MIT Probability System Analysis and Applied Probability.

Lecture 1: Probability Models and Axioms

(1) Sample space Ω : list of possible outcomes

properties: * mutually exclusive
* collectively exhausive

categories: discrete, continuous

(2) Axioms: ① Nonnegativity $P(A) \geq 0$

② Normalization: $P(\Omega) = 1$

③ Additivity: If $A \cap B = \phi$, then $P(A \cup B) = P(A+B)$

(3) Discrete uniform law: $P(A) = \frac{\#\text{ of elements of } A}{\#\text{ of total samples}}$

if all outcomes are equally likely.

(4) Continuous uniform law: Probability = Area

(5) ~~Cout~~ Countable additivity axiom:

if $A_1, A_2 \ldots$ are disjoint events, then:

$$P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$$

Lecture 2: Conditioning and Bayes' Rule

(1) $P(A|B)$ = Probability of A, given B

$$\begin{cases} = \dfrac{P(A \cap B)}{P(B)} & \text{if } P(B) \neq 0 \\ \text{undefined} & \text{if } P(B) = 0 \end{cases}$$

(2) Total probability theorem

sample space are divided into $A_1, A_2 \ldots A_n$

B is an event in the sample space

$$P(B) = P(B|A_1)P(A_1) + \cdots P(B|A_n)P(A_n)$$

(3) Bayes' rule

* prior probabilities $P(A_i)$ – initial ~~beliefs~~ beliefs
* know $P(B|A_i)$ for all i's
* wish to know $P(A_i|B)$ – revise beliefs, given B occured

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i)P(B|A_i)}{P(B)} = \frac{P(A_i)P(B|A_i)}{\sum_j P(A_j)P(B|A_j)}$$

$A_i \xrightarrow[P(B|A)]{\text{cause-effect}} B$

$A_i \xleftarrow[P(A_i|B)]{\text{inference}} B$

Lecture 3: Independence

(1) Definition of independence: $P(A \cap B) = P(A) \cdot P(B)$

(2) Conditioning may affect independence

$$P(A \cap B|C) \{=, \neq\} P(A|C)P(B|C)$$

(3) Independence of a collection of events: $P(A_i \cap A_j \cap \cdots \cap A_q) = \prod P(A_i) \cdots P(A_q)$

(4) If there are A, B, C events.

$$P(A \cap C) = P(A)P(C);\ P(A \cap B) = P(A)P(B);\ P(B \cap C) = P(B)P(C)$$

called pairwise independence

pairwise independence doesn't imply independence: i.e.

$$P(A \cap B \cap C) \{\neq\} P(A)P(B)P(C).$$

(5)

Lecture 4: Counting

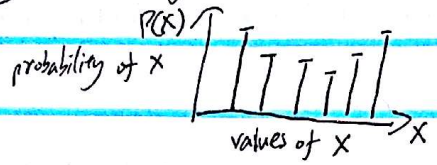
(1) Discrete uniform law

$$P(A) = \frac{\#\text{ of } A}{\text{total } \#\text{ of samples}} = \frac{|A|}{|\Omega|}$$

Lecture 5: Discrete Random Variables; Probability Mass Functions; Expectations

(1) Random variables: functions from the sample space Ω to the real numbers. {discrete / continuous

(2) Probability mass function



(3) Expectation: $E[X] = \sum_x x P_X(x)$

(4) Properties of expectations:

* X is a r.v., $Y = g(X)$

$$E[Y] = \sum_y y P_Y(y) = \sum_x g(x) P_X(x)$$

* $E[\alpha] = \alpha$
* $E[\alpha X] = \alpha E[X]$
* $E[\alpha X + \beta] = \alpha E[X] + \beta$

} α and β are constants

* $E[g(X)] = g(E(X))$ if g is a linear function

(5) Second moment $E[X^2] = \sum_x x^2 P_X(x)$

(6) Variance: $var(X) = E[(X - E[X])^2]$

$$= \sum_x (x - E[X])^2 P_X(x) =$$

$$= E[X^2] - (E[X])^2$$

$\sum_x x^2 P_X(x) - \sum_x 2x E[X] P_X(x) + \sum_x E[X]^2 P_x$

$= E[X^2] - 2E[X]^2 + E[X]$

$$var(\alpha X + \beta) = \alpha^2 var(X)$$

Lecture 6: Discrete Random Variable Examples; Joint PMFs

(1) Conditional expectation

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

where $P_{X|A}(x) = P(X = x|A)$

(2) Total Expectation theorem

$$P(B) = P(A_1)P(B|A_1) + \cdots P(A_n)P(B|A_n)$$

$$P_X(x) = P(A_1)P_{X|A_1}(x) + \cdots P(A_n)P_{X|A_n}(x)$$

$$E[X] = \cancel{P(A_1)P_{X|A_1}(} P(A_1)E[X|A_1] + \cdots P(A_n)E[X|A_n]$$

(3) Joint PMF: $P_{X,Y}(x,y) = P(X = x \text{ and } Y = y)$

Lecture 7: Multiple Discrete Random Variables: Expectations, Conditioning, Independence.

