Статистичний простір. Вибірка, оцінки та їх властивості

1 Теоретичні відомості

Розглянемо статистичний простір $(\Omega, \mathcal{F}, \{\mathbf{P}_\theta, \theta \in \Theta\})$, де Ω є простором елементарних подій, $\mathcal{F}$ – сигма-алгебра, пов'язана з Ω, $\{\mathbf{P}_\theta, \theta \in \Theta\}$ – набір ймовірнісних мір, залежних від параметра $\theta \in \Theta$, Θ – параметричний простір. Форму залежності $\mathbf{P}_\theta$ від θ вважаємо відомою. Невідомим для дослідника є параметр θ.

Вибіркою будемо називати вимірну функцію $X : \Omega \to S$, задану на просторі $(\Omega, \mathcal{F}, \mathbf{P}_\theta)$, зі значеннями у просторі (S, Σ, λ). Простір (S, Σ, λ) будемо називати вибірковим простором.

Якщо вибірковий простір можна подати у вигляді прямого добутку просторів $(R, \mathcal{B}, \nu)$, тобто

$$S = \underbrace{R \times \ldots \times R}_{n},$$

$$\Sigma = \sigma[B_1 \times \ldots \times B_n,\ B_k \in \mathcal{B}],\ \lambda(B_1 \times \ldots \times B_n) = \nu(B_1) \times \ldots \times \nu(B_n),\ B_k \in \mathcal{B}.$$

Простір вище (S, Σ, λ) називають n-кратним добутком простору $(R, \mathcal{B}, \nu)$. Кратною вибіркою будемо називати випадковий вектор на $(\Omega, \mathcal{F}, \mathbf{P}_\theta)$, координати якого є незалежними в сукупності та однаково розподіленими.

Задача статистичного оцінювання така: маючи реалізацію вектора $X(\omega) = x \in S$, дослідник має оцінити невідомий параметр $\theta \in \Theta$ розподілу $\mathbf{P}_\theta$.

Вимірну функцію $T(X(\omega)) : \Omega \to C$ від статистичної вибірки називають статистикою. Також статистикою називають функцію від значень які може приймати вибірка: $T(x) : S \to C$.

Якщо $T(x) \in \Theta$ для всіх $x \in S$, тоді статистику $T(X)$ називають оцінкою.

Декілька базових властивостей, які можуть мати оцінки:

1. Незміщеність: $\mathbf{E}_\theta\left[\hat{\theta}_n\right] = \theta$ для всіх $\theta \in \Theta$.
2. Асимптотична незміщеність: $\mathbf{E}_\theta\left[\hat{\theta}_n\right] \to \theta$ при $n \to \infty$ для всіх $\theta \in \Theta$.
3. Консистентність: $\hat{\theta}_n \to^{P_\theta} \theta$ для всіх $\theta \in \Theta$ (збіжність за ймовірністю).
4. Строга консистентність: $\hat{\theta}_n \to^{P_\theta 1} \theta$ для всіх $\theta \in \Theta$ (збіжність майже напевно).
5. Асимптотична нормальність: існує така послідовність $\{c_n(\theta)\}_{n \geq 1}$, що для всіх $\theta \in \Theta$ (слабка збіжність / збіжність в основному)

$$c_n(\theta)(\hat{\theta}_n - \theta) \to^{W_\theta} \xi \sim N(0, 1),\ n \to +\infty.$$

Пригадайте перелічені види збіжностей та відповідні граничні теореми!

Зауваження. Коли ми займаємося оцінюванням вектора невідомих параметрів θ, як у нормальному розподілі скажімо, то ми можемо будувати окремі оцінки для кожного з параметрів (необов'язково для всіх). У такому разі ми це можемо розглядати як оцінювання функції від векторного параметра $\tau(\theta)$.

Приклад: $X = (X_1, \dots, X_n)$ – кратна вибірка, $X_1 \sim N(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, +\infty)$. Вибіркове середнє $\overline{X}_n = \sum_{j=1}^n X_j/n$ – незміщена консистентна оцінка $\tau(\theta) = \mu$.

