- Correlation Coefficient

= Dimensionless Version of Covariance

$$\rho = E\left[\frac{(X-E[X])}{\sigma_X}\cdot\frac{(Y-E[Y])}{\sigma_Y}\right]$$

$$= \frac{cov(X,Y)}{\sigma_X\sigma_Y} \rightarrow -1 \le \rho \le 1.$$

$|\rho| = 1 \Leftrightarrow (X-E[X]) = c(Y-E[Y])$

↳ linearly Related.

$|\rho| = 0 \rightarrow$ independent

L.12 Iterated Expectations + Sum of Random Number of Random Variables

- conditional expectations

$$E[X|Y=y] = \sum_x x P_{X|Y}(x|y)$$

- Law of Iterated expectations

$$E[E[X|Y]] = \sum_y E[X|Y=y]\,P_Y(y) = E[X]$$

- $Var(X|Y=y) = E[(X-E[X|Y=y])^2 \mid Y=y]$

↳ $\boxed{\text{law of total variance}}$

$$Var(X) = E[var(X|Y)] + Var(E[X|Y])$$

"X의 분산은 X의 Y조건부 분산의 Y평균

+X의 Y조건부 평균의 Y분산."

- Section Means & Variance.

(ex) $Y=1$ (10 students) → mean $score|_{Y=1} = 90 = E[X|Y=1]$

$Y=2$ (20 students) → mean $score|_{Y=2} = 60 = E[X|Y=2]$

then $E(X) = \frac{1}{30}\times(10\times90 + 20\times60) = 70.$

what happend? $E(X) = \sum_x \sum_y x\,P_Y(Y)\,P_X(X|Y)$

$$= \sum_y P_Y(y)\,E(X|y)$$

$$Var(E[X|Y]) = \sum_y P_Y(y)\,(E[X|Y] - \underbrace{E[E[X|Y]]}_{=E[X]})^2$$

$$= \frac{1}{3}(90-70)^2 + \frac{2}{3}(60-70)^2 = 200.$$

$E(Var[X|Y]) = \sum_y P_Y(y)\,Var[X|Y=y] = \frac{50}{3}$ → Var(X)

↳ $\boxed{\text{section별 mean \& variance \& 각 section의 확률 알면 total mean \& variance 구할 수 있다.}}$

- Sum of random number of RN (independent)

↳ N: number of stores visited

X_i: money spent in store i

(assume i.i.d) ↳ independent & identically distributed

↳ $Y = \sum_{i}^{N} X_i$

then $E[Y|N=n] = n\,E[X]$

$E[Y|N] = N\,E[X]$

$$\Rightarrow E[Y] = E[E[Y|N]]$$

$$= E[N\,E[X]]$$

$$= E[N]\,E[X]$$

$E[N]$: expectation of random number from 1 to N

$E[X]$: expectation of i.i.d. RV X

- Variance of sum of random number of independent R.V

$$Var[Y] = E(Var(Y|N)) + Var(E(Y|N))$$

$$= E[N]\,Var[X] + E[X]^2\,Var(N)$$

L.13 Bernoulli Process (Discrete) → P

L.14, 15 Poisson Process (Continuous) → $P(k,\tau) = \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!}$

↳ memoryless

L.16~ Markov Chain

⇒ with memory / dependence across time.

(Lec16~18) Markov Processes

new state = f(oldstate, noise)

- Finite Markov Chain.
  - X_n : state after n transitions belong to a finite set $\{1, \ldots, m\}$
    - X_0 is either given or random.
  - Markov Property / Assumption :
    - $P_{ij} = P(X_{n+1} = j \mid X_n = i)$
      $= P(X_{n+1} = j \mid X_n = i, X_{n-1}, \ldots X_0)$

      assumption that past doesn't matter.

