Достатні та повні статистики.
Зв'язок з оптимальними оцінками

Теоретичні відомості та приклади

Статистики та оцінки. Нагадування

Нехай X – вибірка, розподіл якої залежить від $\theta \in \Theta$.

Довільну вимірну функцію від вибірки $T(X)$ називають статистикою.

Статистика $T(X)(\omega) : \Omega \to C$, яка приймає значення в параметричній множині (тобто $C \subset \Theta$), називається оцінкою параметра θ. Те саме означення узагальнюється на функції від невідомого параметра $\tau(\theta)$.

Статистика може слугувати як опорним інструментом для побудови оцінок, або ж сама слугувати як оцінка, або ж нести в собі інформацію, що може не стосуватися безпосередньо невідомого параметра, необхідну для подальшого аналізу.

Статистика **не** має містити невідомий параметр θ (відповідно й оцінка). Бо як таку рахувати на практиці?

Достатні статистики

Нехай та надалі $X = (X_1, \ldots, X_n)$ – кратна вибірка, розподіл якої залежить від невідомого параметра $\theta \in \Theta \subset \mathbb{R}^d$ та $L(X, \theta)$ – емпірична функція вірогідності.

Ми хочемо інформацію про спостережувані дані подати в більш стиснутому вигляді, тобто використавши статистики меншої розмірності. Такі статистики будемо називати достатніми. Дамо означення достатньої статистики в термінах функції вірогідності вибірки.

Означення. Статистика $T(X)$ називається достатньою, якщо вибіркова функція вірогідності $L(X, \theta)$ допускає розклад вигляду

$$L(X, \theta) = g(T(X), \theta) f(X) j(\theta) \text{ м.н. } \forall \theta \in \Theta,$$

де g, f, j є вимірними невід'ємними функціями.

Достатні статистики завжди існують. Прикладом такої є $T(X) = X$, тобто вибірка. Дійсно, означення вище буде виконуватися при $g = L,\ f \equiv 1,\ j \equiv 1$.

Достатня статистика, яку можна виразити через будь-яку іншу достатню статистику, називають мінімальною достатньою.

Означення. Достатня статистика $T(X)$ називається мінімальною достатньою, якщо для довільної достатньої статистики $S(X)$ існує така вимірна функція f, що $T(X) = f(S(X))$ майже напевно.

Наведемо приклади побудови достатніх статистик.

Приклад 1. Нехай $X = (X_1, \ldots, X_n)$ – кратна вибірка, де спостерження мають розподіл Пуассона, тобто $X_j \sim Pois(\theta)$, $\theta > 0$. Для всіх $\theta > 0$ спостереження X_j приймає значення в $\mathbb{Z}_+$, тому покладемо $S = \mathbb{Z}_+^n$. Побудуємо на S функцію вірогідності (як будувати – див. записи 4-го практичного)

$$L(x, \theta) = \prod_{j=1}^{n} \mathbf{P}_\theta (X_j = x_j) = \prod_{j=1}^{n} \frac{\theta^{x_j}}{x_j!} e^{-\theta} = e^{-n\theta} \theta^{\sum_{j=1}^n x_j} \prod_{j=1}^{n} \frac{1}{x_j!},\ x = (x_1, \ldots, x_n) \in S$$

Відповідно емпірична функція вірогідності (підстановка спостережуваних значень) така:

$$L(X, \theta) = e^{-n\theta} \theta^{\sum_{j=1}^n X_j} \prod_{j=1}^{n} \frac{1}{X_j!}$$

Розглянемо таку статистику $T(X) = \sum_{j=1}^n X_j$ та перевіримо, що вона є достатньою. Дійсно, функцію вірогідності вибірки $L(X, \theta)$ можна розкласти на функції g, f, j, якщо розглянути

$$g(T(X), \theta) = \theta^{T(X)},\ f(X) = \prod_{j=1}^{n} \frac{1}{X_j!},\ j(\theta) = e^{-n\theta}.$$

Отже $L(X, \theta) = g(T(X), \theta) f(X) j(\theta)$, тому статистика $T(X)$ є достатньою. $\square$

Приклад 2. Нехай $X = (X_1, \ldots, X_n)$ – кратна вибірка, де спостереження мають нормальний розподіл, тобто $X_j \sim N(\mu, \theta)$, де $\mu \in \mathbb{R}$ вважається відомим, а $\theta > 0$ – ні. Оскільки для всіх $\theta > 0$, нормальний розподіл зосереджений на $\mathbb{R}$, отже вибірка X приймає значення в $S := \mathbb{R}^n$. Будуємо на S функцію вірогідності:

$$L(x, \theta) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\theta}} e^{-(x_j - \mu)^2/(2\theta)} = (2\pi\theta)^{-n/2} e^{-\sum_{j=1}^n (x_j - \mu)^2/(2\theta)},\ x = (x_1, \ldots, x_n) \in S$$

