The Central Limit Theorem

Tan Do

Vietnamese-German University

Lecture 15

In this lecture

- Linear combinations of normal random variables
- Central limit theorem

Recall the following:

Let $X \sim N(\mu, \sigma^2)$. Then

$$Y = aX + b \sim N(a\mu + b, a^2\sigma^2).$$

Let $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ be independent random variables. Then

$$Y = X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

Let $X_i \sim N(\mu, \sigma^2)$, $i = 1, 2, \ldots, n$, be independent random variables. Then

$$\overline{X} = \frac{X_1 + \ldots + X_n}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Generally, let $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, 2, \ldots, n$, be independent random variables. Then

$$Y = a_1 X_1 + \ldots + a_n X_n + b \sim N(\mu, \sigma^2),$$

where

$$\mu = a_1 \mu_1 + \ldots + a_n \mu_n + b$$

and

$$\sigma^2 = a_1^2 \sigma_1^2 + \ldots + a_n^2 \sigma_n^2.$$

Example Recall that

- the radius of a piston head X_1 has $\mu = 30$ mm and $\sigma = 0.05$ mm
- the inside radius of a cylinder X_2 has $\mu = 30.25$ mm and $\sigma = 0.06$ mm
- the gap between the piston head and the cylinder $Y = X_2 - X_1$ therefore has

$$\mu = 30.25 - 30 = 0.25 \quad \text{and} \quad \sigma^2 = 0.05^2 + 0.06^2 = 0.0061.$$

Suppose $X_1 \sim N(30, 0.05^2)$ and $X_2 \sim N(30.25, 0.06^2)$. Then

$$Y \sim N(0.25, 0.0061).$$

The probability that

- a piston head does not fit within a cylinder is

$$P(Y \leq 0) = \Phi\Big(\frac{0 - 0.25}{\sqrt{0.0061}}\Big) = \Phi(-3.2) = 0.0007.$$

- a piston performs optimally (gap between 0.1 mm and 0.35 mm) is

$$P(0.1 \leq Y \leq 0.35) = \Phi\Big(\frac{0.35 - 0.25}{\sqrt{0.0061}}\Big) - \Phi\Big(\frac{0.1 - 0.25}{\sqrt{0.0061}}\Big) = \Phi(1.28) - \Phi(-1.92) = 0.8723.$$

Example Recall that the height of a tomato plant (in cm) 3 weeks after planting is given by $N(29.4, 2.1^2)$. Suppose that 20 tomato plants are planted.

- What is the distribution of the average tomato plant height after three weeks of growth?
- Which height interval gives 95% chance that the average tomato plant height lies within?

Answer

- The distribution is

$$\overline{X} \sim N\big(29.4, \frac{2.1^2}{20}\big) = N(29.4, 0.2205).$$

- Note $z_{0.025} = 1.96$. So the interval is

$$[29.4 - 1.96 \times \sqrt{0.2205}, 29.4 + 1.96 \times \sqrt{0.2205}] = [28.48, 30.32].$$

Example A chemist has two different methods for measuring the concentration level C of a chemical solution.

- Method A produces a measurement $X_A \sim N(C, 2.97)$.
- Method B produces a measurement $X_B \sim N(C, 1.62)$.

Since $\sigma_B < \sigma_A$, it is clear that B gives a more accurate measurement. But should A be completely ignored?

Answer No. Combining A and B improves the measurement. Let Y be a weighted average of the measurements from A and B. Then

$$Y = pX_A + (1-p)X_B, \quad p \in [0,1].$$

So

$$E(Y) = pE(X_A) + (1-p)E(X_B) = pC + (1-p)C = C$$

and

$$\mathrm{Var}(Y) = p^2\mathrm{Var}(X_A) + (1-p)\mathrm{Var}(X_B) = 2.97p^2 + 1.62(1-p)^2.$$

