03.01 2022

YZV 231E

Probability Theory & Stats

Gü.

Recap: **Limit Theorems:**

We have a large # i.i.d. r.v.s.

eg. polling ; 1000 people $X_1, X_2, \dots, X_n$ (Yes/No)

- $M_n = \frac{X_1 + \dots + X_n}{n}$: estimate of the expected value

(M_n: sample mean)

- M_n : is a random variable b/c the sample you collected is random.

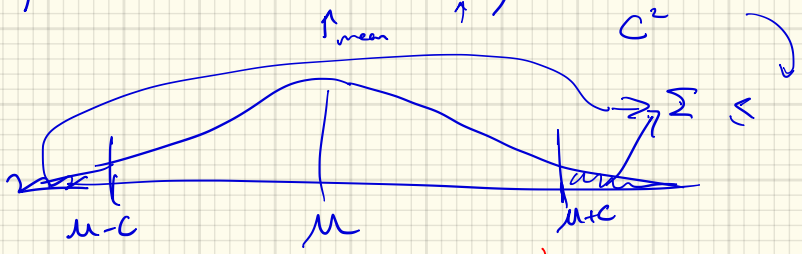
Is M_n representative of the (true) expected value ?

$$M_n \longrightarrow E[X] = \mu \quad \text{in probability}$$

- Convergence in probability : A seq. $Y_1, \dots, Y_n$ of r.v.s converges to a number a in prob. if as $n \to \infty$

$$P(|Y_n - a| > \varepsilon) \xrightarrow[n \to \infty]{} 0$$

- Chebyshev Ineq: $P(|X-\mu| \geq c) \leq \frac{\sigma_x^2}{c^2}$



Weak Law of Large Numbers (WLLN)

$$M_n = \frac{X_1 + \dots + X_n}{n}$$

$M_n \xrightarrow{\text{in prob.}} \mu$

$$\rightarrow E[M_n] = \frac{E[X_1] + \dots + E[X_n]}{n} = \frac{n \cdot \mu}{n} = \mu \rightarrow \text{true mean of the population}$$

$$\rightarrow \text{Var}(M_n) = \frac{1}{n^2} n \cdot \underbrace{\text{Var}(X_i)}_{\sigma^2} = \frac{\sigma_x^2}{n} \xrightarrow[n \to \infty]{} 0$$

Back to Ex: Polling: p: fraction of population that prefers ...

$$X_i = \begin{cases} 1, & \text{if yes (signed)} \\ 0, & \text{o/w} \end{cases}$$

Bernoulli r.v.

$M_n = \dfrac{X_1 + \cdots + X_n}{n}$: prediction of the fraction p.

Sample mean

Goal: $P(|M_n - p| \geq \underbrace{0.01}_{\text{accuracy}}) \leq \underbrace{0.05}_{1 - \text{confidence}}$

Given Specs w/ 95% confidence of $\leq 1\%$ error

i) Collect data, calculate sample mean M_n & check whether you satisfy the specs.

ii) Calculate the sample size n so that the specs are satisfied

→ Chebyshev ineq: $P(|M_n - p| \geq 0.01) \leq \dfrac{\sigma_{M_n}^2}{(0.01)^2} = \dfrac{\sigma_X^2/n}{(0.01)^2} \leq 0.05$

$\sigma_X^2 = ?$

$\sigma_x^2 = p(1-p) \rightarrow \arg\max_p \sigma_x^2(p) \rightarrow p = \frac{1}{2} \qquad \sigma_x^2 \leq \frac{1}{4}$

$$P(|M_n - p| \geq 0.01) \leq \frac{\sigma_x^2}{n \cdot (10^{-4})} \leq \frac{1}{n \cdot 4 \cdot 10^{-4}} \leq 0.05$$

$$\Rightarrow n \geq \frac{1}{4 \cdot 10^{-5} \cdot 5 \cdot 10^{-2}}$$

$n = 50$ K samples :

fine on Twitter, but not practical for "in person" polling.

→ ∴ play w/ specs :

eg. $P(|M_n - p| \geq (0.03)) \leq 0.05$

accuracy less conservative

Chebyshev

$(0.03)^2$

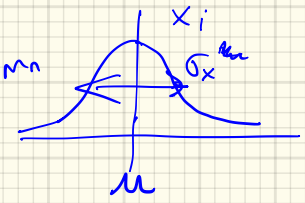
Saves a factor of ~ 10 ← $3^2 \cdot 10^{-4}$

→ $n \sim$ 5000 people.

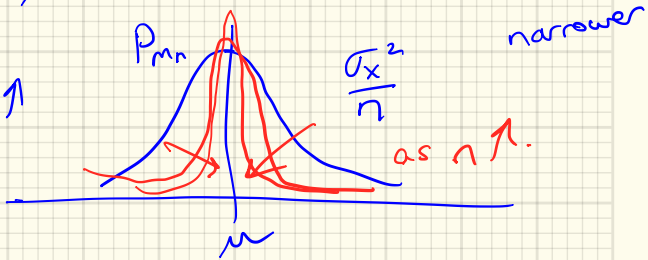
Scaling:

(i) Sample mean $M_n = \frac{X_1 + \cdots + X_n}{n}$, $X_i \to$ mean μ.

