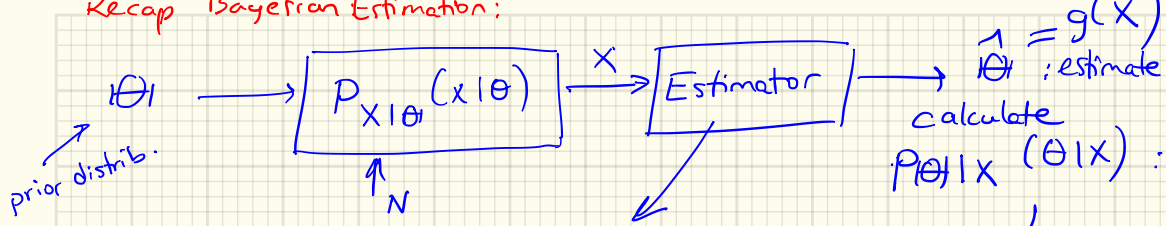
YZV 231E

Probability Theory & Stats

04.01.2022

Gü.

Recap Bayesian Estimation:



→ 1) MAP: $\hat{\theta} = \arg\max_{\theta} P_{\Theta|X}(\theta|x)$

→ 2) LMS: $\hat{\theta} = E_{\Theta|X}[\Theta|X]$ ←

Some properties of LMS Estimator:

- Estimator $\hat{\Theta} = E[\Theta|X]$

Estimation Error: $\tilde{\Theta} \triangleq \hat{\Theta} - \Theta$ (r.v.)

$E[\tilde{\Theta}|X] = E[(\hat{\Theta} - \Theta)|X] = E[\hat{\Theta}|X] - E[\Theta|X]$

$\hat{\Theta} = g(X)$

given X, $\hat{\theta}$ is known

$\hat{\theta} - \hat{\theta} = 0$

$E[\tilde{\Theta}|X] = 0$ (r.v. = r.v.)

$E[E[\tilde{\Theta}|X]] = E[\tilde{\Theta}] = 0$

→ Estimator $\hat{\theta}$ is UNBIASED.

→ $\mathrm{Cov}(\tilde{\theta}, \hat{\theta}) = ?$ (r.v. r.v.)

To find this: $E[\tilde{\theta}\cdot\hat{\theta}] - E[\tilde{\theta}]\cdot E[\hat{\theta}]$

Given X, $h(X)$ is a number

$E[\tilde{\theta}\cdot h(X) \mid X] = ?$

$h(X)\cdot E[\tilde{\theta} \mid X] = 0$

$E[E(\tilde{\theta} h(X) \mid X)] = E[\tilde{\theta} h(X)] = 0$ for any function $h(X)$

we know $\hat{\theta}(X)$: a fn. of X $\therefore E[\tilde{\theta}\cdot\hat{\theta}] = 0$

$\mathrm{Cov}(\tilde{\theta}, \hat{\theta}) = E[\tilde{\theta}\hat{\theta}] - E[\tilde{\theta}]E[\hat{\theta}]$

$\mathrm{Cov}(\tilde{\theta}, \hat{\theta}) = 0$

$\therefore$ The estimation error is uncorrelated wl the estimate.

$\tilde{\Theta} = \hat{\Theta} - \Theta \rightarrow \Theta = \hat{\Theta} - \tilde{\Theta}$; $\hat{\Theta}$ & $\tilde{\Theta}$ are uncorrelated,

— Variance ~ a measure of uncertainty.

$$\underbrace{Var(\Theta)}_{\text{uncertainty in the true r.v. } \Theta} = \underbrace{Var(\hat{\Theta})}_{\text{uncertainty in the estimate}} + \underbrace{Var(\tilde{\Theta})}_{\text{uncertainty in the error.}}$$

Linear LMS estimator:

→ Next, consider a simpler estimator of Θ, of the form:

Let $\hat{\Theta} = g(X) = aX + b$: affine mapping ; parameters a, b are unknown.

