Agenda

* ** Central Limit theorem

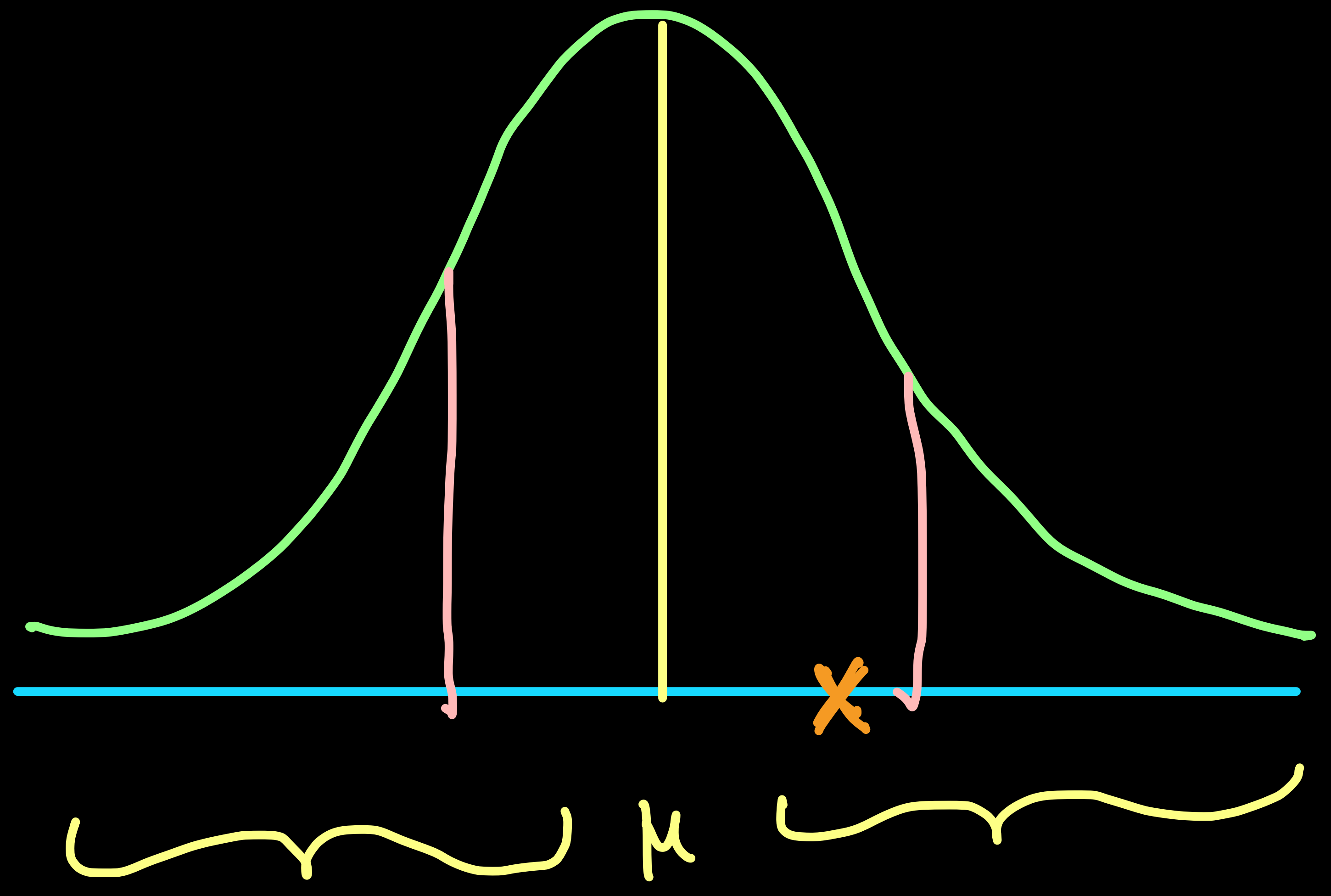
* Application of CLT on real life dataset → Height - weight

* Confidence Intervals
  - using CLT
  - using Bootstraping Assignments

Recap

* Normal | Gaussian distribution

* Empirical rule



$$
\begin{cases}
\mu - \sigma <= x <= \mu + \sigma \rightarrow \boxed{0.68} \\
\mu - 2\sigma <= x <= \mu + 2\sigma \rightarrow 0.95 \\
\mu - 3\sigma <= x <= \mu + 3\sigma \rightarrow 0.997
\end{cases}
$$

$$\boxed{z = \frac{x - \mu}{\sigma}}$$

* ✓✓ z-score

z-table

python → scipy

standard Normal Distribution / z- distribution

The special case, when Normal Distribution has

$\boxed{\mu = 0}$

$\boxed{\sigma = 1}$

standau

✓ Example

Imagine you have two friends, ✓
Alex and Taylor, who are both great at math but have different grading systems. Alex's math scores range from 0 to 100, while Taylors scores range from 0 to 50. You want to know who is performing with more consistency

$$z = \frac{x - \mu}{\sigma}$$

Central Limit Theorem

Q) what is the avg weight of the JB Basket?

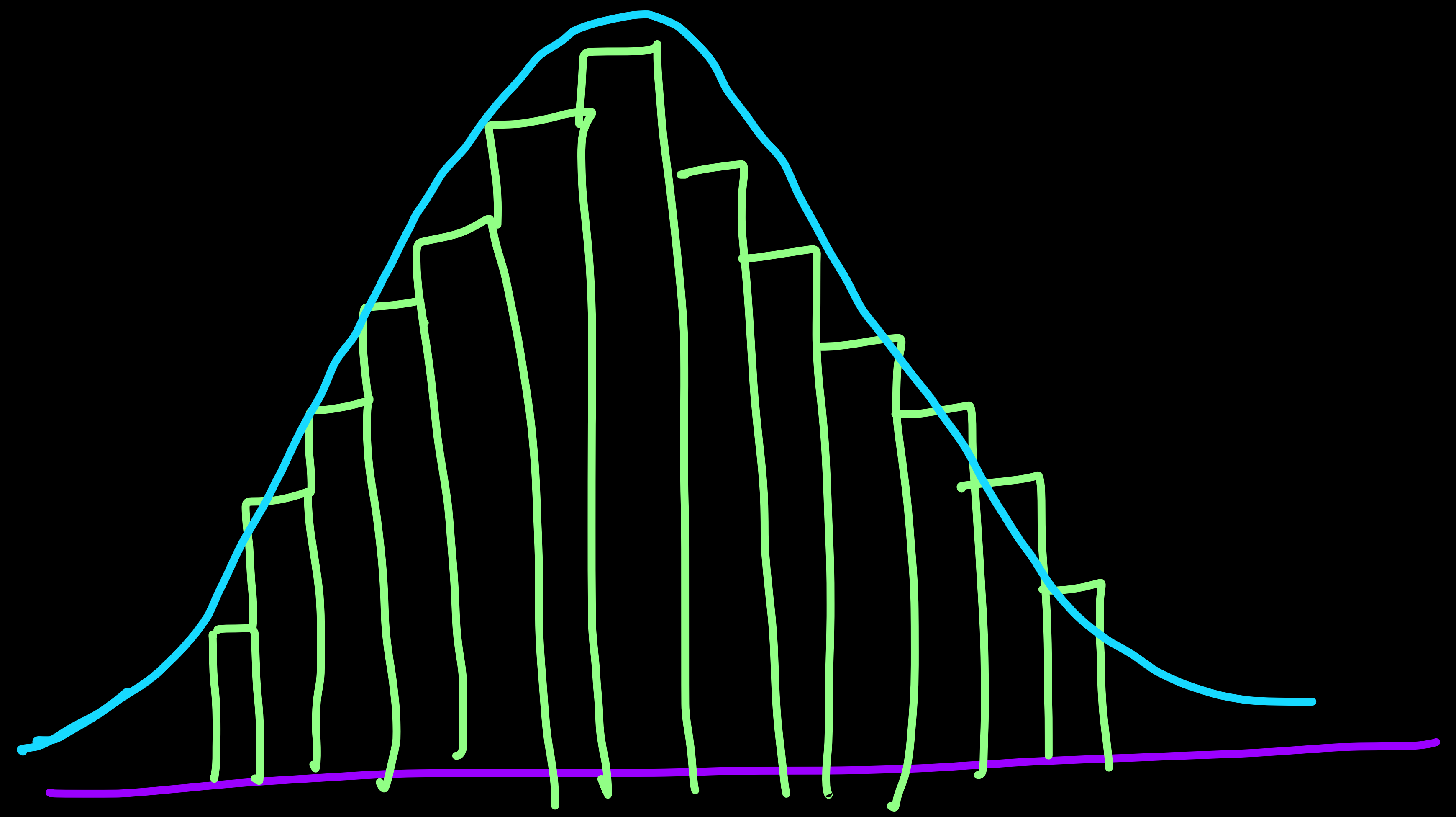
$S_1 = [w_1, w_2, \text{-----}, w_{30}] = \bar{x}_1$

