Foundations of Data Science

Limit Theorems

尹一通、刘明谋 Nanjing University, 2024 Fall

Limit Theorems

Let $X_1, X_2, \ldots$ be *i.i.d.* random variables with $\mu = \mathbb{E}[X_1]$ and $\mathbf{Var}[X_1] = \sigma^2$.

And let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the <u>sample mean</u>.

- <u>Law of large numbers</u> (**LLN**): sample mean → expectation

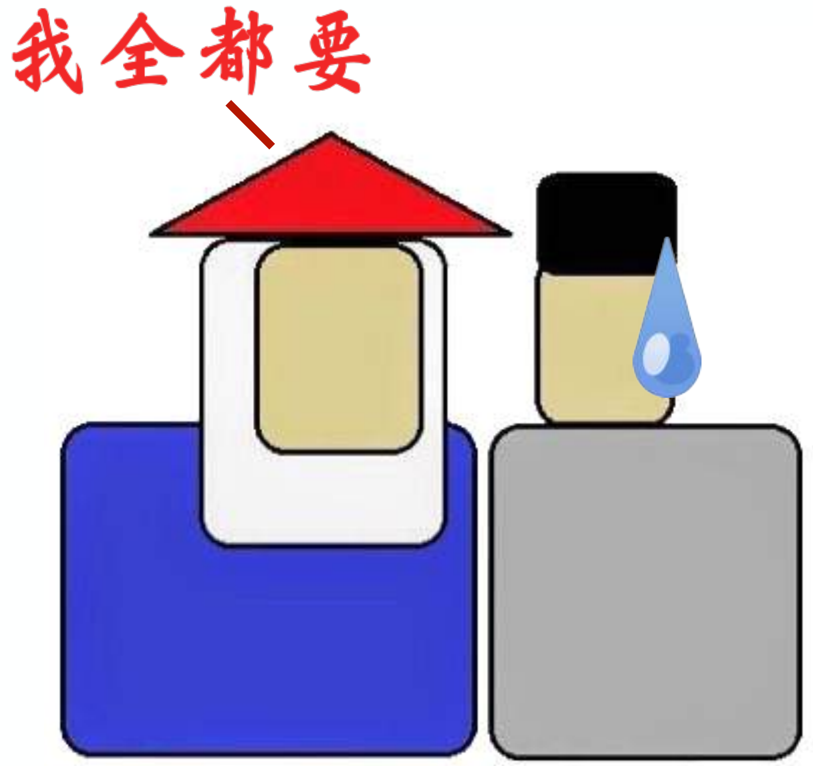
$$\overline{X}_n \longrightarrow \mu \quad \text{as } n \to \infty$$

- <u>Central limit theorem</u> (**CLT**): standardized sample mean → standard normal

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \longrightarrow N(0,1) \quad \text{as } n \to \infty$$

Convergence

- A real sequence $\{a_n\}$ converges to $a \in \mathbb{R}$, denoted $\lim_{n\to\infty} a_n = a$ or $a_n \to a$, if for all $\epsilon > 0$, there is N such that $|a_n - a| < \epsilon$ for all $n > N$
- A sequence $f_1, f_2, \ldots : \Omega \to \mathbb{R}$ is said to <u>**converge pointwise**</u> to $f : \Omega \to \mathbb{R}$, if and only if $\lim_{n\to\infty} f_n(x) = f(x)$ for all $x \in \Omega$
- For random variables $X_1, X_2, \ldots$ and X on probability space $(\Omega, \Sigma, \Pr)$:
  - random variables $X_1, X_2, \ldots : \Omega \to \mathbb{R}$ and $X : \Omega \to \mathbb{R}$ are functions
  - CDFs $F_{X_1}, F_{X_2}, \ldots : \mathbb{R} \to [0,1]$ and $F_X : \mathbb{R} \to [0,1]$ are functions
- Should $X_n \to X$ be: $X_n \to X$ pointwise or $F_{X_n} \to F_X$ pointwise?



Convergence of Random Variables

0.  $\cdots \to U_{[0,1]}$

Modes of Convergence

- Let $X, X_1, X_2, \ldots : \Omega \to \mathbb{R}$ be random variables on prob. space $(\Omega, \Sigma, \Pr)$.

