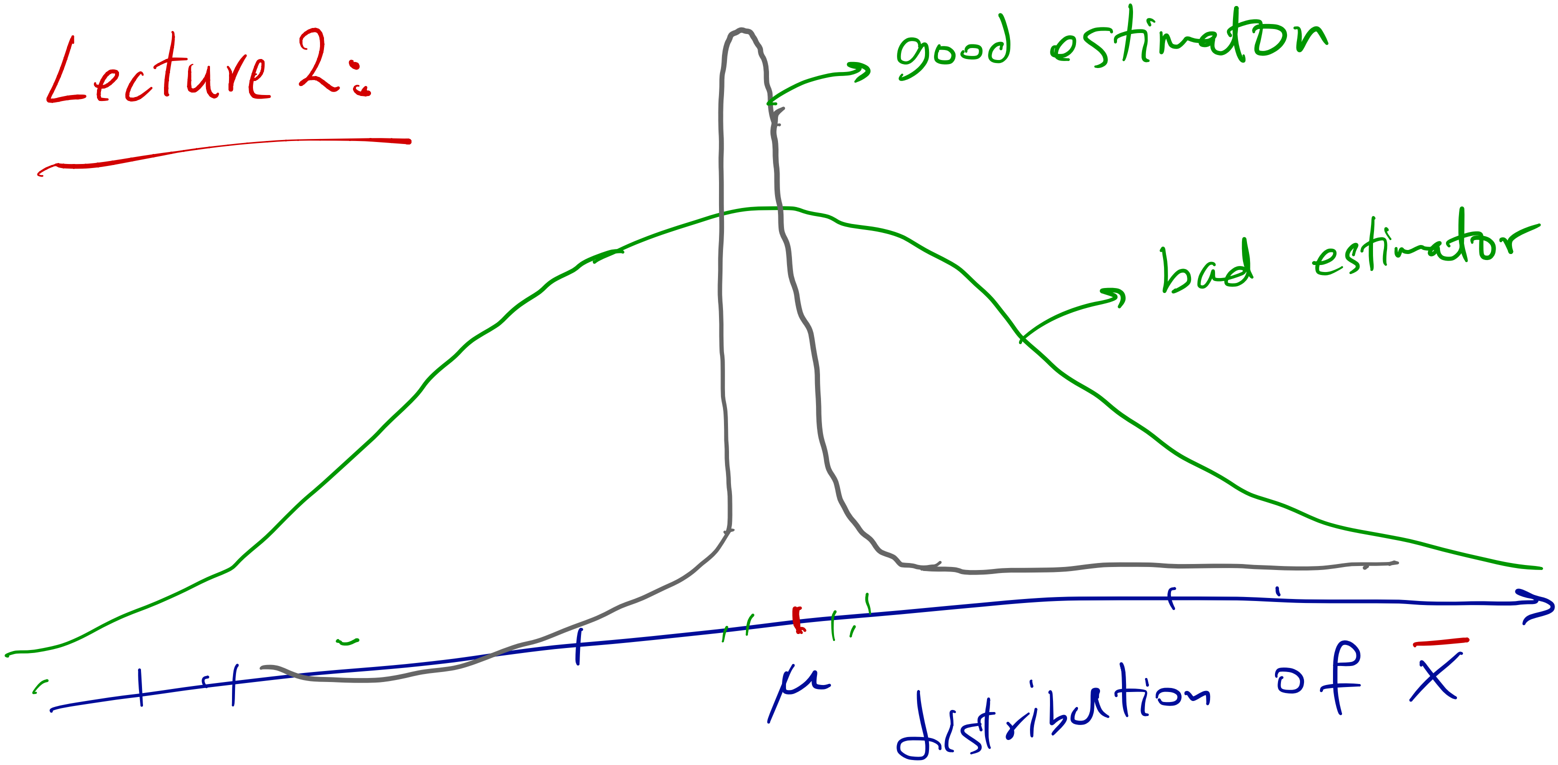
Lecture 2:



In order to see how concentrated $\bar{X}$ is around μ, we need to compute ...