Відповідно емпірична функція вірогідності така:

$$L(X, \theta) = (2\pi\theta)^{-n/2} e^{-\sum_{j=1}^n (X_j - \mu)^2/(2\theta)}$$

Розглянемо статистику $T(X) = \sum_{j=1}^n (X_j - \mu)^2$ (зауважте, що вона **не** містить невідоме θ). Ця статистика є достатньою, оскільки $L(X, \theta)$ можна розкласти на три функції таким чином:

$$g(T(X), \theta) = e^{-T(X)/(2\theta)},\ f(X) = 1,\ j(\theta) = (2\pi\theta)^{-n/2}$$

Отже $L(X, \theta) = g(T(X), \theta) f(X) j(\theta)$. $\square$

Приклад 3. Нехай тепер спостереження мають розподіл $N(\mu, \sigma^2)$, де $\theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$ обидва невідомі. Аналогічно до попереднього, $S = \mathbb{R}^n$ та емпірична функція вірогідності має вигляд

$$L(X, \theta) = (2\pi\sigma^2)^{-n/2} e^{-\sum_{j=1}^n (X_j - \mu)^2/(2\sigma^2)} = (2\pi\sigma^2)^{-n/2} e^{-(\sum_{j=1}^n X_j^2 - 2\mu \sum_{j=1}^n X_j + n\mu^2)/(2\sigma^2)}$$

Статистика $T(X) = (\sum_{j=1}^n X_j, \sum_{j=1}^n X_j^2)$ є достатньою, бо $L(X, \theta)$ допускає розклад на функції $g(T(X), \theta) = \exp(-(\sum_{j=1}^n X_j^2 - 2\mu \sum_{j=1}^n X_j + n\mu^2)/(2\sigma^2))$, $f(X) = 1$ та $j(\theta) = (2\pi\sigma^2)^{-n/2}$. Звісно, що можна підібрати інший вигляд для $g(T(X), \theta)$ та $j(\theta)$. $\square$

Приклад 4. Нехай спостереження мають рівномірний розподіл на відрізку $[a, b]$, де кінці відрізка є невідомими, тобто $\theta = (a, b) \in \Theta = \{(x, y) \in \mathbb{R}^2 \mid x < y\}$. Зауважимо, що для всіх значень θ, спостереження X_j прийматиме значення в $\mathbb{R}$ (відрізок рухається якщо грубо), тому вибірка прийматиме значення в $S := \mathbb{R}^n$. Запишемо функцію вірогідності:

$$L(x, \theta) = \prod_{j=1}^{n} \frac{1}{(b-a)} \mathbf{1}_{(a,b)}(x_j) = \frac{1}{(b-a)^n} \prod_{j=1}^{n} \mathbf{1}_{(a,b)}(x_j) = \frac{1}{(b-a)^n} \mathbf{1}_{a < \min_{1 \le j \le n} x_j} \mathbf{1}_{\max_{1 \le j \le n} x_j < b},$$

де $x = (x_1, \ldots, x_n) \in S$. Зауважимо, що останнє спрощення добутку індикаторів має місце, бо цей добуток ненульовий лише тоді, коли всі x_j належать (a, b), значить найменше серед x_j має бути більше за a, а найбільше серед x_j менше за b.

Відповідно емпірична функція вірогідності така:

$$L(X, \theta) = \frac{1}{(b-a)^n} \mathbf{1}_{a < \min_{1 \le j \le n} X_j} \mathbf{1}_{\max_{1 \le j \le n} X_j < b}$$

Розглянемо статистику $T(X) = (\min_{1 \le j \le n} X_j, \max_{1 \le j \le n} X_j)$. Ця статистика є достатньою, бо $L(X, \theta)$ можна розкласти, наприклад, на $g(T(X), \theta) = L(X, \theta)$, $f(X) = 1$, $j(\theta) = 1$. $\square$

Для доведення того, чи є статистика достатньою, можна скористатися так званим критерієм достатності.

