STAT 3690 Lecture 03

zhiyanggeezhou.github.io

Zhiyang Zhou (zhiyang.zhou@umanitoba.ca)

Jan 28th, 2022

"All models are wrong, but some are useful."

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

Statistical modelling

- What is a statistical model?
  - (Joint) distribution of random variable (RV) of interest
- Rephrase linear regression and logit regression models in terms of distributions

Characterizing/representing univariate distributions

- (scalar-valued) RV X: a real-valued function defined on a sample space Ω
- Cumulative distribution function (cdf): $F_X(x) = \Pr(X \leq x)$
  - right continuous with respect to x
  - non-decreasing w.r.t. x
  - ranging from 0 to 1
- Discrete RV
  - RV X takes countable different values.
  - probability mass function (pmf): $p_X(x) = \Pr(X = x)$
- Continuous RV
  - RV X is continuous iff its cdf F_X is absolutely continuous with respect to x, i.e., $\exists f_X$, s.t.

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

  - probability density function (pdf): $f_X(x) = F'_X(x)$.
- Characteristic function
- Moment-generating function

Characterizing/representing joint/multivariate distributions

- Random vector/vector-valued RV
  - $\mathbf{X} = [X_1, \ldots, X_p]^\top$
- Joint cumulative distribution function (joint cdf): $F_{\mathbf{X}}(x_1, \ldots, x_p) = \Pr(X_1 \leq x_1, \ldots, X_p \leq x_p)$
  - right continuous w.r.t. each x_i
  - non-decreasing w.r.t. each x_i
  - ranging from 0 to 1
- Joint distribution of continuous RVs

- Joint pdf/density:

$$f_{\mathbf{X}}(x_1, \ldots, x_p) = \frac{\partial^p}{\partial x_1 \cdots \partial x_p} F_{\mathbf{X}}(x_1, \ldots, x_p)$$

  - Multivariate normal (MVN) distribution
- Joint distribution of discrete RVs
  - Joint pmf:

$$p_{\mathbf{X}}(x_1, \ldots, x_p) = \Pr(X_1 = x_1, \ldots, X_p = x_p)$$

  - Multinomial distribution
- Multivariate characteristic/moment-generating functions

- Exercise: Suppose that we independently observe an experiment that has m possible outcomes $O_1, \ldots, O_m$ for n times. Let $p_1, \ldots, p_k$ denote probabilities of $O_1, \ldots, O_m$ in each experiment respectively. Let X_i denote the number of times that outcome O_i occurs in the n repetitions. What is the joint pmf of $\mathbf{X} = [X_1, \ldots, X_m]^\top$?

$$\begin{aligned}
&p_X(x_1, \cdots, x_m) \\
&= \Pr(X_1 = x_1, \cdots, X_m = x_m) \\
&= \binom{n}{x_1} \cdots \binom{n - \sum_{i=1}^{m-1} x_i}{x_m} p_1^{x_1} \cdots p_m^{x_m} \\
&= \frac{n!}{x_1! \cdots x_m!} p_1^{x_1} \cdots p_m^{x_m}, \quad x_1, \cdots, x_m \in \mathbb{Z}^+ \cup \{0\}, \ \textstyle\sum_i x_i = n
\end{aligned}$$

and 0 otherwise

Marginalization

- $\mathbf{X} = [X_1, \ldots, X_p]^\top$ $\mathbf{Y} = [X_1, \ldots, X_q]^\top$, and $q < p$.
- Marginal cdf

$$F_{\mathbf{Y}}(x_1, \ldots, x_q) = \lim_{x_i \to \infty \text{ for all } i > q} F_{\mathbf{X}}(x_1, \ldots, x_p)$$

- Marginal pdf of $\mathbf{Y}$ (when $X_1, \ldots, X_p$ are all continous)

$$f_{\mathbf{Y}}(x_1, \ldots, x_q) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(x_1, \ldots, x_p) \mathrm{d}x_{q+1} \cdots \mathrm{d}x_p$$

- Marginal pmf of $\mathbf{Y}$ (when $X_1, \ldots, X_p$ are all discrete)

$$p_{\mathbf{Y}}(x_1, \ldots, x_q) = \sum_{x_{q+1}=-\infty}^{\infty} \cdots \sum_{x_p=-\infty}^{\infty} p_{\mathbf{X}}(x_1, \ldots, x_p)$$

- "marginal" is used to distinguish pdf/pmf of $\mathbf{Y}$ from the joint pdf/pmf of $\mathbf{X}$.

