Оцінювання в схемі випробувань Бернуллі. Емпірична функція розподілу, варіаційний ряд вибірки

1 Теоретичні відомості

Рекламна пауза з теорії ймовірностей

Розглянемо випадкову величину ξ з функцією розподілу $F(x) = \mathbf{P}(\xi < x)$.

Квантилем $Q^{\xi}(p)$ розподілу випадкової величини ξ рівня $p \in (0,1)$ називають таке найменше число $x \in \mathbb{R}$, що $F(x) = p$. Точніше, $Q^{\xi}(p) = \inf\{x \in \mathbb{R} \mid F(x) \geq p\}$. Якщо існує функція $F^{-1}(p)$, обернена до $F(x)$, тоді $Q^{\xi}(p) = F^{-1}(p)$.

Теоретичною медіаною називають квантиль рівня $1/2$, тобто $Q^{\xi}(1/2)$.

Приклад. Нехай $\xi \sim Exp(\lambda)$, $\lambda > 0$. Функція розподілу ξ така:

$$F(x) = \mathbf{1}\{x > 0\}(1 - \exp(-\lambda x)),\ x \in \mathbb{R}.$$

Нехай $x > 0$ та $p \in (0,1)$ знайдемо таке x, щоб

$$p = F(x) = 1 - \exp(-\lambda x) \Rightarrow x = -\frac{\ln(1-p)}{\lambda} =: Q^{\xi}(p).$$

Таким чином знайшли квантильну функцію розподілу ξ на $(0,1)$.

Оцінювання в схемі випробувань Бернуллі

Розглянемо схему з n незалежних випробувань Бернуллі. Результати випробувань можна подати у вигляді вектора $X = (X_1, \ldots, X_n)$, де координати є незалежними в сукупностями та однаково розподіленими випадковими величинами, а $X_1 \sim \mathrm{Bern}(\theta)$, тобто

$$\mathbf{P}_{\theta}(X_1 = 1) = \theta,\ \mathbf{P}_{\theta}(X_1 = 0) = 1 - \mathbf{P}_{\theta}(X_1 = 1) = 1 - \theta$$

Тут θ – невідома ймовірність успіху в одному випробуванні. Вважаємо, що $\theta \in \Theta = [0,1]$.

На основі спостережень X потрібно побудувати оцінку для θ. Для цього можна розглянути відносну емпіричну частоту (вибіркову частку) $\hat{\theta}_n$ вигляду

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

Вибіркова частка є незміщеною, строго консистентною, аисмптотично нормальною оцінкою імовірності успіху θ. Зокрема, оскільки

$$\hat{\theta}_n \to^{P_{\theta}} \theta \text{ та } \sqrt{n} \cdot \frac{\hat{\theta}_n - \theta}{\sqrt{\theta(1-\theta)}} \to^{W_{\theta}} \xi \sim N(0,1),$$

то

$$\sqrt{n}\cdot\frac{\hat{\theta}_n-\theta}{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}}=\sqrt{n}\cdot\frac{\hat{\theta}_n-\theta}{\sqrt{\theta(1-\theta)}}\cdot\frac{\sqrt{\theta(1-\theta)}}{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}}\to^{W_\theta}\xi\sim N(0,1). \tag{1}$$

Остання збіжність (1) має місце внаслідок теореми Слуцького.

Теорема (Слуцького). Нехай (ξ_n) та (η_n) – такі дві послідовності випадкових чисел, що $\xi_n \to^W \xi$ та $\eta_n \to^P c$, де c – невипадкова стала. Тоді

$$\xi_n+\eta_n\to^W \xi+c \text{ та } \xi_n\cdot\eta_n\to^W \xi\cdot c,\ n\to+\infty.$$

Асимптотична нормальність вибіркової частки $\hat{\theta}_n$ дозволяє будувати *довірчі інтервали* для невідомої ймовірності успіху θ. Дамо означення того, що таке довірчий інтервал для скалярного параметра $\theta\in\Theta\subset\mathbb{R}$.

Означення. Довірчим інтервалом рівня довіри $p\in(0,1)$ параметра θ називається така пара статистик $(\tilde{\theta}^-,\tilde{\theta}^+)$, що $\tilde{\theta}^-<\tilde{\theta}^+$ та

$$\mathbf{P}_\theta\left(\tilde{\theta}^-\le\theta\le\tilde{\theta}^+\right)=p,\ \theta\in\Theta.$$

Надалі введемо $x_p=Q^{N(0,1)}((1+p)/2)$ – число, що задовольняє рівність $\mathbf{P}\left(|\xi|<x_p\right)=p$ для $\xi\sim N(0,1)$ (доведіть). Зауважимо, що $x_p\ge 0$, для всіх $p\in(0,1)$.

