12.12.2022

# YZV 231E

## Probability Theory & Stats

Week 12

Gü.

# Recap: Limit Theorems

**WLLN**: Large # i.i.d. r.v.s

$X_1, \ldots, X_n$ : like sampling from a population.

Convergence in prob.

$$P(|X_n - a| \geq \varepsilon) \underset{n\to\infty}{=} 0.$$

$( X_1, X_2, \ldots X_n \xrightarrow[n\to\infty]{} )$

↳ "close" to $a \mp \varepsilon$

r.v.

$$\boxed{M_n} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

: estimate of the expected value.

sample mean

**WLLN**: $M_n \xrightarrow[\text{in probability}]{\text{convergence}} E[X] = \mu$

Recall Tchebyshev Ineq.

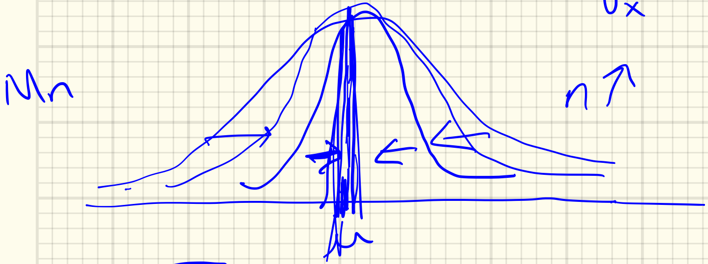
$\mu, \sigma^2$

$$P(|X - \mu| \geq c) \leq \frac{\sigma_X^2}{c^2}$$

$$\rightarrow M_n = \left(\frac{X_1+X_2+\dots+X_n}{n}\right)$$

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

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



Polling ex: p : fraction of population that prefer mt.

$$X_i = \begin{cases} 1, & \text{if yes (prefer)} \\ 0, & \text{if no (does not prefer)} \end{cases}$$

Bernoulli:

$$\underbrace{M_n}_{} = \frac{X_1+\dots X_n}{n}$$ : prediction of the fraction p.

Goal: $P(|\underbrace{M_n}_{} - \overset{\uparrow \text{true fraction}}{p}| \geq \overset{\text{accuracy}}{0.01}) < \boxed{0.05}$ ≡95% confidence

1 − confidence

# Types of Polling Problems : Specify specs

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

i) Collect data (n), Calculate sample mean $M_n$ & check whether we satisfy the specs

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

---

Check the polling ex. from last time :

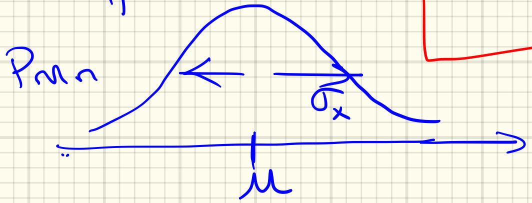
→ n = 50K people/samples were needed to satisfy the specs.

not practical.

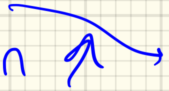
→ 1-2 K, samples → you need to change your specs.

# ☆ Different Scalings of $X_1+X_2+\cdots+X_n$, $X_i$ i.i.d. w/ $\mu, \sigma^2$.

i) Sample Mean r.V. $\boxed{\text{Scale} = \frac{1}{n}}$ sum r.v. $\rightarrow M_n = \frac{X_1+\cdots+X_n}{n}$

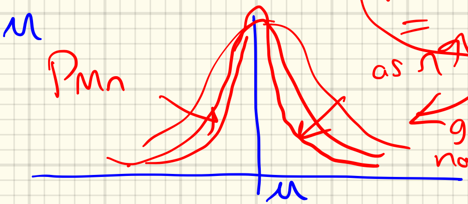


$E[M_n]=\mu$



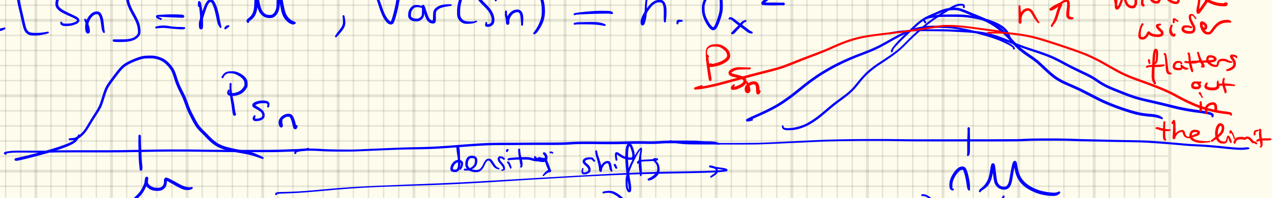
$$Var(M_n) = \frac{\sigma_x^2}{n}$$

as n ⇑ → gets narrower & narrower



ii) Scale = 1 Sum r.V. $S_n = X_1+\cdots+X_n$ ; $X_i$ : i.i.d.