$S_2 = [w_1, w_2 \text{-----} w_{30}] = \bar{x}_2$

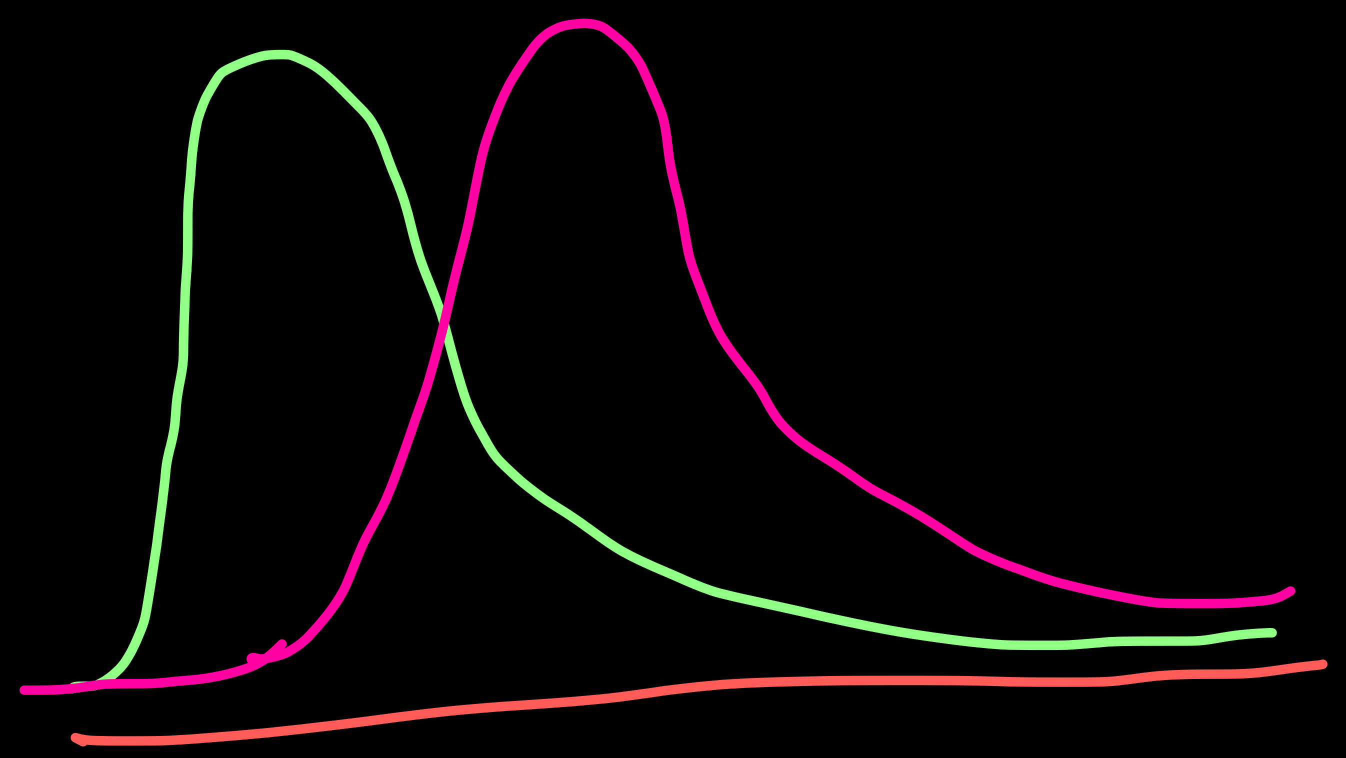
$\vdots$

$S_{1000} = [w_1\ w_2 \text{-----} w_{30}] = \bar{x}_{1000}$

mean of sample means $\rightarrow \boxed{\dfrac{\bar{x}_1 + \bar{x}_2 + \text{----} \bar{x}_{1000}}{1000}} \simeq$ population mean



$\Rightarrow$ { Sample means always follows the Normal Distribution



Central Limit Theorem

CLT states that " the mean of a random sample will resemble even closer to the population mean as the $\boxed{\text{sample size}}$ increases and it will approximate a normal distribution regardless of the shape of the population distribution

mean

mean of Sample means $\simeq$ population mean

$n = 4$

Q) What is the right sample size to take?

jelly beans samples taken

4 times → trails ← 40 times

$n=12$ → sample size ← $n=40$

→ ✓ Larger sample size → A sampling distribution that closely resembles a normal distribution

CLT tends to work well when n is sufficiently large. Typically consider as $\boxed{n \geq 30}$ ✓

sample size = 5 ✓✓

avg

$\boxed{3} > \underline{66.} \Rightarrow \underline{66.3}$ > 66

3 < 66.