- $\{X_n\}$ converges in distribution (依分布收敛) to X, denoted $X_n \xrightarrow{D} X$, if

$$F_{X_n}(x) = \Pr(X_n \le x) \to F_X(x) = \Pr(X \le x) \quad \text{as} \quad n \to \infty$$

  for all $x \in \mathbb{R}$ at which $F_X(x)$ is continuous

- $\{X_n\}$ converges in probability (依概率收敛) to X, denoted $X_n \xrightarrow{P} X$, if

$$\Pr(|X_n - X| > \epsilon) = 0 \quad \text{as} \quad n \to \infty \quad \text{for all } \epsilon > 0$$

- $\{X_n\}$ converges almost surely to X, denoted $X_n \xrightarrow{a.s.} X$, if $\exists A \in \Sigma$ such that

$$\lim_{n\to\infty} X_n(\omega) = X(\omega) \quad \text{for all } \omega \in A, \quad \text{and} \quad \Pr(A) = 1$$

Modes of Convergence

- Let $X_1, X_2, \ldots$ and X be random variables on probability space $(\Omega, \Sigma, \Pr)$.
- $X_n \xrightarrow{D} X$ (convergence <u>in distribution</u> / <u>in law</u> / weak convergence of measure) if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

$F_{X_n} \to F_X$ pointwise on continuous set

for all $x \in \mathbb{R}$ at which $F_X(x)$ is continuous

- $X_n \xrightarrow{P} X$ (convergence <u>in probability</u> / <u>in measure</u>) if

$$\lim_{n\to\infty} \Pr(|X_n - X| > \epsilon) = 0 \quad \text{for all } \epsilon > 0$$

$X_n \to X$ in measure

- $X_n \xrightarrow{a.s.} X$ (convergence <u>almost surely</u> / <u>almost everywhere</u> / <u>w.p. 1</u>) if

$$\Pr\left(\lim_{n\to\infty} X_n = X\right) = 1$$

$X_n \to X$ pointwise on a set of measure 1

Convergence in Distribution

- Let $X_1, X_2, \ldots$ and X be random variables on probability space $(\Omega, \Sigma, \Pr)$.
- $X_n \xrightarrow{D} X$ (convergence in distribution / in law / weak convergence of measure) if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

$F_{X_n} \to F_X$ pointwise on continuous set

for all $x \in \mathbb{R}$ at which $F_X(x)$ is continuous

- The restriction on continuity set is necessary, consider:

uniform X_n on $(0,1/n)$, which satisfies $X_n \xrightarrow{D} X$, where $\Pr(X = 0) = 1$

- $X_n \xrightarrow{D} X$ and $F_X = F_Y \implies X_n \xrightarrow{D} Y$ (convergence in distribution depends only on distribution)
- $X_n \xrightarrow{D} X$ is a weak convergence of measures

Convergence in Probability

- Let $X_1, X_2, \ldots$ and X be random variables on probability space $(\Omega, \Sigma, \Pr)$.
- $X_n \xrightarrow{P} X$ (convergence <u>in probability</u>) if

$$\lim_{n \to \infty} \Pr(|X_n - X| > \epsilon) = 0 \quad \text{for all } \epsilon > 0$$

$X_n \to X$
in measure

- Functions $X_n : \Omega \to \mathbb{R}$ converges to $X : \Omega \to \mathbb{R}$ <u>in measure</u> $\Pr$
- $X_n \xrightarrow{P} X \implies X_n \xrightarrow{D} X$
  - **Counterexample for converse**: X is uniform on $[0,1]$ and $X_n = 1 - X$
- If $X_n \xrightarrow{D} c$, where $c \in \mathbb{R}$ is a constant, then $X_n \xrightarrow{P} c$
  - **Proof**: $\Pr(|X_n - c| > \epsilon) = \Pr(X_n < c - \epsilon) + \Pr(X_n > \epsilon + c) \to 0$ if $X_n \xrightarrow{D} c$