→ minimize $_{a,b}$ $E[(\Theta - (aX+b))^2]$ →

Last time minimize $E[(\Theta - g(X))^2]$: a generic estimator nonlinear.

$$g(X) = E[\Theta | X] = \hat{\Theta}$$

exercise: Derive

$$E[\Theta^2 - (aX+b)^2 - 2\Theta(aX+b)]$$

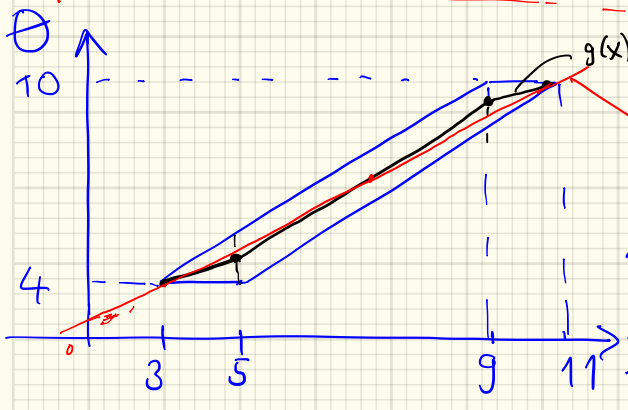
$$= E[\Theta^2] + a^2E[X^2] + 2ab\,E[X] + b^2 - 2aE[\Theta X] - 2bE[\Theta]$$

$\frac{\partial}{\partial a}\{\,\} = 0$, $\frac{\partial}{\partial b}\{\,\} = 0$ $\rightarrow a, b$ ✓

→ "Best" choise of a, b; "best" linear estimator (in the mse sense)

$$\hat{\Theta}_L = E[\Theta] + \frac{cov(X,\Theta)}{Var(X)}(X - E[X])$$

Recall prev ex.



$\hat{\Theta}_L$: $cov(X,\Theta) > 0$

$$\hat{\Theta}_L = E[\Theta] + \frac{cov(X,\Theta)}{Var(X)}(X - E[X])$$

($cov(X,\Theta) > 0$, $Var(X) > 0$, $(X - E[X])$ ↑↑)

i) $\hat{\Theta}_L = E[\Theta]$ when $cov(X,\Theta) = 0$

ii) $cov(X,\Theta) > 0$

→ What is the MSE ? $E\left[(\hat{\theta}_L - \theta)^2\right] = (1-\rho^2)\,\sigma_\theta^2$

exercise: derive this result. $\rho = \dfrac{Cov(X,\theta)}{\sigma_X\,\sigma_\theta}$ correlation coeff

$$\boxed{MSE^{(\hat{\theta}_L)} = (1-\rho^2)\,\sigma_\theta^2}$$

variance of the original r.v. θ.

i) MSE ↑ ✓ σ_θ^2 ↑.

$X \longrightarrow$ [Estimator] $\longrightarrow g(x) = \hat{\theta}$, here $\hat{\theta}_L$ ✓.

ii) When the two r.v.s X, θ are correlated, ⇒ we improve our estimate, MSE ↓

iii) $\rho = 1$ → maximal correlation case $\rightarrow \rho^2 = 1$ $MSE = 0$. ✓

iv) $\rho = 0$: 2 r.v.'s are uncorrelated. $MSE = \sigma_\theta^2$: measurements don't help to improve our estimate.

Linear LMS w/ Multiple Data: We make several measurements $X_1, \dots, X_n$

Linear Estimator is of the form:

$$\hat{\Theta} = a_1 X_1 + \dots + a_n X_n + b$$

→ Find "best" coefficients $a_1, \dots, a_n, b$ → "optimal" Linear LMSE estimator.