  ⇒ possible states
      transition
      probability for each transition

- n-step transition probability
  - $r_{ij}(n) = P(X_n = j \mid X_0 = i)$
    $= \sum_{k=1}^{m} r_{ik}(n-1) P_{kj}$; key recursion
  - With random initial state
    $P(X_n = j) = \sum_{i=1}^{m} P(X_0 = i) r_{ij}(n)$

- Generic Convergence Questions ?
  ⇒ - Does $r_{ij}(n)$ convergence to something?
      ; steady, oscillating, ...
    - Does the limit depend on initial state?

- Recurrent vs Transient States
- Periodic states 
- Steady-state probability
- Visit frequency interpretation

- The phone company problem.

- Calls originates a Poisson process rate λ
  - ↳ call duration(μ) exponentially distributed.
- B lines available

⇒ Discrete time intervals of small length δ

- Balance equations ; $\lambda \pi_{i-1} = i\mu \pi_i$
  (Finding steady states by $n \to \infty$ for key recursion)

⇒ $\pi_i = \pi_0 \dfrac{\lambda^i}{\mu^i i!}$ $\quad \pi_0 = \dfrac{1}{\sum_{i=0}^{B} \dfrac{\lambda^i}{\mu^i i!}}$

- Mean First Passage and Recurrence Times
  - chain with one recurrent class ;

  ⇒ fix s recurrent
  - Mean first time passage time from i to s:
    $t_i = E[\min\{n \geq 0 \text{ such that } X_n = s\} \mid X_0 = i\}$

  ⇒ $t_s = 0$
     $t_i = 1 + \sum_j P_{ij} t_j$
  - Mean recurrence time of s :

    $t_s^* = E[\min\{n \geq 1 \text{ s.t. } X_n = s\} \mid X_0 = s]$

    $t_s^* = 1 + \sum_j P_{sj} t_j$

Lect. 19) Limit Theorem

• Chebyshev's Theorem.

r.v. X w/ (μ, σ^2)

$$\Rightarrow \sigma^2 = \int (x-\mu)^2 f_X(x)dx$$

$$\geq \int_{-\infty}^{-c} (x-\mu)^2 f_X(x)dx + \int_{c}^{\infty} (x-\mu)^2 f_X(x)dx$$

$$\geq c^2 P(|X-\mu| \geq c)$$

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

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$$

• Convergence in probability

For every $\epsilon > 0$, $\lim_{n\to\infty} P(|Y_n - a| \geq \epsilon) = 0$.

example $Y_n = 1 - \frac{1}{n}$.

• Convergence of sample mean.

$X_1, X_2, \cdots$ iid finite mean μ & variance σ^2

$$\Rightarrow M_n = \frac{X_1 + \cdots X_n}{n}$$

$$E[M_n] = \frac{E[X_1] + \cdots + E[X_n]}{n} = \frac{n\mu}{n} = \mu$$

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

$$\hookrightarrow \boxed{P(|M_n - \mu| < \epsilon) \leq \frac{Var(M_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}}$$

• M_n converges to μ in probability

⇓

Weak Law of Large Numbers.

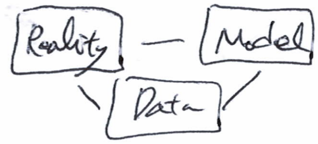
• The Central limit theorem

↳ $\forall c$; $P(Z_n \leq c) \rightarrow P(Z < c)$ ← standard normal CDF

$$Z_n = \frac{S_n - E(S_n)}{\sigma_{S_n}}$$

(zero mean, unit variance), $S_n = X_1 + X_2 + \cdots X_n$;

Lect 21. Baysian Statistical Inference



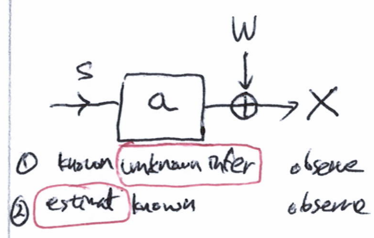
Application: polling, medical / pharmaseutical trials, netflix competition, Finance, signal processing (tracking, detection, speaker identification)

• Types of Inference models / approaches

− Model Building vs Inferring unknown variables

($X = aS + W$)
observe, signal, noise

infera



then...
Estimation in the presence of noise
know a, observe X, estimate S

− Hypothesis ~~the~~ testing

↳ unknown takes one of possible values & aim at small PR of incorrect decision

− Estimation: aim at a small estimation error.