$$E\left[(\bar{X}-\mu)^2\right].$$

$$E[(\bar{X}-\mu)^2] \qquad \left(\bar{X} = \frac{X_1+\cdots+X_n}{n}\right)$$

$$E\left[\left(\frac{X_1+X_2+\cdots+X_n}{n}-\mu\right)^2\right]$$

$$= E\left[\left(\frac{X_1+X_2+\cdots+X_n-n\mu}{n}\right)^2\right]$$

$$= E\left[\left(\frac{X_1-\mu+X_2-\mu+\cdots+X_n-\mu}{n}\right)^2\right]$$

$$= \frac{1}{n^2} E\left[\left((X_1-\mu)+(X_2-\mu)+\cdots+(X_n-\mu)\right)^2\right]$$

$$= \cdots$$

$$\left(\quad - \quad - \quad \right)^2 = \left(\sum_{i=1}^{n} (x_i - \mu) \right)^2$$

$$= \frac{1}{n^2} E\left[\left(\sum_{i=1}^{n} (x_i - \mu) \right)^2 \right]$$

$$= \frac{1}{n^2} E\left\{ \sum_{i=1}^{n} (x_i - \mu)^2 + 2 \sum_{\substack{i,j=1 \\ i \neq j}}^{n} (x_i - \mu)(x_j - \mu) \right\}$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} \underbrace{E\left[(x_i - \mu)^2 \right]}_{1} + \frac{2}{n^2} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \underbrace{E\left[(x_i - \mu)(x_j - \mu) \right]}_{2}$$

$$(\alpha_1 + \alpha_2 + \cdots + \alpha_n)^2 = \alpha_1^2 + \alpha_2^2 + \cdots + \alpha_n^2$$
$$+ 2\alpha_1\alpha_2 + 2\alpha_1\alpha_3 \cdots + 2\alpha_{n-1}\alpha_n$$

(2) $E[(X_i-\mu)(X_j-\mu)]$ $(i \neq j)$

when $i \neq j$, X_i and X_j are independent. Hence

$$E[(X_i-\mu)(X_j-\mu)]$$

$$\underset{\text{independence}}{=} \underbrace{E[(X_i-\mu)]}\,E[(X_j-\mu)]$$

$$= \underbrace{(E[X_i]-\mu)}_{=0}\underbrace{(E[X_j]-\mu)}_{=0}$$

$\Rightarrow$ (2) $= 0$

---

if two random variables U, V are independent, then $E[U \cdot V] = E[U]E[V]$

$$E[g(U)\cdot h(V)] = E[g(U)]\cdot E[h(V)]$$

$$① = \frac{1}{n^2} \sum_{i=1}^{n} E\left[(x_i - \mu)^2\right]$$

$$= \frac{1}{n^2} \left(E[(x_1-\mu)^2] + E[(x_2-\mu)^2] \cdots + E[(x_n-\mu)^2] \right)$$

$$= \frac{1}{n^2} \cdot n \, E\left[(X_1 - \mu)^2\right]$$

X_i's are identically distributed (uniform)

$$= \frac{1}{n} E\left[(X_1 - \mu)^2\right]$$

$$= \frac{1}{n} \left(\frac{1}{N}(y_1 - \mu)^2 + \frac{1}{N}(y_2 - \mu)^2 + \cdots + \frac{1}{N}(y_N - \mu)^2 \right)$$

$$= \frac{1}{n} \cdot \sigma^2_{population}$$

where $\sigma^2_{population} = \frac{1}{N}\left((y_1-\mu)^2 + \cdots + (y_N - \mu)^2 \right)$

Variance of $\bar{X}$:

