Central Limit Theorem

- Central limit theorem
- Confidence Intervals

} Agenda

Central limit theorem :

- one of the most powerful & useful ideas in all of STATISTICS.

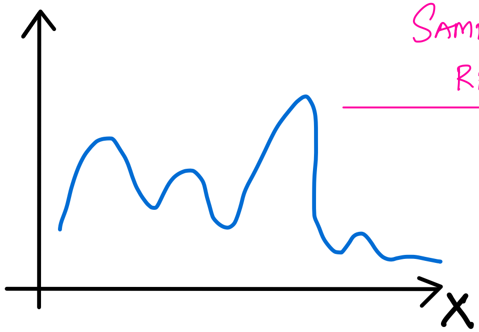
- Inference about population from samples.

(eg. population mean).

Core idea:

Original Pop. db^n

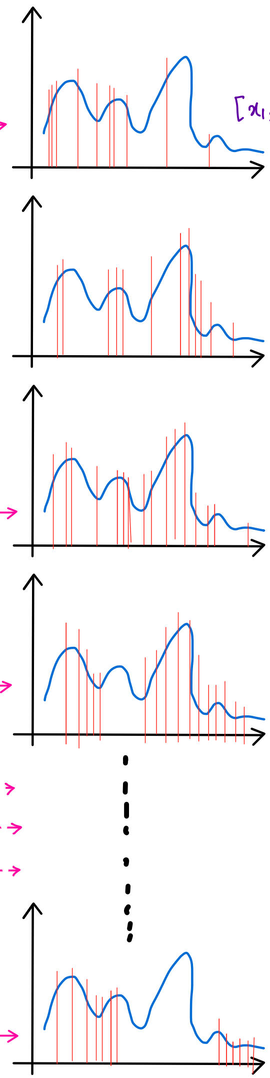
μ, σ



• Orignal population distribution does not matter

SAMPLING WITH REPLACEMENT

$n \to$ Sample size



$S_1 \to \bar{x}_1$ $[x_1, x_2, x_3, \ldots, x_n]$

$S_2 \to \bar{x}_2$

$S_3 \to \bar{x}_3$

$S_4 \to \bar{x}_4$

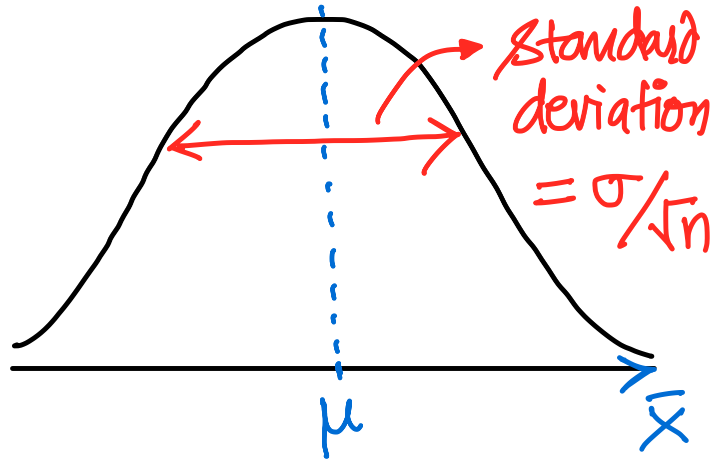
$S_{100} \to \bar{x}_{100}$

Sampling with replacement.

Sample means.

$[\bar{x}_1, \bar{x}_2, \ldots\ldots, \bar{x}_{100}]$

Sampling distribution of means



$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Sample mean $\bar{X}$ is R.V

