f : fraction of population that like pepsi over coke

$$x_i = \begin{cases} 1, & \text{yes} \\ 0, & \text{No} \end{cases}$$

* $m_n = (x_1 + \dots + x_n)/n$

* we want : $P(|m_n - f| \geq 0.1) \leq 0.05$ ↳ confidence

↓

Specification

(Accuracy)

Event of interest: $|m_n - f| \geq 0.01$

$$\left| \frac{x_1 + \dots + x_n - nf}{n} \right| \geq 0.1$$

$$\left| \frac{x_1 + \dots + x_n - nf}{\sqrt{n}\,\sigma} \right| \geq \frac{0.01\sqrt{n}}{\sigma} \rightarrow \text{Standardize}$$

$= |m_1 - f| + |m_2 - f| \dots$

$= |m_1 + m_2 + \dots - nf|$

$$\boxed{\begin{aligned} \sigma &= \sqrt{f(1-f)} \\ &\leq 1/4 \\ &\text{(max case)} \end{aligned}}$$

$$P(|m_n - f| \geq 0.01) \approx P(|Z| \geq 0.01\sqrt{n}/\sigma)$$

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

↓

$$P\left(\frac{|m_n - f|}{\sigma} \geq \frac{0.01}{\sigma}\right) = P\left(\left|\frac{x_1 + x_2 + \dots + x_n - nf}{\sigma\sqrt{n}}\right| \geq \frac{0.01\sqrt{n}}{\sigma}\right)$$

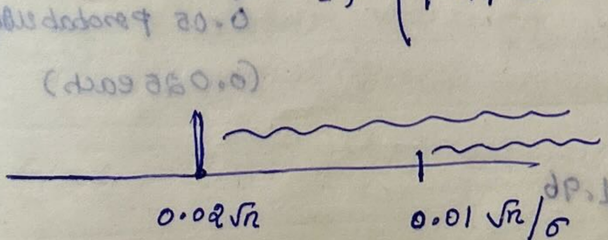
$$= P\left(|Z_n| \geq 0.01\sqrt{n}/\sigma\right)$$

$\sigma = 1/2$

$[\sigma^2 = 1/4]$

↓

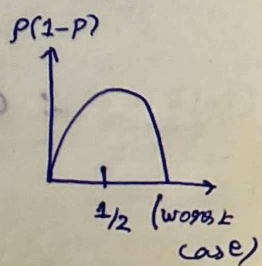
Worst Case



↓

$\sigma = 1/2$

$$\boxed{\therefore \sigma \leq \frac{1}{2}, \quad \frac{1}{\sigma} \geq 2}$$



Prob of falling b/w $0.02\sqrt{n}$ to ∞ will be larger than falling b/w $0.01\sqrt{n}/\sigma$

$$\boxed{\therefore \quad \sigma \le 2, \quad \frac{1}{\sigma} \ge 2 \rightarrow \text{minimum}}$$

∴ But σ may be anything.

Normal tables

$n = 10000$

$$P(|Z| \ge 0.02 \times \sqrt{10000}) = P(|Z| \ge 2)$$

$$= \text{'Normal'}$$

$$= 2\, P(Z \ge 2)$$

$$= 2\,(1 - P(Z \le 2))$$

using CDF of z we can

$$= 2\,(1 - 0.9772)$$

$$= 0.0456$$

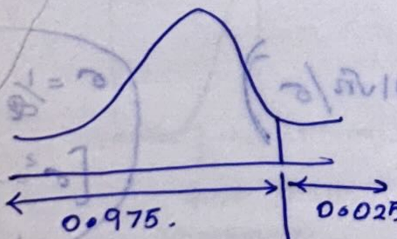
∴ our error is 0.0456 (allowed: 0.1)

'we can go for smaller sample size'

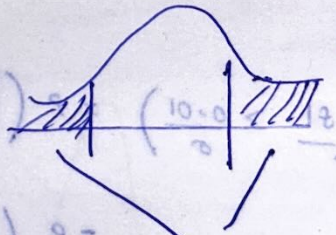
$$P(|Z| \ge 0.02\sqrt{n}) = 0.05$$

↓

Std. normal.



value ⇒ 1.96



$$\therefore 0.02\sqrt{n} = 1.96$$

$$\sqrt{n} = \frac{1.96}{0.02}$$

$$\sqrt{n} = \frac{196}{2}$$

$$\sqrt{n} = 98$$

$$\boxed{n = 9604}$$

R.V → Bernoulli (Here we dealt)

Sum of Bernoulli → Binomial.

Central limit theorem - applied to standardized version of Binomial

---

* $S_n = X_1 + \dots + X_n$

mean $= np$, variance $= np(1-p)$

CDF of $\dfrac{S_n - np}{\sqrt{np(1-p)}}$ → standard normal.

<u>Example</u>

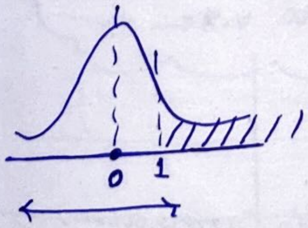
$n = 36$, $p = 0.5$, $P(S_n \le 21)$

$$P\left(\frac{S_n - 36\times0.5}{\sqrt{0.5\times36(1-0.5)}} \le \frac{21 - 36\times0.5}{\sqrt{0.5\times36(1-0.5)}}\right)$$

$$P\left(\frac{S_n - 18}{\sqrt{9}} \le \frac{3}{\sqrt{9}}\right)$$

$$P\left(\frac{S_n - 18}{3} \le 1\right) \approx P(Z \le 1) = 1 - P(Z = 0)$$

$$= 0.843$$



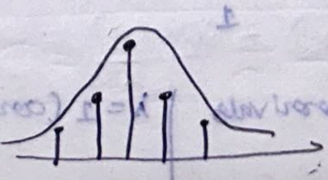
<u>Exact answer:</u>

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

Idea: $P(S_n \le 22) = P(S_n < 22)$

$\therefore S_n \to$ integer

<u>Compromise</u>: Consider $P(S_n \le 21.5)$



$S_n \to$ discrete [CDF $\to$ continuous (Area) $\to$ Don't include 22]

Instead of $22 \to 21.5 \to$ Better numerical results.

$$P(Z_n \leq 1.17) = P(Z \leq 1.17)$$

$$= 0.879 \longrightarrow \text{close to summation answer.}$$

'1/2' correction $\to$ Better approximation.