(1) Random variables X, Y, Z are independent if

$$P_{X,Y,Z}(x,y,z) = p_X(x) \cdot p_Y(y) \cdot p_Z(z)$$

(2) if X and Y are ~~variabl~~ independent random variables

$$E[XY] = \sum_x \sum_y xy\, P_{X,Y}(x,y)$$

$$= \sum_x \sum_y xy\, P_X(x) P_Y(y)$$

$$= \sum_x x P_X(x) \sum_y y P_Y(y)$$

$$= E[X]E[Y]$$

$$E[g(X)h(Y)] = \sum_x \sum_y g(x)h(y) P_{X,Y}(x,y)$$

$$= E[g(X)]E[h(Y)]$$

(3) for general cases, $Var(aX) = a^2 Var(X)$

$$Var(X+a) = Var(X)$$

if X, Y are independent

$$Var(X+Y) = Var(X) + Var(Y)$$

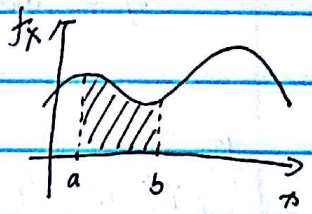
Lecture 8: Continuous Random Variables

(1) Probability density function f_X

$$P(a \le X \le b) = \int_a^b f_X(x)\,dx$$

$$\int_{-\infty}^{+\infty} f_X(x)\,dx = 1$$

$$P(X \in B) = \int_B f_X(x)\,dx$$



(2) $E[X] = \int_{-\infty}^{\infty} x f_X(x)\,dx$

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\,dx$$

$$\text{var}(X) = \sigma_X^2 = \int_{-\infty}^{\infty} (x - E[X])^2 f_X(x)\,dx$$

(3) Cumulative distribution function (CDF)

$$F_X(x) = P(X \le x) = \int_{-\infty}^{x} f_X(t)\,dt$$

$$\frac{dF_X(x)}{dx} = f_X(x)$$

(4) Gaussian (normal) PDF: ~~$N(0,1): f_X(x) =$~~

$$N(\mu, \sigma^2) = f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

$$E[X] = \mu \qquad \text{Var}[X] = \sigma^2$$

(5) No closed-form solution for CDF of normal distribution

If $X \sim N(\mu, \sigma^2)$ $\quad \frac{X-\mu}{\sigma} \sim N(0,1)$

Lecture 9: Multiple Continuous Random Variables

(1) Joint PDF: $P((X,Y) \in S) = \iint_S f_{X,Y}(x,y)\,dx\,dy$

$$P(x \le X \le x+\delta,\ y \le Y \le y+\delta) \approx f_{X,Y}(x,y) \cdot \delta^2$$

$$E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\,dx\,dy$$

$$\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy = f_X(x)$$

(2) X and Y are independent if $f_{X,Y}(x,y) = f_X(x) f_Y(y)$

(3) Conditioning

$$f_{X|Y}(x,y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \quad \text{if } f_Y(y) > 0$$

if X and Y are independent: $f_{X,Y} = f_X f_Y$,

then $f_{X|Y}(x,y) = f_X(x)$

Lecture 10: Continuous Bayes' Rule; Derived Distributions

(1) The Bayes variations

$$\begin{cases} P_{X|Y}(x|y) = \dfrac{P_{X,Y}(x,y)}{P_Y(y)} = \dfrac{P_X(x)P_{Y|X}(y|x)}{P_Y(y)} \\ f_{X|Y}(x|y) = \dfrac{f_X(x)f_{Y|X}(y|x)}{f_Y(y)} \end{cases}$$

$$\begin{cases} \text{discrete } X, \text{ continuous } Y: \quad P_{X|Y}(x|y) = \dfrac{P_X(x)f_{Y|X}(y|x)}{f_Y(y)} \\ \text{continuous } X, \text{ discrete } Y: \quad f_{X|Y}(x|y) = \dfrac{f_X(x)P_{Y|X}(y|x)}{P_Y(y)} \end{cases}$$

(2) Derived distribution: a PMF or PDF of a function of one or more random variables with known probability law.