$\mu = 66.36$

$\sigma = 3.84$

population

10000

Sample size = 5

mean of sample means = $\boxed{66.35}$

$s_1 = 1.735$

Sample size = 20

mean of sa means x_{20} = 66.36

$s_2 = 0.86$

$\sigma / \sqrt{n}$

Sample size n = 100

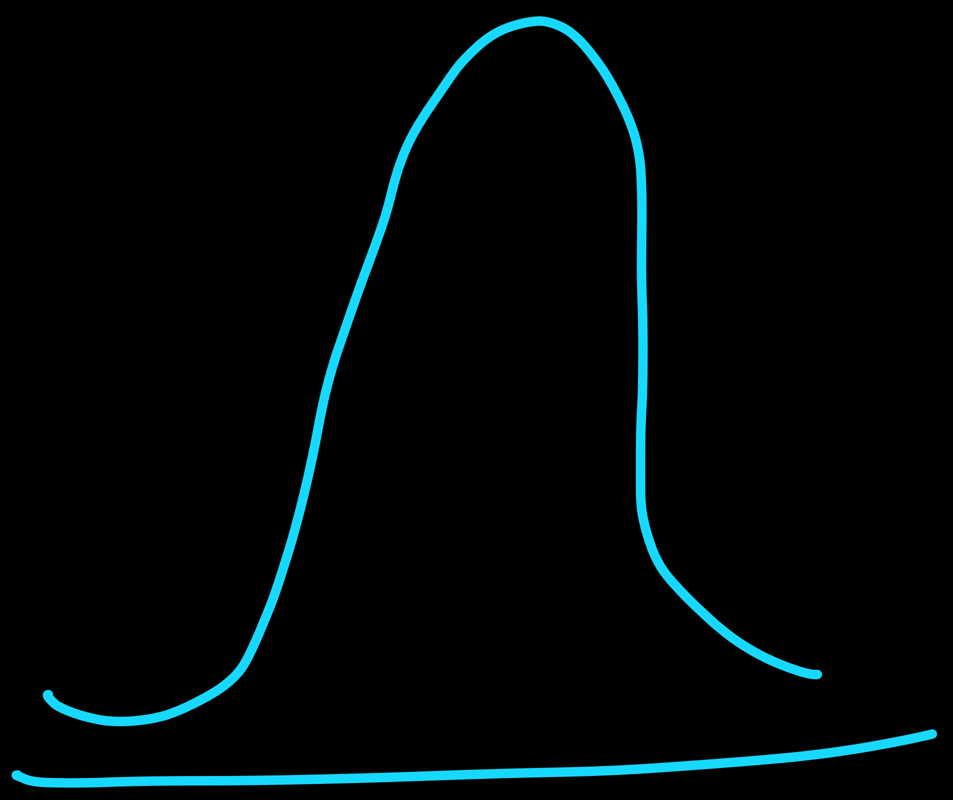
mean of sample mean = 66 36

$s_3 = 0.38$

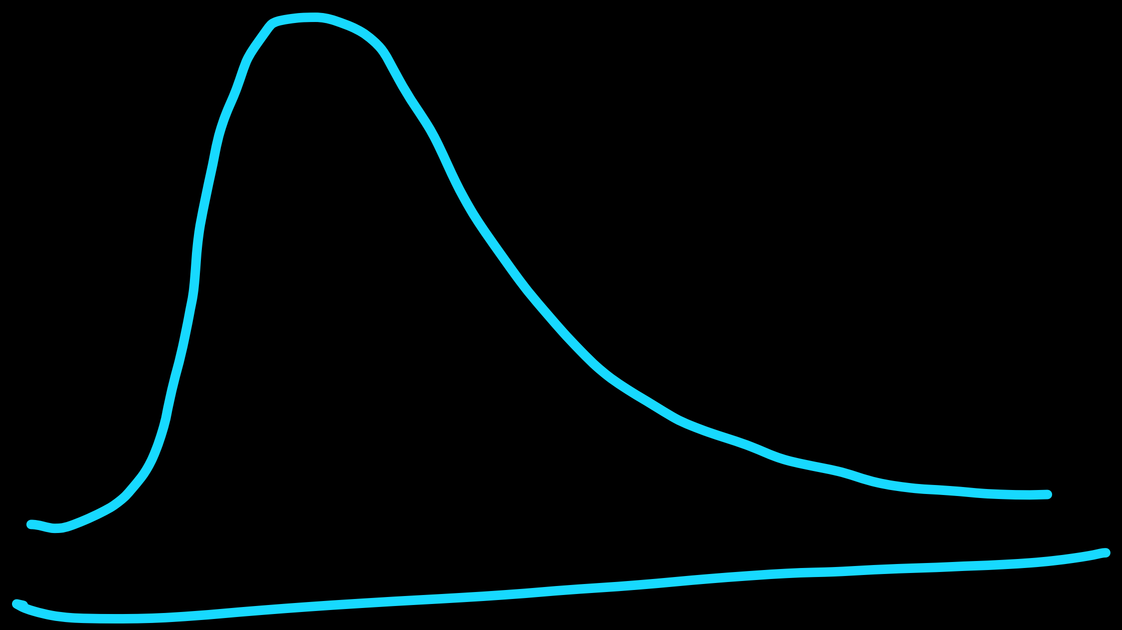
(1) mean of sample mean $\simeq$ population mean

(2) ✓ $\sigma / \sqrt{n} \Rightarrow \frac{3.84}{\sqrt{100}} = \frac{3.84}{10} = 0.384$

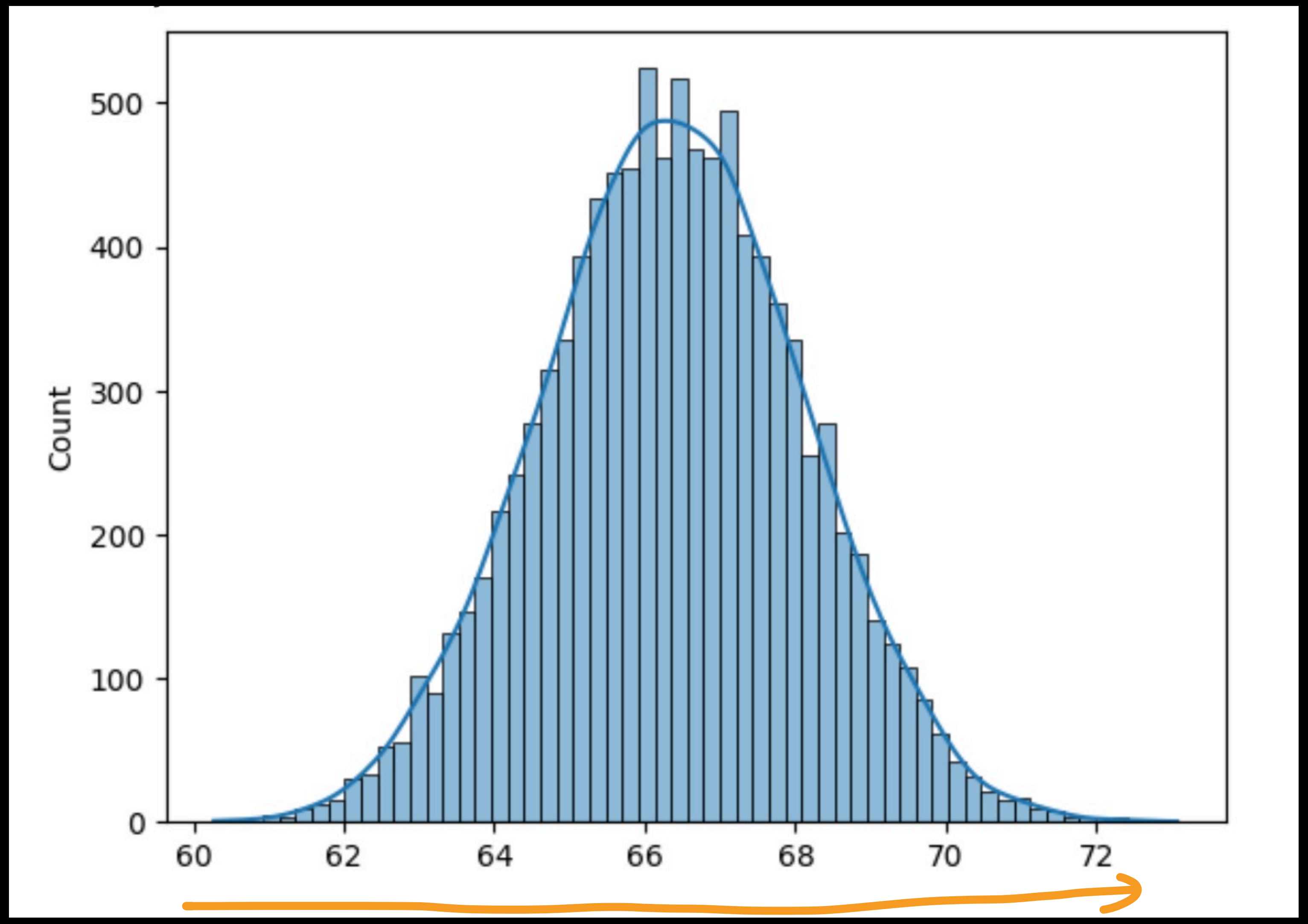
$\sigma / \sqrt{n}$ → SD of sample means