Теорема (Критерій достатності) Статистика $T(X)$ є достатньою тоді і тільки тоді, коли умовний розподіл вибірки відносно значень $T(X)$ не залежить від θ, тобто

$$\mathbf{P}_\theta \left(X \in A \mid T(X) = t \right) = l(A, t), \ \forall \theta \in \Theta.$$

Тобто статистика $T(X)$ достатньо вичерпує інформацію про невідомі параметри розподілу так, що вибірка X не надає більше інформації про θ.

Приклад (На умовний розподіл вибірки) Розглянемо кратну вибірку з розподілом спостержень $X_j \sim Pois(\theta)$, $\theta > 0$ (тобто приклад №1). Ми запропонували таку достатню статистику: $T(X) = \sum_{j=1}^{n} X_j$. Покажемо, що розподіл X відносно $T(X)$ не залежить від θ.

Спочатку доведемо, що розподіл $T(X)$ є Пуассонівським з параметром $n\theta$. Можна це показати через характеристичні функції: для всіх $\lambda \in \mathbb{R}$ (зрозумійте чому наступні кроки справедливі)

$$\mathbf{E}_\theta \left[e^{i\lambda T(X)} \right] = \prod_{j=1}^{n} \mathbf{E}_\theta \left[e^{i\lambda X_j} \right] = \left(\mathbf{E}_\theta \left[e^{i\lambda X_1} \right] \right)^n$$

Залишається знайти характеристичну функцію одного спостереження:

$$\mathbf{E}_\theta \left[e^{i\lambda X_1} \right] = e^{-\theta} \sum_{k=0}^{+\infty} e^{i\lambda k} \frac{\theta^k}{k!} = e^{-\theta} \sum_{k=0}^{+\infty} \frac{(e^{i\lambda}\theta)^k}{k!} = e^{-\theta} e^{e^{i\lambda}\theta} = e^{(e^{i\lambda}-1)\theta}$$

Отже $\mathbf{E}_\theta \left[e^{i\lambda T(X)} \right] = (e^{(e^{i\lambda}-1)\theta})^n = e^{(e^{i\lambda}-1)n\theta}$, що і доводить висунуте припущення про розподіл $T(X)$. Далі, знаходимо умовний розподіл X відносно значень $T(X)$: для всіх $x = (x_1, \ldots, x_n) \in S$ та $t \in \mathbb{Z}_+$

$$\mathbf{P}_\theta \left(X \in A \mid T(X) = t \right) = \frac{\mathbf{P}_\theta \left(X \in A, T(X) = t \right)}{\mathbf{P}_\theta \left(T(X) = t \right)}$$

Зі знаменником все очевидно, оскільки розподіл $T(X)$ відомий, тому $\mathbf{P}_\theta\left(T(X)=t\right)=\frac{(n\theta)^t}{t!}e^{-n\theta}$. Залишається знайти $\mathbf{P}_\theta\left(X\in A, T(X)=t\right)$. Зауважимо, що подія в імовірності значить, що вибірка приймає значення x, при цьому сума всіх спостережень дорівнює t. Якщо $\sum_{j=1}^n x_j\neq t$, то ця ймовірність дорівнює 0. Інакше, імовірність дорівнює $L(x,\theta)$ (пригадайте як вона виглядає та чому це має місце). Тому

$$\mathbf{P}_\theta\left(X\in A, T(X)=t\right)=\mathbf{1}_{\sum_{j=1}^n x_j=t}L(x,\theta)=\mathbf{1}_{\sum_{j=1}^n x_j=t}\prod_{j=1}^n (x_j!)^{-1}\theta^{\sum_{j=1}^n x_j}e^{-n\theta}=$$
$$=\mathbf{1}_{\sum_{j=1}^n x_j=t}\prod_{j=1}^n (x_j!)^{-1}\theta^t e^{-n\theta}$$

Поєднучи результати для чисельника та знаменника маємо, що

$$\mathbf{P}_\theta\left(X\in A\mid T(X)=t\right)=\mathbf{1}_{\sum_{j=1}^n x_j=t}\prod_{j=1}^n (x_j!)^{-1}\theta^t e^{-n\theta}\left(\frac{(n\theta)^t}{t!}e^{-n\theta}\right)^{-1}=\frac{s!}{\prod_{j=1}^n x_j!\cdot n^t}$$

Тобто умовний розподіл X не залежить від θ, тобто критерій достатності виконується, що було б варто очікувати.

Зауваження. В іноземній літературі означення достатньої статистики та критерій достатності можуть формулюватися навпаки. Тобто, означення формулюється через умовний розподіл вибірки, а критерій достатності – через факторизацію емпіричної функції вірогідності. У постановці розкладу функції вірогідності, критерій ще називають факторизаційною теоремою Неймана-Фішера.