(3) Continuous derived distribution: Cookbook procedure

* Get CDF of Y: $F_Y(y) = P(Y \le y)$
* Differentiate to get $f_Y(y) = \frac{dF_Y}{dy}(y)$

(4) The PDF of $Y = aX + b$

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$$

Lecture 11: Derived Distributions; Convolution; Covariance and Correlation

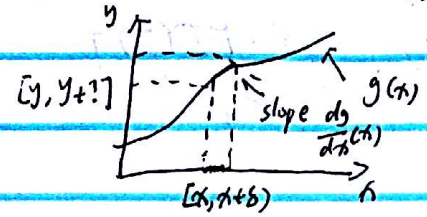
General formula for derived distributions.

(1) Let $Y = g(X)$, g strictly monotonic

Event $x \le X \le x+\delta \Rightarrow g(x) \le Y \le g(x+\delta)$

$\Rightarrow g(x) \le Y \le g(x) + \delta\,|(dg/dx)(x)|$

Hence, $\delta f_X(x) = \delta f_Y(y) \left|\frac{dg}{dx}(x)\right|$



(2) Discrete distribution of $X+Y$, $W = X+Y$

$$P_W(w) = P(X+Y=w) = \sum_x P(X=x) P(Y=w-x)$$
$$= \sum_x p_X(x) P_Y(w-x)$$

(3) Continuous distribution of $X+Y$, $W = X+Y$

$$f_W(w) = \int_{-\infty}^{\infty} f_X(x) f_Y(w-x)\, dx$$

(4) Sum of independent normal random variables is normal distributed.

$$X_i \sim N(\mu_i, \sigma_i^2) \quad i = 1 \dots n$$
$$Z = \sum_i X_i \sim N\left(\sum_i \mu_i, \sum_i \sigma_i^2\right) \quad i = 1 \dots n$$

(5) Covariance: $Cov(X,Y) = E[(X - E[X]) \cdot (Y - E[Y])]$

if $Cov(X,Y) > 0$, X and Y are positive related

< 0, negative related.

$$Cov(X,Y) = E[XY] - E[X]E[Y]$$

(6) $Var\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n Var(X_i) + \sum_{(i,j):i\neq j} Cov(X_i, X_j)$

~~$\because E[(X_1 + \dots X_n)^2] = E[\sum_i X_i^2 + \sum_{i\neq j} X_i X_j]$~~

$$\because E\left[\left(\sum_i X_i - \sum_i E[X_i]\right)^2\right] = E\left[\sum_i X_i^2 + \sum_{i\neq j} X_i X_j - 2\sum_i X_i E[X_i] - 2\sum_i X_i \sum_j E[X_j] + \sum_i E[X_i]^2 + \sum_{i\neq j} E[X_i] E[X_j]\right]$$
$$= E\left[\sum_i X_i^2 - 2\sum_i X_i E[X_i] + \sum_i E[X_i]^2\right] + E\left[\sum_{i\neq j} X_i X_j - 2\sum_i X_i \sum_j E[X_j] + \sum_{i\neq j} E[X_i] E[X_j]\right]$$
$$= E[(X_i - E[X_i])^2] + \sum_{i\neq j} E[X_i X_j] - 2E\left[\sum_i X_i \sum_{i\neq j} E[X_i] E[X_j]\right]$$
$$= \sum_i Var(X_i) + \sum_{i\neq j} Cov(X_i, X_j)$$

if ~~X~~ X, Y independent $E[XY] = E[X]E[Y]$, $Cov(X,Y) = 0$

(7) Correlation Coefficient (Dimensionless version of covariance)

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

$$= \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

$\rho \in [-1,1]$

if $|\rho| = 1 \iff (X-E[X]) = c(Y-E[Y])$ linear related

$\rho = 0 \Leftarrow$ independent. converse is not true

Lecture 12: Iterated Expectations; Sum of a Random Number of Random Variables.

(1) Conditional expectations

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

$$E[X|Y=y] = \int_x x f_{X|Y}(x|y)$$

(2) Law of iterated expectations:

$$E[E[X|Y]] = \sum_y E[X|Y=y] P_Y(y) = E[X]$$ Total Expectation Thm

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

(4) Law of total variance:

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

average Variability of X within each Y groups

variability of X between Y groups

proof: $Var(X) = E[X^2] - (E[X])^2$

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

$\Rightarrow E[Var(X|Y)] = E[X^2] - E[(E[X|Y])^2]$ ①

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

$= E[(E[X|Y])^2] - E[X]^2$ ②

① + ② $\Rightarrow E[Var(X|Y)] + Var(E[X|Y]) = E[X^2] - E[X]^2 = Var(X)$

Lecture 13: Bernoulli Process

memoryless, discrete time

(1) Bernoulli process: a sequence of independent Bernoulli trials,

at each trial, i: $P(\text{success}) = P(X_i = 1) = p$

$P(\text{failure}) = P(X_i = 0) = 1 - p$

(2) Interarrival times: T_1 : number of trials until first success

$$P(T_1 = t) = (1-p)^{t-1} p$$
$$E[T_1] = \frac{1}{p}$$
$$\text{Var}(T_1) = \frac{(1-p)}{p^2}$$

