Lec 1: Recap of Probability & Statistics

Yanjun Han

Sept. 3, 2024

Probability: mathematical model of random outcomes

(Gaussian, Poisson, Markov chains, Brownian motion, $\cdots$)

Statistics: given random outcomes, infer the underlying model

(statistical modeling, parameter estimation, testing, confidence interval, regression, $\cdots$)

Central question for this course: given data $x_1, \cdots, x_n$,

1. How to find a mathematical model of P_θ such that $(x_1, \cdots, x_n) \sim P_\theta$?
2. How should we infer the unknown parameter θ?

Three classes of models we'll cover:

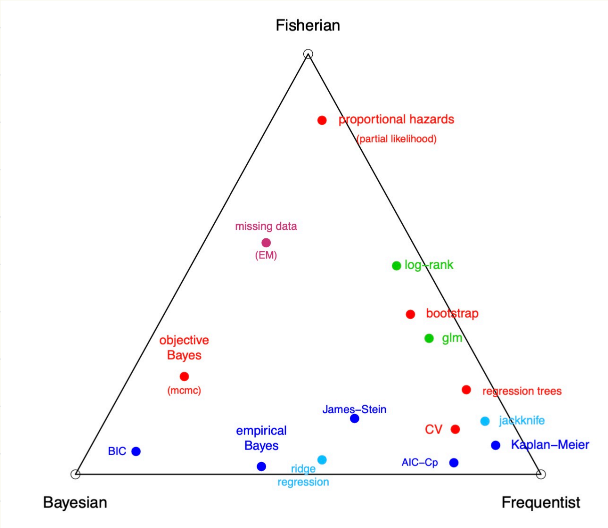
Parametric: $\theta \in \mathbb{R}^d$ finite-dimensional, typically with a linear structure

Semiparametric: $\theta = (\tau, \eta)$, τ: parameter of interest

η: nuisance parameter

Nonparametric: $\theta = f$ is parametrized by a function

An (incomplete) list of topics we'll cover:



Limit theorems in probability

Surprising fact (origins of probability theory): sums of independent, identically distributed (i.i.d.) RVs have universal behavior.

Law of large numbers (LLN)

$X_1, \dots, X_n$ i.i.d, $\mathbb{E}[X]$ exists. then

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i \to \mathbb{E}X \quad \text{as } n \to \infty$$ (in probability & almost surely)

Central limit theorem (CLT)

For iid $X_1, \dots, X_n$ with $\mathbb{E}X = \mu$, $\mathrm{Var}(X) = \sigma^2$,

$$Z_n = \frac{\sqrt{n}(\overline{X_n} - \mu)}{\sigma} \xrightarrow{d} N(0,1) \quad \text{as } n \to \infty$$

→ standard normal distribution

converges in distribution, meaning that

$$\mathbb{P}(Z_n \le t) \longrightarrow \mathbb{P}(N(0,1) \le t) \text{ for every } t \in \mathbb{R}.$$

Delta method

If g is differentiable & $\sqrt{n}(X_n - \mu) \xrightarrow{d} Z \sim N(0,1)$. Then

$$\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{d} g'(\mu)\cdot Z \sim N(0, g'(\mu)^2)$$

Pf (Taylor expansion)

$$g(X_n) = g(\mu) + g'(\mu)(X_n - \mu) + o(|X_n - \mu|)$$

$$\Rightarrow \sqrt{n}(g(X_n) - g(\mu)) = g'(\mu)\cdot \underbrace{\sqrt{n}(X_n - \mu)}_{\xrightarrow{d} Z} + \underbrace{o(\sqrt{n}|X_n - \mu|)}_{\xrightarrow{P} 0}$$

$$\to g'(\mu) Z. \quad \blacksquare$$

Example. $g(x) = x^2$, $g'(\mu) = 2\mu$

$$\Rightarrow \sqrt{n}(\overline{X}_n^2 - \mu^2) \xrightarrow{d} N(0, 4\mu^2\sigma^2).$$

Example of how probability is applied to statistics:

Suppose we're watching the US open, where a player is on 1^{st} serve 100 times, and wins 80 of them.

Statistical model: the wins/losses are independent, with an unknown win rate p

LLN: $\frac{1}{n}\#\text{wins} \to p$ as $n \to \infty$.

Since $n=100$ is large enough, $\hat{p} = 0.8$ is a reasonable estimate of p.

