Probability

Probability Axiom

Probability Space 由以下3个元素定义

1. sample space Ω that contains all possible outcomes
2. σ-algebra $\mathcal{F}$, a collection of subset of Ω
3. probability measure $\mathbb{P}$, a function that assign non-negative value on every set of $\mathcal{F}$

σ-field: $\mathcal{F}$ 被称为是以下条件下 collection of subset of Ω

1. $\phi \in \mathcal{F}$
2. $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$
3. $A_i \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$
   - A: $\mathcal{F}$-measurable set / event
   - $\sigma(\omega)$: 包含 ω 的最小的 σ-field

probability measure: $\mathbb{P}: \mathcal{F} \to [0,1]$ 被以下 axiom 的 function

1. $\mathbb{P}(\Omega) = 1$
2. $\mathbb{P}(\phi) = 0$
3. $\mathbb{P}(\bigcup A_i) = \sum_1^{\infty} \mathbb{P}(A_i)$ countable additivity

Theorem: continuity of probability measure.

① $\mathbb{P}$ is a probiliby measure

② $A_i \subset A_{i+1}$ for all i and $A = \bigcup_1^{\infty} A_i \Rightarrow \lim_{n\to\infty} \mathbb{P}(A_i) = \mathbb{P}(A)$

③ $A_i \supset A_{i+1}$ for all i and $A = \bigcap_1^{\infty} A_i \Rightarrow \lim_{n\to\infty} \mathbb{P}(A_i) = \mathbb{P}(A)$

Measurable Function

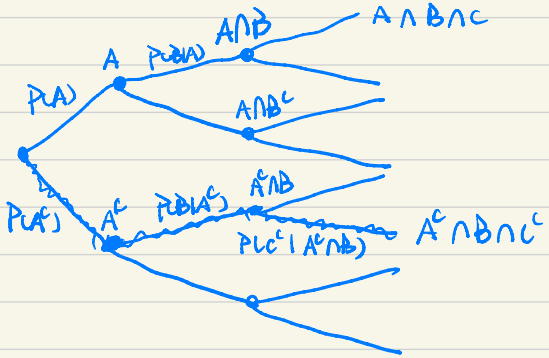
$(\Omega_1, \mathcal{F}_1)$ 与 $(\Omega_2, \mathcal{F}_2)$ be 2 measurable space. A function $f: \Omega_1 \to \Omega_2$ is called $(\mathcal{F}_1, \mathcal{F}_2)$ measurable if $f^{-1}(B) \in \mathcal{F}_1$ for every $B \in \mathcal{F}_2$

Conditional probability

- Condition lead to revised (conditional) probabilities that take into account parcial information.
- Useful in 'divide and conquer' for complex problem

1. Multiplication Rule

$$P(A_1 \cap A_2 \cdots A_n) = P(A_1) \prod_{i=2}^{n} P(A_i | A_1 \cap \cdots A_{i-1})$$



2. Bayes Rule

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

3. Independence

Events $A_1, \cdots A_n$ are independent if

$P(A_i \cap A_j \cdots \cap A_m) = P(A_i) \cdot P(A_j) \cdots P(A_m)$ for any distinct indices

↗ & the pairwise independence

- Conditional PMF

$$P_{X|Y}(x|y) = \frac{P_{XY}(x,y)}{P_Y(y)} \quad \text{defined for } y \text{ such that } P_Y(y) > 0$$

- Conditional PDF

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

- Mixed

$$\checkmark\ P_{K|Y}(k|y) = \frac{P_K(k) f_{Y|K}(y|k)}{f_Y(y)}$$

$$\checkmark\ f_{Y|K}(y|k) = \frac{f_Y(y) P_{K|Y}(k|y)}{P_K(k)}$$

- Multiplication

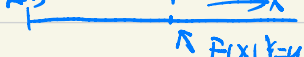
$$f_{X,Y,Z}(x,y,z) = f_Z(z) f_{Y|Z}(y|z) f_{X|Y,Z}(x|y,z)$$

- Moments

$E(X|Y) = g(Y)$ a function of Y

$g(y) = E[X|Y=y] \Rightarrow E[E(X|Y)] = E(X)$

i.e. Forecast Revision



$E(X|Y=y)$: revised forecast

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

$$\text{Var}(X) = \underbrace{E[\text{Var}(X|Y)]}_{\text{① average variability within sectors}} + \underbrace{\text{Var}[E(X|Y)]}_{\text{② variability between sectors}}$$

Counting Problem

- Permutation : # of ways of ordering n elements
  $n! = n(n-1)\cdots 1$

- # of subsets of $\{1, \cdots, n\}$
  2^n

- # of k-element subsets of a given n-element set
  $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ Binomial

  $\sum_{k=0}^{n} \binom{n}{k} = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n} =$ # of all subsets $= 2^n$

- Partitions
  # of partitions $= \frac{n!}{n_1! n_2! \cdots n_r!}$ multi-nomial

Derived Distribution

- Discrete:

$$Y = g(X) \Rightarrow P_Y(y) = P(g(X)=y) = \sum_{x: g(x)=y} P_X(x)$$

i.e. when $y = ax+b$ $\quad P_Y(y) = P_X\left(\frac{y-b}{a}\right)$

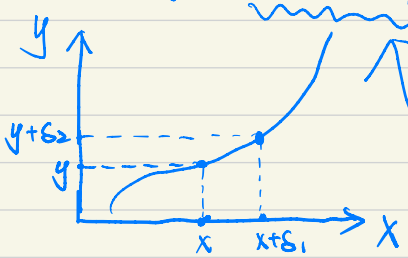
- Continuous

$Y = g(X)$

1. CDF: $F_Y(y) = P(Y \le y) = P(g(X) \le y)$
2. Take derivative: $f_Y(y) = \frac{dF_Y}{dy}(y)$

When g is monotonic, $x_1 > x_2 \Rightarrow g(x_1) > g(x_2)$

1. $F_Y(y) = P(Y \le y) = P(g(X) \le y) = P(X \le g^{-1}(y)) = F_X(g^{-1}(y))$
2. $f_Y(y) = f_X(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right|$



$y = g(x) \rightarrow \delta_2 \approx \delta_1 \frac{dg}{dx}(x)$

$x = g^{-1}(y) \rightarrow \delta_1 \approx \delta_2 \frac{dg^{-1}}{dy}(y)$

$\delta_2 f_Y(y) \approx P(y \le Y \le y+\delta_2) = P(x \le X \le x+\delta_1)$

$\approx \delta_1 f_X(x) \approx \delta_2 \frac{dg^{-1}}{dy}(y) f_X(x)$

Sum of Random Variables

Convolution $Z = X + Y$ (X, Y independent)

1. discrete

$$P_Z(z) = \sum_x P_X(x)\, P_Y(z-x)$$

2. continuous

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

$$f_{Z|X}(z|x) = f_Y(z-x)$$

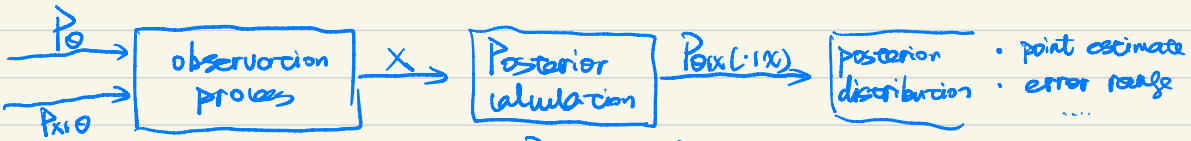
$$\mathrm{Var}(X_1 + \cdots + X_n) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{cov}(X_i, X_j)$$

$$\rho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = E\left[\frac{(X-\mu_X)}{\sigma_X}\,\frac{(Y-\mu_Y)}{\sigma_Y}\right]$$

- $\rho = 1 \Rightarrow X - \mu_X = c(Y - \mu_Y)$
- $\rho(aX+b, Y) = \mathrm{sign}(a)\, \rho_{XY}$

Bayesian Inference framework

- Unknown Θ
  - treat as random variable
  - prior distribution P_Θ or f_Θ from
    1. symmetry
    2. known range
    3. earlier study
    4. arbitrary
- Observation X
  - Observation Model $P_{X|\Theta}$ or $f_{X|\Theta}$



estimate: $\hat{\theta} = g(x)$ (number)

estimator: $\hat{\Theta} = g(X)$ (R.V.)

- Max Posterior Probability (MAP)
  $P_{\Theta|X}(\theta^*|x) = \max_\theta P_{\Theta|X}(\theta|x)$
  $f_{\Theta|X}(\theta^*|x) = \max_\theta f_{\Theta|X}(\theta|x)$
- Conditional Expectation (LMS)
  $E[\Theta|X=x]$

Performance evaluation

1. conditional prob of error
   $P(\hat{\Theta} \neq \Theta | X = x)$ smallest under MAP
2. Overall prob of error
   $P(\hat{\Theta} \neq \Theta) = \sum_x P(\hat{\Theta} \neq \Theta | X = x) P_X(x)$
   $= \sum_\theta P(\hat{\Theta} \neq \Theta | \Theta = \theta) P_\Theta(\theta)$
3. Conditional mean square error
   $E[(\hat{\Theta} - \Theta)^2 | X = x]$
4. Mean square error
   $E[(\hat{\Theta} - \Theta)^2]$

LMS Estimation

- Without Observation

  $\min_{\hat{\theta}} E[(\Theta-\hat{\theta})^2]$ → Mean square error

  $\Rightarrow \hat{\theta} = E[\Theta]$

  the $E[(\Theta-\hat{\theta})^2] = \mathrm{Var}(\Theta) \le E[(\Theta-c)^2]$ for all c

- LMS estimate of Θ based on X

  ① unknown Θ: prior $P_\Theta(\theta)$

  ② observation X: model $P_{X|\Theta}(x|\theta)$

  $\min_{\hat{\theta}} E[(\Theta-\hat{\theta})^2 | X=x]$

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

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

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

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

  $\hat{\Theta}_{LMS} = E[\Theta|X]$ minimize $E[(\Theta-g(X))^2]$ over all estimator g

- Performance

  LMS estimate: $\hat{\theta} = E[\Theta|X=x]$

  LMS estimator: $\hat{\Theta} = E[\Theta|X]$

  ① Expected performance once we have observation

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

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

  ② Expected performance of the design (over all possible x)

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

  $= E[\mathrm{Var}(\Theta|X)]$

  property:

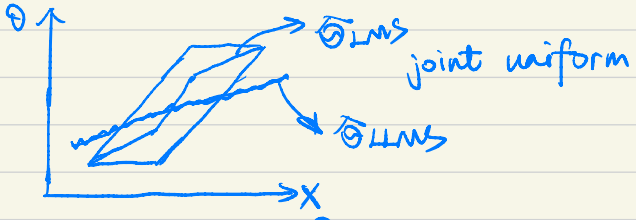
  ① $\mathrm{cov}(\hat{\Theta}, \hat{\Theta}-\Theta) = 0$

  ② $\mathrm{Var}(\Theta) = \mathrm{Var}(\hat{\Theta}) + \mathrm{Var}(\hat{\Theta}-\Theta)$

LLMS estimation

The conditional expectation may be hard to compute. In that case, we can restrict to estimators $\hat{\Theta} = aX + b$ that minimize the mean square error.

