Central Limit Theorem (CLT): Analytical and Simulation Proofs

桑慧敏
清華大學工業工程系

2015.09.30

Central Limit Theorem (CLT)

- Q: 中央極限定理是: ______ 的極限分配是標準常態;
  The limiting distribution of ______ is standard normal.

Central Limit Theorem (CLT)

- Q: 中央極限定理是: ______ 的極限分配是標準常態; The limiting distribution of ______ is standard normal.
- A: 不論什麼分配 (錯)

Central Limit Theorem (CLT)

- Q: 中央極限定理是: ______ 的極限分配是標準常態; The limiting distribution of ______ is standard normal.
- A: 不論什麼分配 (錯)
- A: 樣本平均數, $\overline{X}$ (錯)

Central Limit Theorem (CLT)

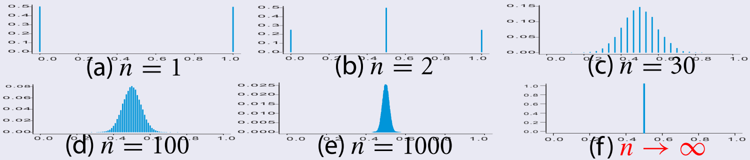
- Q: 中央極限定理是: ______ 的極限分配是標準常態; The limiting distribution of ______ is standard normal.
- A: 不論什麼分配 (錯)
- A: 樣本平均數, $\overline{X}$ (錯)
- A: 標準化後的樣本平均數: $\frac{\overline{X}_{(n)} - \mu}{\mathrm{se}(\overline{X}_{(n)})}$ (正確)

Central Limit Theorem (CLT)

- Q: 中央極限定理是: ______ 的極限分配是標準常態; The limiting distribution of ______ is standard normal.
- A: 不論什麼分配 (錯)
- A: 樣本平均數, $\overline{X}$ (錯)
- A: 標準化後的樣本平均數: $\frac{\overline{X}_{(n)}-\mu}{\mathrm{se}(\overline{X}_{(n)})}$ (正確)
- Definition of CLT: $X_1, \ldots, X_n$, a random sample with mean μ and variance σ^2.
  $Z = \frac{\overline{X}_{(n)}-\mu}{\mathrm{se}(\overline{X}_{(n)})} \sim N(0, 1)$ as $n \to \infty$

How do we show CLT via Plots?

Most Books: x-axis: $\overline{x}$

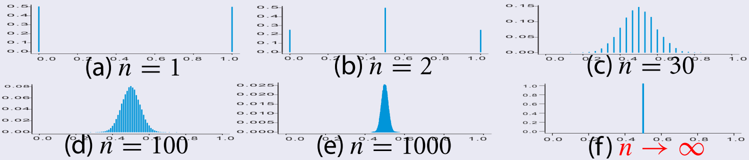


(a) $n = 1$ (b) $n = 2$ (c) $n = 30$

(d) $n = 100$ (e) $n = 1000$ (f) $n \to \infty$

How do we show CLT via Plots?

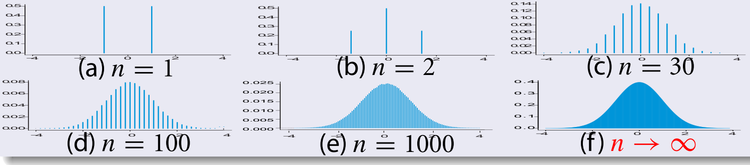
Most Books: x-axis: $\overline{x}$



(a) $n = 1$ (b) $n = 2$ (c) $n = 30$

(d) $n = 100$ (e) $n = 1000$ (f) $n \to \infty$

Song's book: x-axis: $\frac{\overline{x}_{(n)} - \mu}{\sigma/\sqrt{n}}$



(a) $n = 1$ (b) $n = 2$ (c) $n = 30$

(d) $n = 100$ (e) $n = 1000$ (f) $n \to \infty$

Learning CLT

My Ways: Q, Plots, Play with data, Poem

- 1. Ask questions. (The limiting distribution of ________ is Standard normal.)
- 2. Plots. (see Plot $Z = \frac{\overline{x}_{(n)} - \mu}{\sigma/\sqrt{n}}$ vs. $\overline{x}_{(n)}$)
- 3. “Prove CLT” via simulation approach. (next pages)
- 4. 打油詩 (博君一笑):「將相本無種, 人人當自強」