$\therefore$ we can use it for individual prob:

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

$$= P(0.17 \leq Z_n \leq 0.5)$$

First central limit theorem was proved for $p = 1/2$ for binomials

$$= P(Z \leq 0.5) - P(Z \leq 0.17)$$

$$= 0.6915 - 0.5675$$

$$= 0.124.$$

$\binom{36}{19}\left(\frac{1}{2}\right)^{36} = 0.1251) \to$ Exact formula

---

For CLT $\to$ For binomial distribution (with 1/2 correction) $\to$ is even for individual PMF.

(Not in other case $\to$ CDF alone)

Puzzle

Poisson process

$0 \quad\quad\quad\quad 1$

$X =$ No. of arrivals | $\lambda = 1$ (arrival rate)

mean $= 1$

Seperate : n channels $\left(\frac{p}{n}\right)$

when $n \to \infty \to$ appr $\to$ Bernoulli

X_i's are Independent

$X = X_1 + \dots + X_n$

Appr CLT $\to$ Sum of ind, identically dist $\to$ R.V $\to$ lot of R.V $\to$ behaves like normal r.v.

$\therefore$ X is a random [By gn, X is a poisson]

Central mean theorem

1) IID

2) Finite mean & variance

3) consider : Fix a prob distri $\to$ Let X_i's be distributed according to that prob distribution & add larger & larger no. of X_i's. $\to$ we fix distribution of X_i.

In CLT $\to$ we fix distribution of X_i

As $n \to \infty$, the statistics of each X_i don't change.

In larger n,
$X_i \to$ R.V are with different mean & variance.
↓
poisson (don't suit for poisson)

In Poisson $\to$ Distributions may vary (mean & var too)

↓

Against CLT

↓

CLT doesn't apply.

Solution:

p fixed, $n \to \infty$: normal

$p = 1/100$, $n = 500$: normal $\longrightarrow$ use CLT

np fixed, $n \to \infty$, $p \to 0$: poisson

$p = 1/100$, $n = 100$: poisson. $\longrightarrow$ use poisson

$\therefore$ Large numbers of interval (small time) $\to$ prob of success very low

$\boxed{\text{As } n \to \infty,\ P \to 0}$

Two cases of binomial

$\swarrow$ poisson $\searrow$ CLT

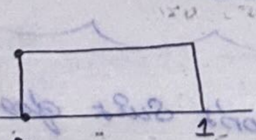
Recitation

$X_1 \sim X_{10} \sim \text{Uniform}[0, 1]$

markov : $P(X \geq a) \leq \frac{E[X]}{a}$

$P\left(\sum_{i=1}^{10} X_i \geq 7\right) \to$ Give bound

solve:



$E[X_i] = \frac{a+b}{2} = \frac{1}{2}$

$X = \sum_{i=1}^{10} X_i$

$E[X] = 10\, E[X_i] = 10\left(\frac{1}{2}\right) = 5$

$P(X \geq 7) = \frac{E[X]}{7} = \frac{5}{7}$

b) Chebyshev : $P(|X - E[X]| \geq a) \leq \frac{Var(X)}{a^2}$

$P(|X - E[X]| \geq a) \leq \frac{var(X)}{a^2}$

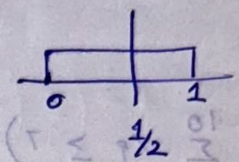
$P(|X-5| \geq 2) \leq \frac{var(X)}{a^2}$ $\quad [\, P(X \geq 7) = P(X-5 \geq 2)$

$\hookrightarrow$ transformed.

$\boxed{7 = 5+2}$

$P(X-5 \geq 2)$ or $P(X-5 \leq -2)$

$X \to$ +ve & negative.



$\cancel{P(X \geq 2)}$ $\quad$ $\cancel{P(X \geq 7)}$

$\cancel{P(X-5 \geq 2)}$ $\quad$ $P(-X \geq -7)$

$P(-X-$

$-2 \qquad 2$

$\therefore \cancel{P(|X-5| \geq 2)}$

<u>Two cases</u>

$P(X-5 \geq 2) = P(X-5 \leq -2)$

$P(|X-5| \geq 2) = P(X-5 \geq 2) + P(X-5 \leq 2)$

Both are equal

$\therefore P(X \geq 7) = \frac{1}{2} P(|X-5| \geq 2) \leq \frac{1}{2} \cdot \frac{var(X)}{a^2}$

$\leq \frac{1}{2} \cdot \left(\frac{var(X)}{2^2}\right)$

$\leq \frac{1}{8} \left(10 \cdot \frac{1}{12}\right)$

$\leq \frac{5}{48}$

$\sigma = \sqrt{\frac{(b-a)^2}{12}}$

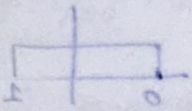
$\boxed{\sigma^2 = \frac{(b-a)^2}{12}}$

$\therefore$ using chebyshev we got stronger upper bound

$$\left(\frac{\sum_{i=1}^{n} x_i - \text{mean}}{\text{Standard deviation}}\right)_{n\to\infty} = \text{CLT}. \quad \left[N \to (0,1)\right]$$

looklike as $n \to \infty$

$\boxed{n=10}$



$$P\left(\sum_{i=1}^{10} x_i \ge 7\right) = 1 - P\left(\sum_{i=1}^{10} x_i \le 7\right)$$

$$= 1 - P\left(\frac{\sum_{i=1}^{10} x_i - 5}{\sqrt{\frac{10}{12}}} \le \frac{7-5}{\sqrt{\frac{10}{12}}}\right)$$

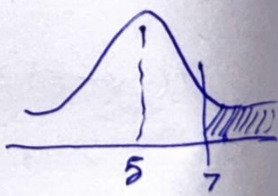
$$= 1 - P\left(Z_n \le \frac{2}{\sqrt{\frac{10}{12}}}\right)$$