"optimal" LMSE estimator

$$\hat{\Theta} = E[\Theta \mid X_1, \dots, X_n]$$

Minimize $E\left[(\hat{\Theta} - \Theta)^2\right] = E\left[(a_1 X_1 + \dots + a_n X_n + b - \Theta)^2\right]$

$a_1, a_2 \dots, a_n, b$

$a_1^2 E[X_1^2] + 2 a_1 a_2 E[X_1 X_2] + \dots$

$\frac{\partial}{\partial a_1} = 0$ $\frac{\partial}{\partial a_2} = 0 \dots \frac{\partial}{\partial b} = 0$

— we need full posterior distrib. of $\Theta | X_1, \dots, X_n$ to calculate $\hat{\Theta}$.

But w/ the Linear LMSE estimator: we need to know expectations only: means, covariances, variances, ...

We don't need to know the whole distribution.

— Linear LMS example : θ : unknown (random) parameter.

measurement model: $X_i = \theta + W_i$:

$W_i \sim 0, \sigma_i^2$

prior → $\theta \sim \mu, \sigma_0^2$

Assumption:

$\theta, W_1, \dots, W_n$ are independent

X_i: measurement in i^{th} experiment/measurement

$X_1 = \theta + W_1$

$X_2 = \theta + W_2$

⇒ The form of the "optimal" linear estimator is very nice/neat form.

$$\hat{\theta}_L = \frac{\mu/\sigma_0^2 + \sum_{i=1}^{n} X_i/\sigma_i^2}{\sum_{i=0}^{n} \frac{1}{\sigma_i^2}}$$

→ sum of all coeffs in the numerator

a weighted average of $\mu, X_1, \dots, X_n$

weights are interesting!

— $\sigma^2 \approx$ measure of uncertainty, $\frac{1}{\sigma^2} \approx$ reliability.

★ Weights are inversely proportional to the (variance) uncertainty in the measurement.

★ The prior mean is treated the same way as the X_i's.

— Here, we did not impose any certain shape of the distributions of $W \& \theta$.

★ For normal r.v.s; (all normal, $\Theta \propto W$):

$$\hat{\Theta}_L = E[\Theta \mid X_1, \ldots, X_n]$$

∴ The optimal (LMSE) estimate ∝ the optimal linear (LMSE) estimate turn out to be the same.

→ If your measurement device measures X^2 or X^3's instead of X; want to estimate Θ.

→ Choosing X_i in the Linear LMS vs LMS estimate

For LMS estimator ⇒ b/c $X \propto X^2$ or X^3 have the same info posterior distribution of $\Theta|X \propto \Theta|X^3$

$\Theta|X$ or $\Theta|X^3$; → do not matter are the same

For the Linear LMS estimator

$\hat{\Theta} = aX + b$ vs $\hat{\Theta} = aX^3 + b$

$\hat{\Theta} = a_1 X + a_2 X^2 + a_3 X^3 + b$

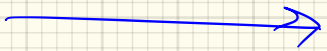
In Linear LMS estimator, which features/functions of observations to choose from matters.

Bayesian Estimation methods: Use a prior and Bayes' thm to find posterior distributions.

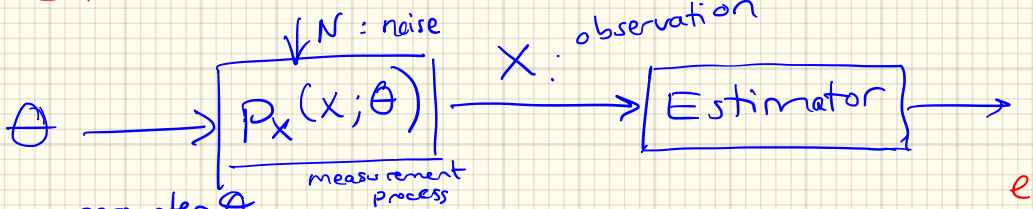
- MAP
- MSE
- Linear MSE

- Standard Models : $X_i = \Theta + W_i$;

- X_i : uniform on $[0, \theta]$; uniform prior on Θ
- X_i : Bernouilli (p) : uniform (Beta) prior on p.
- X_i : Normal : $\mathcal{N}(\mu, \sigma^2)$: normal prior on Θ



Classical Estimation :



$p_X(x;\theta)$: a family of distributions, parameterized by θ.