Learning CLT

My Ways: Q, Plots, Play with data, Poem

- 1. Ask questions. (The limiting distribution of ________ is Standard normal.)
- 2. Plots. (see Plot $Z = \frac{\overline{x}_{(n)} - \mu}{\sigma / \sqrt{n}}$ vs. $\overline{x}_{(n)}$)
- 3. “Prove CLT” via simulation approach. (next pages)
- 4. 打油詩 (博君一笑):「將相本無種, 人人當自強」
- 將相: N(0,1)

Learning CLT

My Ways: Q, Plots, Play with data, Poem

- 1. Ask questions. (The limiting distribution of ___________ is Standard normal.)
- 2. Plots. (see Plot $Z = \frac{\overline{x}_{(n)} - \mu}{\sigma/\sqrt{n}}$ vs. $\overline{x}_{(n)}$)
- 3. “Prove CLT” via simulation approach. (next pages)
- 4. 打油詩 (博君一笑):「將相本無種, 人人當自強」
- 將相: N(0,1)
- 無種, 祖先不必是皇親貴族: $X_1, \ldots, X_n$ can follow any distribution

Learning CLT

My Ways: Q, Plots, Play with data, Poem

- 1. Ask questions. (The limiting distribution of ___________ is Standard normal.)
- 2. Plots. (see Plot $Z = \frac{\overline{x}_{(n)} - \mu}{\sigma/\sqrt{n}}$ vs. $\overline{x}_{(n)}$)
- 3. “Prove CLT” via simulation approach. (next pages)
- 4. 打油詩 (博君一笑):「將相本無種, 人人當自強」
- 將相: N(0,1)
- 無種, 祖先不必是皇親貴族: $X_1, \ldots, X_n$ can follow any distribution
- 自強: $n \to \infty$

Learning CLT

My Ways: Q, Plots, Play with data, Poem

- 1. Ask questions. (The limiting distribution of __________ is Standard normal.)
- 2. Plots. (see Plot $Z = \frac{\overline{x}_{(n)} - \mu}{\sigma/\sqrt{n}}$ vs. $\overline{x}_{(n)}$)
- 3. "Prove CLT" via simulation approach. (next pages)
- 4. 打油詩 (博君一笑):「將相本無種, 人人當自強」
- 將相: N(0,1)
- 無種, 祖先不必是皇親貴族: $X_1, \ldots, X_n$ can follow any distribution
- 自強: $n \to \infty$

- Comparing the way you learned CLT with the ways list above. Comment.

Analytical Proof of CLT

Prove CLT via mgf (動差母函數)

Proof:

$$\begin{aligned}
\lim_{n\to\infty} \mathrm{M}_{Z_n}(t) &= \lim_{n\to\infty} [\mathrm{M}_{Y_i}(t/\sqrt{n})]^n && \text{Hint 1}\\
&= \lim_{n\to\infty} [\mathrm{M}_{Y_i}(0) + \frac{t}{n^{1/2}}\mathrm{M}'_{Y_i}(0) + \frac{t^2}{2!n}\mathrm{M}''_{Y_i}(0) + \frac{t^3}{3!n^{3/2}}\mathrm{M}'''_{Y_i}(0) + \cdots]^n && \text{Hint 2}\\
&= \lim_{n\to\infty} [1 + \frac{t^2}{2n} + \frac{t^3}{3!n^{3/2}}\mathrm{M}'''_{Y_i}(0) + \cdots]^n && \text{Hint 3}\\
&= \lim_{n\to\infty} [1 + (\frac{t^2}{2} + \frac{t^3}{3!n^{1/2}}\mathrm{M}'''_{Y_i}(0) + \cdots)/n]^n && \text{Algebra}\\
&= e^{t^2/2} && \text{Hint 4}
\end{aligned}$$

- Hint 1: $\mathrm{M}_{Z_n}(t) = [\mathrm{M}_{Y_i}(t/\sqrt{n})]^n$, where $Y_i = (X_i - \mu)/\sigma$, $Z_n = \dfrac{\overline{X}_{(n)} - \mu}{\sigma/\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i$
- Hint 2 (泰勒展開式): $\mathrm{M}_{Y_i}(\frac{t}{n^{1/2}}) = \mathrm{M}_{Y_i}(0) + \frac{t}{n^{1/2}}\mathrm{M}'_{Y_i}(0) + \frac{t^2}{2!n}\mathrm{M}''_{Y_i}(0) + \frac{t^3}{3!n^{3/2}}\mathrm{M}'''_{Y_i}(0) + \cdots$
- Hint 3: $\mathrm{M}_{Y_i}(0) = 1, \mathrm{M}'_{Y_i}(0) = 0, \mathrm{M}''_{Y_i}(0) = \mathrm{E}[Y_i^2] = 1$
- Hint 4: If $\lim_{n\to\infty} a_n = b$, then $\lim_{n\to\infty}(1 + \frac{a_n}{n})^n = e^b$, where $a_n = \frac{t^2}{2} + \frac{t^3}{3!n^{1/2}}\mathrm{M}'''_{Y_i}(0) + \cdots, \quad b = t^2/2$