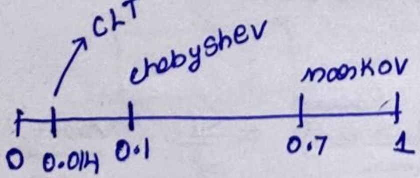
$$= 1 - \left(Z_n \le 2.19\right)$$

'If we beleive $10 = n$ is large enough'

$$= 1 - \Phi(2.19)$$

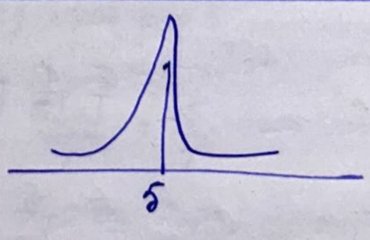
$$= 0.014. = \frac{7}{500}$$





Prob$(X \ge 7)$

Reality:



AS: $n=10$ concentration much higher on 5. so $P(X \ge 7) \to$ low

CLT

$X \to$ # gadgets produced on day n

$X_n \sim N(\mu = 5, \sigma^2 = 9)$

X_n · iid

a) $P\left(\text{tot gadgets in 100 days} < 440\right) = ?$

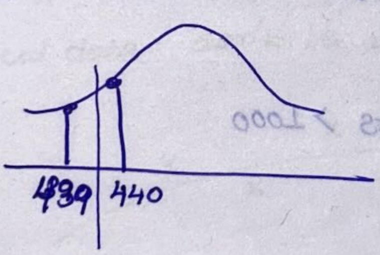
$$= P\left(\frac{\sum_{i=1}^{100} X_i - 500}{\sqrt{900}} < \frac{440-500}{\sqrt{900}}\right)$$

$$= P\left(Z_n < -2\right)$$

$$E\left[\sum_{i=1}^{n} X_i\right] = n\,E[X_i]$$
$$= 100\,E[X_i]$$
$$= 100 \times 5 \quad [5n]$$
$$= 500$$

$$\operatorname{var}\left[\sum_{i=1}^{n} X_i\right] = n \operatorname{var}(X_i)$$
$$= 100 \times 9$$
$$= 900.$$
$$[9n]$$

∴ our problem: discrete



$\frac{1}{2}$ approximation

$$= P\left(Z_n < \frac{439.5 - 500}{\sqrt{900}}\right)$$

$$= P(Z_n < -2.016)$$

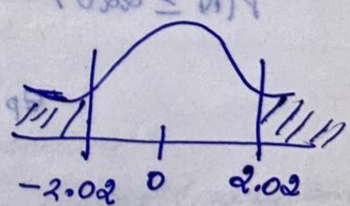
$= P(Z_n < -2.02)$

$= P(Z_n > 2.02)$

$= 1 - \Phi(2.02)$

$= 1 - 0.9783$

$= 0.0217.$



Both area equal.

b) appr: largest n $P\left(\sum_{i=1}^{n} X_i \geq 200 + 5n\right) \leq 0.05$

solu:

$$P\left(\sum_{i=1}^{n} X_i \geq 200 + 5n\right)$$

$$= P\left(Z_n \geq \frac{200+5n - 5n}{\sqrt{9n}} \right)$$

$$\approx P\left(Z_n \geq \frac{200}{3\sqrt{n}} \right) \leq 0.05$$

$$\approx P\left(Z_n \geq \frac{66.67}{\sqrt{n}} \right) \leq 0.05$$

$$\approx 1 - \left(P\left(Z_n \leq \frac{200}{3\sqrt{n}} \right) \right) \leq 0.05$$

$$P\left(Z_n \leq \frac{200}{3\sqrt{n}} \right) \geq 0.95$$

only when

$$\frac{200}{\sqrt{9n}} \geq 1.65$$

when $n \leq \frac{1}{9}\left(\frac{200}{1.65} \right)^2$

c) N = 1st day when total gadets > 1000

$P(N \geq 220) = ?$

$$P(N \geq 220) = P\left(\sum_{i=1}^{219} X_i \leq 1000 \right)$$

up to 219 it has not exceeded 1000

$$= P\left(Z_n \geq \frac{1000 - 219 \times 5}{\sqrt{9 \times 219}} \right)$$

$$= P(Z_n \geq -2.1398)$$

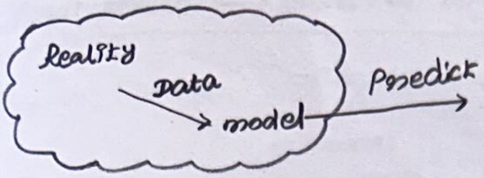
$$\approx 1 - P(Z_1 \leq -2.14)$$

↳ same area bothsides

$$= 1 - P(Z_1 \leq 2.14)$$

$$\approx 1 - 0.9838$$

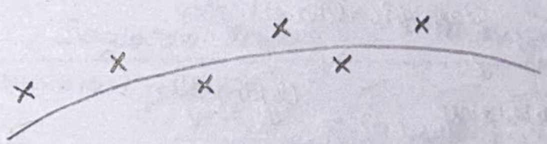
Bayesian Statistical Inference - I



eg: polling, medical trials

* 'we can make prediction and inferences'. (preference)
* 'New drug effective or not'
* 'Netflix Competition' → 'recommendation system'
  ↓
  * 'movies & people - people rated but may not watched'
  * 'relevant - watch & rated (similar ratings of people)'
* financial data: S&P, NIFTY index. [model → predict]
* signal processing: Track, detect, Speaker identification.

Astronomical data: don't be fit → curve fitting

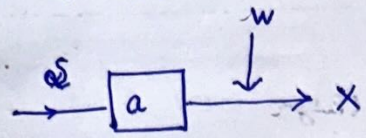


Relation b/w app & probability

* Statistical inference: Applied exercise of probability
* Probability problem → No ambiguity, unique solution
* Data → Best way estimating of motion of a planet (many come with many models): Based on experiments, we can tell
  ↓
  one can come with really bad models (possible)
* 'often Conclusions people made' → false
* Instead of just copying models, test → we need to understand 'assumptions of a test & guarantees they have if any.