CLT: $\frac{1}{\sqrt{np(1-p)}}(\#\text{wins} - pn) \xrightarrow{d} N(0,1)$ as $n \to \infty$

$\Rightarrow$ the estimation error $\hat{p} - p \approx N(0, 0.04^2)$, so a 95% confidence interval

for p is $p \in [\hat{p} - 2\cdot 0.04, \hat{p} + 2\cdot 0.04] = [0.72, 0.86]$

Estimation: given $X_1, \cdots, X_n \overset{i.i.d.}{\sim} P_\theta$ with a known distribution family $\theta \mapsto P_\theta$ but an unknown parameter θ, the target of estimation is to find θ.

Approach I: estimating equation

Suppose one can find functions $F_1, F_2, \cdots, F_p$ s.t.

$$\mathbb{E}_{X\sim P_\theta}[F_j(\theta, X)] = 0 \qquad \forall j = 1, \cdots, p.$$

Then a reasonable estimator $\hat{\theta}_n$ is defined as the solution to

$$\begin{cases} \frac{1}{n}\sum_{i=1}^{n} F_1(\theta, X_i) = 0 \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} F_p(\theta, X_i) = 0 \end{cases}$$

Analysis: ① By LLN, $0 = \frac{1}{n}\sum_{i=1}^{n} F_j(\hat{\theta}_n, X_i) \approx \mathbb{E}_{X\sim P_\theta}[F_j(\hat{\theta}_n, X)]$

$\Rightarrow \hat{\theta}_n \approx \theta$ (i.e. "consistency")

② CLT can also be used to establish the asymptotic normality of $\sqrt{n}(\hat{\theta}_n - \theta)$

Example: $X_1, \cdots, X_n \overset{i.i.d}{\sim} N(\mu, \sigma^2)$ with unknown (μ, σ^2)

then $\mathbb{E}_{\mu,\sigma^2}[X - \mu] = 0$
$\mathbb{E}_{\mu^2,\sigma^2}[X^2 - \mu^2 - \sigma^2] = 0$
$\Longrightarrow$ $F_1(X, (\mu, \sigma^2)) = X - \mu$
$F_2(X, (\mu, \sigma^2)) = X^2 - \mu^2 - \sigma^2$

The estimator $(\hat{\mu}, \hat{\sigma}^2)$ solves

$$\begin{cases} \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}) = 0 \\ \frac{1}{n}\sum_{i=1}^{n}(X_i^2 - \hat{\mu}^2 - \hat{\sigma}^2) = 0 \end{cases} \Longrightarrow \begin{cases} \hat{\mu} = \overline{X} \\ \hat{\sigma}^2 = \overline{X^2} - \overline{X}^2 \end{cases}$$

Approach II: MLE (maximum likelihood estimator)

Def (likelihood).

Suppose $X_1, \cdots, X_n$ have joint pdf $f_\theta(X_1, \cdots, X_n)$ (or pmf $p_\theta(X_1, \cdots, X_n)$).

The likelihood function is

$$L_n(\theta) = f_\theta(X_1, \cdots, X_n) \quad \text{(pdf viewed as function of } \theta\text{)}$$

The log likelihood function is

$$\ell_n(\theta) = \log L_n(\theta) = \log f_\theta(X_1, \cdots, X_n).$$

MLE: $\hat{\theta}_n = \underset{\theta}{\operatorname{argmax}}\ L_n(\theta) = \underset{\theta}{\operatorname{argmax}}\ \ell_n(\theta)$

Example (cont'd): $f_{\mu,\sigma^2}(X_1, \cdots, X_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(X_i - \mu)^2}{2\sigma^2}\right)$

$$\Longrightarrow \ell_n(\mu, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 - \frac{n}{2}\log(2\pi\sigma^2)$$

F.O.C.: $\begin{cases} \frac{\partial \ell_n}{\partial \mu} = 0 \\ \frac{\partial \ell_n}{\partial \sigma^2} = 0 \end{cases} \Longrightarrow \begin{cases} \hat{\mu} = \overline{X} \\ \hat{\sigma}^2 = \overline{X^2} - \overline{X}^2 \end{cases}$

(more on both topics in future lectures)

<u>Testing</u>: given $X_1, \cdots, X_n \sim p_\theta$, we'd like to test between

$H_0: \theta \in \Theta_0$ vs. $H_1: \theta \in \Theta_1$ $\quad (\Theta_0 \cap \Theta_1 = \phi)$