Almost Sure Convergence

- Let $X_1, X_2, \ldots$ and X be random variables on probability space $(\Omega, \Sigma, \Pr)$.
- $X_n \xrightarrow{a.s.} X$ (convergence almost surely / almost everywhere / **w.p. 1**) if

$$\Pr\left(\lim_{n\to\infty} X_n = X\right) = 1$$

$X_n \to X$ pointwise on a set of measure 1

- $X_n : \Omega \to \mathbb{R}$ converges to $X : \Omega \to \mathbb{R}$ **almost everywhere** except a null set
- $X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{P} X$
  - **Counterexample for converse**: $\{X_n\}$ are independent $\text{Bernoulli}(1/n)$. We have $X_n \xrightarrow{P} 0$, but we do not have $X_n = 0$ almost everywhere as $n \to \infty$.

*

Strength of Convergence

- $(X_n \xrightarrow{a.s.} X) \implies (X_n \xrightarrow{P} X) \implies (X_n \xrightarrow{D} X)$

Proof* $(X_n \xrightarrow{a.s.} c \implies X_n \xrightarrow{P} c)$: Let event $C \triangleq \{X_n \to c\}$, then $\Pr(C) = 1$.

For any $\epsilon > 0$, let event $A_k \triangleq \{\forall n \geq k, |X_n - c| < \epsilon\}$.

Assume $X_n \xrightarrow{a.s.} X$, then $\exists k$, such that $\forall n \geq k, C \subseteq A_n$. Therefore $C \subseteq \bigcup_{k=1}^{\infty} A_k$.

Since $A_1 \subseteq A_2 \subseteq \ldots$, and $A_k \subseteq \{|X_n - c| < \epsilon\}$,

$$\lim_{n\to\infty} \Pr(|X_n - c| < \epsilon) \geq \lim_{n\to\infty} \Pr(A_k) = \Pr(\cup_{k=1}^{\infty} A_k) \geq \Pr(C) = 1.$$

Therefore,

$$\lim_{n\to\infty} \Pr(|X_n - c| \geq \epsilon) = 0, \text{ i.e. } X_n \xrightarrow{P} X.$$

Other Convergence Modes*

- $X_n \xrightarrow{1} X$ (convergence <u>in mean</u>) if

$$\lim_{n\to\infty} \mathbb{E}\left[|X_n - X|\right] = 0$$

- $X_n \xrightarrow{r} X$ (convergence <u>in rth mean</u> / <u>in the L^r-norm</u>) if