} geometric distribution

(3) Time of the k-th arrival

$$P_{Y_k}(t) = P(Y_k = t) = P(\underbrace{k-1 \text{ arrivals in } [1, t-1]}_{\text{independent}}, \underbrace{\text{arrival at time } k}_{\text{independent}})$$

$$= C_{t-1}^{k-1} p^{k-1} (1-p)^{t-k} \cdot p \Rightarrow \text{Pascal PMF}$$

$$E[Y_k] = \frac{k}{p} \quad \Rightarrow \because Y_k = \underbrace{T_1 + T_2 + \cdots T_k}_{\text{independent to each other}}$$

$$\text{Var}[Y_k] = \frac{k(1-p)}{p^2}$$

$$\therefore E[Y_k] = \sum_{i=1}^{k} E[T_i] \qquad \text{Var}[Y_k] = \sum_{i=1}^{k} \text{Var}[T_i]$$

(4) Merging of independent Bernoulli Processes yields a Bernoulli Process

Lecture 14: Poisson Process I

memoryless, continuous time

(1) Time homogeneity:

$P(k, \tau)$ = Prob. of k arrivals in the interval of duration τ

* # of arrivals in disjoint time intervals are independent.

(2) Small interval probabilities:

For very small δ:

$$P(k, \delta) \simeq \begin{cases} 1 - \lambda\delta, & \text{if } k = 0 \\ \lambda\delta, & \text{if } k = 1 \\ 0, & \text{if } k > 1 \end{cases}$$

where λ is "arrival rate"

(3) Approximate Poisson process by Bernoulli process

$$P(k,\tau) = \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!}, \quad k = 0, 1, \ldots$$

$E[N_t] = \lambda t$
$Var[N_t] = \lambda t$
} based on Bernoulli

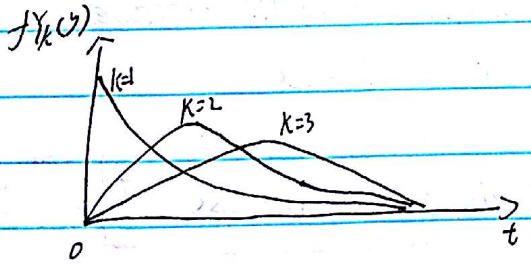
$E[N_t] = np$,
$Var[N_t] = np(1-p)$, $n = \frac{t}{\delta}$, $p = \lambda\delta$
let $\delta \to 0$, $p \to 0$

(4) Interarrival times : time to the first arrival

* Y_k time of kth arrival
* Erlang distribution :

$$f_{Y_k}(y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!}, \quad y \geq 0$$ depends on k and λ

$$\therefore \delta \cdot f_{Y_k}(t) = P(t \leq Y_k \leq t+\delta)$$
$$= P(k-1 \text{ arrivals in } [0, t]) \cdot \lambda\delta$$
$$= \frac{(\lambda t)^{k-1} e^{-\lambda t}}{(k-1)!} \cdot \lambda\delta$$



when $k=1$
$f_{Y_1}(y) = \lambda e^{-\lambda y}$
exponential distribution

given a specific period of time, the number of arrivals

$$E[Y_k] = k \cdot E[Y_1] = k \cdot \int_0^\infty \lambda e^{-\lambda y} \cdot y \, dy$$

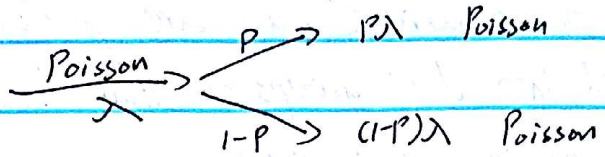
(5) Sum of independent Poisson random variables is Poisson.