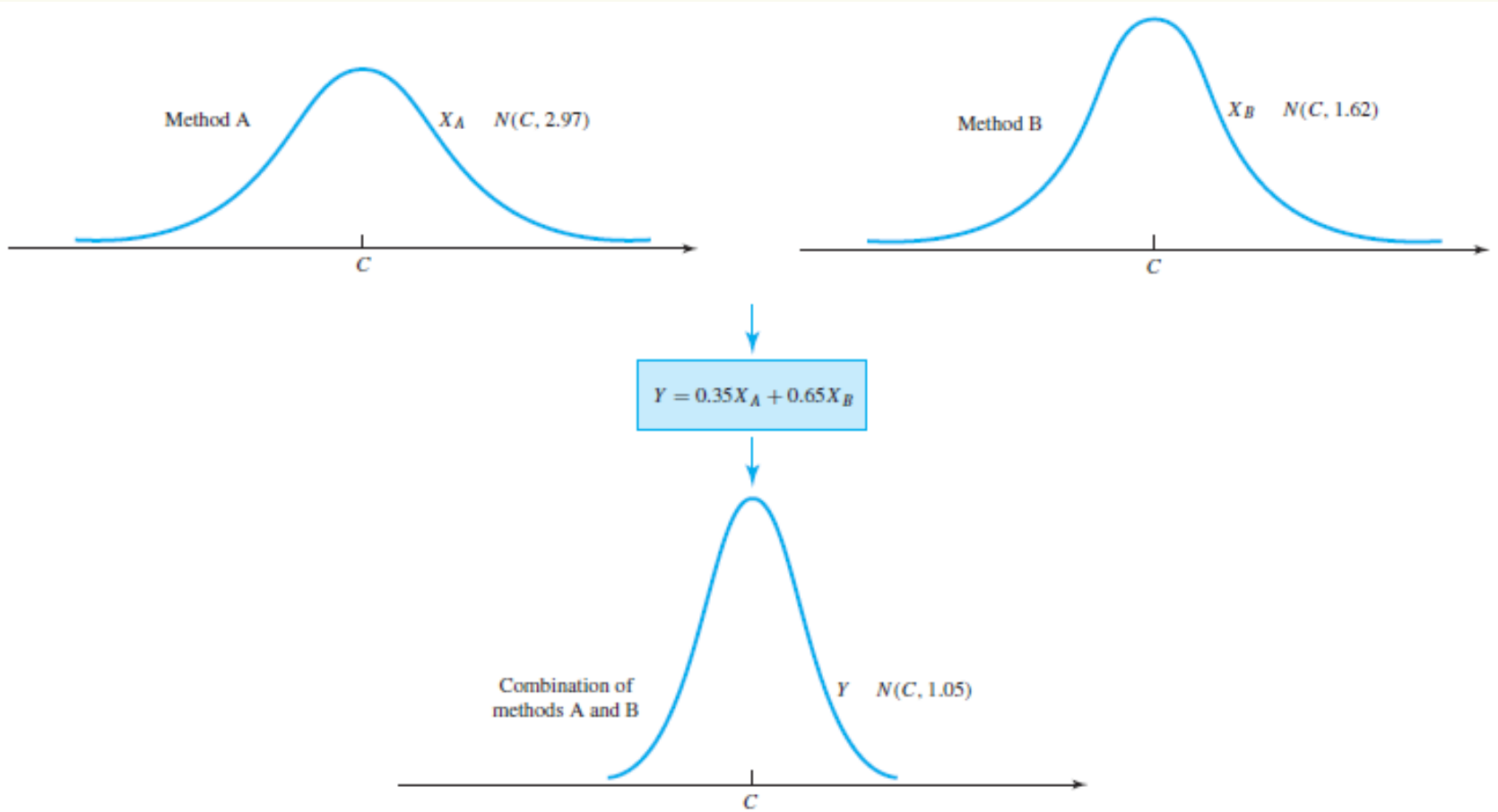
Y is best when $\mathrm{Var}(Y)$ is minimal:

$$\big(\mathrm{Var}(Y)\big)' = 5.94p - 3.24(1-p) = 0 \quad \Longrightarrow \quad p = 0.35.$$

In this case

$$\mathrm{Var}(Y) = 2.97 \times 0.35^2 + 1.62 \times (1-0.35)^2 = 1.05$$

which is smaller than $\mathrm{Var}(X_B)$.



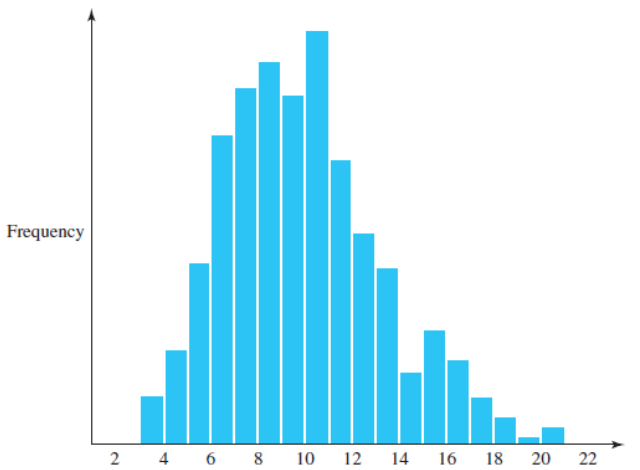
The central limit theorem

Motivation Let $X_1, \ldots, X_n$ be identically distributed (i.e. they have the same distribution). We investigate the distribution of the average