$$\lim_{n\to\infty} \mathbb{E}\left[|X_n - X|^r\right] = 0$$

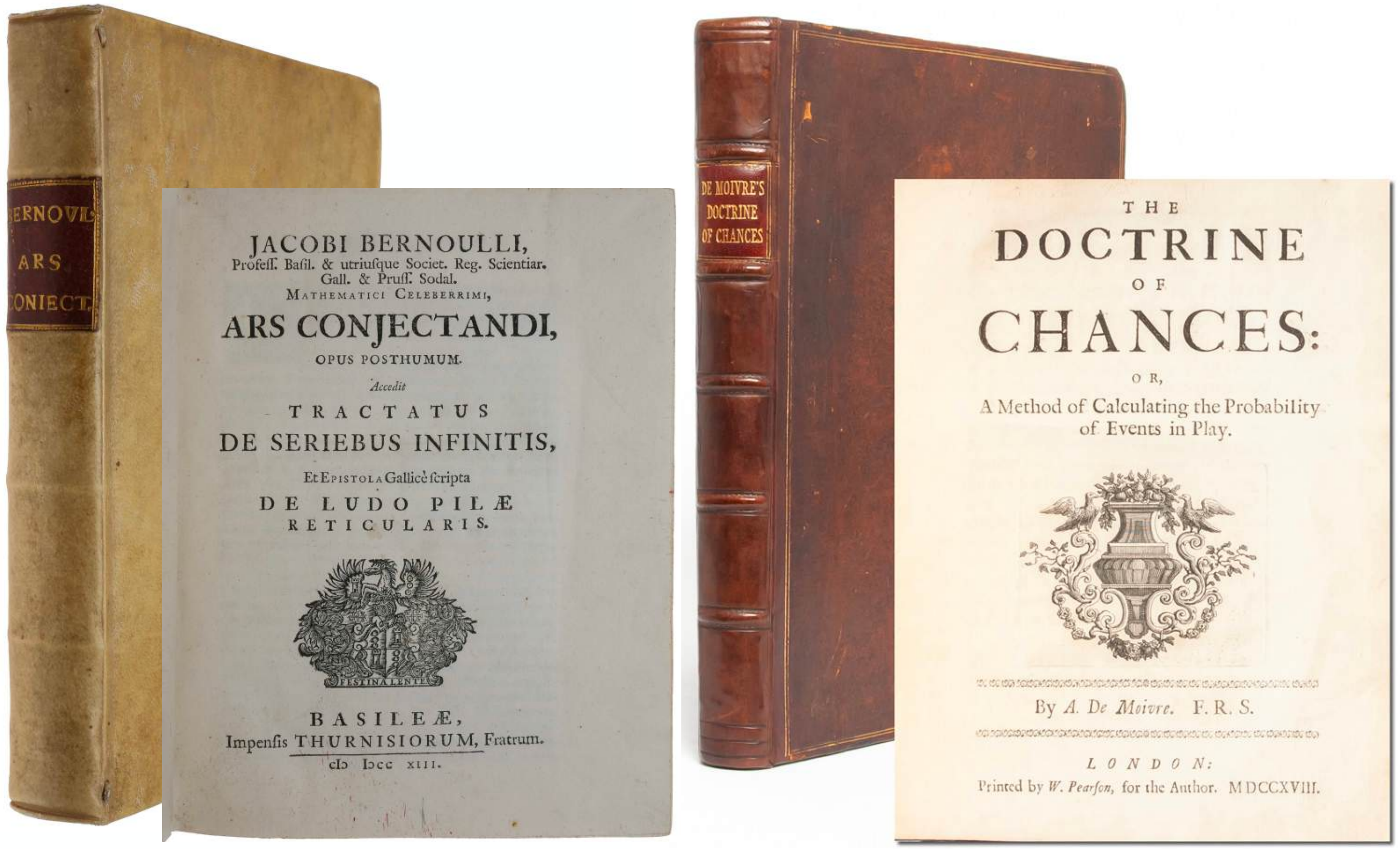
$$(X_n \xrightarrow{a.s.} X) \implies (X_n \xrightarrow{P} X) \implies (X_n \xrightarrow{D} X)$$

$$\Uparrow$$

$$(X_n \xrightarrow{s} X) \implies (X_n \xrightarrow{r} X) \implies (X_n \xrightarrow{1} X)$$

(for $s \geq r \geq 1$)

LLN and CLT



JACOBI BERNOULLI,
Profess. Basil. & utriusque Societ. Reg. Scientiar. Gall. & Pruss. Sodal.
MATHEMATICI CELEBERRIMI,
ARS CONJECTANDI,
OPUS POSTHUMUM.
Accedit
TRACTATUS
DE SERIEBUS INFINITIS,
Et EPISTOLA Gallicè scripta
DE LUDO PILÆ
RETICULARIS.
BASILEÆ,
Impensis THURNISIORUM, Fratrum.
ARS
DE MOIVRE'S
DOCTRINE
OF CHANCES
THE
DOCTRINE
OF
CHANCES:
OR,
A Method of Calculating the Probability of Events in Play.
By A. De Moivre. F. R. S.
LONDON:
Printed by W. Pearson, for the Author. MDCCXVIII.

Bernoulli's Law of Large Number

In *Ars Conjectandi* (1713)

- Let $X_1, X_2, \ldots$ be *i.i.d.* Bernoulli trials with $\mathbb{E}[X_1] = p \in [0,1]$. Then

$$\Pr\left(\left| \frac{X_1 + X_2 + \cdots + X_n}{n} - p \right| > \epsilon \right) \to 0 \quad \text{as } n \to \infty \quad \text{for all } \epsilon > 0$$

i.e. $\overline{X}_n \xrightarrow{P} p$, where $\overline{X}_n$ is the sample mean $\overline{X}_n = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$

Proof: By Chebyshev's inequality, $\Pr(|\overline{X}_n - p| > \epsilon) \le \dfrac{p(1-p)}{n\epsilon^2} \to 0$ as $n \to \infty$

(This is of course not the original proof of Bernoulli.)

