y) = \sum_y \sum_{x \in A} \Pr_{XgivenY}\{()\, x, y) p_Y(y)$$

This is the "Theorem of total probability". Recall: Theorem of total probability: $\{B_j\}$ *events with* $B_i \cap B_j = \varnothing$ *for* $i \neq j$ *and* $\Omega = \bigcup_j B_j$. *Then* $\Pr\{A\} = \sum_j \Pr\{A \mid B_j\} \Pr\{B_j\}$.

2.1 Conditional Density

Let (X, Y) be a pair of random variables with joint pdf $f(x, y)$. The conditional probability of $X \in A$ given $\{Y = y\}$ is

$$\Pr_{XgivenY}\{X \in A \mid Y = y\} = \frac{\Pr\{X \in A, Y = y\}}{f_Y(y)} = \frac{\int_{x \in A} f(x,y)dx}{\int_{-\infty}^{\infty} f(x,y)dx}.$$

Therefore, the conditional density of X given $Y = y$ is

$$f_{XgivenY}(xgiveny) = \frac{f(x,y)}{f_Y(y)}.$$

Definition 6 (Conditional Density)**.** *Let X and Y be absolutely continuous random variables with joint pdf $f(x, y)$. The conditional density of X given $Y = y$ is*

$$f_{XgivenY}(xgiveny) = \frac{f(x,y)}{f_Y(y)}.$$

Remark 4. *On conditional density: (We are concerned about dividing by $f_Y(y)$.)*
 (1) Compute the conditional probability of $X \in A$ given the event $\{y_0 - h \le Y < y_0 + h\}$.

$$\begin{aligned}
\Pr\{X \in A \mid y_0 - h < Y \le y_0 + h\} &= \frac{\int_A \int_{y_0-h}^{y_0+h} f(x,y)dydx}{\int_{y_0-h}^{y_0+h} f_Y(y)dy} \\
&\approx \frac{2h \int_A f(x,y_0)dx}{2h f_Y(y_0)} && \text{(let } h \to 0\text{)} \\
&\to \frac{f(x,y_0)}{f_Y(y)}
\end{aligned}$$

(2) $f_{X givenY}(x, y)$ *only defines on the set where* $f_Y(y) > 0$. *Consider* $S = \{(x, y) : f_Y(y) = 0\}$:

$$\Pr\{(X,Y) \in S\} = \iint\limits_{(x,y)\in S} f(x,y)dxdy = \int\limits_{\{(x,y):f_Y(y)=0\}} f(x,y)dxdy$$
$$= \int\limits_{\{f_Y(x)=0\}} f_Y(y)dy = 0.$$

(3) ***Theorem of total probability***: *Compute* $\{X \in A\}$.

$$\Pr\{X \in A\} = \int_{x\in A}\int_{-\infty}^{\infty} f(x,y)dxdy = \int_{a\in A}\int_{-\infty}^{\infty} \frac{f(x,y)}{f_Y(y)}dyf_Y(y)dx$$
$$= \int_{-\infty}^{\infty}\left[\int_{x\in A}\frac{f(x,y)dx}{f_Y(y)}\right]f_Y(y)dy$$
$$= \int_{-\infty}^{\infty} \Pr\{X \in A \mid Y = y\}\, f_Y(y)dy.$$

Example. (Student's t-test.) Let $Z \sim \mathcal{N}(()0,1)$ and $Y \sim \chi^2(n)$ where Z and Y are independent. Define $T = \frac{Z}{\sqrt{Y/n}} = \sqrt{n}\frac{Z}{\sqrt{Y}}$. We want to find the pdf $f_T(t)$.

Sol:

$$f_T(t) = \int_{-\infty}^{\infty} f_{Y,T}(y,t)dy = \int_{-\infty}^{\infty}\Big[f_{T givenY}(tgiveny)f_Y(y)\Big]dy \qquad \text{(by definition above)}$$

Recall that we have

$$f_Y(y) = \frac{e^{-y/2}y^{n/2-1}}{2^{n/2}\Gamma(n/2)} \quad \text{and} \quad f_{T givenY}(tgiveny) = \frac{1}{\sqrt{2\pi n/y}}e^{-(t^2y)/(2n)}.$$
(scaling of normal distribution)

Therefore, we have

$$f_{Y,T}(y,t) = \frac{1}{\sqrt{\pi n}2^{(n+1)/2}\Gamma(\frac{n}{2})}e^{-\frac{t^2+n}{2n}y}y^{(n-1)/2} \qquad (y \geq 0)$$
$$\Longrightarrow f_T(t) = \int_0^{\infty} f_{Y,T}(y,t)dy = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}\left(1+\frac{t^2}{n}\right)^{-(n+1)/2}. \qquad (-\infty < t < \infty)$$

Example. Two buses A and B. Arrival time are the station $R_1 \sim \mathcal{U}(0, t_A)$ and $R_2 \sim \mathcal{U}(0, t_B)$ where $t_A < t_B$. R_1 and R_2 are independent. Find the probability of $R_1 < R_2$.

Sol:

(1) (Direct method) The join pdf of (R_1, R_2) is

$$f_{R_1,R_2}(x,y) = \frac{1}{t_At_B}\chi_{(0,t_A)}\chi_{(0,t_B)}.$$

Therefore,

$$\Pr\{R_1 < R_2\} = \int\limits_{x<y} \frac{1}{t_At_B}\chi_{(0,t_A)}\chi_{(0,t_B)}dxdy = 1 - \frac{t_A}{2t_B}.$$

(2) (Conditioning)

$$f_{R_1 givenR_2}(ygivenx) = \frac{f_{R_1,R_2}(x,y)}{f_{R_1}(x)} = \frac{f_{R_1}(x)f_{R_2}(y)}{f_{R_1}(x)} = f_{R_2}(y). \qquad \text{(independence)}$$

By theorem of total probability, we have

$$\Pr\{R_1 < R_2\} = \int_0^{t_A} \Pr\{x < R_2 \mid R_1 = x\}\, f_{R_1}(x)dx = \int_0^{t_A}\frac{1}{t_B}(t_B - x)\frac{1}{t_A}dx = 1 - \frac{t_A}{2t_B}.$$

2.2 Conditional Expectation

Let (X, Y) have the joint pdf $f(x, y)$. The conditional pdf of X given $Y = y$ is

$$f_{XgivenY}(xgiveny) = \frac{f(x,y)}{f_Y(y)}.$$

Therefore, the conditional distribution function of X given $Y = y$ is

$$F_{XgivenY}(xgiveny) = \Pr\{X \leq x \mid Y = y\} = \int_{-\infty}^{x} f_{XgivenY}(zgiveny)dz.$$

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

$$\mathbb{E}[X \mid Y = y] = \int_{-\infty}^{\infty} x f_{XgivenY}(xgiveny)dx = \int_{-\infty}^{\infty} \frac{xf(x,y)}{f_Y(y)}dx.$$

Observation: $\mathbb{E}[X \mid Y = y]$ is a function of y. Hence, $\mathbb{E}[X \mid Y] = g(Y)$ for some $g(\cdot)$. Therefore, $\mathbb{E}[X \mid Y]$ is a random variable!

Remark 5. *(On conditional expectation)*

1. $\mathbb{E}[X \mid Y]$ *is a random variable of* Y*, thus, it is* Y*-measurable. That is, it is measurable by* Y*.*
2. $\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X]$*. The interpretation of this is that additional information does not alter the original expectation.*

   ▶*Proof.* Compute

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[X \mid Y]] &= \int \mathbb{E}[X \mid Y = y]\, f_Y(y)dy && \text{(theorem of total expectation)} \\
&= \iint x f_{XgivenY}(xgiveny)dx f_Y(y)dy \\
&= \iint x \frac{f(x,y)}{f_Y(y)} f_Y(y)dxdy \\
&= \int x f_X(x)dx = \mathbb{E}[X]\,.
\end{aligned}$$

■

Example. N = number of customers entering a store with $\mathbb{E}[N] = 50$. $X_1, X_2, \ldots$ = amount of money spent by those customers iid with $\mathbb{E}[X_i] = 8$. N and X_i are independent. Let $R = \sum_{i=1}^{N} X_i$. We want to find $\mathbb{E}[R]$.

Sol:

$$\mathbb{E}[R] = \mathbb{E}[\mathbb{E}[RgivenN]]$$

$$\text{where} \quad \mathbb{E}[R \mid N = n] = \mathbb{E}\left[\sum_{i=1}^{N} X_i \mid N = n\right] = \mathbb{E}\left[\sum_{i=1}^{n} X_i \mid N = n\right] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = n\,\mathbb{E}[X_i]\,.$$

Hence,

$$\mathbb{E}[R \mid N] = N\,\mathbb{E}[X_i] = 18N \implies \mathbb{E}[R] = \mathbb{E}[18N] = 18 \times 50.$$

2.3 General Conditional Expectation

$\mathbb{E}[X \mid Y]$ is a random variable of Y, i.e., we can denote $\mathbb{E}[X \mid Y] = g(Y)$. Consider $(\Omega, \mathscr{F}, \Pr) \xrightarrow{Y} (\mathbb{R}, \mathcal{B})$ where $\mathcal{B}$ is a Boreal field. We have

$$Y^{-1}(\mathcal{B}) \in \mathscr{F} \implies \sigma(Y^{-1}(\mathcal{B})) \subseteq \mathscr{F}.$$

Condition X on $\sigma(f^{-1}(\mathcal{B}))$: $\mathbb{E}[X \mid Y]$

Definition 7 (Conditional Expectation)**.** $(\Omega, \mathscr{F}, \Pr)$: *probability space. Consider* $\mathscr{F}_0 \subseteq \mathscr{F}$ *is a* σ*-field (could be smaller). Assume that* $\mathbb{E}[|X|] < \infty$*. We say that* $\mathbb{E}[X|\mathscr{F}_0]$ *is a* version *of the conditional expectation of* X *on* $\mathscr{F}_0$ *if it satisfies*

(i) $\mathbb{E}[X|\mathscr{F}_0]$ *is* $\mathscr{F}_0$*-measurable, i.e.,* $\forall B \in \mathcal{B}$,

$$\Big(\mathbb{E}[(X|\mathscr{F}_0)]\Big)^{-1}(B) \in \mathscr{F}_0.$$

Note that X *is* $\mathscr{F}$*-measurable, i.e.,* $\forall B \in \mathcal{B}$, $X^{-1}(B) \in \mathscr{F}$.

(ii) $\forall A \in \mathscr{F}_0 (\subseteq \mathscr{F})$*, we have*

$$\int_A X d\Pr = \mathbb{E}[\chi_A X] = \int_A \mathbb{E}[X|\mathscr{F}_0]\, d\Pr.$$

Lemma 1 (Uniqueness)**.** *There exists at most "one" conditional expectation ("one" up to to a set of probability zero).* [i]

▶*Proof.* Let Y and $\tilde{Y}$ be two versions of $\mathbb{E}[X|\mathscr{F}_0]$. Y and $\tilde{Y}$ are $\mathscr{F}_0$-measurable. By (ii), $\forall A \in \mathscr{F}_0$ we have

$$\int_A Y d\Pr = \int_A X d\Pr = \int_A \tilde{Y} d\Pr \implies \int_A \left(Y - \tilde{Y}\right) d\Pr = 0,$$

that is, $\Pr\Big\{\omega \in \Omega : (Y - \tilde{Y}) \neq 0\Big\} = 0$. ■

Lemma 2 (Existence)**.** *Consider* $X \geq 0$ *(For general* X*, we write* $X = X^+ - X^-$*). Define*

$$\nu(A) = \int_A X d\Pr \quad \forall A \in \mathscr{F}_0.$$

Observe that ν *is absolutely continuous with respect to* $\Pr$*. By the Radom-Nikodym theorem,* $\exists Y \geq 0$ *s.t.*

$$\begin{aligned} \nu(A) &= \int_A Y d\Pr && \text{(where } Y \text{ is } \mathscr{F}_0\text{-measurable)} \\ &= \int_A X d\Pr \implies Y = \mathbb{E}[X|\mathscr{F}_0]. \end{aligned}$$ □

For general $X = X^+ - X^-$*. We have* Y^+ *and* Y^- *from above are* $\mathscr{F}_0$*-measurable. Therefore,*

$$\int_A X^{\pm} d\Pr = \int_A Y^{\pm} d\Pr \implies \int_A X d\Pr = \int_A X^+ d\Pr - \int_A X^- d\Pr = \int_A (Y^+ - Y^-) d\Pr.$$

Example. Of conditional expectation:

1. If X itself is $\mathscr{F}_0$-measurable, then $X = \mathbb{E}[X|\mathscr{F}_0]$. (easily checkable from definition)

[i] If two random variables only differ on a set of measure zero, then we say that the two random variables are the same.

2. If X is independent of $\mathscr{F}_0$, i.e., $\forall B \in \mathcal{B}$ and $A \in \mathscr{F}_0$,

$$\Pr\{\{X \in B\} \cap A\} = \Pr\{\{X \in B\}\} \Pr\{A\} \quad \forall A \in \mathscr{F}_0$$

then $\mathbb{E}[X|\mathscr{F}_0] = \mathbb{E}[X]$. Check: (i) $\mathbb{E}[X]$ is $\mathscr{F}_0$-measurable; (ii) $\forall A \in \mathscr{F}_0$, [ii]

$$\int_A X d\Pr = \int \chi_A X d\Pr = \mathbb{E}[\chi_A X] = \mathbb{E}[X] \Pr\{A\}.$$

3. Let $\Omega = \bigcup_{i=1}^n \Omega_i$ where n can be infinite and $\Omega_i \cap \Omega_j = \varnothing \ \forall i \neq j$. Let $\mathscr{F}_0 = \sigma(\Omega_1, ..., \Omega_n)$. Then,

$$\mathbb{E}[X|\mathscr{F}_0] = \text{constant on } \Omega_i \ \forall i = \frac{\mathbb{E}[\chi_\Omega X]}{\Pr\{\Omega_i\}}.$$

Check: (i) $\mathbb{E}[X \mathscr{F}_0] = \text{constant on } \Omega_i \implies \mathbb{E}[X|\mathscr{F}_0]$ is $\mathscr{F}_0$-measurable. (ii) It suffices to let $A = \Omega_i$ for some i. Then,

$$\int_{\Omega_i} X d\Pr = \mathbb{E}[\chi_\Omega X] = \frac{\chi_{\Omega_i} X}{\Pr\{\Omega_i\}} \Pr\{\Omega_i\} = \int_{\Omega_i} \frac{\chi_{\Omega_i} X}{\Pr\{\Omega_i\}} d\Pr.$$

Extreme case: $\mathscr{F}_0 = \{\varnothing, \Omega\}$, then $\mathbb{E}[X|\mathscr{F}_0] = \mathbb{E}[X]$.

Proposition 1. *Let $\mathscr{F}_1 \subset \mathscr{F}_2 \subset \mathscr{F}$, then*

$$\mathbb{E}[\mathbb{E}[X|\mathscr{F}_1] \,|\mathscr{F}_2] = \mathbb{E}[X|\mathscr{F}_1] \tag{1}$$
$$\mathbb{E}[\mathbb{E}[X|\mathscr{F}_2] \,|\mathscr{F}_1] = \mathbb{E}[X|\mathscr{F}_1] \tag{2}$$

►*Proof.*

(i) Observe that $\mathbb{E}[X|\mathscr{F}_1]$ is $\mathscr{F}_1$-measurable $\implies \mathscr{F}_2$-measurable. $\implies \mathbb{E}[\mathbb{E}[X|\mathscr{F}_1] \,|\mathscr{F}_2] = \mathbb{E}[X|\mathscr{F}_1]$.

(ii) $\mathbb{E}[X|\mathscr{F}_1]$ is $\mathscr{F}_1$-measurable. $\forall A \in \mathscr{F}_1$, we have

$$\int_A \mathbb{E}[\mathbb{E}[X|\mathscr{F}_2] \,|\mathscr{F}_1] \, d\Pr = \int_A \mathbb{E}[X|\mathscr{F}_2] \, d\Pr = \int_A X d\Pr = \int_A \mathbb{E}[X|\mathscr{F}_1] \, d\Pr.$$

■

Proposition 2. *Let Y be $\mathscr{F}_0$-measurable, then $E[YX|\mathscr{F}_0] = Y\,\mathbb{E}[X|\mathscr{F}_0]$.*

►*Proof.* $Y = \chi_B$ and $B \in \mathscr{F}_0$. Then

$$\int_A \mathbb{E}[\chi_B X|\mathscr{F}_0] \, d\Pr = \int_A \chi_B X d\Pr = \int_{A\cap B} X d\Pr = \int_{A\cap B} \mathbb{E}[X|\mathscr{F}_0] \, d\Pr = \int_A \chi_B \, \mathbb{E}[X|\mathscr{F}_0] \, d\Pr.$$

■

2.4 Another Characterisation of Conditional Expectation

Motivation: What is $\mathbb{E}[X]$? $\mathbb{E}[X]$ is the minimiser of $\mathbb{E}\left[(X-b)^2\right]$ for any $b \in \mathbb{R}$:

$$\mathbb{E}\left[(X-b)^2\right] = \mathbb{E}\left[(X - \mathbb{E}[X] + \mathbb{E}[X] - b)^2\right] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] + \mathbb{E}\left[(\mathbb{E}[X] - b)^2\right].$$

Theorem 1. *Let $\mathbb{E}\left[X^2\right] < \infty$, then $\mathbb{E}[X|\mathscr{F}_0]$ is the minimiser of the mean square error $\mathbb{E}\left[(X-Y)^2\right]$ $\forall Y$ is $\mathscr{F}_0$-measurable.*

[ii]Observe that if $X = \chi_B$, $B \in \mathscr{F}$, then $\int_A \chi_B d\Pr = \int_A \chi_A \chi_B d\Pr = \int_{A\cap B} d\Pr = \Pr\{A \cap B\} = \Pr\{B\} \Pr\{B\} = \mathbb{E}[\chi_B] \Pr\{A\}$.