; notation → be careful!

Note: not a conditional distrib. : that's why we do not use $p_X(x|\theta)$ as in Bayesian setting.

→ Design an estimator $\hat{\theta}$ to keep estimation error $\hat{\theta} - \theta$ small

- $\hat{\theta}(X)$: a fn. of an r.v. ∴ $\hat{\theta}$ is an r.v.

→ Let's check out plausible estimator using $p_X(x;\theta)$ →

Maximum Likelihood Estimation: (MLE)

- $X \sim p_X(x;\theta)$: a model w/ unknown parameter θ.
- pick θ that "makes the data X most likely":

$$\hat{\theta}_{ML} = \arg\max_{\theta} p_X(x;\theta)$$

Recall / compare to the Bayesian MAP estimator

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p_{\Theta|X}(\theta|x) = \frac{p_{X|\Theta}(x|\theta)\, p_\Theta(\theta)}{p_X(x)}$$

($p_\Theta(\theta)$: prior)

Θ is an r.v. → we can have a prior $p_\Theta(\theta)$.

model of the measurement process was $p_{X|\Theta}$.

★ If the prior p_Θ is constant (uniform → all θ's are equally likely), then the ML estimation takes exactly the same form as the Bayesian MAP estimation. →

Ex: Let $X_1, \ldots, X_n$; i.i.d. exponential r.v.'s w/ a certain parameter θ.

$\exp(\theta)$

$X_i \sim \theta \exp(-\theta \cdot x_i)$.

joint distrib. $p_{X_1,\ldots,X_n}(x_1 \ldots, x_n ; \theta) = \prod_{i=1}^{n} \theta \cdot e^{-\theta x_i}$

: prob. of observing a particular instance $(x_1, \ldots, x_n)$ given a particular θ value.

MLE: $\max_{\theta} \prod_{i=1}^{n} \theta \cdot e^{-\theta x_i}$: what's the value of θ that makes the X's that we observed most likely?

$p_X(x;\theta)$

maximizing this expression $\equiv$ maximizing its logarithm. → b/c log is monotonous fn

log

$\max_{\theta} \left(n \log \theta - \theta \sum_{i=1}^{n} x_i \right)$ ↓ $\frac{\partial(\cdot)}{\partial \theta} = 0$

$\frac{n}{\theta} - \sum_i x_i = 0$

$\Rightarrow \hat{\theta}_{ML} = \frac{n}{x_1 + \ldots + x_n}$: = reciprocal of the sample mean : $\frac{x_1 + \ldots + x_n}{n}$.

Recall:

→ In an exponential distrib ; the mean $= \frac{1}{\theta}$:

$\therefore \hat{\theta}_{ML}$ is a reasonable estimate. ✓

$X_1, \dots, X_n \longrightarrow$ [ML Estimator] $\longrightarrow \hat{\theta}_n(X_1, \dots, X_n)$: r.v.

$$\hat{\theta}_{ML} = \frac{n}{X_1 + \dots + X_n}$$

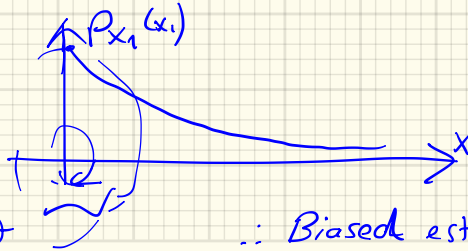
Desirable Properties of Estimators

① **Unbiased**: $E[\hat{\theta}_n] = \theta$ ✓

We don't want the estimator to have a systematic error on the +ve or -ve side of the true parameter θ.

ex: exponential example w/ $n=1$

$$\hat{\theta}_{ML} = \frac{1}{X_1} \rightarrow E[\hat{\theta}_{ML}] = E\left[\frac{1}{X_1}\right] = \infty \neq \theta$$



$\therefore$ Biased estimator. in this case.