$E[S_n] = n.\mu$ , $Var(S_n) = n.\sigma_x^2$

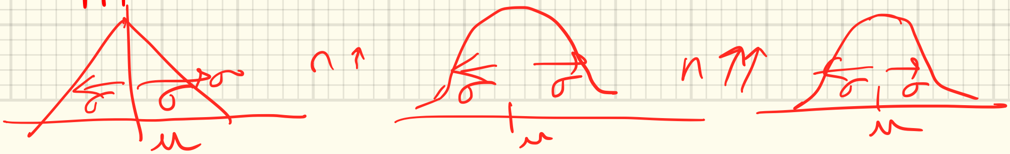


distrib. wider & wider flattens out in the limit

iii) Scale $= \frac{1}{\sqrt{n}}$ $\quad \frac{S_n}{\sqrt{n}} \rightarrow Var\left(\frac{S_n}{\sqrt{n}}\right) = Var\left(\frac{X_1+\cdots+X_n}{\sqrt{n}}\right) = \left(\frac{1}{\sqrt{n}}\right)^2 n.\sigma_x^2 = \sigma_x^2$

constant variance.

shape of the distrib. may change



distrib. width stays the same!

Different scalings of $S_n$ : $X_1, X_2, \dots, X_n$ i.i.d. w/ mean $\mu$, var $\sigma^2$.

3 variants : 1) $S_n = X_1 + \dots + X_n$ : Variance $n\sigma^2 \to \infty$.

2) $M_n = \frac{S_n}{n}$ : Variance $\frac{\sigma^2}{n} \to 0$

3) $\frac{S_n}{\sqrt{n}}$ : Constant Variance $\sigma^2$ ✓

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

$$Z_n = \frac{S_n - E[S_n]}{\sigma_{S_n}} = \frac{S_n - n.E[X]}{\sqrt{\underbrace{n\sigma_x^2}_{=\sigma_{S_n}^2}}} = \frac{S_n - nE[X]}{\sqrt{n}.\sigma}$$

$\sigma_{S_n}$ : standard deviation

This is called standardization of an r.v.: → Zero mean, unit variance.

Now
$$\begin{cases} E[Z_n] = 0 \\ Var(Z_n) = 1 \end{cases}$$

Compare $Z_n$ to a standard Normal r.v.,

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

Let $Z$ be a standard normal r.v.

# CENTRAL LIMIT THEOREM (CLT) : statement about CDFs.

We have a large # i.i.d. r.v.s $X_i$'s ( $X_i$'s could be any r.v., any distribution! )

→ Sum them $(S_n)$ → standardize them to $Z_n$

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

For every c,

$$\underbrace{P(Z_n \le c)}_{\text{cdf of } Z_n} \longrightarrow \underbrace{P(Z \le c) = \Phi(c)}_{\text{standard Normal}}$$

≡ Gaussian w/ 0 mean & unit var

useful b/c we have CDF of std Normal available from tables.

Given

$X_1, X_2, \dots, X_n$ → normalized $\sum X_i$ distrib. (CDF) → Normal Distrib. CDF

$X_i$'s w/ mean $\mu$, var : $\sigma$ finite.

$$Z_n = \frac{S_n - nE[X]}{\sqrt{n}\,\sigma} \longrightarrow S_n = \sqrt{n}\,\sigma Z_n + n.E[X]$$

$S_n$ is Normal r.v.

$Z_n \sim$ Normal when n is large

Exercise Check why $\underline{S_n}$ is also a normal r.v.

(b/c $S_n \leftarrow$ linear transformation of $\underline{\bar{Z}_n}$ is Normal r.v. when n is large enough.

we ~~assume~~ pretend

- CLT is a limit theorem : $n \nearrow \infty$.
- In practice, maybe $n = 30 \rightarrow$ gives accurate results.

---

Ex: Let $X_i \sim \mathcal{N}(0,1)$ i.i.d.

Define $\boxed{Y_N = \sum_{i=1}^{N} X_i^2}$

Approximate $Y_N$ distrib by a Gaussian.