Scale $\frac{1}{n}$ — WLLN



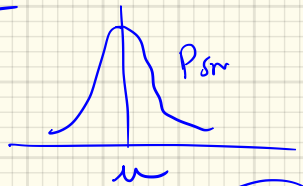
$\frac{1}{n}$ scaling, $n \uparrow$

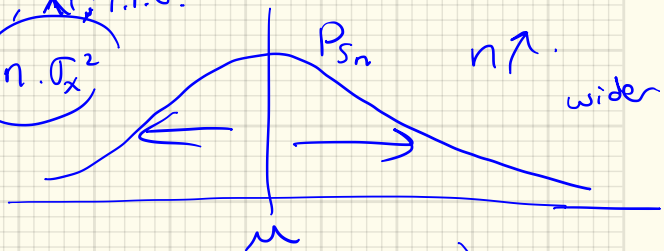


(ii) Sum: Scale = 1 — No-scale

$S_n = X_1 + \cdots + X_n$, X_i: i.i.d.

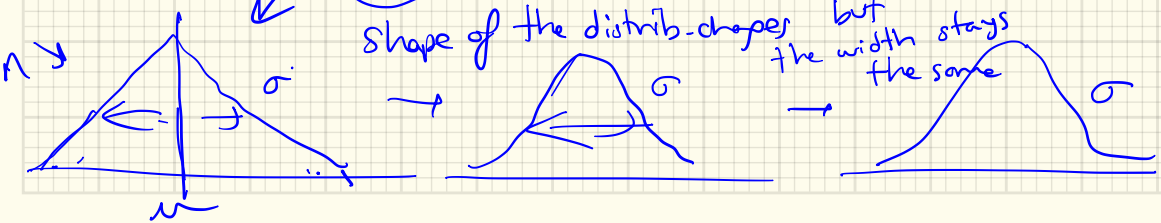
$\mathrm{Var}(S_n) = n \cdot \sigma_x^2$





(iii) Scale by $\frac{1}{\sqrt{n}}$?: $\frac{S_n}{\sqrt{n}}$

$\mathrm{Var}\left(\frac{S_n}{\sqrt{n}}\right) = \mathrm{Var}\left(\frac{X_1 + \cdots + X_n}{\sqrt{n}}\right) = \left(\frac{1}{\sqrt{n}}\right)^2 \cdot n\sigma_x^2 = \sigma_x^2$.



Different scalings of M_n; $X_1, \dots, X_n$ i.i.d w/ μ, σ^2

3 variants of the sum:

1) $S_n = X_1 + \dots + X_n$: variance $\boxed{n\sigma^2}$

2) $M_n = \frac{S_n}{n}$: variance $\frac{\sigma^2}{n}$

3) $\frac{S_n}{\sqrt{n}}$: constant variance σ^2

Central Limit Theorem (CLT)

Standardize $S_n = X_1 + \dots + X_n$ by

$$Z_n = \frac{S_n - E[S_n]}{\sigma_{S_n}} = \frac{S_n - n \cdot E[X]}{\sqrt{n}\,\sigma}$$

σ_{S_n} : standard deviation

$\sqrt{n\sigma^2}$

Z_n r.v.

This is STANDARDIZATION of an r.v.

$E[Z_n] = 0$
$Var(Z_n) = 1$

$Z_n \xrightarrow{\text{in distribution}} Z \sim N(0,1)$

Let Z be a standard normal r.v.

Theorem (CLT): For every c,

$$P(Z_n \le c) \to P(Z \le c)$$

$P(Z_n \le c)$: cdf of Z_n r.v.

$P(Z \le c)$: standard normal cdf

$\Phi(c) = P(Z \le c)$: standard normal CDF is available from normal tables.

$X_1, \ldots, X_n$ i.i.d. X_i: any distribution $\to$ Normal Distribution as $n \nearrow$

X_i:

$$Z_n = \frac{S_n - n E[X]}{\sqrt{n}\cdot\sigma}$$

$$\rightarrow S_n = \sigma\sqrt{n}\, Z_n + n E[X]$$

when n is large $\sim$ Normal. $\therefore S_n$ is $\sim$ Normal

b/c this is a linear transform of Z_n.

Ex: Let $X_i \sim \mathcal{N}(0,1)$ i.i.d. Define $Y_N = \sum_{i=1}^{N} X_i^2$

Approximate Y_N distrib. by a Gaussian. $\rightarrow$ justified by CLT.

b/c X_i i.i.d $\rightarrow X_i^2$ i.i.d. $Var(Y_N) = N \cdot Var(X^2)$

CLT says: $\tilde{Y}_N = \frac{Y_N - E[Y_N]}{\sigma_{Y_N}} = \frac{Y_N - N\cdot E[X^2]}{\sqrt{N}\cdot Var(X^2)} \xrightarrow{CLT.} \mathcal{N}(0,1)$

$E[X^2] = \mathrm{Var}(X) = 1$

$\mathrm{Var}(X^2) = \underbrace{E[X^4]}_{= 3 \text{ (check)}} - \underbrace{(E[X^2])^2}_{1} = 2$

$$y_N = \sum_{i=1}^{N} x_i^2$$

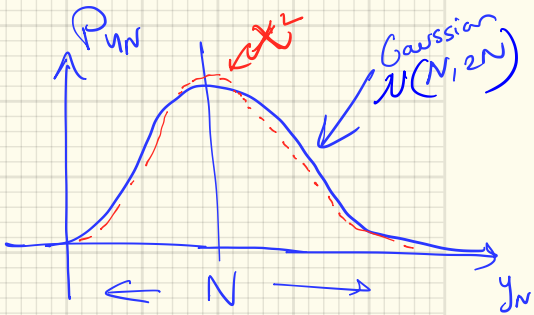
$$\rightarrow \tilde{y}_N = \frac{y_N - N}{\sqrt{2N}} \underset{N \to \infty}{\approx} \mathcal{N}(0,1) \quad \text{by CLT.}$$

$\rightarrow$ Approx. $y_N = \sqrt{2N}\,\tilde{y}_N + N \rightarrow y_N \sim \mathcal{N}(N, 2N)$

y_N: Gaussian; $\tilde{y}_N$: Gaussian w/ mean 0, var 1

$E[y_N] = E[\tilde{y}_N] + N$

$\mathrm{Var}(y_N) = 2N \cdot 1 + 0$



y_N : χ^2 : chi-squared.