(6) Merging of independent Poisson processes is Poisson

merged rate $\lambda = \sum_i \lambda_i$

Lecture 15: Poisson Process II

(1) Splitting of Poisson process: still Poisson process



(2) Random incidence in "renewal process"

Lecture 16: Markov Chains I

(1) Finite state Markov chains

* X_n: state after n transitions
  - belongs to a finite set, e.g., $\{1, \dots, m\}$
  - X_0 is either given or random
* Markov property / assumption: (given current state, the past does not matter)

$$P_{ij} = P(X_{n+1} = j \mid X_n = i) = P(X_{n+1} = j \mid X_n = i, X_{n-1}, \dots, X_0)$$

(2) State occupancy probabilities (given initial state i):

$$r_{ij}(n) = P(X_n = j \mid X_0 = i)$$

Key recursion: $r_{ij}(n) = \sum_{k=1}^{m} r_{ik}(n-1) P_{kj}$

with random initial state: $P(X_n = j) = \sum_{i=1}^{m} P(X_0 = i)\, r_{ij}(n)$

(3) Recurrent and transient states

state i is ~~recurre~~ recurrent if:

starting from i, and from whenever you can go, there is a way of returning to i.

If not ~~recu~~ recurrent, called transient

Lecture 17: Markov Chains II

(1) Recurrent Class: collection of recurrent states that "communicate" to each other and to no other state

(2) Periodic states

The states in a recurrent class are periodic if they can be grouped into $d > 1$ groups so that all transitions from one group lead to the next group

(3) Steady-State Probabilities

$r_{ij}(n)$ converge to π_j (independent of the initial state i)

if * recurrent states are all in a single class

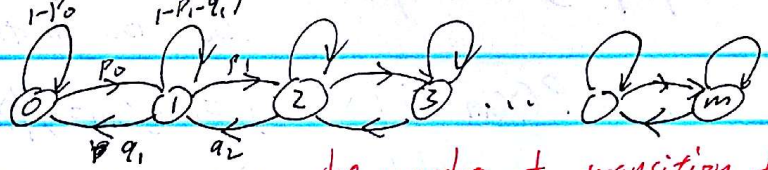
and * single recurrent class is not periodic

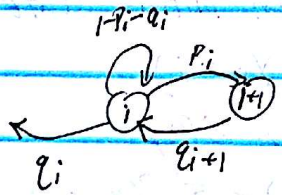
(4) from $r_{ij}(n) = \sum_k r_{ik}(n-1) p_{kj}$

$\Rightarrow \pi_j = \sum_k \pi_k p_{kj}$, for all j Balance Equation

solve π_j and $\sum_j \pi_j = 1$

(5) Birth-death process





the number of transition from i to $i+1$, $t_{i,i+1}$ must be $\simeq t_{i+1,i}$. Because $t_{i,i+1} = t_{i+1,i}$, or

$t_{i,i+1} = t_{i+1,i} - 1$ or $t_{i,i+1} = t_{i+1,i} + 1$

so $\pi_i p_i = \pi_{i+1} q_{i+1}$

Lecture 18: Markov Chains III

(1) Absorption probability.

a_i is the probability that process eventually settles in transient states given the initial state i.

$a_i = \sum_j a_j P_{ij}$ for all other i

(2) Expected time to absorption

μ_i is the expected number of transitions until reaching the absorbing state given the initial state i

$$\mu_i = 1 + \sum_j P_{ij} \mu_j$$

Lecture 19: Weak Law of Large Numbers

(1) Markov inequality (discrete variable)

$X \geq 0$, $E[X] = \sum_x x P_X(x) \geq \sum_{x \geq a} x P_X(x)$

$$\geq \sum_{x \geq a} a P_X(x) = a \sum_{x \geq a} P_X(x)$$

$$= a P(X \geq a)$$

$\mu = E[X]$

$$\text{Var}(X) = E[(X - \mu)^2] \geq P((X-\mu)^2 \geq a^2)\, a^2$$

$$= P(|X - \mu| \geq a)\, a^2$$

(2) Chebyshev's inequality (continuous variable)

$x \in [\mu + c, \infty]$ when $(x-\mu)^2 \geq c^2$

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

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

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

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

(3) Convergence of the sample mean

$X_1, X_2 \ldots$ i.i.d. with finite mean μ and variance σ^2

$M_n = \frac{X_1 + \cdots + X_n}{n}$ is a random variable

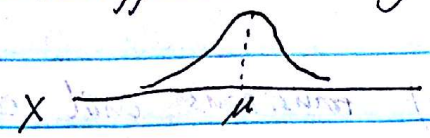
$$E[M_n] = \frac{E[X_1] + \cdots E[X_n]}{n} = \frac{n \cdot \mu}{n} = \mu$$

$$\text{Var}(M_n) = \frac{n \cdot \sigma^2}{n^2} = \frac{\sigma^2}{n}$$

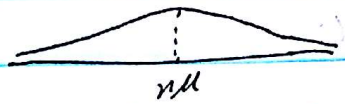
$$P(|M_n - \mu| \geq \varepsilon) \leq \frac{\text{Var}(M_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}$$

for $\forall \varepsilon > 0$, $P(|M_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} \xrightarrow{n \to \infty} 0$

