

PH 712 Probability and Statistical Inference

Part I: Random Variable

Zhiyang Zhou (zhou67@uwm.edu, zhiyanggeezhou.github.io)

2024/Sep/04 13:09:59

“All models are wrong, but some are useful.”

— G. E. P. Box. (1976). *Journal of the American Statistical Association*, 71:791–799.

What is a statistical model?

- Two types of statistical models (Breiman, 2001)
 - Stochastic model vs. machine learning model (PH 812 Statistical Learning & Data Mining)
- Stochastic model: the distribution of random variables (RVs) of interest
 - Recall the linear regression and logit regression (PH 711 Intermediate Biostatistics)
 - Parametric vs non-parametric vs semi-parametric

Statistical modelling

Confirmed RVs of interest → Data collection and cleaning → Specified models → Model fitting and inference
→ Interpretation

Statistical inference

- What is the underlying true model?
 - E.g., is the RV distributed as $\mathcal{N}(0, 1)$?

Characterization the distribution of an RV (CB/HMC Chp. 1)

- An RV is a real-valued function.
- The cumulative distribution function (cdf) of RV X , say F_X , is defined as

$$F_X(x) = \Pr(X \leq x), \quad x \in \mathbb{R}.$$

- F_X satisfies following three properties:
 - * (Right continuous) $\lim_{x \rightarrow x_0^+} F_X(x) = F_X(x_0)$;
 - * (Non-decreasing) $F_X(x_1) \leq F_X(x_2)$ for $x_1 \leq x_2$;
 - * (Ranging from 0 to 1) $F_X(-\infty) = 0$ and $F_X(\infty) = 1$.
- Reversely, a function satisfying the three above properties must be a cdf for certain RV.
 - * Indicating an one-to-one correspondence between the set of all the RVs and the set of all the cdfs
- Knowing the distribution of an RV \Leftrightarrow knowing the cdf
- Discrete RV
 - RV X merely takes countably different values

- Probability mass function (pmf): $p_X(x) = \Pr(X = x)$
 - * $F_X(x) = \sum_{u \leq x} p_X(u)$
 - * $p_X(x) = F_X(x) - \lim_{u \rightarrow x^-} F_X(u)$
- Knowing the distribution of a discrete RV \Leftrightarrow knowing the pmf
- Continuous RV
 - RV X is continuous \Leftrightarrow its cdf F_X is (absolutely) continuous, i.e., there exists f_X such that

$$F_X(x) = \int_{-\infty}^x f_X(u) du, \quad \forall x \in \mathbb{R}.$$

- * Probability density function (pdf): $f_X(x) = dF_X(x)/dx = \lim_{\delta \rightarrow 0^+} \Pr(x < X \leq x + \delta)/\delta$.
- Knowing the distribution of a continuous RV \Leftrightarrow knowing the pdf

Example Lec1.1

- Given $p \in (0, 1)$, suppose

$$F_X(x) = \begin{cases} 1 - (1 - p)^{\lfloor x \rfloor}, & x \geq 1, \\ 0, & \text{otherwise,} \end{cases}$$

where $\lfloor x \rfloor$ represents the integer part of x .

- Show that F_X is a cdf. (Hint: Check all the three properties of cdf, especially the right-continuity of F at positive integers.)
- What is the type of X , discrete or continuous?

Support of RV (CB pp. 50 & HMC pp. 46)

- For discrete RV X with pmf p_X
 - $\text{supp}(X) = \{x \in \mathbb{R} : p_X(x) > 0\}$
 - E.g., support of $\text{Binom}(n, p)$ is $\{0, \dots, n\}$
- For continuous RV X with pdf f_X
 - $\text{supp}(X) = \{x \in \mathbb{R} : f_X(x) > 0\}$
 - E.g., support of $\mathcal{N}(0, 1)$ is \mathbb{R}

Example Lec1.2

- Revisit F_X defined in Example Lec1.1, i.e.,

$$F_X(x) = \begin{cases} 1 - (1 - p)^{\lfloor x \rfloor}, & x \geq 1, \\ 0, & \text{otherwise,} \end{cases}$$

where $\lfloor x \rfloor$ represents the integer part of x .

- What is the support of X ?

Characterization the distribution of an RV (con'd)

- Moment generating function (MGF, CB Sec. 2.3)
 - $M_X(t) = E\{\exp(tX)\}$
 - * Continuous X : $M_X(t) = \int_{-\infty}^{\infty} \exp(tx) f_X(x) dx$
 - * Discrete X : $M_X(t) = \sum_{\{x: x \in \text{supp}(X)\}} \exp(tx) p_X(x)$
 - The MGF of X is $M_X(t)$, $t \in A$, $\Leftrightarrow M_X(t)$ is finite for t in a neighborhood of 0, say A ; otherwise the MGF does NOT exist or is NOT well defined.
 - $M_{aX+b}(t) = \exp(bt) M_X(at)$
 - Knowing the distribution of an RV \Leftrightarrow knowing the MGF (if any)
- Characteristic function (CF, optional)
 - $\varphi_X(t) = E\{\exp(itX)\}$

- * Continuous X : $\varphi_X(t) = \int_{-\infty}^{\infty} \exp(itx) f_X(x) dx$
- * Discrete X : $\varphi_X(t) = \sum_{\{x: x \in \text{supp}(X)\}} \exp(itx) p_X(x)$
- Always well-defined
- $\varphi_{aX+b}(t) = \exp(bt) \varphi_X(at)$
- Knowing the distribution of an RV \Leftrightarrow knowing the CF

Example Lec1.3

- Find the MGFs of following distributions
 - $\mathcal{N}(\mu, \sigma^2)$, i.e., $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
 - (Cauchy distribution) $f_X(x) = \{\pi(1+x^2)\}^{-1}$, $x \in \mathbb{R}$

Indicator function

Given a set A , the indicator function of A is

$$\mathbf{1}_A(x) = \begin{cases} 1, & x \in A, \\ 0, & \text{otherwise.} \end{cases}$$

Example Lec1.4

- Revisit F_X defined in Example Lec1.1, i.e.,

$$F_X(x) = \begin{cases} 1 - (1-p)^{\lfloor x \rfloor}, & x \geq 1, \\ 0, & \text{otherwise,} \end{cases}$$

where $\lfloor x \rfloor$ represents the integer part of x .

- Please reformulate F_X with the indicator function of $A = \{x : x \geq 1\}$.