Повні статистики. Зв'язок з оптимальністю

Означення. Статистика $T(X)$ називається повною, якщо для кожної вимірної функції φ з тотожності $\mathbf{E}_\theta\left[\varphi(T(X))\right]=0$ для всіх $\theta\in\Theta$ випливає, що $\varphi(T(X))=0$ майже напевне.

Тепер питання: навіщо розглядати достатні та повні статистики? Принаймні для того, щоб побудувати оптимальні оцінки. Означення оптимальної оцінки див. у записах практичного заняття №4. Тут розглядаємо клас незміщених оцінок Γ_τ для функції від невідомого параметра $\tau(\theta)$.

Наступна теорема встановлює зв'язок між оптимальною оцінкою та достатньою статистикою.

Теорема (Рао-Блекуелла-Колмогорова) Якщо існує оптимальна оцінка $T^*\in\Gamma_\tau$, а $T=T(X)$ – деяка достатня статистика, тоді T^* є вимірною функцією від цієї статистики, тобто $T^*=H^*(T)$ майже напевно.

А що вийде, якщо статистика буде одночасно достатньою та повною? Відповідь на це дає

Теорема (Про оптимальність повної достатньої статистики) Якщо існує повна достатня статистика, то довільна функція від неї є оптимальною оцінкою свого математичного сподівання.

Остання теорема дає натяк на те, як можна будувати оптимальні оцінки. Це може допомогти в тому випадку, коли ефективну оцінку побудувати не можна.

Наведемо приклади побудови таких оцінок, використовуючи повні достатні статистики.

Приклад (Про зв'язок між оптимальними оцінками та достатніми статистиками) Покажемо, що $S(X) = \sum_{k=1}^{m} c_k \pi_k(T) n^{-k}$, де $T = \sum_{j=1}^{n} X_j$ та $\pi_k(t) = t(t-1)\dots(t-k+1)$, є оптимальною оцінкою для $P(\theta) = \sum_{k=1}^{m} c_k \theta^k$. Оскільки $T \sim Pois(n\theta)$, то

$$\mathbf{E}_\theta\left[\pi_k(T)\right] = \sum_{t=k-1}^{+\infty} t(t-1)\dots(t-k+1)\frac{(n\theta)^t}{t!}e^{-n\theta} = e^{-n\theta}\sum_{t=k-1}^{+\infty}(t-k+1)\frac{(n\theta)^t}{(t-k+1)!} =$$

$$= |s = t-(k-1)| = e^{-n\theta}\sum_{s=0}^{+\infty} s\frac{(n\theta)^{s+(k-1)}}{s!} = (n\theta)^{k-1}(n\theta) = (n\theta)^k,\ k \geq 1$$

Отже $\mathbf{E}_\theta\left[S(X)\right] = \sum_{k=1}^{m} c_k n^{-k}\mathbf{E}_\theta\left[\pi_k(T)\right] = P(\theta)$. Виходить, що оцінка $S(X)$ незміщена для $P(\theta)$. Але ж T – достатня, тому за теоремою Рао-Блекуелла-Колмогорова оцінка $S(X)$ є оптимальною для $P(\theta)$.

Приклад (Про повноту статистики та оптимальність)

Продовжуємо розглядати кратну вибірку, де спостереження мають розподіл Пуассона.

Попередньо ми доводили, що статистика $T(X) = \sum_{j=1}^{n} X_j$ є достатньою. Доведемо, що вона є повною. Нехай φ – довільна вимірна функція, причому

$$\mathbf{E}_\theta\left[\varphi(T(X))\right] = \sum_{k=0}^{+\infty}\varphi(k)\frac{(n\theta)^k}{k!}e^{-n\theta} = 0,\ \forall\theta > 0 \Leftrightarrow \sum_{k=0}^{+\infty}\varphi(k)\frac{(n\theta)^k}{k!} = 0,\ \forall\theta > 0$$

Зверху можна побачити степеневий ряд:

$$f(\theta) = \sum_{k=0}^{+\infty}\varphi(k)\frac{(n\theta)^k}{k!} = \sum_{k=0}^{+\infty}\psi_k\theta^k,\ \psi_k = \varphi(k)\cdot\frac{n^k}{k!},\ k \geq 0$$