Theorem 2. *Let $\mathscr{F}_0 \subset \mathscr{F}$ be a σ-field. For any X with $\mathbb{E}\left[|X|^2\right] \leq \infty$, then $\mathbb{E}[X|\mathscr{F}_0]$ is the minimiser of*

$$\mathbb{E}\left[(X-Y)^2\right] \quad \text{where } Y \text{ is } L^2(\mathscr{F}_0) \text{ variable.}$$

where $L^2(\mathscr{F}_0) = \left\{\mathbb{E}\left[|Z|^2\right] < \infty \text{ and } Z \text{ is } \mathscr{F}_0\text{-measurable}\right\}$.
Remark 6. *Why is $\mathbb{E}[X|\mathscr{F}_0]$ $L^2(\mathscr{F}_0)$?*

$$\mathbb{E}\left[|\mathbb{E}[X|\mathscr{F}_0]|^2\right] < \mathbb{E}\left[|X|^2\right] < \infty$$

This is called Jensen's inequality.
►*Proof.* (of Jensen's inequality) Consider the following:

$$\begin{aligned}
\mathbb{E}\left[(X-Y)^2\right] &= \mathbb{E}\left[(X-\mathbb{E}[X|\mathscr{F}_0]+\mathbb{E}[X|\mathscr{F}_0]-Y)^2\right] \\
&= \mathbb{E}\left[(X-\mathbb{E}[X|\mathscr{F}_0]+Z)^2\right] && (Z=\mathbb{E}[X|\mathscr{F}_0]-Y) \\
&= \mathbb{E}\left[(X-\mathbb{E}[X|\mathscr{F}_0])^2\right] + 2\,\mathbb{E}[ZX - Z\,\mathbb{E}[X|\mathscr{F}_0]] + \mathbb{E}\left[Z^2\right].
\end{aligned}$$

<u>Claim</u>: $\mathbb{E}[ZX] = \mathbb{E}[Z\,\mathbb{E}[X|\mathscr{F}_0]]$. Since Z is $\mathscr{F}_0$-measurable, the result is clear. □
Therefore, we have

$$\begin{aligned}
\mathbb{E}\left[(X-Y)^2\right] &= \mathbb{E}\left[(X-\mathbb{E}[X|\mathscr{F}_0])^2\right] + \mathbb{E}\left[Z^2\right] \\
&= \mathbb{E}\left[(X-\mathbb{E}[X|\mathscr{F}_0])^2\right] + \mathbb{E}\left[(\mathbb{E}[X|\mathscr{F}_0]-Y)^2\right] \geq \mathbb{E}\left[(X-\mathbb{E}[X|\mathscr{F}_0])^2\right]
\end{aligned}$$

where the equality holds if $\mathbb{E}[X|\mathscr{F}_0] = Y$. ■
Definition 8 (Conditional Variance)**.**

$$\begin{aligned}
\operatorname{Var}(X|Y) &\equiv \mathbb{E}\left[(X-\mathbb{E}[X|Y])^2\middle|Y\right] \\
&= \mathbb{E}\left[X^2 - 2X\,\mathbb{E}[X|Y] + (\mathbb{E}[X|Y])^2\middle|Y\right] \\
&= \mathbb{E}\left[X^2|Y\right] - 2\,\mathbb{E}[X\,\mathbb{E}[X|Y]\,|Y] + \mathbb{E}\left[(\mathbb{E}[X|Y])^2\middle|Y\right].
\end{aligned}$$

Remark 7. *Observation:*
(i) $\mathbb{E}[X\,\mathbb{E}[X|Y]\,|Y] = \mathbb{E}[X|Y]^2$ *since $\mathbb{E}[X|Y]$ is Y-measurable.*
(ii) $\mathbb{E}\left[(\mathbb{E}[X|Y])^2|Y\right] = \mathbb{E}[X|Y]^2$.
Hence, we have

$$\begin{aligned}
\operatorname{Var}(X|Y) &= \mathbb{E}\left[X^2|Y\right] - 2\,\mathbb{E}[X\,\mathbb{E}[X|Y]\,|Y] + \mathbb{E}\left[\mathbb{E}[X|Y]^2\middle|Y\right] \\
&= \mathbb{E}\left[X^2|Y\right] - \mathbb{E}[X|Y]^2.
\end{aligned}$$

And since on the right-hand side it is clear that $\operatorname{Var}(X|Y)$ is Y-measurable, we take the expectation on both sides:

$$\mathbb{E}[\operatorname{Var}(X|Y)] = \mathbb{E}\left[\mathbb{E}\left[X^2|Y\right] - \mathbb{E}[X|Y]^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[\mathbb{E}[X|Y]^2\right].$$

Similarly consider its variance:

$$\begin{aligned}
\operatorname{Var}(\mathbb{E}[X|Y]) &= \mathbb{E}\left[\mathbb{E}[X|Y]^2\right] - \mathbb{E}[\mathbb{E}[X|Y]]^2 && \text{(by definition of variance)} \\
&= \mathbb{E}\left[\mathbb{E}[X|Y]^2\right] = \mathbb{E}[X]^2.
\end{aligned}$$

Therefore, we have

$$\mathbb{E}[\operatorname{Var}(X|Y)] + \operatorname{Var}(\mathbb{E}[X|Y]) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \operatorname{Var}(X).$$

Example. Let $N \in \mathbb{N}$ is a random variable. $\{X_1, X_1, ...\}$ are iid with $\mathbb{E}[X_i] = \mathbb{E}[X]$. Assume that $\{X_1, ..., X_N, N\}$ are independent. Compute $\operatorname{Var}\left(\sum_{i=1}^{N} X_i\right)$.

$$\operatorname{Var}\left(\sum_{i=1}^{N} X_i\right) = \mathbb{E}\left[\operatorname{Var}\left(\sum_{i=1}^{N} X_i \middle| N\right)\right] + \operatorname{Var}\left(\mathbb{E}\left[\sum_{i=1}^{N} X_i \middle| N\right]\right)$$

Observe that since $\{X_1, ..., X_N, N\}$ are independent, we have

$$\begin{aligned} \mathbb{E}\left[\sum_{i=1}^{N} X_i \middle| N\right] &= N\,\mathbb{E}[X] \\ \operatorname{Var}\left(\sum_{i=1}^{N} X_i \middle| N\right) &= N\operatorname{Var}(X)\,. \end{aligned} \implies \begin{aligned} \mathbb{E}\left[\operatorname{Var}\left(\sum_{i=1}^{N} X_i \middle| N\right)\right] &= \mathbb{E}[N\operatorname{Var}(X)] = \mathbb{E}[N]\operatorname{Var}(X) \\ \operatorname{Var}\left(\mathbb{E}\left[\sum_{i=1}^{N} X_i \middle| N\right]\right) &= \operatorname{Var}(N\,\mathbb{E}[X]) = \mathbb{E}[X]^2\operatorname{Var}(N) \end{aligned}$$

Therefore, we have

$$\operatorname{Var}\left(\sum_{i=1}^{N} X_i\right) = \mathbb{E}[N]\operatorname{Var}(X) + \mathbb{E}[X]^2\operatorname{Var}(N)\,.$$

3 Moment Generating Function (MGF)

Motivation: Assume that $\mathbb{E}\left[X^k\right]$ exist for all $k \in \mathbb{N}$. Consider the following "formally":

$$1 + t\,\mathbb{E}[X] + \frac{t^2}{2!}\mathbb{E}\left[X^2\right] + \frac{t^3}{3!}\mathbb{E}\left[X^3\right] + \cdots = \mathbb{E}\left[1 + tX + \frac{(tX)^2}{2!} + \cdots\right] = \mathbb{E}\left[e^{tX}\right].$$

Observe that

$$M_X(t) \equiv \mathbb{E}\left[e^{tX}\right] = \sum_x e^{tx}p(x) = \int e^{tx}f(x)dx.$$

Question:

(i) If $\mathbb{E}\left[X^k\right]$ exist for all k, does $M(t)$ exists for some t in an interval of $\mathbb{R}$? For any X, $M(X) = \mathbb{E}\left[e^{tX}\right]$ always exists at $t = 0$. There are examples that $M(t)$ only exists at $t = 0$.
Consider $p(n) = p(-n) = \frac{c}{n}$ where $n = 1, 2, ...$ Choosing c appropriately such that $\{p(n)\}_{n=\mathbb{Z}\setminus\{0\}}$ is a pmf.

$$M(t) = \sum_{n \neq 0} e^{nt}\frac{c}{n^2} \quad \text{only converges at } t = 0.$$

(ii) Now if $M_X(t)$ exists for $t \in (-t_0, t_0)$ where $t_0 > 0$, i.e., $M_X(t)$ exists in a neighborhood of 0. Note that

$$\begin{aligned} & e^{|tX|} \le e^{tX} + e^{-tX} \\ \implies & \int_{-\infty}^{\infty} e^{|tx|}f(x)dx \le \int_{-\infty}^{\infty} e^{tx}f(x)dx + \int_{-\infty}^{\infty} e^{-tx}f(x)dx \\ \implies & \text{finite if } t \in (-t_0, t_0). \end{aligned}$$

Hence,

$$
\begin{aligned}
M(t) = \int_{-\infty}^{\infty} e^{tx} f(x)dx &= \int_{-\infty}^{\infty} \sum_{k=0}^{\infty} \frac{(tx)^k}{k!} f(x)dx \\
&= \sum_{k=0}^{\infty} \frac{t^k}{k!} \int_{-\infty}^{\infty} x^k f(x)dx \qquad \text{(dominated convergence theorem)} \\
&= \sum_{k=0}^{\infty} \frac{t^k\,\mathbb{E}\left[X^k\right]}{k!} \\
&\implies M_X(t) \text{ is an analytical function for } t \in (-t_0, t_0).
\end{aligned}
$$

Therefore, if $M_X(t)$ exists in some neighborhood of 0, then $\mathbb{E}\left[X^k\right]$ exists $\forall k$ and $M_X^{(k)}(0) = \mathbb{E}\left[X^k\right]$.
Question: If $\mathbb{E}\left[X^k\right]$ exists for all k, does $M_X(t)$ exists in some neighborhood of 0? **No**. If $f(x) = 0$ if $x \le 0$, then $M_X(t)$ exists if $t \le 0$.

$$
M_X(t) = \int_0^{\infty} e^{tx} f(x)dx \le \int_0^{\infty} f(x)dx = 1 \text{ if } t \le 0.
$$

Example. Here are some common MGF:

1. $X \sim \text{Binomial}(n, p)$.

$$
\begin{aligned}
M_X(t) = \mathbb{E}\left[e^{tX}\right] = \sum_{k=0}^{n} e^{tk} \binom{n}{k} p^k (1-p)^{n-k} &= \sum_{k=0}^{n} \binom{n}{k} (pe^t)^k (1-p)^{n-k} \\
&= \left(pe^t + 1 - p\right)^n \quad t \in \mathbb{R}
\end{aligned}
$$

2. $X \sim \text{Poisson}(\lambda)$. $p(k) = (\lambda^k e^{-\lambda})/k!$.

$$
\begin{aligned}
M_X(t) = \mathbb{E}\left[Xe^{tX}\right] = \sum_{k=0}^{\infty} e^{tk} \frac{\lambda^k e^{-\lambda}}{k!} = \sum_{k=0}^{\infty} \frac{e^{-\lambda}(\lambda e^t)^k}{k!} &= e^{-\lambda} \sum_{k=1}^{\infty} \frac{(\lambda e^t)^k}{k!} \\
&= e^{-\lambda} e^{\lambda e^t} \\
&= \exp\left\{\lambda(e^t - \lambda)\right\}.
\end{aligned}
$$

3. $X \sim \mathcal{N}(0, 1)$,

$$
M_X(t) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{tx} e^{-x^2/2} dx = e^{t^2/2}.
$$

For any normal distribution $Y \sim \mathcal{N}(\mu, \sigma^2)$,

$$
M_Y(t) = \mathbb{E}\left[e^{tY}\right] = \mathbb{E}\left[e^{t\sigma X + t\mu}\right] = e^{t\mu}\,\mathbb{E}[(t\sigma)X] = \exp\left\{\frac{(t\sigma)^2}{2} + t\mu\right\}.
$$

4. $X \sim \text{Cauchy}$, $f(x) = 1/(\pi(1 + x^2))$ $\forall x \in \mathbb{R}$. The moment generating function of X only exists at $t = 0$.

Theorem 3. *Let X and Y be two random variables. Moreover, $M_X(t)$ and $M_Y(t)$ exist in a neighborhood of 0 and $M_X(t) = M_Y(t)$. Then $X \overset{d}{=} Y$.*

►*Proof.* For simplicity, consider discrete random variables. Let X takes values $\{x_1, ..., x_n\}$ where $p(x_i) = \Pr\{X = x_i\} = p_i$; Y takes values $\{y_1, ..., y_m\}$ where $p(y_j) = \Pr\{Y = y_j\} = q_j$. Then

$$
M_X(t) = \sum_{i=1}^{n} e^{tx_i} p_i = M_Y(t) = \sum_{j=1}^{n} e^{ty_j} q_j
$$

Assume that $M_X(t)$ and $M_Y(t)$ exists $\forall t$. Define $x_{i_0} = \max\{x_1, ..., x_n\}$ and $y_{j_0} = \max\{y_1, ..., y_m\}$ then as $t \to \infty$, we have

$$\left.\begin{aligned} \lim_{t\to\infty} M_X(t) &= e^{tx_{i_0}} p_{i_0} \\ \lim_{t\to\infty} M_Y(t) &= e^{ty_{j_0}} q_{j_0} \end{aligned}\right. \implies x_{i_0} = y_{j_0}, p_{i_0} = q_{j_0}.$$

Therefore, we prove that one of the events are the same. For other events, just remove the max event and repeat the argument above. ■

3.1 Conditioning (MGF)

Let $\{U_1, U_2, ...\} \overset{\text{iid}}{\sim} \mathcal{U}(0,1)$. Let $N = \min\{n : \sum_{i=1}^n U_i > 1\}$. We have $\mathbb{E}[N] = e$. Let $x \in (0,1)$, define $N_x = \min\{n : \sum_{i=1}^n U_i > x\}$. Set $m(x) := \mathbb{E}[N_x]$. Derive $m(x)$.
Sol: Compute $m(x) = \mathbb{E}[N_x]$. We condition $\mathbb{E}[N_x]$ on U_1. That is,

$$\mathbb{E}[N_x] = \mathbb{E}[\mathbb{E}[N_x | U_1]] = \int_0^1 \mathbb{E}[N_x | U_1 = y]\, dy.$$

Observe that

$$\mathbb{E}[N_x | U_1 = y] = \begin{cases} 1 & \text{if } y \geq x \\ 1 + m(x-y) & \text{if } 0 < y < x \end{cases}$$

Therefore, we have

$$\begin{aligned} m(x) = \mathbb{E}[N_x] &= \int_0^x (1 + m(x-y))dy + \int_x^1 dy = 1 + \int_0^x m(x-y)dy \\ &\implies m'(x) = m(x) \\ &\implies m(x) = ce^x \text{ with } m(0) = 1. \end{aligned}$$

3.2 Independence (MGF)

Consider $X_1, ..., X_n$ are independent random variables. Set $Y = \sum_{i=1}^n X_i$.

$$M_Y(t) = \mathbb{E}\left[e^{t(X_1 + \cdots + X_n)}\right] = \mathbb{E}\left[e^{tX_1} \cdots e^{tX_n}\right] = \mathbb{E}\left[e^{tX_1}\right] \cdots \mathbb{E}\left[e^{tX_n}\right] = M_{X_1}(t) \cdots M_{X_n}(t).$$

4 Characteristic functions (chf)

Definition 9. *Let X be a random variable. Then the characteristic function (chf) of X,*

$$\begin{aligned} \phi_X(t) &:= \mathbb{E}\left[e^{itX}\right] \\ &= \int_\Omega e^{itX} d\Pr && (X : \Omega \to \mathbb{R}) \\ &= \int_{\mathbb{R}} e^{itx} d\mu && (\mu \text{ is probability measure on } (\mathbb{R}, \mathcal{B})) \\ &= \int_{-\infty}^{\infty} e^{itx} f(x) dx. && (\text{if } \mu \text{ is absolutely continuous w.r.t. Lebesgue measure}) \\ &= \hat{f}(t). && (\text{Fourier transform of } f(x)) \end{aligned}$$

where $\mu = \Pr \circ X^{-1}$ is a probability measure on $(\mathbb{R}, \mathcal{B})$. Note that if X is real-valued, then $\phi_X(t)$ always exists since $|\phi_X(t) < 1|$ $\forall t$.