Доведемо незміщеність:

$$\mathbf{E}_\theta \left[\overline{X}_n\right] = \frac{1}{n} \sum_{j=1}^n \mathbf{E}_\theta \left[X_j\right] = \frac{1}{n} n\mu = \mu,\ \theta \in \Theta.$$

Оцінка є строго консистентною (і, як наслідок, консистентною) оцінкою μ. Дійсно, оскільки $\mathbf{E}_\theta \left[X_1\right] < \infty$, то за критерієм Колмогорова про ПЗВЧ маємо

$$\overline{X}_n \to^{P_\theta 1} \mu,\ n \to \infty,\ \theta \in \Theta.$$

Доведемо асимптотичну нормальність. Зауважимо, що X_j задовольняють класичній ЦГТ (внаслідок інтегровності перших двох теоретичних моментів), а тому

$$\frac{\sum_{j=1}^n X_j - n\mu}{\sqrt{n\sigma^2}} \to^{W_\theta} \xi \sim N(0,1),\ n \to +\infty$$

З іншого боку,

$$\frac{\sum_{j=1}^n X_j - n\mu}{\sqrt{n\sigma^2}} = \frac{n \cdot (\overline{X}_n - \mu)}{\sqrt{n\sigma^2}} = \sqrt{\frac{n}{\sigma^2}} \cdot (\overline{X}_n - \mu)$$

Тобто нормуюча послідовність c_n справді існує і має вигляд $c_n = K\sqrt{n}$. Якщо $K = 1$, тоді

$$\sqrt{n}(\overline{X}_n - \mu) \to^{W_\theta} N(0, V(\theta)),\ n \to +\infty,$$

де $V(\theta) = \sigma^2$ – асимптотична дисперсія оцінки $\overline{X}_n$.

Нормуюча послідовність c_n намагається вхопити швидкість збіжності оцінки до невідомого параметра. Якщо c_n буде занадто 'швидкою', тоді нормована оцінка прямуватиме до нескінченності. Інакше, якщо обрати занадто 'повільну' послідовність, то нормована послідовність буде прямувати до нуля. У прикладі вище дослідимо дисперсію нормованої оцінки, якщо вважати що $c_n > 0$:

$$\begin{aligned}\mathbf{D}_\theta \left[c_n(\overline{X}_n - \mu)\right] &= c_n^2 \frac{1}{n^2} \mathbf{D}_\theta \left[\sum_{j=1}^n X_j\right] = |\textbf{чому?}| = c_n^2 \frac{1}{n^2} \sum_{j=1}^n \mathbf{D}_\theta \left[X_j\right] = \\ &= c_n^2 \frac{1}{n} \mathbf{D}_\theta \left[X_1\right] = \left(\frac{c_n}{\sqrt{n}}\right)^2 \sigma^2 \to \begin{cases} 0, & c_n/\sqrt{n} \to 0, \\ \sigma^2, & c_n/\sqrt{n} \to K \in (0, \infty), \\ \infty, & c_n/\sqrt{n} \to \infty. \end{cases}\end{aligned}$$

2 Задачі

2.1 Задача 1

Статистик має асиметричну монету зі сторонами А та Б. Для цієї монети він хоче оцінити характер асиметрії виходячи з ймовірностей результатів підкидання. Для цього бідний статистик підкидає 1000 разів монету, кожне з підкидань є незалежним, після чого чого має набір з результатів підкидання монети.

Описати статистичний простір, вибірку. У рамках заданої моделі, запропонувати потрібну оцінку.

Розв'язання

Якщо вибірка містить всю інформацію про експеримент, то для спрощення ототожнюють простори, тобто $(\Omega, \mathcal{F}) = (S, \Sigma)$.

Результат підкидання монети в 1000 незалежних випробувань можна змоделювати як випадковий вектор з 1000 незалежних в сукупності координат з однаковим розподілом (оскільки процедуру експерименту можна вважати незмінною для всіх випробувань).