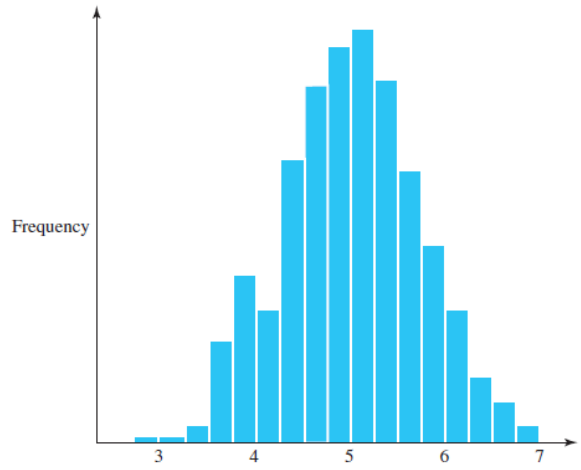
$$\overline{X} = \frac{X_1 + \ldots + X_n}{n}.$$

Simulation Let $n = 10$. So $\overline{X} = \frac{X_1 + \ldots + X_{10}}{10}$.

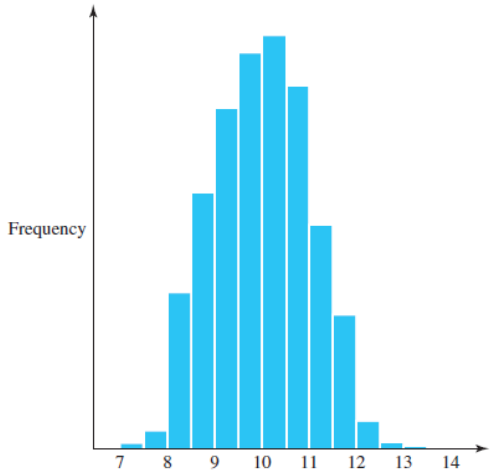
We use histograms to see the shape of the p.d.f of $\overline{X}$ when X_i's take on a certain type of distribution.



Histogram of the averages of simulated exponential random variables



Histogram of the averages of simulated beta random variables



Histogram of the averages of simulated Poisson random variables

The Central Limit Theorem (CLT) Let $X_1, \ldots, X_n$ be identically distributed with mean μ and variance σ^2 (regardless of distribution type).
Then the distribution of $\overline{X} = \dfrac{X_1 + \ldots + X_n}{n}$ can be approximated by

$$N\Big(\mu, \frac{\sigma^2}{n}\Big).$$

Corollary Let $X_1, \ldots, X_n$ be identically distributed with mean μ and variance σ^2. Then the distribution of $X = X_1 + \ldots + X_n$ can be approximated by

$$N(n\mu, n\sigma^2).$$

In particular,

$$B(n,p) \approx N\big(np, np(1-p)\big).$$

Let $X \sim B(n,p)$. Then

$$P(X \le x) \approx \Phi\Big(\frac{x + 0.5 - np}{\sqrt{np(1-p)}}\Big) \quad \text{and} \quad P(X \ge x) \approx 1 - \Phi\Big(\frac{x - 0.5 - np}{\sqrt{np(1-p)}}\Big).$$

These approximations work well if $np \ge 5$ and $n(1-p) \ge 5$.

Example Recall that there is a probability of 0.261 that a milk container is underweight. Consequently, the number of underweight containers X in a box of 20 containers has a $B(20, 0.261)$ distribution. By the CLT,

$$X \sim B(20, 0.261) \approx Y \sim N(20 \times 0.261, 20 \times 0.261 \times (1 - 0.261)) = N(5.22, 3.86).$$

We can double-check the approximation as follows:

- The probability that a box contains no more than three underweight containers was calculated before:

$$P(X \leq 3) = 0.1935.$$

- Using the normal approximation,

$$P(X \leq 3) \approx P(Y \leq 3.5) = \Phi\Big(\frac{3.5 - 5.22}{\sqrt{3.86}}\Big) = 0.1922$$

  which is very close.

Example Recall that there is a probability of 0.6 that an oyster produces a pearl with a diameter of at least 4 mm, which is of commercial value. How many oysters does an oyster farmer need to farm in order to be 99% confident of having at least 1000 pearls of commercial value?

Answer Let n be the number of oysters to be farmed. Then the distribution of the number of pearls of commercial value is $X \sim B(n, 0.6)$. The CLT gives

$$X \approx Y \sim N(0.6n, 0.24n).$$

The probability of having at least 1000 pearls of commercial value is then

$$P(X \geq 1000) \approx P(Y \geq 999.5) = 1 - \Phi\Big(\frac{999.5 - 0.6n}{\sqrt{0.24n}}\Big) = 1 - \Phi(t).$$

We want

$$1 - \Phi(t) \geq 0.99 \iff \Phi(t) \leq 0.01 \iff t \leq -2.33 \iff n \geq 1746.$$