CEE 260/MIE 273: Probability and Statistics in Civil Engineering

# Lecture M4b: Sampling Variability and Central Limit Theorem

**Jimi Oke**

UMass Amherst

College of Engineering

October 28, 2026

# Outline

1. Method of moments
2. CLT
3. Outlook

# Sample moments

- The moments of a random variable are its key descriptors.
- Parameters of the distribution of a random variable are usually related to the first and second moments (mean and variance, respectively)

Given a sample $x_1, x_2, \ldots, x_n$, the point estimates of the population mean $\mu$ and variance $\sigma^2$ are:

## Sample mean

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{1}$$

## Sample variance

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 \tag{2}$$

# Sample mean and variance

## Example 1: Elastic modulus of alloys

The elastic modulus (GPa) of a sample of alloy specimens from a die-casting process is:

$$X = 44.2, 43.9, 44.7, 44.2, 44.0, 43.8, 44.6, 43.1$$

(a) Estimate the population mean using the estimator $\overline{x}$ (sample mean)

(b) Estimate the population variance using the estimator $s^2$ (sample variance)

(c) Now, estimate the variance replacing the denominator $(n-1)$ with $n$ in the estimator $s^2$. What do you notice?

# Sample mean and variance

## Example 1: Elastic modulus of alloys (cont.)

$X = 44.2, 43.9, 44.7, 44.2, 44.0, 43.8, 44.6, 43.1$

(a) $\hat{\mu} = \overline{x} = \frac{1}{8}\sum_{i=1}^{8} x_i \approx \boxed{44.063}$

(b) $s^2 = \frac{1}{7}\left[\sum_{i=1}^{8} x_i^2 - 8(44.063^2)\right] \approx \boxed{0.251}$

(c) Biased estimate of $\sigma^2$:

$\hat{\sigma}^2 = \frac{1}{8}\left[\sum_{i=1}^{8} x_i^2 - 8(44.063^2)\right] = \frac{7}{8}(0.251) = \boxed{0.220}$

$\hat{\sigma}^2$ underestimates $\sigma^2$ by 0.031 squared units.

# The Central Limit Theorem (CLT)

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean $\mu$ and variance $\sigma^2$. If $n$ is sufficiently large, then the sample mean $\overline{X}$ has approximately a **normal distribution** with

$$\mu_{\overline{X}} = \mu \tag{3}$$

$$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n} \tag{4}$$

and the sample total, $S_n = X_1 + X_2 + \ldots + X_n$, has approximately a normal distribution with

$$\mu_S = n\mu \tag{5}$$

$$\sigma^2_S = n\sigma^2 \tag{6}$$

Implications:

- The sum of a **large number** of random components approaches a **normal/Gaussian distribution**
- The product of large number of random components approaches the lognormal distribution

# Central limit theorem (cont.)

## Sample mean

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} \tag{7}$$

## Sum of sample observations

$$S_n = X_1 + X_2 + \cdots + X_n \tag{8}$$

If $n$ is sufficiently large for **any** sample:

$$\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \tag{9}$$

$$S_n \sim \mathcal{N}(n\mu, n\sigma^2) \tag{10}$$

Note that the quantity $\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$ is also known as the **sampling error** (SE) or the **standard error of the mean** (SEM)

# Sample proportion and the CLT

If the observations in a given sample are a Bernoulli sequence with a constant proportion (or probability) $p$, then if $n$ is large, the sample proportion $\hat{p}$ follows a normal distribution (according to the CLT):

$$\hat{p} \sim \mathcal{N}(\mu_{\hat{p}}, SE_{\hat{p}}^2) = \mathcal{N}\left(p, \frac{p(1-p)}{n}\right) \tag{11}$$

where

$$\text{Sample mean proportion: } \mu_{\hat{p}} = p$$

$$\text{Sampling error/standard error of } \hat{p}\text{: } SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

One rule of thumb for determining whether $n$ is large enough is to check that both $np$ and $n(1-p)$ are $\geq 10$ (also known as the success-failure condition).

# Success-failure condition

In the case of a proportion $p$, the CLT holds only if:

- The observations are independent (i.e. random)
- The sample size $n$ is **sufficiently large**

The second condition is typically observed via the **success-failure condition**, i.e.:

$$np \geq 10 \tag{12}$$

$$n(1-p) \geq 10 \tag{13}$$

# CLT application: sample proportion

## Example 2: Solar energy expansion (CLT)

Suppose the proportion of American adults who support the expansion of solar energy is $p = 0.88$, which is our parameter of interest. If we were to take a poll of 1000 American adults on this topic, the estimate would not be perfect, but how close might we expect the sample proportion in the poll would be to 88%?

