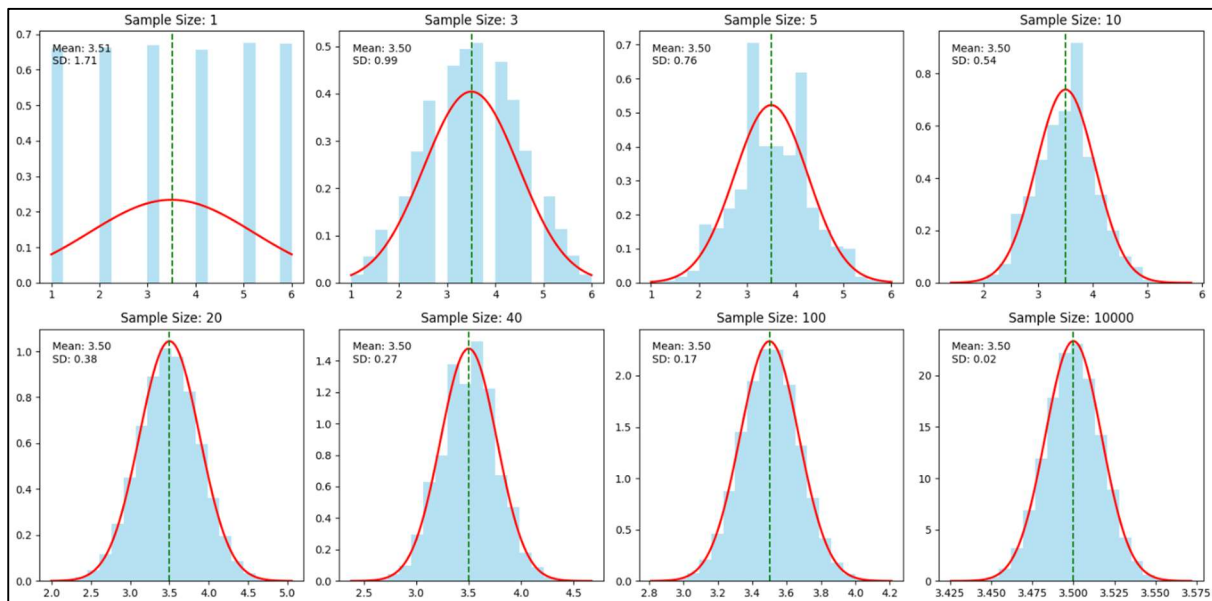


การพิสูจน์: ทฤษฎีแนวโน้มนำสู่ส่วนกลาง

ทฤษฎีแนวโน้มนำสู่ส่วนกลาง (The Central Limit Theorem; CLT) กล่าวว่า

สำหรับประชากรใด ๆ แล้ว หากเราเก็บตัวอย่างในจำนวนที่มากพอ การกระจายของค่าเฉลี่ยตัวอย่างดังกล่าวจะมีแนวโน้มใกล้เคียงกับการกระจายแบบปกติ (Normal distribution) เสมอ โดยไม่คำนึงถึงรูปร่างการกระจายที่แท้จริงของประชากรนั้น

สรุปโดยย่อ ถ้าเราสุ่มตัวอย่างซ้ำเรื่อย ๆ และบันทึกค่าเฉลี่ยที่ได้จากการสุ่มตัวอย่างแต่ละครั้ง แล้วนำค่าเหล่านั้นมาสร้างแผนภูมิฮิสโตแกรม (Histogram) จะได้ว่าข้อมูลมีการแจกแจงแบบปกตินั่นเอง



ตัวอย่าง CLT โดยให้ X แทนค่าเฉลี่ยของหมายเลขที่เกิดจากการทอยลูกเต๋า 6 หน้าเพียงตรง

ที่มา: [Statistics-and-ML/Central_Limit_Theorem.ipynb at main · SKY-TKP/Statistics-and-ML](#)