$X \rightarrow$ Random Variable { Known/Unknown distribution }

- Mean of $X = \mu$
- standard deviation of $X = \sigma$

If we draw random samples of size 'n'

then as 'n' increases,

then Random Variable $\bar{X}$ which consists of sample means, tends to be Normally distributed :

$$\boxed{\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)}$$

μ : mean of $\bar{X}$

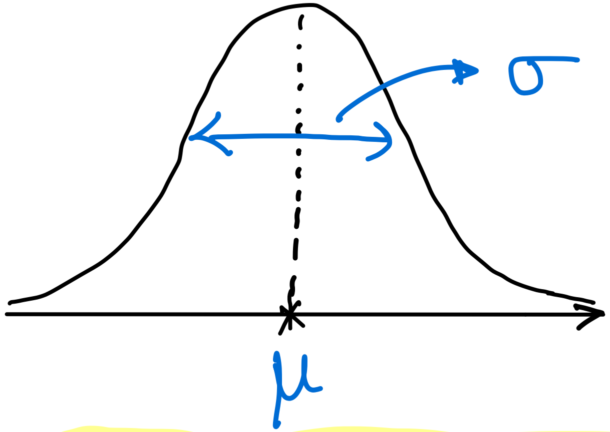
$\frac{\sigma}{\sqrt{n}}$: standard deviation of $\bar{X}$

also known as

Standard Error

Original distribution of X.

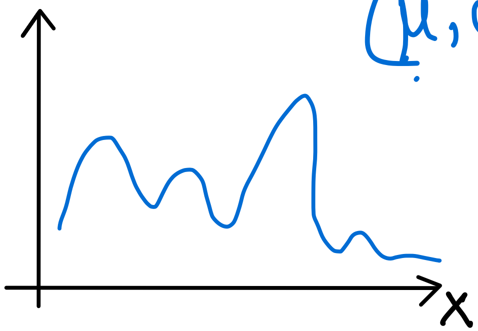
★ $X \sim N(\mu, \sigma)$



→ Sample size 'n' does not matter

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

★ $X \sim$ Any random distribution



(μ, σ) → As sample size $n \uparrow$, usually when $n > 30$

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

The amount of impurity in a batch of a chemical product is a random variable with mean value 4.0 g and standard deviation 1.5 g. (unknown distribution)

A) If 50 batches are independently prepared, what is the probability that the average amount of impurity in these 50 batches is between 3.5 and 3.8 g?

B) If we sample 100 independent batches, what should be the 95% Confidence Interval for average amount of impurity?

The amount of impurity in a batch of a chemical product is a random variable with mean value 4.0 g and standard deviation 1.5 g. (unknown distribution)

A) If 50 batches are independently prepared, what is the probability that the average amount of impurity in these 50 batches is between 3.5 and 3.8 g?

B) If we sample 100 independent batches, what should be the 95% Confidence Interval for average amount of impurity?

Sol^n: $\mu = 4.0\,g$, $\sigma = 1.5\,g$ } Population mean & standard deviation

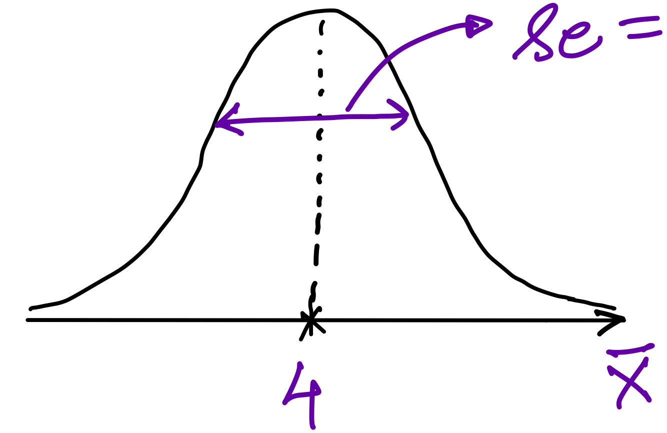
A) $n = 50$, $\bar{X}$ = sample mean → $\bar{x} = \dfrac{x_1 + x_2 + \cdots + x_{50}}{50}$

using CLT, $\bar{X} \sim \text{Normal}\left(\mu, \dfrac{\sigma}{\sqrt{n}}\right)$