Law of Large Numbers (LLN)

Let $X_1, X_2, \ldots$ be *i.i.d.* random variables with finite mean $\mathbb{E}[X_1] = \mu$.

And let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the **sample mean.**

- **<u>Weak law</u>** (Khinchin's law) **<u>of large number</u>**:

$$\overline{X}_n \xrightarrow{P} \mu \quad \text{as} \quad n \to \infty$$

- **<u>Strong law</u>** (Kolmogorov's law) **<u>of large number</u>**:

$$\overline{X}_n \xrightarrow{a.s.} \mu \quad \text{as} \quad n \to \infty$$

(*The deviation* $|\overline{X}_n - \mu|$ *is always small for all sufficiently large* n)

Weak LLN Assuming Bounded Variance

- Let $X_1, X_2, \ldots$ be independent random variables with finite mean $\mathbb{E}[X_i] = \mu$ and finitely bounded variance $\mathbf{Var}[X_i] \leq \sigma^2$.

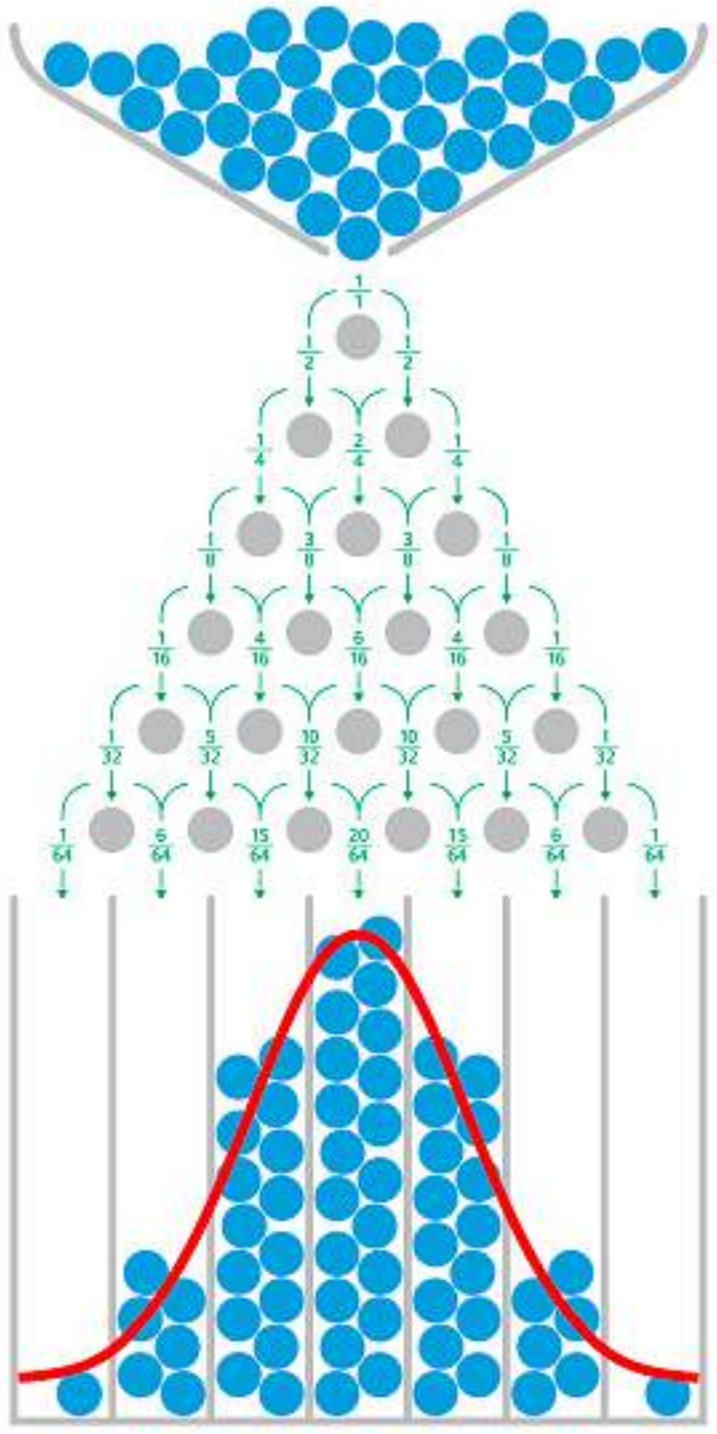
  Then the sample mean $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ has

$$\overline{X}_n \xrightarrow{P} \mu \text{ as } n \to \infty$$

Proof: By **Chebysev's inequality**, $\Pr(|\overline{X}_n - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \to 0$ as $n \to \infty$