- Estimator $\hat{\Theta} = g(X)$
- $\min E[(\hat{\Theta} - \Theta)^2] \Rightarrow \hat{\Theta}_{LMS} = E[\Theta | X]$
- $\min_{a,b} E[(\Theta - aX - b)^2] \Rightarrow \hat{\Theta}_{LLMS}$
- If $E[\Theta|X]$ is linear in $X \Rightarrow \hat{\Theta}_{LMS} = \hat{\Theta}_{LLMS}$



$$\hat{\Theta}_{LLMS} = E[\Theta] + \frac{cov(\Theta, X)}{var(X)}(X - E[X])$$

$$= E[\Theta] + \rho \frac{\sigma_\Theta}{\sigma_X}(X - E[X])$$

- $\rho = 0 \Rightarrow \hat{\Theta}_{LMS} = E[\Theta]$
- $E[(\hat{\Theta}_{LMS} - \Theta)^2] = (1 - \rho^2) var(\Theta)$

Multiple Observation case:

$$\min_{a_1 \cdots a_n b} E[(a_1 X_1 + \cdots + a_n X_n + b - \Theta)^2]$$

Inequality

- bound $P(X \geq a)$ based on limited information of the Distrib
- Markov (based on mean)
- Chebyshev (based on mean and variance)

Markov Inequality

If $X \geq 0$ and $a > 0$, $P(X \geq a) \leq \frac{EX}{a}$

(If $X > 0$ and EX small, X is unlikely to be large)

Chebyshev Inequality

X with finite mean and variance. $P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}$

(If variance small, X unlikely to be far away from mean)

Jensen's Inequality

g is convex $\rightarrow$ $g(E(X), E(Y)) \leq E[g(X, Y)]$

R.V. convergence

WLLN:

$X_1, X_2, \cdots$ iid; finite mean and variance σ^2

- sample mean: $M_n = \frac{X_1 + \cdots + X_n}{n}$
- $E(M_n) = \mu$.
- $\mathrm{Var}(M_n) = \frac{\sigma^2}{n}$

$$P(|M_n - \mu| \geq \varepsilon) \leq \frac{\mathrm{Var}(M_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \xrightarrow{n\to\infty} 0 \quad (\text{fix } \varepsilon)$$

WLLN: $P(|M_n - \mu| \geq \varepsilon)$

$$= P\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \varepsilon\right) \xrightarrow{n\to\infty} 0$$

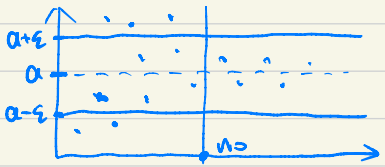
Interpretation: many observations with noise, the sample mean M_n is unlikely to be far from the true μ.

Convergence in probability

ordinary convergence

- sequence of number $a_n \to a$

(a_n eventually sets and stay close to a)

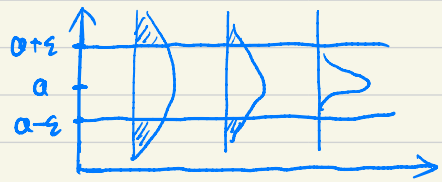


- For every $\varepsilon > 0$, there exists n_0 such that for every $n > n_0$ $|a_n - a| < \varepsilon$

convergence in probability

- Sequence of R.V. $Y_n \to a$

(almost all of the pdf/pmf of Y_n eventually get concentrated close to a)



for any $\varepsilon > 0$ $P(|Y_n - a| \geq \varepsilon) \to 0$

does not mean convergence in Expectation (ie:)

$$Y_n = \begin{cases} n^2, & \frac{1}{n} \\ 0, & 1 - \frac{1}{n} \end{cases} \qquad Y_n \to 0, \quad E(Y_n) \to \infty$$

Bernoulli Process

A sequence of Bernoulli trials

$\begin{cases} P(X_i = 1) = p \\ P(X_i = 0) = 1-p \end{cases}$ independence + Time-homogeneity

number of success/arrival

- $S = X_1 + \cdots + X_n$
- $P(S = k) = \binom{n}{k} p^k (1-p)^{n-k}$ for $k = 0, \cdots, n$
- $E(S) = np$
- $\text{Var}(S) = np(1-p)$

Time until first arrival

- $T_1 = \min\{i : X_i = 1\}$
- $P(T_1 = k) = (1-p)^{k-1} p$ for $k = 1, \cdots$
- $E(T_1) = 1/p$
- $\text{Var}(T_1) = (1-p)/p^2$

Memoryless

- Fresh start after a random time N

  N: a causally determined (without future information)

  The process $X_{N+1}, X_{N+2}, \cdots$ is

  ① Bernoulli Process

  ② independent of $N, X_1, \cdots X_N$

Time of the k-th arrival/success

Y_k: time of k-th arrival

T_k: k th inter-arrival time $= Y_k - Y_{k-1}$ $(k \geq 2)$

$Y_k = T_1 + \cdots + T_k$

T_2 is independent of $T_1 \sim \text{Geo}(p)$

$Y_k = T_1 + \cdots + T_k$

T_i are i.i.d Geometric(p)

$E(Y_k) = kE(T_i) = \frac{k}{p}$ $\quad Var(Y_k) = \frac{k(1-p)}{p^2}$

$P_{Y_k}(t) = \binom{t-1}{k-1} p^k (1-p)^{t-k}$ $\quad t = k, k+1, \cdots$

Merging / Split

- $X_t \sim$ Poisson(p) $\quad Y_t \sim$ Poisson(q)

  $Z_t = g(X_t, Y_t)$

- $X_t \sim$ Poisson(p) → q: Poisson(pq); $1-q$: Poisson$(p(1-q))$

Poisson approximation to binomial

- when n large, p small, $\lambda = np$ moderate
- Number of arrivals S in n slots:

$$P_S(k) = \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

$$= \frac{n}{n} \cdot \frac{n-1}{n} \cdot \cdots \cdot \frac{n-k+1}{n} \cdot \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k}$$

$$\xrightarrow{n \to \infty} 1 \cdot 1 \cdot \cdots \cdot 1 \cdot \frac{\lambda^k}{k!} e^{-\lambda} \cdot 1$$

Poisson Process

- Prob of k arrivals in interval of duration δ

$$P(k,\delta) = \begin{cases} 1-\lambda\delta+o(\delta^2) & \text{if } k=0 \\ \lambda\delta+o(\delta^2) & \text{if } k=1 \\ o(\delta^2) & \text{if } k>1 \end{cases}$$

λ: arrival rate

- N_τ: arrival in $[0,\tau]$

$n = \tau/\delta$ = interval / length of slot

$N_\tau \approx$ binomial

$p = \lambda\delta + o(\delta^2)$

$np = \lambda\tau + o(\delta) \approx \lambda\tau$

- $N_\tau \approx$ Binomial (n,p)

$n = \tau/\delta$, $p = \lambda\delta + o(\delta^2)$

$P(k,\tau) = P(N_\tau = k) = \frac{(\lambda\tau)^k}{k!} e^{-\lambda\tau}$

$E(N_\tau) = \lambda\tau$

$\text{var}(N_\tau) = \lambda\tau$

T_1: time until first arrival

$P(T_1 < t) = 1 - P(T_1 \geq t) = 1 - P(0,t) = 1 - e^{-\lambda t}$

$T_1 \sim$ exponential

Y_k: time of the k-th arrival

$P(Y_k \leq y) = \sum_{n=k}^{\infty} P(n,y)$

- $f_{Y_k}(y)\delta \approx P(y \leq Y_k \leq y+\delta)$

scenario 1: $k-1$ arrivals in $[0,y]$, k-th in $[y,y+\delta]$ $\approx P(k-1,y)\lambda\delta$

scenario 2: $k-2$ arrivals in $[0,y]$, $k-1,k$ in $[y,y+\delta]$ $+ P(k-2,y)\,o(\delta^2)$

$+ P(k-3,y)\,o(\delta^3) + \cdots$

$\xrightarrow{\text{as } \delta \to 0} \approx P(k-1,y)\lambda\delta$

Erlang: $f_{Y_k}(y) = \lambda P(k-1,y) = \lambda \frac{(\lambda y)^{k-1} e^{-\lambda y}}{(k-1)!}$

$Y_k = T_1 + \cdots + T_k$ → a equivalent definition

- sum of iid exponencials
- $E(Y_k) = k/\lambda$
- $\text{Var}(Y_k) = k/\lambda^2$

Poisson v.s. Bernoulli

$n = \tau/\delta$
$p = \lambda\delta$
$np = \lambda\tau$

	Poisson	Bernoulli
Time of arrival	Continuous	Discrete
Arrival rate	λ/unit time	p/per trial
PMF of # of arrivals	Poisson	Binomial
Interarrival time	Exponencial	Geometric
Time to k-th arrival	Erlang	Pascal

Merge of 2 independant Poisson Process

M : Poisson(u) → $M+N$: Poisson($u+v$)
N : Poisson(v)

- P(arrival is M | arrival at time t) = $\frac{u}{u+v}$
- P(k out of first n arrival is M) = $\binom{n}{k}\left(\frac{u}{u+v}\right)^k\left(\frac{v}{u+v}\right)^{n-k}$

Time of first / last arrivals of 3 independent Poisses

first arrival : $\min(X, Y, Z) \sim \text{Exp}(3\lambda)$

last arrival : $\max(X, Y, Z)$

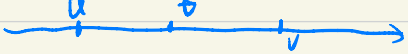
$$E[\max(X, Y, Z)] = \frac{1}{3\lambda} + \frac{1}{2\lambda} + \frac{1}{\lambda}$$

Split

$\text{Poisson}(\lambda)$ $<$ $\text{Poisson}(\lambda q)$, $\text{Poisson}(\lambda(1-q))$ independent

Random Incidence

more likely to show up at a longer length of interval

$\text{Poisson}(\lambda)$ 

- u: last arrival
- v: next arrival
- t^*: arrival time

$$v - u = \underbrace{v - t^*}_{\text{Exp}(\lambda)} + \underbrace{t^* - u}_{\text{Exp}(\lambda)}$$

$$E(v-u) = E(v-t^*) + E(t^*-u) = 2/\lambda$$

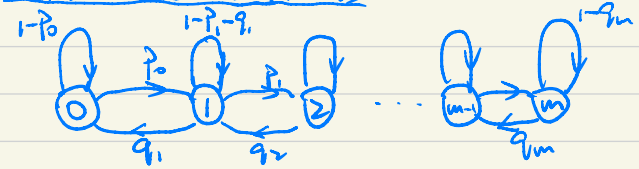
sampling method matters

Markov Process

State(t+1) = f(State(t), noise)

- X_n : state after n transitions
  - belong to a finite set
  - initial state X_0 either given or random
  - homogeneous transition probability
    $P_{ij} = P(X_{n+1}=j \mid X_n=i) = P(X_1=j \mid X_0=i)$
- **Key assumption**
  Given current state, the past does not matter
  $P_{ij} = P(X_{n+1}=j \mid X_n=i)$
  $= P(X_{n+1}=j \mid X_n=i, X_{n-1}, \cdots X_0)$
- **N-Step transition**
  $r_{ij}(n) = P(X_n=j \mid X_0=i)$
  $= P(X_{n+s}=j \mid X_s=i)$
  recursion: $r_{ij}(n) = \sum_k r_{ik}(n-1) P_{kj}$
  $= \sum_k P_{ik} r_{kj}(n-1)$
- **Recurrent / transient**
  - state i is recurrent if starting from i, wherever you go, there is a way back to i
  - not recurrent → transient
  - recurrent class: a collection of recurrent states communicating only between each other
- **Steady state probability**
  - $\lim_{n\to\infty} r_{ij}(n) = \pi_j$ converge if
    1. recurrent states are in one class
    2. single recurrent class is not periodic
  - calculate by solving linear system
    $\pi_j = \sum_k \pi_k P_{kj}$ → balance equation
    $\sum \pi_j = 1$