(null) (alternative)

<u>Test</u>: a function $\underset{\text{set of outcomes}}{\Omega} \longrightarrow \{\underset{\text{accept } H_0}{H_0}, \underset{\text{reject } H_0}{H_1}\}$

Truth \ Test	H_0	H_1
H_0	Correct	Type I error (false positive)
H_1	Type II error (false negative)	Correct

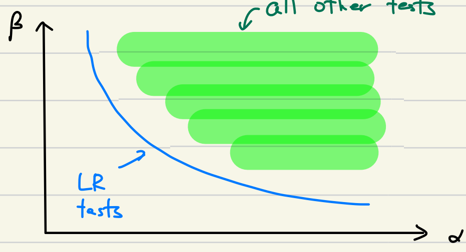
Truth \ Probability	$\mathbb{P}$(output H_0)	$\mathbb{P}$(output H_1)
H_0	$1-\alpha$	α (prob. of type I error, or (significance) level)
H_1	β (prob. of type II error)	$1-\beta$ (power of the test)

<u>Fundamental tradeoff between two types of error</u>

making smaller type I error

$\Rightarrow$ setting higher bar for rejecting H_0

$\Rightarrow$ larger type II error



<u>Simple hypothesis testing</u>: $H_0: \theta = \theta_0$ vs. $H_1: \theta = \theta_1$

<u>Neyman-Pearson lemma</u>: the <u>likelihood ratio test</u> is optimal

$$\text{output} \begin{cases} H_0 & \text{if } \dfrac{P(X_1,\cdots,X_n \mid H_0)}{P(X_1,\cdots,X_n \mid H_1)} > c \\ H_1 & \text{if } \dfrac{P(X_1,\cdots,X_n \mid H_0)}{P(X_1,\cdots,X_n \mid H_1)} \le c \end{cases}$$

<u>Composite hypothesis testing</u> : $|\Theta_0| > 1$ and/or $|\Theta_1| > 1$

Unfortunately, no complete picture here. So statisticians have made:

① compromise I: focus only on significance level

② compromise II: focus only on asymptotic tests ($n \to \infty$)

<u>Idea</u>: find a function $F(\theta, X)$ s.t. $F(\theta, X)$ (asymptotically) follows a known distribution P for every $\theta \in \Theta_0$ (e.g. by CLT);
given a significant level α, find A s.t. $P(A) = 1-\alpha$;

level-α test: reject $H_0: \theta = \theta_0$ if $F(\theta_0, X) \notin A$

$(1-\alpha)$-confidence interval for θ: $C = \{\theta : F(\theta, X) \in A\}$.

<u>Example</u>. Given $X \sim B(n,p)$, then CLT gives

$$\frac{X-np}{\sqrt{np(1-p)}} \xrightarrow{d} N(0,1) \quad \text{as} \quad n \to \infty.$$

level-α test for $H_0: p = p_0$: reject H_0 iff $\left|\frac{X-np_0}{\sqrt{np_0(1-p_0)}}\right| > z_{\alpha/2}$

$(1-\alpha)$-confidence interval for p: $C = \left\{p : \left|\frac{X-np}{\sqrt{np(1-p)}}\right| \le z_{\alpha/2}\right\}$.

<u>p-value</u>: every test can be equivalently represented by a p-value $\in [0,1]$.
and rejects H_0 iff p-value $\le \alpha$

(p-value contains more information than yes/no answers)

likelihood function

<u>Generalized likelihood ratio test</u>:
$$LR = \frac{\max_{\theta \in \Theta_0} L_n(\theta)}{\max_{\theta \in \Theta_0 \cup \Theta_1} L_n(\theta)}$$

<u>Wilk's thm</u>, under mild conditions, $-2\log LR \xrightarrow{d} \chi^2_d$ under H_0
with $d = \dim(\Theta_0 \cup \Theta_1) - \dim(\Theta_0)$

Regression (one of the great ideas in statistics)

Goal: prediction (Given independent/explanatory/predictor variables $x_1, \cdots, x_{p-1}$,
predict dependent/response/outcome variable y)

(p: feature dimension; we use $p-1$ as there's an additional intercept term)

Idea: least-squares: $\underset{\beta_0, \cdots, \beta_{p-1}}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{p-1} x_{i,p-1}) \right)^2$

Much more general than it appears by data transformation:

Ex. 1 Suppose $y \approx \sum_{k=0}^{d} a_k x^k$ has a polynomial relationship with x, then for each data point x_i we can define

$$z_{i,1} = x_i, \quad z_{i,2} = x_i^2, \quad \cdots, \quad z_{i,d} = x_i^d.$$

Then

$$y \approx \sum_{k=0}^{d} a_k x^k = a_0 + \sum_{k=1}^{d} a_k z_k$$

Ex. 2 Suppose $y \approx C_0 e^{C_1 x}$, then by defining $w_i = \log y_i$,

$$y_i \approx C_0 e^{C_1 x_i} \implies w_i \approx \log C_0 + C_1 x_i$$

So (x_i, w_i) has a linear relationship $w_i \approx \beta_0 + \beta_1 x_i$, with
$\beta_0 = \log C_0$, $\beta_1 = C_1$.

Solution:

$$X = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p-1} \end{bmatrix} \in \mathbb{R}^{n \times p}, \quad \beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \in \mathbb{R}^p$$

least squares solution:

$$\min_{\beta_0, \cdots, \beta_{p-1}} \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{p-1} x_{i,p-1}) \right)^2 \longrightarrow \min_{\beta \in \mathbb{R}^p} \| y - X\beta \|^2$$

Theorem The minimizer $\hat{\beta}$ is

$$\hat{\beta} = (X^TX)^{-1}X^Ty,$$

provided that $X^TX \in \mathbb{R}^{p\times p}$ is invertible (requiring $n \geq p$)

Statistical analysis

Statistical modeling: $y = X\beta + e$, $\mathbb{E}[e] = 0$, $\mathrm{Cov}(e) = \sigma^2 I_n$

Then:

- $\mathbb{E}[\hat{\beta}] = \beta$
- $\mathrm{Cov}(\hat{\beta}) = \sigma^2 (X^TX)^{-1}$
- $\mathbb{E}\left[\dfrac{\|y - X\hat{\beta}\|^2}{n-p}\right] = \sigma^2$ (motivating the estimator $\hat{\sigma}^2 = \dfrac{\|y - X\hat{\beta}\|^2}{n-p}$ for σ^2)
- $$\frac{\hat{\beta}_j - \beta_j}{\underbrace{\sqrt{\hat{\sigma}^2[(X^TX)^{-1}]_{jj}}}_{\text{estimated variance of } \hat{\beta}_j}} \xrightarrow{d} N(0,1) \text{ as } n \to \infty.$$

Pf.

$$\hat{e} = y - X\hat{\beta} = y - X(\beta + (X^TX)^{-1}X^Te)$$
$$= y - X\beta - X(X^TX)^{-1}X^Te$$
$$= (I - \underbrace{X(X^TX)^{-1}X^T}_{\text{call it } P})e.$$

Note that P is a projection matrix: $P^T = P$, $P^2 = P$, since

$$P^T = [X(X^TX)^{-1}X^T]^T = X(X^TX)^{-1}X^T = P$$
$$P^2 = X\underbrace{(X^TX)^{-1}X^T \cdot X}_{= I_p}(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T = P.$$

$\Rightarrow (I_n - P)^2 = I_n - 2P + P^2 = I_n - 2P + P = I_n - P$

$\Rightarrow \|\hat{e}\|^2 = \hat{e}^T\hat{e} = e^T(I_n - P)^T(I_n - P)e = e^T(I_n - P)^2 e = e^T(I_n - P)e$

$\Rightarrow \mathbb{E}[\|\hat{e}\|^2] = \mathrm{Tr}(\mathrm{Cov}(e)(I_n - P)) = \sigma^2\,\mathrm{Tr}(I_n - P)$

$(\mathbb{E}[z^TAz] = \mathrm{Tr}(\mathrm{Cov}(z)A))$

Since $\mathrm{Tr}(I_n) = n$ check $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$!

$\mathrm{Tr}(P) = \mathrm{Tr}(X(X^TX)^{-1}X^T) = \mathrm{Tr}(X^TX(X^TX)^{-1}) = \mathrm{Tr}(I_p) = p$

$\Rightarrow \mathrm{Tr}(I_n - P) = \mathrm{Tr}(I_n) - \mathrm{Tr}(P) = n - p$

$\Rightarrow \mathbb{E}[\|\hat{e}\|^2] = (n-p)\sigma^2.$ ∎