* model building (vs) Inferring unknown variables.

$X = aS + W$



$$\boxed{X = aS + W}$$

* Inference: we don't know about a (attenuation)
* Design a certain signal S, observe X → make a guess about a
* System identification.

"we know S, a & X" → guess noise [reconstruct X]

"model of medium which through our signal is propagating".

"Noise Cancellation"

[Reconstruct noisy version of X]

* Same math in both cases: different approach.

Classification

↙ ↘

Discrete

eg: Airplane (make decisions)

Interest: prob of making decisions.

* prob of error.

Estimation (continuous) problems

* close as possible
* eg: polling problem (size of error)

Philosophical point of view (how we model)

↙ ↘

Definite value (numbers)

* Nothing random

eg: estimate mass of an e^-

Data $\xrightarrow[\text{apparatus}]{\text{measuring}}$ estimator → θ

θ (lowercase) → real numbers (Not a R.V)

Don't know

* model as RV

[mass of e^- is constant not random] → we have prior data of others.

"we have prior distribution of others"

↓

Not meaning random (nature)

↓

Subjective beliefs (where it may fall on.

'Even though not random - model it as a random variable'

'Initial beliefs → observe R.V (measurement) → estimate.

↙ unknown numbers

↘ Random variables (Bayes rule)

Based on x: Find θ

Classical $\xrightarrow{\theta}$ $\boxed{P_X(x;\theta)}$ $\xrightarrow{X}$ $\boxed{\text{estimator}}$ → $\bar{\Theta}$ (with noise N)

Bayesian $\xrightarrow{\theta}$ $\boxed{P_X(x|\theta)}$ $\xrightarrow{X}$ $\boxed{\text{estimator}}$ → $\bar{\Theta}$ (with noise N)

Both cases $\theta \rightarrow$ R.V

why? → ① $\theta \rightarrow$ constant, data → Random, we calculate function of data → o/p will be a function of R.V (affected by noises)

$\bar{\Theta} \rightarrow$ R.V → estimated.

Bayesian inference : use Bayes rule

Hypothesis testing:

$$P_{\Theta|X}(\theta|x) = \frac{P_\Theta(\theta)\, P_{X|\Theta}(x|\theta)}{P_X(x)}$$

→ Initial beliefs

→ model of exp apparatus

→ descrete data.

$$P_{\Theta|X}(\theta|x) = \frac{P_\Theta(\theta)\, f_{X|\Theta}(x|\theta)}{f_X(x)}$$

→ continuous data

Estimation : Continuous data

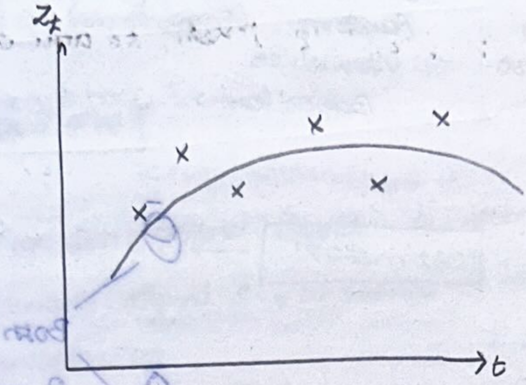
$$f_{\Theta|X}(\theta|x) = \frac{f_\Theta(\theta)\, f_{X|\Theta}(x|\theta)}{f_X(x)}$$

→ Data observed'

$$Z_t = \theta_0 + t\,\Theta_1 + t^2\,\Theta_2 \longrightarrow$$ [flying in air : Because of gravity: parabolic path]

General eqn of Parabola

* we don't know which parabola
* measure at diff times - measurement are noisy.



$x_t = z_t + W_t$ [measure position at different times]

priors: probability of distribution of θ

End result

$$f_{\theta_0, \theta_1, \theta_2 | x_1, \ldots, x_n}(\theta_0, \theta_1, \theta_2 | x_1, x_2, \ldots, x_n)$$

↓

multiple dimensional R.v

θ's and x's are often vectors of R.v

Estimation

$$f_{\theta|x}(\theta|x) = \frac{f_\theta(\theta)\, P_{x|\theta}(x|\theta)}{P_x(x)}$$

$$P_x(x) = \int f_\theta(\theta)\, P_{x|\theta}(x|\theta)\, d\theta$$

* coin flip (parameter θ)
* observe X heads in n tosses

↙ classical ↘ Bayesian

Bias $\hat{\theta}_n = \frac{X}{n}$ $\frac{\text{(heads in n trials)}}{\text{n trials (number}}$ → Tries to prove the properties

* classical → weak law of large numbers [converge to true parameters]

Bayesian

* prior distribution of θ [Assume distribution of θ]
* Any value of bias - possible likely to any other value. (uniform distribution): we don't anything.

* Little more faith: centered around 1/2. (Tried to give a fair coin)

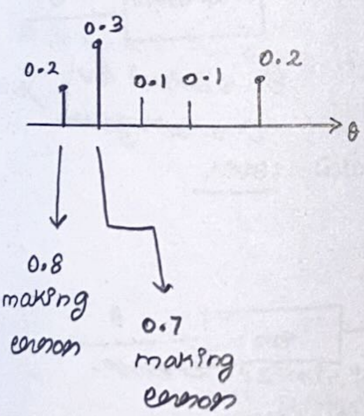
↓

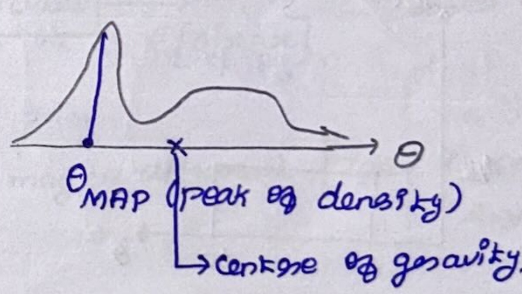
approx 1/2

* we will choose a prior
* use Bayes $f_{\theta|X}(\theta|x)$

Plots

pmf of $p_{\theta|X}(\theta|x)$ or pdf $f_{\theta|X}(\cdot|x)$





→ maximize: Being correct $\boxed{\theta = 0.3}$ → most likely.