$y_N \sim \chi^2_N$ exact

as $N \uparrow$ approx. by a Gaussian becomes better. Standard dev. $\sqrt{2N}$

Ex: Pollster's problem using CLT.

- p : fraction of the population that "prefers" - - .
- X_i : i^{th} randomly selected person, $X_i = \begin{cases} 1, & \text{if yes} \\ 0, & \text{o/w} \end{cases}$

$M_n = \frac{X_1 + \dots + X_n}{n}$: this is the estimate for the fraction of the population that preferred . . .

→ We define 2 specifications for the poll : ≡ 2 parameters

$\Rightarrow P(\underbrace{|M_n - p| \geq 0.01}_{\text{accuracy}}) \leq 0.05$ → confidence

≡ Specs

Want probability 95% that our estimate M_n is within 1% of the true value p.

Event of interest : $|M_n - p| \geq 0.01$

standardize → $\left| \frac{X_1 + \dots + X_n - np}{n} \right| \geq 0.01$

(≡ same event) $\left| \frac{X_1 + \dots + X_n - np}{\sqrt{n}\,\sigma} \right| \geq \frac{0.01\sqrt{n}}{\sigma}$

standardized r.v → Z_n ≡

Now: $P(|M_n - p| \geq 0.01) = P(|Z_n| \geq \frac{0.01\sqrt{n}}{\sigma})$ (M_n: sample mean, p: true)

Using CLT → $\approx P(|Z| \geq 0.01\sqrt{n}/\sigma)$

where Z is a standard normal r.v.

Q: what is σ? We know an upper bound $\sigma^2 \leq \frac{1}{4}$ → $\sigma \leq \frac{1}{2}$

$\frac{1}{\sigma} \geq 2$

$$P(|Z| \geq \frac{0.01\sqrt{n}}{\sigma}) \leq P(|Z| \geq 0.02\sqrt{n})$$

(i) Given $n = 10{,}000$

$P(|Z| \geq \frac{0.02\sqrt{10^4}}{=2}) = P(|Z| \geq 2)$

$= 2P(Z \geq 2)$

$= 2(1 - P(Z \leq 2))$

$\Phi(2)$ — use Normal table

$= 2(1 - 0.9772)$

$= 0.0456$ → prob of error $< 5\%$ $= 4.6\%$

(ii) Find n s.t. $P(|Z| \geq 0.02\sqrt{n}) = 0.05$

$\Phi(c) = 0.975$

↓

1.96 from the table

$0.02\sqrt{n} = 1.96 \Rightarrow n = 9604$

→ W/ $n = 9604$ people in the poll
our probability of error is 0.05. ✓

→ We use CLT in different ways. → approximate distributions.

Apply to Binomial:

X_i: Bernoulli (p), indep. → Fix p, $0 < p < 1$

$S_n = X_1 + \cdots + X_n$: Binomial (n, p)

Binomial → mean np, variance $= np(1-p)$ ✓

CDF of $\dfrac{S_n - np}{\sqrt{np(1-p)}} \longrightarrow$ Standard Normal Distrib.
$n \uparrow$

Check whether this approx. is good!

$n = 36$, $p = 0.5$; find $P(S_n \leq 21)$

CLT: standardize $\dfrac{S_n - 18}{3} \leq \dfrac{21-18}{3} = 1$

mean $= 18 = np$
var $= 9 = np(1-p)$

$Z_n \leq 1 \approx Z \leq 1$ (std normal)

$\rightarrow \Phi(1) = P(Z \le 1) = 0.843$ $\rightarrow$ an approximate answer.

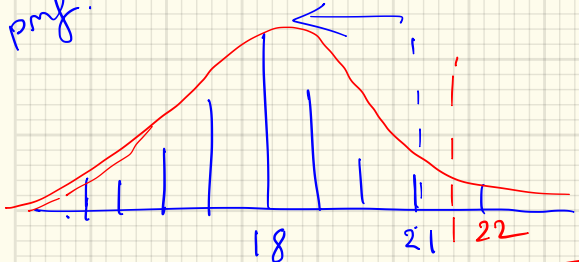
from the table

ok. but b/c S_n : a discrete r.v

Exact answer : $\sum_{k=0}^{21} \binom{36}{k} \left(\frac{1}{2}\right)^{36} = 0.8785$

$P(S_n = k)$

pmf.



W/ CLT : we treated S_N as normal

$P(S_n \le 21) = P(S_n < 22)$

$\therefore$ use $P(S_n \le 21.5)$

$\frac{S_n - 18}{3} \le \frac{21.5 - 18}{3} = 1.17$

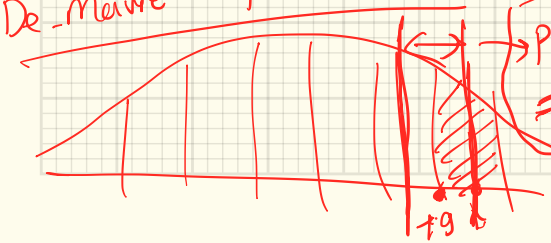
Table $P(Z \le 1.17) = 0.879$

a much better approximation

De-Moivre - Laplace CLT.

