FoDA L6 : Central Limit Theorem

Data Set $P = \{p_1, p_2, \dots, p_n\}$

iid : Independently and Identically Distributed

polling : each $p_i \Leftarrow$ call to someone w/ landline

R.V. $X \underset{iid}{\sim} f \Longrightarrow$ sample mean $\bar{P} = \frac{1}{n} \sum_{i=1}^{n} p_i$

observations
data

$\bar{P} = \{p_i\} \xleftarrow{\text{realize}} \{X_i\} \underset{iid}{\sim} f$

$\frac{1}{n}\sum$

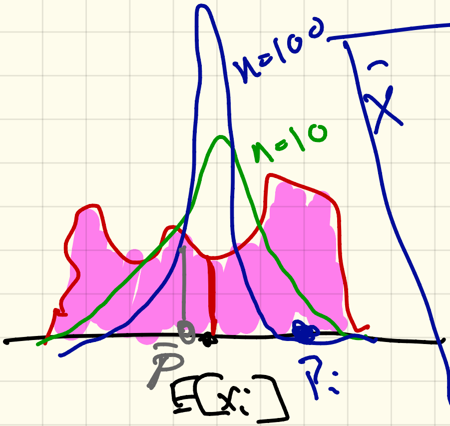
reason
unknown
distribution

Central Limit Theorem

R.V. $\{X_1, X_2, \ldots, X_n\}$ $\quad X_i \underset{\text{iid}}{\sim} f$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \leftarrow \text{another random variable}$$

$$\mu = E[X_i] \qquad \sigma^2 = Var[X_i]$$



$\bar{X}$ converges to normal distribution

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

$$Var[\bar{X}] = \frac{\sigma^2}{n}$$

n gets larger $Var[\bar{X}]$ smaller

How close is $\bar{P}$ to $E[\bar{X}] = E[X_i]$

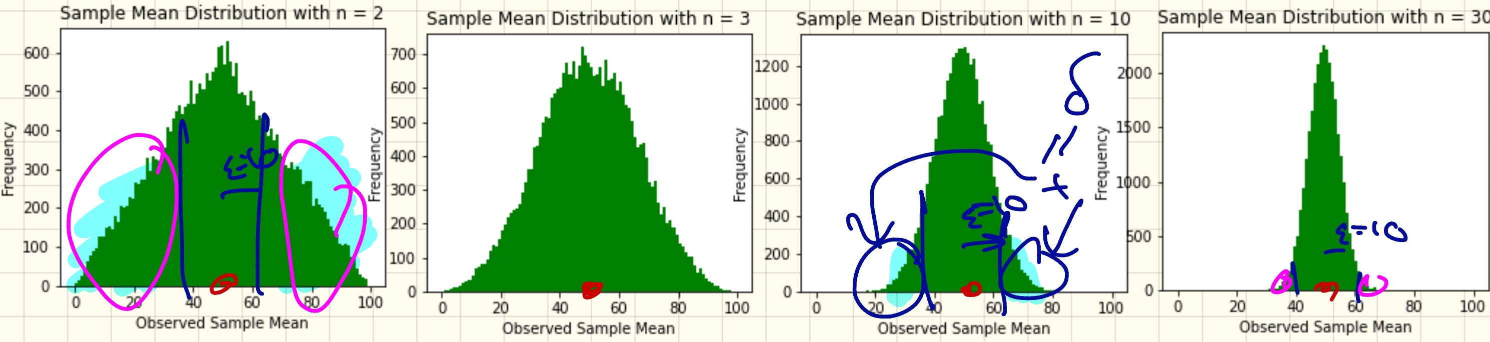
How close $\bar{X}$ to $E[\bar{X}]$?

R.V. 50

ε = error tolerance $\quad \left| \bar{X} - E[\bar{X}] \right| \leq \varepsilon$

δ = probability of failure

that is not true



$$\bar{p} \xleftarrow[\text{realization}]{} \bar{X}$$

Probably Approximately Correct (PAC)

$$\Pr\left[\left|\bar{X} - E[\bar{X}]\right| > \underbrace{\varepsilon}_{\text{error tolerance}}\right] < \underbrace{\delta}_{\text{probability of failure}}$$

so

$$\Pr\left[\bar{x} \in [40, 60]\right] \quad \text{e.g. } \varepsilon = 10$$