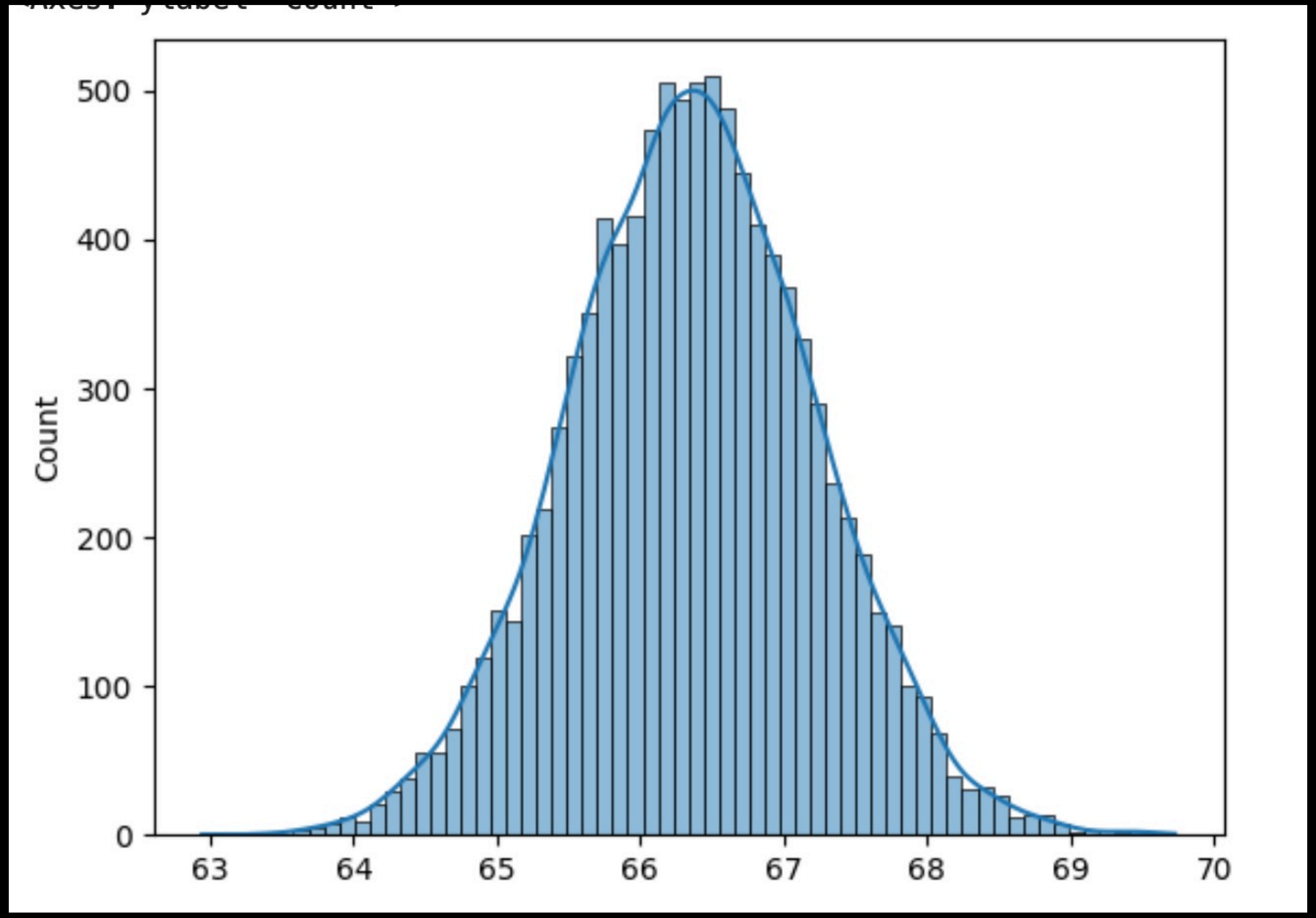


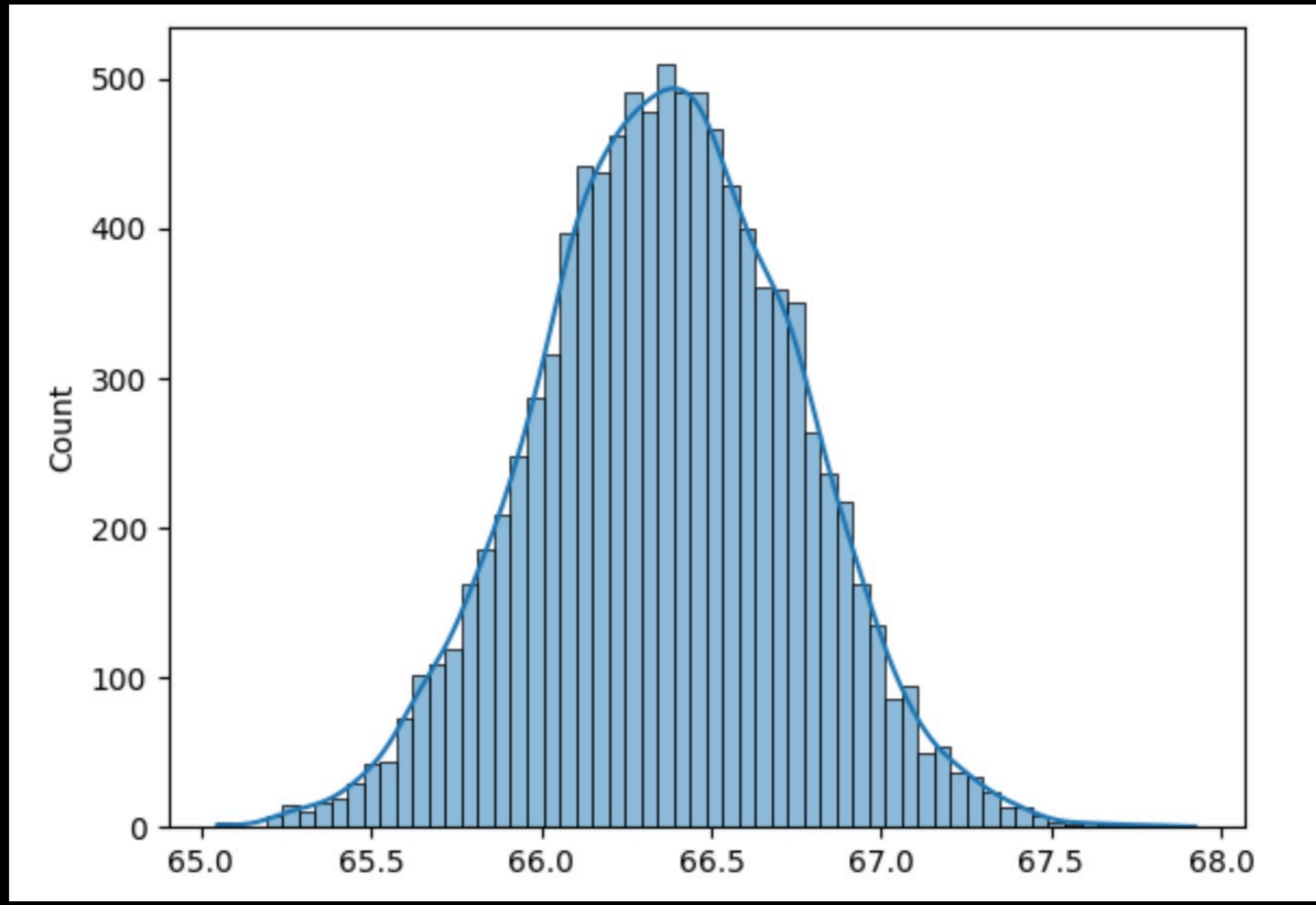
$\Rightarrow \begin{cases} n = 5 \\ n = 20 \end{cases} \rightarrow$



$\boxed{n \geq \boxed{30}}$ ✓







Sample-5

20

100

$\sigma / \sqrt{n}$

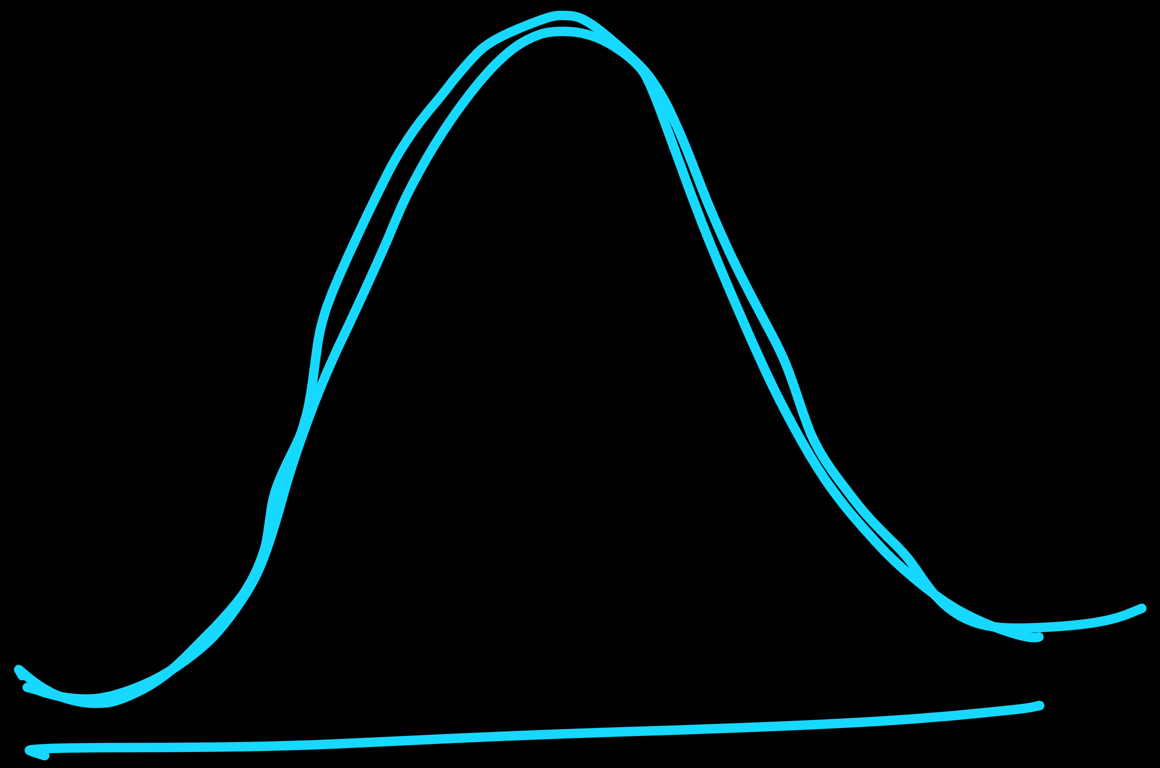
$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Follows Normal Distribution

atleast 30

larger sample

more accurate



$\Rightarrow$

$n < 30$

$n \geq 30$

Conditions of the CLT

To apply the central limit theorem, the following conditions must be met:

1. Randomization: Data should be randomly sampled, ensuring every population member has an equal chance of being included.