$$E[(\bar{X}-\mu)^2] = 1 + 2$$

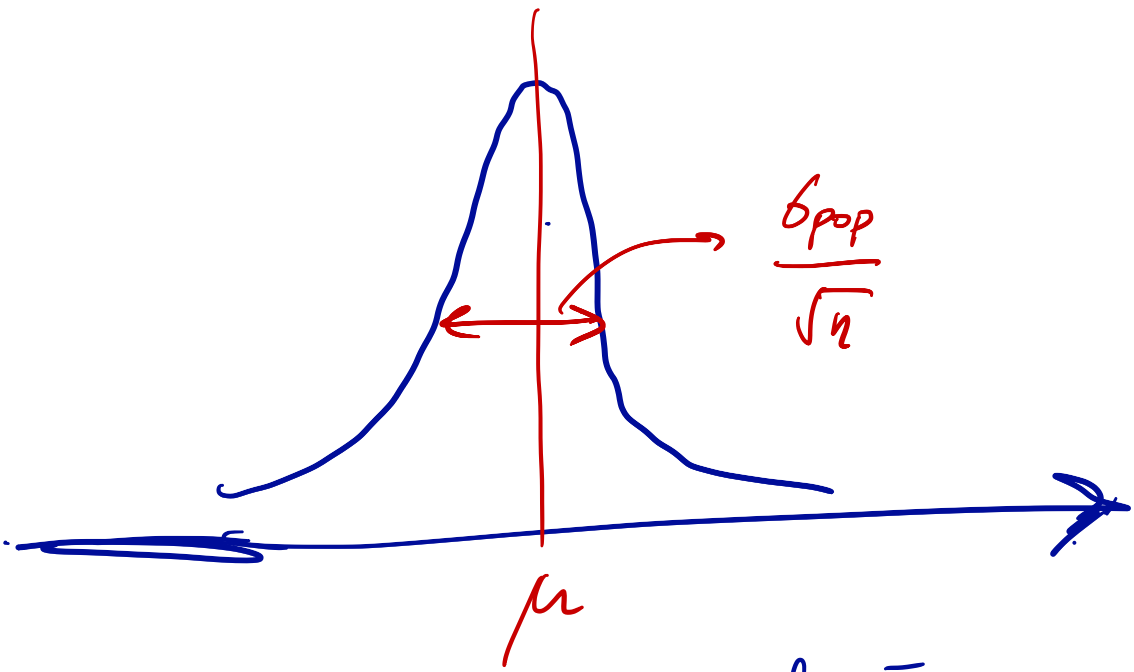
$$= \frac{1}{n} \sigma^2_{\text{population}} \longrightarrow \text{fixed number}$$

So the variance of $\bar{X}$ decays wrt the number of data poins, n, like $\frac{1}{n}$.

Standard deviation of $\bar{X}$:

$$\sqrt{E(\bar{X}-\mu)^2} = \frac{\sigma_{pop}}{\sqrt{n}}$$

Conclusion:



distribution of $\bar{x}$

real-world pheno $\rightarrow$ Data acquisition/Storage $\rightarrow$ $X_1, \dots, X_n$ are iid $\rightarrow$ Data Analysis $\rightarrow$ knowledge/pattern/insight

$$\{y_1, \dots, y_N\}$$

N = very large

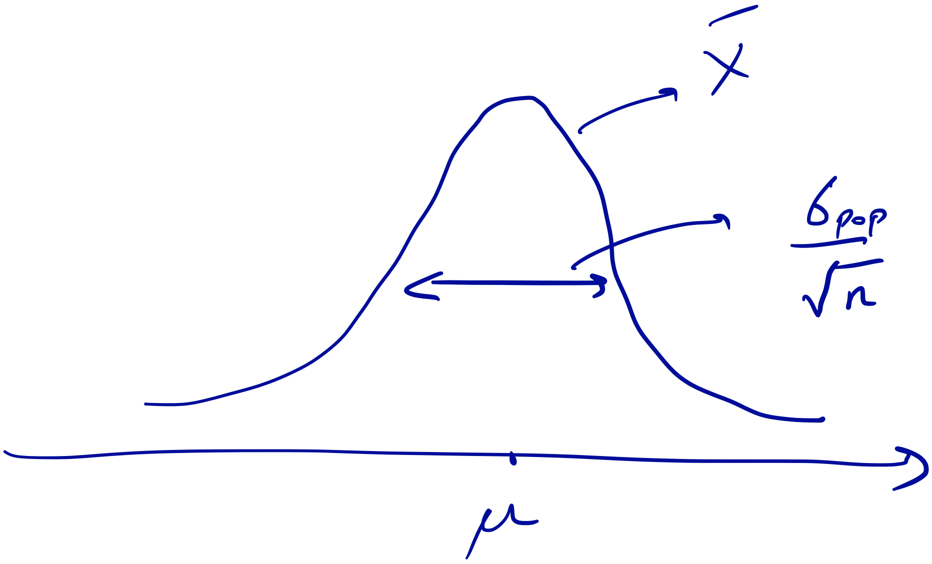
$$\mu = \frac{\sum_{i=1}^{N} y_i}{N}, \quad \sigma_{pop}^2 = \frac{\sum_{i=1}^{N} (y_i - \mu)^2}{N}$$