↓

maximum a posteriori prob (MAP)

(Need to answer a single result)

Continuous:

* Any individual point has zero probability.
* Report average over the distribution.

1) * $f_{\theta|X}(\theta|x) = \max_\theta f_{\theta|X}(\theta|x)$ → Individual prob will be zero
1

Conditional expectation (Average) —— 2 methods

$$\boxed{E[\theta \mid X=y] = \int \theta f_{\theta|X}(\theta|x)\, d\theta}$$

* single answers can be misleading.

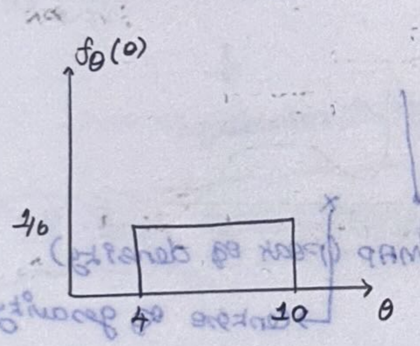
---

* some may choose one over other
↓
'single point doesn't tell anything'
'show entire plot - most likely in that range'

Rules:

'Estimation - without information'

Least mean squares estimation



θ → Actual R.V } error
c → our guess

Find estimate c,

minimize $E\left[(\theta - c)^2\right]$ → minimize error. (Square)
↓
Random value.

optical estimate: $c = E[\theta]$
↓ [Once known: what it is - what I guessed]

Solu: → minimum error.

$$E\left[(\theta - c)^2\right] = E[\theta^2] - 2c\,E[\theta] + c^2$$

$$\frac{d}{dc} E\left[(\theta-c)^2\right] = -2E[\theta] + 2c$$

$$2c = 2E[\theta]$$

$\boxed{c = E[\theta]}$ → Thing we need to report is $E[\theta]$.

How big will be error - How good your estimate

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

↓

mean square error.

* expectation is the best way to estimate a quantity if we are interested in mean squared error. [variance]

Now we have data → we can compute posterior/conditional distribution

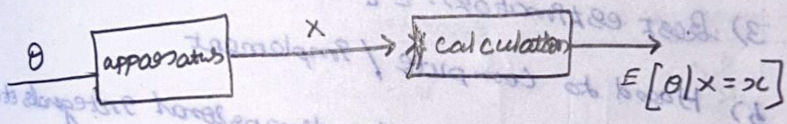
cond prob ≈ ordinary prob" → new universe.

$E\left[(\theta - c)^2 \mid X = x\right]$ is minimized

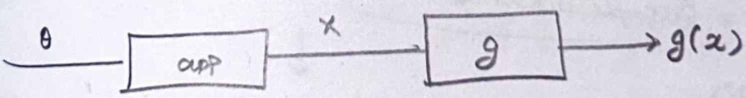
i.e. $\boxed{c = E[\theta \mid X = x]}$ → once you have data (minimize mean square error)

↓

report cond. estimate based on your data



Note: Calculations may be different [Idea of calculating R.V]



which one is best → ?

Assume: $E[\theta \mid X = x]$ → optimal in universe where $X = x$"

$$E\left[(\theta - E[\theta \mid X = x])^2 \mid X = x\right] \leq E\left[(\theta - g(x))^2 \mid X = x\right]$$

$$E\left[(\theta - E[\theta \mid X])^2 \mid X\right] \leq E\left[(\theta - g(X))^2 \mid X\right]$$

$$E\left[(\theta - E[\theta \mid X])^2\right] \leq E\left[(\theta - g(x))^2\right]$$

For any particular value → my $E[\theta \mid X = x]$ must be better

(Average — I need to be better) → Data itself random

$$\boxed{E[\theta \mid X] \text{ minimizes } E\left[(\theta - g(X))^2\right] \text{ over all estimates } g(\cdot)}$$

'Conditional expectation estimator'

↓

Ultimate estimating machine

(If you are forced to give a single value)

Bayesian: calcu & reporting conditional expectations?

LMS estimation with several measurement

1) unknown r.v θ

2) observe values of r.v $X_1, X_2, \dots X_n$

3) Best estimator: $E[\theta \mid X_1, \dots, X_n]$

4) Hard to compute / implement

Involve multi-dimensional integrals etc

Deal with vectors

Lot of computation

↓

Simple alternative

Bayesian Statistical Inference II

Problems:

we don't know θ, but we know $f_\Theta(\theta)$

↓

Distribution

$$\xrightarrow[f_\Theta(\theta)]{\theta} \boxed{f_{X|\theta}(x|\theta)} \xrightarrow{X} \boxed{\text{Estimation } g(\cdot)} \rightarrow \hat{\Theta} = g(X)$$

X is Affected by θ [Captured by $f_{X|\theta}(x|\theta)$]

Joint pdf = multiply $f_\theta(\theta)$ and $f_{X|\theta}(x|\theta)$

what about θ (Given X)

$f_{\theta|X}(\theta|x) \longrightarrow \theta$

(Now we have x)

↳ plot as a function of θ.

'Posterior distribution: Tell about θ' $\rightarrow \theta \rightarrow$ still random (with probabilities)

single value of θ

* Apply a certain function $\rightarrow$ estimator estimates $\bar{\theta}$'

* In general, θ & X are vectors.
$(\theta_1, \dots), (X_1, X_2, \dots)$

choose θ to report

↙ max pmf (maximum a posteriori estimate)

↓ maximizes $f_{\theta|X}(\theta|x)$

↘ mean squared error

↓ LMS estimation

$\bar{\theta} = E[\theta|X]$ minimizes $E[(\theta - g(X))^2]$ over all estimators $g(\cdot)$

Square. Estimation error (make as small as possible)

* Conditional expectation $\bar{\theta} = E[\theta|X] \longrightarrow$ Best estimator.

* Objective: mean square error as small as possible'

→ for any x, $\bar{\theta} = E[\theta \mid X = x]$ minimizes $E\left[(\theta - \bar{\theta})^2 \mid X = x\right]$

over all estimates $\bar{\theta}$.