Why $\phi_X(t)$?

(1) If X_1 and X_2 are independent, then $Y = X_1 + X_2$ has

$$\phi_Y(t) = \phi_{X_1}(t)\phi_{X_2}(t) = \mathbb{E}\left[e^{it(X_1+X_2)}\right] = \mathbb{E}\left[e^{itX_1}\right]\mathbb{E}\left[e^{itX_2}\right] = \phi_{X_1}(t)\phi_{X_2}(t).$$

or if $f_{X_1}(x)$ and $f_{X_2}(x)$ are pdf's of X_1 and X_2, then $f_Y(t) = (f_{X_1} * f_{X_2})(x)$.

$$\phi_Y(t) = \hat{f}_Y(t) = \widehat{f_{X_1} * f_{X_2}}(t) = \hat{f}_{X_1}(x)\hat{f}_{X_2}(x).$$

(2) Assume that $\{\mu_n\}$ is a sequence of probability measures, e.g., $\mu_n = \Pr \circ X_n^{-1}$ where $\{X_n\}$ is a sequence of random variables. We have

$$\mu_n \to \mu \iff \phi_{\mu_n}(t) \to \phi_\mu(t) \quad \forall t$$

4.1 Properties of Characteristic function

1. $|\phi_X(t)| < 1$.
2. $$\overline{\phi_X(t)} = \int_\Omega e^{-itX} d\Pr = \int_\Omega e^{i(-t)X} d\Pr = \phi_X(-t).$$

   provided X is real.
3. $\phi_X(t)$ is uniform continuous, i.e.,

   $$|\phi_X(t_h) - \phi_X(t)| = \left|\int_{-\infty}^{\infty}\left[e^{i(t+h)x}f(x) - e^{itx}f(x)\right]dx\right| \le \int_{-\infty}^{\infty}|e^{ihx} - 1|f(x)dx < \infty.$$

   Note that $\lim_{n\to 0}|e^{ihx} - 1| = 0$. Therefore, by bounded convergence theorem,

   $$\lim_{h\to 0}|\phi_X(t+h) - \phi_X(t)| = 0. \qquad \text{(independent of } t\text{)}$$

   That is, $\phi_X(t)$ is uniform continuous.

4.2 Moments

By Taylor's formula:

$$e^{ix} = \sum_{k=0}^{n}\frac{(ix)^k}{k!} + \frac{i^{n+1}}{n!}\int_0^x (x-s)^n e^{is} ds \tag{3}$$

Observe that

$$\int_0^x (x-s)^n e^{is} ds = \int_0^x e^{is} d\left(\frac{-1}{n+1}(x-s)^{n+1}\right) = \frac{x^{n+1}}{n+1} + \frac{i}{n+1}\int_0^x (x-s)^{n+1}e^{is}ds.$$

Let $n = n-1$,

$$\begin{aligned}
\int_0^x (x-s)^{n-1}e^{is}ds &= \frac{x^n}{n} + \frac{i}{n}\int_0^x (x-s)^n e^{is}ds. \\
\implies \int_0^x (x-s)^n e^{is}ds &= \frac{n}{i}\left[\int_0^x (x-s)^{n-1}e^{is}ds - \frac{x^n}{n}\right] \\
&= \frac{n}{i}\int_0^x (x-s)^{n-1}e^{is}ds - \frac{n}{i}\int_0^x (x-t)^{n-1}ds \\
&= \frac{n}{i}\int_0^x (x-s)^{n-1}(e^{is}-1)ds
\end{aligned}$$

Hence, we have

$$\begin{aligned}
e^{ix} &= \sum_{k=0}^{n} \frac{(ix)^k}{k!} + \frac{i^{n+1}}{n!}\frac{n}{i}\int_0^x (x-s)^{n-1}(e^{is}-1)ds \\
&= \sum_{k=0}^{n} \frac{(ix)^k}{k!} + \frac{i}{(n-1)!}\int_0^x (x-s)^{n-1}(e^{is}-1)ds \qquad (4)
\end{aligned}$$

<u>Conclusion</u>:

$$\left| e^{ix} - \sum_{i=0}^{n} \frac{(ix)^n}{k!} \right| \leq \min\left\{ \frac{|x|^{n+1}}{(n+1)!}, \frac{2|x|^n}{n!} \right\}.$$

where the first remainder term comes from (3) and the second terms comes from (4). The reason for this calculation is that when x is less than one, we can use the first approximation; when x is larger, we can use the second error approximation. For example:

$$\begin{aligned}
|e^{ix}-1| &\leq \min\{|x|, 2\} \\
|e^{ix}-(1+ix)| &\leq \min\left\{\frac{1}{2}x^2, 2|x|\right\} \\
\left|e^{ix}-\left(1+ix-\frac{1}{2}x^2\right)\right| &\leq \min\left\{\frac{1}{6}|x|^3, x^2\right\}
\end{aligned}$$

Now if $\mathbb{E}[|x|^n] < \infty$, then

$$\left| \phi_X(t) - \sum_{k=1}^{\infty} \frac{(it)^k}{k!}\,\mathbb{E}\left[X^k\right] \right| \leq \mathbb{E}\left[\min\left\{ \frac{|tX|^{n+1}}{(n+1)!}, \frac{2|tX|^n}{n!} \right\}\right] \qquad (5)$$

Therefore, for any t satisfying

$$\lim_{n\to\infty} \frac{|t|^n\,\mathbb{E}[|X|^n]}{n!} = 0, \qquad (6)$$

we have that

$$\phi_X(t) = \sum_{k=0}^{\infty} \frac{(it)^k}{k!}\,\mathbb{E}\left[X^k\right], \qquad (7)$$

which is the same as the “formal differentiation”.

To guarantee (6), we have the sufficient condition:

$$\mathbb{E}\left[e^{|tX|}\right] < \infty.$$

If (7) holds, then

$$\phi_X^{(k)}(0) = i^k\,\mathbb{E}\left[X^k\right] \quad \forall k = 0, 1, 2, ... \qquad (8)$$

But (8) holds as long as $\mathbb{E}\left[|X|^k\right] < \infty$. For example, $k = 1$. Consider

$$\begin{aligned}
\frac{\phi_X(t+h)-\phi_X(t)}{h} - \mathbb{E}\left[iXe^{itX}\right] &= \mathbb{E}\left[\frac{e^{i(t+h)X}-e^{itX}}{h} - iXe^{itX}\right] \\
&= \mathbb{E}\left[e^{itX}\frac{e^{ihX}-1-ihX}{h}\right] \\
&= \int e^{itx}\frac{e^{ihx}-1-ihx}{h} f_X(x)dx
\end{aligned}$$

if X has pdf $f_X(x)$.

From the estimates (5) derived above, this term is bounded by

$$\frac{1}{n}\int \min\left\{\frac{1}{2}|hx|^2, 2|hx|\right\} f_X(x)dx = \int \min\left\{\frac{1}{2}|h||x|^2, 2|x|\right\} f_X(x)dx < \infty \text{ if } \mathbb{E}[|X|] < \infty.$$

Moreover,

$$\lim_{n\to\infty} \min\left\{\frac{1}{2}|h||x|^2, 2|x|\right\} = 0$$

Hence, taking $h \to 0$, we have

$$\phi'_X(t) = \mathbb{E}\left[iXe^{itX}\right], \text{ i.e. } \phi'_X(0) = \mathbb{E}[iX] .$$

In particular, if $\mathbb{E}\left[|X|^2\right] < \infty$, then (by Taylor series)

$$\phi_X(t) = 1 + it\,\mathbb{E}[X] - \frac{1}{2}t^2\,\mathbb{E}\left[X^2\right] + t^2\,\mathbb{E}\left[\min\left\{\frac{1}{5}|t||X|^3, |X|^2\right\}\right]$$

Theorem 4. *$\phi_X(t)$ uniquely determines the distribution of X and vice versa. (An inversion formula exists)*

5 Convergence of Infinite Sequence of Random Variables

Assume $\{X_n\}_{n=1}^{\infty}$ is a sequence of random variables.

Definition 10 (Converge almost surely)**.** *We say that $X_n \xrightarrow{as} X$ if*

$$\Pr\{\{\omega \in \Omega : X_n(\omega) \to X(\omega)\}\} = 1 \iff \Pr\{\{\omega \in \Omega : X_n(\omega) \not\to X(\omega)\}\} = 0\,.$$

Definition 11 (Converge in probability)**.** *$X_n \xrightarrow{p} X$ if*

$$\forall \varepsilon > 0,\ \Pr\{|X_n - X| \geq \varepsilon\} \to 0 \text{ as } n \to \infty.$$

Theorem 5. *Convergence almost surely $\implies$ Convergence in probability.*

Remark 8. *The converse does not hold. Consider the following counterexample:*

$$\begin{aligned}
&A_1 = (0, 1] \\
&A_2 = (0, 1/2],\ A_3 = (1/2, 1] \\
&A_4 = (0, 1/4],\ A_5 = (1/4, 1/2], \ldots
\end{aligned}$$

Let $X_n = \chi_{A_n}(\omega)$, $\omega \in (0, 1]$. Notice that $X = 0$ is one possible limit. Actually, we have $X_n \xrightarrow{p} 0$ since $\forall \varepsilon > 0$,

$$\Pr\{|X_n - 0| \geq \varepsilon\} = \Pr\{|X_n| \geq \varepsilon\} = 0 \text{ as } n \to \infty.$$

However, $X_n \overset{as}{\not\to} 0$.

5.1 Another Characterisation of convergence almost surely

$$\omega \in \Omega,\ X_n(\omega) \not\to X(\omega). \iff \exists \varepsilon > 0 \text{ s.t. } |X_n(\omega) - X(\omega)| > \varepsilon \text{ occurs infinitely often}$$

$$X_n \xrightarrow{as} X \iff \Pr\left\{ \bigcup_{\varepsilon \in \mathbb{Q}^+} \{|X_n - X| \geq \varepsilon \text{ i.o.}\} \right\} = 0.$$

Assume that $\{A_n\}_{n=1}^{\infty}$ is a sequence of events.

$$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{k\geq n} A_k \in \mathscr{F} \quad \liminf_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} \bigcap_{k\geq n} A_k \in \mathscr{F}$$

Lemma 3. *$\omega \in \limsup_{n\to\infty} A_n \iff \omega \in A_n$ infinitely often. That is,*

$$\{A_n \text{ i.o.}\} = \left\{ \limsup_{n\to\infty} A_n \right\}.$$

Lemma 4. *$\omega \in \liminf_{n\to\infty} A_n \iff \omega \in A_n$ for all but finitely many n's. That is,*

$$\{A_n \text{ occurs but finitely many } n\text{'s}\} = \left\{ \liminf_{n\to\infty} A_n \right\}.$$

(前面可能沒有發生, 但是後面一定要全部都發生。)

Theorem 6 (Continuity of probability measure)**.**

$$\Pr\left\{ \liminf_{n\to\infty} A_n \right\} \leq \liminf_{n\to\infty} \Pr\{A_n\} \leq \limsup_{n\to\infty} \Pr\{A_n\} \leq \Pr\left\{ \limsup_{n\to\infty} A_n \right\}$$

►*Proof.* Let $B_n = \bigcap_{k\geq n} A_k$.

$$\text{let } B_n = \bigcap_{k\geq n} A_k \implies B_n \uparrow \liminf_{n\to\infty} A_n$$

$$\text{let } C_n = \bigcup_{k\geq n} A_k \implies C_n \uparrow \limsup_{n\to\infty} A_n.$$

Then

$$\Pr\{A_n\} \geq \Pr\{B_n\} \implies \liminf_{n\to\infty} \Pr\{A_n\} \geq \liminf_{n\to\infty} \Pr\{B_n\} = \lim_{n\to\infty} \Pr\{B_n\} = \Pr\left\{ \liminf_{n\to\infty} A_n \right\}.$$

Next,

$$\Pr\{A_n\} \leq \Pr\{C_n\} \implies \limsup_{n\to\infty} \Pr\{A_n\} \leq \limsup_{n\to\infty} \Pr\{C_n\} = \lim_{n\to\infty} \Pr\{C_n\} = \Pr\left\{ \limsup_{n\to\infty} A_n \right\}.$$

■

Now we return to the proof of Theorem 5:

►*Proof.* (of Theorem 5)

$$X_n \xrightarrow{as} X \iff \Pr\left\{ \bigcup_{\varepsilon \in \mathbb{Q}^+} \{|X_n - X| \geq \varepsilon \text{ i.o.}\} \right\} = 0$$

$$\iff \Pr\{\{|X_n - X| \geq \varepsilon \text{ i.o.}\}\} = 0 \quad \forall \varepsilon > 0$$

Let $A_n = \{|X_n - X| \geq \varepsilon\}$,

$$\begin{aligned} \limsup_{n\to\infty} \Pr\{A_n\} &\leq \Pr\left\{ \limsup_{n\to\infty} A_n \right\} \\ &= \Pr\{\{A_n \text{ i.o.}\}\} = 0 \\ &\implies \lim_{n\to\infty} \Pr\{\{|X_n - x| \geq \varepsilon\}\} = 0 \text{ i.e. } X_n \xrightarrow{as} X. \end{aligned}$$

■

5.2 Borel-Cantelli's Lemmas

Lemma 5 (First Borel-Cantelli's Lemma)**.**

$$\sum_n \Pr\{A_n\} < \infty \implies \Pr\{\{A_n \textit{ i.o.}\}\} = 0.$$

►*Proof.*

$$\{A_n \text{ i.o.}\} = \limsup_{n\to\infty} A_n = \bigcap_{n\geq 1} \bigcup_{k\geq n} A_k \subseteq \bigcup_{k\geq n} A_k \quad \forall n$$

Now take probability on both sides:

$$\Pr\{\{A_n \text{ i.o.}\}\} \leq \Pr\left\{\bigcup_{k\geq n} A_k\right\} \leq \sum_{k\geq n} \Pr\{A_k\} \to 0 \text{ as } n \to \infty.$$

Therefore, $\Pr\{\{A_n \text{ i.o.}\}\} = 0$. ■

Lemma 6 (Second Borel-Cantelli's Lemma)**.** *If the sequence of events $\{A_n\}$ are independent and $\sum_n \Pr\{A_n\} = \infty$, then*

$$\Pr\{\{A_n \textit{ i.o.}\}\} = 1.$$

►*Proof.* Recall that $\{A_n \text{ i.o.}\} = \limsup_{n\to\infty} A_n$. It suffices to show that

$$\Pr\Big\{\{A_n \text{ i.o.}\}^{\complement}\Big\} = 0.$$

Consider the event $\{A_n \text{ i.o.}\}^{\complement}$:

$$\{A_n \text{ i.o.}\} = \left(\bigcap_{n\geq 1} \bigcup_{k\geq n} A_k\right)^{\complement} = \bigcup_{n\geq 1} \bigcap_{k\geq n} A_k^{\complement} \qquad \text{(De Morgan)}$$

So it suffices to show that $\Pr\Big\{\bigcap_{k\geq n} A_k^{\complement}\Big\}$ $\forall n$. Consider

$$\begin{aligned}
\Pr\left\{\bigcap_{k=n}^{n+j} A_k^{\complement}\right\} = \prod_{k=n}^{n+j} \Pr\Big\{A_k^{\complement}\Big\} = \prod_{k=n}^{n+j} (1 - \Pr\{A_k\}) &\leq \prod_{k=n}^{n+j} \exp\Big\{e^{-\Pr\{A_k\}}\Big\} \\
&= \exp\left\{-\sum_{k=n}^{n+j} \Pr\{A_k\}\right\} \quad \forall n \\
\implies \lim_{j\to\infty} \Pr\left\{\bigcap_{k=n}^{n+j} A_k^{\complement}\right\} = \Pr\left\{\bigcap_{k=n}^{n+j} A_k^{\complement}\right\} \\
&\leq \lim_{j\to\infty} \exp\left\{-\sum_{k=n}^{n+j} \Pr\{A_k\}\right\} = 0.
\end{aligned}$$

■

5.3 Necessary and Sufficient condition for Convergence in probability

Theorem 7. *A necessary and sufficient condition for $X_n \xrightarrow{p} X$ is that for any subsequence $\{X_{n_k}\}$ of $\{X_n\}$, $\exists$ another subsequence $\{X_{n_{k_j}}\}$ of $\{X_{n_k}\}$ s.t. $X_{n_{k_j}} \xrightarrow{as} X$ as $n_{k_j} \to \infty$.*

►*Proof.*

($\Rightarrow$) Assume that $X_n \xrightarrow{p} X$, i.e., $\forall \varepsilon > 0$,

$$\Pr\{|X_n - X| \geq \varepsilon\} \to 0 \text{ as } n \to \infty.$$

Given any subsequence $\{X_{n_k}\}$, there exists a n_{k_j} s.t.

$$\Pr\left\{|X_{n_k} - X| \geq \frac{1}{j}\right\} \leq 2^{-j} \quad \forall k \geq k_j.$$

In particular,

$$\Pr\left\{|X_{n_{k_j}} - X| \geq \frac{1}{j}\right\} \leq 2^{-j} \quad \forall j.$$

Hence,

$$\sum_{j=1}^{\infty} \Pr\left\{|X_{n_{k_j}} - X| \geq \frac{1}{j}\right\} \leq \sum_{j=1}^{\infty} 2^{-j} < \infty.$$

By the first Borel-Cantelli Lemma (Lemma 5), we have

$$\Pr\left\{\left\{|X_{n_{k_j}} - X| \geq \frac{1}{j} \text{ i.o.}\right\}\right\} = 0$$

That is,

$$\Pr\left\{\left\{|X_{n_{k_j}} - X| \leq \frac{1}{j}, \forall j \text{ but finitely many } j\text{'s}\right\}\right\} = 1.$$

Therefore, $X_{n_{k_j}} \xrightarrow{as} X$ as $j \to \infty$ □

($\Leftarrow$) Assume the contrary, that $X_n \not\xrightarrow{p} X$. There exists an $\varepsilon > 0$ and a subsequence $\{X_{n_k}\}$ s.t.

$$\Pr\{|X_{n_k} - X| \geq \varepsilon\} > \varepsilon.$$

That is, no subsequence of $\{X_{n_k}\}$ can converge to X in probability. Hence, no subsequence of $\{X_{n_k}\}$ can converge to X almost surely. $\rightarrow\!\leftarrow$

■

5.4 Determination of Convergence of Random Variables (Cauchy)

Question: given a sequence of random variables $\{X_n\}$, can we determine whether $\{X_n\}$ converges to something in probability?