ประเด็นสำคัญ

- ทฤษฎีบทนี้ระบุถึงความสำคัญที่ว่า การกระจายของค่าเฉลี่ยตัวอย่างจะใกล้เคียงกับการแจกแจงปกติเมื่อตัวอย่างมีขนาดใหญ่โดยไม่คำนึงถึงการแจกแจงของประชากรที่แท้จริง
- การสุ่มตัวอย่างต้องเป็นไปอย่างสุ่ม (Random Sampling)
- เราจะเรียก ค่าเฉลี่ยจากการสุ่มตัวอย่างเหล่านั้นว่า “ค่าเฉลี่ยกลุ่มตัวอย่าง (Sample Mean; \bar{x})” และเรียก ส่วนเบี่ยงเบนมาตรฐานของค่าเฉลี่ยเหล่านั้นว่า “ส่วนเบี่ยงเบนมาตรฐานของค่าเฉลี่ยกลุ่มตัวอย่าง” (Sample Standard Deviation; $s_{\bar{x}}$)

ทบทวนคณิตศาสตร์ที่จำเป็น

[1] ตัวแปรสุ่ม (Random Variable) และคุณสมบัติพื้นฐาน

ตัวแปรสุ่ม X คือ ฟังก์ชันจากปริภูมิตัวอย่างของการทดลองสุ่ม (S) ไปยัง เซตของจำนวนจริง (\mathbb{R})

$$X: S \rightarrow \mathbb{R}$$

โดยที่

- $E[X]$ = ค่าเฉลี่ยของตัวแปรสุ่ม $= \frac{1}{N} \sum_{i=1}^N x_i$
- $Var[X]$ = ความแปรปรวนของตัวแปรสุ่ม $= \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$
- สมบัติเพิ่มเติม
 1. $E[aX + b] = aE[X] + b$
 2. $E[a_1X_1 + a_2X_2 + \dots + a_nX_n] = a_1E[X_1] + a_2E[X_2] + \dots + a_nE[X_n]$
 3. $Var[aX + b] = a^2Var[X]$
 4. $Var[a_1X_1 + a_2X_2 + \dots + a_nX_n]$
 $= a_1^2Var[X_1] + a_2^2Var[X_2] + \dots + a_n^2Var[X_n] + 2 \sum_{j \neq i} \sum_i a_i a_j Cov(X_i, X_j)$

[2] ฟังก์ชันความหนาแน่นความน่าจะเป็นของการแจกแจงปกติ (Probability Density Func.)

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

เมื่อ $\mu = E[X]$ และ $\sigma = \sqrt{Var[X]}$

เขียนแทนด้วย $X \sim Normal(\mu, \sigma^2)$

หมายเหตุ: เมื่อ $\mu = 0$ และ $\sigma = 1$ เราจะเรียกการแจกแจงดังกล่าวเป็น การแจกแจงปกติมาตรฐาน (Standard Normal Distribution)

[3] ฟังก์ชันก่อกำเนิดโมเมนต์ (Moment Generating Function; MGF)

$$M_X(t) = E[e^{tx}] = \int_{-\infty}^{\infty} e^{tx} f(x) dx$$

โดยที่

- เมื่อ Z เป็นตัวแปรสุ่มที่มีการแจกแจงปกติมาตรฐาน จะได้ว่า

$$M_Z(t) = E[e^{tx}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{z^2}{2} + tz} dz = e^{\frac{t^2}{2}}$$
- $M_X^{(n)}(t=0) = E[X^n] = \int_{-\infty}^{\infty} x^n f(x) dx$
- ถ้า X และ Y เป็นตัวแปรสุ่มที่เป็นอิสระต่อกันแล้ว จะได้ว่า $M_{X+Y}(t) = M_X(t)M_Y(t)$
- ถ้า X และ Y มีฟังก์ชันก่อกำเนิดโมเมนต์เดียวกันแล้ว X และ Y มีการแจกแจงเดียวกัน

$$M_X(t) = M_Y(t); \forall t \in \mathbb{R} \implies X, Y \text{ are same distribution.}$$

[4] เพิ่มเติม

- นิยามของเลขออยเลอร์ (Euler's number; e)

$$\lim_{n \rightarrow \infty} \left(1 + \frac{k}{n}\right)^n = e^k; \quad e \approx 2.71828$$