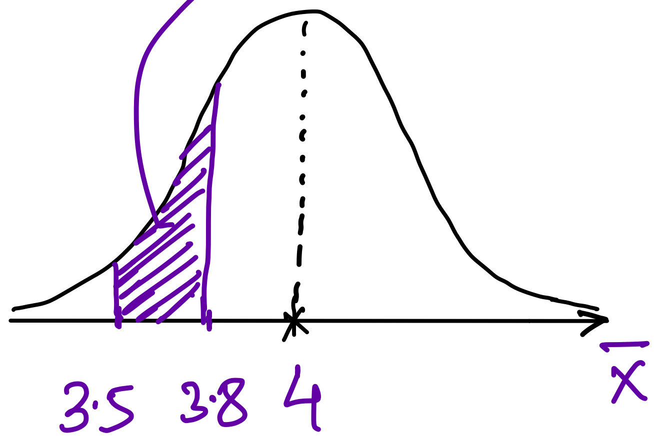
$\bar{X} \sim \text{Normal}\left(4, \dfrac{1.5}{\sqrt{50}}\right)$

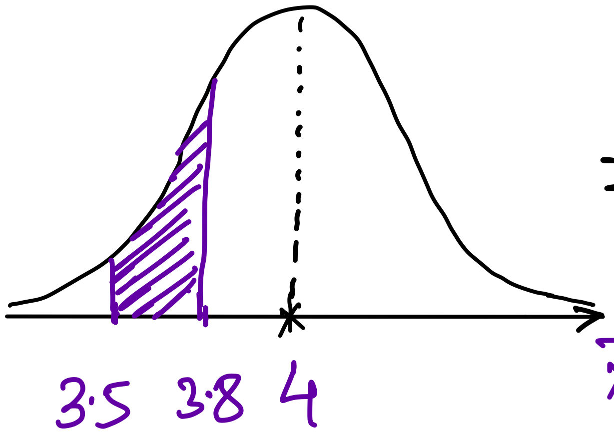
$std \cdot dev = se$

$se = \frac{1.5}{\sqrt{50}}$

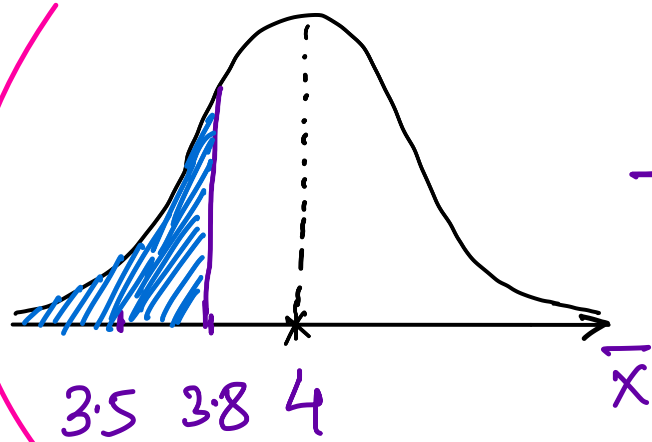


$P(3.5 < \bar{x} < 3.8) = ?$

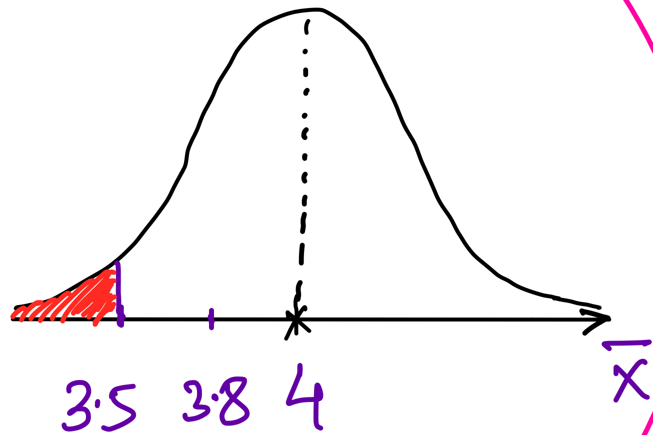




$=$ (



$-$



)

$P(3.5 < \bar{x} < 3.8) \quad P(\bar{x} < 3.8) \quad P(\bar{x} < 3.5)$

$(A \quad - B)$

* A : z score of 3.8 $\Rightarrow$ $z_{3.8} = \frac{3.8-4}{(1.5/\sqrt{50})} = -0.94$