З припущення відомо, що $f(\theta) = 0$ для всіх θ. Отже, для всіх $n \geq 0$: $f^{(n)}(\theta) = 0$, $\theta > 0$. З іншого боку, відомо, що коефіцієнти ψ_k степеневого ряду можна обчислити через похідні f

$$\psi_k = \frac{f^{(k)}(0)}{k!},\ k \geq 0,$$

де $f^{(n)}(0) := 0$ – довизначили за неперервністю. Тому $\psi_k = 0$ для всіх $k \geq 0$, тобто

$$\psi_k = \varphi(k)\cdot\frac{n^k}{k!} = 0 \Leftrightarrow \varphi(k) = 0,\ \forall k \geq 0$$

Тобто φ дорівнює нулю на цілих невід'ємних числах. Але $T(X) \in \mathbb{Z}_+$, отже $\varphi(T(X)) = 0$ майже напевне.

Таким чином показали повноту статистики $T(X)$.

Оскільки $T(X)$ – повня достатня, то для довільної вимірної функції f маємо, що $f(T(X))$ – оптимальна оцінка $\tau(\theta) = \mathbf{E}_\theta\left[f(T(X))\right]$. Побудуємо оптимальну оцінку для θ: для цього розглянемо $f(x) = x/n$. Тоді $f(T(X)) = \sum_{j=1}^{n} X_j/n$ – оптимальна оцінка параметра $\mathbf{E}_\theta\left[f(T(X))\right] = \theta$.

Приклад (Нерегулярна модель) Розглянемо кратну вибірку з розподілом спостережень $X_j \sim U[0, \theta]$, де $\theta > 0$ вважається невідомим. Оскільки для довільного $\theta > 0$, $X_j > 0$, то $S := (0, \infty)^n$. Покажемо, що умови регулярності порушуються.

Запишемо функцію вірогідності:

$$L(x, \theta) = \prod_{j=1}^{n} \frac{1}{\theta} \mathbf{1}_{(0,\theta)}(x_j) = \frac{1}{\theta^n} \mathbf{1}_{\max_{1 \le j \le n} x_j < \theta},\ x = (x_1, \ldots, x_n) \in S$$

Якщо зафіксувати числа x_j та роглядати $L(x, \theta)$ як функцію по θ, то можна побачити, що $L(x, \theta) = 0$ при $\theta < \max_{1 \le j \le n} x_j$, та $L(x, \theta) = 1$ інакше. Значить, строга додатність $L(x, \theta)$ залежить від значень θ, що порушує першу умову регулярності. Отже застосувати теорему Крамера-Рао не можна, відповідно ефективної оцінки не існує. З іншого боку, можна спробувати знайти оптимальну.

Емпірична функція вірогідності така:

$$L(X, \theta) = \frac{1}{\theta^n} \mathbf{1}_{\max_{1 \le j \le n} X_j < \theta},$$

з якої легко бачити, що статистика $T(X) = \max_{1 \le j \le n} X_j$ є достатньою. Зокрема, ця статистика повною. Дійсно, якщо взяти довільну вимірну функцію φ, для якої $\mathbf{E}_\theta \left[\varphi(T(X))\right] = 0$ для всіх $\theta > 0$, то

$$\mathbf{E}_\theta \left[\varphi(T(X))\right] = \int_0^\theta \varphi(t) \left(\frac{t}{\theta}\right)^{n-1} \frac{n}{\theta} dt = 0,\ \theta > 0,$$

де пригадали щільність розподілу максимума з попередніх занять (справедливо згадайте як виводити, якщо забули). Далі маємо, що

$$F(t) = \int_0^\theta \varphi(t) t^{n-1} dt = 0,\ \theta > 0$$

Зауважимо, що $F(t)$ неперервна на $(0, \infty)$. Значить, існує майже скрізь на $(0, \infty)$ похідна $F'(t) = \varphi(t) t^{n-1}$. Але $0 = F'(\theta) = \varphi(\theta)\theta^{n-1}$, звідки маємо що $\varphi(\theta) = 0$ майже скрізь на $(0, \infty)$. Оскільки $T(X) \in (0, \infty)$, то $\varphi(T(X)) = 0$ майже напевно, що і доводить повноту.

В результаті маємо, що $T(X)$ – повна достатня статистика. Тепер побудуємо оптимальну оцінку для θ. Відомо, що оцінка $S(X) = (1 + 1/n) T(X)$ є незміщеною для θ. Але ж $S(X)$ можна розглядати як вимірну функцію від $T(X)$: $S(X) = f(T(X))$, де $f(x) = (1 + 1/n)x$. Внаслідок властивостей статистики $T(X)$, оцінка $S(X)$ є оптимальною для невідомого параметра θ.