Definition 12. *A sequence of random variables $\{X_n\}$ is called **Cauchy in probability** if $\forall \varepsilon > 0$,*

$$\Pr\{|X_n - X_m| \geq \varepsilon\} \to 0 \text{ as } n, m \to \infty.$$

Theorem 8. $X_n \xrightarrow{p} X \iff \{X_n\}$ *is Cauchy in probability.*

►*Proof.*

(⇒) Consider the following:

$$\{|X_n - X_m| \geq \varepsilon\} \subseteq \{|X_n - X| \geq \varepsilon/2\} \cup \{|X_m - X| \geq \varepsilon/2\}.$$

Take the probability on both sides:

$$\Pr\{\{|X_n - X_m| \geq \varepsilon\}\} \leq \Pr\{\{|X_n - X| \geq \varepsilon/2\}\} + \Pr\{\{|X_m - X| \geq \varepsilon/2\}\} \to 0$$

as $n, m \to \infty$. □

(⇐) Since $\{X_n\}$ is Cauchy in probability, $\forall j = 1, 2, 3, \ldots\ \exists k_j \in \mathbb{N}$ s.t.

$$\Pr\left\{|X_n - X_m| \geq 2^{-j}\right\} \leq 2^{-j} \quad \forall m, n \geq k_j \tag{9}$$

We can assume that $k_j \nearrow \infty$ as $j \to \infty$. By equation (9), we have

$$|X_{k_{j+1}} - X_{k_j}| < 2^{-j} \quad \text{except on a set } E_j \text{ with } \Pr\{E_j\} \leq 2^{-j}$$

Define $H_i = \bigcup_{j=i}^{\infty} E_j$. Then

$$\left|X_{k_{j+1}}(\omega) - X_{k_j}(\omega)\right| < 2^{-j} \quad \forall j \geq i \text{ for } \omega \notin H_i.$$

Hence,

$$\sum_{j=i}^{\infty} \left(X_{k_{j+1}}(\omega) - X_{k_j}(\omega)\right) \leq \sum_{j=i}^{\infty} 2^{-j} = 2^{-i+1} \quad \forall \omega \in H_i^{\complement}.$$

That is, $X_{k_{j+1}}$ converges uniformly to some X on $H_i^{\complement}$ $\forall i$.

$$\Pr\{H_i\} = \Pr\left\{\bigcup_{j=i}^{\infty} E_j\right\} \leq 2^{-i+1} \quad \forall i$$

$$\Pr\left\{\bigcup_{i=1}^{\infty} H_i^{\complement}\right\} = \Pr\left\{\left(\bigcap_{i=1}^{\infty} H_i\right)^{\complement}\right\} = 1 - \Pr\left\{\bigcap_{i=1}^{\infty} H_i\right\} = 1 \implies X_{k_j} \xrightarrow{as} X$$
$$\implies X_{k_j} \xrightarrow{p} X.$$

Next, check

$$\Pr\{|X_n - X| \geq \varepsilon\} \leq \underbrace{\Pr\left\{|X_n - X_{k_j}| \geq \varepsilon/2\right\}}_{\text{Cauchy in probability}} + \Pr\left\{|X_{k_j} - X| \geq \varepsilon/2\right\}.$$

■

5.5 Weak Law of Large Number (WLLN)

Let $X_1, X_2, \ldots$ be iid. Consider the mean

$$\bar{X} := \frac{S_n}{n} := \frac{\sum_{k=1}^{n} X_k}{n}.$$

Theorem 9 (WLLN)**.** *$\bar{X} \xrightarrow{p} \mathbb{E}[X_i]$ if $\mathrm{Var}(X) < \infty$. That is, $\Pr\left\{|\bar{X}_n - \mathbb{E}[X_i]| \geq \varepsilon\right\} \to 0$ as $n \to \infty$.*

Theorem 10 (Markov Inequality)**.** *Assume that* $X \geq 0$, *then* $\forall \alpha > 0$, *we have*

$$\Pr\{X \geq \alpha\} \leq \frac{1}{\alpha}\,\mathbb{E}[X]\,.$$

►*Proof.* Consider $\mathbb{E}[X]$:

$$\begin{aligned}\mathbb{E}[X] = \int_{\Omega} X\,d\Pr = \int_{\{X<\alpha\}} X\,d\Pr + \int_{\{X\geq\alpha\}} X\,d\Pr \geq \int_{\{X\geq\alpha\}} X\,d\Pr \\ \geq \alpha \int_{\{X\geq\alpha\}} d\Pr = \alpha\Pr\{X \geq \alpha\}\end{aligned}$$

■

Theorem 11 (Chebyshev's Inequality)**.** *For any random variable* X

$$\Pr\{|X - \mathbb{E}[X]\,| \geq \varepsilon\} \leq \frac{\operatorname{Var}(X)}{\varepsilon^2}$$

►*Proof.* By Markov inequality, we have

$$\Pr\{|X - \mathbb{E}[X]\,| \geq \varepsilon\} = \Pr\big\{(X - \mathbb{E}[X])^2 \geq \varepsilon^2\big\} \leq \frac{1}{\varepsilon^2}\,\mathbb{E}\big[(X - \mathbb{E}[X])^2\big] = \frac{\operatorname{Var}(X)}{\varepsilon^2}.$$

■

►*Proof.* (of WLLN) We have the expectation of $\bar{X}_n$ as

$$\bar{X}_n = \frac{1}{n}S_n, \quad \mathbb{E}\big[\bar{X}_n\big] = \mathbb{E}[X_1]\,.$$

Also, we have the variance as

$$\operatorname{Var}\big(\bar{X}_n\big) = \operatorname{Var}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{\operatorname{Var}(X_1)}{n}.$$

Therefore, by Chebyshev's inequality,

$$\Pr\big\{|\bar{X}_n - \mathbb{E}[X_1]\,| \geq \varepsilon\big\} \leq \frac{\operatorname{Var}\big(\bar{X}_n\big)}{\varepsilon^2} = \frac{\operatorname{Var}(X_1)}{n\varepsilon^2} \to 0$$

as $n \to \infty$ ■

Example. Let $X_1, X_2, \ldots \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$. By WLLN,

$$\Pr\big\{|\bar{X}_n - p| \geq \varepsilon\big\} \leq \frac{\operatorname{Var}(X_1)}{n\varepsilon^2} = \frac{p(1-p)}{n\varepsilon^2}.$$

That is,

$$\Pr\big\{|\bar{X}_n - p| < \varepsilon\big\} = \Pr\left\{p - \varepsilon < \frac{S_n}{n} < p + \varepsilon\right\} \geq 1 - \frac{p(1-p)}{n\varepsilon^2} \geq 1 - \frac{1}{4n\varepsilon^2}.$$

If $p = 1/2$. Now take $\varepsilon = 1/100$. If we want

$$\Pr\left\{p - 0.01 < \frac{S_n}{n} < p + 0.01\right\} \geq 0.95 \implies 1 - \frac{10000}{4n} \geq 0.95 \implies n \geq 50000.$$

Note that the number of trials needed is relatively large. When central limit theorem is used, we can cut the number of trial by a lot.

5.6 Central Limit Theorem

Poisson approximation of a Binomial(n,p) under the condition $p \lll 1$, $n \ggg 1$ and $\lambda \approx np$. Now consider that p is fixed (difference from Poisson approximation), but $n \ggg 1$. We want to approximate Binomial(n,p).

Under this assumption, we have the following theorem:

Theorem 12 (DeMoivre-Laplace)**.** *Let $X \sim$ Binomial(n,p). $\forall a, b \in \mathbb{R}$, we have*

$$\Pr\left\{a \le \frac{X-np}{\sqrt{npq}} < b\right\} \to \frac{1}{\sqrt{2\pi}} \int_a^b \exp\left\{-\frac{x^2}{2}\right\} dx \quad \text{as} \quad n \to \infty$$

where $q = 1-p$. That is, it will converge to "normal" distribution.

Question: Is DeMoivre-Laplace "better" then WLLN? Recall that under WLLN, in the example on page 20, to have a confidence level of 95%, we need 50000 trials. If we want 99% of confidence level, we need over 250000 trials! If we let $\bar{X} = S_n/n$, then the theorem implies

$$\begin{aligned} \Pr\left\{-r \le \frac{\bar{X}-p}{\sqrt{(pq)/n}} < r\right\} &\to \int_{-r}^{r} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} dx && (\text{as } n \to \infty) \\ &= 2\int_0^r \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} dx \overset{\text{let}}{=} 0.99 && (\text{if confidence level } 99\%) \\ &\implies r \approx 2.57. \end{aligned}$$

Now we reconsider "how many trials we need to get to 99% confidence?"

$$\begin{aligned} \left|\frac{\bar{X}-p}{\sqrt{pq/n}}\right| \le 2.57 &\implies |\bar{X}-p| \le 2.57\sqrt{\frac{pq}{n}} \le \frac{2.57}{2}\frac{1}{\sqrt{n}} \\ \text{we want } |\bar{X}-p| = |S_n/n - p| \le \frac{1}{100} &\implies \frac{2.57}{2}\frac{1}{\sqrt{n}} \le \frac{1}{100} && (\varepsilon = 1/100) \\ &\implies n \ge 16513. \end{aligned}$$

Theorem 13. *$X \sim$ Binomial(n,p) with p fixed. Assume that α_n is a non-decreasing function of n satisfying $\lim_{n\to\infty} \alpha_n/n^{1/6} = 0$. Then*

$$\max_{k:|k-np|<\sqrt{n}\cdot\alpha_n} \left|\frac{\Pr\{X=k\}\sqrt{npq}}{\varphi\left(\frac{k-np}{\sqrt{npq}}\right)} - 1\right| = O\left(\frac{\alpha_n^3}{\sqrt{n}}\right) \quad \text{where} \quad \varphi(x) = \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{x^2}{2}\right\}.$$

►*Proof.* (of DeMoivre-Laplace) By triangular inequality, we have

$$\begin{aligned} &\left|\Pr\left\{a \le \frac{X-np}{\sqrt{npq}} < b\right\} - \frac{1}{\sqrt{2\pi}}\int_a^b \exp\left\{-\frac{x^2}{2}\right\} dx\right| \\ \le &\left|\sum_{k=\lceil a\sqrt{npq}+np\rceil}^{\lceil b\sqrt{npq}-1+np\rceil} \left(\Pr\{X=k\} - \frac{1}{\sqrt{npa}}\varphi\left(\frac{k-np}{\sqrt{npq}}\right)\right)\right| \qquad (\mathrm{I}) \\ + &\underbrace{\left|\sum_{k=\lceil a\sqrt{npq}+np\rceil}^{\lceil b\sqrt{npq}-1+np\rceil} \frac{1}{\sqrt{npq}}\varphi\left(\frac{k-np}{\sqrt{npq}}\right) - \frac{1}{\sqrt{2\pi}}\int_a^b \exp\left\{-\frac{x^2}{2}\right\} dx\right|}_{\text{Riemann sum } \to 0} \end{aligned}$$

Choose $\alpha_n = c > \max\{b\sqrt{pq}, a\sqrt{pq}\}$. Now from Theorem 13,

$$\begin{aligned}
\max_{|k-np|<c\sqrt{n}} \left| \Pr\{X=k\} - \frac{1}{\sqrt{npq}}\varphi\left(\frac{k-np}{\sqrt{npq}}\right) \right| &= \max_{|k-np|<c\sqrt{n}} \left| \frac{\Pr\{X=k\}\sqrt{npq}}{\varphi\left(\frac{k-np}{\sqrt{npq}}\right)} - 1 \right| \frac{1}{\sqrt{npq}}\varphi\left(\frac{k-np}{\sqrt{npq}}\right) \\
&= O\left(\frac{1}{\sqrt{n}}\right) O\left(\frac{1}{\sqrt{n}}\right) \\
&= O\left(\frac{1}{n}\right).
\end{aligned}$$

Since there are about $\sqrt{n}$ terms, the total error in (I) is about $O(1/\sqrt{n})$. ■

►*Proof.* (of Theorem 13) Idea: Stirling's formula.

$$n! \approx \frac{n^n}{e^n}\sqrt{2\pi n} \quad \text{i.e.} \quad \left| \frac{n!}{\frac{n^n}{e^n}\sqrt{2\pi n}} - 1 \right| = O(1/n)$$

and plug this into the Binomial probability mass function. ■

5.7 Convergence in Distribution or Convergence in Law

Assume that $\{X_n\}_{n=1}^{\infty}$ and X are random variables. The associated distribution functions $\{F_n\}_{n=1}^{\infty}$ and F. Also, let $\{\mu_n\}_{n=1}^{\infty}$ and μ be corresponding probability measures of $\{X_n\}_{n=1}^{\infty}$ and X.

Definition 13. *A sequence of distribution $\{F_n\}$ is said to converge* ***weakly*** *to another distribution F if*

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

at every continuity point x of F. Denoted by $F_n \Rightarrow F$. That is, we only consider points-wise convergence on continuity points of F and disregard the discontinuity points. Since discontinuity points are countable, we don't care.

Definition 14. *A sequence of probability measures $\{\mu_n\}_{n=1}^{\infty}$ is said to converge* ***weakly*** *to a probability measure μ if $\forall A = (-\infty, x]$,*

$$\mu_n\{A\} \to \mu\{A\} \quad \text{where} \quad \mu\{x\} = 0.$$

Denoted by $\mu_n \Rightarrow \mu$.

Definition 15. *A sequence of random variables $\{X_n\}_{n=1}^{\infty}$ is said to be converge to X in distribution or in law if $F_n \to F$ weakly. Denoted by $X_n \xrightarrow{d} X$.*

Definition 16 (equivalent definition of Definition 15)**.** $\mu_n \Rightarrow \mu \iff \forall f \in C_b(\mathbb{R})$,

$$\int f\,d\mu_n \to \int f\,d\mu$$

as $n \to \infty$ where $C_b(\mathbb{R})$ denotes all bounded continuous function on $\mathbb{R}$.

Remark 9. $F_n \Rightarrow F \implies \exists\{X_n\}_{n=1}^{\infty}, X$ s.t. $X_n \xrightarrow{d} X$.

Theorem 14 (Skorohod's)**.** *Assume that $\mu_n \Rightarrow \mu$, then there exists random variables X_n and X on a common probability space $(\Omega, \mathscr{F}, \Pr)$ s.t. X_n has distribution μ_n, X has distribution μ and*

$$X_n(\omega) \to X(\omega) \quad \forall \omega \in \Omega$$

point-wise (almost surely).

