STA 355 Lecture 1

logistics.

monwenlong.github.io/teaching/sta355f25.

Quercus. Piazza.

Grading.

$$\max\left\{\begin{array}{l} 30\% \times \text{midterm} + 40\% \times \text{final} + 30\% \times \text{homework}, \\ \qquad (15\% \text{ each}) \\ 70\% \times \text{final} + 30\% \times \text{homework} \end{array}\right\}$$

Homework marks by completion only.

Poll.

1. 50 + 10 + 50 + 10 + 50 ends 8pm
2. 75 + 10 + 75 ends 7:50pm.

TBD on piazza

Textbook : "All of Statistics" by L. Wasserman.

Overview.

- Probability review (quick), Statistical models.
- Methodology
  - Empirical distributions & bootstrap.
  - Parametric inference, M-estimators. MLE, etc.
  - Testing: p-values, likelihood ratio tests.
  - Bayesian inference.

(*) means optional.

- Theory
  - Asymptotics
  - Basics of decision theory
  - Uniform convergence (*).
- Models
  - Linear and generalized linear models.
  - Nonparametric models.
  - classification (*), causal inference (*).

Probability review.

- Basic concepts.

  Sample space Ω, $\omega \in \Omega$ sample outcome.

  $\mathbb{P}$: mapping from subsets of Ω to non-negative reals.

(sometimes some sets are non-measurable, and we restrict attention to measurable subsets)

Axioms of probability.

(i) $\mathbb{P}(A) \geq 0 \quad \forall A$

(ii) $\mathbb{P}(\Omega) = 1.$

(iii). $\mathbb{P}\left(\bigcup_{i=1}^{+\infty} A_i\right) = \sum_{i=1}^{+\infty} \mathbb{P}(A_i)$ If the subsets $(A_i)_{i=1}^{+\infty}$ are disjoint.

(An event is a subset of Ω).

Independence of events.

A, B independent $\iff \mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B).$

Conditional probability.

If $\mathbb{P}(B) > 0$, $\quad \mathbb{P}(A|B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$

For conditioning on zero-probability events.

$B = \{X = x\}$ where X is a continuous r.v.

Bayes theorem. let $(A_i)_{i=1}^{+\infty}$ be a partition of Ω,

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

Random variables.

Sample space Ω.

$$X : \Omega \to \mathbb{R}.$$

$$(\forall \omega \in \Omega, \quad X(\omega) \in \mathbb{R}).$$

- pmf, cdf, and pdf.

Describing the probability distribution of an r.v.

— X takes values in $\{x_1, x_2, x_3, \cdots\}$

(Can be infinite, but needs to be countable).

Probability mass function defined on $\{x_1, x_2 \cdots\}$

$$p_X(x_i) = \mathbb{P}(X = x_i).$$

(By countable summability, $\sum_{i=1}^{+\infty} p_X(x_i) = 1$).

$$p_X(x_i) \geq 0.$$

This does not work for continuous cases.

$\mathbb{P}(X = x) = 0$ for any $x \in \mathbb{R}$.

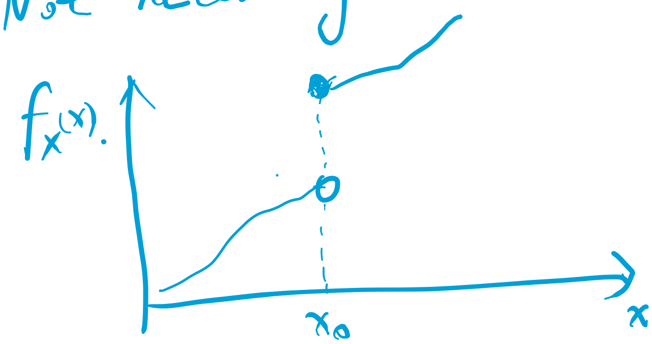
— Alternative approach.

$\forall x \in \mathbb{R}$, define $F_X(x) = \mathbb{P}(X \leq x)$.

" cumulative distribution function".

- Always well-defined.
- Monotonically non-decreasing.
- $\lim_{x \to +\infty} F_X(x) = 1, \quad \lim_{x \to -\infty} F_X(x) = 0.$

- Not necessarily continuous.



This means

$\mathbb{P}(X = x_0) > 0$
"
the jump level

- Probability density function (pdf)

Suppose F_X satisfies $\int_a^b p_X(x)\,dx = F_X(b) - F_X(a)$ $(\forall a, b)$.

$F_X(b) - F_X(a) = \mathbb{P}(X \in [a, b])$

then we call p_X the pdf of X.

- May not exist for certain r.v.

e.g. $X \begin{cases} \sim \mathrm{Unif}([0,1]) & \text{w.p. } 1/2 \\ = \frac{1}{2} & \text{w.p. } 1/2. \end{cases}$

- "Almost" requires F_X to be differentiable, but it still exists under weaker conditions.

· pdf can be extended to multivariate cases.