↓

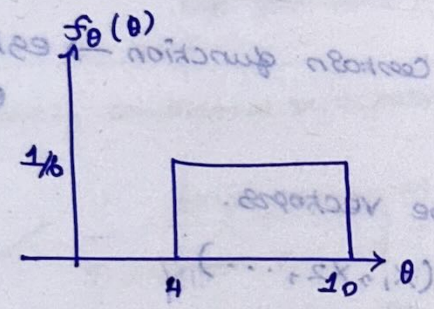
'No matter what we observe — best estimation'

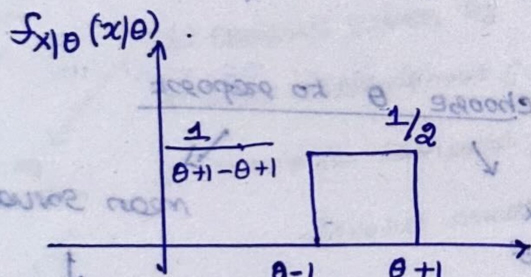
↓

cond. expectation on specific value observed.

optimal way of producing estimate

unknown R.V ⇝ uniform distributed b/w 4 and 10





X → noisy measurement of θ.

observation model.

$X = \theta + u$

u ↓ uniform $[-1, 1]$

prior

conditional density.

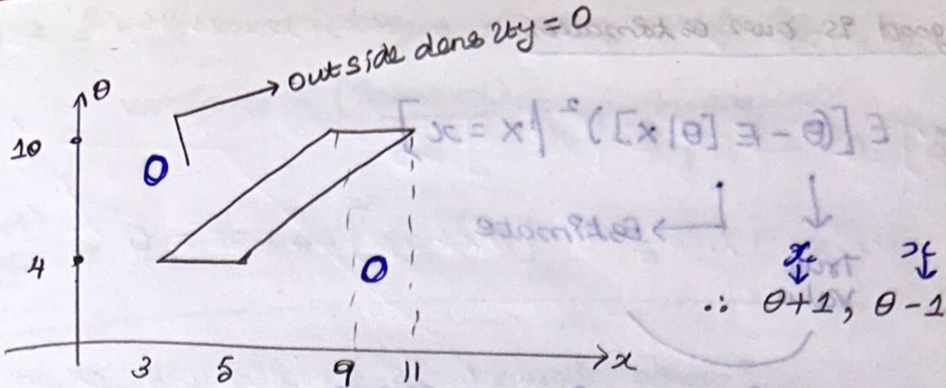
$$\boxed{f_{X,\theta} = f_\theta \cdot f_{X|\theta}}$$

Joint distribution density.

$f_{X,\theta} = \frac{1}{6} \times \frac{1}{2} = \frac{1}{12}$ [Not everywhere only possible X & θ]

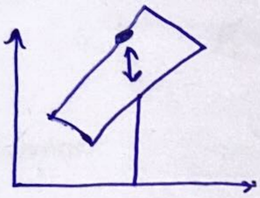
$\theta \rightarrow$ 4 & 10 (b/w)

$X \rightarrow \theta - 1$ up to $\theta + 1$.



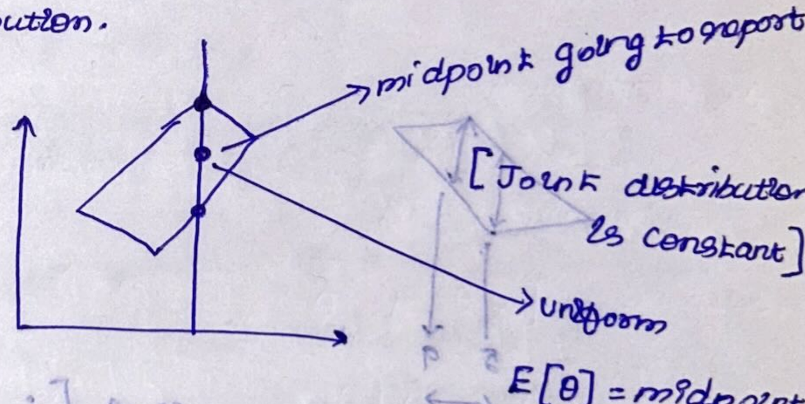
θ in terms of x → That's why θ in y axis.

Optimal estimator: Conditional expectation:



If x is fixed, we know the limit of θ.

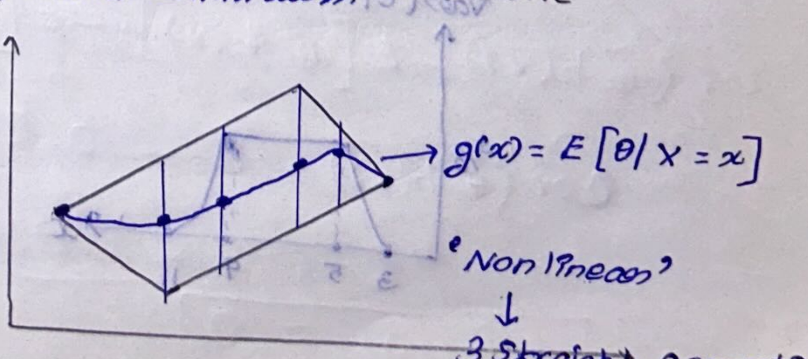
↳ Conditional distribution is a part of joint distribution.



∴ θ uniform distribution.
So midpoint.

$g(x) = E[\theta \mid X = x]$

↳ optimal estimator → curve



How good is our estimation:

$$E[(\Theta - E[\Theta|X])^2 | X = x]$$

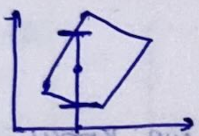
True Value

Estimate

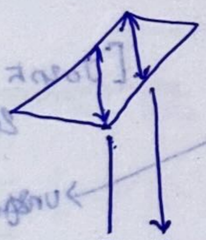
Expected square error.

* Same as $\text{var}(\Theta|X=x)$: variance of the conditional distribution of θ.