(2) Consistent : $\hat{\theta}_n \xrightarrow{\text{in probability}} \theta$

ex : exponential example : $X_i \sim \exp(\theta)$

$$\left(\frac{X_1 + \dots + X_n}{n}\right) \xrightarrow{\text{WLLN, in prob.}} E[X] = \frac{1}{\theta}$$

sample mean $\longrightarrow \mu$: the mean

$$\hat{\theta} = \frac{n}{X_1 + \dots + X_n} \xrightarrow{\text{in prob}} \frac{1}{E[X]} = \theta \quad , \quad \forall \theta$$

$$\hat{\theta} \longrightarrow \theta$$

MLE here is a consistent estimator

(3) "Small" Mean_Squared Error (MSE)

$$E\left[(\hat{\theta} - \theta)^2\right] = \mathrm{Var}(\hat{\theta} - \theta) + \left(E[\hat{\theta} - \theta]\right)^2$$

w.r.t. $p_{\hat{\theta}}$ for a particular θ value

$$= \mathrm{Var}(\hat{\theta}) + (\text{Bias})^2$$

$\text{Bias} = \hat{\theta} - \theta$

For an unbiased estimator Bias $= 0$

$\mathrm{Var}(\hat{\theta})$: uncertainty in the estimate

our desire : small $\mathrm{Var}(\hat{\theta})$

⇒ We desire both Small Variance & Small Bias.

$$MSE = E[(\hat{\theta}-\theta)^2] = \underbrace{Var(\hat{\theta})}_{\text{Variance of the estimator}} + \underbrace{(Bias)^2}_{\text{Bias of the estimator.}}$$

But we typically have a Bias/Variance Trade-off:

eg. $X \sim \mathcal{N}(\theta, 1)$ (θ: parameter is the mean)

say $\hat{\theta} = 100$: constant output. : Naive estimator

$Var(\hat{\theta}) = 0$: smallest variance ✓

$Bias = \hat{\theta} - \theta$: θ^{true} : 0, 1 .. small

$Bias = 100 - \theta$

$(Bias)^2 = (100-\theta)^2$ Bias ↗↗.

For this naive estimator, MSE is huge!

Conclusion: You may decrease the variance, but ∃ a tradeoff → your bias explodes!!

So to come up w/ low Bias & low Variance in your estimator

→ you have to come up w/ sophisticated estimators.

For parameter estimation (classical estimators): we study:

1) MLE : Maximum Likelihood estimation ✓

2) Sample Mean Estimator.

Sample Mean Estimator:

$X_1, \dots, X_n$: i.i.d, mean θ, variance σ^2.

$X_i = \underbrace{\theta}_{\text{mean}} + \underbrace{w_i}_{\text{noise}}$.

w_i : i.i.d. mean 0, variance σ^2

$\hat{\theta} = \dfrac{X_1 + \dots + X_n}{n}$

$\hat{\theta}$ ↓ Sample mean Estimator

Properties of the Sample Mean Estimator.

- $E[\hat{\theta}_n] = E\left[\frac{X_1 + \dots + X_n}{n}\right] = \theta$ (Unbiased) $\therefore$ Bias $= 0$.