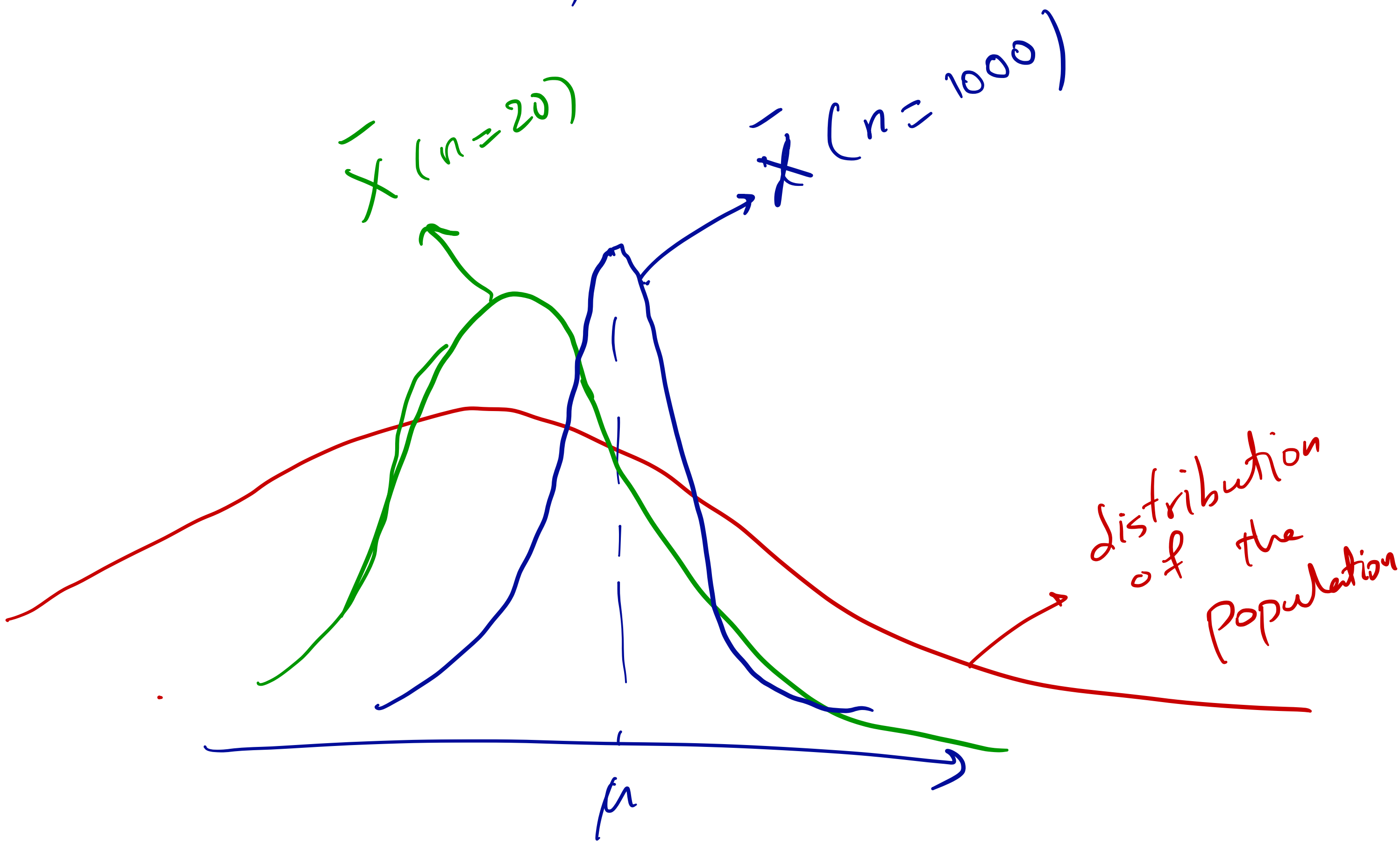
$$X_i = \begin{cases} y_1 & \frac{1}{N} \\ \vdots & \\ y_N & \frac{1}{N} \end{cases}$$