1/2 correction

$P(S_n \le 19) = P(18.5 \le S_n \le 19.5)$



STATISTICAL INFERENCE :

This is what we do w/ probability theory.

Real World phenomenon → Data → Model

Model → parameters

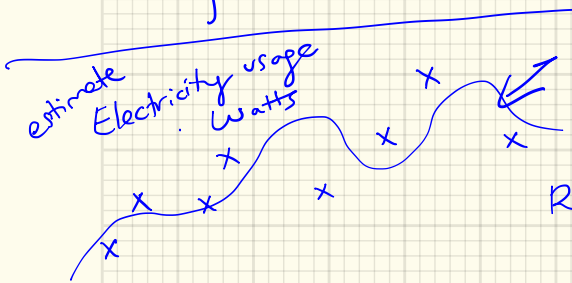
eg. customer arrivals

→ eg. Poisson r.v.

→ Come up w/ a probability model

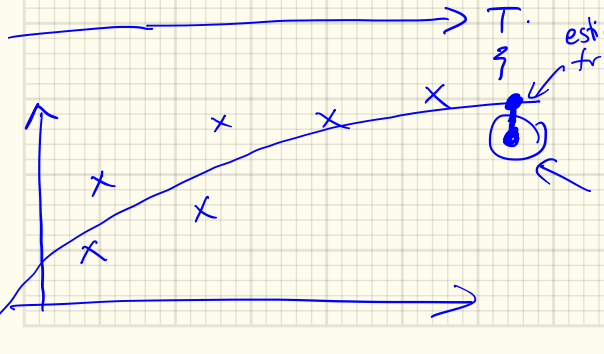
→ then calculate its parameters.

eg. here : the arrival rate λ.



Real world → Estimation Problems

→ Hypothesis Testing Problems



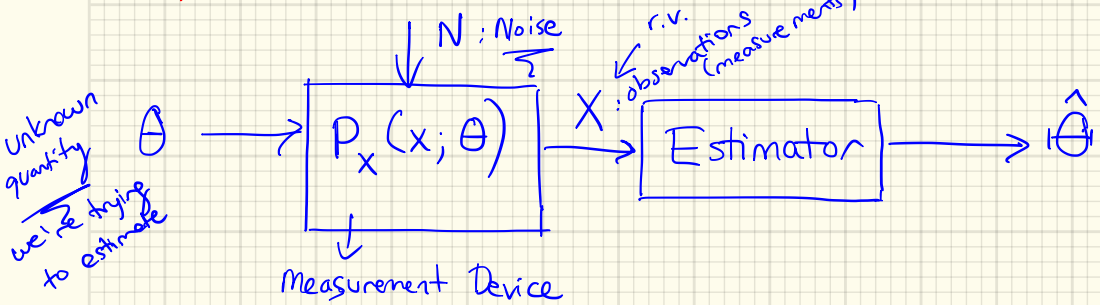
1) Bayesian Statistical Estimation

2) Classical Statistical Estimation

Estimation:

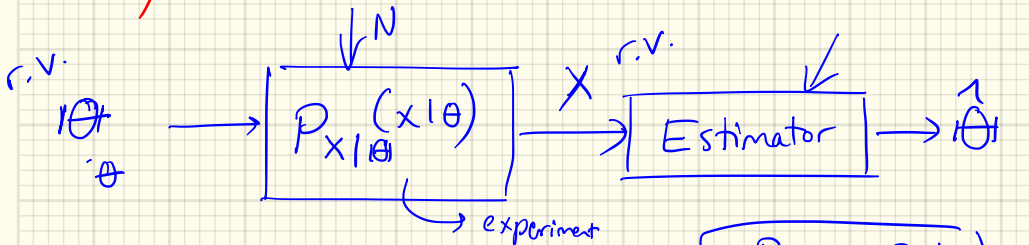
Classical vs Bayesian Statistics
ongoing debate.
whichever is better?

1) Classical Statistics



θ: unknown parameter (not an r.v.)

2) Bayesian Statistics : Use priors on θ



$P_\Theta(\theta)$: prior distribution.
our initial beliefs about θ before the experiment

Bayes Rule → come up w/ a posterior distrib.

$P_{\Theta|X}(\theta|x)$

our revised beliefs once we obtain data X

Bayesian inference: Use Bayes' rule

$$p_{\Theta|X}(\theta|x) = \frac{p_{X|\Theta}(x|\theta)\cdot p_{\Theta}(\theta)}{p_X(x)}$$

conditional / model of the experimental process

prior

we want to calculate posterior distrib. of Θ given X

p : pmfs or pdfs

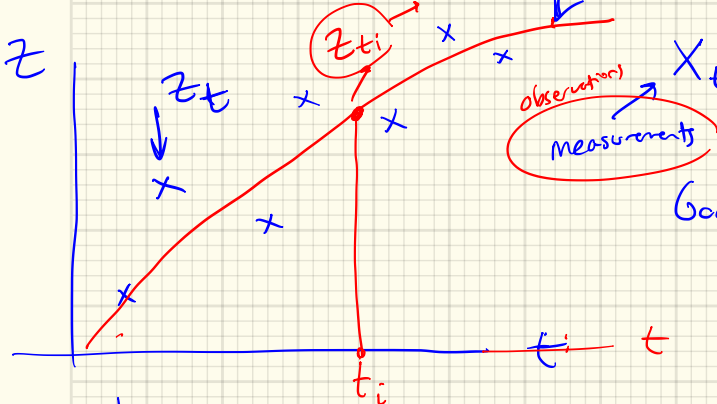
trajectory model

$$z_t = \theta_0 + \theta_1 t + \theta_2 t^2$$

$\Theta = [\theta_0\ \theta_1\ \theta_2]$

observations / measurements

$$X_t = z_t + W_t, \quad t = 1, 2, \ldots, n$$



Goal: Estimate $\theta_0, \theta_1, \theta_2$.