- อนุกรมเทย์เลอร์ของฟังก์ชันเลขชี้กำลัง (Taylor's series of exponential function)

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

- อนุกรมแมคคลอรินของฟังก์ชันลอการิทึม (Maclaurin's series of logarithmic function)

$$\ln(1+x) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} x^n$$

ขั้นตอนการพิสูจน์

ข้อกำหนด: ¹

- [1] พิจารณาจากตัวแปรสุ่ม X ที่มีการกระจายแบบหนึ่ง ๆ (ที่เราไม่รู้) โดยมี $\mu_X = \mu = E[X]$ และ $\sigma_X^2 = \text{Var}[X]$
- [2] การเลือกตัวอย่างของเราจากค่าที่เป็นไปได้ทั้งหมดในตัวแปรสุ่ม X จะต้องมีความสมบัติต่อไปนี้
- (2.1) การเลือกตัวอย่างแบบสุ่ม (Random Sampling)
 - (2.2) การสุ่มแต่ละครั้งต้องเป็นอิสระต่อกัน (Independent)
 - (2.3) การสุ่มแต่ละครั้งต้องมาจากประชากรในการแจกแจงเดิม
- [3] จากผลของ [2] เขียนรวบรัดเป็น X_i เป็น *i. i. d.* (*independent and identically distributed*) โดย X_i คือ ตัวแปรสุ่มของค่าที่เป็นไปได้จากปริภูมิตัวอย่างในทั้งหมดในตัวแปรสุ่ม X ตัวที่ $i = 1, 2, 3, \dots, n$ จากการสุ่มมา n ครั้ง
- [4] การสุ่มตัวอย่าง ควรใช้ขนาดกลุ่มตัวอย่าง n จำนวนมาก ๆ

การพิสูจน์:

[1] กำหนดให้ $Y_n = \frac{1}{n} (X_1 + X_2 + X_3 + \dots + X_n) = \frac{1}{n} \sum_{i=1}^n X_i$

ดังนั้น

เราบอกได้ว่า Y_n คือ ตัวแปรสุ่มค่าเฉลี่ยกลุ่มตัวอย่างจากการสุ่มในประชากร X

จะได้ว่า

$$\begin{aligned} E[Y_n] &= E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} (E[X_1] + E[X_2] + \dots + E[X_n]) \\ &= \frac{1}{n} (\mu + \mu + \dots + \mu) \\ &= \frac{1}{n} (n\mu) = \mu \dots (1.1) \end{aligned}$$

$$\begin{aligned} \text{Var}[Y_n] &= \text{Var}\left[\frac{1}{n} \sum_{i=1}^n X_i\right] \\ &= \frac{1}{n^2} \left[\text{Var}[X_1] + \text{Var}[X_2] + \dots + \text{Var}[X_n] + 2 \sum_i \sum_{i \neq j} \text{Cov}(X_i, X_j) \right] \end{aligned}$$

จาก ข้อกำหนด (2.2) ทำให้ $\text{Cov}(X_i, X_j) = 0$;

$$\text{Var}[Y_n] = \frac{1}{n^2} [\sigma^2 + \sigma^2 + \dots + \sigma^2] = \frac{1}{n^2} (n\sigma_X^2) = \sigma_X^2/n \dots (1.2)$$

¹ ด้วยข้อกำหนดที่จะนำไปสู่การพิสูจน์เหล่านี้จึงเป็นสิ่งสืบเนื่องที่ต้องตรวจสอบเสมอ
หลังทำการใช้สถิติเชิงอนุมานในหลาย ๆ ส่วน

จากข้อ [1]