►*Proof.* Let $(\Omega, \mathscr{F}, \Pr)$ where $\Omega = (0,1)$, $\mathscr{F}$ be Borel field of $(0,1)$, Pr be Lebesgue measure on $(0,1)$.
Construction of X_n and X: $\forall 1 < \omega < 1$,

$$\begin{aligned} X_n(\omega) &= \inf\{x : \omega \le F_n(x)\} \\ X(\omega) &= \inf\{x : \omega \le F(x)\} \end{aligned} \quad \omega \le F_n(x) \iff X_n(\omega) \le x.$$

Claim: $X_n(\omega) \to X(\omega)$ $\forall \omega \in \Omega$. (Idea: X_n and X are essentially inverse functions of F_n and F. So $F_n(x) \to F(x) \implies X_n(\omega) \to X(\omega)$.)
$\forall \omega \in (0,1)$, $\forall \varepsilon > 0$, choose x s.t.

$$X(\omega) - \varepsilon < x < X(\omega) \quad \text{and} \quad \mu\{x\} = 0.$$

That is, x is a continuity point of F. So

$$F(x) < \omega.$$

By the assumption,

$$\begin{aligned} F_n(x) \to F(x) \quad \text{as} \quad n \to \infty &\implies \text{when } n \ggg 1, F_n(x) < \omega \\ &\implies X(\omega) - \varepsilon < x < X_n(\omega) \\ &\implies X(\omega) - \varepsilon < \liminf_{n\to\infty} X_n(\omega), \quad \forall \varepsilon \\ &\implies X(\Omega) \le \liminf_{n\to\infty} X_n(\omega). \end{aligned}$$

Next, if $\omega < \omega'$, choose y s.t.

$$X(\omega') \le y < X(\omega') + \varepsilon,\ \mu\{y\} = 0 \implies \omega < \omega' \le F(y).$$

$F_n(y) \to F(y) \implies \omega \le F_n(y)$ $\forall n \ggg 1$.

$$\implies X_n(\omega) \le y < X(\omega') + \varepsilon \implies \limsup_{n\to\infty} X_n(\omega) \le X(\omega') + \varepsilon \quad \varepsilon > 0$$

That is, $\limsup_{n\to\infty} X_n(\omega) \le X(\omega')$ $(\omega < \omega')$. Combined with the previous result, if $X(\omega)$ is "continuous" at ω, then $\lim_{n\to\infty} X_n(\omega) = X(\omega)$. By definition, X is non-decreasing. The set of discontinuity points of $X(\omega)$ is countable. (probability measure $= 0$) We now define $X_n(\omega) = X(\omega) = 0$ for all discontinuity points ω of X. Then

$$\lim_{n\to\infty} X_n(\omega) = X(\omega).$$

■

Theorem 15. *$\mu_n \Rightarrow \mu$ (or $F_n \Rightarrow F$) iff $\forall f \in C_b(\mathbb{R})$,*

$$\int f d\mu_n \to \int f d\mu$$

as $n \to \infty$.
►*Proof.*
($\Rightarrow$) Assume $\mu_n \Rightarrow \mu$. By Skorohod's Theorem, $\exists$ random variables X_n and X s.t.

$$X_n(\omega) \to X(\omega) \quad \forall \omega \in \Omega$$

Observe that

$$\int f d\mu_n = \mathbb{E}[f(X_n)] \quad \int f d\mu = \mathbb{E}[f(X)].$$

Since f is a continuous function, $f(X_n(\omega)) \to f(X(\omega))$ $\forall \omega \in \Omega$. By bounded convergence theorem,

$$\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)] \iff \int f d\mu_n \to \int f d\mu.$$

($\Leftarrow$) Let $F_n(X) = \mu_n(-\infty, x]$, $F_n(X) = \mu(-\infty, x]$. We want to show that $F_n(X) \to F(X)$ for any continuity point of F.
Suppose $x < y$. Define

$$f(t) = \begin{cases} 1 & \text{if } t \le x \\ (y-t)/(y-x) & \text{if } x \le t \le y \\ 0 & \text{if } t \ge y \end{cases}$$

So $f \in C_b(\mathbb{R})$. Now $F_n(X) = \mu_n(-\infty, x] \le \int f d\mu_n$ and $\int f d\mu \le F(y)$. Since $\int f d\mu_n \to \int f d\mu \implies \limsup_{n\to\infty} X_n(x) \le F(y) \downarrow F(x)$ as $y \downarrow x$. (F is right-continuous)
On the other hand, if $u < x$, then by the same argument, we have

$$F(u) \le \liminf_n F_n(x).$$

If x is a continuity point of F, then $F(u) \to F(x)$ as $u \uparrow x$ ($u < x$).
Combining both $\lim_{n\to\infty} F_n(x) = F(x)$ at continuity point x of F.

∎

Theorem 16. *Convergence in probability $\implies$ convergence in distribution.*
Remark 10. *Convergence almost surely $\implies$ convergence in distribution.*
▶ *Proof.* By the equivalence definition, it suffices to show that

$$\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)] \quad \forall f \in C_b(\mathbb{R}).$$

Then $X_n \xrightarrow{d} X$.
Assume the contrary, that the statement is false, i.e., we can choose a subsequence N' s.t.

$$\inf_{n\in N'} |\,\mathbb{E}[f(X_n)] - \mathbb{E}[f(X)]\,| > 0. \tag{10}$$

Since $X_n \xrightarrow{p} X$, for the subsequence $\{X_n\}_{n\in N'}$, $\exists$ another subsequence $N'' \subset N'$ s.t. $X_n \xrightarrow{as} X$ as $n \to \infty$, $n \in N''$. That is, we have

$$\mathbb{E}[f(X_n)] \to \mathbb{E}_{f(X)}[\quad] \text{ as } n \to \infty,\ n \in N'' \text{ where } f \in C_b(\mathbb{R}).$$

This contradicts with (10). ∎

Theorem 17 (Continuity)**.** *$X_n \xrightarrow{d} X$ iff $\phi_n(t) = \mathbb{E}\left[e^{itX_n}\right] \to \phi(t) = \mathbb{E}\left[e^{itX}\right]$ $\forall t$.*
Corollary 1 (Central Limit Theorem)**.** *Let $X_1, \ldots$ be iid with $\mathbb{E}[X_1] = \mu$ and $\mathbb{E}\left[X_1^2\right] < \infty$ (i.e., $\mathrm{Var}(X_1) < \infty$). Define $S_n = \sum_{i=1}^n X_i$, then*

$$\frac{S_n - n\mu}{\sqrt{n}\sqrt{\mathrm{Var}(X_1)}} \to \mathcal{N}(0,1).$$

That is,

$$\Pr\left\{\frac{S_n - n\mu}{\sqrt{n}\sqrt{\mathrm{Var}(X_1)}} \le x\right\} \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left\{-\frac{y^2}{2}\right\} dy.$$

▶ *Proof.* (of Continuity Theorem)
($\Rightarrow$) $X_n \xrightarrow{d} X$, i.e., $\mu_n \Rightarrow \mu$, where $\mu_n = p \cdot X_n^{-2}$ and $\mu = p \cdot X^{-1}$. (probability measures corresponding to X_n and X). By the equivalence definition of weak convergence, choose $f(t) = e^{itX} \in C_n(\mathbb{R})$. Hence,

$$\mathbb{E}\left[e^{itX}\right] = \phi_n(t) = \int f d\mu_n \to \int f d\mu = \mathbb{E}\left[e^{itX}\right] = \phi(t) \quad \forall t.$$

($\Leftarrow$) Omitted. (We need Helly's selection Theorem & the tightness of probability measures) ■

►*Proof.* (of Central Limit Theorem) It suffices to consider $X_1, X_2, ...$ iid with $\mathbb{E}[X_1] = 0$ and $\mathrm{Var}(X_1) = 1$. We want to show that

$$\frac{S_n}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1) \quad \text{where} \quad S_n = X_1 + \cdots + X_n. \iff \phi_{S_n/\sqrt{n}}(t) \to \phi_{\mathcal{N}(0,1)}(t) = \exp\left\{-\frac{t^2}{2}\right\}.$$

Observe that

$$\begin{aligned}
\phi_{S_n/\sqrt{n}}(t) = \mathbb{E}\left[\exp\left\{it\frac{S_n}{\sqrt{n}}\right\}\right] &= \mathbb{E}\left[\exp\left\{it\frac{X_1 + \cdots + X_n}{\sqrt{n}}\right\}\right] \\
&= \left(\mathbb{E}\left[\exp\left\{it\frac{X_1}{\sqrt{n}}\right\}\right]\right)^n \\
&= \left(\phi_{X_1/\sqrt{n}}(t)\right)^n.
\end{aligned}$$

By Taylor's formula,

$$\phi_{X_1/\sqrt{n}} = 1 + i\left(\frac{t}{\sqrt{n}}\right)\mathbb{E}[X_1] - \frac{1}{2}\left(\frac{t}{\sqrt{n}}\right)^2 \mathbb{E}[X_1]^2 + o\left(\frac{t^2}{n}\right) = 1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right).$$

So we have

$$\left[\phi_{X_1/\sqrt{n}}(t)\right]^n = \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n \to \exp\left\{-\frac{t^2}{2}\right\}$$

as $n \to \infty$. ■

Example.

1. Ten fair dice are rolled. Find an approximation of probability that the sum is between 30 and 40 using CLT.

   Sol: $X_1, ..., X_{10} \overset{\text{iid}}{\sim} (\mathbb{E}[X_1], \mathrm{Var}(X_1)) = (7/2, 35/12)$. Compute

$$\begin{aligned}
\Pr\{30 \le S_{10} \le 40\} &= \Pr\left\{\frac{30 - 10\cdot\frac{7}{2}}{\sqrt{10}\sqrt{35/12}} \le S_{10} \le \frac{40 - 10\cdot\frac{7}{2}}{\sqrt{10}\sqrt{35/12}}\right\} \\
&\approx \frac{1}{\sqrt{2\pi}}\int_{-5/\sqrt{350/12}}^{5/\sqrt{350/12}} \exp\left\{-\frac{t^2}{2}\right\} dt
\end{aligned}$$

2. $X_1, ... \overset{\text{iid}}{\sim} \mathcal{U}(0,1)$. Approximate $\Pr\left\{\sum_{i=1}^{10} X_i > 6\right\}$. Sol: $\mathbb{E}[X_1] = 1/2$ and $\mathrm{Var}(X_1) = 1/12$.

$$\begin{aligned}
\Pr\{S_{10} > 6\} &= \Pr\left\{\frac{S_{10} - 10\cdot\frac{1}{2}}{\sqrt{10}\sqrt{1/12}} > \frac{6 - 10\cdot\frac{1}{2}}{\sqrt{10}\sqrt{1/12}}\right\} \\
&\approx \frac{1}{\sqrt{2\pi}}\int_{\sqrt{1.2}}^{\infty} \exp\left\{-\frac{t^2}{2}\right\} dt.
\end{aligned}$$

6 Central Limit Theorem (Lindeberg Theorem)

Definition 17 (Triangular array)**.** *Consider a array of random variables*

$$\begin{matrix}
X_{11} & X_{12} & \cdots & X_{1r_1} \\
X_{21} & X_{22} & \cdots & X_{2r_2} \\
\vdots & \vdots & & \vdots \\
X_{n1} & X_{n2} & \dots & X_{nr_4} \\
\vdots & \vdots & \vdots & \vdots
\end{matrix}$$

where r_n is the length of each row. Each row is independent.

Assume that

- $\mathbb{E}[X_{nk}] = 0,\ 1 \le k \le r_n$.
- $\sigma_{nk}^2 = \mathbb{E}\left[X_{nk}^2\right] < \infty$
- $t_n^2 = \sum_{k=1}^{r_n} \sigma_{nk}^2$

Definition 18 (Lindeberg Condition)**.** *For any $\varepsilon > 0$,*

$$\lim_{n\to\infty} \sum_{k=1}^{r_n} \frac{1}{t_n^2} \int_{|X_{nk}|=\varepsilon t_n} X_{nk}^2 dp = 0. \tag{11}$$

Theorem 18. *Assume that for each n, $X_{n1}, X_{n2}, ..., X_{nr_n}$ are independent satisfying the assumptions. Suppose that (11) holds. Then*

$$\frac{S_n}{t_n} \xrightarrow{d} \mathcal{N}(0,1)$$

where $S_n = \sum_{k=1}^{r_n} X_{nk}$.

Remark 11. *This theorem implies the CLT we used previously. Suppose that $r_n = n$ and $X_{nk} = X_k$, then $S_n = X_1 + \cdots + X_n$.*
Consider the case $\mathbb{E}[X_k] = 0$ and $\mathbb{E}\left[X_k^2\right] = n\sigma^2\ \forall 1 \le k \le n$. So

$$t_n^2 = \sum_{k=1}^{n} \mathbb{E}\left[X_k^2\right] = n\sigma^2$$

Check condition (11): In this case, (11) reduces to

$$\lim_{n\to\infty} \frac{1}{n\sigma^2} \sum_{k=1}^{n} \int_{|X_1|\ge\varepsilon\sigma\sqrt{n}} X_1^2 dp = \lim_{n\to\infty} \frac{1}{\sigma^2} \int_{|X_1|\ge\varepsilon\sigma\sqrt{n}} X_1^2 dp$$

►*Proof.* (of Lindeberg) Replacing X_{nk} by X_{nk}/t_n. Then

$$\sum_{k=1}^{r_n} \frac{\mathbb{E}\left[X_{nk}^2\right]}{t_n^2} = \frac{1}{t_n^2} \sum_{k=1}^{r_n} \sigma_{nk}^2 = 1.$$

Observe that

$$\left| e^{itX} - \left(1 + itX - \frac{1}{2} t^2 X^2\right) \right| \le \min\{|tX|^2, |tX|^3\},$$

this implies that the CHF $\phi_{nk}(t)$ of X_{nk} satisfies

$$\left| \phi_{nk}(t) - \left(1 - \frac{1}{2} t^2 \sigma_{nk}^2\right) \right| \le \mathbb{E}\left[\min\{|tX_{nk}|^2, |tX_{nk}|^3\}\right]. \tag{12}$$

For any $\varepsilon > 0$, the right-hand side of (12) is bounded by

$$\begin{aligned}
&\int_{|X_{nk}|<\varepsilon} |tX_{nk}|^3 dp + \int_{|X_{nk}|\ge\varepsilon} |tX_{nk}|^2 dp \le \varepsilon |t|^3 + t^2 \int_{|X_{nk}|\ge\varepsilon} X_{nk}^2 dp \\
\implies \sum_{k=1}^{r_n} &\left[\int_{|X_{nk}|<\varepsilon} |tX_{nk}|^3 dp + \int_{|X_{nk}|\ge\varepsilon} |tX_{nk}|^2 dp \right] \le \varepsilon |t|^3 \sum_{k=1}^{r_n} \sigma_{nk}^2 + t^2 \sum_{k=1}^{r_n} \int_{|X_{nk}|\ge\varepsilon} |X_{nk}|^2 dp \\
&\le \varepsilon |t|^3 + t^2 \sum_{k=1}^{r_n} \int_{|X_{nk}|\ge\varepsilon(t_n)} |X_{nk}|^2 dp \qquad (t_n = 1) \\
&\le \varepsilon |t|^3 \quad \forall \varepsilon \qquad \text{(by Lindeberg Condition)}
\end{aligned}$$

Hence,

$$\lim_{n\to\infty}\sum_{k=1}^{r_n}\left|\phi_{nk}(t)-\left(1-\frac{1}{2}t^2\sigma_{nk}^2\right)\right|\le\varepsilon|t|^3 \quad \varepsilon>0$$

implies that for any fixed $t\in\mathbb{R}$,

$$\lim_{n\to\infty}\sum_{k=1}^{r_n}\left|\phi_{nk}(t)-\left(1-\frac{1}{2}t^2\sigma_{nk}^2\right)\right|=0.$$

<u>Claim</u>: As $n\to\infty$, for any $t\in\mathbb{R}$,

$$\prod_{k=1}^{r_n}\phi_{nk}(t)\to\exp\left\{-\frac{t^2}{2}\right\}.$$

Write

$$\left|\prod_{k=1}^{r_n}\phi_{nk}(t)-\prod_{k=1}^{r_n}\left(1-\frac{1}{2}t^2\sigma_{nk}^2\right)+\prod_{k=1}^{r_n}\left(1-\frac{1}{2}t^2\sigma_{nk}^2\right)-\prod_{k=1}^{r_n}\exp\left\{-\frac{t^2\sigma_{nk}^2}{2}\right\}\right| \tag{13}$$

$$\le\left|\prod_{k=1}^{r_n}\phi_{nk}(t)-\prod_{k=1}^{r_n}\left(1-\frac{1}{2}t^2\sigma_{nk}^2\right)\right| \tag{14}$$

$$+\left|\prod_{k=1}^{r_n}\left(1-\frac{1}{2}t^2\sigma_{nk}^2\right)-\prod_{k=1}^{r_n}\exp\left\{-\frac{t^2\sigma_{nk}^2}{2}\right\}\right| \tag{15}$$

It remains to show that (14) and (15) go to zero as $n\to\infty$.