9 2 10000

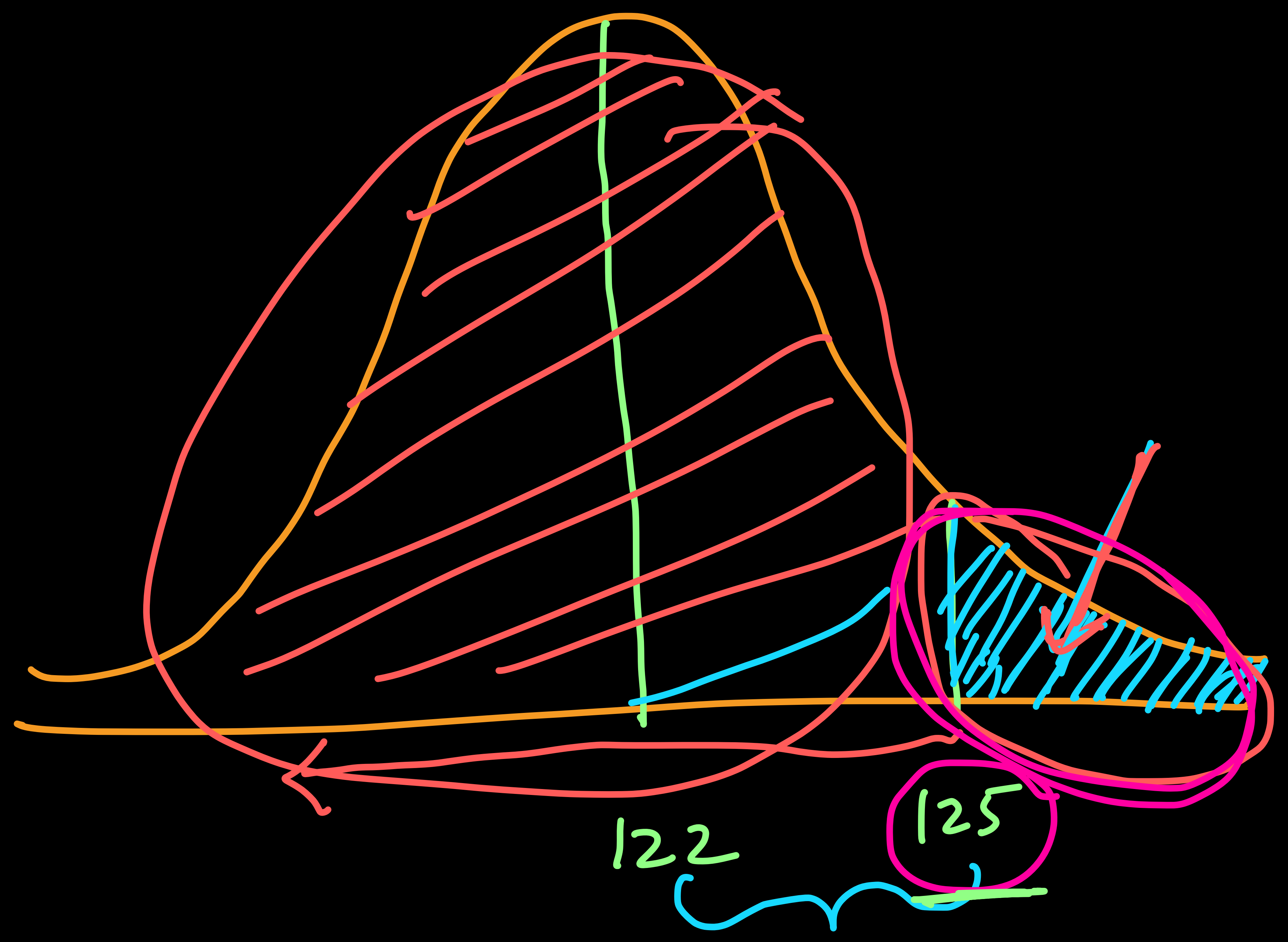
2. Independence: Each sample value should be independent, with one event's occurrence not affecting another. Commonly met in probability sampling methods, which independently select observations.

3. Large Sample Condition: A sample size of 30 or more is generally considered "sufficiently large." This threshold can vary slightly based on the population distribution's shape.

Problem

Systollic blood Pressure of a group of people is known to have an average of 122 mmHg and standard deviation of 10 mmHg Calculate the probability that the average blood pressure of 16 people will be greater than 125 mm Hg.

$$\begin{cases} \mu = 122 \text{ mm Hg} \\ \sigma = 10 \text{ mm Hg} \\ n = 16 \end{cases}$$



S.D of Sample $= \sigma / \sqrt{n}$

S. Error $= 10/4 = \boxed{2.5}$

$1 - \frac{x - \mu}{}$

$$\boxed{Z = \frac{125 - 122}{\sigma \; 2.5}} = \frac{3}{2.5} = 1.2$$

$1 - \text{norm.cdf}(1.2)$ ✓ $\Rightarrow 0.115$

$\boxed{8:08 \rightarrow 8:15}$

In an ecommerce website, the average purchase amount per customer is \$80 with a standard deviation of \$15.
If we randomly select a sample of 50 customers,
what is the probability that the average purchase amount in the sample will be less than \$75?