  $= \frac{n \cdot E[X]}{n} = \frac{n \cdot \theta}{n} = \theta$

- $\hat{\theta}_n \xrightarrow{\text{w/ WLLN}} \theta$ (Consistent)

- MSE : $E[(\hat{\theta} - \theta)^2] = \text{Var}(\hat{\theta}) + (\text{bias})^2$ ($\text{bias}^2 \to 0$)

  $\text{Var}(\hat{\theta}) = \frac{1}{n^2} \cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}$

  $= \frac{\sigma^2}{n} \xrightarrow[n \to \infty]{} 0$

Note: If $X_i \sim \mathcal{N}(0, \sigma^2)$, i.i.d

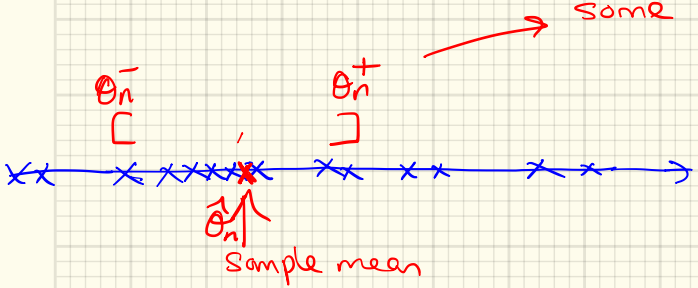
Sample mean Estimator = ML estimate.

exercise: write $p_X(x; \theta)$ $\Leftarrow$ maximize w.r.t. θ.

$\frac{\partial}{\partial \theta} = 0$?

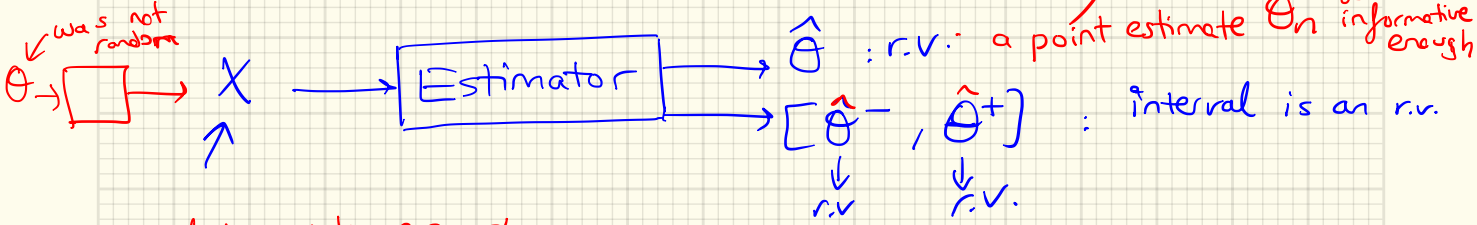
Now, → You report your sample mean : 3.27

Q. How reliable is that number? Can we trust it?



→ some interval w/ interval limits $\hat{\theta}_n^-$, $\hat{\theta}_n^+$ where our true parameter may lie.

→ CONFIDENCE INTERVALS (CI)



— We pick an α →

— A $(1-\alpha)$ confidence interval : $[\hat{\theta}_n^-, \hat{\theta}_n^+]$

s.t. $P(\hat{\theta}_n^- \leq \theta \leq \hat{\theta}_n^+) \geq 1-\alpha$, $\forall \theta$

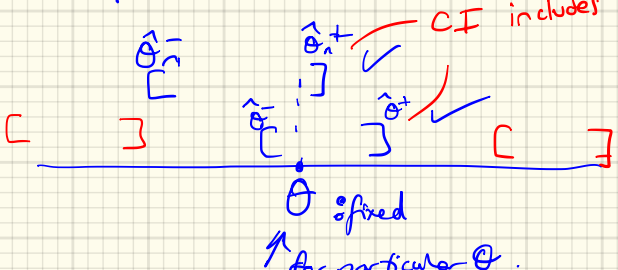
eg. $\alpha = 0.05$

→ 95% confidence Interval interpretation:

Rather than [the statement that θ, w/ prob 95% falls in $[\hat{\theta}_n^-, \hat{\theta}_n^+]$, b/c θ: not random,

Say this:

w/ prob 95%, the interval falls on the true value of θ



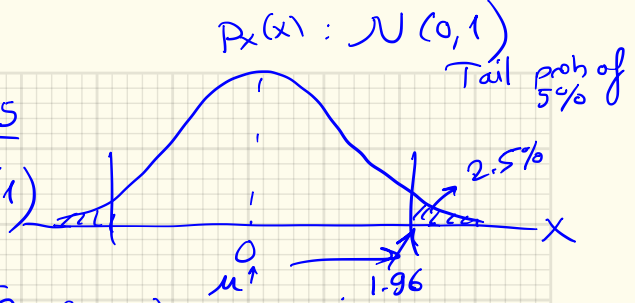
Q. How do we construct a 95% CI?