Birth – Death Process



- $\lambda_i p_i = \lambda_{i+1} q_{i+1} \Rightarrow \lambda_0 + \lambda_0 \frac{p_0}{q_1} + \lambda_0 \frac{p_0 p_1}{q_1 q_2} + \cdots = 1$
- $\sum \lambda_i = 1$

If p_i, q_i fixed

- $\lambda_i = \lambda_0 \left(\frac{p}{q}\right)^i$
- $p = q$ 时 $\Rightarrow \lambda_i = \lambda_0$
- $p < q$ 时 $\Rightarrow \lambda_0 = 1 - \frac{p}{q}$ $\quad E(X_n) = \frac{p/q}{1-p/q}$

Absorption

absorbing state: recurrent state/class k with $P_{kk} = 1$

- probability a_i that eventually settle if it start from i

$$a_j = \sum_{j=1}^{m} P_{ij} a_j$$

- Expected time to absorption

$$u_i = 1 + \sum_j P_{ij} u_j$$

Statistic

Introduction

Statistic: use data to gather insight and make decisions

- computational view

  Data (sequence of number)

  ↕

  Algorithm (application of statistical principle)

  sometimes will skip insights

- statistical view

  Data: comes from a random process

  Goal: learn how the process work and further } predict / find determinant

Statistical Modeling

complicated process

"="

simple process + random noise } ① real randomness ② deterministic but too complex

Modeling consists in choosing } ① plausible simple process ② noise distribution

noise is what we do not understand. It is a way of modeling lack of information

Probability Tool

- Law of large number (LLN)

  When size n of the experiment becomes larger, sample mean is a good (consistent) estimator of expectation

- Central limit Theorem (CLT)

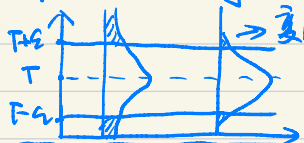
  How good the estimator is

  $\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)$

Convergence

Random Variable Convergence

- **Almost Sure** $T_n \xrightarrow{a.s.} T$ iff
  $P[\{\omega : T_n(\omega) \to T(\omega)\}] = 1$
  意味着 # of $|T_n - T| > \varepsilon$ 是 finite 的, 除一些 prob 0 的 case 都会收敛
  converge, 表示 limit of number sequence
- **In Probability** $T_n \xrightarrow{i.p.} T$ iff
  $P[|T_n - T| \geq \varepsilon] \to 0$



- **In distribution** $T_n \xrightarrow{d} T$ iff
  $E[f(T_n)] \to E[f(T_n)]$ for all continuous and bounded f
  $\lim F_{X_n}(x) \to F_X(x)$ pointwise for all x

Convergence operation

$T_n \xrightarrow{a.s./i.p.} T \quad U_n \xrightarrow{a.s./i.p.} U$

- $T_n + U_n \xrightarrow{a.s./i.p.} T + U$
- $T_n U_n \xrightarrow{a.s./i.p.} TU$
- if $U \neq 0$, $T_n / U_n \xrightarrow{a.s./i.p.} T/U$
- In general does not apply to (d)

Slutsky: partial operation hold for (d)

$T_n \xrightarrow{d} T \quad U_n \xrightarrow{P} u$ (deterministic)

- $T_n + U_n = T + u$
- $T_n U_n = uT$
- If $u \neq 0$, $T_n / U_n = T/u$

Continuous Mapping

For continuous function f: $T_n \xrightarrow{a.s./i.p./d} T \Rightarrow f(T_n) \xrightarrow{a.s./i.p./d} f(T)$

Delta Method

Let Z_n be asymptotical normal $\sqrt{n}(Z_n - \theta) \xrightarrow[n\to\infty]{(d)} N(0, \sigma^2)$

Let g be continuous differentiable at the point θ

$g(Z_n)$ is also asymptotic normal:

$$\sqrt{n}(g(Z_n) - g(\theta)) \xrightarrow[n\to\infty]{(d)} N(0, g'(\theta)^2\sigma^2)$$

Taylor Expansion approximation:

$$g(Z_n) - g(\theta) = (Z_n - \theta) g'(\theta) + \frac{1}{2}(Z_n - \theta)^2 g''(w) \quad w \in [Z_n, \theta]$$
$$\overset{n}{\approx} (Z_n - \theta) g'(\theta)$$

Multivariate:

$\sqrt{n}(T_n - T) \to N(0, \Sigma)$ then

$\sqrt{n}(f(T_n) - f(T)) \to N(0, \nabla g^T(\theta) \Sigma \nabla g(\theta))$

Statistical Inference

Goal: Use reasonable assumptions to find tractable models

1. estimation ⟶ single number
2. confidence interval → error bar around the number
3. Hypothesis test ⟶ answer to yes/no question

Statistical Model

Let the observed outcome of a statistical experiment be a sample $X_1, \dots X_n$ of n i.i.d random variables in some measurable space $\mathcal{E}$ and common distribution P

$$(\mathcal{E}, (P_\theta)_{\theta \in \Theta})$$

- $\mathcal{E}$: sample space
- P : probability measure on $\mathcal{E}$
- Θ : parameter set
  - parametrical : can be defined with finite unknowns
  - Nonparametrical : otherwise

Identifiability

Parameter θ is identified iff

$$\theta \neq \theta' \Rightarrow P_\theta \neq P_{\theta'}$$
$$P_\theta = P_{\theta'} \Rightarrow \theta = \theta'$$

Parameter Estimation

Asymptotic normal : $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n \to \infty]{(d)} \mathcal{N}(0, \sigma^2)$

σ^2 is the asymptotic variance

Quadratic risk : $R(\hat{\theta}_n) = E[(\hat{\theta}_n - \theta)^2]$

$$R(\hat{\theta}_n) = \underbrace{E[(\hat{\theta}_n - E(\hat{\theta}_n))^2]}_{\text{var}(\hat{\theta}_n)} + \underbrace{E[E(\hat{\theta}_n) - \theta]^2}_{\text{bias}^2}$$

Confidence Interval

- Confidence Interval of level $1-\alpha$ for θ
  Any random interval I (depend on $X_1, \cdots X_n$) whose boundary do not depend on θ such that
  $$P_\theta[I \ni \theta] \geq 1-\alpha, \quad \forall \theta \in \Theta$$
- Asymptotic variance
  $$\lim_{n\to\infty} P_\theta[I \ni \theta] \geq 1-\alpha, \quad \forall \theta \in \Theta$$

Get Confidence Interval

① Find a pivotal statistic (a statistic without unknown)
   - → Non-asymptotic CI
   - CLT / Delta Method → Asymptotic CI

② Solve for CI
   - Two side symmetrical
     $$P(\theta \in [\hat{\theta} - s, \hat{\theta} + s]) = 1-\alpha$$
     $$\Rightarrow I = \hat{\theta} \pm \sigma q_{\frac{\alpha}{2}}/\sqrt{n}$$
   - One side
     $$P(\theta \leq \hat{\theta} + s)$$
     $$\Rightarrow I = (-\infty, \hat{\theta} + \sigma q_{\alpha}/\sqrt{n}$$

Interpretation:
- probability of containing the true parameter if you do many experiments.
- Once plug in observations. It either contains the true parameter (100%) or not (0%)

Estimation

Goal: Given $X_1, \cdots X_n$, find an estimator $\hat{\theta} = \hat{\theta}(X_1, \cdots X_n)$ such that P_θ is close to P_{θ^*}

$\Updownarrow$

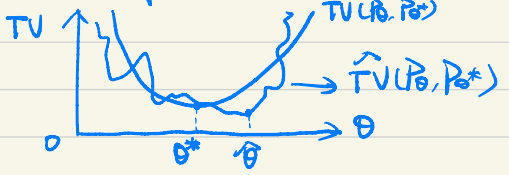
$|P_\theta(A) - P_{\theta^*}(A)|$ is small for all $A \subset E$

Total Variation Distance

- $TV(P_\theta, P_{\theta'}) = \max_{A \subset E} |P_\theta(A) - P_{\theta'}(A)|$

$$= \begin{cases} \frac{1}{2} \sum_{x \in E} |P_\theta(x) - P_{\theta'}(x)| \\ \frac{1}{2} \int_E |f_\theta(x) - f_{\theta'}(x)| \, dx \end{cases}$$

- estimation strategy: Build an estimator $\widehat{TV}(P_\theta, P_{\theta^*})$ for all θ. Then find $\hat{\theta}$ that minimize the function $\theta \to \widehat{TV}(P_\theta, P_{\theta^*})$



- How to estimate $TV(P_\theta, P_{\theta^*})$. Nice if TV is expectation so that we can use LLN. TV is not

KL ↙

Kullback-Leibler (KL) divergence

$$KL(P_\theta, P_{\theta'}) = \begin{cases} \sum_{x \in E} P_\theta(x) \log\left(\frac{P_\theta(x)}{P_{\theta'}(x)}\right) & E \text{ discrete} \\ \int_E f_\theta(x) \log\left(\frac{f_\theta(x)}{f_{\theta'}(x)}\right) dx & E \text{ continuous} \end{cases}$$

$$= E_\theta\left[\log\left(P_\theta(x) / P_{\theta'}(x)\right)\right]$$

Maximum Likelihood Estimation

KL Divergence to ML estimator

$$KL(P_{\theta^*}, P_\theta) = E_{\theta^*}\left[\log\left(\frac{P_{\theta^*}(x)}{P_\theta(x)}\right)\right] = \underbrace{E_{\theta^*}[\log P_{\theta^*}(x)]}_{\text{constant}} - E_{\theta^*}[\log P_\theta(x)]$$

$$\downarrow$$

$$\widehat{KL}(P_{\theta^*}, P_\theta) = \text{"constant"} - \frac{1}{n}\sum_{i=1}^{n} \log P_\theta(X_i)$$

$$\downarrow$$

$$\min_{\theta\in\Theta} \widehat{KL}(P_{\theta^*}, P_\theta) \iff \max_{\theta\in\Theta} \underbrace{\prod_{i=1}^{n} P_\theta(X_i)}$$

Likelihood

$$L_n = \begin{cases} (X_1, \cdots X_n, \theta) \to P_\theta[X_1 = x_1, \cdots, X_n = x_n] & \text{for discrete} \\ (X_1, \cdots X_n, \theta) \to \prod_{i=1}^{n} f_\theta(x_i) & \text{for continuous} \end{cases}$$

ML Estimator

$$\hat{\theta}_n^{ML} = \underset{\theta\in\Theta}{\operatorname{argmax}}\ L(X_1, \cdots X_n; \theta)$$

Property under mild condition (continuous, identifiable, ... etc)

① Consistency

$$\hat{\theta}_n^{ML} \xrightarrow[n\to\infty]{i.P} \theta^* \qquad \text{(X axis converge)}$$

$$\uparrow\uparrow \qquad\qquad \uparrow\uparrow \text{ identifiable, etc...}$$

$$\frac{1}{n} L(X_1, \cdots X_n; \theta) \xrightarrow[n\to\infty]{i.P} KL(P_{\theta^*}, P_\theta) \qquad \text{(y axis converge)}$$

② Asymptotic normality

$$\sqrt{n}(\hat{\theta}_n^{ML} - \theta^*) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, I(\theta^*)^{-1})$$

- Fisher Information

Define $l(\theta) = \log L_1(X, \theta)$ for one observation

$$I(\theta) = \underbrace{E[\nabla l(\theta)\, \nabla l(\theta)^T] - E[\nabla l(\theta)]\, E[\nabla l(\theta)]^T}_{\text{variance of gradient}} = -E[H l(\theta)]$$

$$\text{cov}(\nabla l(\theta)), \quad l(\theta) = \ln f_\theta(x)$$

EM Algorithm

Iterative process to find ML estimator in presence of latent variable
- E: compute lower bound of log likelihood for current parameter
- M: maximize the lower bound to set new parameter

Goal: Maximize likelihood

$P(X|\theta) = \sum_Z P(X,Z|\theta)$ — X: observed variable, Z: latent variable

$\ln P(X|\theta) = L(q,\theta) + KL(q,P) \Leftarrow \ln P(X,Z|\theta) = \ln P(Z|X,\theta) + \ln P(X|\theta)$

$= \sum_Z q(Z) \ln\left(\frac{P(X,Z|\theta)}{q(Z)}\right) + \left[-\sum_Z q(Z) \ln\left(\frac{P(Z|X,\theta)}{q(Z)}\right)\right]$ for any $q(Z)$

$\Downarrow KL \geq 0$

$L(q,\theta) \leq \ln P(X|\theta)$ is a lower bound

Assumption

Optimization of $P(X|\theta)$ is difficult but optimization of $P(X,Z|\theta)$ is easier.