$\therefore P(\bar{X} < 3.8) = \text{norm.cdf}(Z_{3.8})$

$= \text{norm.cdf}(-0.94)$

* **B:**

Z score of 3.5 $\Rightarrow Z_{3.5} = \dfrac{3.5-4}{(1.5/\sqrt{50})} = -2.36$

$P(\bar{X} < 3.5) = \text{norm.cdf}(Z_{3.5})$

$= \text{norm.cdf}(-2.36)$

$\Rightarrow P(3.5 < \bar{X} < 3.8) = \text{norm.cdf}(-0.94) - \text{norm.cdf}(-2.36)$

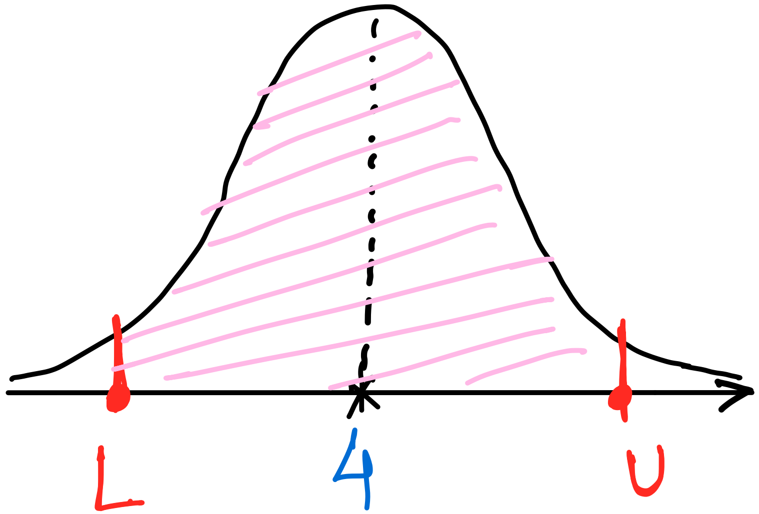
$= 0.16$

$\simeq 16\%$

B) If we sample 100 independent batches, what should be the 95% Confidence Interval for average amount of impurity?