**(a)** According to the CLT, what is the distribution of $\hat{p}$?

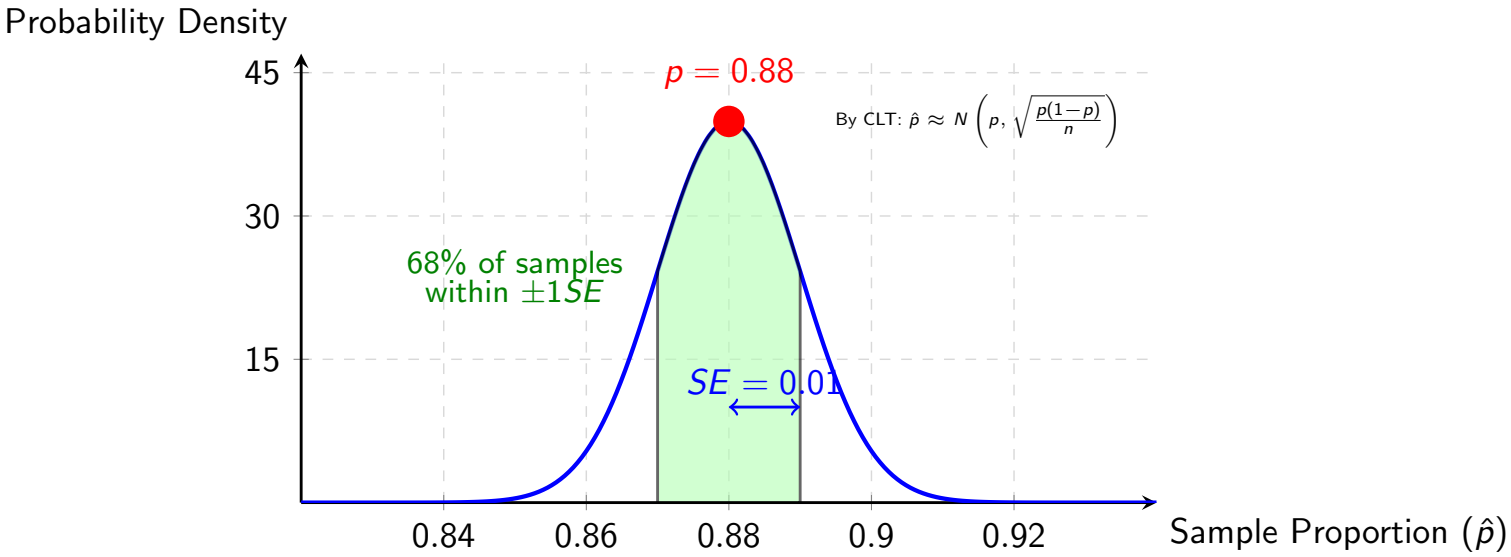
**(b)** According to the CLT, what are $\mu_{\hat{p}}$ and $SE_{\hat{p}}$, respectively?

# CLT application: sample proportion (cont.)

## Example 2: Solar energy expansion (CLT)

**(a)** First, we note that the response of each American adult in the entire population is part of a Bernoulli sequence with $p = 0.88$. According to the CLT, the distribution of $\hat{p}$ (sample proportion) is normal/Gaussian. We can denote this as:

$$\hat{p} \sim \mathcal{N}\left(p, \frac{\sigma^2}{n}\right) \text{ OR } \mathcal{N}\left(\mu_p, \frac{\sigma_p^2}{n}\right) \tag{14}$$



# CLT application: sample proportion (cont.)

## Example 2: Solar energy expansion (CLT)

**(b)** $\mu_{\hat{p}}$ denotes the mean estimate of $p$, which is 0.88 (according to the CLT, the mean of the sample is the population mean if $n$ is large).
$SE_{\hat{p}}$ denotes the sampling error, which is the the square root of the variance of the sample mean: $\sqrt{\sigma^2/n}$. Given that the sample is governed by the Binomial distribution with $\sigma^2 = p(1-p)$. Thus:

$$SE_{\hat{p}}^2 = \frac{\sigma^2}{n} = \frac{p(1-p)}{n} = \frac{0.88(0.12)}{1000}$$