① E-Step: maximizing $L(q,\theta^{old})$ holding θ^{old} constant

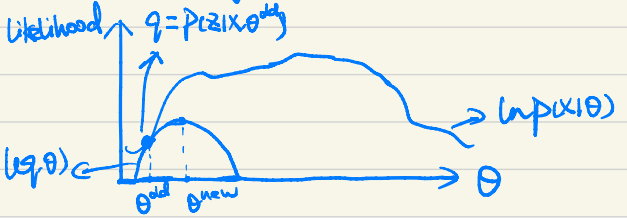
Since $\ln P(X|\theta)$ not dependent on $q(Z)$

$\max L(q,\theta) \Leftrightarrow \min KL(q,P)$

$q = P(Z|X,\theta^{old})$

② M-Step: maximizing $L(q,\theta)$ holding q constant

Maximizing with complete data $\Rightarrow \theta^{new}$



$L(q,\theta)$ 是 $\ln P(X|\theta)$ 在 θ^{old} 附近的一个 convex 的下界

$\ln P(X|\theta) = \ln \sum_Z P(X,Z|\theta) = \ln \sum_Z q(Z) \frac{P(X,Z|\theta)}{q(Z)}$

$\geq \sum_Z q(Z) \ln \frac{P(X,Z|\theta)}{q(Z)}$ (Jensen Inequality)

EM for Gaussian Mixture

1. Initialize μ_k, Σ_k and mix coefficient π_k

2. E-step

evaluate $P(z|x, \theta^{old}) = P(j|i)$ point i is generated by cluster j

$$P(j|i) = \frac{\pi_j \mathcal{N}(x_i|\mu_j, \Sigma_j)}{\sum_{m=1}^{k} \pi_m \mathcal{N}(x_i|\mu_m, \Sigma_m)}$$

M-step

$$\hat{\mu}_j = \frac{\sum_{i}^{n} P(j|i)\, x_i}{\sum_{i}^{n} P(j|i)}$$

$$\hat{\pi}_j = \frac{1}{n} \sum_{i}^{n} P(j|i)$$

$$\Sigma_j = \frac{\sum_{i}^{n} P(j|i)(x_i - \hat{\mu}_j)(x_i - \hat{\mu}_j)^T}{\sum_{i}^{n} P(j|i)}$$

Evaluate for convergence

$$\ln p(x|\mu, \Sigma, \pi) = \sum_{i}^{n} \ln\left(\sum_{k}^{k} \pi_k \mathcal{N}(x_i|\mu_j, \Sigma_j) \right)$$

M - Estimation

M estimator

1. Loss function ρ $\mathcal{X} \times \mathcal{M} \to \mathbb{R}$ ($\mathcal{M}$: possible value of u)
   $Q(u) = \mathbb{E}[\rho(X, u)]$ against true distribution
2. can estimate with $\frac{1}{n}\sum \rho(x_i, u)$
3. choose ρ so $Q(u)$ is minimized at u^*

$$\hat{u} = \operatorname{argmin}_{u \in \mathcal{M}} \frac{1}{n}\sum \rho(x_i, u)$$

Asymptotic Normality

$$\begin{cases} J(u) = \mathbb{E}[H\rho] = \mathbb{E}[\frac{\partial^2 \rho}{\partial u^2}\rho(x_i, u)] \\ K(u) = \operatorname{var}[\nabla \rho] = \operatorname{var}[\frac{\partial \rho}{\partial u}(x_i, u)] \end{cases}$$

Assume:

1. u^* is the only minimizer of Q
2. $J(u)$ is invertible for all $u \in \mathcal{M}$
3. A few technical conditions

we have:

$$\begin{cases} \hat{u} \xrightarrow[n\to\infty]{i.p} u^* \\ \sqrt{n}(\hat{u} - u^*) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, J(u^*)^{-1} K(u^*) J(u^*)) \end{cases}$$

- ρ is log-likelihood if $J(u) = K(u) = I(u)$

MLE is also a m-estimator
MLE tend to be best with smallest asymptotic variance
while MLE need assumption about distribution

ρ function

- $\rho(x, u) = \|x - u\|_2^2 \Rightarrow u^* = \mathbb{E}(x)$
- $\rho(x, u) = |x - u| \Rightarrow u^* = \text{median}$
- $\rho(x, u) = C_\alpha(x - u) \Rightarrow u^*$: α-quantile

$$C_\alpha(x) = \begin{cases} -(1-\alpha)x & \text{if } x < 0 \\ \alpha x & \text{if } x \geq 0 \end{cases}$$

Method of Moment Estimation

Moments

X be random variable with P_θ, k-th moment is

$$m_k = m_k(\theta) = E_\theta(X^k)$$

Moment Generating Function (MGF)

$$M_X(t) = E[e^{tX}], \quad t \in R$$

$$m_k = M_X^{(k)}(t)\Big|_{t=0}$$

Method of Moments

sample mean

$$\hat{m}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^{(k)} \xrightarrow{i.p.} m_k$$

$$\psi: \theta \to (m_1(\theta), \cdots m_d(\theta))$$

Estimator

$$m_1(\hat{\theta}_n) = \hat{m}_1$$
$$m_2(\hat{\theta}_n) = \hat{m}_2 \quad \xrightarrow{\text{solve linear system}} \quad \hat{\theta}_n = \psi^{-1}\left(\frac{1}{n}\Sigma X_i, \frac{1}{n}\Sigma X_i^2, \cdots, \frac{1}{n}\Sigma X_i^d\right)$$
$$\vdots$$
$$m_d(\hat{\theta}_n) = \hat{m}_d$$

$$\Downarrow \text{ CLT + Delta Method}$$

$$\sqrt{n}(\hat{\theta}^{MM} - \theta) \xrightarrow[n\to\infty]{d} N\left(0, \left[\frac{\partial m^{-1}}{\partial \theta}(m(\theta))\right]^T \Sigma(\theta) \left[\frac{\partial m^{-1}}{\partial \theta}(m(\theta))\right]\right)$$

When to use: don't know the function working with

Hypothesis Testing

Statistical Formulation

IID sample: $X_1, \dots X_n$ with statistical model $(\mathcal{X}, (P_\theta)_{\theta \in \Theta})$

consider 2 disjoint hypothesis $\begin{cases} \text{Null} & H_0: \theta \in \Theta_0 \\ \text{Alternative} & H_1 \quad \theta \in \Theta_1 \end{cases}$

If we believe that true θ is either in Θ_0 or Θ_1, we can test H_0 against H_1. H_0 is the status quo. Data can only be used to reject H_0. (about scientific method)

★ "Innocent till proven guilty" ★

Test

A test is a statistic $\psi \in \{0,1\}$ that does not depend on unknowns

$\psi = \begin{cases} 0 \Leftrightarrow \text{not reject } H_0 \\ 1 \Leftrightarrow \text{reject } H_0 \end{cases}$ can always be written in "reject region"

$= \mathbb{1}(R_\psi)$, $R_\psi = \{x \in \mathcal{X}^n : \psi(x) = 1\}$

Errors

	Fail to reject H_0	Reject H_1
H_0 is true: $\theta \in \Theta_0$	✓	Type I
H_1 is true: $\theta \in \Theta_1$	Type II	✓

beta function $\beta(\theta) = P_\theta(\psi = 1)$

$\begin{cases} ① & H_0 \text{ is true}: \quad \beta(\theta) = P(\psi \text{ make type I error}) \\ ② & H_1 \text{ is true}: \quad 1 - \beta(\theta) = P(\psi \text{ make type II error}) \end{cases}$

The Neyman-Pearson paradigm

- Make sure $P(\text{Type I error}) \le \alpha$ level of the test
- Minimize $P(\text{Type II error})$ subject to the constraint

A test ψ has level α if $\max_{\theta \in \Theta_0} P_\theta(\psi = 1) \le \alpha$

Power of a test ψ

$\pi_\psi = \inf_{\theta \in \Theta_1} (1 - \beta_\psi(\theta))$

Build tests from CI

① $P_\theta(\theta \in [A, B]) \geq 1-\alpha$, confidence interval at level $1-\alpha$

② $H_0: \theta = \theta_0$

$H_1: \theta \neq \theta_0$

③ Natural candidate

$\Psi = \mathbb{1}\{\theta_0 \notin [A, B]\}$

④ $P_{\theta_0}(\Psi = 1) = P_{\theta_0}[\theta_0 \notin I] = 1 - P_{\theta_0}[\theta_0 \in I] \leq 1 - (1-\alpha) = \alpha$

Ψ is a test with level α.