you have the model for X_t, measurements

$$\underline{\Theta} = \begin{bmatrix} \Theta_0 \\ \Theta_1 \\ \Theta_2 \end{bmatrix}$$

$$p_{\Theta_0,\Theta_1,\Theta_2|X_1,\ldots,X_n}(\theta_0,\theta_1,\theta_2|x_1,\ldots,x_n) = \frac{p(x_1,\ldots,x_n|\underline{\theta})\, p_{\underline{\Theta}}(\underline{\theta})}{p(x_1,\ldots,x_n)}$$

conditional pmf

continuous r.v.s Θ_i.

ex: Coin w/ unknown parameter θ (eg. $\theta = \frac{1}{2}$)

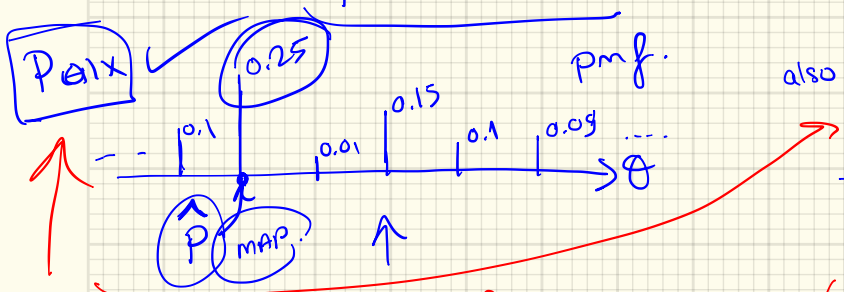
→ In classical statistics: we flip the coin many times $X_1, \ldots, X_n$, form S_n.

$$\hat{\theta}_n = \frac{S_n}{n} \xrightarrow{\text{WLLN}} \theta.$$? Check this w/ WLLN.

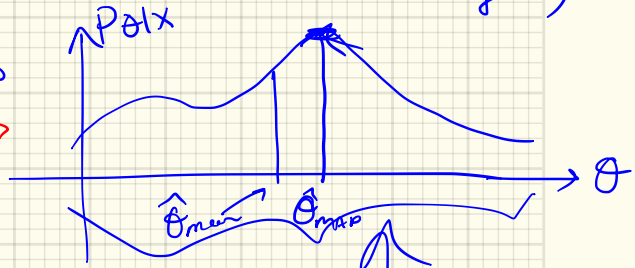
Now:

→ Bayesian approach: Assume a prior on θ (eg. uniform prior if you don't know anything :-)

Goal: $p_{\Theta|X}(\theta|x) = ?$ use Bayes rule to find:

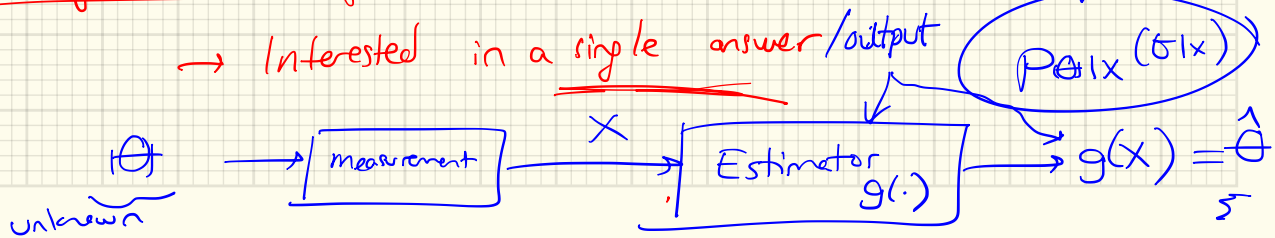


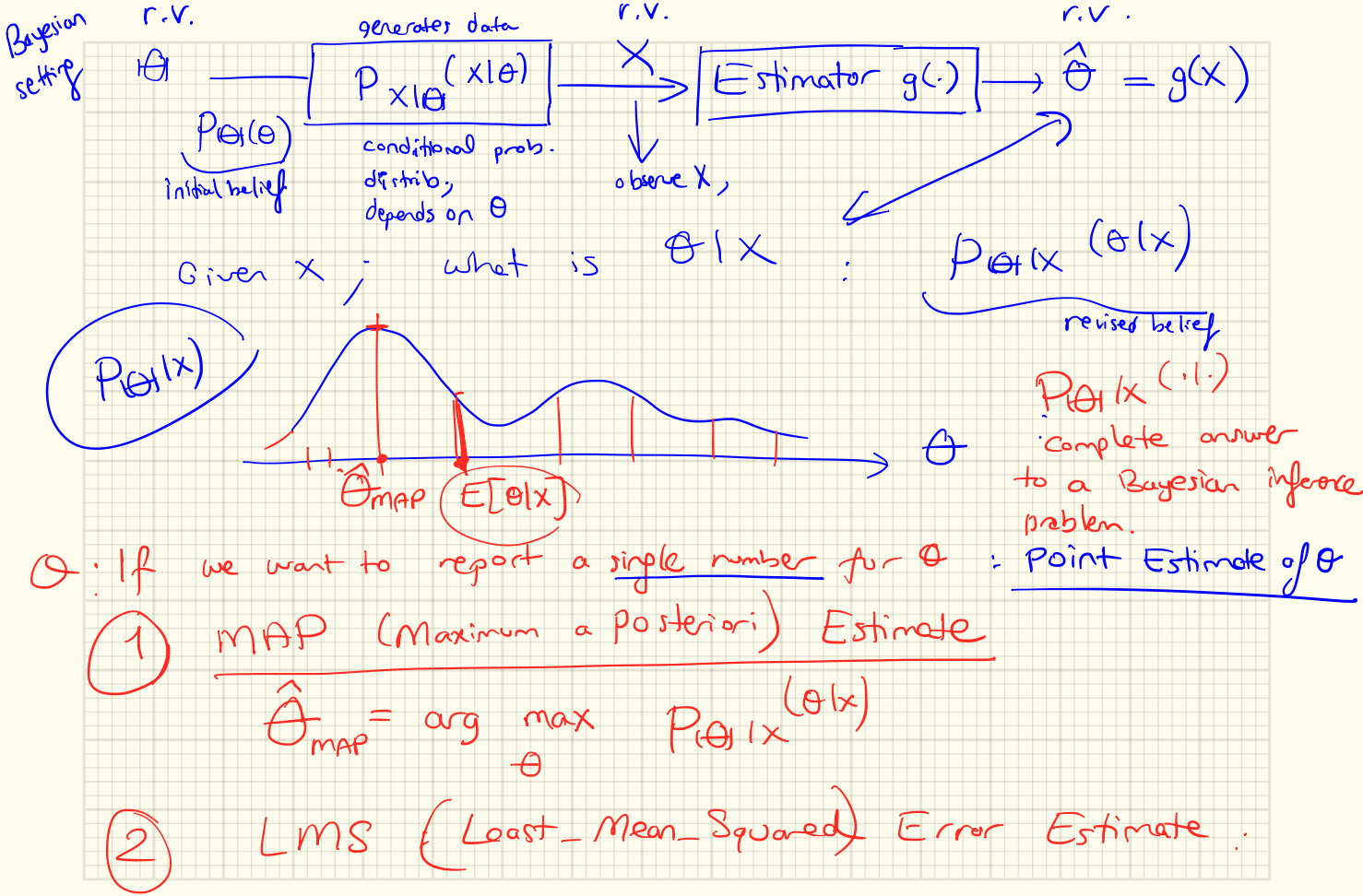
also



Output of Bayesian Inference: $p_{\Theta|X}(\theta|x)$

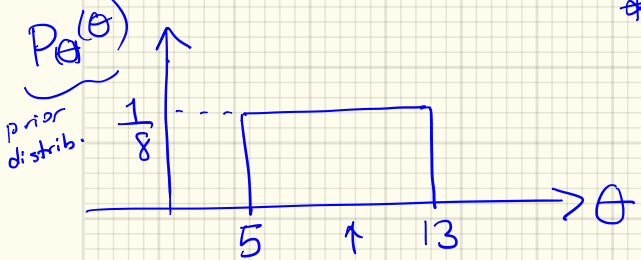
→ Interested in a single answer/output




Bayesian setting
r.v.
generates data
r.v.
r.v.
Θ
$p_{X|\Theta}(x|\theta)$
X
Estimator $g(\cdot)$
$\hat{\Theta} = g(X)$
$p_\Theta(\theta)$
initial belief
conditional prob. distrib; depends on θ
observe X,
Given X, what is $\Theta|X$:
$p_{\Theta|X}(\theta|x)$
revised belief
$p_{\Theta|X}(\theta|x)$
$\hat{\theta}_{MAP}$
$E[\Theta|X]$
θ
$p_{\Theta|X}(\cdot|\cdot)$
complete answer to a Bayesian inference problem.
θ : If we want to report a single number for θ : Point Estimate of θ
1 MAP (Maximum a posteriori) Estimate
$\hat{\theta}_{MAP} = \arg\max_\theta p_{\Theta|X}(\theta|x)$
2 LMS (Least_Mean_Squared) Error Estimate .

Least Mean Squares Estimation : (LMS)

* Estimation in the absence of information (no X observed)



→ Goal : Come up w/ a point estimate.

using LMS criterion

point estimate c : minimize

$$\min \; E\left[(\Theta - c)^2\right]$$

: make a small error ; squaring the error we penalize large errors even more.

This is called the Least Mean Squares Estimation

Optimal estimate $c = ?$

$$E\left[\Theta^2\right] - 2E[\Theta]\cdot c + c^2$$

min. by taking $\frac{\partial}{\partial c} = 0$ →

$$\boxed{c = E[\Theta]}$$

; above ex. $c = 9. = E[\Theta]$

$\frac{d}{dc} = -2E[\Theta] + 2c = 0$

$\rightarrow c = E[\Theta]$

To judge how good is our estimate?

"optimal" = mean-squared error

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

→ Expectation is the "best" estimate if you're minimizing the least-mean-squared error.

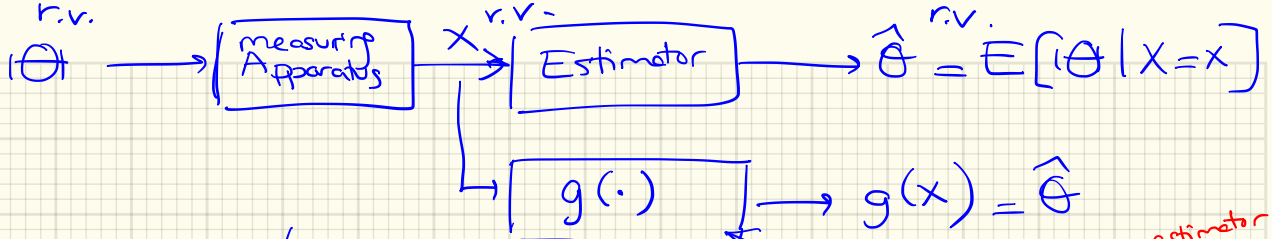
→ Next: we have data X, how will this estimate change?