Q. Is this justified?

Yes

b/c $X_i$ are i.i.d. w/ finite mean & var ✓

$\Rightarrow$ $X_i^2$ are i.i.d ✓

$$Var(Y_N) = N\,Var(X_i^2)$$

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

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

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

Kurtosis of $N(0,1)$ $E[X^4] = 3$ (check derived ✓ known ✓)

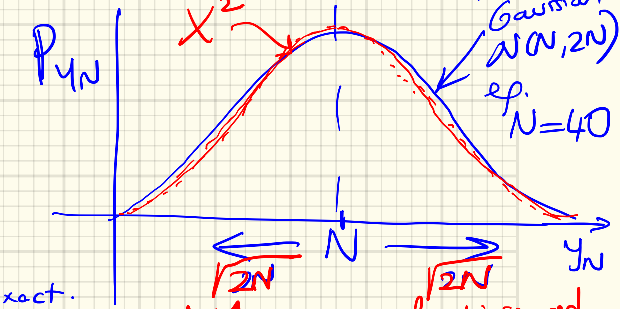
$$\tilde{Y}_N = \frac{Y_N - N\cdot 1}{\sqrt{2N}} \approx \underbrace{\mathcal{N}(0,1)}_{\text{as } n\to\infty} \text{ by CLT.}$$

$$\Rightarrow \underbrace{Y_N}_{\text{Gaussian ? yes}} = \sqrt{2N}\,\underbrace{\tilde{Y}_N}_{\text{Gaussian} \to \text{w/ mean}=0,\ \text{std}=1} + N \to Y_N \sim \mathcal{N}(N, 2N)$$

$$E[Y_N] = \sqrt{2N}\,\cancel{E(\tilde{Y}_N)}^{\to 0} + N \quad ✓$$

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

$Y_N$ is: $\chi^2$: chisquared distrib. $Y_N \sim \chi^2_N$ exact.



as N↑ approx. of chi-squared by a Gaussian becomes better.

# Ex: Pollster's problem using CLT.

- p: fraction of the population that prefers "something"
- $X_i$: ith randomly selected person, $\underset{\text{Bernoulli}}{X_i} = \begin{cases} 1 & \text{if yes} \\ 0, & \text{if no.} \end{cases}$

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

- We define 2 specifications for the poll $\equiv$ $\underline{\text{2 parameters}}$

$$P(|M_n - \overset{\uparrow}{p}| \geq \underbrace{0.01}_{\text{accuracy}}) \leq 0.05 = 1 - \underbrace{\text{confidence}}_{95\%}$$

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

$\underline{\text{event of interest}}$: $|M_n - p| \geq 0.01$

rewrite

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

standardizing:

$$= \left| \boxed{\frac{X_1 + \dots + X_n - np}{\sqrt{n}\,\sigma}} \right| \geq \frac{0.01\sqrt{n}}{\sigma}$$

standardized r.v. $\triangleq Z_n$.

$$\sigma^2_{M_n} = \frac{\sigma^2}{n}$$

$$\boxed{\sigma_{M_n} = \frac{\sigma}{\sqrt{n}}}$$

divide by this.

result of the poll — the value

→ $P(|M_n - p| \geqslant 0.01) = P\left(|Z_n| \geqslant \frac{0.01\sqrt{n}}{\sigma}\right)$

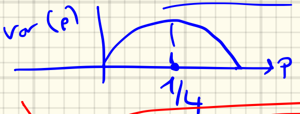
$\approx P\left(|Z| \geqslant \frac{0.01\sqrt{n}}{\sigma}\right)$

Using CLT, we calculate this prob

where Z is a standard normal r.v.: $Z \sim N(0,1)$

→ Note one difficulty: we don't know $\sigma$? But we know an upper bound

For Bernoulli: $Var(X) = p(1-p)$. Recall last time: $\sigma^2 \leqslant \frac{1}{4}$



→ $\sigma \leqslant \frac{1}{2}$ → we'll use this upper bound.

$= P\left(|Z| \geqslant \frac{0.01\sqrt{n}}{\sigma}\right) \leqslant P(|Z| \geqslant 0.02\sqrt{n})$

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

lowerbound.

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

$P(|Z| \geqslant \underbrace{0.02\sqrt{10^4}}_{2}) = P(|Z| \geqslant 2) = 2P(Z \geqslant 2)$

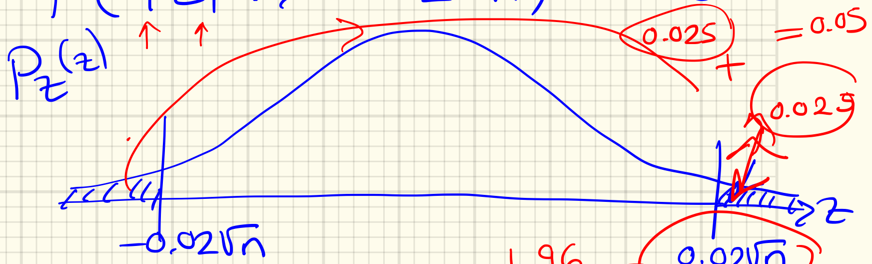
$= 2(1 - \underbrace{P(Z<2)}_{\Phi(2)}) = 2(1 - 0.9772)$

Use Normal Table

$= 0.0456$ → prob. of error $< 5\%$

$\simeq \%4.6$

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



$\underset{\sim}{1.96} = 0.02\sqrt{n}$

$\Phi(c) = 1 - 0.025$

$\Phi(c) = 0.975$

$\rightarrow c = 1.96$ from the table

$0.02\sqrt{n} = 1.96$

$\rightarrow n = 9604$

w/ $n = 9604$ people in the poll our prob. of error is 0.05

CLT: in polling

i) start w/ $n$ & calculate probabilities

ii) start w/ specs (probs) & calculate $n$.

we use CLT in different ways

① polling

② approximating distribution by a standard Gaussian, Ex: CLT: Apply to Binomial approx.

