PH 712 Probability and Statistical Inference

Recap for Midterm

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

2024/Nov/06 21:33:21

Probability

- Probability: A function quantifying the occurrence likelihood of an event
  - Event: a subset of the sample space (Ω) which is the set of all the possible outcomes
- Conditional probability of B given A (with $\Pr(A) > 0$): the occurrence probability of B, given that A has already occurred
  - (Bayes' theorem) $\Pr(A_i \mid B) = \Pr(A_i)\Pr(B \mid A_i) / \sum_{n=1}^{N} \Pr(A_n)\Pr(B \mid A_n)$ if $\{A_n\}_{n=1}^{N}$ are mutually exclusive and $\Omega = \bigcup_{n=1}^{N} A_n$
- Independence between events B and A (i.e., $B \perp A$): $\Pr(B \cap A) = \Pr(A)\Pr(B)$, or equiv. $\Pr(B \mid A) = \Pr(B)$

Random variable (RV)

- RV: encoding the entries of the sample space (i.e., all the possible outcomes)
- Knowing the distribution of an RV $\Leftrightarrow$ knowing one of the following functions
  - Cumulative distribution function (cdf): $\Pr(X \leq x)$
    - "$X \leq x$" short for the event $\{\omega \in \Omega : X(\omega) \leq x\}$
  - Probability mass function (pmf, specifically for discrete RVs: $\Pr(X = x)$
    - "$X = x$" short for the event $\{\omega \in \Omega : X(\omega) = x\}$
  - Probability density function (pdf, specifically for continuous RVs): derivative of the cdf with respect to x
  - Moment generating function (mgf, not always existing)
- Support: $\operatorname{supp}(X) = \{x \in \mathbb{R} : p_X(x) \text{ or } f_X(x) > 0\}$
- Expectation

$$\mathrm{E}\{g(X)\} = \begin{cases} \int_{x \in \operatorname{supp}(X)} g(x) f_X(x) \mathrm{d}x & \text{for continuous } X \\ \sum_{x \in \operatorname{supp}(X)} g(x) p_X(x) & \text{for discrete } X \end{cases}$$

  - Examples
    - Taking $g(X) = X$

$$\mathrm{E}(X) = \begin{cases} \int_{x \in \operatorname{supp}(X)} x f_X(x) \mathrm{d}x & \text{for continuous } X \\ \sum_{x \in \operatorname{supp}(X)} x p_X(x) & \text{for discrete } X \end{cases}$$

      - $\mathrm{E}(aX + b) = a\mathrm{E}(X) + b$ for constants a and b
    - Taking $g(X) = \exp(tX)$, $\mathrm{E}\{g(X)\}$ is the mgf if it is finite at least for t in a neighborhood of 0
    - Taking $g(X) = X^k$ with positive integer k:

$$\mathrm{E}(X^k) = \begin{cases} \int_{x \in \operatorname{supp}(X)} x^k f_X(x) \mathrm{d}x & \text{for continuous } X \\ \sum_{x \in \operatorname{supp}(X)} x^k p_X(x) & \text{for discrete } X \end{cases}$$

· $\mathrm{E}(X^k) = M^{(k)}(0)$ if the mgf $M(t)$ is well-defined

∗ Taking $g(X) = \{X - \mathrm{E}(X)\}^2$:

$$\mathrm{var}(X) = \mathrm{E}[\{X - \mathrm{E}(X)\}^2] = \begin{cases} \int_{x \in \mathrm{supp}(X)} \{x - \mathrm{E}(X)\}^2 f_X(x) \mathrm{d}x & \text{for continuous } X \\ \sum_{x \in \mathrm{supp}(X)} \{x - \mathrm{E}(X)\}^2 p_X(x) & \text{for discrete } X \end{cases}$$

· $\mathrm{var}(X) = \mathrm{E}(X^2) - \{\mathrm{E}(X)\}^2$

· $\mathrm{var}(aX + b) = a^2 \mathrm{var}(X)$

· $\mathrm{sd}(X) = \sqrt{\mathrm{var}(X)}$: the standard deviation of X

∗ Taking $g(X) = \mathbf{1}_A(X)$:

$$\mathrm{E}\{\mathbf{1}_A(X)\} = \Pr(X \in A)$$

Univariate transformation: finding the distribution of $Y = g(X)$, given the distribution of X

- Figure out $\mathrm{supp}(Y) = \{y : y = g(x), x \in \mathrm{supp}(X)\}$
- For discrete Y with discrete X: $p_Y(y) = \Pr(Y = y) = \Pr(g(X) = y)$
- For continuous Y
  - $F_Y(y) = \Pr\{g(X) \le y\}$
  - $f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} \int_{\{x: g(x) \le y\}} f_X(x) \mathrm{d}x$
    - ∗ Integration region $\{x : g(x) \le y\}$ may be expressed in terms of a series of intervals with endpoints as functions of y, say $[a(y), b(y)]$, $[c(y), d(y)]$, etc.
    - ∗ The integration of f_X is often avoidable by employing the Leibniz Rule (CB Thm. 2.4.1):

$$\frac{\mathrm{d}}{\mathrm{d}y} \int_{a(y)}^{b(y)} f(x) \mathrm{d}x = f\{b(y)\} \frac{\mathrm{d}}{\mathrm{d}y}\{b(y)\} - f\{a(y)\} \frac{\mathrm{d}}{\mathrm{d}y}\{a(y)\}$$

      with $a(y)$ and $b(y)$ both differentiable with respect to y.

