STA 103 Lecture 8: The Central Limit Theorem

Instructor: Wookyeong Song

Department of Statistics, University of California, Davis

Aug 20th, 2025

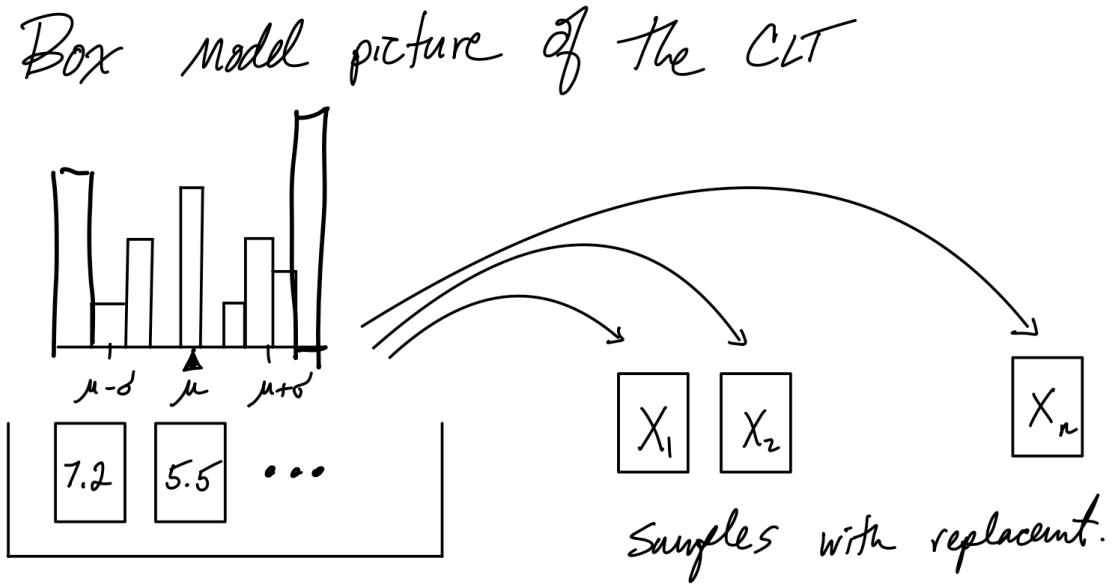
The Central Limit Theorem (CLT)

- **Theorem (CLT)**: Let $X_1, X_2, \ldots, X_n$ are independent random variables all with the same PMF or PDF, and the sample mean $\bar{X} = \frac{X_1 + \cdots + X_n}{n}$. Then for large n ($n \geq 30$ usually),

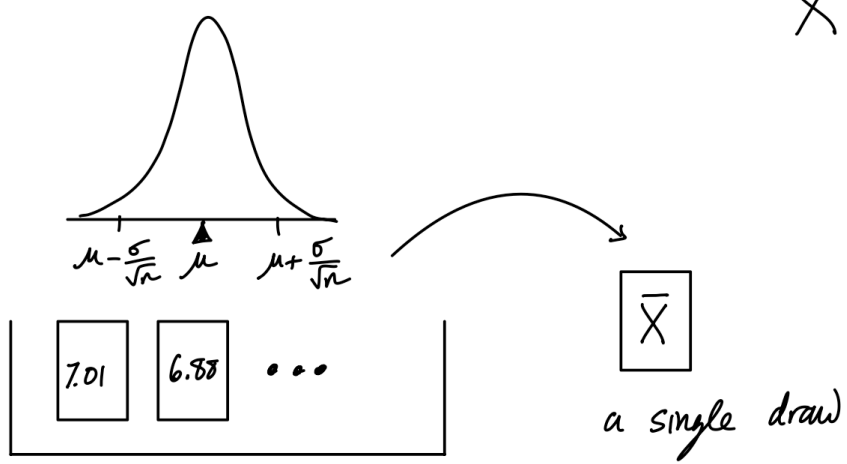
$$\bar{X} \approx N(a, b),$$

  where $a = E(\bar{X})$ and $b = \mathrm{Var}(\bar{X})$.

The Central Limit Theorem (CLT)



$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$



The Central Limit Theorem (CLT)

- **Example**: Let $X_1, X_2, \ldots, X_{100}$ be independent RVs all with the following PMF:

$$p(x) = \begin{cases} \binom{22}{x} \left(\frac{1}{3}\right)^x \left(\frac{2}{3}\right)^{22-x} & \text{for } x = 0, 1, \ldots, 22 \\ 0 & \text{otherwise.} \end{cases}$$

- Let

$$\bar{X} = \frac{X_1 + \cdots + X_{100}}{100}.$$

- **Question**: Find $P(\bar{X} > 7)$.

- Unfortunately, calculating $P(X_1 > 7)$ directly is already tricky, and $P(\bar{X} > 7)$ is even more so.

The Central Limit Theorem (CLT)

- But the CLT lets us approximate $P(\bar{X} > 7)$ easily.