$X_i$: Bernoulli $(p)$, i.i.d. $0 < p < 1$.

$$S_n = X_1 + \cdots + X_n \equiv \text{Binomial}(n, p)$$

Binomial r.v.: mean : $np$

Variance : $np(1-p)$ ✓

CDF of

standardized → $$\boxed{\frac{S_n - np}{\sqrt{np(1-p)}}} \xrightarrow{\text{CLT}} \text{Std Normal Distrib.}$$

Check whether this approx is good

Let $n = 36$, $p = 0.5$, Find $P(S_n \le 21)$?

mean $= np = 18$

$\text{Var} = np(1-p) = 9 \to \sigma = 3$

$$\frac{S_n - 18}{3} \le \frac{21-18}{3} = 1$$

$$\boxed{Z_n \le 1}$$

CLT:

$$\approx Z \le 1$$

std normal.

$\rightarrow \Phi(1) = P(\cancel{Z} \le 1) = \boxed{0.843}$ approx. answer $\overset{?}{=}$

from the table.

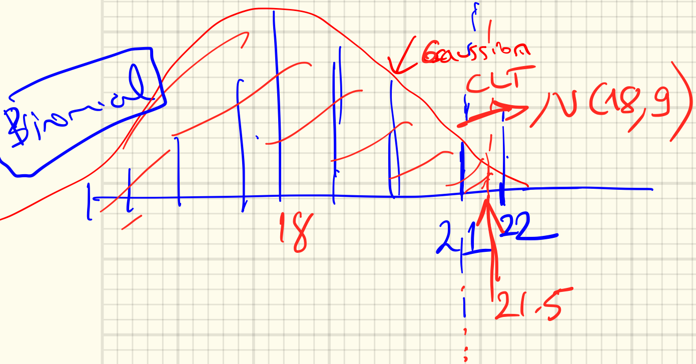
Exact answer : $\displaystyle\sum_{k=0}^{21} \underbrace{\binom{36}{k}\left(\frac{1}{2}\right)^{36}}_{P(S_n=k)} = \boxed{0.8785}$

ok but not good!

b/c $S_n$ is a discrete r.v.

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

$\therefore$ use $P(S_n \le 21.5)$ → called $\frac{1}{2}$ correction for Binomial approx.

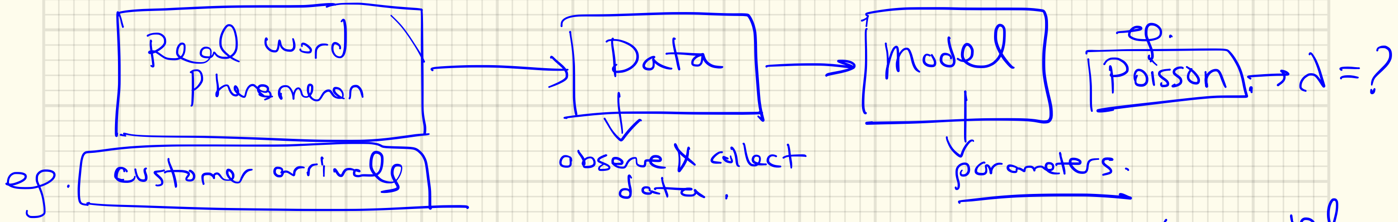


$$\frac{S_n - 18}{3} \le \frac{\boxed{21.5} - 18}{3} = 1.17$$

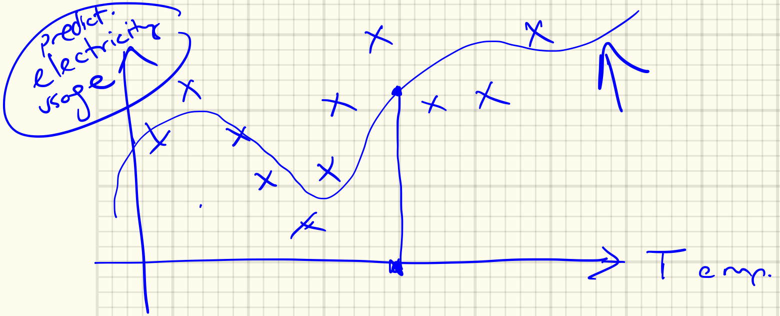
Table : $P(Z \le 1.17) = \boxed{0.879}$

a much better approx. w/ the $\frac{1}{2}$ correction.