(4) different scalings

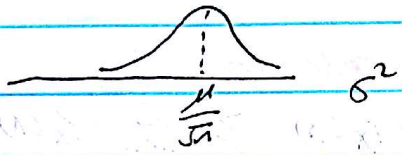


$S_n = \sum_{i=1}^{n} X_i$



$\frac{S_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$

$\frac{S_n}{\sqrt{n}}$



(5) The central limit theorem

"standardized" $S_n = X_1 + \dots X_n$

zero mean, unit variance → $Z_n = \frac{S_n - E[S_n]}{\sigma_{S_n}} = \frac{S_n - nE[X]}{\sqrt{n}\sigma}$

~~$\sigma_{S_n} = \sqrt{n}\sigma$~~ $\sigma^2_{S_n} = n\sigma^2$

Theorem: for $\forall c$: $P(Z_n \le c) \to P(Z \le c)$

Z is a standard normal r.v.

$P(Z \le c)$ is the standard normal CDF, available from the normal tables.

$S_n = \sqrt{n}\sigma Z_n + nE[X]$

Lecture 20: Central Limit Theorem

(1) The pollster's problem using the CLT

* f: fraction of population that "..."
* ith (randomly selected) person polled:

$$X_i = \begin{cases} 1 & \text{if yes} \\ 0 & \text{if no} \end{cases}$$

* $M_n = (X_1 + \cdots + X_n)/n$

suppose we want $P(|M_n - f| \geq 0.01) \leq 0.05$

$|M_n - f| = \left|\frac{X_1 + \cdots X_n - nf}{n}\right| \geq 0.01$

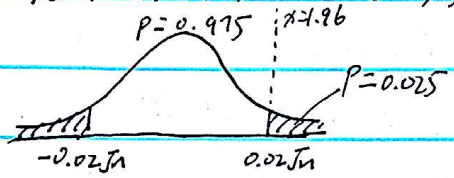
$\left|\frac{X_1 + \cdots X_n - nf}{\sqrt{n}\sigma}\right| \geq \frac{0.01\sqrt{n}}{\sigma}$

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

$\because \sigma = \sqrt{f(1-f)}$ see Bernoulli variable

$\therefore \sigma \leq \frac{1}{2}$

$\therefore P(|Z| \geq 0.01\sqrt{n}/\sigma) \leq P(|Z| \geq 0.02\sqrt{n})$, and then check table for ~~$P \leq 0.05$~~



$\therefore n = 9604$

(2) CLT apply to Binomial

* Fix p, where $0 < p < 1$
* X_i : Bernoulli (p)
* $S_n = X_1 + \cdots + X_n$: ~~Ber~~ Binomial (n, p), with mean np, variance $np(1-p)$ ∵ * $Var(X) = p(1-p)^2 + (1-p)(0-p)^2 = p(1-p)$, ~~$Var(X)$~~ $Var(X_1 + \cdots X_n) = nVar(X)$
* CDF of $\frac{S_n - np}{\sqrt{np(1-p)}}$ → standard normal

(3) De Moivre-Laplace CLT (for binomial)

* using the $\frac{1}{2}$ correction for binomial approximation

e.g. $n = 36$, $p = 0.5$, mean np, variance $np(1-p)$

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

$\iff \frac{18.5 - np}{\sqrt{np(1-p)}} \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq \frac{19.5 - np}{\sqrt{np(1-p)}}$

Lecture 21: Bayesian Statistical Inference I

(1) Bayesian inference: Use Bayes rule

* Hypothesis testing (discrete data)

- discrete data

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

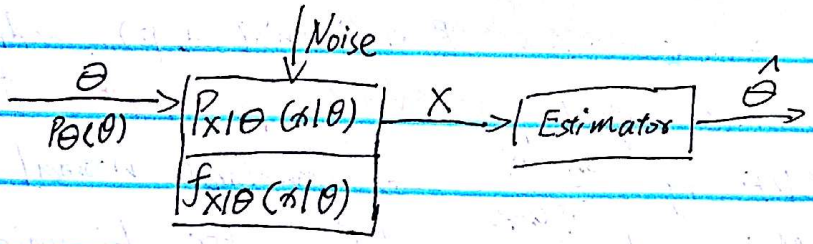
- continuous data

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

* Estimation (continuous data)