Variance



var of uniform interval $= \left(\frac{(b-a)^2}{12}\right) = \frac{2^2}{12} = \frac{1}{3}$

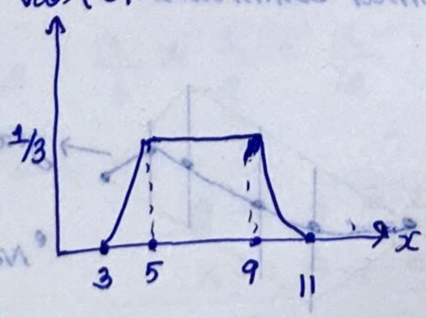


uniform interval of 2 [∴ 1/3 variance]

variance in $x = 3$ is 0
variance in $x = 11$ is 0

b/w 5 & 9 → 1/3

$\text{var}(\theta|X=x)$



H/w 3 & 5 → Interval changes linearly but in

$$\text{variance} = \frac{(\text{Interval})^2}{12} \rightarrow \text{quadratic.}$$

$$\text{variance} \propto (\text{interval})^2$$

'certain measurements are better than others'

In $x = 3 \longrightarrow \theta = 4 \longrightarrow$ certain.

'Based on info of x → Some x are information others not'

Prop of LMS estimation

$\bar{\theta} \rightarrow$ estimator

Estimation error : $\tilde{\theta} = \bar{\theta} - \theta$

expected estimation error:

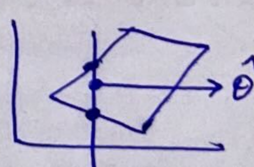
$$E[\tilde{\theta} | x] = E[\hat{\theta} - \theta | x]$$

$$= E[\hat{\theta} | x] - E[\theta | x]$$

$\hat{\theta}$ function of θ → itself function of x

once we know $X = x$, we know θ → then $\hat{\theta}$

(only one $\hat{\theta}$ per x)



$$E[\hat{\theta} | x] = \hat{\theta}$$

$$= \hat{\theta} - \underbrace{\hat{\theta}}_{E[\theta|x]}$$

$$\boxed{E[\tilde{\theta} | x] = 0}$$

'on average my error is 0'

(No matter of x cond. exp errors will be ≈ 0) → on average

* $E[\tilde{\theta}] = 0 \rightarrow E[\tilde{\theta} \mid X = x] = 0$ —— $\hat{\theta}$ is unbiased. (No tendency to be in high/low)

* This means likely 0 [Average] ↪ not = 0

*

How θ is related with other functions

$$E[\tilde{\theta}\, h(X) \mid X]$$

error: related with an arbitrary function of data.

when X is known, h(X) is known:

$$h(X)\, E[\tilde{\theta} \mid X]$$

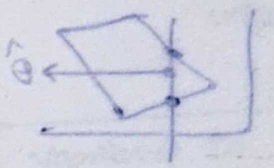
$$E[\tilde{\theta}\, h(X) \mid X = x] = h(X)\,(0) = 0 \longrightarrow \times$$

Iterative expectation

$E[\tilde{\theta}\, h(X) \mid X = x]$ → iterative expectation of a conditional expectation is unconditional expectation.

$$E[\tilde{\theta}\, h(X) \mid X = x] = E[\tilde{\theta}\, h(X)]$$

$$\boxed{0 = E[\tilde{\theta}\, h(X)]}$$



Covariance b/w $\tilde{\theta}$ & $\hat{\theta}$

(error & estimate)

$$\text{Cov}(\tilde{\theta}\, h(X)) = E[\tilde{\theta}\, h(X)] - E[\tilde{\theta}]\, E[h(X)]$$

$$\text{Cov}(\tilde{\theta}\, h(X)) = 0 - 0\, E[h(X)]$$

↳ Already proved

$$\boxed{\text{Cov}(\tilde{\theta}\, h(X)) = 0}$$

Covariance b/w error & any function of x is 0.

$\boxed{cov(\tilde{\theta}, \tilde{\theta}) = 0}$ ← using previous.

$\boxed{\therefore\ \theta = \hat{\theta} - \tilde{\theta}}$

$$var(\theta) = var(\hat{\theta}) + var(\tilde{\theta})$$

$\hookrightarrow (-)^2$

1) Condition R.v → any function on than R.v is constant (x is known, f(x) known) → pulled out of expectation (cond.)

2) Iterative expectation.

'Error is uncorrelated with estimation'

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

when $\hat{\theta} > \theta \rightarrow +ve$

$\hat{\theta} < \theta \rightarrow (-)ve$

But

By correlation → Tells error doesn't depend on estimation

violate. ←

'hypothetical situation'

∴ Based on expectation: No matter of x, Your error will be 0. (on average): don't expect it on +ve or (-)ve.'

↳ 'That's why $\tilde{\theta}$ uncorrelated with $\hat{\theta}$.'

If

If we think,

$\hat{\theta} < \theta$ ($\tilde{\theta} \to$ negative)

our estimate is not a optimal one → Need to be Corrected

for $\boxed{\hat{\theta} = 0}$

Already we are using LMS → No way of improving it

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

↓

∴ $\hat{\theta}$ & $\tilde{\theta}$ are uncorrelated

$\text{Var}(\theta) = \text{Var}(\hat{\theta}) + \text{Var}(\tilde{\theta})$ → law of total variance (case)

∴ $\hat{\theta}$ tries close to θ

random (more uncertainity it removes – lesser the error)

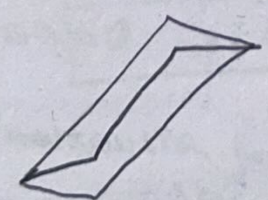
⋆ $\boxed{\tilde{\theta} = \hat{\theta} - \theta}$

we want

$\hat{\theta} \approx \theta$ $\boxed{\tilde{\theta} \approx 0}$

Best case

Linear Mean Square



'Non linear' curve - optimal estimation

↓

[Complicated] → multi dimension integration.

Cond. expectation

many R.v

'make linear' — estimate constrained.