P value

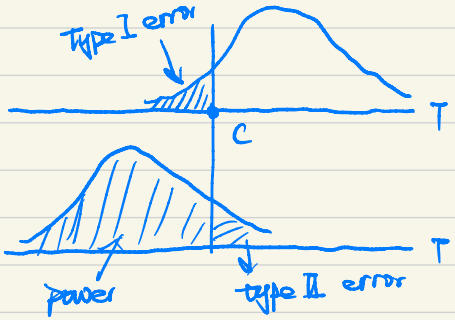
The (asymptotic) p-value of a test is the smallest (asymptotic) level α at which Ψ reject H_0

p-value $\leq \alpha \iff H_0$ is rejected by Ψ at the level α

p-value $> \alpha \iff H_0$ is not rejected by Ψ at the level α.

small p-value $\iff$ strong evidence against H_0

Tradeoff between type I and type II



→ distribution under H_0

$\Psi = \mathbb{1}\{T \geq c\}$

→ distribution under H_1

Also why usually one-sided test is more powerful than 2-sided test: the distribution under H_1 is usually tail shape not 2-peak

Wald Test

For the estimator $\hat{\theta}$

we have $\frac{\hat{\theta}-\theta}{\sqrt{\widehat{\text{var}}(\hat{\theta})}} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)$ Define $W := \frac{\hat{\theta}-\theta_0}{\sqrt{\widehat{\text{var}}(\hat{\theta})}}$

	$H_0: \theta=\theta_0$ $H_1: \theta\neq\theta_0$	$H_0: \theta\leq\theta_0$ $H_1: \theta>\theta_0$	$H_0: \theta\geq\theta_0$ $H_1: \theta<\theta_0$
Wald Test	$\mathbb{1}\{\lvert W\rvert > q_{\frac{\alpha}{2}}\}$	$\mathbb{1}\{W > q_\alpha\}$	$\mathbb{1}\{W < -q_\alpha\}$
P-value	$P(\lvert W\rvert > \lvert w^{obs}\rvert)$	$P(W > w^{obs})$	$P(W < w^{obs})$
Asym P	$P(\lvert Z\rvert > \lvert w^{obs}\rvert)$	$P(Z > w^{obs})$	$P(Z < w^{obs})$

- $\widehat{\text{var}}(\hat{\theta})$ can be any consistent variance estimator (Slusky)
- $Z \sim \mathcal{N}(0,1)$
- For mle estimator $\hat{\theta}^{mle}$, $\widehat{\text{var}}(\hat{\theta}) = \frac{I(\theta)^{-1}}{n}$

  $W := \sqrt{n I(\theta^{mle})}(\hat{\theta}^{mle} - \theta_0)$

Two Sample Wald Test

$E(X) = \mu_1$, $\text{var}(X) = \sigma_1^2$ $\Rightarrow$ $H_0: \mu_1 = \mu_2$, $\theta = \mu_1 - \mu_2 = 0$

$E(Y) = \mu_2$, $\text{var}(Y) = \sigma_2^2$ $H_1: \mu_1 \neq \mu_2$, $\theta = \mu_1 - \mu_2 \neq 0$

$\hat{\theta} = \bar{X}_n - \bar{Y}_m \Rightarrow \frac{\hat{\theta}-\theta}{\sqrt{\text{var}(\hat{\theta})}} \xrightarrow{(d)} \mathcal{N}(0,1)$

$\Downarrow$ $\text{var}(\hat{\theta}) = \text{var}(\bar{X}_n) + \text{var}(\bar{Y}_m) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}$

Slusky: replace σ_1^2, σ_2^2 with $\hat{\sigma}_1^2, \hat{\sigma}_2^2$

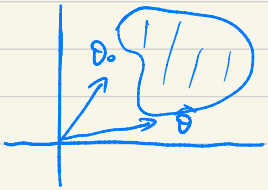
$$\frac{\hat{\theta}-\theta}{\sqrt{\frac{\hat{\sigma}_1^2}{n} + \frac{\hat{\sigma}_2^2}{m}}} \xrightarrow[m\to\infty]{n\to\infty,\,(d)} \mathcal{N}(0,1)$$

Multi-variate Wald Test

$\theta \in \mathbb{R}^d$ $H_0: \theta = \theta_0$ Intuition: test $\lVert\hat{\theta} - \theta_0\rVert$ is small enough

$H_1: \theta \neq \theta_0$

$W_n := n(\hat{\theta}_n^{mle} - \theta_0)\, I(\hat{\theta}^{mle})(\hat{\theta}_n^{mle} - \theta_0) \xrightarrow{(d)} \chi_d^2$



if $\|\theta_0 - \hat{\theta}\|$ small enough

Implicit hypotheses

Let $g: \mathbb{R}^d \to \mathbb{R}^k$ be continuous differentiable

Suppose $\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \Sigma(\theta^*))$, $\Sigma(\theta^*) \in \mathbb{R}^{d\times d}$

$\Downarrow$ Multi-variate Delta

$\sqrt{n}(g(\hat{\theta}_n) - g(\theta^*)) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_k(0, \Gamma(\theta))$

$\Gamma(\theta) = \nabla g^T(\theta)\, \Sigma(\theta)\, \nabla g(\theta) \in \mathbb{R}^{k\times k}$

Assume $\Sigma(\theta)$ invertible and $\nabla g(\theta)$ has rank k

$\sqrt{n}\,\Gamma^{-1}(\theta)(g(\hat{\theta}_n) - g(\theta^*)) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_k(0, I_k)$

$\Downarrow$ Slutsky

$\Gamma(\theta)$ is continuous at θ^* replace with $\Gamma(\hat{\theta})$

$\sqrt{n}\,\Gamma^{-1}(\hat{\theta})(g(\hat{\theta}_n) - g(\theta^*)) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_k(0, I_k)$

$\Downarrow$ $H_0: g(\theta) = 0$ $\quad$ $H_1: g(\theta) \neq 0$ $\quad$ under-null $g(\theta^*) = 0$

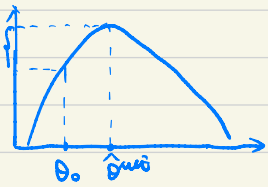
$W_n = n\, g^T(\hat{\theta}_n)\, \Gamma^{-1}(\hat{\theta}_n)\, g(\hat{\theta}_n) \xrightarrow[n\to\infty]{(d)} \chi^2_k$

Usually used with MLE estimator

Likelihood Test

Basic Form

If large enough to reject θ_0.



Consider iid sample $X_1, \dots X_n$ with statistical model $(E, (P_\theta)_{\theta\in\Theta})$, $\Theta \subseteq \mathbb{R}^d$

$H_0: (\theta_{r+1}, \dots, \theta_d) = (\theta_{r+1}^{(0)}, \dots \theta_d^{(0)})$

$H_1: (\theta_{r+1}, \dots \theta_d) \neq (\theta_{r+1}^{(0)}, \dots, \theta_d^{(0)})$

$\Downarrow$

MLE: $\hat{\theta}_n = \underset{\theta\in\Theta}{\operatorname{argmax}} \ln(\theta)$

constrained MLE: $\hat{\theta}_n^c = \underset{\theta\in\Theta_0}{\operatorname{argmax}} \ln(\theta)$

$\Downarrow$

$T_n = 2(\ln(\hat{\theta}_n) - \ln(\hat{\theta}_n^c))$

$\Downarrow$ Wilk's Theorem

$T_n \xrightarrow[n\to\infty]{(d)} \chi^2_{d-r}$ $\quad d-r = \dim(\Theta) - \dim(\Theta_0)$ ($\dim(\Theta_0 \cup \Theta_1)$)

$\psi = \mathbb{1}\{T_n > q_\alpha\}$ $\quad q_\alpha$: $(1-\alpha)$ quantile of χ^2_{d-r}

Neyman-Pearson Lemma

Among all tests for the hypothesis $H_0: \theta=\theta_0$, $H_1: \theta=\theta_1$ at significant level α. Likelihood ratio test is the most powerful: highest probability to reject the null.

T-test

Cochran's Theorem

If $X_1, \cdots X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, $S_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

1. $\bar{X}_n$ is independent with S_n
2. $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$

$\Downarrow$ (lose 1 df to replace μ with $\bar{X}_n$)

$$\frac{\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}}{\sqrt{S_n^2/\sigma^2}} \sim \chi(0,1) \sim t_{n-1}$$

T test

- $X_1 \cdots X_n \sim N(\mu, \sigma^2)$
- $\sqrt{n}\frac{\bar{X}_n - \mu}{S_n} \sim t_{n-1}$ for all n

$T := \sqrt{n}\frac{\bar{X}_n - \mu_0}{S_n}$	$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	$H_0: \mu \leq \mu_0$ $H_1: \mu > \mu_0$	$H_0: \mu \geq \mu_0$ $H_1: \mu < \mu_0$
T-test	$\mathbb{1}\{\lvert T\rvert > q^{t_{n-1}}_{\frac{\alpha}{2}}\}$	$\mathbb{1}\{T > q^{t_{n-1}}_{\alpha}\}$	$\mathbb{1}\{T < -q^{t_{n-1}}_{\alpha}\}$
P-value	$P(\lvert T\rvert > \lvert T^{obs}\rvert)$	$P(T > T^{obs})$	$P(T < T^{obs})$

- $q^t > q^{normal}$: heavy tail
- Wald test p-value $\overset{d}{\to}$ Unif, more likely to reject H_0

Two Sample Test

$X_1 \cdots X_n \overset{iid}{\sim} N(\mu_1, \sigma_1^2)$

$Y_1 \cdots Y_m \overset{iid}{\sim} N(\mu_2, \sigma_2^2)$

$H_0: \mu_1 - \mu_2 \leq 0$

$H_1: \mu_1 - \mu_2 > 0$

$$T = \frac{\bar{X}_n - \bar{Y}_m - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n + S_2^2/m}} \sim t_N$$

$$N = \frac{(S_1^2/n + S_2^2/m)^2}{S_1^2/n^2(n-1) + S_2^2/m^2(m-1)} \geq \min(m, n)$$

Welch-Satterthwaite (WS) formula

Multiple Hypothesis Test

Issue: try to make aggregate decision but only control error individually

	Not reject H_0^i (non-sig)	Reject H_0^i (significant)	
H_0^i true	A (# of true non-significant)	V (# of type I error)	m_0
H_1^i true	C (# of type II error)	D (# of true significant)	m_1
	N_0 (# of non-significant)	N_1 (# of significant)	m

FWER: family wise error rate

- Probability of at least one false discovery (type I)
- $FWER = P(V \geq 1) = P(\sum_{i=1}^{m_0} V_i \geq 1)$
  $= 1 - P(V=0) = 1 - (1-\alpha)^m \xrightarrow{m \to \infty} 1$

  Usually too strict for large m

FDR: False Discovery Rate

- Expected fraction of false significant results among all significant results
- $FDR = E[V/N_1]$
- FDR has higher power compared to FWER. FWER is stricter

  $E[V/N_1] \leq E[\mathbb{1}(V \geq 1)] = P(V \geq 1)$

  $V/N_1 \leq \mathbb{1}(V \geq 1)$ when $V \geq 1$

  $V/N_1 = \mathbb{1}(V \geq 1) = 0$ when $V = 0$

  $FDR \leq FWER$
- Power of a series of tests with FDR are more powerful

Bonferroni Correction → FWER

Use α/m instead of α to reject

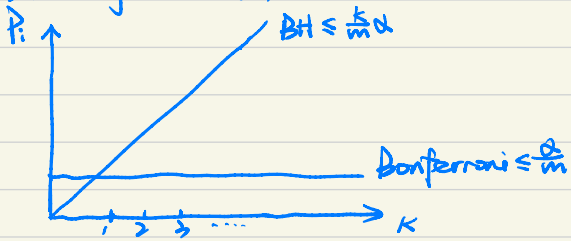
- $P_i \le \alpha/m \iff FWER \le \alpha$
- $FWER = P(W \ge 1) = P(U_{i \in I_0} \{\Psi_i = 1\})$
  $\le \sum P(\Psi_i = 1) = \sum \alpha/m = m_0 \alpha/m \le \alpha$

其中 $I_0 = \{i = 1, \cdots, m \mid H_0^i \text{ true}\}$

Benjamini–Hochberg → FDR

For a series of m independent test

1. sort the p-value : $P^{(1)} \le P^{(2)} \le \cdots \le P^{(m)}$
2. Find the maximum k that
   $P^{(k)} \le \frac{k}{m}\alpha$
3. Reject $H_0^{(1)}, H_0^{(2)}, \cdots H_0^{(k)}$



Goodness of Fit - Discrete

$X_1, \cdots X_n$ n discrete P.