LMS estimation of Θ based on X: Given X, we are in a conditional universe

$$\min \; E\left[(\Theta - c)^2 \mid X = x\right]$$

is minimized by $c = E[\Theta \mid X = x]$

now, you use conditional expectation



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

min $E_{X,\Theta}\left[(\Theta - g(X))^2 \right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (\Theta - g(X))^2 \, p_{X,\Theta}(x,\theta)\, dx\, d\theta$

min $= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} (\Theta - g(x))^2 \, p_{\Theta|X}(\theta|x)\, d\theta \right) p_X(x)\, dx.$

$p_X(x) \geq 0$

we can minimize $\int_{-\infty}^{\infty} (\theta - c)^2 \, p_{\Theta|X}(\theta|x)\, d\theta = f(c)$ (c unknown.)

$\frac{\partial f}{\partial c} = 0$

$2\int_{-\infty}^{\infty} (\theta - c)\, p_{\Theta|X}(\theta|x)\, d\theta = 0$

$\int \theta \, p_{\Theta|X}(\theta|x)\, d\theta = c \cdot \int p_{\Theta|X}(\theta|x)\, d\theta$ ($\int p_{\Theta|X}(\theta|x)\, d\theta = 1$.) $\Rightarrow$

$$\rightarrow c = E[\theta|x] = \int \theta \cdot p_{\theta|x}(\theta|x)\, d\theta$$

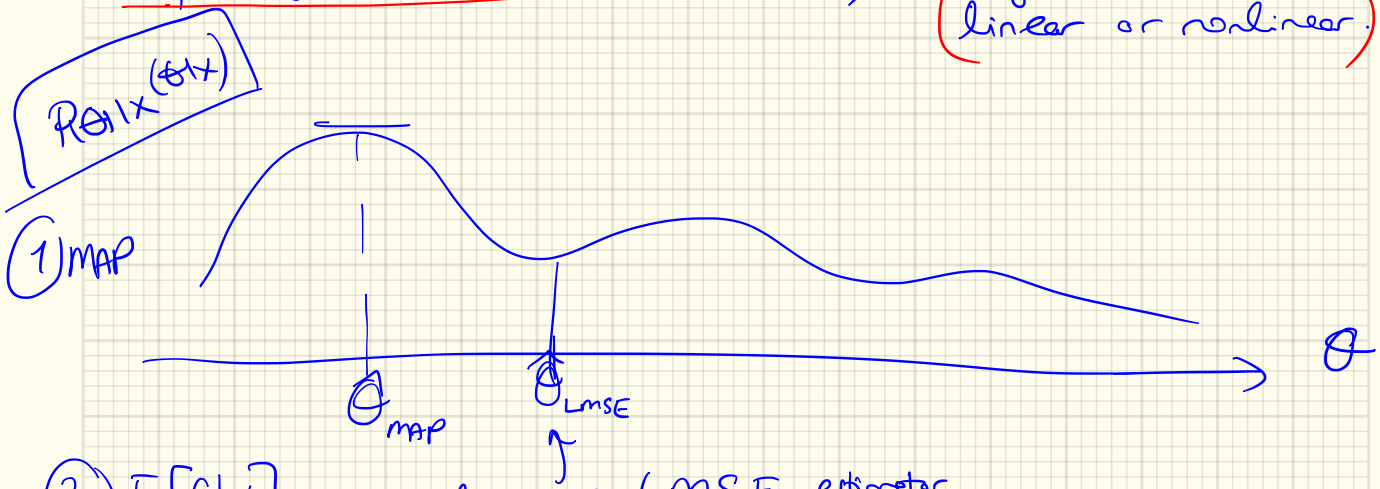
$$= g(x) \equiv \hat{\theta}_{\theta|x}$$

$$\Rightarrow \therefore \boxed{\hat{\theta}(x) = E[\theta|x]}$$

$= g(x)$ $\quad$ $\hat{\theta}_{\theta|x}$

in the MSE sense. (LMSE) sense

"Optimal" estimator (LMSE sense) among all estimators (linear or nonlinear)



(2) $E[\theta|x]$ = conditional mean of the posterior density $\rightarrow$ LMSE estimator

Bayesian estimation w/ several measurements

Unknown r.v. Θ

Observe values of r.v.s $X_1, \ldots, X_n$

"Best" estimator

- LMS: $E[\Theta \mid X_1, \ldots, X_n]$
- MAP: $\arg\max_{\theta} p_{\Theta \mid X_1 \ldots X_n}(\theta \mid x_1, \ldots x_n)$

Calculate using Bayes' rule: $p_{\Theta \mid X_1, \ldots X_n}(\theta \mid x_1, \ldots x_n)$

Issues:

- Have to come up w/ a reasonable prior distrib. for Θ.
- Have to calculate the posterior distrib $p_{\Theta \mid X}$ & its expectations:

Those calculations can be really hard !! → even intractable !!

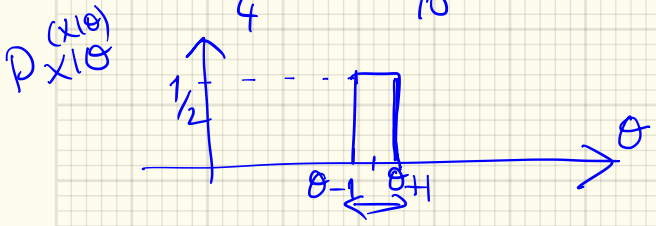
w/ multi-dimensional integrals . .