Тобто, $X = (X_1, \ldots, X_{1000})$, де X_j відповідатиме за результат підкидання в j-му експерименті. Нехай $\mathbf{P}_\theta\,(X_1 = 1) = \theta$, де $X_1 = 1$ якщо монета впала стороною А, та $\mathbf{P}_\theta\,(X_1 = 0) = 1 - \theta$ (надайте інтерпретацію для $X_1 = 0$). Тут можна розглянути $\Theta = [0, 1]$. Статистичний простір можна визначити так: $(\Omega, \mathcal{F}, \mathbb{P}_\theta)$, де

- $\Omega = \{0, 1\}^{1000}$, **(чому?)**
- $\Sigma = 2^\Omega$,
- $\mathbf{P}_\theta(\{\omega\}) = \theta^{k(\omega)}(1 - \theta)^{n-k(\omega)}$, де $k(\omega) = \sum_{j=1}^{n} \mathbf{1}\{\omega_j = 1\}$. **(чому? подумати!)**

Ясно, що в такій постановці $X(\omega) = \omega$ **(чому?)**. В якості оцінки невідомої ймовірності успіху можна розглянути таку статистику:

$$\hat{\theta} = \frac{1}{n} \sum_{j=1}^{n} X_j,$$

що є вибірковим середнім.

2.2 Задача 2

Розглядається кратна вибірка $X = (X_1, \ldots, X_n)$, розподіл спостережень якої є рівномірним на відрізку $[a, b]$, тобто $X_1 \sim U[a, b]$, $a < b$. Значення кінців відрізка, a та b, вважаються невідомими. Перевірити, для яких параметрів наступні оцінки

$$\hat{\theta}_n^{(1)} = \max_{1 \le j \le n} X_j,\ \hat{\theta}_n^{(2)} = \frac{1}{n} \sum_{j=1}^{n} X_j$$

будуть (асимптотично) незміщеними, консистентними. Чи будуть задані оцінки асимптотично нормальними для відповідних параметрів?

Розв'язання

$\theta = (a, b)$. Позначимо

$$F(t) = \mathbf{P}_\theta\left(X_1 < t\right) = \begin{cases} 0, & t < a, \\ \frac{t-a}{b-a}, & a \le t < b, \\ 1, & b \le t \end{cases}$$
$$f(t) = F^{'}(t) = \mathbf{1}_{[a,b]}(t) \cdot \frac{1}{b-a}$$

Перевіримо властивості для $\hat{\theta}_n^{(1)}$. Спочатку знайдемо у явному вигляді розподіл цієї оцінки:

$$G(t) = \mathbf{P}_\theta\left(\hat{\theta}_n^{(1)} < t\right) = \mathbf{P}_\theta\left(\cap_{j=1}^{n}\{X_j < t\}\right) = \prod_{j=1}^{n} \mathbf{P}_\theta\left(X_j < t\right) = (F(t))^n,$$

звідси знаходимо щільність розподілу:

$$g(t) = G^{'}(t) = n(F(t))^{n-1} f(t).$$

Знайдемо математичне сподівання оцінки:

$$\mathbf{E}_\theta\left[\hat{\theta}_n^{(1)}\right] = \int\limits_{-\infty}^{+\infty} t g(t) dt = \int\limits_{-\infty}^{+\infty} t n (F(t))^{n-1} f(t) dt = \int\limits_{a}^{b} t n((t-a)/(b-a))^{n-1} 1/(b-a) dt =$$
$$= \frac{n}{(b-a)^n} \int\limits_{a}^{b} t(t-a)^{n-1} dt = \Big| u = t - a \Big| = \frac{n}{(b-a)^n} \int\limits_{0}^{b-a} (u+a) u^{n-1} du =$$
$$= \frac{n}{(b-a)^n} \left(\int\limits_{0}^{b-a} u^n du + \int\limits_{0}^{b-a} a u^{n-1} du \right) = \frac{n}{(b-a)^n} \left(\frac{(b-a)^{n+1}}{n+1} + \frac{a(b-a)^n}{n} \right) =$$
$$= \frac{n}{n+1}(b-a) + a \to b,\ n \to +\infty.$$

для всіх θ. Отже, $\hat{\theta}_n^{(1)}$ – асимптотично незміщена оцінка параметра b.