> <u>Recall</u>: let $z_1,\dots,z_m,w_1,\dots,w_m$ be complex numbers and $|z_i|\le 1$, $|w_i|\le 1$ $\forall 1\le i\le m$, we have
>
> $$|z_1\cdots z_m-w_1\cdots w_m|\le\sum_{j=1}^{m}|z_j-w_j|. \tag{16}$$

To use (16), we need to verify that

$$\left|1-\frac{1}{2}t^2\sigma_{nk}^2\right|\le 1 \tag{17}$$

▶*Proof.* (of (17)) $\forall\varepsilon>0$,

$$\begin{aligned}\sigma_{nk}^2=\int X_{nk}^2dp&=\int_{|X_{nk}|\le\varepsilon}X_{nk}^2dp+\int_{|X_{nk}|\ge\varepsilon}X_{nk}^2dp\\&\le\varepsilon^2+\int_{|X_{nk}|\ge\varepsilon}X_{nk}^2dp\end{aligned}$$

By (L)

$$\max_{1\le k\le r_n}\sigma_{nk}^2\le\varepsilon^2,\ \forall\varepsilon\implies\max_{1\le k\le r_n}\sigma_{nk}^2\to 0$$

Thus, $0<1-\frac{1}{2}t^2\sigma_{nk}^2<1$ for $n\gg 1$. ■ By (16), we have

$$\left|\prod_{k=1}^{r_n}\phi_{nk}(t)-\prod_{k=1}^{r_n}\left(1-\frac{1}{2}t^2\sigma_{nk}^2\right)\right|\le\sum_{k=1}^{r_n}\left|\phi_{nk}(t)-\left(1-\frac{1}{2}t^2\sigma_{nk}^2\right)\right|\to 0\quad\text{as } n\to\infty.$$

Now consider (15), we have (from (16))

$$\left|\prod_{k=1}^{r_n}\left(1-\frac{1}{2}t^2\sigma_{nk}^2\right)-\prod_{k=1}^{r_n}\exp\left\{-\frac{t^2\sigma_{nk}^2}{2}\right\}\right|\le\sum_{k=1}^{r_n}\left|\phi_{nk}(t)-\left(1-\frac{1}{2}t^2\sigma_{nk}^2\right)\right|.$$

Recall that

$$|e^z-(\le|z|1+z)|\le|z|^2\sum_{k=1}^{\infty}\frac{|z|^{k-2}}{k!}\le|z|^2e^{|z|}.$$

Now take $z=-t^2/2\sigma_{nk}^2$.

$$\begin{aligned}\left|\exp\left\{-\frac{t^2}{2}\sigma_{nk}^2\right\}-\left(1-\frac{t^2}{2}\sigma_{nk}^2\right)\right|&\le\frac{t^4}{4}\sigma_{nk}^4\exp\left\{\frac{t^2}{2}\sigma_{nk}^2\right\}\\ \implies\sum_{k=1}^{r_n}\left|\exp\left\{-\frac{t^2}{2}\sigma_{nk}^2\right\}-\left(1-\frac{t^2}{2}\sigma_{nk}^2\right)\right|&\le\frac{t^4}{4}\sum_{k=1}^{r_n}\sigma_{nk}^4\exp\left\{\frac{t^2}{2}\sigma_{nk}^2\right\}\\ &\le\frac{t^4}{4}e^{1/2}\left(\max_{1\le k\le r_n}\sigma_{nk}^2\right)\sum_{k=1}^{r_n}\sigma_{nk}^2\end{aligned}$$

where $\max_{1\le k\le r_n}\sigma_{nk}^2\to 0$ as $n\to\infty$. Therefore, (15) converges to 0 as $n\to\infty$. ∎

7 Strong Law of Large Number (SLLN)

Recall that WLLN says that $X_1,\ldots,X_n\overset{\text{iid}}{\sim}(\mathbb{E}[X_1]=\mu,\operatorname{Var}(X_1)<\infty)$. $\forall\varepsilon>0$,

$$\Pr\left\{\left|\frac{S_n}{n}-\mu\right|\ge\varepsilon\right\}\to 0\text{ as }n\to\infty.$$

That is, $S_n/n\xrightarrow{p}\mu$. Also recall that in the proof of WLLN, we used Chebyshev's inequality, which requires finite variance. To proof SLLN, we require that the fourth moment is finite.

Theorem 19 (Strong Law of Large Number)**.** *Assume that $X_1,X_2,\ldots$ are independent random variables. Suppose that $\exists M>0$ s.t.*

$$\mathbb{E}\left[|X_i-\mathbb{E}[X_i]|^4\right]\le M\quad\forall i.\tag{fourth moment}$$

Denote $S_n=X_1+\cdots+X_n$. Then

$$\frac{S_n-\mathbb{E}[S_n]}{n}\xrightarrow{as}0.$$

In particular, if $\mathbb{E}[X_i]=\mu\ \forall i$, then $S_n/n\xrightarrow{as}\mu$.

Remark 12. *Replace X_i by $X_i-\mathbb{E}[X_i]$. It suffices to assume $\mathbb{E}[X_i]=0\ \forall i$.*
Recall that $X_n\xrightarrow{as}X$ as $n\to\infty$ iff $\forall\varepsilon>0$,

$$\Pr\{|X_n-X|\ge\varepsilon\text{ i.o.}\}=0,$$

which is implied by the first Borel-Cantelli Lemma. Therefore, we only need to check

$$\sum_{n=1}^{\infty}\Pr\{|X_n-X|\ge\varepsilon\}<\infty\quad\forall\varepsilon>0.\tag{from Borel-Cantelli}$$

Lemma 7. *Assume that $\sum_{k=1}^{\infty}\mathbb{E}[|S_k/k|^\alpha]<\infty$ for some $\alpha>0$, then $S_n/n\xrightarrow{as}0$ as $n\to\infty$.*

►*Proof.* By Chebyshev's inequality,

$$\Pr\left\{\left|\frac{S_k}{k}\right| \geq \varepsilon\right\} = \Pr\left\{\left|\frac{S_k}{k}\right|^\alpha \geq \varepsilon^\alpha\right\} \leq \frac{\mathbb{E}[|S_k/k|^\alpha]}{\varepsilon^2}$$
$$\implies \sum_{k=1}^{\infty} \Pr\left\{\left|\frac{S_k}{k}\right| \geq \varepsilon\right\} \leq \frac{1}{\varepsilon^\alpha} \sum_{k=1}^{\infty} \mathbb{E}[|S_k/k|^\alpha] < \infty$$

■

►*Proof.* (of SLLN) Check:

$$\sum_{n=1}^{\infty} \mathbb{E}\left[\left|\frac{S_n}{n}\right|^4\right] < \infty.$$

Consider the following:

$$\begin{aligned} S_n^4 &= (X_1 + \cdots + X_n)^4 \\ &= \sum_{k=1}^{n} X_k^4 + \sum_{j,k=1;\, j<k}^{n} \frac{4!}{2!2!} X_j^2 X_k^2 \\ &+ \sum_{j\neq k, j\neq l, k<l} \frac{4!}{2!2!} X_j^2 X_k X_l + \sum_{j<k<l<m} 4! X_j X_k X_l X_n + \sum_{j\neq k} \frac{4!}{3!1!} X_j^3 X_k \qquad (\mathbb{E}[\cdot] = 0) \end{aligned}$$

Hence, we have

$$\mathbb{E}\left[S_n^4\right] = \sum_{k=1}^{n} \mathbb{E}\left[X_k^4\right] + \sum_{j,k=1;j<k} 6\, \mathbb{E}\left[X_j^2\right] \mathbb{E}\left[X_k^2\right].$$

By the Cauchy-Schwarz inequality,

$$\mathbb{E}\left[X_j^2\right] \leq \left(\mathbb{E}\left[X_j^4\right]\right)^{1/2} \cdot 1 \leq M^{1/2}$$
$$\mathbb{E}\left[X_k^2\right] \leq \left(\mathbb{E}\left[X_k^4\right]\right)^{1/2} \cdot 1 \leq M^{1/2}$$

Therefore,

$$\begin{aligned} \mathbb{E}\left[S_n^4\right] &\leq nM + 6\frac{n(n-1)}{2} M^{1/2} M^{1/2} = (3n^2 - 2n)M < 3n^2 M \\ \implies \mathbb{E}\left[\left(\frac{S_n}{n}\right)^4\right] &\leq \frac{3}{n^2} M \\ \implies \sum_{n=1}^{\infty} \mathbb{E}\left[\left(\frac{S_n}{n}\right)^4\right] &\leq 3 \cdot \left(\sum_{n=1}^{\infty} \frac{1}{n^2}\right) M < \infty \end{aligned}$$

■

7.1 Monte-Carlo Integration

Assume that $f(x)$ is defined on $[0, 1]$. We want to calculate the integral

$$\int_0^1 f(x)dx.$$

Assume that $U_1, U_2, \ldots \overset{\text{iid}}{\sim} \mathcal{U}(0,1)$. Then

$$\mathbb{E}[f(U_i)] = \int_0^1 f(x)dx$$

Hence, by SLLN,

$$\frac{1}{n}\left[f(U_1) + \cdots + f(U_n)\right] \xrightarrow{as} \mathbb{E}[f(U_1)] = \int_0^1 f(x)dx$$

Therefore, we can use computer simulation to find the integral.

7.2 Jensen's Inequality

Definition 19. *A C^2-function $f(x)$ is convex if $f''(x) \geq 0\ \forall x$.*
Proposition 3 (Jensen's inequality)**.** *Assume that f is convex. Then*

$$f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)]\,.$$

►*Proof.* By Taylor's formula,

$$f(x) = f(\mu) + f'(\mu)(x-\mu) + \frac{f''(\xi)}{2}(x-\mu)^2$$

where ξ is between x and μ. Since $f(x)$ is convex,

$$\begin{aligned} f(x) \geq f(\mu) + f'(\mu)(x-\mu) \quad \forall x \implies \mathbb{E}[f(x)] &\geq f(\mu) + f'(\mu)[\mathbb{E}[X] - \mu] \\ &= f(\mathbb{E}[X]) + 0. \end{aligned}$$

■

Remark 13. *We have the following properties of moments*

$$\mathbb{E}[X]^p \leq \mathbb{E}[X^p] \quad \forall p \geq 2.$$

8 Some Estimates

8.1 One-sided Chebyshev's inequality

Assume that X is a random variable with exponential μ and $\sigma^2 < \infty$. For any $a > 0$, we want to bound $\Pr\{X - \mu \geq a\}$. Note that

$$\Pr\{X - \mu \geq 0\} \leq \Pr\{|X-\mu| \geq a\} \leq \frac{\sigma^2}{a} \qquad \text{(by Chebyshev)}$$

We can do better.
Proposition 4 (One-sided Chebyshev's inequality)**.** *Assume that $\mathbb{E}[X] = 0$ and $\operatorname{Var}(X) = \sigma^2$. Then for any $a > 0$,*

$$\Pr\{X \geq a\} \leq \frac{\sigma^2}{\sigma^2 + a^2}$$

►*Proof.* Consider the following:

$$\Pr\{X \geq a\} = \Pr\{X + b \geq a + b\} \leq \frac{\sigma^2 + b^2}{(a+b)^2} \quad \forall b > 0$$

We want to minimise $(\sigma^2+b^2)/(a+b)^2$ in $b \in (0,\infty)$.

$$g(b) := \frac{\sigma^2+b^2}{(a+b)^2} \quad g'(b) = \frac{2(ab-\sigma^2)}{(a+b)^2} \overset{\text{let}}{=} 0$$

Hence, we have $b = \sigma^2/a$, which implies

$$\frac{\sigma^2+b^2}{(a+b)^2} = \frac{\sigma^2}{\sigma^2+a^2}.$$

■

Remark 14. *We have*

$$\begin{aligned} \Pr\{X-\mu \geq a\} &\leq \frac{\sigma^2}{\sigma^2+a^2} \\ \Pr\{\mu - X \geq a\} &\leq \frac{\sigma^2}{\sigma^2+a^2} \end{aligned} \implies \begin{aligned} \Pr\{X \geq a+\mu\} &\leq \frac{\sigma}{\sigma^2+a^2} \\ \Pr\{X \leq \mu - a\} &\leq \frac{\sigma}{\sigma^2+a^2} \end{aligned}$$

8.2 Chernoff Bounds

Assume that the MGF of X, $M_X(t)$ exists $\forall t$. Observe that $M_X(t) = \mathbb{E}\left[e^{tX}\right]$, hence we have

$$\begin{aligned} \Pr\{X \geq a\} &= \Pr\left\{e^{tX} \geq e^{ta}\right\} \leq \frac{\mathbb{E}\left[e^{tX}\right]}{e^{ta}} = M_X(t)e^{-ta} \quad \text{for } t > 0 \\ \Pr\{X \leq a\} &= \Pr\left\{e^{tX} \geq e^{ta}\right\} \leq \frac{\mathbb{E}\left[e^{tX}\right]}{e^{ta}} = M_X(t)e^{-ta} \quad \text{for } t < 0 \end{aligned}$$

Example. Assume $Z \sim \mathcal{N}(0,1)$ with $M_Z(t) = \exp\left\{t^2/2\right\}$. Consider the Chernoff bound for $a > 0$:

$$\Pr\{Z \geq a\} \leq M_Z(t)e^{-ta} = \exp\left\{\frac{t^2}{2} - ta\right\} \quad \text{for } t > 0$$

Minimise $\frac{t^2}{2} - ta$ in $t \in (0,\infty)$. At $t = a$, we have

$$\min_{t\in(0,\infty)} \left\{\frac{t^2}{2} - ta\right\} = -\frac{a^2}{2}.$$

Hence $\Pr\{Z \geq a\} \leq \exp\left\{a^2/2\right\}$.
Now consider $a < 0$. Since Z is symmetric, $-Z \sim \mathcal{N}(0,1)$. Then,

$$\Pr\{-Z \geq -a\} \leq \exp\left\{-\frac{a^2}{2}\right\} \iff \Pr\{Z \leq a\} \leq \exp\left\{-\frac{a^2}{2}\right\} \quad a > 0.$$

So

$$\begin{aligned} \Pr\{Z \geq a\} &\leq \exp\left\{-a^2/2\right\} \text{ for } a > 0 \\ \text{and } \Pr\{Z \leq a\} &\leq \exp\left\{-a^2/2\right\} \text{ for } a < 0 \end{aligned}$$

Example. Gambling with $\Pr\{X_i = 1\} = \Pr\{X_i = -1\} = 1/2$ where X_i is the winning of each play (independent). Consider $S_n = X_1 + \cdots + X_n$, which denotes the winnings after n plays. We want to bound

$$\Pr\{S_n \geq a\} \text{ for } a > 0.$$

Note that

$$M_{S_n}(t) = \mathbb{E}[\exp\{tS_n\}] = \mathbb{E}[\exp\{tX_1\}]^n$$

since X_i are independent. We have the following:

$$\mathbb{E}[\exp\{tX_1\}] = \frac{e^t + e^{-t}}{2} \implies M_{S_n}(t) = \left(\frac{e^t + e^{-t}}{2}\right)^n.$$

We can estimate

$$\frac{e^t + e^{-t}}{2} \le \exp\{t^2/2\} \quad \forall t \in \mathbb{R}.$$

Hence,

$$M_{S_n}(t) \le \exp\{nt^2/2\}.$$

So, by Chernoff bounds, we have (for $a > 0$)

$$\begin{aligned} \Pr\{S_n \ge a\} &\le M_{S_n}(t)\exp\{-ta\} && (t > 0) \\ &\le \exp\left\{\frac{nt^2}{2} - ta\right\}. \end{aligned}$$

If we choose $t = a/n$, we have

$$\Pr\{S_n \ge a\} \le \exp\left\{-\frac{a^2}{2n}\right\}$$

9 Markov Chain

One of the most famous uses of Markov Chain: Random walk. Consider a one-dimensional random walk. Let X_1 be a random variable describing the position of the person at the first step. That is, $\Pr\{X_1 = 1\} = p$ and $\Pr\{X_1 = -1\} = 1 - p = q$. Question: We want to know whether the probability of getting back to the origin is zero.

Let $X_0, X_1, ...$ be iid Bernoulli trials. Let $\Pr\{X_k = 1\} = \Pr\{X_k = -1\} = 1/2$ for $k = 1, 2, ...$ Let $\Pr\{X_0 = 0\} = 1$. Define

$$S_n = X_0 + X_1 + \cdots + X_n = X_1 + \cdots + X_n$$

This is a one-dimensional random walk.

$$\begin{aligned} &\Pr\{S_n = k | S_0 = i_0, S_1 = i_1, ..., S_{n-1} = i_{n-1}\} \text{ where } i_0, i_1, ..., i_{n-1} \text{ are integers} \\ = &\Pr\{S_n = k | S_{n-1} = i_{n-1}\} \end{aligned}$$

This is the Markov property (Markov Process).