- Sample space $E = \{a_1, \cdots, a_k\}$
- probability measure $P = [P_1, \cdots, P_k]^T$ $\sum P_i = 1$, $P_i \geq 0$

GoF - Chi square

$H_0: P = P^0 \Rightarrow T_n := n \sum_{j=1}^{k} \frac{(\hat{P}_j - P_j^0)^2}{P_j^0} \xrightarrow[n\to\infty]{(d)} \chi^2_{k-1}$ $[\psi = \mathbb{1}(T_n > q_\alpha)]$

$H_1: P \neq P^0$

$\Downarrow$ MLE $\qquad \Uparrow$

$\hat{P}_j = \frac{N_j}{n}$, $\sqrt{n}(\hat{P} - P^0)$ is asymptotical normal in dim $k-1$

Degenerate since $\sqrt{n} \sum (\hat{P}_i - P_i^0) = 0$

$\sqrt{n}\, I^{\frac{1}{2}}(P^0)(\hat{P} - P^0) \xrightarrow{n\to\infty} \mathcal{N}_{k-1}(0, I_{k-1})$ no variance along 1 direction

GoF test on Family of Distribution

$H_0: P \in \{P_\theta\}_{\theta \in \mathbb{R}^d}$

$H_1: P \notin \{P_\theta\}_{\theta \in \mathbb{R}^d}$

- $\{a_1, \cdots, a_k\}$, $f_\theta(a_i) = \mathbb{P}(X = a_i)$
- Observe $X_1, \cdots X_n \Rightarrow$ MLE $\hat{P} = [\frac{N_0}{n}, \cdots \frac{N_k}{n}]^T$
- $T_n := n \sum_{i=0}^{k} \frac{(\frac{N_i}{n} - f_\theta(a_i))^2}{f_\theta(a_i)} \xrightarrow{(d)} \chi^2_{(k+1)-1-d}$

$\vee$ $k+1$: the support size of P_θ (for all θ)

Goodness of Fit - Continuous

Convergence of Function

1. Convergence pieciewise : for each x $\lim_{n\to\infty} g_n(x) = g(x)$
2. Convergence uniformly : for every $\varepsilon > 0$, exists n_ε such that $\sup_x |g_n(x) - g(x)| < \varepsilon$ for all $n \geq n_\varepsilon$

- Example of pieciewise but not uniformly
  $g_n(x) = \frac{1}{x^n}$ $\quad \sup_{x<1} g_n(x) = \sup_{x<1} \frac{1}{x^n} = 1$

CDF

CDF : $F(t) = P(X \leq t) = E[\mathbb{1}(X \leq t)]$

empirical CDF : $F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \leq t)$

1. Convergence pieciewise from LLT
   $F_n(t) \xrightarrow[n\to\infty]{a.s.} F(t)$ for all $t \in \mathbb{R}$
2. Convergence uniformly from GC (Glivenko-Cantelli) Theorem
   $\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \xrightarrow[n\to\infty]{a.s.} 0$

- Example of 1 but not 2
  $F_n(t) = F(t) + \frac{t}{n}$ $\quad \forall t \; F_n(t) = F(t)$
  $\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| = \sup_{t\in\mathbb{R}} |\frac{t}{n}| = \infty$

$\mathbb{1}\{X_i \leq t\} \sim Ber(F(t))$

1. CLT : $\sqrt{n}(F_n(t) - F(t)) \xrightarrow[n\to\infty]{(d)} N(0, F(t)(1-F(t)))$
2. Donsker's Theorem
   $\sqrt{n}\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \xrightarrow[n\to\infty]{(d)} \sup_{0\leq t\leq 1} |B(t)|$ Brownian bridge

worst possible case of all t — A pivotal distribution that does not depend on $F(t)$

Goodness of continuous distribution

$X_1, \cdots X_n \sim$ CDF F

$H_0 : F = F^*$

$H_1 : F \neq F^*$

Kolmogorov-Smirnov test

$$T_n = \sup_{t\in\mathbb{R}} \sqrt{n}\,|F_n(t) - F^0(t)|$$
$$= \sqrt{n} \sup_{t\in\mathbb{R}} \left|\frac{1}{n}\sum \mathbb{1}(X_i \le t) - F^0(t)\right|$$
$$\Downarrow \text{ Let } U_i = F^0(t), \quad \forall i \sim \text{Unif}[0,1]$$
$$= \sqrt{n} \sup_{t\in[0,1]} \left|\left(\frac{1}{n}\sum \mathbb{1}(U_i \le U_t) - U_t\right)\right|$$

Calculation $\Downarrow$ observation $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$

$$T_n = \sqrt{n} \max_{i=1,\cdots,n} \left\{ \max\left(\left|\frac{i-1}{n} - F^0(X_{(i)})\right|, \left|\frac{i}{n} - F^0(X_{(i)})\right|\right)\right\}$$

No analytical solution for quantile of Brownian
→ Simulation M T_n : $T_n^1, \cdots T_n^M$. Use sample quantile as estimation

$\Psi = \mathbb{1}\{T_n > \hat{Q}_{1-\alpha}^{(n,M)}\}$ (or table)

p-value $\approx \dfrac{\#\{j=1,\cdots M : T_n^j > T_n\}}{M}$

test based on function distance measure:

- sup-norm : $d(\hat{F}_n, F) = \sup_{t\in\mathbb{R}} |\hat{F}_n(t) - F(t)|$
- L2 distance : $d^2(\hat{F}_n, F) = \int_{\mathbb{R}} (\hat{F}_n(t) - F(t))^2\, dt$
- scaled L2 : $\int_{\mathbb{R}} (\hat{F}_n(t) - F(t))^2 / (F(t)(1-F(t)))\, dt$

Kolmogorov-Lilliefors test

Check if from Gaussian distribution

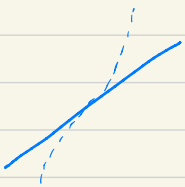
$\sup_{t\in\mathbb{R}} |\hat{F}_n(t) - \Phi_{\hat{\mu},\hat{\sigma}}(t)|$ → k-L table

QQ plot

A visual way for GoF test

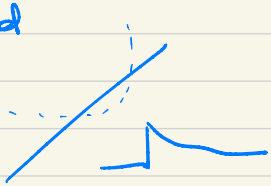
1. $X_{(1)} \le \cdots \le X_{(n)}$
2. $\left(F^{-1}\left(\frac{1}{n+1}\right), X_{(1)}\right), \left(F^{-1}\left(\frac{2}{n+1}\right), X_{(2)}\right) \cdots \left(F^{-1}\left(\frac{i}{n+1}\right), X_{(i)}\right), \cdots \left(F^{-1}\left(\frac{n-1}{n+1}\right), X_{(n-1)}\right)$

① heavy tail
sample ↑

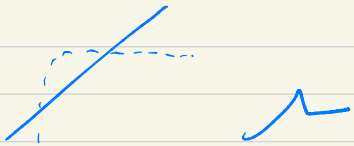


→ theoretical

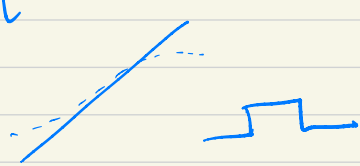
② right skewed



③ left skewed



④ light tail



Bayesian Statistic

Most effective when experts have a lot of background knowledge
→ can bring a lot of prior information

Bayes Formula

Prior $\pi(\theta)$
- ✓ a function by which the likelihood function is weighted by in order to produce posterior
- ✓ is to be specified by the researcher in order to take into account previous knowledge about possible values of the parameter

$$\pi(\theta|x_1,\cdots x_n) = \frac{\pi(\theta) L_n(x_1,\cdots x_n|\theta)}{\int_\Theta L_n(x_1,\cdots x_n|t)\pi(t)dt} \quad \forall \theta\in\Theta$$

$$\propto \pi(\theta) L_n(x_1,\cdots x_n|\theta) \quad \forall \theta\in\Theta$$

$\propto$ proportional, not depend on θ

Priors

1. Whether we can specify the parameters to approximate belief
2. Whether it is realistic
3. Whether computationally tractable
   i.e. conjugate prior of the likelihood distribution
   ⇔ prior and posterior distribution from same family

We can still use Bayesian without prior information

Noninformative prior

$\pi(\theta) \propto 1$
- If Θ is bounded ⇒ uniform prior on Θ
- If Θ is unbounded ⇒ improper prior ⇔ $\int \pi(\theta)d\theta = \infty$ can still use

Jeffreys Prior

Reparameterize Distribution

- $\pi(\theta)$, $\theta \in \mathbb{R}$

  ⇓ reparameterize with $\eta = \phi(\theta)$ $\quad \phi$ strict monotone

- $\tilde{\pi}(\eta)$

$$\int_{\theta_0}^{\theta_0+\Delta\theta} \pi(\theta)\,d\theta = \int_{\eta_0}^{\eta_0+\Delta\eta} \tilde{\pi}(\eta)\,d\eta$$

$$\Delta\theta \to 0 \qquad\qquad \Delta\eta \to 0$$

$$\Delta\theta\, \pi(\theta_0) = \Delta\eta\, \tilde{\pi}(\eta_0)$$

$$\tilde{\pi}(\eta_0) = \pi(\theta_0) \left|\frac{\Delta\theta}{\Delta\eta}\right|_{\theta=\theta_0}$$

⇓ $\theta = \phi^{-1}(\eta)$ $\quad \phi'(\theta) = d\eta/d\theta$

$$\boxed{\tilde{\pi}(\eta) = \frac{\pi(\theta)}{|\phi'(\theta)|} = \frac{\pi(\phi^{-1}(\eta))}{|\phi'(\phi^{-1}(\eta))|}}$$

Jeffreys Prior

A non-informative prior. The probability assigned to a volume of the probability space is the <u>same</u> under all parameterization.

$$\pi_J(\theta) \propto \sqrt{\det(I(\theta))}$$

$I(\theta)^{-\frac{1}{2}}$: standard error of estimator

- $I(\theta)$ ↑ ⇒ more uncertainty ⇒ smaller weight
- $I(\theta)$ ↓ ⇒ less uncertainty ⇒ larger weight

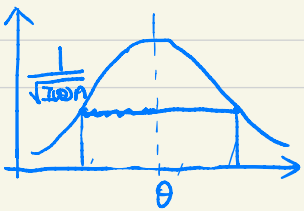
$$I(\theta) = E\left[\left(\frac{\partial}{\partial\theta} \ln L(x;\theta)\right)^2\right] = -E\left[\frac{\partial^2}{\partial\theta^2} \ln L(x;\theta)\right]$$

① marginal change of θ to log-likelihood

$$E\left[\left(\Delta \ln L(x;\theta)\right)^2\right] = I(\theta)\,(\Delta\theta)^2$$

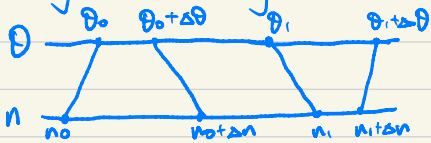
② curvature around θ

more curved ⇔ higher $I(\theta)$



Reparameterization Invariance of Jeffreys

taking sensitivity into account



$$\left.\begin{array}{l} \ln L_0(X_i|\theta) = \ln L_\eta(X_i|\eta) \\ \ln L_0(X_i|\theta+\Delta\theta) = \ln L_\eta(X_i|\eta+\Delta\eta) \end{array}\right\} \Rightarrow \Delta\theta \frac{\partial}{\partial\theta}\ln L_0(X_i|\theta) = \Delta\eta \frac{\partial}{\partial\eta}\ln L_\eta(X_i|\eta)$$

When θ_0 stretched: $\phi'(\theta) > 1$, marginal change in log-likelihood is larger than marginal change caused by $\Delta\eta$

$\pi_J(\theta)$ gives higher weight to

1. θ whose MLE estimate is more certain
2. θ where data has more information in deciding param
3. θ where marginal shifts have larger effect to X_i