Працюємо далі $\hat{\theta}_n^{(1)}$ як з оцінкою b. Перевіримо консистентність оцінки, тому треба дослідити збіжність оцінки за ймовірністю для всіх можливих θ. Дійсно, якщо $\varepsilon \in (0, b-a)$, то

$$\begin{aligned}
\mathbf{P}_\theta\left(|\hat{\theta}_n^{(1)} - b| \geq \varepsilon\right) &= \mathbf{P}_\theta\left(\{\hat{\theta}_n^{(1)} \geq b+\varepsilon\} \cup \{\hat{\theta}_n^{(1)} \leq b-\varepsilon\}\right) = \\
&= \left|\{\hat{\theta}_n^{(1)} \geq b+\varepsilon\} \text{ та } \{\hat{\theta}_n^{(1)} \leq b-\varepsilon\} \text{ несумісні}\right| = \\
&= \mathbf{P}_\theta\left(\hat{\theta}_n^{(1)} \geq b+\varepsilon\right) + \mathbf{P}_\theta\left(\hat{\theta}_n^{(1)} \leq b-\varepsilon\right) = \\
&= \left|\mathbf{P}_\theta\left(\hat{\theta}_n^{(1)} \geq b+\varepsilon\right) = 0 \text{ оскільки } \hat{\theta}_n^{(1)}(\omega) \in [a,b]\right| = \\
&= \mathbf{P}_\theta\left(\hat{\theta}_n^{(1)} \leq b-\varepsilon\right) = G(b-\varepsilon) = \left(\frac{(b-\varepsilon)-a}{b-a}\right)^n = \left(1 - \frac{\varepsilon}{b-a}\right)^n \to 0,
\end{aligned}$$

оскільки вираз, що береться до степеня, за модулем менше одиниці. Випадок $\varepsilon \geq b-a$ очевидний, оскільки

$$\mathbf{P}_\theta\left(\hat{\theta}_n^{(1)} \leq b-\varepsilon\right) \leq \mathbf{P}_\theta\left(\hat{\theta}_n^{(1)} \leq b-(b-a)\right) = \mathbf{P}_\theta\left(\hat{\theta}_n^{(1)} \leq a\right) = 0.$$

Отже $\hat{\theta}_n^{(1)} \to^{P_\theta} b$ при $n \to +\infty$ та всіх θ, тобто оцінка є консистентною для параметра b.

Асимптотична нормальність для $\hat{\theta}_n^{(1)}$ не справджується. Дійсно, нехай $\{c_n\}_{n\geq 1}$ – довільна числова послідовність. Оскільки $\hat{\theta}_n^{(1)}$ приймає значення лише на відрізку $[a,b]$, тому $c_n(\hat{\theta}_n^{(1)} - b)$ прийматиме значення на відрізку $[c_n(a-b), 0]$. Тобто розподіл нормованої величини $c_n(\hat{\theta}_n^{(1)} - b)$ завжди буде зосереджений на одній з піввісей $(-\infty, 0]$ або $[0, +\infty)$. Але нормальний розподіл зосереджений на усій дійсній прямій, тобто $\mathbb{R} = (-\infty, +\infty)$. Тому стверджувати, що гранична поведінка буде нормальною не можна.

Вибіркове середнє $\hat{\theta}_n^{(2)}$ буде незміщеною, консистентною та асимптотично нормальною оцінкою математичного сподівання (функції від невідомих параметрів): $\tau(\theta) = (a+b)/2$:

- Для доведення незміщеності скористайтеся властивостями математичного сподівання та умовою задачі,
- Для доведення консистентності можна скористатися критерієм Колмогорова про ПЗВЧ. Тоді вийде строга консистентність, з якої випливає звичайна консистентність. **Подивитися що це за критерій та чому зі строгої консистентності випливає звичайна.**
- Для доведення асимптотичної нормальності достатньо скористатися ЦГТ у класичній формі. **Подивитися що то за звір такий, ЦГТ.**

2.3 Задача 3

Нехай вивчається кратна вибірка $X = (X_1, \ldots, X_n)$ з рівномірним розподілом спостережень $X_1 \sim U[0, \theta]$, де $\theta > 0$ вважається невідомим параметром, який потрібно оцінити. Дослідник розглядає наступні оцінки:

$$\hat{\theta}_n^{(1)} = (n+1) \min_{1 \le j \le n} X_j,\ \hat{\theta}_n^{(2)} = (1 + 1/n) \max_{1 \le j \le n} X_j.$$

Довести, що вищенаведені оцінки є незміщеними для параметра θ.