$\mu = \$80$

$\sigma = \$15$

$n = 50$

$$z = \frac{75 - 80}{\left(\frac{15}{\sqrt{50}}\right)}$$

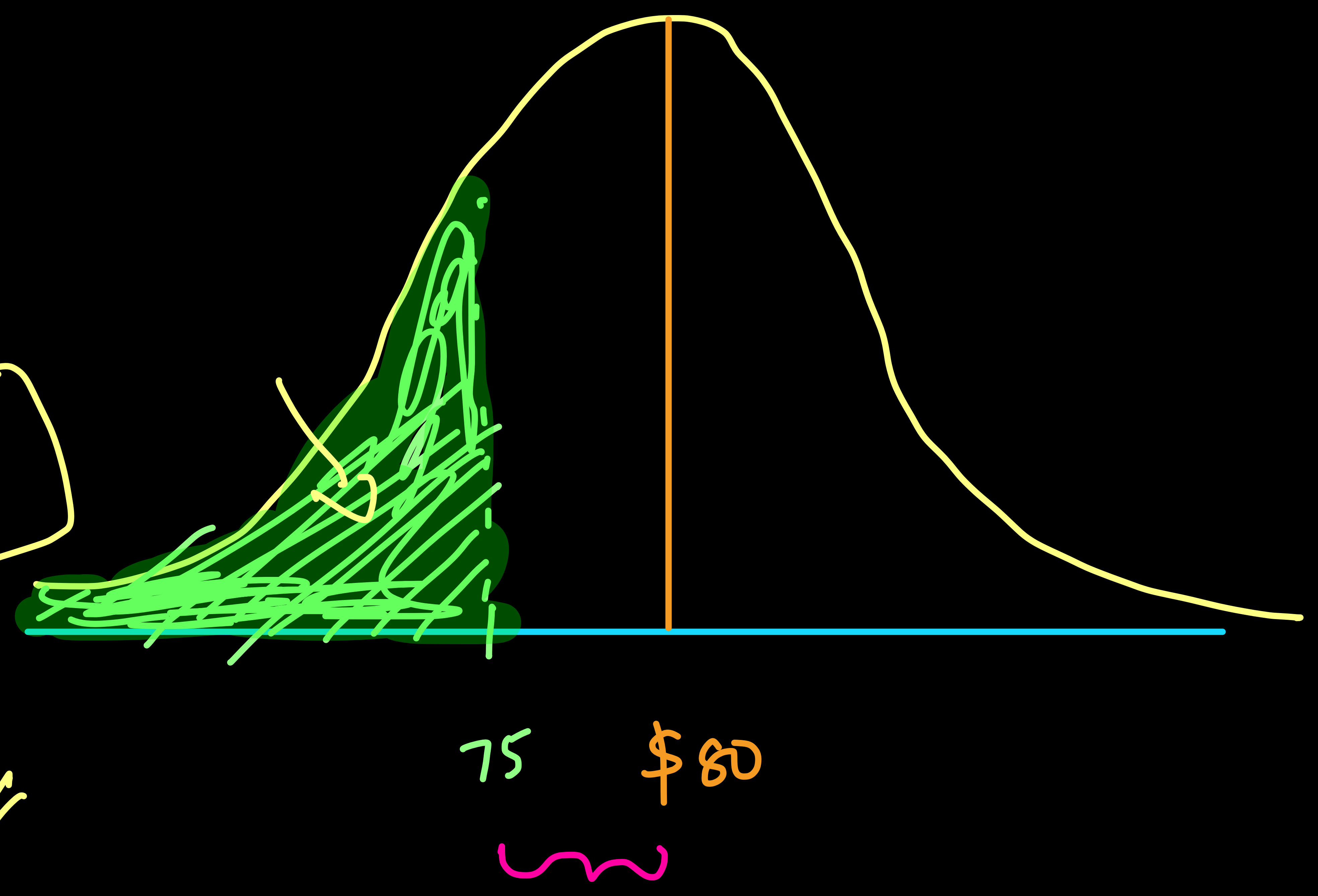
$$\downarrow$$

$$\boxed{z = -2.35}$$

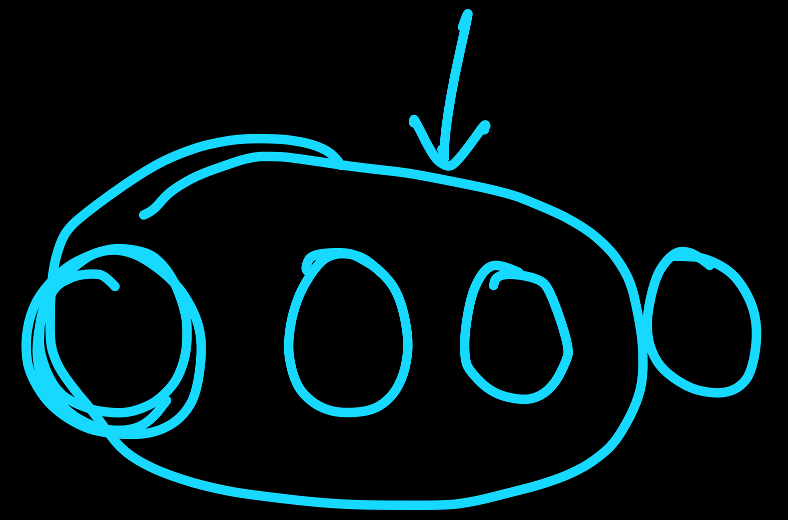
$$S.D = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{50}}$$

$=$

$P[x < 75] = 0.009 \leftarrow$ norm.cdf(-2.35)



Weekly toothpaste sales have a mean of 1000 and std dev of 200. What is the probability that the avg weekly sales next month is more than 1110?



$\mu = 1000$

population std dev $\leftarrow$ $\sigma = 200$

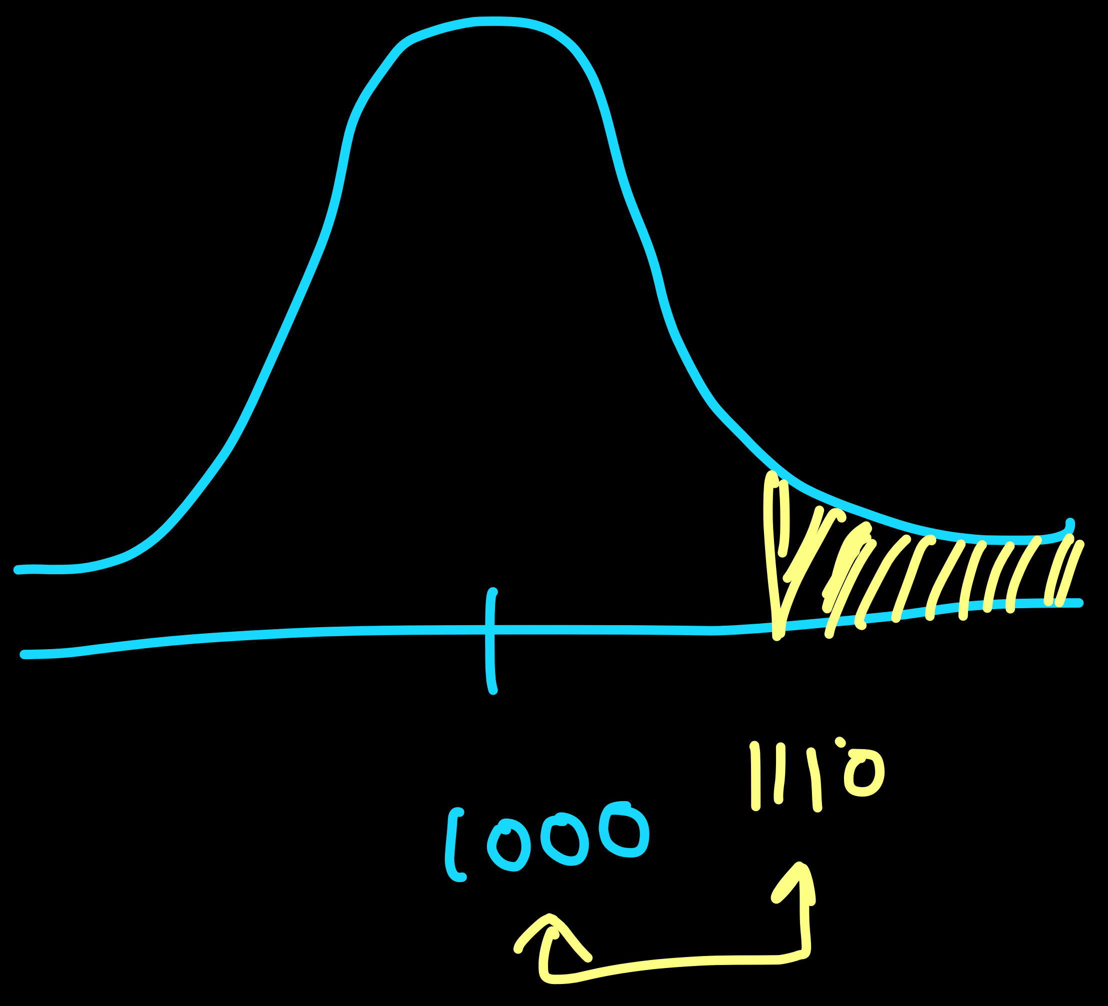
$n = 4$

$\sigma / \sqrt{n}$

S.D of Samp $= \dfrac{\sigma}{\sqrt{n}} = \dfrac{200}{2} = 100$

$$z = \frac{1110 - 1000}{100} = \frac{110}{100} = 1.1$$

$1 - \text{normcdf}(1.1)$



Confidence Interval

Exit polls

$P_1 \rightarrow 30 \rightarrow 0 - 100$ ✓

$P_2 \rightarrow 0 - 100$

$P_3 \rightarrow 0 - 100$

Range → 25 - 35 → How do They decide the range

→ point estimate

sample mean [point estimate] ± ~~SD~~ some error

C I

10

9 11

95% C.I →

99% CI → [,]

95% [ ]

90% → [ ]

95% [ ]

[64 66]

$\boxed{95}$

Q How to calculate C.I ?

(1) Using CLT [central Limit theorem]