# STATISTICAL INFERENCE ~ Applied Probability.



- come up w/ a model
- calculate its parameters
- make predictions about the real world.

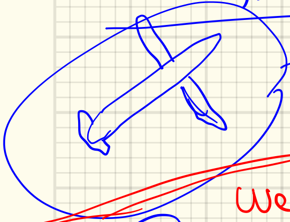


- Polling Problems: estimation of preferences of populations
- Finance :

1) Bayesian Statistical Inference
2) Classical Statistical Inference.

Two types of problems

(i) Hypothesis Testing: discrete quantities.



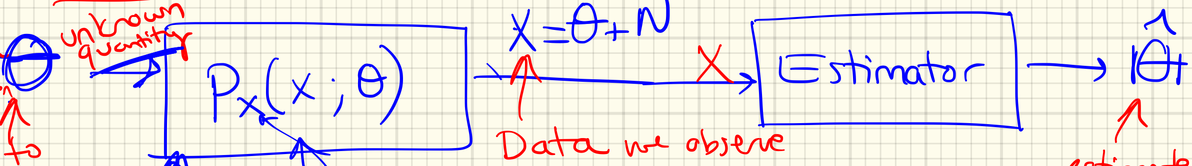
Radar. measurement → Detect. a signal or not → interested in prob-of error.

1/0.

We start w/ problem

(ii) Estimation! Want to measure a voltage $\theta$, in $(-V,V)$ volts

1st approach to Estimation



$\hat{\theta}$ is an r.v.

we try to estimate $\theta$

Data we observe

estimate of the unknown $\theta$.

Measurement device.

N: noise

observations

X: r.v.

In Estimation unknown is a continuous quantity

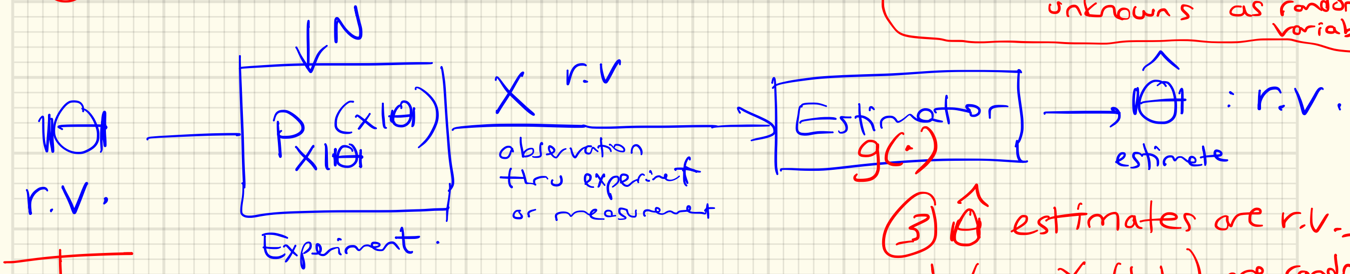
① Classical Statistics approach

$\theta$: unknown number: not a r.v.!

# ② Bayesian Statistics Approach:

Note ① Classical stat. treats unknowns as unknown numbers

② Bayesian Stats treats unknowns as random variables



$P_{\Theta}(\theta)$ : prior distribution:

☆ our initial belief about θ before the experiment

③ $\hat{\Theta}$ estimates are r.v.s. b/c. X (data) are random variables

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

Recall: functions of r.v.s are r.v.s.

We rely on Bayes rule → come up w/ a posterior distrib.

$P_{\Theta|X}(\theta|x)$ : our revised beliefs once we obtain data X, i.e. after the experiments

☆ prior → posterior ☆

# Bayesian Inference

$$\underbrace{P_{\Theta|X}(\theta|x)} = \frac{\overbrace{P_{X|\Theta}(x|\theta)}^{\text{conditional model of the experimental process = likelihood of observing the data}} \cdot \overbrace{P_{\Theta}(\theta)}^{\text{prior: initial belief on } \theta \text{ before the experiment}}}{P_X(x)}$$

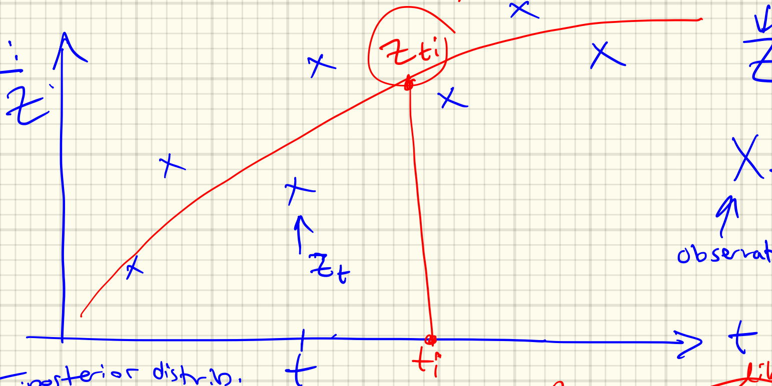
$P_X(x)$ ; evidence : can be calculated based on likelihood

$$\int_{\{\theta\}} \underbrace{P_{X|\theta}(x|\theta)\, p_{\theta}(\theta)}\, d\theta$$

valid for $P$ : pmfs or pdfs ✓
both discrete & continous r.v.s.