- เมื่อเราสุ่มตัวอย่างจากประชากร X แล้วนำมาหาค่าเฉลี่ย เราเชื่อว่าค่าเฉลี่ยของกลุ่มตัวอย่างที่ได้มานั้นมีโอกาสเป็นค่าเฉลี่ยที่แท้จริงของประชากรดังกล่าว ในขณะที่ถ้าหาความแปรปรวนแทนจะกลายเป็น ความแปรปรวนของประชากรหารขนาดกลุ่มตัวอย่างที่เราเลือกมาสุ่ม
- $E[Y_n] = \bar{x}_n = \mu =$ ค่าเฉลี่ยกลุ่มตัวอย่างจากการสุ่มในตัวแปรสุ่ม X
- $Var[Y_n] = s_n^2 = \sigma_X^2/n =$ ความแปรปรวนตัวอย่างจากการสุ่มในตัวแปรสุ่ม X^2

[2] กำหนดให้

$$Z_n = (Y_n - \mu)/s_x$$

$$= \frac{\sqrt{n}}{\sigma_X} \left(\frac{1}{n} \sum_{i=1}^n X_i - \mu \right)$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma_X} \right)$$

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n S_i \quad \text{โดย} \quad S_i = \frac{X_i - \mu}{\sigma_X}$$

$$\text{ทั้งนี้ } E[S_i] = \frac{1}{\sigma_X} E[X_i - \mu] = 0$$

$$Var[S_i] = \left(\frac{1}{\sigma_X} \right)^2 [Var[X_i] - Var[\mu]] = 1$$

$$\text{และ } E[Z_n] = \frac{1}{\sqrt{n}} \sum_{i=1}^n E[S_i] = 0$$

$$Var[Z_n] = \left(\frac{1}{\sqrt{n}} \right)^2 \sum_{i=1}^n Var[S_i] = 1 = E[Z_n^2] - E[Z_n]^2 = E[Z_n^2]$$

² ภายหลัง เรามีอีกชื่อให้เป็น Standard Error ของตัวแปรสุ่ม X จากการสุ่มตัวอย่างขนาด n

[3] พิสูจน์ว่า

$$\begin{aligned}
M_{Z_n}(t) &= M_{\sum_{i=1}^n S_i} \left(\frac{t}{\sqrt{n}} \right) \\
&= M_{S_1+S_2+\dots+S_n} \left(\frac{t}{\sqrt{n}} \right) \\
&= M_{S_1} \left(\frac{t}{\sqrt{n}} \right) M_{S_2} \left(\frac{t}{\sqrt{n}} \right) \dots M_{S_n} \left(\frac{t}{\sqrt{n}} \right) \quad \because S_i \text{ เป็นอิสระต่อกัน} \\
&= \left[M_{S_n} \left(\frac{t}{\sqrt{n}} \right) \right]^n
\end{aligned}$$

สนใจพจน์ $M_{S_n}(t)$

$$\begin{aligned}
M_{S_n}(t) &= M_{(X_i - \mu)}(t) \\
&= E[e^{t(X_i - \mu)}] \\
&= E \left[1 + t(X_i - \mu) + \frac{(t)^2(X_i - \mu)^2}{2!} + \dots + \frac{(t)^n(X_i - \mu)^n}{n!} \right] \\
&= 1 + tE[X_i - \mu] \\
&\quad + \frac{t^2}{2!} [E[(X_i - \mu)^2] - E[X_i - \mu] + E[X_i - \mu]] \quad \because E[X_i - \mu] = 0 \\
&\quad + \frac{t^3 E[(X_i - \mu)^3]}{3!} \dots + \frac{t^n E[(X_i - \mu)^n]}{n!} \\
&= 1 + \frac{t^2}{2!} + \frac{t^3 E[(X_i - \mu)^3]}{3!} \dots + \frac{t^n E[(X_i - \mu)^n]}{n!}
\end{aligned}$$

เพราะฉะนั้น

$$M_{Z_n}(t) = \left[1 + \frac{(t/\sqrt{n})^2}{2!} + \frac{(t/\sqrt{n})^3 E[(X_i - \mu)^3]}{3!} \dots + \frac{(t/\sqrt{n})^n E[(X_i - \mu)^n]}{n!} \right]^n$$

ทำการ Take \ln ;

$$\ln[M_{Z_n}(t)] = n \cdot \ln \left[1 + \frac{(t/\sqrt{n})^2}{2!} + \frac{(t/\sqrt{n})^3 E[(X_i - \mu)^3]}{3!} \dots + \frac{(t/\sqrt{n})^n E[(X_i - \mu)^n]}{n!} \right]$$