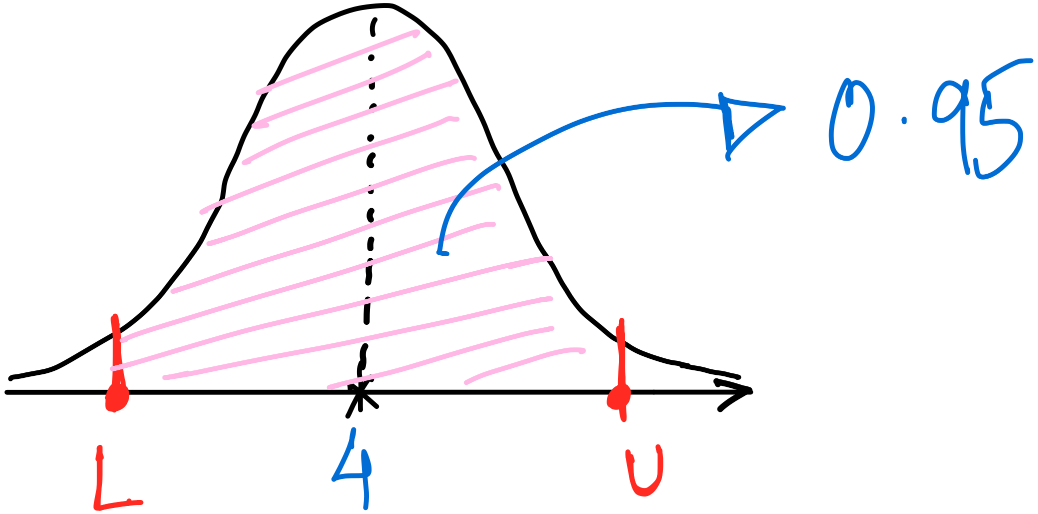
$n = 100$

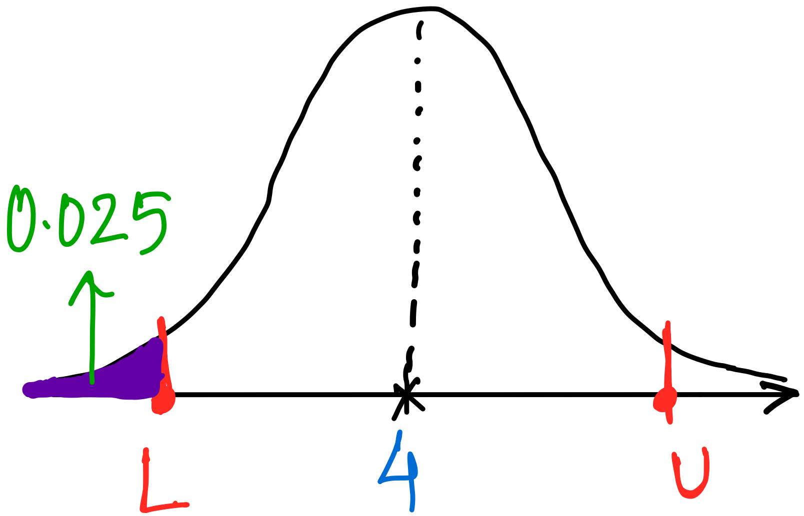
$\bar{X} \sim N\left(4, \frac{2.5}{\sqrt{100}}\right)$

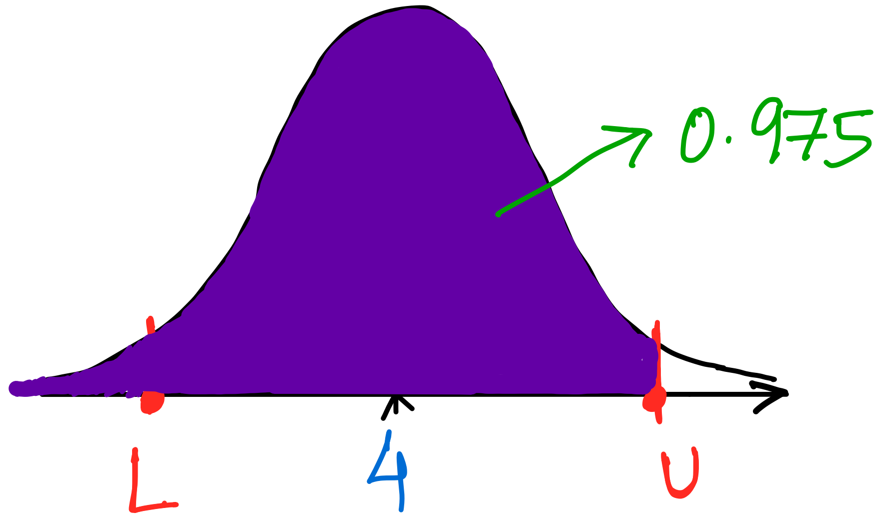


for 95% CI, the area b/w lower & upper limit should be $= \underline{0.95}$

95% CI $= [L, U]$







$z_L =$ norm.ppf(0.025)

$z_L = -1.96$

$z_U =$ norm.ppf(0.975)

$z_U = +1.96$

L : lower limit

↳ $Z_L = -1.96$; i.e. L is 1.96 std. dev. away f^m the mean – towards left.

$$Z = \left(\frac{x - \text{mean}}{\text{std. dev}}\right)$$

$$\therefore \quad -1.96 = \frac{L - 4}{\left(2.5/\sqrt{100}\right)} \Rightarrow L = 4 - 1.96 \times \frac{2.5}{\sqrt{100}}$$

$$\Rightarrow \boxed{L = 3.51}$$

Similarly;

U : Upper limit

↳ $Z_U = +1.96$; i.e. L is 1.96 std. dev. away f^m the mean – towards Right

$$\therefore \quad +1.96 = \frac{U-4}{(2.5/\sqrt{100})} \Rightarrow U = 4 + 1.96 \times \frac{2.5}{\sqrt{100}}$$

$$\Rightarrow \boxed{U = 4.49}$$

∴ The interval which captures true mean (= 4.0 gm) with 95% Confidence = [L, U] = [3.51, 4.49]