Example. A machine has two components. There are three states:

$$\text{Three states: } \begin{cases} 0 & \text{No components work} \\ 1 & \text{One component works} \\ 2 & \text{Two components work} \end{cases}$$

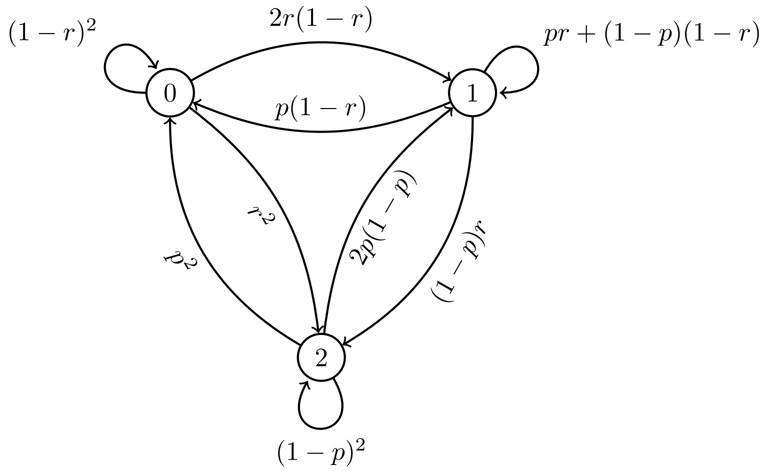
As time $t = n$, any given component has probability p of failing before the next inspection; a component that is not operating at time $t = n$ has probability r of being repaired at $t = n + 1$.

Let R_n be the number of components in operation at $t = n$.

$$\begin{cases} \Pr\{R_{n+1} = 0 | R_n = 0\} = (1-r)^2 \\ \Pr\{R_{n+1} = 1 | R_n = 0\} = 2r(1-r) \\ \Pr\{R_{n+1} = 2 | R_n = 0\} = r^2 \end{cases}$$

$$\begin{cases} \Pr\{R_{n+1} = 0 | R_n = 1\} = p(1-r) \\ \Pr\{R_{n+1} = 1 | R_n = 1\} = pr + (1-p)(1-r) \\ \Pr\{R_{n+1} = 2 | R_n = 1\} = (1-p)r \end{cases}$$

$$\begin{cases} \Pr\{R_{n+1} = 0 | R_n = 2\} = p^2 \\ \Pr\{R_{n+1} = 1 | R_n = 2\} = 2p(1-p) \\ \Pr\{R_{n+1} = 2 | R_n = 2\} = (1-p)^2 \end{cases}$$



Define $P_{ij} = \Pr\{R_{n+1} = j | R_n = i\}$ where $i, j \in \{0, 1, 2\}$. Write the P_{ij}'s into a matrix Π:

$$\Pi = (P_{ij}) = \begin{bmatrix} (1-r^2) & 2r(1-r) & r^2 \\ p(1-r) & pr + (1-p)(1-r) & (1-p)r \\ p^2 & 2p(1-p) & (1-p)^2 \end{bmatrix}$$

Note that $\sum_{j=0}^{2} = P_{ij} = 1 \ \forall i$. Π is called a **stochastic matrix** and P_{ij} is called **transition probability**. Check:

$$\Pr\{R_{n+1} = i_{n+1} | R_0 = i_0, ..., R_n = i_n\} = \Pr\{R_{n+1} = i_{n+1} | R_n = i_n\}$$

This satisfies the Markov property. $\square$

Let S be a finite or countable set (state space). Suppose that to each pair i and j in S, there is associated with non-negative number P_{ij} satisfying $\sum_j P_{ij} = 1 \ \forall i \in S$.

Let $X_0, X_1...$ be a sequence of random variables whose ranges are contained in S. Then this sequence is called a **Markov chain** or **Markov process** if

$$\Pr\{X_{n+1} = j \mid X_0 = i_0, ..., X_n = i_n\} = \Pr\{X_{n+1} = j | X_n = i_n\} = P_{ij} \quad \forall i_0, ..., i_n, j \in S.$$

And $T = (P_{ij})_{i,j \in S}$ is the **Transition Matrix**. For example, a *random walk* is a special example of this where $S = \{0, \pm 1, \pm 2, ...\}$. If

$$\Pr\{X_{n+1} = j \mid X_n = i\} = P_{ij}, \text{ independent of } n$$

In this case, this transition probability is called **stationary**.

Definition 20. *$\alpha_i = \Pr\{X_0 = i\}$ for $i \in S$. $\sum_{i\in S} \alpha_i = 1$. This is called an* ***initial distribution****.*

9.1 Higher Order Transitions

Observe that $\Pr\{X_0 = i_0, X_1 = i_1, X_2 = i_2\}$ where $i_0, i_1, i_2 \in S$.

$$\begin{aligned}
&\Pr\{X_0 = i_0, X_1 = i_1, X_2 = i_2\} \\
&= \Pr\{X_0 = i_0\} \Pr\{X_1 = i_1 \mid X_0 = i_0\} \Pr\{X_2 = i_2 \mid X_0 = i_0, X_1 = i_1\} \\
&= \Pr\{X_0 = i_0\} \Pr\{X_1 = i_1 \mid X_0 = i_0\} \Pr\{X_2 = i_2 \mid X_1 = i_1\} \qquad \text{(Markov property)} \\
&= \alpha_{i_0} P_{i_0 i_1} P_{i_1 i_2}
\end{aligned}$$

In general, we have

$$\Pr\{X_k = i_k, 0 \le k \le m\} = \alpha_{i_0} P_{i_0 i_1} \cdots P_{i_{m-1} i_m}$$

For staring at X_m for any m, we have

$$P_{ij}^{(n)} := \Pr\{X_{n+m} = j \mid X_m = i\} = \sum_{k_1,\ldots,k_{n-1}\in S} P_{ik_1} P_{k_1 k_2} \cdots P_{k_{n-1} j}.$$

This is called the n-th order transition probability. Observe that $P_{ij}^{(n)}$ is the (i,j)-entry of the matrix T^n. In general, we have

$$P_{ij}^{(n+m)} = \sum_{k\in S} P_{ik}^{(n)} P_{kj}^{(m)} \quad \forall n, m$$

This is called **Chapman-Kolmogorov's identity**.

9.2 Classification of States

We want to classify two different states: *persistent* and *transient*. Define $\Pr_i\{A\} = \Pr\{A|X_0 = i\}$. Observe that

$$\begin{aligned}
&\Pr\{X_1 = i_1, ..., X_m = i_m, X_{m+1} = j_1 ..., X_{m+n} = j_n \mid X_0 = i\} \\
&= \Pr\{X_1 = i_1, ..., X_m = i_m \mid X_0 = i\} \Pr\{X_1 = j_1, ..., X_n = j_n \mid X_0 = i_m\}
\end{aligned}$$

$$\text{Let } f_{ij}^{(n)} := \Pr\{X_1 \neq j, X_2 \neq j, ..., X_{n-1} \neq j, X_n = j \mid X_0 = i\}$$

which is the probability of a first visit j at time n for a chain that starts at i.

$$\text{Let } f_{ij} := \Pr\left\{ \bigcup_{n=1}^{\infty} (X_n = j) \,\middle|\, X_0 = i \right\} = \sum_{n=1}^{\infty} f_{ij}^{(n)}$$

be the probability of an eventual visit.

Definition 21. *A state i is called* persistent *if a system starting at i is certain that sometime it returns to i, i.e., $f_{ii} = 1$. The state i is called* transient *if $f_{ii} < 1$.*

Theorem 20. *We have the following equivalent definition of persistent and eventually.*

- *The state i is transient iff $\Pr_i\{X_n = i \text{ i.o.}\} = 0$ and iff $\sum_n P_{ii}^{(n)} < \infty$.*
- *The state i is persistent iff $\Pr_i\{X_n = i \text{ i.o.}\} = 1$ and iff $\sum_n P_{ii}^{(n)} = +\infty$.*

Observations: Let $n_1, ..., n_k$ be integers satisfying that $1 \le n_1 < \cdots < n_k$. Consider the following event:

$$A = \left\{ \begin{matrix} X_1 \neq j, X_2 \neq j, ..., X_{n_1 - 1} \neq j, X_{n_1} = j, \\ X_1 \neq j, X_2 \neq j, ..., X_{n_2 - 1} \neq j, X_{n_2} = j, \\ \vdots \\ X_1 \neq j, X_2 \neq j, ..., X_{n_k - 1} \neq j, X_{n_k} = j \end{matrix} \right\}$$

Then we have

$$\begin{aligned} \Pr_i\{A\} &= \Pr_i\{X_1 \neq j, ..., X_{n_1-1} \neq j, X_{n_1} = j\} \\ &\quad \cdot \Pr_j\{X_1 \neq j, ..., X_{n_2-n_1-1} \neq j, X_{n_2-n_1} = j\} \\ &\quad\quad \vdots \\ &\quad \cdot \Pr_j\big\{X_1 \neq j, ..., X_{n_k-n_{k-1}-1} \neq j, X_{n_k-n_{k-1}} = j\big\} \\ &= f_{ij}^{(n_1)} f_{jj}^{(n_2-n_1)} \cdots f_{jj}^{(n_k-n_{k-1})} \end{aligned}$$

Hence

$$\sum_{n_1 < \cdots < n_k} f_{ij}^{(n_1)} f_{jj}^{(n_2-n_1)} \cdots f_{jj}^{(n_k-n_{k-1})}$$

Let $k \to \infty$, then

$$\Pr_i\{X_n = j \text{ i.o.}\} = \lim_{k\to\infty} \left(\sum_{n_1 < \cdots < n_k} f_{ij}^{(n_1)} f_{jj}^{(n_2-n_1)} \cdots f_{jj}^{(n_k-n_{k-1})} \right)$$

Note that

$$\begin{aligned} \sum_{n_1 < \cdots < n_k} f_{ij}^{(n_1)} f_{jj}^{(n_2-n_1)} \cdots f_{jj}^{(n_k-n_{k-1})} &= \sum_{n_1=1}^{\infty} \sum_{n_2=n_1+1}^{\infty} \cdots \sum_{n_k=n_{k-1}+1}^{\infty} f_{ij}^{(n_1)} f_{jj}^{(n_2-n_1)} \cdots f_{jj}^{(n_k-n_{k-1})} \\ &= f_{ij} \underbrace{f_{jj} \cdots f_{jj}}_{k-1 \text{ times}}. \end{aligned}$$

So we have the following as $k \to \infty$:

$$\Pr_i\{X_n = j \text{ i.o.}\} = \begin{cases} 0 & \text{if } f_{jj} < 1 \\ f_{ij} & \text{if } f_{jj} = 1 \end{cases}$$

In particular,

$$\Pr_j\{X_n = j \text{ i.o.}\} = \begin{cases} 0 & \text{if } f_{jj} < 1 \\ 1 & \text{if } f_{jj} = 1 \end{cases}$$

►*Proof.*

- $\Pr_i\{X_n = i \text{ i.o.}\} = 0 \iff \sum_n P_{ii}^{(n)} < \infty$.
  ($\Leftarrow$) By the first Borel-Cantelli Lemma, this direction is trivial. Hence $f_{ii} < 1$.

($\Rightarrow$) It remains to show that $f_{ii} < 1 \implies \sum_n P_{ii}^{(n)} < \infty$. To do this, we write

$$\begin{aligned}
P_{ij}^{(n)} = \Pr_i\{X_n = j\} &= \sum_{k=0}^{n-1} \Pr_i\{X_1 \neq j, ..., X_{n-k-1} \neq j, X_{n-k} = j, X_n = j\} \\
&= \sum_{k=0}^{n-1} \Pr_i\{X_1 \neq j, ..., X_{n-k-1} \neq j, X_{n-k} = j\} \Pr_j\{X_k = j\} \\
&= \sum_{k=0}^{n-1} f_{ij}^{(n-k)} P_{jj}^{(k)}
\end{aligned}$$

Take $j = i$, note that $P_{ii}^{(n)} = 1 = \Pr\{X_0 = i \mid X_0 = i\} = 1$.

$$\begin{aligned}
\sum_{s=1}^{n} P_{ii}^{(n)} = \sum_{s=1}^{n}\sum_{k=0}^{s-1} f_{ii}^{(s-k)} P_{ii}^{(k)} = \sum_{k=0}^{n-1} P_{ii}^{(k)} \sum_{s=k+1}^{n} f_{ii}^{(s-k)} &\leq \sum_{k=0}^{n} P_{ii}^{(k)} \sum_{l=1}^{\infty} f_{ii}^{(l)} \\
&= \sum_{k=0}^{n} P_{ii}^{(k)} f_{ii} + P_{ii}^{(n)} f_{ii} \\
&= \sum_{k=1}^{n} P_{ii}^{(n)} f_{ii} + f_{ii} \\
\implies (1 - f_{ii}) \sum_{s=1}^{n} P_{ii}^{(n)} \leq f_{ii}
\end{aligned}$$

Hence if $f_{ii} < 1$, we have

$$\sum_{s=1}^{n} P_{ii}^{(n)} \leq \frac{f_{ii}}{1 - f_{ii}} < \infty \quad \forall n.$$

- From the results of the proof above , we have

$$\begin{aligned}
f_{ii} = 1 &\iff \Pr_i\{X_n = i \text{ i.o.}\} = 1 \\
&\iff \sum_n P_{ii}^{(n)} = \infty.
\end{aligned}$$

■

9.3 Pólya Theorem

Consider the symmetric random walk in d dimension. Let state

$$S = \mathbb{Z}^d = \{x = (x_1, ..., x_d) : x_i \in \mathbb{Z}\}$$

For each $x \in S$, x has $2d$ neighbors of the form $y = (x_1, ..., x_i \pm 1, ..., x_d)$ and $P_{xy} = (2d)^{-1}$.
Theorem 21. *When $d \in \{1, 2\}$, all states are persistent. When $d \geq 3$, all states are transient.*
▶*Proof.*

($d = 1$) We want to show that $\sum_n P_{ii}^{(n)} = \infty$. $P_{ii}^{(n)}$ is independent of i. We denote $P_{ii}^{(n)}$ when $d = 1$ as $a_n^{(1)}$. Observe that $a_{2n+1}^{(1)} = 0$. (If we take an odd number of steps, we cannot get back) Thus, it suffcies to estimate $a_{2n}^{(1)}$. Hence

$$a_{2n}^{(1)} = \binom{2n}{n} \left(\frac{1}{2}\right)^{2n} = \frac{(2n)!}{(n!)^2} \frac{1}{2^{2n}}.$$

By Stirling's formula, we have

$$n! \approx n^n e^{-n}\sqrt{2\pi n} \implies a_{2n}^{(1)} \approx \frac{(2n)^{2n}e^{-2n}\sqrt{2\pi(2n)}}{\left(n^n e^{-n}\sqrt{2\pi n}\right)^2}\frac{1}{2^{2n}} = (\pi n)^{-1/2}$$
$$\implies \sum_n a_n^{(1)} = \sum_n a_{2n}^{(1)} \approx \sum_n (\pi n)^{-1/2} = \infty \qquad \square$$

$(d=2)$ $a_{2n+1}^{(2)} = 0$. So we compute $a_{2n}^{(2)}$.

$$\begin{aligned}
a_{2n}^{(2)} = \sum_{u=0}^{n}\frac{(2n)!}{u!u!(n-u)!(n-u)!}\left(\frac{1}{4}\right)^{2n} &= \left(\frac{1}{4}\right)^{2n}\binom{2n}{n}\sum_{u=0}^{n}\frac{n!n!}{u!u!(n-u)!(n-u)!} \\
&= \left(\frac{1}{4}\right)^{2n}\binom{2n}{n}\underbrace{\sum_{u=0}^{n}\binom{n}{u}\binom{n}{n-u}}_{\text{Vandermonde Identity}=\binom{2n}{n}} \\
&= \left(\frac{1}{2}\right)^{2n}\binom{2n}{n}\left(\frac{1}{2}\right)^{2n}\binom{2n}{n} = (a_{2n}^{(1)})^2 \approx (\pi n)^{-1}
\end{aligned}$$

Therefore, $\sum_n (\pi n)^{-1} = \infty$. $\square$

$(d=3)$ Similar to above

$$a_{2n}^{(3)} = \sum_{u+v=n}\frac{(2n)!}{(u!)^2(v!)^2[(n-u-v)]^2}\left(\frac{1}{6}\right)^{2n} = \sum_{l=0}^{n}\binom{2n}{2l}\left(\frac{1}{3}\right)^{2n-2l}\left(\frac{2}{3}\right)^{2l}a_{2n-2l}^{(1)}a_{2l}^{(2)}$$

In the above equation, the terms $l=0$, $l=n \le O(n^{-3/2})$. It is enough to estimate

$$\sum_{l=1}^{n-1}\binom{2n}{2l}\left(\frac{1}{3}\right)^{2n-2l}\left(\frac{2}{3}\right)^{2l}a_{2n-2l}^{(1)}a_{2l}^{(1)} \le C\sum_{l=1}^{n-1}\binom{2n}{2l}\left(\frac{1}{3}\right)^{2n-2l}\left(\frac{2}{3}\right)^{2l}(2n-2l)^{-1/2}(2l)^{-1}$$

for some constant C. Observe that when $1 \le l \le n-1$,

$$\begin{aligned}
&\begin{cases}(2n-2l)^{-1/2} \le 2n^{1/2}(2n-2l)^{-1} \le 4n^{1/2}(2n-2l+1)^{-1} \\ (2n)^{-1} \le 2(2l+1)^{-1}\end{cases} \\
&\implies C\sum_{l=1}^{n-1}\binom{2n}{2l}\left(\frac{1}{3}\right)^{2n-2l}\left(\frac{2}{3}\right)^{2l}(2n-2l)^{-1/2}(2l)^{-1} \\
&\le Cn^{1/2}\frac{(2n)!}{(2n+2)!}\sum_{l=1}^{n-1}\binom{2n+2}{2l+1}\left(\frac{1}{3}\right)^{2n-2l+1}\left(\frac{2}{3}\right)^{2l+1} \\
&\le \tilde{C}n^{1/2}\frac{1}{(2n+2)(2n+1)} = O(n^{-3/2})
\end{aligned}$$

Thus, $a_{2n}^{(3)} = O(n^{-3/2})$, implying $\sum_n a_{2n}^{(3)} < \infty$.