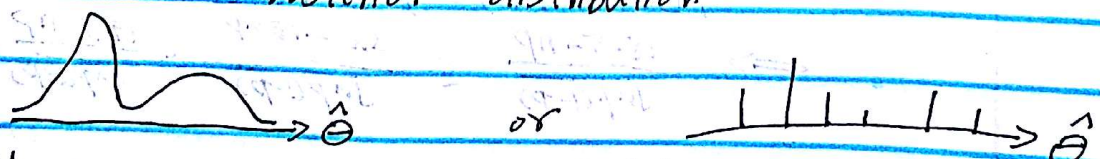
- continuous $f_{\Theta|X}(\theta|x) = \frac{f_\Theta(\theta) f_{X|\Theta}(x|\theta)}{f_X(x)}$

- discrete $f_{\Theta|X}(\theta|x) = \frac{f_\Theta(\theta) P_{X|\Theta}(x|\theta)}{P_X(x)}$



(2) * $f_\Theta(\theta)$ or $P_\Theta(\theta)$ is called prior distribution.

* pdf $f_{\Theta|X}(\cdot|x)$ or pmf $P_{\Theta|X}(\cdot|x)$ is called Posterior distribution



(3) Maximum a posteriori probability (MAP):

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

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

Conditional expectation: $E[\Theta|X=x] = \int \theta f_{\Theta|X}(\theta|x)\,d\theta$

(4) Least Mean Squares Estimation (LMS)

Knowing distribution of Θ, find ~~estimation~~ estimate c, to minimize $E[(\Theta-c)^2]$

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

$$\frac{dE[(\Theta-c)^2]}{dc} = -2E[\Theta] + 2c = 0$$

$$\therefore c = E[\Theta]$$

Optimal mean squared error: $E[(\Theta-c)^2] = Var(\Theta)$

(5) Conditional LMS Estimation

* Two r.v.'s Θ, X
* observe that $X = x$
* $E[(\Theta-c)^2|X=x]$ is minimized by $c = E[\Theta|X=x]$
* ~~E[Θ~~ assume there is another estimator ~~g(x)~~ $g(X)$

we can have that: ~~$E[(\Theta - E[\Theta|X=x])^2]$~~

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

because x is random, so if take average for all x

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

take expectation of two sides

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

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

Lecture 22: Bayesian Statistical Inference II

(1) Some properties of LMS estimation

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

- Estimation error: $\tilde{\Theta} = \hat{\Theta} - \Theta$

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

proof: $E[\tilde{\Theta}|X] = E[\hat{\Theta} - \Theta | X]$

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

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

* $E[\tilde{\Theta} h(X)] = 0$, for any function h

proof: $E[\tilde{\Theta} h(X)|X] = h(X) E[\tilde{\Theta}|X] = 0$

take expectation to both sides: $E[\tilde{\Theta} h(X)] = 0$

* $Cov(\tilde{\Theta}, \hat{\Theta}) = 0$

proof: $Cov(\tilde{\Theta}, \hat{\Theta}) = \underbrace{E[\tilde{\Theta}\hat{\Theta}]}_{0} - \underbrace{E[\tilde{\Theta}]}_{0} E[\hat{\Theta}] = 0$

* $Var(\Theta) = Var(\hat{\Theta}) + Var(\tilde{\Theta})$

proof: $\because \Theta = \hat{\Theta} - \tilde{\Theta}$

(2) Linear LMS

* $\hat{\Theta} = aX + b$
* minimize $E[(\Theta - aX - b)^2]$
* $\hat{\Theta}_L = E[\Theta] + \frac{Cov(X,\Theta)}{Var(X)}(X - E[X])$

optimal $a = \frac{Cov(X,\Theta)}{Var(X)}$

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

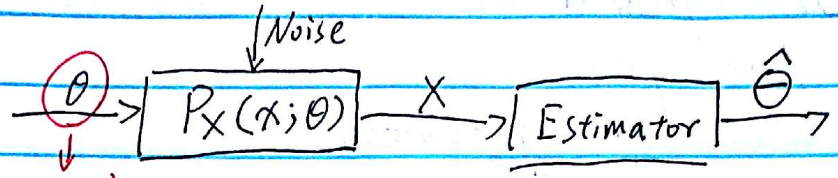
* properties:

$E[(\hat{\Theta}_L - \Theta)^2] = (1 - \rho^2)\sigma_\Theta^2$ → $Var(\Theta)$

ρ: correlation coefficient $= Corr(X, \Theta)$

Lecture 23: Classical Statistical Inference I

(1)



θ is a number, not random

(2) Maximum Likelihood Estimation

* Model with unknown parameter(s):
  $X \sim P_X(x;\theta)$ or $X \sim f_X(x;\theta)$
* Pick θ that "makes data most likely"
  $\hat{\theta}_{ML} = \arg\max_\theta P_X(x;\theta)$ or $= \arg\max_\theta f_X(x;\theta)$
* Compare to Bayesian MAP estimation:
  $\hat{\theta}_{MAP} = \arg\max_\theta \frac{P_{X|\Theta}(x|\theta)\,P_\Theta(\theta)}{P_X(x)}$ ($P_X(x)$: constant)

same as $\hat{\theta}_{MAP}$ when $\hat{\theta}_{ML}$

∴ ML estimation is MAP with uniform distribution of prior ~~proba~~ distribution.