- In particular, the CLT says

$$\bar{X} \approx N(a, b).$$

  Then, we can use Z-score

$$Z = \frac{\bar{X} - E(\bar{X})}{\mathrm{sd}(\bar{X})} \approx N(0, 1).$$

- We know that $X_i \sim \text{Binomial}(22, \frac{1}{3})$ so

$$E(\bar{X}) = E(X_1) = 22 \times \frac{1}{3} = \frac{22}{3},$$

  and

$$\mathrm{Var}(\bar{X}) = \frac{\mathrm{Var}(X_1)}{100} = \left(22 \times \frac{1}{3} \times \frac{2}{3}\right) \times \frac{1}{100} = \frac{44}{900}.$$

The Central Limit Theorem (CLT)

- Now the problem can be solved by standardization:

$$P(\bar{X} > 7) = P\left(\frac{\bar{X} - E(\bar{X})}{\mathrm{sd}(\bar{X})} > \frac{7 - E(\bar{X})}{\mathrm{sd}(\bar{X})}\right)$$
$$= P\left(Z > \frac{7 - \frac{22}{3}}{\sqrt{\frac{44}{900}}}\right) = P(Z > -1.508).$$

- Using the Z-table:

$$P(Z > -1.508) = 1 - P(Z \leq -1.508) = 0.9342.$$

Applying CLT to Sum of Random Variables

- **Example 1**: Suppose the average height of California residents is 170.18 (cm) with a standard deviation of 6.35 (cm). Randomly pick 57 people from California and let $X_1, X_2, \ldots, X_{57}$ be their heights. Let Y be the sum of 57 California residents, $Y = X_1 + X_2 + \cdots + X_{57}$. Use the CLT to approximate

$$P(Y < 9600).$$

- What is the underlying distribution of heights in California? We don't know, it wasn't specified as normal or any particular form.

- It does not matter. We can still apply the CLT (what a powerful tool). The CLT says

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_{57}}{57} \approx N\left(E(\bar{X}), \mathrm{sd}(\bar{X})^2\right).$$

Applying CLT to Sum of Random Variables

- Since

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_{57}}{57} \approx N(\cdot, \cdot),$$

  we know that

$$Y = 57 \times \bar{X} \approx N\left(E(Y), \mathrm{sd}(Y)^2\right).$$

- **(Why?)** We know that linear combination of Normal distribution is also Normal distribution.

- Then, using Z-score standardization to Y, we have

$$Z = \frac{Y - E(Y)}{\mathrm{sd}(Y)} \approx N(0, 1).$$

- **Question**: Find $E(Y)$ and $\mathrm{sd}(Y)$.

Applying CLT to Sum of Random Variables

- **Answer**: We have

$$\begin{aligned} E(Y) &= E(X_1) + E(X_2) + \cdots + E(X_{57}) \\ &= 57 \times E(X_1) = 57 \times 170.18. \end{aligned}$$

- For variance, we have

$$\begin{aligned} \mathrm{Var}(Y) &= \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_{57}) \\ &= 57 \times \mathrm{Var}(X_1) = 57 \times (6.35)^2. \end{aligned}$$

- Then we have

$$\begin{aligned} P(Y < 9600) &= P\left(Z < \frac{9600 - 57 \times 170.18}{\sqrt{57 \times (6.35)^2}}\right) \\ &= P(Z < -2.09). \end{aligned}$$

- **Takeaway**: We can apply the CLT not only to the sample mean $\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$, but also the sum of independent and identically distributed (i.i.d.) RVs $Y = X_1 + X_2 + \cdots + X_n$.

More Practice

- **Example 2**: Suppose $X_1, X_2, \ldots, X_{92}$ are i.i.d. where $E(X_1) = -\frac{1}{2}$ and $\mathrm{Var}(X_1) = 10$. Use the CLT to approximate

$$P(|\bar{X}| > 0.1),$$

where $\bar{X} = \frac{X_1 + X_2 + \cdots + X_{92}}{92}$.

- **Answer for Example 2**: We know that

$$E(\bar{X}) = E(X_1) = -\frac{1}{2}, \quad \mathrm{Var}(\bar{X}) = \frac{\mathrm{Var}(X_1)}{92} = \frac{10}{92}.$$

More Practice

- **Answer for Example 2 (Conti.)**: By applying the CLT, we have

$$\bar{X} \approx N(E(\bar{X}), \mathrm{Var}(\bar{X})) = N\left(-\frac{1}{2}, \frac{10}{92}\right),$$

we have Z-score:

$$Z = \frac{\bar{X} - E(\bar{X})}{\mathrm{sd}(\bar{X})} = \frac{\bar{X} + \frac{1}{2}}{\sqrt{10/92}} \sim N(0, 1).$$