De Moivre—Laplace Theorem

(棣莫弗–拉普拉斯定理)



- Let $p \in (0,1)$ and $X_n \sim B(n,p)$. Then its standardization

$$\frac{X_n - np}{\sqrt{np(1-p)}} \xrightarrow{D} N(0,1) \quad \text{as} \quad n \to \infty$$

- For any $p \in (0,1)$ and any $\epsilon > 0$, there is an n_0 such that for all $n > n_0$ and all k,

$$\binom{n}{k} p^k (1-p)^{n-k} \in (1 \pm \epsilon) \frac{1}{\sqrt{2\pi np(1-p)}} \mathrm{e}^{-\frac{(k-np)^2}{2np(1-p)}}$$

By Stirling's formula $n! \simeq n^n e^{-n}\sqrt{2\pi n}$ and Maclaurin series $\ln(1+x) \simeq x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots$

Central Limit Theorem (CLT)

- Let $X_1, X_2, \ldots$ be *i.i.d.* random variables with $\mathbb{E}[X_1] = \mu$ and $\mathbf{Var}[X_1] = \sigma^2$.

  And let $\overline{X}_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ be the **sample mean.**

- **Classical** (Lindeberg–Lévy) **central limit theorem**:

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} N(0,1) \quad \text{as } n \to \infty$$

Central Limit Theorem (CLT)

- Let $X_1, X_2, \ldots$ be *i.i.d.* random variables with $\mathbb{E}[X_1] = \mu$ and $\mathbf{Var}[X_1] = \sigma^2$.

$$Z_n = \frac{\sum_i (X_i - \mu)}{\sigma\sqrt{n}} \xrightarrow{D} N(0,1) \quad \text{as } n \to \infty$$

Proof: $M_X(t) = \sum_{k \geq 0} \frac{t^k \mathbb{E}[X^k]}{k!} = \frac{t^0 \mathbb{E}[X^0]}{0!} + \frac{t^1 \mathbb{E}[X^1]}{1!} + \frac{t^2 \mathbb{E}[X^2]}{2!} + o\left(\frac{t^2 \mathbb{E}[X^2]}{2!}\right)$

$$\implies M_{X_1 - \mu}(t) = 1 + t^2\sigma^2/2 + o(t^2)$$

$$M_{Z_n}(t) = \mathbb{E}[e^{tZ_n}] = \mathbb{E}\left[e^{\frac{t}{\sigma\sqrt{n}}\sum_i (X_i - \mu)}\right] = \prod_i \mathbb{E}\left[e^{\frac{t}{\sigma\sqrt{n}}(X_i - \mu)}\right] = \left(M_{X_1 - \mu}\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n$$

Central Limit Theorem (CLT)

- Let $X_1, X_2, \ldots$ be *i.i.d.* random variables with $\mathbb{E}[X_1] = \mu$ and $\mathbf{Var}[X_1] = \sigma^2$.

$$Z_n = \frac{\sum_i (X_i - \mu)}{\sigma\sqrt{n}} \xrightarrow{D} N(0,1) \quad \text{as } n \to \infty$$

Proof:

$$M_{X_1-\mu}(t) = 1 + t^2\sigma^2/2 + o(t^2)$$

$$M_{Z_n}(t) = \left(M_{X_1-\mu}\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n$$

$$M_{Z_n}(t) = \left(1 + \left(\frac{t}{\sigma\sqrt{n}}\right)^2 \sigma^2/2 + o\left(\left(\frac{t}{\sigma\sqrt{n}}\right)^2\right)\right)^n$$

$$= \left(1 + t^2/(2n) + o(t^2/n)\right)^n$$

$$\lim_{n\to\infty} M_{Z_n}(t) = \lim_{n\to\infty} \left(1 + t^2/(2n) + o(t^2/n)\right)^n = e^{t^2/2}$$