If there exists a function p_X. s.t.

$$\mathbb{P}(X \in A) = \int_A p_X(x)\, dx$$

d: dimension. for any (measurable) subset $A \subseteq \mathbb{R}^d$

Then we call p_X to be the pdf of X.

For this class: focus on discrete and continuous r.v.'s
(probably also their mixtures).

· Conditional distribution

(X, Y) jointly distributed $\begin{cases} \text{discrete} \\ \text{continuous.} \end{cases}$

Discrete
$$\mathbb{P}(Y = y_i \mid X = x_i) = \frac{\mathbb{P}(X = x_i, Y = y_i)}{\mathbb{P}(X = x_i)}$$

(Assuming $\mathbb{P}(X = x_i) > 0$).

$$= \frac{\mathbb{P}(X = x_i, Y = y_i)}{\sum_{j=1}^{+\infty} \mathbb{P}(X = x_i, Y = y_j)}$$

$\mathbb{P}(\cdot \mid X = x_i)$ defines
a probability distribution
on the discrete set $\{y_1, y_2, \cdots\}$.

Cts case:

$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)} = \frac{p_{X,Y}(x,y)}{\int p_{X,Y}(x,y')\,dy'}.$$

"conditional density of Y at y, given that $X=x$".

e.g. $\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\right)$

$\rho \in (-1, 1)$.

$$p_{X_1,X_2}(x_1, x_2) = \frac{1}{2\pi\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}\begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix}\right)$$

where $\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$ is the covariance matrix.

Conditional distribution.

$$X_2 | X_1 = x_1 \sim N\left(\mu_2 + \frac{\rho\sigma_2}{\sigma_1}(x_1-\mu_1),\ \sigma_2^2 - \rho^2\sigma_2^2\right).$$

Expectation.

X { Discrete r.v. $\mathbb{E}[X] := \sum_{i=1}^{+\infty} x_i\, p_X(x_i)$

(Assuming that $\sum_{i=1}^{+\infty} |x_i| \cdot p_X(x_i) < +\infty$)

Continuous r.v. $\mathbb{E}[X] := \int_{\mathbb{R}} x\, p_X(x)\, dx$

(Assuming that $\int |x| \cdot p_X(x)\, dx < +\infty$)

- Expectations are linear.

$$\mathbb{E}[aX+bY] = a\mathbb{E}[X] + b\mathbb{E}[Y].$$

($a, b \in \mathbb{R}$ are deterministic quantities),

does not require (in)dependence between X, Y.

- There exists r.v. s.t. expectations do not exist.

e.g. "Cauchy distribution".

$$p(x) = \frac{1}{\pi(1+x^2)} \qquad (x \in \mathbb{R})$$

$$\int_0^{+\infty} |x| p(x) dx = \frac{1}{\pi}\int_0^{+\infty} \frac{x\,dx}{1+x^2}$$

$$= \frac{1}{2\pi} \log(1+x^2)\Big|_0^{+\infty} = +\infty.$$

Inequalities.

From expectation to tail behavior.

- Markov ineq. $X \geq 0$, r.v. $\mathbb{E}[X] < +\infty$.

then $$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t} \qquad (\forall t > 0)$$

Tail probability decaying at $1/t$ rate (or faster)

Proof:

$$\mathbb{P}(X \geq t) = \mathbb{E}[\mathbb{1}_{X \geq t}] \leq \mathbb{E}\left[\frac{X}{t} \cdot \mathbb{1}_{X \geq t}\right]$$

$$= \frac{1}{t}\mathbb{E}[X \cdot \mathbb{1}_{X \geq t}] \leq \frac{1}{t}\mathbb{E}[X].$$

Chebyshev ineq. Suppose $\mathbb{E}[|X|^2] < +\infty$

(stronger requirement than Markov ineq).

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

($1/t^2$, faster decaying tail).

Proof: Apply Markov ineq to $Y = |X - \mathbb{E}[X]|^2$.

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) = \mathbb{P}(Y \geq t^2) \leq \frac{\mathbb{E}[Y]}{t^2} = \frac{\mathrm{var}(X)}{t^2}.$$

In general, if $\mathbb{E}[|X - \mathbb{E}[X]|^p] < +\infty$.

then $$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{\mathbb{E}[|X - \mathbb{E}[X]|^p]}{t^p}.$$

Proof: same as chebyshev.

Rmk: tighter tail bound (eg. exponentially decaying) is possible, assuming $\exists$ moment generating functions.

Convergence of random variables.

Notions of convergence.

- almost sure convergence

$$X_1, X_2, \dots \quad X_n, \dots$$

we say $X_n \xrightarrow{a.s.} X$ if $\mathbb{P}(\lim_{n\to+\infty} X_n = X) = 1$.

- convergence in probability.

we say $X_n \xrightarrow{\mathbb{P}} X$ when

$$\forall \varepsilon > 0, \quad \lim_{n\to+\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0.$$

- L^p convergence.

we say $X_n \xrightarrow{L^p} X$ when

$$\lim_{n\to+\infty} \mathbb{E}[|X_n - X|^p] = 0.$$