Want to calculated posterior distribution of $\Theta$ given $X$.

Ex:



↓position of the object.

$$Z_t = \Theta_0 + \Theta_1 t + \Theta_2 t^2$$

$$\underline{\Theta} = [\theta_0\ \theta_1\ \theta_2]^T$$

$$X_t = Z_t + W_t$$

↑ observations ↑ noise.

(measurements we make are noisy).

**Goal:** Estimate $\boxed{\Theta_0, \Theta_1, \Theta_2,}$

Goal calculate posterior distrib.

$$p_{\Theta_0,\Theta_1,\Theta_2|X_1,\dots,X_n}(\Theta_0, \Theta_1, \Theta_2 | X_1, \dots, X_n) \overset{\text{Bayes}}{=} \frac{\overbrace{P(X_1, \dots, X_n|\underline{\Theta})}^{\text{likelihood.}} P_{\Theta}(\Theta)}{P(X_1, \dots, X_n)}$$

prior is given
continuous r.v. (pdfs.)

Ex: Coin w/ unknown parameter $\Theta$ $(\overset{eg.}{=} p)$.

Estimate $\Theta$.

→ Classical statistics approach: Flip the coin many times $X_1, X_2, \ldots, X_n$ → form $S_n$ → sum $X_i$'s

$$\hat{\Theta}_m = \frac{S_n}{n} \xrightarrow{WLLN} \theta$$

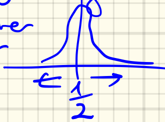
$S_n$ → how many heads in n tosses

$n$ → no. of trials

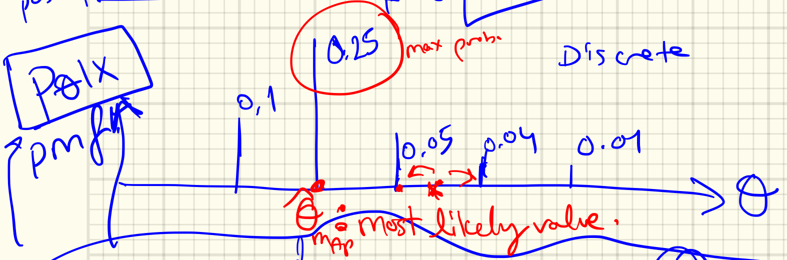
→ Bayesian approach: Assume a prior on $\Theta$. $\boxed{P_\Theta(\theta)}$

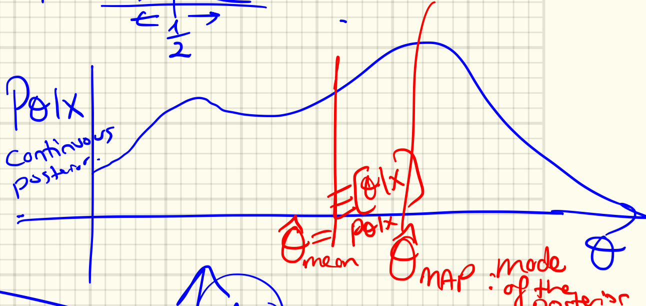
(eg. uniform prior if you don't know anything about $\theta$).

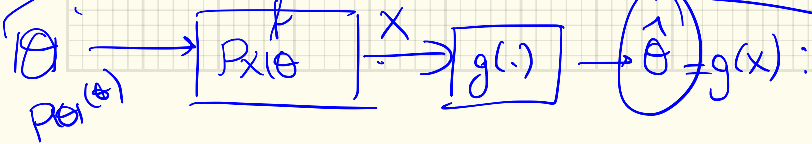
(eg. another prior $N(0.5, \sigma)$)



calculate this posterior $\boxed{P_{\Theta|X}(\theta|x)} = ?$ w/ Bayes rule.