• Classical Statistics

different philosophical approach

random variable due to X (not θ)

unknown ex) ~~electron mass~~ fixed number (No random variable)

• Bayesian: Use priors & Bayes rule

prior knowledge (= personal belief) "I view the electron mass as distribution P_Θ"

observe another related random variable X by using an aparatus

create distribution of theta

Model "electron mass" as random variable Θ

Bayesian Inference : Use Bayes rule

- Hypothesis testing
  - discrete data

posterior, prior, model of experiment of aparatus (likelihood)

$$\boxed{p_{\Theta|X}(\theta|x)} = \frac{\boxed{p_\Theta(\theta)}\,\boxed{p_{X|\Theta}(x|\theta)}}{p(x)}$$

  - continuous data

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

- Estimation ; continuous data

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

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

$$X_t = Z_t + W_t$$

⇒ Bayes rule gives

$$f_{\Theta_0,\Theta_1,\Theta_2|X_1,X_2,\dots X_n}(\theta_0,\theta_1,\theta_2|x_1,\dots x_n)$$

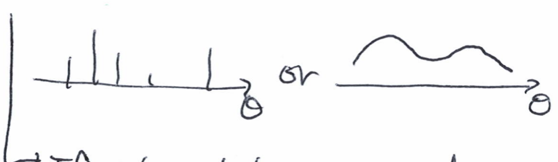
(See Monty Hall Problem)

- Estimation w/ ~~continuous~~ discrete data

$$\boxed{\begin{aligned} f_{\Theta|X}(\theta|x) &= \frac{f_\Theta(\theta)\, p_{X|\Theta}(x|\theta)}{p_X(x)} \\ p_X(x) &= \int f_\Theta(\theta)\, p_{X|\Theta}(x|\theta)\, d\theta \end{aligned}}$$

- Output of Bayesian Inference

⇒ Posterior distribution



↳ If interested in a single answer,

- Maximum a posteriori probability (MAP)

$$\Rightarrow p_{\Theta|X}(\theta^*|x) = \max_\theta p_{\Theta|X}(\theta|x)$$

minimizes PR of error; often used in hypothesis testing.

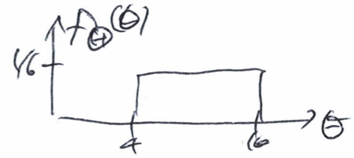
& $f_{\Theta|X}(\theta^*|x) = \max_\theta f_{\Theta|X}(\theta|x)$

& conditional expectation

$$E[\Theta|X=x] = \int \theta f_{\Theta|X}(\theta|x)\, dx$$

& single answers can be misleading!