Prove CLT via simulation approach

- Generating data from MSExcel or Minitab

Prove CLT via simulation approach

- Generating data from MSExcel or Minitab
- Let $X_{i,j}$ denote the data in the (i, j) position of the data-matrix; $i = 1, 2, \ldots, 10000$; $j = 1, 2, \ldots, 100$

Prove CLT via simulation approach

- Generating data from MSExcel or Minitab
- Let $X_{i,j}$ denote the data in the (i, j) position of the data-matrix; $i = 1, 2, \ldots, 10000$; $j = 1, 2, \ldots, 100$
- Plot the histogram of $(\bar{X}(n) - \mu_{\bar{X}})/\mathrm{se}(\bar{X}), n = 1, 2, 30, 100$

Prove CLT via simulation approach

- Generating data from MSExcel or Minitab
- Let $X_{i,j}$ denote the data in the (i, j) position of the data-matrix; $i = 1, 2, \ldots, 10000$; $j = 1, 2, \ldots, 100$
- Plot the histogram of $(\bar{X}(n) - \mu_{\bar{X}})/\mathrm{se}(\bar{X}), n = 1, 2, 30, 100$
- Discussion:
  - Distinguish histogram and pdf

Prove CLT via simulation approach

- Generating data from MSExcel or Minitab
- Let $X_{i,j}$ denote the data in the (i, j) position of the data-matrix; $i = 1, 2, \ldots, 10000$; $j = 1, 2, \ldots, 100$
- Plot the histogram of $(\bar{X}(n) - \mu_{\bar{X}})/\mathrm{se}(\bar{X}), n = 1, 2, 30, 100$
- Discussion:
  - Distinguish histogram and pdf
  - Distinguish data in row and column from Excel or Minitab

Prove CLT via simulation approach

- Generating data from MSExcel or Minitab
- Let $X_{i,j}$ denote the data in the (i, j) position of the data-matrix; $i = 1, 2, \ldots, 10000$; $j = 1, 2, \ldots, 100$
- Plot the histogram of $(\bar{X}(n) - \mu_{\bar{X}})/\mathrm{se}(\bar{X}), n = 1, 2, 30, 100$
- Discussion:
  - Distinguish histogram and pdf
  - Distinguish data in row and column from Excel or Minitab
  - Distinguish data from different distributions

Prove CLT via simulation approach

- Generating data from MSExcel or Minitab
- Let $X_{i,j}$ denote the data in the (i, j) position of the data-matrix; $i = 1, 2, \ldots, 10000$; $j = 1, 2, \ldots, 100$
- Plot the histogram of $(\bar{X}(n) - \mu_{\bar{X}})/\mathrm{se}(\bar{X}), n = 1, 2, 30, 100$
- Discussion:
  - Distinguish histogram and pdf
  - Distinguish data in row and column from Excel or Minitab
  - Distinguish data from different distributions

Generating Data from MSExcel

- A1: No.
- B1: Xbar(2)
- C1: (Xbar(2)-mu)/SE
- D1: X1: expon (mean=1)
- E1: X2: expon (mean=1)
- A2-A10001: $1, 2, \ldots, 10000$
- D2-D10001: = - ln (1-RAND())
- E2-E10001: = - ln (1-RAND())
- B2: =AVERAGE(D2:E2)
- C2: = (B2-1)/(1/SQRT(2))
- Plot the histogram of data in C2

- Calc > Random Data > Exponential
- Exponential Distribution:
  - Number of rows of data: 10000
  - Store in column(s): C6-C106
  - Scale: 1 (=Mean when Threshold =0.0)
  - Threshold: 0.0
- Calc > Row Statistics
  - Mean
  - Input Variable: C7-C8
  - Store result in: C1
- Calc >Calculator
  - Store result in variable: C2
  - Expression: (C1-1)/(1/sqrt(2))
- Plot the histogram of data in C2 via Minitab

HW: Comparing Different ways of Learning CLT

- 選擇任意兩本統計學書. 最好其中一本是你大學時統計學課本.
- 比較書中解釋 CLT 與本講義解釋 CLT 的不同. 哪一種方法學習效果比較好?
- 3-4 人為一組, 上台報告你們的看法與發現 (Power Point file or Tex Beamer)