$$SE_{\hat{p}} = \sqrt{\frac{0.88(0.12)}{1000}} = \boxed{0.01}$$

# CLT application: sample proportion (cont.)

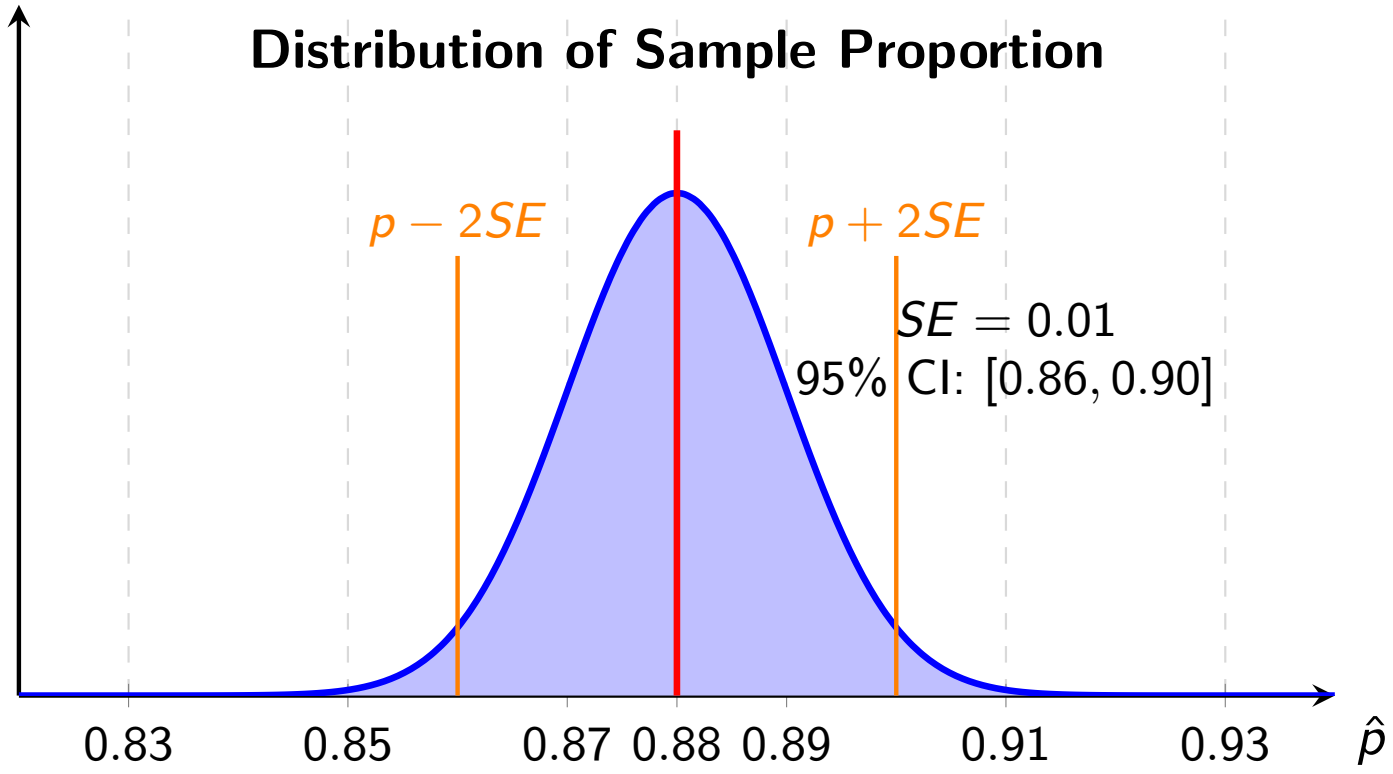



Figure: Sample proportion distribution: most samples fall within $\pm 2SE$ of the true proportion

# Another application of the CLT

## Example 3: Mean batch weight

A certain brand of cement is shipped in batches of 40 bags. Previous records indicate the weight of a randomly selected bag of this brand has a mean of 2.5 kg and an SD of 0.1 kg. The exact distribution is unknown.

(a) What is the mean weight of one batch of this brand of cement?

(b) If the shipping company charges an overweight fee if a batch exceeds the mean batch weight by more than 1 kg, what is the probability that a batch will be charged?

# Another application of the CLT

## Example 3: Mean batch weight (cont.)

Let $B$ be the total weight of one batch.

(a) The mean weight of one batch is thus

$$\mu_B = 40 \times 2.5 = 100 \text{ kg} \tag{15}$$

(b) By the CLT, $B$ is approximately normal with $\mu_B = 100$ and $\sigma_B^2 = 40(0.1)^2$.The probability a batch will be charged is:

$$\begin{aligned} P(B > 101) &= 1 - \Phi\left(\frac{101 - 100}{0.1\sqrt{40}}\right) \\ &= 1 - \Phi(1.581) \\ &= 1 - 0.9431 \approx \boxed{5.69\%} \end{aligned}$$

# Summary

- Desired properties of point estimates: unbiasedness and efficiency
- Distribution of sample proportions (or other parameters) is called a sampling distribution
- When $n$ is sufficiently large and observations are independent, the sample proportion (or other parameter) follows a normal distribution
- The success-failure condition can be used to determine if $n$ is large enough for the CLT to hold (for a sample proportion)

# CLT

If $n$ is sufficiently large for **any** sample:

$$\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \tag{16}$$

$$S_n \sim \mathcal{N}(n\mu, n\sigma^2) \tag{17}$$

For proportions:

$$\hat{p} \sim \mathcal{N}(\mu_{\hat{p}}, SE_{\hat{p}}^2) = \mathcal{N}\left(p, \frac{p(1-p)}{n}\right) \tag{18}$$

where

$$\text{Sample mean proportion: } \mu_{\hat{p}} = p$$

$$\text{Sampling error/standard error of } \hat{p}\text{: } SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

**Temporary page!**

LaTeX was unable to guess the total number of pages correctly. As the
unprocessed data that should have been added to the final page this e
has been added to receive it.
If you rerun the document (without altering it) this surplus page will g
because LaTeX now knows how many pages to expect for this documen