เนื่องจากเทอมฝั่งขวาของสมการสามารถเขียนในรูปอนุกรมแมคลอริน ได้เป็น

$$\ln(1 + x) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} x^n$$

จึงได้เป็น

$$\ln[M_{Z_n}(t)] = n \cdot \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} \left(\frac{(t/\sqrt{n})^2}{2!} + \frac{(t/\sqrt{n})^3 E[(X_i - \mu)^3]}{3!} \dots + \frac{(t/\sqrt{n})^n E[(X_i - \mu)^n]}{n!} \right)^k$$

แยกอนุกรมออกเป็น 2 ส่วน คือ $k = 1$ และ 2 เป็นต้นไป

$$\begin{aligned} \ln[M_{Z_n}(t)] = n \cdot & \left[\left(\frac{(t/\sqrt{n})^2}{2!} + \frac{(t/\sqrt{n})^3 E[(X_i - \mu)^3]}{3!} + \dots + \frac{(t/\sqrt{n})^n E[(X_i - \mu)^n]}{n!} \right) \right. \\ & \left. + \sum_{k=2}^{\infty} \frac{(-1)^{k+1}}{k} \left(\frac{(t/\sqrt{n})^2}{2!} + \frac{(t/\sqrt{n})^3 E[(X_i - \mu)^3]}{3!} + \dots + \frac{(t/\sqrt{n})^n E[(X_i - \mu)^n]}{n!} \right)^k \right] \end{aligned}$$

กระจายพจน์ n ออกสุดไปในวงเล็บใหญ่

$$\begin{aligned} \ln[M_{Z_n}(t)] = & \left(\frac{t^2}{2!} + \frac{nt^3 E[(X_i - \mu)^3]}{3! n^{\frac{3}{2}}} + \dots + \frac{nt^n E[(X_i - \mu)^n]}{n! n^{\frac{n}{2}}} \right) \\ & + \sum_{k=2}^{\infty} \frac{(-1)^{k+1}}{k} \left(\frac{t^2}{2!} + \frac{nt^3 E[(X_i - \mu)^3]}{3! n^{\frac{3}{2}}} + \dots + \frac{nt^n E[(X_i - \mu)^n]}{n! n^{\frac{n}{2}}} \right)^k \end{aligned}$$

พิจารณาที่ $n \rightarrow \infty$

$$\begin{aligned} \lim_{n \rightarrow \infty} \ln[M_{Z_n}(t)] &= \lim_{n \rightarrow \infty} \left(\frac{t^2}{2!} + \frac{nt^3 E[(X_i - \mu)^3]}{3! n^{\frac{3}{2}}} + \dots + \frac{nt^n E[(X_i - \mu)^n]}{n! n^{\frac{n}{2}}} \right) \\ &+ \lim_{n \rightarrow \infty} \sum_{k=2}^{\infty} \frac{(-1)^{k+1}}{k} \left(\frac{t^2}{2!} + \frac{nt^3 E[(X_i - \mu)^3]}{3! n^{\frac{3}{2}}} + \dots + \frac{nt^n E[(X_i - \mu)^n]}{n! n^{\frac{n}{2}}} \right)^k \\ &= \frac{t^2}{2} \dots \blacksquare \end{aligned}$$

เพราะฉะนั้น

$$M_{Z_n}(t) = e^{\frac{t^2}{2}}$$

[4]. เนื่องจาก $M_{Z_n}(t)$ มี MGF เดียวกับ MGF ของการแจกแจงปกติมาตรฐาน

เพราะฉะนั้น $Z_n \sim \text{Normal}(0, 1)$

กล่าวคือ $\frac{Y_n - \mu}{s_x}$ หรือตัวแปรสุ่มค่าเฉลี่ยกลุ่มตัวอย่างนี้มีการแจกแจงปกติมาตรฐาน

[5]. จากข้อ [4]. ดังนั้น

การแจกแจงของค่าเฉลี่ยตัวอย่างมีการแจกแจงปกติ