(3) Desirable properties of estimators

* Unbiased: $E[\hat{\theta}_n] = \theta$
* Consistent: $\hat{\theta}_n \to \theta$ (converge in probability)
* small mean squared error

$$E[(\hat{\theta}-\theta)^2] = \mathrm{var}(\hat{\theta}-\theta) + (E[\hat{\theta}-\theta])^2$$
$$= \mathrm{var}(\hat{\theta}) + (\mathrm{bias})^2$$

∵ $E[X^2] = \mathrm{Var}(X) + (EX)^2$

(4) Confidence intervals (CIs)

An $1-\alpha$ confidence interval is a (random) interval $[\hat{\theta}_n^-, \hat{\theta}_n^+]$ s.t. $P(\hat{\theta}_n^- \le \theta \le \hat{\theta}_n^+) \ge 1-\alpha$, $\forall\theta$

Lecture 24: Classical Inference II

(1) Linear Regression

* Least squares

$Y = \theta_0 + \theta_1 X$

$$\min_{\theta_0, \theta_1} \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2$$

* Maximum Likelihood

$Y = \theta_0 + \theta_1 X + W$; $W \sim N(0, \sigma^2)$ i.i.d

$f_W(w) = c e^{-\frac{w^2}{2\sigma^2}}$

$\therefore$ likelihood function $f_{X,Y|\theta}(x, y; \theta) = c \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2\right\}$

take logs, same as least squares

$\therefore$ Least squares $\longleftrightarrow$ pretend W i.i.d. normal

(2) Solution to $Y = \theta_0 + \theta_1 X$

$$\min_{\theta_0, \theta_1} \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2$$

is $$\hat{\theta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}$$

interpretation: $Y = \theta_0 + \theta_1 X + W$, multiply X to both sides

$\longrightarrow$ $YX = \theta_0 X + \theta_1 X^2 + WX$; take expectation to both sides

$\longrightarrow E[YX] = \theta_0 E[X] + \theta_1 E[X^2] + E[W]E[X]$

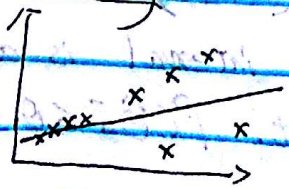
plug in $\theta_0 = E[Y] - \theta_1 E[X]$

$\longrightarrow E[YX] = E[X]E[Y] - \theta_1 E[X]^2 + \theta_1 E[X^2] + E[W]E[X]$

$E[W] = 0 \longrightarrow Cov(X, Y) = \theta_1 Var(X)$

$$\hat{\theta}_1 = \frac{Cov(X, Y)}{Var(X)}$$

(3) Heteroskedasticity



(4) Hypothesis testing

two possible hypotheses : H_0 and H_1

$X \sim P_X(x; H_0)$ [or $f_X(x; H_0)$]

or

$X \sim P_X(x; H_1)$ [or $f_X(x; H_1)$]

* Bayesian case : choose H_1 if (likelihood ratio test)

$$P(H_1 | X = x) > P(H_0 | X = x)$$

$$\Rightarrow \frac{P(X = x | H_1) P(H_1)}{P(X = x)} > \frac{P(X = x | H_0) P(H_0)}{P(X = x)}$$

$$\Rightarrow \frac{P(X = x | H_1)}{P(X = x | H_0)} > \frac{P(H_0)}{P(H_1)}$$

* Nonbayesian version: choose H_1 if

$$\frac{P(X = x; H_1)}{P(X = x; H_0)} > \xi \quad \text{or} \quad \frac{f_X(x; H_1)}{f_X(x; H_0)} > \xi$$

ξ: trade off FN and FP

Lecture 25 : Classical Inference III ; Course Overview

(1) Chi-square test :

$$\sum_i \frac{(N_i - np_i)^2}{np_i} > \xi$$

(2) Test the correctness of PDF

- partition the range into bins
- np_i : expected incidence of bin i from the PDF
- N_i : observed incidence of bin i
- use chi-square test

(3) Kolmogorov-Smirnov test :

form empirical CDF, $\hat{F}_X$, from data

$$D_n = \max_x |F_X(x) - \hat{F}_X(x)|$$

$$P(\sqrt{n} D_n \geq 1.36) \approx 0.05$$