- $I(\theta)$ is proxy of how much, at a particular θ, would would equivalent shifts to the parameter influence the data.

$$\left.\begin{array}{l} \pi_J(\theta) \propto \sqrt{I_0(\theta)} \\ \tilde{\pi}_J(\theta) \propto \sqrt{I(\theta)} \end{array}\right\} \xrightarrow{\eta=\phi(\theta)} \tilde{\pi}_J(\eta) \text{ is the reparameterization of } \pi_J(\theta) \text{ through } \eta=\phi(\theta)$$

I.I.D: $X_i \sim \text{Ber}(p^{10})$

① $I(p) = -E\left[\frac{\partial^2}{\partial p^2}\ln L(X_i|p)\right] = -E\left[\frac{\partial^2}{\partial p^2}\ln\left(p^{10x}(1-p^{10})^{(1-x)}\right)\right]$

$= 100\,p^8/(1-p^{10})$

② $p = p'^{10} \quad \phi(p) = p^{\frac{1}{10}} \quad \phi'(p) = \frac{1}{10}p^{-\frac{9}{10}} \quad \phi^{-1}(p) = p^{10}$

$\pi_J(p) \propto \frac{1}{p^{\frac{9}{10}}\sqrt{1-p}}$

$\tilde{\pi}_J(p) = \frac{\pi(\phi^{-1}(p))}{|\phi'(\phi^{-1}(p))|} = \sqrt{\frac{100p^8}{1-p^{10}}}$

Linear Regression

Goal: Given data point $\{(x_i, Y_i)\}$, fit a model to predict Y given a value for X

Model Set Up

(x_i, Y_i), $i = 1, \dots, n$ are iid from unknown joint distribution $\mathbb{P}$

$\mathbb{P}$ can be entirely described by:

① joint PDF $h(x,y)$ or

② marginal of X: $h(x) = \int_y h(x,y)\,dy$

conditional: $h(y|x) = h(x,y)/h(x)$

can also describe the distribution partially i.e: expectation

regression function of Y with respect to X:

$$u(x) = E[Y|X=x] = \int_{\mathbb{R}} y f(y|x)\,dy$$

linear regression function (Σ representation)

$$u(x) = E[Y|X=x] = a + bx$$

$$\Downarrow \quad \underset{a,b}{\operatorname{argmin}} E[(Y-a-bX)^2] \Rightarrow$$

$$b^* = \frac{\operatorname{cov}(X,Y)}{\operatorname{var}(X)}, \quad a^* = E(Y) - \frac{\operatorname{cov}(X,Y)}{\operatorname{var}(X)} E(X)$$

$\varepsilon = Y - (a^* + b^* X)$ is noise

with $E(\varepsilon) = 0$, $\operatorname{cov}(X, \varepsilon) = 0$

$\Downarrow$ replace expectation with average

LSE of (a^*, b^*) is $\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2$

Multi-variate Linear model

$$\begin{matrix} Y_1 = X_1^T\beta + \varepsilon_1 \\ \vdots \\ Y_n = X_n^T\beta + \varepsilon_n \end{matrix} \Leftrightarrow \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \underset{n\times p}{\begin{pmatrix} X_1^T \\ \vdots \\ X_n^T \end{pmatrix}} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$Y = X\beta + \varepsilon$$

$Y = X^T\beta + \varepsilon$ is an assumption on the model. Equivalently, we can assume that the regression function is linear: $\mu(x) = E(Y|X=x) = x^T\beta$ with the understanding that $E(\varepsilon) = 0$

(can always do linear regression even when model is misspecified)

⇓ linear regression

Find the best $\hat{\beta}$

$$\hat{\beta}_{LSE} = \underset{\beta}{\operatorname{argmin}} \|Y - X\beta\|_2^2 \xrightarrow{\text{1st order}} X^TX\hat{\beta} = X^TY$$

$$= (X^TX)^{-1}X^TY \quad \xleftarrow{\operatorname{rank}(X) = p \Leftrightarrow \text{unique } \hat{\beta}}$$

$\hat{Y} = X\hat{\beta} = X(X^TX)^{-1}X^TY$ is linear combination of column of X. The orthogonal projection of Y on the space spanned by column of X

Fisher Information

$$f(y;\beta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(Y - X^T\beta)^2\right)$$

$$\Rightarrow \nabla_\beta l(y;\beta) = -\frac{1}{2\sigma^2}(-2YX + 2X^TX\beta)$$

$$= \frac{1}{\sigma^2}(YX - X^TX\beta)$$

$$\Rightarrow H_\beta l(y;\beta) = -XX^T/\sigma^2$$

$$I(\beta) = -\sum H_\theta \ln f(x_i;\theta) = -\frac{n}{n}E_Y\left[-\frac{1}{\sigma^2}xx^T\right]$$

$$= \frac{1}{\sigma^2}X^TX$$

LSE property under deterministic X and Gaussian ε

$\hat\beta = (X^TX)^{-1}X^TY$ assume
1. X is deterministic, $\text{rank}(X) = p$
2. homoscedastic, $\varepsilon_1, \cdots \varepsilon_n$ iid
3. $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$

1. **LSE = MLE**

$Y_i \sim \mathcal{N}(X_i^T\beta^*, \sigma^2)$

$\log L(Y_1, \cdots Y_n, \beta) = -n\ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - x_i^T\beta)^2$

2. **$\hat\beta \sim \mathcal{N}_p(\beta^*, \sigma^2(X^TX)^{-1})$**

$\hat\beta = (X^TX)^{-1}X^TY = (X^TX)^{-1}X^T(X\beta^* + \varepsilon)$

$= \beta^* + (X^TX)^{-1}X^T\varepsilon \quad \mathcal{N}_p(0, (X^TX)^{-1}\sigma^2 I_n X(X^TX)^{-1})$

$= \beta^* + \mathcal{N}(0, \sigma^2(X^TX)^{-1})$

- X^TX more spread $\rightarrow$ $\text{Var}(\hat\beta)$ smaller

3. **Quadratic risk $E[\|\hat\beta - \beta^*\|_2^2] = \sigma^2 \text{tr}((X^TX)^{-1})$**

$E[\|\hat\beta - \beta^*\|_2^2]$ $\quad \Downarrow \quad \|X\|_2^2 = \text{tr}(X^TX) = \text{tr}(XX^T)$

$= E[\text{tr}((\hat\beta - \beta^*)(\hat\beta - \beta^*)^T)]$

$= \text{tr}\, E[(\hat\beta - \beta^*)(\hat\beta - \beta^*)^T] = \sigma^2 \text{tr}((X^TX)^{-1})$

4. **Prediction Error $E[\|Y - X\hat\beta\|_2^2] = \sigma^2(n-p)$**

$\|Y - X\hat\beta\|_2^2 = \sum \hat\varepsilon_i^2$

$= \|Y - PY\|_2^2 = \|(I_n - P)Y\|_2^2 = \|P^\perp Y\|_2^2$

5. **Unbiased estimator of $\hat\sigma^2 = \frac{1}{n-p}\|Y - X\hat\beta\|_2^2$**

$\hat\sigma^2 = \frac{1}{n-p}\|Y - X\hat\beta\|_2^2 = \frac{1}{n-p}\sum \hat\varepsilon_i^2$

$\Downarrow$ Cochran's Theorem

$(n-p)\frac{\hat\sigma^2}{\sigma^2} \sim \chi^2_{n-p}$

$\hat\beta \perp \hat\sigma^2$

Hypothesis Test

Test: $H_0: u^T\beta^* \le 0$ $\quad u \in \mathbb{R}^p$

$H_1: u^T\beta^* > 0$

Assume X full rank $\Rightarrow$ identifiable

1. MLE estimator

$$P_{\beta,\sigma^2}(y_1 \cdots y_n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - x_i^T\beta)^2\right)$$

$$l_{\beta,\sigma^2} = \log P_{\beta,\sigma^2}(y_1 \cdots y_n)$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\|Y - X\beta\|_2^2$$

$\nabla_\beta l = 0 \Rightarrow X^TY - X^TX\beta = 0 \Rightarrow$ $\boxed{\hat\beta = (X^TX)^{-1}X^TY}$

$\nabla_\sigma l = 0 \Rightarrow -\frac{n}{2\sigma^2} - \frac{1}{2\sigma^4}\|Y - X\beta\|_2^2 = 0 \Rightarrow$ $\boxed{\hat\sigma_{mle}^2 = \frac{1}{n}\|Y - X\beta\|_2^2}$

2. Pivotal Statistic

$$u^T\hat\beta = u^T(X^TX)^{-1}X^TY = u^T(X^TX)^{-1}X^T(X\beta^* + \varepsilon)$$

$$= u^T\beta^* + u^T(X^TX)^{-1}X^T\varepsilon$$

$$\Rightarrow \boxed{u^T\hat\beta - u^T\beta^* \sim \mathcal{N}(0,\ \sigma^2 u^T(X^TX)^{-1}u)}$$

$$\|Y - X\hat\beta\|_2^2 = \|Y - X(X^TX)^{-1}X^TY\|_2^2 = \|\varepsilon - \underbrace{X(X^TX)^{-1}X}_{H}\varepsilon\|_2^2$$

$$= \|(I_n - H)\varepsilon\|_2^2$$

$\exists$ orthogonal V and diagonal Λ such that

$$H = V\Lambda V^T = V\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}V^T = \underbrace{[V_1}_{P}\ \underbrace{V_2]}_{n-p}\begin{bmatrix} I_p & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix}$$

$$I_n - H = I_n - V\Lambda V^T = VV^T - V\Lambda V^T = V_2V_2^T$$

$$= \|V_2^T\varepsilon\|_2^2$$

$$\Rightarrow \boxed{\|Y - X\hat\beta\|_2^2/(n-p)\sigma^2) \sim \chi^2_{n-p}}$$

$$\boxed{\frac{u^T\hat\beta - u^T\beta^*}{\sigma\sqrt{u^T(X^TX)^{-1}u}} \sim \mathcal{N}(0,1)} \qquad \boxed{\frac{u^T\hat\beta - u^T\beta^*}{\hat\sigma\sqrt{u^T(X^TX)^{-1}u}} \sim t_{n-p}}$$

Multiple Hypothesis Test

$\underset{n}{Y} = \underset{n\times p}{X}\underset{p}{\beta} + \underset{n}{\varepsilon}$, β, σ^2 unknown

Test $H_0: 0 \le \beta_1 \le \beta_2$ $\Leftrightarrow$ $H_0: 0 \le \beta_1$ and $\beta_1 \le \beta_2$

$H_1:$ not $0 \le \beta_1 \le \beta_2$ $H_1: 0 > \beta_1$ or $\beta_1 > \beta_2$

$$H_0 = \begin{cases} H_0^{(1)}: 0 \le \beta_1 \Rightarrow \psi_{\beta_1}^{(1)} \\ H_0^{(2)}: \beta_1 \le \beta_2 \Rightarrow \psi_{\beta_2}^{(2)} \end{cases} \Rightarrow \psi_\alpha = \psi_{\beta_1}^{(1)} \text{ or } \psi_{\beta_2}^{(2)} \quad \max(\psi_{\beta_1}^{(1)}, \psi_{\beta_2}^{(2)})$$

$P(\psi \text{ reject}) = P(\psi^{(1)} \text{ reject or } \psi^{(2)} \text{ reject}) \ge P(\psi^{(1)} \text{ reject})$

$\Downarrow$ Union bound

$P(A_1 \cup \cdots \cup A_n) \le \sum_{i=1}^{k} P(A_i)$

Boferroni test: Test whether a group of explanatory variable is significant in linear regression