(ex1) Least Mean Square Estimation



→ find estimate c to minimize $E[(\Theta-c)^2]$

→ $c = E(\Theta)$

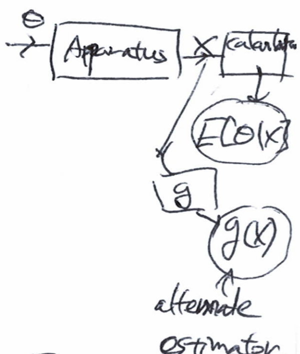
↳ Optimal mean squared error

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

if we observe that $X=x$

$E[(\Theta-c)^2|X=x]$ is minimized by

$c = [\Theta|X=x]$



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

$$E[(\Theta-E[\Theta|X])^2|X] \le E[(\Theta-g(X))^2|X]$$

$$E[(\Theta-E[\Theta|X])^2] \le E[(\Theta-g(X))^2]$$

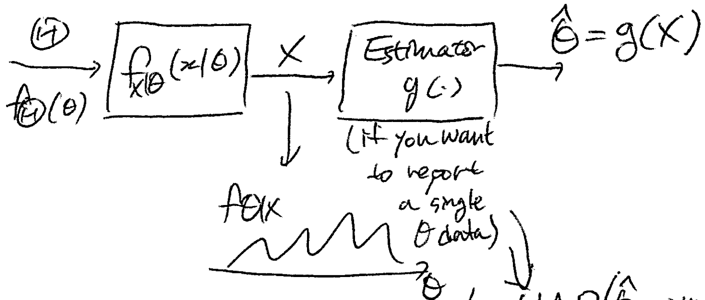
$E[\Theta|X]$ minimizes $E[(\Theta-g(X))^2]$ over all estimators $g(\cdot)$

★ LMS Estimation w. several measurement

- unknown r.v Θ
- Observe values of r.v's $X_1, X_2, \dots X_n$
- Best Estimator : $E[\Theta|X_1, \dots X_n]$

Bayesian

• Linear LMS estimation



• MAP ($\hat\Theta_{map}$ = max $f_{\Theta|X}(\theta|x)$)

or

• LMS Estimation

$\hat\Theta = E[\Theta|X]$ minimizes variation of Θ overall $g(x)$

• Linear LMS

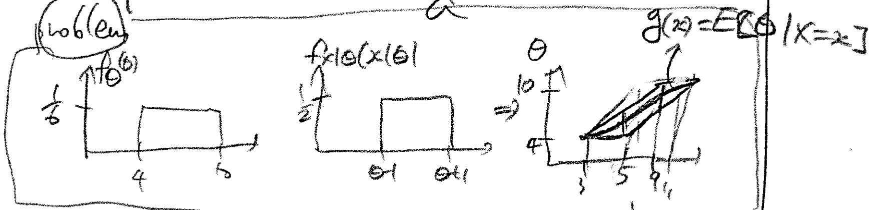
Consider estimator of Θ, of the form $\hat\Theta = aX + b$

↳ minimize$_{a,b}$ $E[(\Theta - aX - b)^2] = h(a,b)$: quad

↳ Best choice of a, b; best linear estimator

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

a



• Linear LMS with multiple data → general estimator $E[\Theta|X_1, \dots X_n]$

$\hat\Theta = a_1X_1 + \cdots a_nX_n + b$

↳ Find best choices of $a_1 \sim a_n$, b

↳ minimize $E[(a_1X_1 + \cdots a_nX_n + b - \Theta)^2]$

↳ set derivative to zero. linear system in b and the a_i

↳ only means, variances, covariances matter!!

• The clearest linear LMS example

$X_i = \Theta + W_i$ $\quad \Theta, W_1, \cdots W_n$ independent

$\Theta \sim \mu, \sigma_0^2$ $\quad W_i \sim 0, \sigma_i^2$

$$\Rightarrow \hat\Theta_L = \frac{\mu/\sigma_0^2 + \left(\sum_{i=1}^{n} X_i/\sigma_i^2\right)}{\left(\sum_{i=0}^{n} 1/\sigma_i^2\right)}$$

If all normal, $\hat\Theta_L = E[\Theta | X_1, X_2, \cdots X_n]$

Big picture

• Standard examples

– X_i uniform on $[0, \theta]$:

↳ uniform prior θ.

– X_i Bernoulli(p)

↳ uniform (or Beta) prior on p

– X_i normal w/ mean θ, known variance σ^2

↳ $X_i = \Theta + W_i$

normal prior on θ;

• Estimation Method

- MAP
- MSE
- Linear MSE

PRK

Lect. 23 Classical Inference

* Maximum likelihood estimation

Model w/ unknown parameters

$X \sim P_X(x;\theta)$

→ Pick θ that "makes data most likely"

$\hat{\Theta}_{ML} = \arg\max_{\theta} P_X(x;\theta)$ → likelihood

→ Compare to Bayesian MAP ?

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

likelihood, prior

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