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

$$E[\bar{X}] = \mu$$

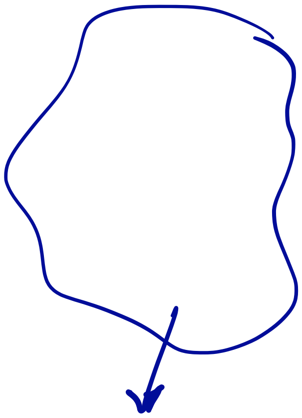
$$Var(\bar{X}) = \frac{\sigma_{pop}^2}{n}$$





Another example:

we'd like to find the precentage of people who would vote for a certain candidata in a state.



$\{y_1, \ldots, y_N\}$

$N = \#$ of people in the state

$$y_i = \begin{cases} 1 & \text{if } i \text{ would vote for the candidate} \\ 0 & \text{o.w.} \end{cases}$$

$$\mu = \frac{y_1 + \cdots + y_N}{N} \rightarrow \text{\% of people who'd vote for the candidate}$$

Let's summarize what we've learned so far. (and this is how we think of data acquisition / gathering procedure throughout the course).

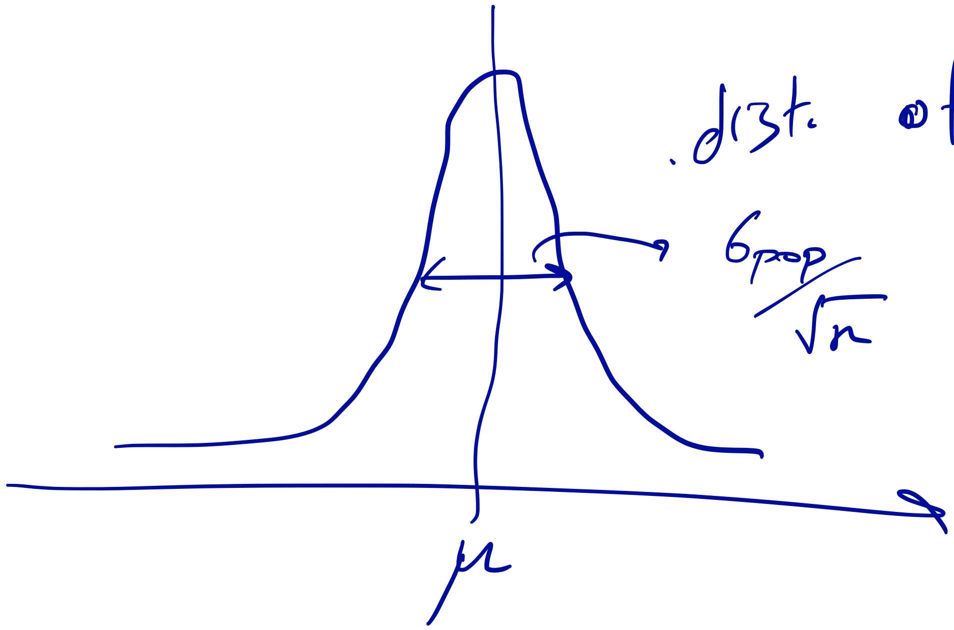
In order to learn about a physical/real-world phenomenon, we will gather sample data.

Random Sample: statistically, we model the sample data as $X_1, X_2, \ldots, X_n$ where X_i's are assumed to be independent and identically distributed (i.i.d.) as the samples are generated from the same phenomenon.

↳ X_i's are random observations

- A <u>statistic</u> is any quantity whose value can be computed from sample data. Prior to obtaining data, there is uncertainty as to what value any particular statistic will result. Therefore, a statistic is typically a random variable.

e.g. sample mean $\bar{X}$ is a statistic.



In order to understand more about the distribution of $\bar{X}$, we need to study the powerful framework of central limit theorem.