$\psi_{S,\alpha} = \bigcup_{j\in S} \psi_{j,\alpha/k}$

① $\hat\beta^{MLE} = (X^TX)^{-1}X^TY \sim \mathcal{N}_p(\beta, \sigma^2(X^TX)^{-1})$

$\hat\beta_1 = e_1^T\hat\beta^{MLE}$, $e_1 = (1,0,\cdots,0)^T$

$\hat\beta_1 \sim \mathcal{N}(\underbrace{e_1^T\beta}_{\beta_1}, \underbrace{e_1^T[\sigma^2(X^TX)^{-1}]e_1}_{\sigma^2 M_{11}})$

$T^{(1)} = \dfrac{e_1^T\hat\beta}{\sqrt{\hat\sigma^2 M_{11}}} \Rightarrow \mathbb{1}\{T^{(1)} < q_{1-\frac{\alpha}{k}}^{t_{np}}\}$

② $\hat\beta_2 - \hat\beta_1 = u^T\hat\beta^{MLE}$ $u = (1,-1,0\cdots0)^T$

$\sim \mathcal{N}(\underbrace{u^T\beta}_{\beta_2-\beta_1}, u^T[\sigma^2(X^TX)^{-1}]u)$

$\sigma^2(M_{11}+M_{22}-2M_{12})$

$\Rightarrow T^{(2)} = \dfrac{\hat\beta_2 - \hat\beta_1}{\sqrt{\hat\sigma^2(M_{11}+M_{22}-2M_{12})}} \Rightarrow \mathbb{1}\{T^{(2)} < q_{1-\frac{\alpha}{k}}^{t_{np}}\}$

Ridge Regression

$y_i = x_i^T \beta + \varepsilon_i$, $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

$Y = X\beta + \varepsilon$, $\varepsilon \sim N(0, \sigma^2 I_n)$

$n \quad n \times p \quad p \quad n$

OLS: $\hat{\beta}^{OLS} = \operatorname{argmin} \|Y - X\beta\|_2^2 = (X^T X)^{-1} X^T Y$

1. Bayesian — Assume $\pi \sim N(0, \tau I_p)$

$$\pi(\beta|y) = \frac{\pi(\beta) P(Y|\beta)}{P(Y)} \propto \pi(\beta) P(Y|\beta)$$

$$= \frac{1}{\sqrt{2\pi\tau^2}^p} \frac{1}{\sqrt{2\pi\sigma^2}^n} \exp\left[-\frac{1}{2\sigma^2}\left(\underbrace{\|Y - X\beta\|_2^2 + \frac{\sigma^2}{\tau^2}\|\beta\|_2^2}_{\star}\right)\right]$$

$\Downarrow$ Let $\lambda = \frac{\sigma^2}{\tau^2}$

$$\star = (Y - X\beta)^T (Y - X\beta) + \lambda \beta^T \beta$$

$$= \sigma^2 (\beta - u)^T \Sigma^{-1} (\beta - u) + Y^T Y - \sigma^2 u^T \Sigma u$$

$$\pi(\beta|y) \sim N(u, \Sigma) \Rightarrow \begin{cases} u = (X^T X + \lambda I_p)^{-1} X^T Y \\ \sigma^2 (X^T X + \lambda I_p)^{-1} \end{cases}$$

2. $\hat{\beta}^{(R)} = \operatorname{argmin}_\beta \|Y - X\beta\|_2^2 + \lambda \|\beta\|_2^2$ R: ridge regression

with $\lambda = \frac{\sigma^2}{\tau^2}$, $\hat{\beta}^{(R)} = \hat{\beta}^{(\lambda)}$

3. Special case: assume $p = n$, $X^T X = I_p$

$$\hat{\beta}^{(R)} = (X^T X + \lambda I_p)^{-1} X^T Y = (1 + \lambda) I_p)^{-1} X^T Y$$

$$= X^T Y / (1 + \lambda)$$

$$E[\|\hat{\beta}^{(R)} - \beta\|_2^2] = \frac{\lambda^2}{1+\lambda^2}\|\beta\|_2^2 + \frac{\sigma^2}{1+\lambda^2}$$

Remark

Especially useful when X is rank deficit ($p \gg n$). When X deficit, it looks flat, there's no unique solution or hard to get. $\lambda\|\beta\|$ add some curvature at $\vec{0}$ so quadratic risk looks like "ridge"

Generalized Linear Model

GLM – relax the linear constraint In general no closed form

1. Random component
   $Y|X \sim$ some distribution
2. Regression function
   $g(\mu(x)) = x^T\beta$ $g \circ \mu$ is linear

g: link function
μ: regression function

g need to be monotone increasing and differentiable

Exponential Family Distribution

$\{P_\theta : \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^k$ is k-parameter exponential family on $\mathbb{R}^q$ iff

$$f_\theta(y) = h(y)\exp\left[\sum_{i=1}^{k}\eta_i(\theta)T_i(y) - B(\theta)\right]$$

$k \to k$ $\quad (\eta_1(\theta) \cdots \eta_k(\theta))$ $\begin{pmatrix} T_1(y) \\ \vdots \\ T_k(y) \end{pmatrix}$ $q \to k$

One-parameter canonical exponential family – another parameterization

$$f_\theta(y) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right)$$

$b(\theta)$: log-partition function
ϕ: dispersion parameter

likelihood: $l(\theta) = \log f_\theta(Y)$ meets the following identity

① First Identity $E\left(\frac{\partial l}{\partial \theta}\right) = 0$
$\int f_\theta(y)\,dy = 1 \Rightarrow \int \frac{\partial}{\partial\theta} f_\theta(y)\,dy = 0$

② Second Identity $E\left(\frac{\partial^2 l}{\partial\theta^2}\right) + E\left(\frac{\partial l}{\partial\theta}\right)^2 = 0$

$E(Y) = b'(\theta)$
$\mathrm{Var}(Y) = b''(\theta)\phi$

Link Function

Map range of expectation of Y to $\mathbb{R}$ (range of $x^T\beta$)

Y: Bernoulli $\Rightarrow$
- logit: $\log\left(\frac{\mu(x)}{1-\mu(x)}\right) = x^T\beta$
- Probit: $\Phi^{-1}(\mu(x)) = x^T\beta$

canonical link

The function g that links the mean μ to the canonical parameter θ is call canonical link.

$$g(\mu(x)) = \theta = (b')^{-1}(\mu)$$

I.E : Bernoulli Y.

$$P_{mf} = p^y (1-p)^{1-y}$$

$$= \exp\left(y \underbrace{\log\frac{p}{1-p}}_{\theta} + \underbrace{\log(1-p)}_{-b(\theta)} \right)$$

$$\Rightarrow p = \frac{e^\theta}{1+e^\theta}, \text{ canonical link is the logit link}$$

parameter the model with θ

$$Y_i | X_i = x_i \sim f_\theta(y_i) = \exp\left((y_i\theta_i - b(\theta_i))/\phi + c(y_i, \phi) \right)$$

$$\begin{cases} \mu_i = b'(\theta_i) \\ \theta_i = g(\mu_i) = x_i\beta \end{cases}$$

$$\theta_i = (b')^{-1}(\mu_i)$$

$$= (b')^{-1}(g^{-1}(x_i^T\beta)) \equiv h(x_i^T\beta)$$

$$h = (b')^{-1} \circ g^{-1} = \underbrace{(g \circ b')^{-1}}$$ for canonical link $g = (b')^{-1} \Rightarrow h = I$

$$\Downarrow$$

$$\theta_i = x_i^T\beta$$

$$\Downarrow \text{Log-likelihood}$$

$$\log L(x_1, Y_1, \cdots x_n, Y_n; \theta_1, \cdots \theta_n)$$

$$\Downarrow f_\theta(y_i)$$

$$= \sum_{i=1}^n (Y_i\theta_i(x_i) - b(\theta_i(x_i)))/\phi + \text{const}$$ complex in β but simple in θ

$$= \frac{1}{\phi}\sum_{i=1}^n (Y_i h(x_i^T\beta) - b(h(x_i^T\beta))) + \text{const}$$

$$\Downarrow \text{canonical link } h = I$$

$$\ln(Y, X, \beta) = \frac{1}{\phi}\Sigma(Y_i x_i^T\beta - b(x_i^T\beta))$$

1. Given exponential family there exist param that canonical $\uparrow$ + invertible
2. $l(Y, X\beta)$ is strictly concave if $\phi > 0$
3. Gaussian $\Rightarrow \hat\beta^{ISO} = \hat\beta^{MO} = (X^TX)^{-1}X^TY$

Asymptotic Normality

(X_i, Y_i) be iid with a distribution from Exponential family

$$g(\mu_i) = X_i^T \beta$$

where $\mu_i = E(Y_i|X_i)$, $\mu_i = b'(\theta_i)$ $g(\cdot)$ is link function. θ_i be canonical parameter for each i: $\theta_i = h(X_i^T \beta^*)$

$\Downarrow$ some conditions

$$\sqrt{n}(\hat{\beta}_n^{mle} - \beta^*) \xrightarrow{(d)} N(0, I^{-1}(\beta))$$

Hypothesis Test for Logit

$Y_i \in \{0,1\}$, $X_i \in \mathbb{R}^p$, (X_i, Y_i) independent, $Y_i \sim Ber(p_i)$

$$P(Y_i = y_i | p_i) = \begin{cases} p_i & \text{if } y_i = 1 \\ 1 - p_i & \text{if } y_i = 0 \end{cases} = p_i^{y_i}(1-p_i)^{1-y_i}$$

$$= \exp\left(y_i \underbrace{\log \frac{p_i}{1-p_i}}_{=\theta_i = X_i^T\beta \Leftrightarrow p_i = \frac{e^{\theta_i}}{1+e^{\theta_i}}} + \log(1-p_i)\right)$$

$$\Rightarrow P(Y_i = y_i | X_i^T\beta) = \exp(y_i X_i^T \beta - \log(1 + e^{X_i^T\beta}))$$

X_i iid distribution $\quad Y_i | X_i^T\beta \sim Ber\left(\frac{e^{X_i^T\beta}}{1+e^{X_i^T\beta}}\right)$

$$\ell_n(\beta | y, X) = \frac{1}{n}\left[y_i X_i^T \beta - \log(1 + e^{X_i^T\beta})\right]$$

$$\hat{\beta} = \arg\max \ell_n(\beta | y, X)$$

Test: $H_0: \hat{\beta}_j = \beta_j^0$

$H_1: \hat{\beta}_j \neq \beta_j^0$

Wald Test

$$W_n = n(\hat{\theta} - \theta^0) \underbrace{I(\hat{\theta})}_{= -E[\nabla^2 \ell(\theta | X_i)]} (\hat{\theta} - \theta^0) \xrightarrow[n \to \infty]{(d)} \chi^2_d \quad \text{under } H_0$$

MLE: $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow[n \to \infty]{(d)} N(0, I^{-1}(\theta))$

$$\Rightarrow n(\hat{\theta} - \theta) I(\theta) (\hat{\theta} - \theta) \xrightarrow[n \to \infty]{(d)} \chi^2_d$$

Slutsky: $\Rightarrow n(\hat{\theta} - \theta) I(\hat{\theta}) (\hat{\theta} - \theta) \xrightarrow[n \to \infty]{(d)} \chi^2_d$ ($\hat{\theta}$ is consistent)

$$\psi = \mathbb{1}\{T_n > q_{\alpha}^{\chi^2_d}\}$$