■

9.4 Classification of Markov Chain

We want to also classify a Markov chain to be either *transient* or *persistent*, not only locally for each state. Therefore, we are motivated to define

Definition 22. *A Markov chain is called **irreducible** if* $\forall i, j \in S$ $\exists n \in \mathbb{N}$ *s.t.*

$$P_{ij}^{(n)} > 0 \iff f_{ij} > 0$$

Theorem 22. *Assume that the Markov chain is irreducible, then one of the following holds:*

(i) *All states are transient.* $\Pr_i\left\{\bigcup_j [X_n = j \text{ i.o.}]\right\} = 0$ $\forall i$ *and* $\sum_n P_{ij}^{(n)} < \infty$ $\forall i, j$.

(ii) *All states are irreducible.* $\Pr_i\left\{\bigcap_j [X_n = j \text{ i.o.}]\right\} = 1$ $\forall i$ *and* $\sum_n P_{ij}^{(n)} = \infty$ $\forall i, j$.

▶*Proof.* Consider the two parts separately:

(i) Let $i, j \in S$. From the irreducibility, $\exists r > 0, s > 0$ s.t. $P_{ij}^{(r)} > 0$ and $P_{ji}^{(s)} > 0$. Observe

$$P_{ii}^{(r+s+m)} \geq P_{ij}^{(r)} P_{jj}^{(m)} P_{ji}^{(s)}.$$

So if i is transient , i.e. $\sum_n P_{ii}^{(n)} < \infty$, then $\sum_n P_{jj}^{(m)} < \infty \implies j$ is transient. That is, if one state is transient, then all states are transient. Since j is transient, we have $f_{jj} < 1$,

$$f_{jj} < 1 \iff \Pr_i\{X_n = j \text{ i.o.}\} = 0 \quad \forall i, j \iff \Pr_i\left\{\bigcup_j \{X_n = j \text{ i.o.}\}\right\} = 0$$

Now, observe that

$$\begin{aligned}
\sum_{n=1}^{\infty} P_{ij}^{(n)} = \sum_{n=1}^{\infty}\left(\sum_{s=0}^{n-1} f_{ij}^{(n-s)} P_{jj}^{(s)}\right) = \sum_{n=1}^{\infty}\sum_{\nu=1}^{n} f_{ij}^{(\nu)} P_{jj}^{(n-\nu)} &= \sum_{\nu=1}^{\infty}\sum_{n-\nu=0}^{\infty} f_{ij}^{(\nu)} P_{jj}^{(n-\nu)} \\
&= \sum_{\nu=1}^{\infty} f_{ij}^{(\nu)} \sum_{m=0}^{\infty} P_{jj}^{(m)} \\
&= f_{ij} \sum_{m=0}^{\infty} P_{jj}^{(m)} \\
&\leq \sum_{m=0}^{\infty} P_{jj}^{(m)} < \infty.
\end{aligned}$$

Therefore, $\sum_n P_{ij}^{(n)} < \infty$.

(ii) Assume i is persistent, then any state is persistent, otherwise, it will contradict with (i). Since j is persistent, $\Pr_j\{X_n = j \text{ i.o.}\} = 1$. But we want to prove $\Pr_i\{X_n = j \text{ i.o.}\} = 1$ $\forall i, j$. If this holds, then $\Pr_i\left\{\bigcap_j \{X_n = j \text{ i.o.}\}\right\} = 1$ by continuity. Observe that

$$\begin{aligned}
P_{ji}^{(m)} = \Pr_j\{X_m = i\} &= \Pr_j\{[X_m = i] \cap [X_n = j \text{ i.o.}]\} \\
&= \sum_{n>m} \Pr_j\{X_m = i, X_{m+1} \neq j, ..., X_{n-1} \neq j, X_n = j\} \\
&= \sum_{n>m} \Pr_j\{X_m = i\} \Pr_i\{X_1 \neq j, ..., X_{n-m-1} \neq j, X_{n-m} = j\} \\
&= P_{ji}^{(m)} \sum_{n>m} f_{ij}^{(n-m)} \\
&= P_{ji}^{(m)} \sum_{n=1}^{\infty} f_{ij}^{(n)} = P_{ji}^{(m)} f_{ij} \quad \forall m.
\end{aligned}$$

Since S is irreducible, $\exists m > 0$ s.t.

$$P_{ji}^{(m)} > 0 \implies f_{ij} = 1.$$

Recall that $\Pr_i\{X_n = j \text{ i.o.}\} = f_{ij} = 1$ if $f_{jj} = 1$. Therefore, $\Pr_i\left\{\bigcap_j [X_n = j \text{ i.o.}]\right\} = 1$. It remains to show that $\sum_n P_{ij}^{(n)} = \infty\ \forall i, j$. If $\sum_n P_{ij}^{(n)} < \infty$ for some i, j, by the first Borel Cantelli Lemma,

$$\Pr_i\{X_n = j \text{ i.o.}\} = 0 \rightarrow\!\!\!\leftarrow .$$

■

Remark 15. *If a Markov is irreducible, we say that a Markov chain is transient or persistent.*
Remark 16. *If S is finite and irreducible, then S is always persistent. Note that $\sum_j P_{ij}^{(n)} = 1$. If $\sum_n P_{ij}^{(n)} < \infty$ (transient case), then $\sum_{j\in S}\sum_n P_{ij}^{(n)} < \infty$ since S is finite. But $\sum_n \sum_j P_{ij}^{(n)} = \sum_n 1 = \infty$.*

9.5 Stationary Distribution

Let the Markov chain has an initial distribution, i.e., $\Pr\{X_0 = i\} = \pi_i\ \forall i \in S$ and $\sum_{i\in S} = 1$. Assume that

$$\sum_{i\in S} \pi_i P_{ij} = \pi_j \tag{18}$$

then

$$\begin{aligned}\sum_{j\in S}\sum_{i\in S} \pi_i P_{ij} P_{jk} = \sum_{j\in S} \pi_j P_{jk} = \pi_k \implies & \sum_{i\in S} \pi_i P_{ij}^{(2)} = \pi_j \\ \implies & \sum_{i\in S} \pi_i P_{ij}^{(n)} = \pi_j \quad \forall n.\end{aligned}$$

Note that $\sum_{i\in S} \pi_i P_{ij}^{(n)} = \Pr\{X_n = j\}$. Hence we have $\Pr\{X_n = j\} = \pi_j\ \forall n$.
Definition 23. *All probabilities $\{\pi_i\}_{i\in S}$ satisfying (18) are called* ***stationary probability*** *or distribution of S.*
Definition 24. *The state i has period $t = \gcd\{n \geq 1, P_{ii}^{(n)} > 0\}$, i.e., if t is the period of u and $P_{ii}^{(n)} > 0$ then n can be divided by t. Such i is the largest number satisfying this property.*

Assume that the chain is irreducible. $\forall i, j\ \exists r > 0, s > 0$ s.t. $P_{ij}^{(r)} > 0$, $P_{ji}^{(s)} > 0$. Note that

$$P_{ii}^{(r+s+n)} \geq P_{ij}^{(r)} P_{jj}^{(n)} P_{ji}^{(s)} \quad \forall n \geq 0$$

Consider $P_{jj}^{(0)} = 1$. We have

$$P_{ii}^{(r+s)} \geq P_{ij}^{(r)} P_{ji}^{(s)} > 0$$

Assume t_i and t_j are the periods of i and j respectively. So t_i divides $r + s$. Now if $P_{jj}^{(n)} > 0$ for some $n > 0$, then $P_{ii}^{(r+s+n)} > 0$. So t_i divides $r + s + n \implies t_i$ divides n. On the other hand, t_j divides n. Therefore, $t_i \leq t_j$. Since i and j can be switched, we have $t_i = t_j$.

Hence, in the irreducible case, one can speak of the periods of the chain since every state has the same period.
Definition 25. *An irreducible chain is called "aperiodic" if the period is 1.*
Theorem 23. *Suppose that the chain is irreducible and aperiodic. Also assume that there exists a stationary distribution $\{\pi_i\}$ where $\pi_i \geq 0$ and $\sum_i \pi_i = 1$. Then the chain is persistent and*

$$\lim_{n\to\infty} P_{ij}^{(n)} = \pi_j \quad \forall i, j$$

and all π_i are positive and the stationary distribution is unique.

Remark 17. *In this theorem, the effect of the initial distribution wears off as $n \to \infty$. Note that*

$$\Pr\{X_n = j\} = \sum_{i \in S} \alpha_i P_{ij}^{(n)}$$

where α_i is the initial distribution. Then

$$\sum_{i \in S} \alpha_i P_{ij}^{(n)} = \pi_i \text{ as } n \to \infty.$$

Theorem 24. *If an irreducible and aperiodic chain has no stationary distribution, then*

$$\lim_{n \to \infty} P_{ij}^{(n)} = 0 \quad \forall i, j \tag{19}$$

Remark 18. *If the chain is transient, then (19) is trivial. This theorem holds even in the case of persistent chain. If the chain is persistent, then*

$$\sum_n P_{ij}^{(n)} = +\infty \quad \forall i, j \text{ but } \lim_{n \to \infty} P_{ij}^{(n)} = 0$$

That is, $P_ij^{(n)}$ decays to 0 slowly. In this case, the chain is called ***null persistent****.*
Remark 19. *Let $i, j \in S$. Recall $f_{ij}^{(n)} = \Pr_i\{X_1 \neq j, ..., X_{n-2} \neq j, X_n = j\}$. Define $\mu_j = \sum_n n f_{jj}^{(n)}$. (mean of first return times)*
Theorem 25. *Suppose that j is persistent and $\lim_{n \to \infty} P_{ii}^{(n)} = u$, then*

$$u > 0 \iff \mu_j < \infty$$

in which case $\mu_j = 1/u$ or $u = 1/\mu_j$. ($u = 0 \iff \mu_j = \infty$)
Corollary 2. *Suppose that the Markov chain is irreducible.*
 (i) The chain is positive persistent iff $\mu_j < \infty \ \forall j$.
 (ii) The chain is null persistent iff $\mu_j = \infty$.
Theorem 26. *Let S be finite. Assume that the chain is irreducible and aperiodic. Then there exists a unique stationary distribution $\{\pi_j\}$. Moreover, $\exists A \geq 0$ and $0 \leq \rho < 1$ s.t.*

$$\left| P_{ij}^{(n)} - \pi_j \right| \leq A \rho^n$$

Theorem 27. *All states in S (finite states) are positive persistent.*

9.6 Monte Carlo Methods

Consider X: random variable. Compute $\mathbb{E}[g(X)]$.

$$\mathbb{E}[g(X)] = \int g(x) f(x) dx$$

where $f(x)$ is the pdf of X.
Idea: Let $X_1, ..., X_n \overset{\text{iid}}{\sim} X$. Then by SLLN,

$$\frac{1}{n} \sum_{k=1}^{n} g(X_k) \xrightarrow{as} \mathbb{E}[g(X)]$$

assuming that the assumptions for SLLN holds. Question: How to simulate X? If X is a simple random variable, then we may be able to simulate X by the exact method. We can use the **MCMC** method (Markov Chain Monte Carlo).

Let $\pi(x)$ be the distribution of X or $\pi(x) = \dfrac{\tilde{\pi}(x)}{\int \tilde{\pi}(x)dx}$ where the denominator is called the *normalizing constant*, which may be unknown.

We want to construct a Markov chain suck that $\pi(x)$ is its stationary distribution. Recall that $\{\pi(x)\}$ is a stationary distribution if

$$\sum_{x \in S} \pi(x) P_{xy} = \pi(y) \quad \forall y \in S \tag{20}$$

where P_{xy} is the transition probability. We want to construct P_{xy} satisfying (20). But this is too hard since (20) is a **global balance equation**.

A Markov chain with transition probability P_{xy} is said to satisfy the **detailed balance equation** if

$$\pi(x) P_{xy} = \pi(y) P_{yx} \quad \forall x, y \in S$$

<u>Claim</u>: detailed balance equation $\implies$ global balance equation.

►*Proof.* Consider the following:

$$\begin{aligned}\sum_{x \in S} \pi(x) P_{xy} &= \sum_{x \in S} \pi(y) P_{yx} && \text{(detailed balance equation)} \\ &= \pi(y) \sum_{x \in S} P_{yx} = \pi(y).\end{aligned}$$

■

Metropolis-Hastings algorithm

Let $X_n = x$ be given. We perform the following two steps repeatedly.

1. Generate $Y \sim \{P_{xy}\}$. Let y be the generated state. That is,

$$\Pr\{Y = y \mid X_n = x\} = P_{xy}$$

2. In general, the $\{P_{xy}\}$ chosen in step one does not satisfy the detailed balance equation. Setting $X_{n+1} = y$ with probability

$$\alpha(x, y) = \min\left\{\frac{\pi(y) P_{yx}}{\pi(x) P_{xy}}, 1\right\}$$

otherwise, setting $X_{n+1} = x$. (One way to simulate step 2 is to sample $u \sim \mathcal{U}(0, 1)$. If $u < \alpha(x, y)$, then $X_{n+1} = y$; else $X_{n+1} = x$.)

The ratio

$$\begin{aligned} r(x, y) &= \frac{\pi(y) P_{yx}}{\pi(x) P_{xy}} \ : \ \text{acceptance ratio} \\ p_r(x) &= \left[1 - \sum_{y \in S, x \neq y} P_{xy} \alpha(x, y)\right] \ : \ \text{rejection ratio}\end{aligned}$$

Hence the new transition probability construct in step 2 is

$$\tilde{P_{xy}} = \alpha(x, y) P_{xy} + p_r(x) \delta_x(y)$$

where $\delta_x(y) = 1$ is $x = y$ and 0 otherwise. <u>Claim</u>: $\{\tilde{P}_{cy}\}$ satisfies the detailed balance equation.

►*Proof.* If $x \neq y$, then

$$\begin{aligned}\tilde{P}_{xy} &= \min\left\{\frac{\pi(y) P_{yx}}{\pi(x) P_{xy}}, 1\right\} P_{xy} \\ &= \begin{cases} P_{yx} \frac{\pi(y)}{\pi(x)} & \text{if } \pi(y) P_{yx} < \pi(x) P_{xy} \\ P_{xy} & \text{if } \pi(y) P_{yx} \geq \pi(x) P_{xy} \end{cases}\end{aligned}$$

If $x = y$, then

$$\tilde{P}_{xx} = 1 - \sum_{x \neq y} \tilde{P}_{xy}.$$

Observe that if $\{P_{xy}\}$ satisfies the detailed balance equation, then $\tilde{P}_{xy} = P_{xy}\ \forall x, y$. However, in general, $\{P_{xy}\}$ does not satisfy the detailed balance equation. If $x = y$, then the detailed balance equation always holds. Thus, we only check for the case where $x \neq y$.

$$\begin{aligned}
\pi(x)\tilde{P}_{xy} &= \begin{cases} \pi(y)P_{yx} & \text{if } \pi(y)P_{yx} < \pi(x)P_{xy} \\ \pi(x)P_{xy} & \text{if } \pi(y)P_{yx} \geq \pi(x)P_{xy} \end{cases} \\
\pi(y)\tilde{P}_{yx} &= \begin{cases} \pi(x)P_{xy} & \text{if } \pi(x)P_{xy} \geq \pi(y)P_{yx} \\ \pi(y)P_{yx} & \text{if } \pi(x)P_{xy} < \pi(y)P_{yx} \end{cases} \\
&\implies \pi(x)\tilde{P}_{xy} = \pi(y)\tilde{P}_{yx}
\end{aligned}$$

■