- CI in estimation of the mean:

$$\hat{\theta}_n = \frac{X_1 + \dots + X_n}{n}$$ ⟶ want to calculate 95% confidence interval

From the normal table:

$$\Phi(1.96) = 0.975 = 1 - \frac{0.05}{2}$$



Use CLT; standardize $\hat{\Theta}_n$ $\sim \mathcal{N}(0,1)$

$$P\left(\frac{|\hat{\Theta}_n - \theta|}{\sigma/\sqrt{n}} \leq 1.96\right) \approx 0.95 \quad \text{(CLT)}$$

rewrite

$$P\left(\hat{\Theta}_n - \frac{1.96\sigma}{\sqrt{n}} \leq \theta \leq \hat{\Theta}_n + \frac{1.96\sigma}{\sqrt{n}}\right) \approx 0.95$$

construct the CI

$\hat{\Theta}_n^-$: lower end of the CI $\qquad$ $\hat{\Theta}_n^+$: upper end

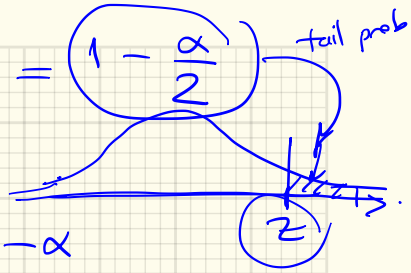
i) as $n \nearrow$; $[\hat{\Theta}_n^-, \hat{\Theta}_n^+]$: CI $\searrow$

we're more confident that our interval captures the true θ.

ii) If $\sigma \searrow$ (our data has low uncertainty) : CI $\searrow$

→ More generally: how to construct the CI?

Let z be s.t. $\Phi(z) = 1 - \frac{\alpha}{2}$



$$P\left(\underbrace{\hat{\theta}_n - \frac{z \cdot \sigma}{\sqrt{n}}}_{\hat{\theta}_n^-} \leq \theta \leq \underbrace{\hat{\theta}_n + \frac{z \cdot \sigma}{\sqrt{n}}}_{\hat{\theta}_n^+}\right) \approx 1 - \alpha$$

Here ⇒
- n is given ✓
- $\sigma = ?$

For Unknown σ; options:

1) Use an upper bound on σ;
   eg. $X_i \sim$ Bernoulli $\quad \sigma \leq \frac{1}{2}$

   } more conservative CI estimates ie larger.

2) Estimate σ from the data → use heuristic estimates
   eg. $X_i \sim$ Bernoulli(θ) → $\sigma = \sqrt{\theta(1-\theta)}$
   → $\hat{\theta}$ ✓ → $\sigma \approx \sqrt{\hat{\theta}(1-\hat{\theta})}$: estimate for the standard deviation

3) Use a generic estimate of the variance : Sample Variance:

$$\sigma^2 = E\left[(X_i - \theta)^2\right] :$$

$$\hat{\sigma}_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \theta)^2 \xrightarrow{\text{WLLN}} \sigma^2$$

Good if we knew θ.

→ We don't know θ (the mean)

→ Insert the sample mean estimate

$$\tilde{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\theta}_n)^2 \longrightarrow \sigma^2$$

$$E\left[\tilde{S}_n^2\right] = \sigma^2 \quad : \text{unbiased}.$$

Now use → $\sigma \approx \tilde{S}_n = \sqrt{\tilde{S}_n^2}$ in constructing your Confidence Interval limits