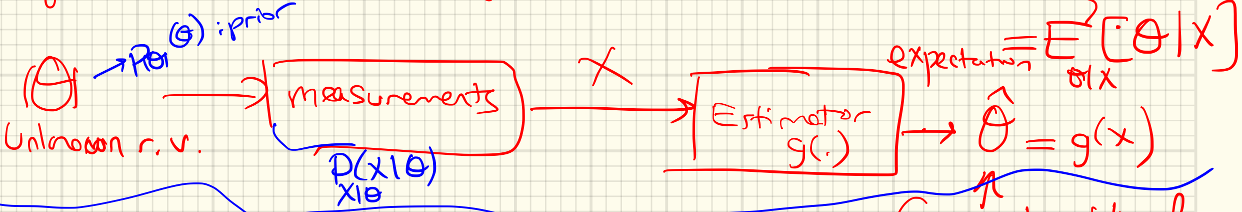




$\hat{\theta} = g(X)$: pick a $\hat{\theta}$ from $P_{\Theta|X}$ density.

→ Output of Bayesian inference : $P_{\Theta|X}(\theta|x)$ ; whole distribution, posterior

→ If interested in a single output : $\hat{\theta}_{MAP}$ or $\hat{\theta}_{MLE}$ :



can be picked from the posterior.

Once we estimate the posterior distrib $P_{\Theta|X}(\theta|x)$ : complete answer to the Bayesian inference problem.

point estimates of $\theta$

① MAP : (Maximum a Posteriori Estimate)

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

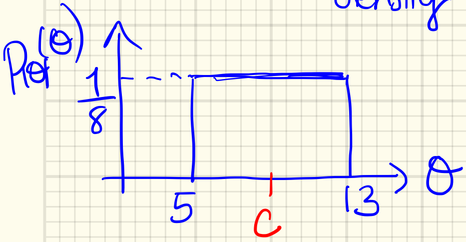
② Conditional Expectation (Least mean Squared Error LMS estimate)

$$E[\Theta|X=x] = \int \theta \, P_{\Theta|X}(\theta|x)\, d\theta$$ : average of the posterior distrib.

# Least Mean-Squared Estimation (LMS) :

Given a prior density on $\Theta$ :

Goal : come up w/ a point estimate using LMS .



(w/ no observed x) we minimize LMS criterion:

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

point estimate

: This is called the Least mean-Squared Estim.

What is "optimal" estimate c ?

in LMS sense

$$E[\Theta^2] - 2E(\Theta)\cdot c + c^2$$

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

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

$$\rightarrow c = E[\Theta] = 9$$ for this specific example.

To judge how good is our estimate ?

"optimal" mean-squared error:

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

★ When we're minimizing the least mean-squared error, Expectation is the "best" estimate ⟶

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

LMS estimation of $\Theta$ based on $X$ : Given $X$ r.v.

$\Theta$, $X$ are r.v.s

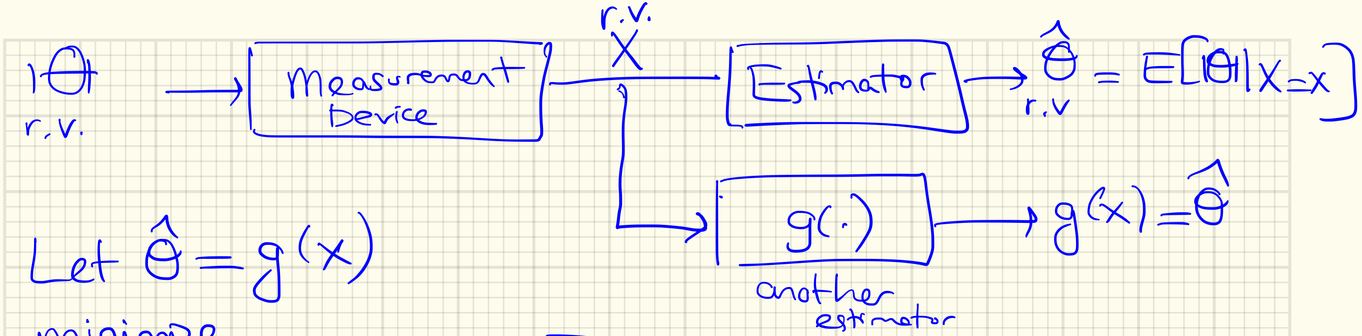
Given $X=x$, evaluate in a conditional universe.

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

is minimized by

$$\boxed{c = E[\Theta \mid X = x]}$$

: now we use conditional expectation (exp. of the posterior density).



Let $\hat{\Theta} = g(x)$

minimize

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

$$= \int_{-\infty}^{\infty} \underbrace{\left(\int_{-\infty}^{\infty} (\theta - \underbrace{g(x)}_{c})^2 p_{\Theta|X}(\theta|x)\,d\theta\right)}\ \underbrace{p_X(x)}_{p_X(x) \geq 0}\,dx \quad ✓$$

We can minimize

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

unknown.