Розв'язання

Для $\hat{\theta}_n^{(2)}$ можна скористатися результатом попередньої задачі. Тут

$$\mathbf{E}_\theta \left[\max_{1 \le j \le n} X_j \right] = \frac{n}{n+1} \cdot \theta \Rightarrow \mathbf{E}_\theta \left[\hat{\theta}_n^{(2)} \right] = \theta \ \ \textbf{(а ви бачите чому це так?)}$$

Для $\hat{\theta}_n^{(1)}$ робимо схожі кроки. Знаходимо розподіл мінімума:

$$H(t) = \mathbf{P}_\theta \left(\min_{1 \le j \le n} X_j < t \right) = 1 - \mathbf{P}_\theta \left(\min_{1 \le j \le n} X_j \ge t \right) = 1 - \mathbf{P}_\theta \left(\cap_{j=1}^n \{X_j \ge t\} \right) = 1 - (1 - \mathbf{P}_\theta (X_j < t))^n$$

Далі знаходимо $h(t) = H'(t)$, обчислюємо $\mathbf{E}_\theta \left[\min_{1 \le j \le n} X_j \right] = \int_{-\infty}^{+\infty} t h(t) dt$ та використовуємо це для того, щоб показати незміщеність $\hat{\theta}_n^{(1)}$ для θ.

2.4 Задача 4

Нехай $X = (X_1, \dots, X_n)$ – кратна вибірка з нормального розподілу, тобто $X_1 \sim N(\theta, \sigma^2)$. Припустимо, що потрібно оцінити невідомий параметр $\theta \in \mathbb{R}$, знаючи $\sigma > 0$. Пропонується використати таку оцінку:

$$\hat{\theta}_n = \frac{X_1 + X_n}{2}.$$

Чи буде $\hat{\theta}_n$ незміщеною? Консистентною? Асимптотично нормальною?

Чи зміниться ситуація, якщо $X_1 \sim Exp(1/\theta)$, де $\theta > 0$ потрібно оцінити?

Розв'язання

Оскільки $X_1 \sim N(\theta, \sigma^2)$, то $\mathbf{E}_\theta [X_1] = \theta$ **(перевірте самостійно!)**. Тому, з властивостей математичного сподівання, маємо:

$$\mathbf{E}_\theta \left[\hat{\theta}_n\right] = \frac{1}{2}(\mathbf{E}_\theta [X_1] + \mathbf{E}_\theta [X_n]) = \theta,\ \theta \in \mathbb{R}.$$

Оцінка $\hat{\theta}_n$ не буде консистентною. Дійсно, для цього потрібно побачити, що

$$\hat{\theta}_n \sim N(\theta, \sigma^2/2)$$

як сума незалежних нормальних випадкових величин. Це доводиться напряму (через згортку розподілів, **це ви маєте пам'ятати**), або ж методом характеристичних функцій (х.ф.) **(що це таке?)**, що власне й зробимо:

$$\begin{aligned}\mathbf{E}_\theta \left[\exp(i\lambda\hat{\theta}_n)\right] &= (\mathbf{E}_\theta [\exp(i(\lambda/2)X_1)])^2 = (\exp(i\theta(\lambda/2) - \sigma^2(\lambda/2)^2/2))^2 = \\ &= \exp(i\theta\lambda - (\sigma^2/2)\lambda^2/2),\ \lambda \in \mathbb{R}.\end{aligned}$$

Вище ми скористалися тим, що $\exp(i\theta(\lambda/2) - \sigma^2(\lambda/2)^2/2)$ задає х.ф. $N(\theta, \sigma^2)$, а останній вираз у ланцюжку рівностей є х.ф. $N(\theta, \sigma^2/2)$.

Тобто, для довільного $n \geq 1$, розподіл $\hat{\theta}_n$ не змінюється, тобто не залежить від n. Тому збіжності до θ при $n \to +\infty$ для $\hat{\theta}_n$ не може бути.

Далі, для асимптотичної нормальності зауважимо, що для всіх $n \geq 1$

$$\hat{\theta}_n - \theta \sim N(0, \sigma^2).$$

Отже, для асимптотичної нормальності в даному випадку достатньо покласти $c_n := 1$ для всіх $n \geq 1$.

Якщо $X_1 \sim Exp(1/\theta)$, оцінка $\hat{\theta}_n$ буде незміщеною для θ, не буде консистентною та не буде асимптотично нормальною для θ. **Розберіться чому це так!**