Estimator: linear curve. [choose best possible] minimize $E[(\theta - aX - b)^2]$ $\hat{\theta} = aX + b.$

over all a & b

mean square error

quadratic function.

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

θ

* Best choice of a, b: best linear estimator:

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

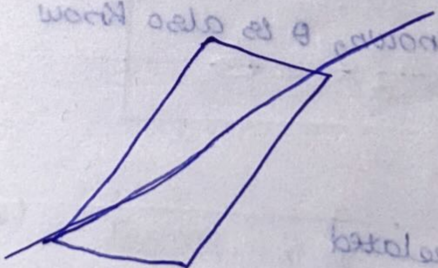
minimize quadratic function:

$E[(\theta - aX - b)^2] \longrightarrow h(a,b) \longrightarrow$ quadratic

Set the derivative (a,b) to 0:

$$\underbrace{\hat{\theta}_L = E[\theta]}_{\text{optimal } b} + \underbrace{\frac{\operatorname{cov}(X, \theta)}{\operatorname{var}(X)}}_{\text{optimal } a} (X - E[X])$$

we will have a curve



In this: X and θ are +vely correlated)

Big X $\longrightarrow$ Big θ.

Estimate:

$$\hat{\theta}_L = E[\theta] + \frac{\overset{\to +ve}{\operatorname{cov}(X, \theta)}}{\operatorname{var}(X)} (X - E[X])$$

$\hat{\theta}_L = E[\theta]$

↳ If we don't have info [not correlated]

★ $X \rightarrow$ Bigger than $E[X]$

$\theta \rightarrow$ Bigger than $E[\theta]$

$$\boxed{\therefore\ \hat{\theta}_L = E[\theta] + \text{Some positive}}$$

↓

Increment.

(Allows us to make correction)

Obtaining info on $X \rightarrow$ allows to correct $\hat{\theta}_L$

+ve correlation

($X \rightarrow$ used to make correction)

0 correlation

X info useless

Formula for MSE: (mean square error)

$$E\left[(\hat{\theta}_L - \theta)^2\right] = (1-\rho^2)\,\sigma_\theta^2$$

$\because$ we know θ_L

sub in this will.

ρ ↓ correlation co.eff

when ρ is large $\rightarrow$ MSE is much small

when $\rho = 1 \rightarrow$ perfectly correlated

when X is known, θ is also know

$$\boxed{MSE = 0}$$

$\rho = 0 \rightarrow$ uncorrelated

measurement doesn't help me to figure out θ'

uncertainity in θ estimation.

↓

(variance of θ)

Multiple data (vectors)

$$\hat{\theta} = a_1 x_1 + a_2 x_2 + \dots + a_n x_n + b.$$

Best choice of $a_1, \ldots a_n, b$.

Estimator: $E[\theta | X_1, \ldots X_n] \rightarrow$ we want to constrain to linear function.

Minimize:

$$E[(a_1 X_1 + \cdots + a_n X_n + b - \theta)^2]$$

$\downarrow$

Mean square error minimized.

* set to zero.

* only means, variances & covariances matter.

$\downarrow$

we will do vector algebra.

Expands to $a_1^2 E[X_1^2] + 2a_1 a_2 E[X_1 X_2] + \cdots$

If we know mean & variance we find this

covariance we know this.

Previous: $E[(\hat{\theta}_L - \theta)^2] = (1 - \rho^2)\sigma_\theta^2$

$\downarrow$

only mean, var, covariance enough not complete distribution.

Example

Random.V with noises (measured)

$X_i = \theta + W_i$ → Noises

$\theta, W_1, \ldots, W_n \rightarrow$ Independent

$\theta \sim \mu_0, \sigma_0^2$

mean & variance

$W_i \sim 0, \sigma_i^2$

Linear combination weighted average

$$\hat{\theta}_L = \frac{\mu_0/\sigma_0^2 + \sum_{i=1}^{n} X_i/\sigma_i^2}{\sum_{i=0}^{n} 1/\sigma_i^2}$$

Prior mean

(weighted average of $\mu, x_1, \ldots, x_n$

* If all normal, $\hat{\theta}_L = E[\theta \mid x_1, \ldots, x_n]$

$$\boxed{\text{Any } x_i \text{ is weighted } \propto \frac{1}{\text{Variance}}}$$

(Linear estimation ~ Normal)

$\frac{x_i}{\sigma_i^2}$ → weighted (very noisy: σ_i^2 large → $\frac{x_i}{\sigma_i^2}$ → small weight)

(Not so useful) as huge randomness

Small σ_i^2 → x_i (useful) → large weight.

measuring device: <u>measure x^3 instead of x</u>

Telling x is same as x^3 (In terms of info) → Same mean

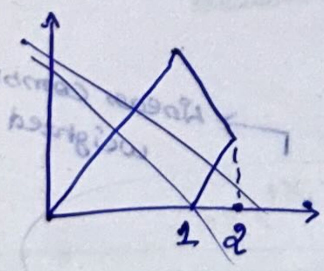
$E[\theta|x]$ same as $E[\theta|x^3]$

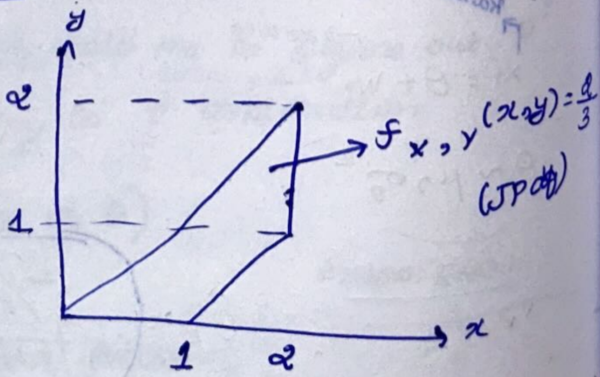
<u>Linear LMS</u> → restrict estimator

$$\hat{\theta} = ax+b \neq \hat{\theta} = ax^3+b$$

$$\boxed{\hat{\theta} = a_1 x + a_2 x^2 + a_3 x^3 + b}$$

<u>Recitation</u>





a) $g(X) = E[Y|X]$