The Central Limit Theorem:

Take $X_1, \dots, X_n \overset{iid}{\sim} \text{dist}(\mu, \sigma^2)$

Then (informally):

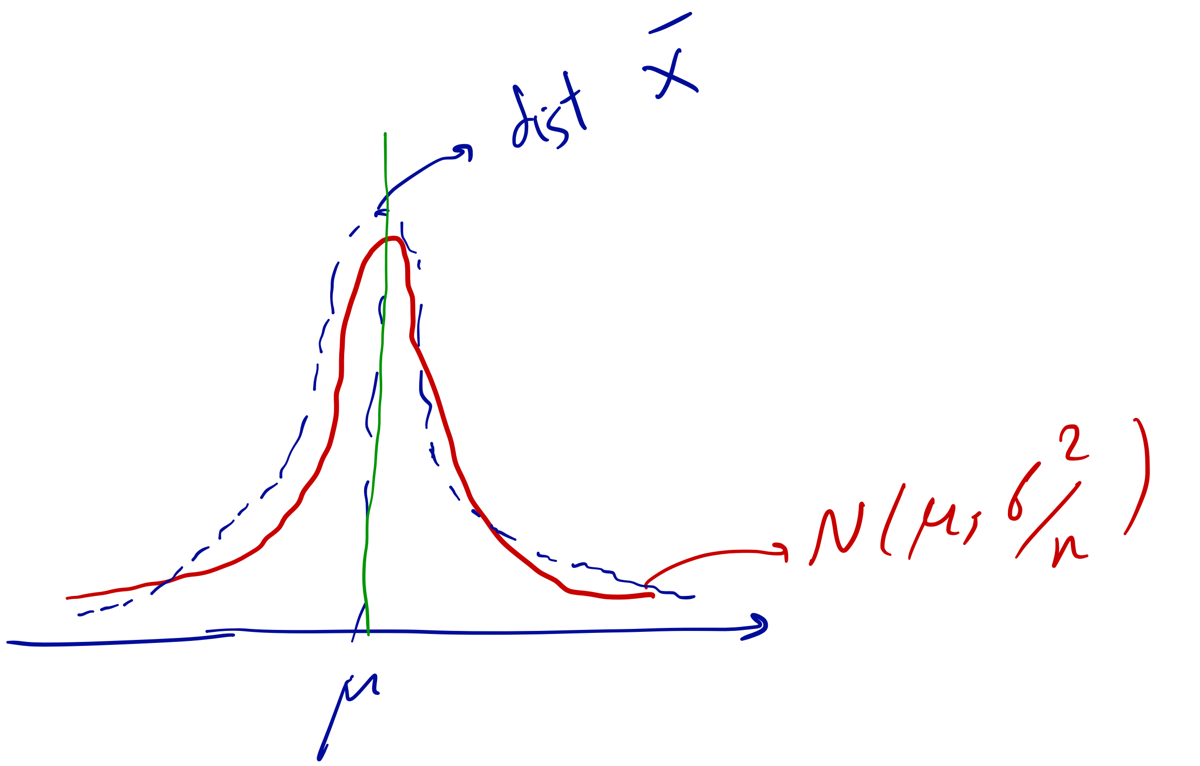
$$\bar{X} \sim \text{gaussian} + \text{small error}$$

we know that :

$$\begin{cases} E[\bar{X}] = \mu \\ Var(\bar{X}) = \sigma^2/n \end{cases}$$

$$\bar{X} = \underbrace{\text{gaussian}}_{\downarrow} + \underbrace{\text{small error}}_{\frac{c}{n}}$$