$c = g(x)$

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

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

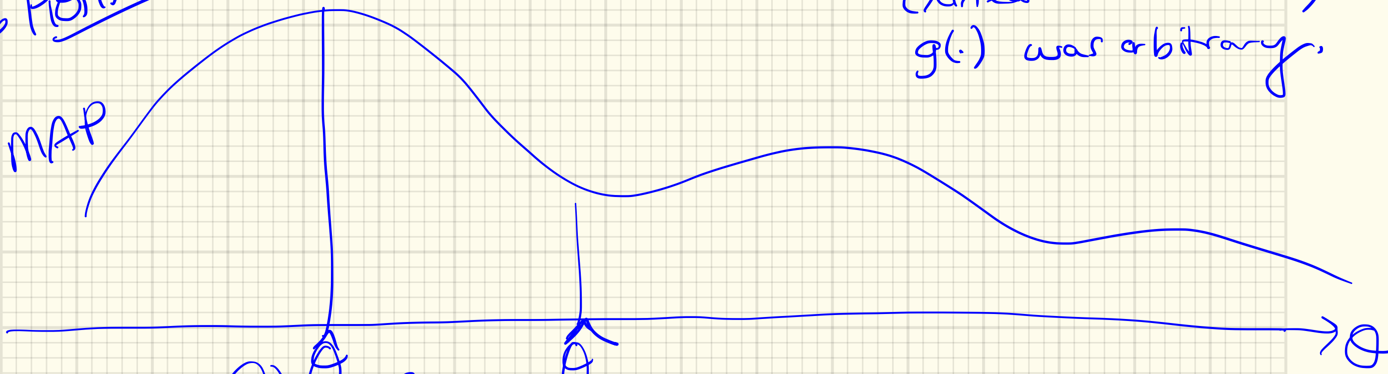
$$\rightarrow \int \theta\, p_{\Theta|X}(\theta|x)\,d\theta = c \cdot \underbrace{\int_{-\infty}^{\infty} p_{\Theta|X}(\theta|x)\,d\theta}_{1} \Rightarrow$$

$$\underset{\tilde{g}(x)}{c} = E_{\Theta|X}[\Theta | X] = \int \theta \, P_{\Theta|X}(\theta|x) \, d\theta$$

$\rightarrow \hat{\Theta}(x) = E[\Theta|X]$ : is the "optimal" estimator (in sense of LMSE) among all estimators (linear or nonlinear) g(.) was arbitrary,

posterior distrib $P_{\Theta|X}(\theta|x)$ : complete answer to a Bayesian inference.

① MAP



① $\hat{\theta}_{MAP}$

② $\hat{\theta}_{LMSE} = E[\Theta|X]$ : conditional mean of the posterior density

Issues

- Unknown $\underline{\Theta}$ : vector of $\Theta$'s.
- Observations $\underline{X} = (X_1, \dots, X_n)$ ← several measurements

Calculate w/ Bayes rule

$P_{\underline{\Theta} | X_1, \dots, X_n}(\underline{\theta} | x_1, \dots x_n) \rightarrow$ LMSE $\underline{E[\Theta | X_1, \dots X_n]}$

$$\rightarrow P_{\Theta | X_1, X_2, \dots X_n} = \frac{P_{X_1, \dots, X_n | \Theta} \quad P_{\underline{\Theta}}}{P_{X_1, \dots X_n}}$$

$$P_{X_1, \dots X_n} := \iiiint P_{X_1, \dots x_n | \Theta} \, p_{\underline{\theta}}(\theta) \, d\underline{\theta}$$

① → Calculations may become intractable ← multi-dimensional integrals

$p_{\theta | X}$ & its expectation

② Come up w/ a plausible prior

# Ex 8.11 from Bertsekas book.

## Linear LMS estimator:

Now, consider a simpler estimator of $\Theta$ !

Let $g(X) = \boxed{\hat{\Theta}_L = aX + b}$ : affine mapping w/ parameters a & b unknown.

minimize $E\left[(\Theta - \underbrace{\hat{\Theta}_L}_{})^2\right]$

$$= E\left[(\Theta - \underbrace{(aX+b)}_{g(X)})^2\right]$$

$\xrightarrow{}$ generic estimator $\quad g(X) = E[\Theta | X] = \hat{\Theta}_{LMS}$

exercise Derive:

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

minimize w.r.t a & b.

$$\left\{E[\Theta^2] + a^2E[X^2] + 2abE[X] + b^2 - 2aE[\Theta \cdot X] - 2bE[\Theta]\right\}$$

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

$$a = \frac{\mathrm{cov}(X,\theta)}{\mathrm{var}(X)}, \quad b = E[\theta] - \frac{\mathrm{cov}(X,\theta)}{\mathrm{var}(X)} E[X]$$

"Best" choice of a & b w.r.t. LMS criterion:

$$\boxed{\underbrace{\hat{\theta}_L}_{} = E[\theta] + \frac{\mathrm{cov}(X,\theta)}{\mathrm{var}(X)} \cdot (X - E[X])}$$

Linear LMS estimator of $\theta$.