MGF of Normal Distribution

- The moment generating function of standard normal $X \sim N(0,1)$ is

$$M_X(t) = \mathbb{E}[\mathrm{e}^{tX}] = \mathrm{e}^{t^2/2}$$

Proof: $M_X(t) = \mathbb{E}[\mathrm{e}^{tX}] = \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{-\infty}^{\infty} \mathrm{e}^{tx}\mathrm{e}^{-x^2/2}\, \mathrm{d}x = \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{-\infty}^{\infty} \mathrm{e}^{-x^2/2+tx}\, \mathrm{d}x$

$= \dfrac{1}{\sqrt{2\pi}}\mathrm{e}^{t^2/2} \displaystyle\int_{-\infty}^{\infty} \mathrm{e}^{(x-t)^2/2}\, \mathrm{d}x = \mathrm{e}^{t^2/2}$ (because $\dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{-\infty}^{\infty} \mathrm{e}^{(x-t)^2/2}\, \mathrm{d}x = 1$)

Central Limit Theorem (CLT)

- Let $X_1, X_2, \ldots$ be *i.i.d.* random variables with $\mathbb{E}[X_1] = \mu$ and $\mathbf{Var}[X_1] = \sigma^2$.

$$Z_n = \frac{\sum_i (X_i - \mu)}{\sigma\sqrt{n}} \xrightarrow{D} N(0,1) \quad \text{as } n \to \infty$$

$$M_X(t) = \mathbb{E}[\mathrm{e}^{tX}] = \mathrm{e}^{t^2/2}$$

Proof:

$$M_{X_1-\mu}(t) = 1 + t^2\sigma^2/2 + o(t^2)$$

$$M_{Z_n}(t) = \left(M_{X_1-\mu}\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n$$

$$M_{Z_n}(t) = \left[1 + \left(\frac{t}{\sigma\sqrt{n}}\right)^2 \sigma^2/2 + o\left[\left(\frac{t}{\sigma\sqrt{n}}\right)^2\right]\right]^n$$

$$= \left(1 + t^2/(2n) + o(t^2/n)\right)^n$$

$$\lim_{n\to\infty} M_{Z_n}(t) = \lim_{n\to\infty} \left(1 + t^2/(2n) + o(t^2/n)\right)^n = e^{t^2/2}$$

$M_X(t) = \mathbb{E}[\mathrm{e}^{tX}] = \mathrm{e}^{t^2/2}$, for $X \sim N(0,1)$

CLT for Non-Identically Distributed RVs*

- Let $X_1, X_2, \ldots$ be independent random variables with $\mathbb{E}[X_i] = \mu_i$ and $S_n^2 = \sum_{i=1}^{n} \mathbf{Var}[X_i]$.

  Assume:

  - $\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} \mathrm{E}\left[\left|X_i - \mu_i\right|^{2+\delta}\right] = 0$ for some $\delta > 0$ (Lyapunov's condition)

  -or, $\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{i=1}^{n} \mathrm{E}\left[(X_i - \mu_i)^2 \cdot \mathbf{1}_{\left\{\left|X_i - \mu_i\right| > \varepsilon s_n\right\}}\right] = 0$ for every $\epsilon > 0$ (Lindeberg's condition)

  Then

$$\frac{\sum_{i=1}^{n} (X_i - \mu_i)}{S_n} \xrightarrow{D} N(0,1)$$

Convergence Rate of CLT

(Berry–Esseen theorem)

- Berry–Esseen theorem: Let $X_1, X_2, \ldots$ be *i.i.d.* random variables with $\mathbb{E}[X_1] = \mu$, $\mathbf{Var}[X_1] = \sigma^2$, and $\rho = \mathbb{E}[\,|X_1 - \mu|^3\,]$. And let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

  There is an absolute constant C, such that for any z

$$\left| \Pr\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leq z \right) - \Phi(z) \right| \leq \frac{C\rho}{\sigma^3\sqrt{n}}$$

  where Φ stands for the CDF for standard normal distribution $N(0,1)$