(2) Using Bootstrapping

Example Height

Mean height of sample of 100 adults = 65 inches

Sample mean

→ [population S.D = 2.5

S.D of sample → $\begin{cases} n = 100 \\ \sigma = 2.5 \end{cases}$

$\bar{x} = 65$

$\sigma / \sqrt{n}$

S.D of sample $= \dfrac{\sigma}{\sqrt{n}}$

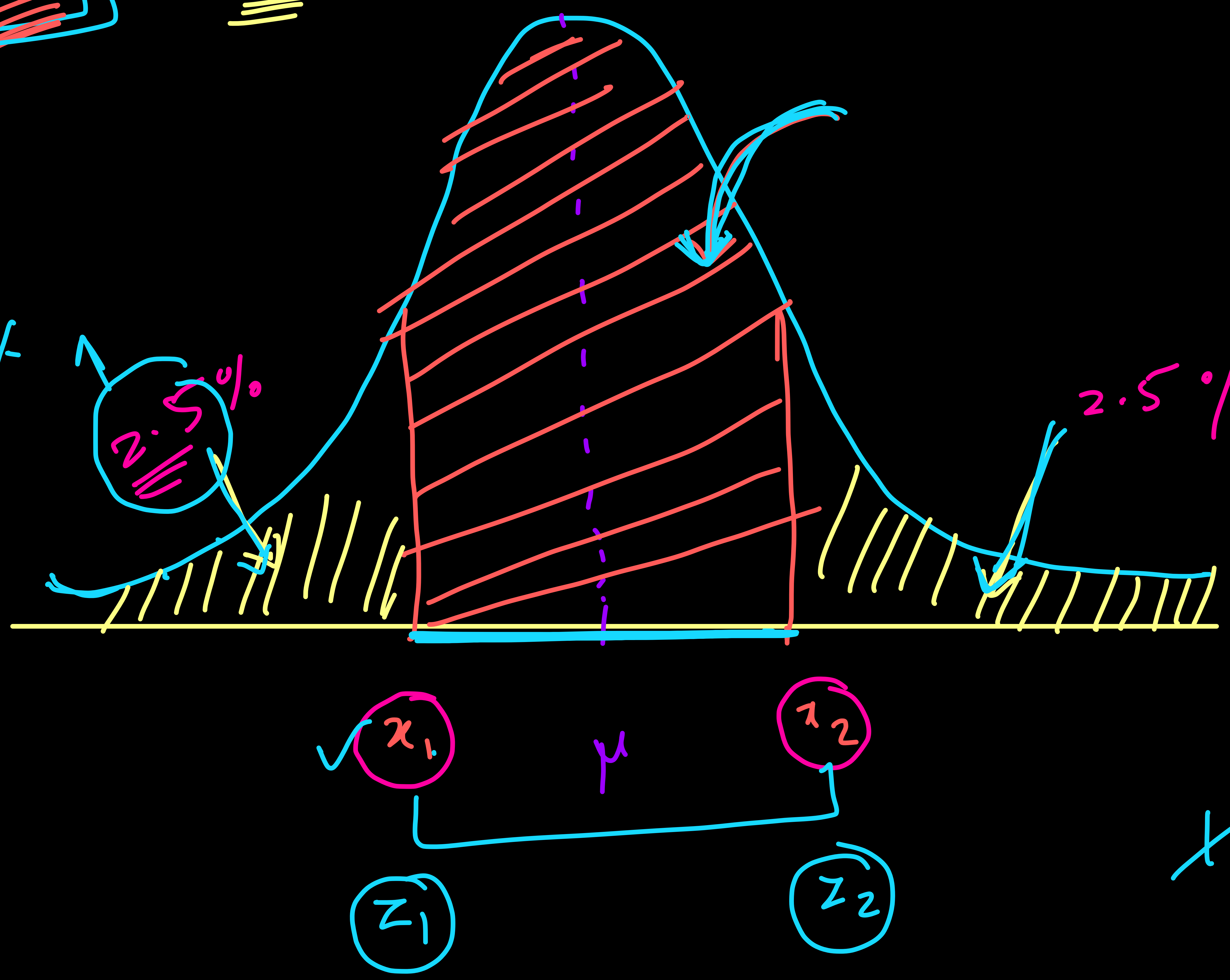
$= \dfrac{2.5}{10} = \boxed{0.25}$

95%

C I

99%

0.5 %



2.5% 0.5%

$z_1 =$ norm ppf(0.025)

~~norm.ppf(~~

norm. ppf(1 − 0.025)

$$S.D = \frac{25}{\sqrt{100}}$$

$$= \boxed{0.25}$$

$$x_1 = 65 + (-1.95) * 0.25$$

$=$

$=$

$$\Rightarrow [64.51 \qquad 65.48]$$

$$C.I = \bar{x} \pm z * \left(\frac{\sigma}{\sqrt{n}}\right)$$

$$\left[\bar{x} - \boxed{z} * \frac{\sigma}{\sqrt{n}} \;,\; \bar{x} + z * \frac{\sigma}{\sqrt{n}}\right]$$

$\bar{x} \rightarrow$ sample mean

$z \rightarrow$

$\sigma/\sqrt{n} \rightarrow$ Standard Error

The sample mean recovery time of 100 patients after taking a drug was seen to be 10.5 days with a σ standard deviation of $\boxed{2}$ days
Find the 95% confidence interval of the true mean.

population mean

$n = 100$
$\bar{x} = 10.5$
$\sigma = 2$

$$S.E = \frac{\sigma}{\sqrt{n}} = \frac{2}{10} = 0.2$$

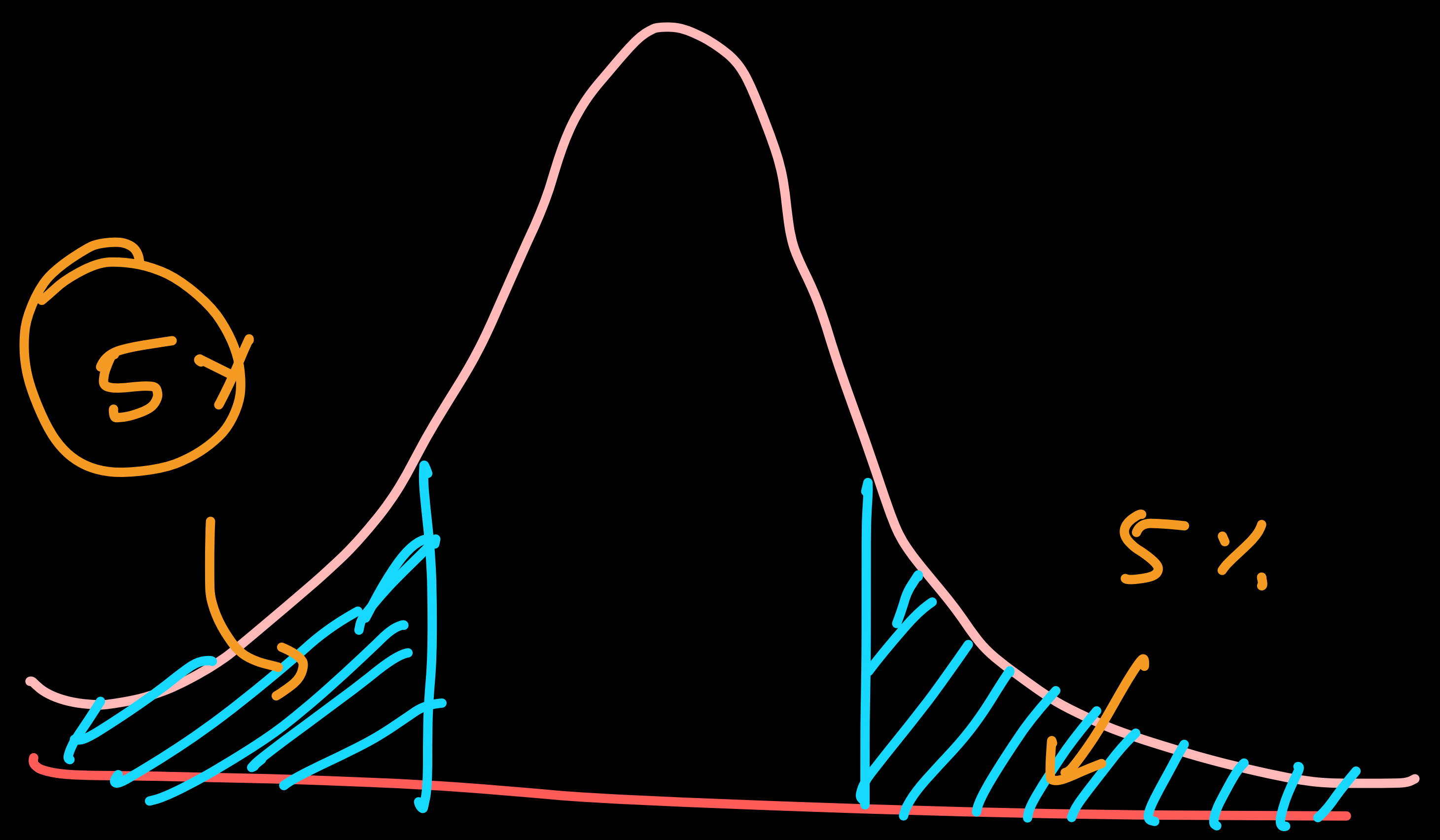
From a sample of 80 endangered birds,
the average wingspan was found to be 45 cm, with a population standard deviation of 10 cm.
What is the correct confidence interval of the mean wingspan of the entire population with 90% confidence.