Conditioning = joint/marginal

$\mathbf{Y} = [y_1, \ldots, y_q]^\top$ and $\mathbf{X} = [x_1, \ldots, x_p]^\top$

- Conditional pdf of $\mathbf{Y}$ given $\mathbf{X}$

$$f_{\mathbf{Y}|\mathbf{X}}(y_1, \ldots, y_q \mid x_1, \ldots, x_p) = \frac{f_{\mathbf{X},\mathbf{Y}}(x_1, \ldots, x_p, y_1, \ldots, y_q)}{f_{\mathbf{X}}(x_1, \ldots, x_p)}$$

- Conditional pmf of $\mathbf{Y}$ given $\mathbf{X}$

$$p_{\mathbf{Y}|\mathbf{X}}(y_1, \ldots, y_q \mid x_1, \ldots, x_p) = \frac{p_{\mathbf{X},\mathbf{Y}}(x_1, \ldots, x_p, y_1, \ldots, y_q)}{p_{\mathbf{X}}(x_1, \ldots, x_p)}$$

Transformation of random variables (p-dimentional case)

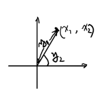
- Let $g = (g_1, \ldots, g_p)$: $\mathbb{R}^p \to \mathbb{R}^p$ be a one-to-one map with inverse $g^{-1} = (g_1^{-1}, \ldots, g_p^{-1})$.
- $\mathbf{Y} = g(\mathbf{X})$ and $\mathbf{X} = g^{-1}(\mathbf{Y})$ are both continuous p-random vectors.
- Jacobian matrix of g^{-1} is $\mathbf{J} = [\partial g_i^{-1}(y_1, \ldots, y_p)/\partial y_j]_{p \times p} = [\partial x_i/\partial y_j]_{p \times p}$.
  - $|\det(\mathbf{J})| = |\det([\partial y_i/\partial x_j]_{p \times p})|^{-1}$ if replace x_j with $g^{-1}(y_1, \ldots, y_p)$
- $f_{\mathbf{X}}$ is known. Then

$$f_{\mathbf{Y}}(y_1, \ldots, y_p) = f_{\mathbf{X}}(h_1^{-1}(y_1, \ldots, y_p), \ldots, h_p^{-1}(y_1, \ldots, y_p))|\det(\mathbf{J})|$$

- Exercise: Let $\mathbf{X} = [X_1, X_2]^\top$ follow the standard bivariate normal, i.e., its pdf is

$$f_{\mathbf{X}}(x_1, x_2) = (2\pi)^{-1} \exp\{-(x_1^2 + x_2^2)/2\}, \quad (x_1, x_2) \in \mathbb{R}^2.$$

  Find out the joint pdf of $\mathbf{Y} = [Y_1, Y_2]^\top$, where $Y_1 = \sqrt{X_1^2 + X_2^2}$ and $0 \le Y_2 < 2\pi$ is angle from the positive x-axis to the ray from the origin to the point (X_1, X_2), that is, Y is X in polar co-ordinates.



① figure out support of Y (i.e., the part of domain for pdf/cdf to be non-zero)

$(x_1, x_2) \in \mathbb{R}^2 \iff \begin{cases} y_1 > 0 \\ y_2 \in [0, 2\pi) \end{cases}$

② find out J

$x_1 = y_1 \cos(y_2)$

$x_2 = y_1 \sin(y_2)$

$\Rightarrow J = \begin{bmatrix} \cos(y_2) & -y_1 \sin(y_2) \\ \sin(y_2) & y_1 \cos(y_2) \end{bmatrix}$

$\Rightarrow |\det(J)| = y$

③ pdf of Y

$$f_Y(y_1, y_2) = \begin{cases} (2\pi)^{-1} \exp(-y_1^2/2) \cdot y_1, & y_1 > 0, \ 0 \le y_2 < 2\pi \\ 0, & \text{otherwise} \end{cases}$$