- If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Y = aX + b$, $a \neq 0$, is also normally distributed. Specifically, $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

Normal sampling theory

- RVs $X_1, \ldots, X_n$: a random sample of size n
  - Independent and identically distributed (iid) sample: $X_1, \ldots, X_n$ are iid
  - iid normal sample: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$
- Statistic: any function of a random sample, e.g.,
  - Sample mean: $\bar{X} = n^{-1} \sum_{i=1}^n X_i$
  - Sample variance: $S^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$
  - Sample standard deviation: $S = \sqrt{(n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X})^2}$
- Identities for an iid normal sample
  - $\sum_{i=1}^n X_i^2 \sim \chi^2(n)$ if $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$
    - ∗ $Q \sim \chi^2(n) \Rightarrow \mathrm{E}(Q) = n$ and $\mathrm{var}(Q) = 2n$
  - $Z/\sqrt{Q/n} \sim t(n)$ if $Z \sim \mathcal{N}(0, 1)$ and $Q \sim \chi^2(n)$ are independent of each other
  - $(P/m)/(Q/n) \sim F(m, n)$ if $P \sim \chi^2(m)$ and $Q \sim \chi^2(n)$ are independent of each other
  - $n^{1/2}(\bar{X} - \mu)/\sigma \sim \mathcal{N}(0, 1)$
  - $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$
  - $\bar{X} \perp S^2$
  - $n^{1/2}(\bar{X} - \mu)/S \sim t(n-1)$

Parametric model

- The true model assumed to be an element of a set of pdfs/pmfs $\{f(\cdot \mid \theta) : \theta \in \Theta\}$
  - Finding the true model reducing to locating the true parameter $\theta_0 \in \Theta$
  - θ_0 unknown but believed to be fixed (frequentist statistics)

Point estimation

- Method of moments (MM) (for iid sample)
  1. Equate the kth-order RAW moments ($\mathrm{E}(X_1^k)$) to its empirical counterpart ($n^{-1}\sum_{i=1}^n X_i^k$).
     - Better to work with a small k
  2. Solve the resulting equation(s) for θ.
- Maximum likelihood (ML)
  - $L(\theta)$ is the joint pdf/pmf of the sample (consisting of n RVs) with emphasis on $\theta \in \Theta$
    * For an independent sample $L(\theta) = \prod_{i=1}^n f_{X_i}(X_i \mid \theta)$, $\theta \in \Theta$
  - $\hat{\theta}_{\mathrm{ML}}$ is the maximizer of $L(\theta)$ (or $\ell(\theta) = \ln L(\theta)$) within Θ
    * For discrete Θ: compare $L(\theta)$ (or $\ell(\theta)$) over all the possible values of θ
    * For continuous Θ:
      · If $\ell'(\theta) = 0$ has no solution: utilize the monotonicity of $L(\theta)$ (or $\ell(\theta)$)
      · If $\ell'(\theta) = 0$ has at least one solution: get the solution(s) (i.e., stationary point(s)) and then compare $L(\theta)$ (or $\ell(\theta)$) over all the stationary points and boundary points of Θ
  - Invariance property: $\widehat{g(\theta)}_{\mathrm{ML}} = g(\hat{\theta}_{\mathrm{ML}})$

Evaluating estimators

- $\mathrm{MSE}(\hat{\theta}) = \mathrm{Bias}^2(\hat{\theta}) + \mathrm{var}(\hat{\theta})$
- Cramér-Rao lower bound (CRLB) for the variance of any unbiased estimator of $g(\theta)$: if $\mathrm{E}(T_n) = g(\theta)$, then $\mathrm{var}(T_n) \geq \{g'(\theta)\}^2 / I_n(\theta)$
  - Fisher information $I_n(\theta) = \mathrm{var}\{\ell'(\theta)\} = \mathrm{E}[\{\ell'(\theta)\}^2] = -\mathrm{E}\{\ell''(\theta)\}$
  - If $g(\theta) = \theta$, then CRLB becomes $I_n^{-1}(\theta)$.
- If $\mathrm{E}(T_n) = g(\theta)$, then $\mathrm{Efficency}(T_n) = \mathrm{CRLB}/\mathrm{var}(T_n)$.
  - The higher efficiency the better (typically up to 1);
  - T_n is an efficient estimator for $g(\theta) \iff \mathrm{E}(T_n) = g(\theta)$ and its efficiency = 1.