$n = 80$

$\bar{x} = 45 \text{ cm}$

$\sigma = 10 \text{ cm}$

$\boxed{90\%}$



$S.E = \sigma/\sqrt{n} = \dfrac{10}{\sqrt{80}}$

In a software project, the team estimates bug resolution time at an average of 6 hours with a standard deviation of 2 hours.
To estimate the mean resolution time with 99% confidence, the project manager samples 25 resolved bugs.
What is the correct confidence interval?

$z_1 = \text{norm ppf}(0.005)$

$z_2 = \text{norm.ppf}(1 - 0.005)$

$n = 25$

99%

$\bar{x} = 6$

$x_1 = 6 + \boxed{z_1} * 0.4$

$\sigma = 2$

$x_2 = 6 + z_2 * 0.4$

$$S.E = \frac{2}{\sqrt{25}} = \frac{2}{5} = \boxed{0.4}$$

CLT → 2-3

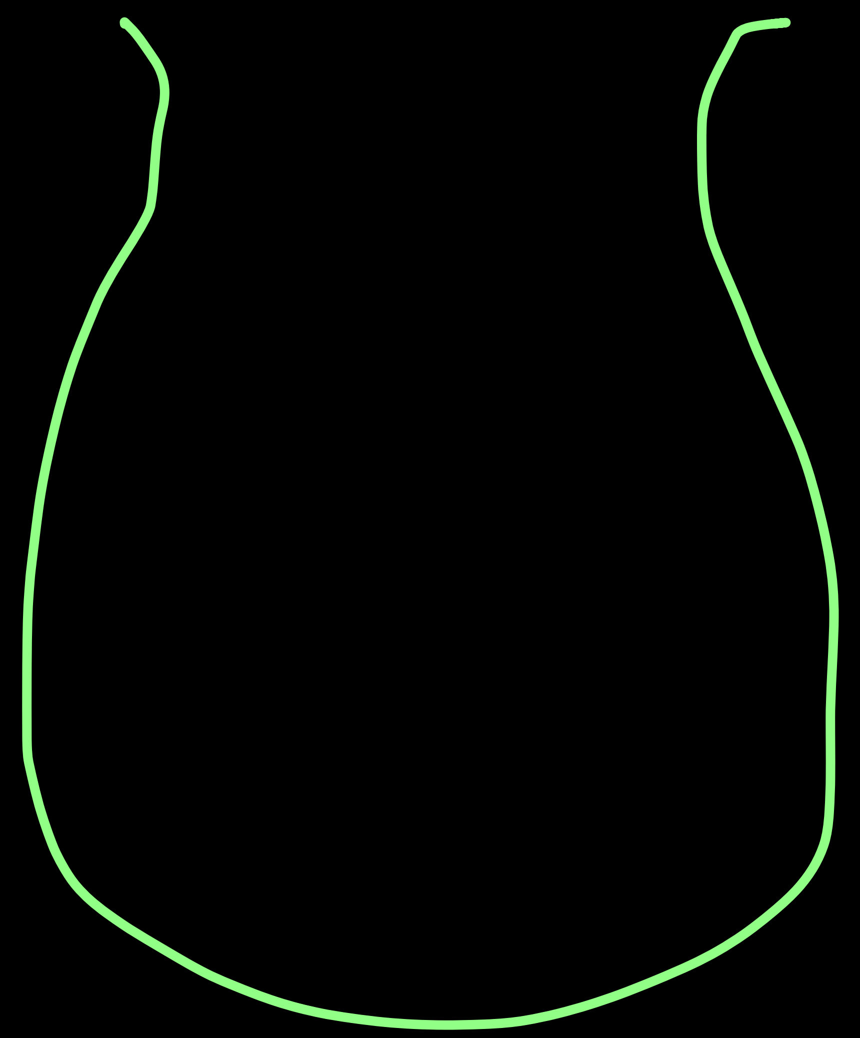
CI → 2-3

Normal → 2-3

$\boxed{\text{CLT}}$

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

$\boxed{\text{C.I}}$ → using CLT ✓



10 time

$$\left\{\begin{array}{l} x_1 = [\quad] \Rightarrow \boxed{\bar{x}_1} \\ x_2 = [\quad] = \bar{x}_2 \\ \vdots \\ x_{100} = \bar{x}_{100} \end{array}\right.$$

Sample means ~ Normal distribution

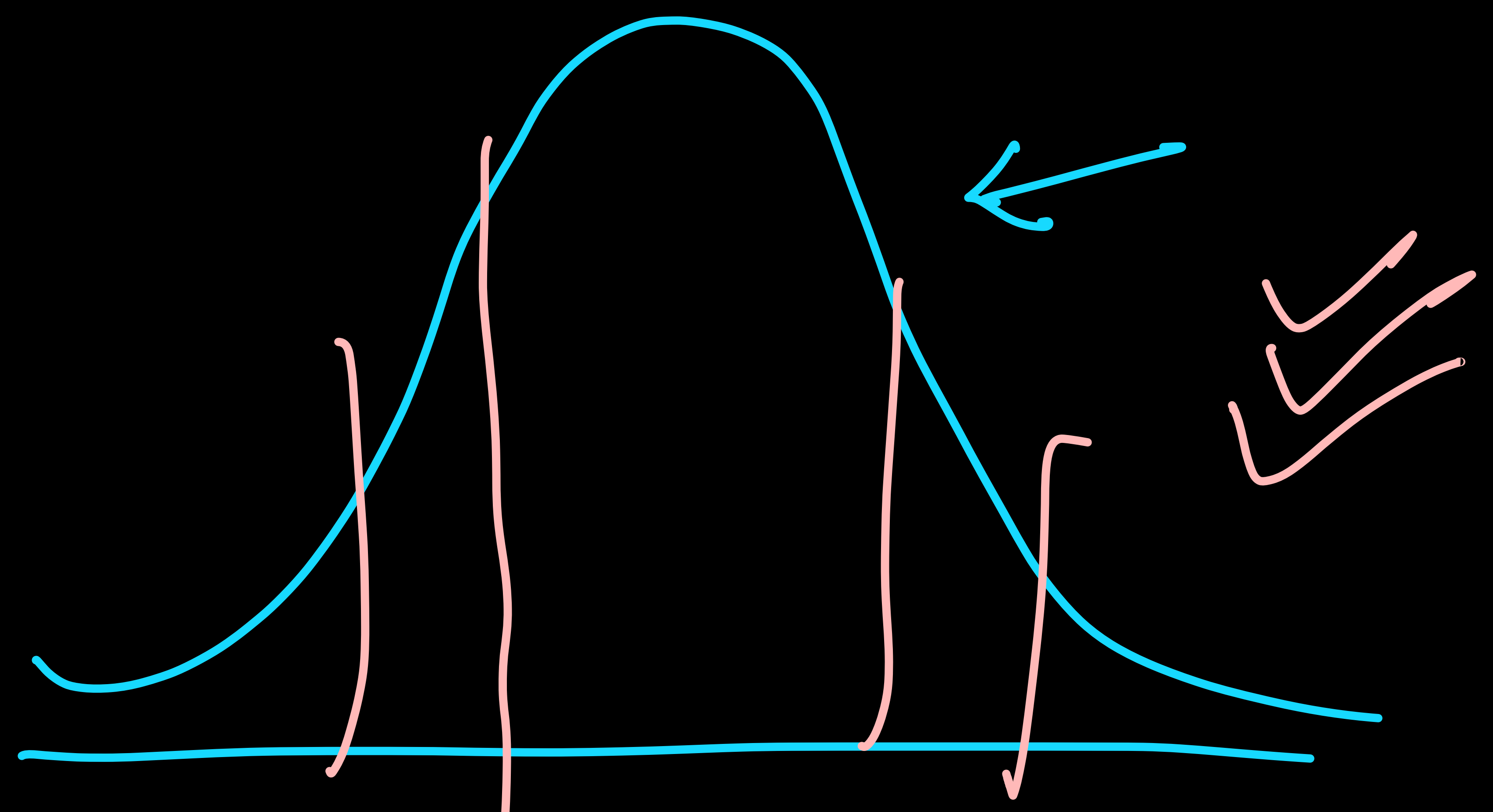
→ Sample

$n = 30$ → $\boxed{100}$

$n =$ $\quad$ $\sigma =$

$\sigma \Rightarrow S.D = \sigma/\sqrt{n}$

$\Rightarrow$ S.D of Sample



$\boxed{\sigma} =$

$\Rightarrow \sigma/\sqrt{n}$

$$\frac{3.84}{\sqrt{100}} \rightarrow 0.384$$