We know that

$$\begin{aligned} P(|\bar{X}| > 0.1) &= P(\bar{X} > 0.1 \text{ or } \bar{X} < -0.1) \\ &= P(\bar{X} > 0.1) + P(\bar{X} < -0.1) \\ &= P\left(\frac{\bar{X} + \frac{1}{2}}{\sqrt{10/92}} > \frac{0.1 + \frac{1}{2}}{\sqrt{10/92}}\right) + P\left(\frac{\bar{X} + \frac{1}{2}}{\sqrt{10/92}} < \frac{-0.1 + \frac{1}{2}}{\sqrt{10/92}}\right) \\ &= P(Z > 1.82) + P(Z < 1.21) = 0.9213. \end{aligned}$$

More Practice

- **Business Application**: Suppose a coffee shop's daily revenue X is a random variable with mean 1000 (in dollars) and standard deviation 300 (in dollars). Over 40 randomly chosen days, what is the probability that the average revenue is greater than 1050 (in dollars)?

- **Answer**: Average revenue is $\bar{X}$. Applying the CLT,

$$\bar{X} \approx N(E(\bar{X}), \mathrm{Var}(\bar{X})) = N(1000, \frac{300^2}{40}) = N(1000, 2250).$$

To find the probability of average revenue greater than 1050,

$$\begin{aligned} P(\bar{X} > 1050) &= P\left(\frac{\bar{X} - 1000}{\sqrt{2250}} > \frac{1050 - 1000}{\sqrt{2250}}\right) = P(Z > 1.05) \\ &= 1 - P(Z \leq 1.05) = 1 - 0.8531 = 0.1469. \end{aligned}$$

Empirical Rule

- The CLT explains why so many RVs observed in nature look bell shaped because many of the random X values we encounter are sums of nearly independent random variables, so the CLT applies $X \approx N(\cdot, \cdot)$.

- **For example**, Binomial Random Variables $Y \sim \text{Binomial}(n, p)$.

- **(Lecture 3)** A binomial random variable is the sum of n i.i.d. Bernoulli random variables with the same success probability p:

$$Y = X_1 + X_2 + \cdots + X_n, \quad X_i \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p).$$

- The PMF of Binomial RV, getting exactly y successes in n trials, is

$$P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}.$$

Applying CLT to Binominal Random Variables

- **Expectation**:

$$\begin{aligned} E(Y) &= E(X_1 + X_2 + \cdots + X_n) \\ &= E(X_1) + E(X_2) + \cdots + E(X_n) \\ &= n \times E(X_1) = np. \end{aligned}$$

- **Variance**:

$$\begin{aligned} \mathrm{Var}(Y) &= \mathrm{Var}(X_1 + X_2 + \cdots + X_n) \\ &= \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n) \\ &= n \times \mathrm{Var}(X_1) = np(1-p). \end{aligned}$$

- Then, we can approximate $Y \sim N(np, np(1-p))$, then

$$P(Y < k) \approx P\left(Z < \frac{k - np}{\sqrt{np(1-p)}}\right),$$

$$P(Y > k) \approx P\left(Z > \frac{k - np}{\sqrt{np(1-p)}}\right).$$

Applying CLT to Binominal Random Variables

- **Example** (Market Research Analyst): You are analyzing customer response behavior in an online survey conducted by a retail economics team. From past data, it is known that the probability a randomly selected customer responds to the survey is $p = 0.2$. You randomly contact $n = 50$ customers. Let the random variable X denote the number of customers who respond.

- **Question**: What is the probability that at least 9 customers respond?

- **Answer**: Since $X \sim \text{Binomial}(50, 0.2)$, we can apply the CLT,

$$X \approx N(50 \times 0.2,\ 50 \times 0.2 \times 0.8) = N(10, 8).$$

Then, we have

$$\begin{aligned} P(X \geq 9) &= P\left(\frac{X - 10}{\sqrt{8}} \geq \frac{9 - 8}{\sqrt{8}}\right) \\ &= P(Z \geq 0.35) = 1 - P(Z \leq 0.35) = 1 - 0.6368 = 0.3632. \end{aligned}$$

Empirical Rule

- The CLT explains why so many RVs observed in nature look bell shaped because many of the random X values we encounter are sums of nearly independent random variables, so the CLT applies $X \approx N(\cdot, \cdot)$.

- This explains the empirical rule, even the underlying distribution of X is **not** Normal:
  - The chance that X falls within $E(X) \pm \mathrm{sd}(X)$ is approximately 68%.
  - The chance that X falls within $E(X) \pm 2\ \mathrm{sd}(X)$ is approximately 95%.
  - The chance that X falls within $E(X) \pm 3\ \mathrm{sd}(X)$ is approximately 99.7%.

- This gives you **Rationale**: In Lecture 7, here is very informal guideline when underlying distribution is unknown, especially not assumed to be Normal distribution:
  - Z-score within $\pm$ 2 is typical.
  - Z-score 3 or -3 is rare but possible.
  - Z-score greater than $|4|$ is extremely rare.