$$= N(\mu, \sigma^2/n) + \text{small error}$$



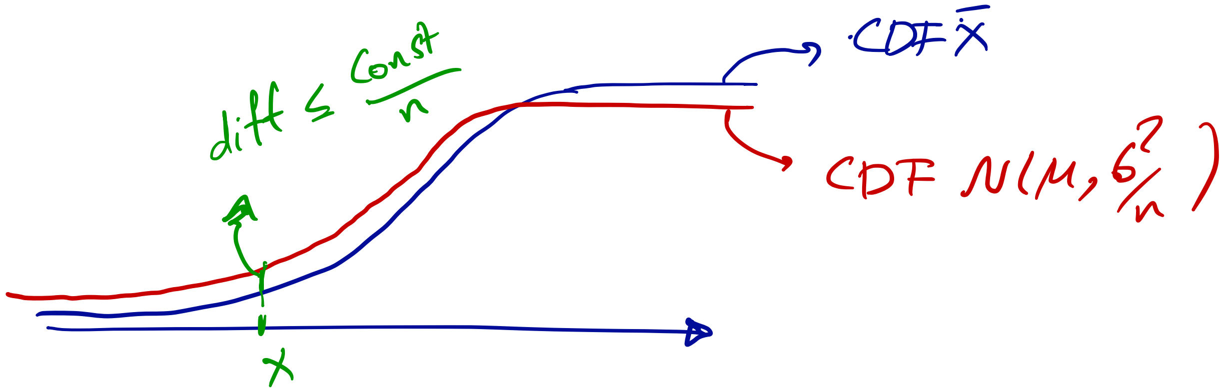
Formal Statement (CLT):

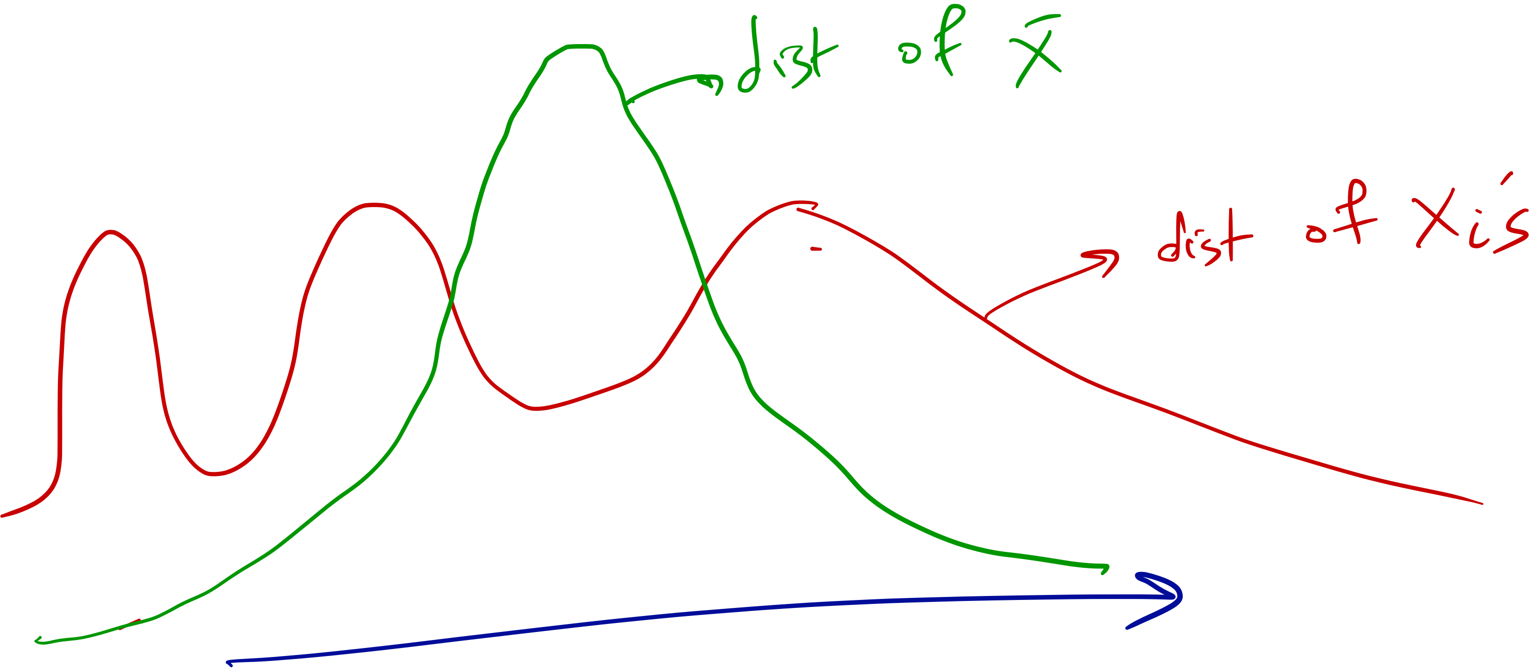
Let $F_{\bar{X}}(x)$ be the CDF of $\bar{X}$:

$$F_{\bar{X}}(x) = \Pr\{\bar{X} \le x\}$$

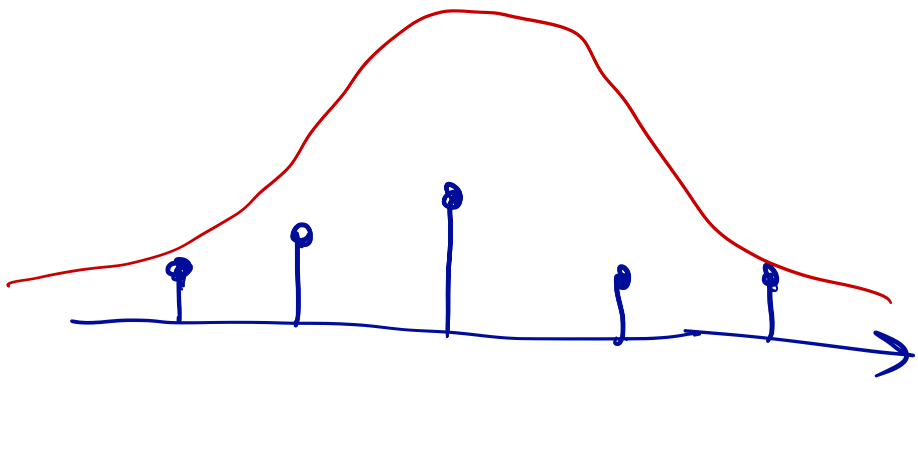
and Let $\phi_{\mu,\sigma}(x)$ be the CDF of $N(\mu, \sigma^2/n)$. Then:

$$\left| F_{\bar{X}}(x) - \phi_{\mu,\sigma}(x) \right| \le \frac{\underbrace{\text{Constant}}_{\text{depends on the dis. } X_i\text{'s}}}{n}$$





The theorem does not hold for Pdfs. (e.g. the dist of X_i could be discrete).



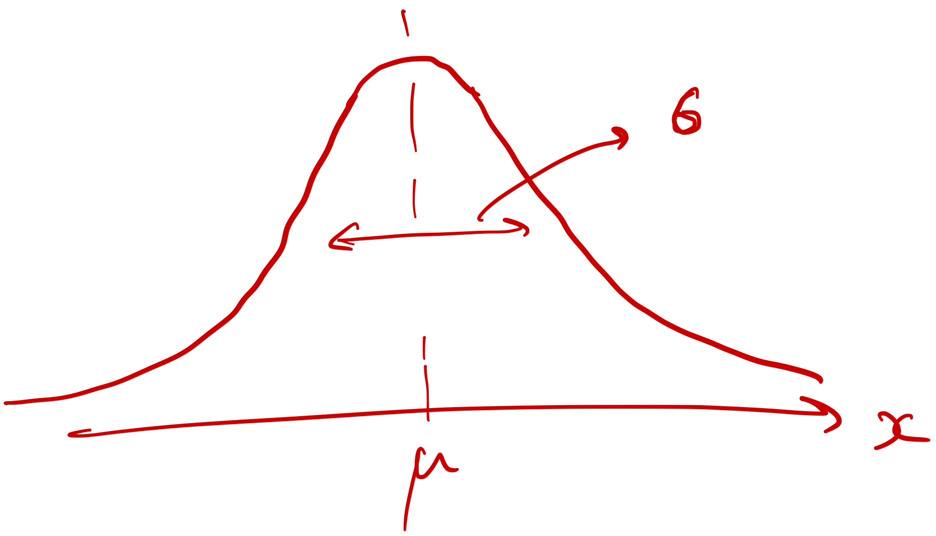
Gaussian Distribution:

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

mean

variance

Pdf: $\dfrac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$



The standard gaussian or the normal distribution:

$$Z = N(0, 1)$$