Ex: Suppose $\Theta \sim U[4,10]$;

Observation model: $X = \Theta + U$; $U \sim U[-1,1]$ (uniform)

$\Theta \rightarrow$ [Measurement] $\rightarrow X$; $U \rightarrow$ Measurement

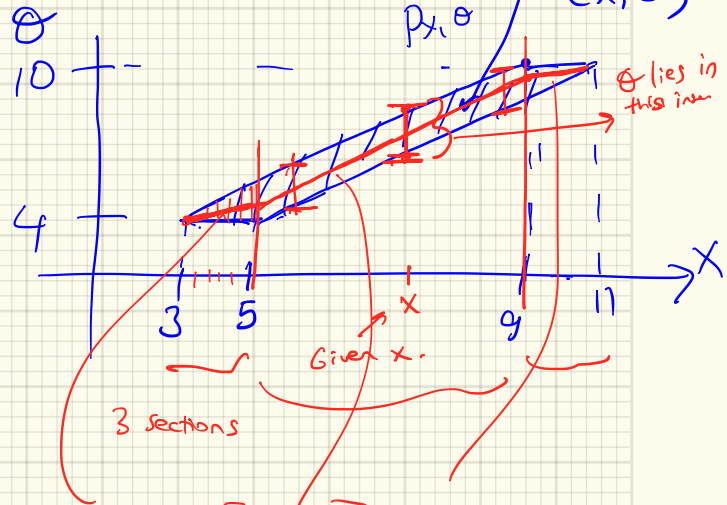
$X \in U(\Theta-1, \Theta+1)$



construct the joint density $p_{X,\Theta}$

$p_{X,\Theta} = p_\Theta \cdot p_{X|\Theta}$

$= \frac{1}{6} \cdot \frac{1}{2} = \frac{1}{12}$, on the support (X,Θ)



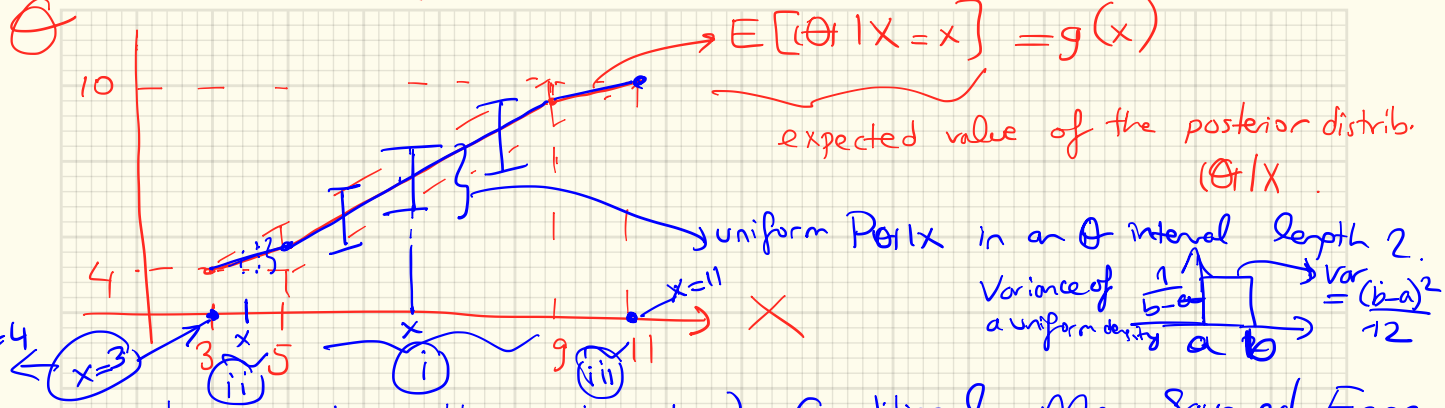
Q. A plausible estimator.

"optimal" estimator in LMS-sense

$E[\Theta|X]$; $p_{\Theta|X}$ →

$E[\Theta|X]$

LMS_estimator.



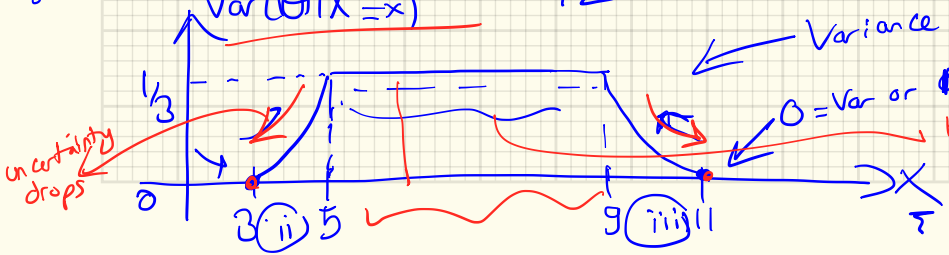
How good is this estimate? Conditional Mean Squared Error.

$$E\left[\left(\Theta - E[\Theta|X]\right)^2 \,\middle|\, X=x\right]$$: variance of the posterior distrib.

Θ: true value of Θ; $E[\Theta|X]$: estimate we made

Region (i): $\text{Var}(\Theta|X) = \frac{2^2}{12} = \frac{1}{3}$, for $5 \le x < 9$

$\text{Var}(\Theta|X=x)$



Variance ~ uncertainty in the estimator

0 = Var or No uncertainty for x=11 or 3.

Highest uncertainty.

Uncertainty in the estimate depends on the observation we make.