Розглянемо два підходи побудови інтервалів для ймовірності успіху, що базуються на асимптотичній нормальності вибіркової частоти. Для достатньо великих n, подія

$$\left\{\sqrt{n}\cdot\frac{|\hat{\theta}_n-\theta|}{\sqrt{\theta(1-\theta)}}\le x_p\right\} \tag{2}$$

наближено має ймовірність p. Для побудови інтервалу скористаємося нерівністю в (2).

1. **Метод Вілсона.** Нерівність (2) підносимо до квадрату. Тоді відносно θ розв'язуємо нерівність

   $$n\cdot\frac{(\hat{\theta}_n-\theta)^2}{\theta(1-\theta)}\le x_p^2,$$

   звідки знаходимо кінці наближеного інтервалу:

   $$\frac{(n\hat{\theta}_n+x_p^2/2)\pm x_p\sqrt{n\hat{\theta}_n(1-\hat{\theta}_n)+x_p^2/4}}{n+x_p^2}$$

2. **Метод Вальда.** В нерівності (2) розкриваємо модуль, множимо нерівності на $\sqrt{\theta(1-\theta)/n}$ та позбуваємося від $\hat{\theta}_n$ між нерівностями. Тоді вийде наближений *симетричний* довірчий інтервал з кінцями

   $$\hat{\theta}_n\pm x_p\sqrt{\frac{\theta(1-\theta)}{n}} \tag{3}$$

   Проте на практиці значення θ є невідомим. Тому для використання (3) можна замість θ використати вибіркову частоту $\hat{\theta}_n$, або скористатися наявною до експерименту інформацією (наприклад, результати обстеження, що проводилися до поточного / значення оцінок з пов'язаних обстежень).

Метод Вілсона хороший тим, що ми будуємо інтервал 'напряму', тобто нам не потрібно конкретні величини ще якось дооцінювати, уникаючи додаткового збурення. Зокрема інтервал не вийде за межі $[0, 1]$

Емпірична функція розподілу, варіаційний ряд вибірки

Розглянемо кратну вибірку $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ з невідомою функцією розподілу $F(x) = \mathbf{P}(X_1 < x)$ спостережень. Емпіричною функцією розподілу називається оцінка

$$\hat{F}_n(x) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}\{X_j < x\},\ x \in \mathbb{R}.$$

Поточково $\hat{F}_n(x)$ є незміщеною, строго консистентною, асимптотично нормальною оцінкою $F(x)$ (аналогічні кроки доведення як і властивостей вибіркової частки). Зокрема має місце рівномірна збіжність до невідомої функції розподілу даних (теорема Глівенко-Кантеллі):

$$\sup_{x \in \mathbb{R}} |F(x) - \hat{F}_n(x)| \to^{P1} 0,\ n \to \infty.$$

Якщо зафіксувати деяку реалізацію вибірки $X(\omega) = (x_1, \ldots, x_n)$, то на $\hat{F}_n(x)$ можна дивитися як на функцію дискретного рівномірного розподілу в точках $\{x_1, \ldots, x_n\}$.

Впорядковану вибірку за зростанням $\mathrm{sort}(X) = (X_{(1)}, \ldots, X_{(n)})$ називають варіаційним рядом. Елементи варіаційного ряду називають порядковими статистиками. Зокрема, k-ий елемент варіаційного ряду $X_{(k)}$ називають k-ою порядковою статистикою. Для неперервного розподілу спостережень вибірки,

$$X_{(1)} < X_{(2)} < \ldots < X_{(n-1)} < X_{(n)} \text{ м.н.}$$

В термінах $X_{(k)}$, $\hat{F}_n(x) = \sum_{j=1}^{n} \frac{j}{n} \cdot \mathbf{1}\{X_{(j-1)} < x \le X_{(j)}\}$, де $X_{(0)} := -\infty$ (та відповідна нерівність в індикаторі стає строгою).

Наведемо приклади статистик, що використовують елементи варіаційного ряду.

Вибірковою медіаною називають середину варіаційного ряду, тобто

$$\mathrm{Median}(X) = \begin{cases} X_{((n+1)/2)}, & n - \text{непарне}, \\ (X_{(n/2)} + X_{(n/2+1)})/2, & n - \text{парне}. \end{cases}$$

Вибіркова медіана є статистикою середнього положення, зокрема береться в якості оцінки теоретичної медіани розподілу спостережень.

Нижнім $Q_1(X)$ та верхнім $Q_3(X)$ квартилем варіаційного ряду називають медіани підвибірок, утворених поділом початкової вибірки медіаною, тобто:

$$Q_1(X) = \mathrm{Median}(\{X_j \in X \mid X_j \le \mathrm{Median}(X)\}),$$
$$Q_3(X) = \mathrm{Median}(\{X_j \in X \mid X_j \ge \mathrm{Median}(X)\}).$$

Також медіану $\mathrm{Median}(X) = Q_2(X)$ називають другим квартилем. Квартилі розбивають вибірку на чотири частини приблизно однакового розміру.

Вибіркові квартилі можна брати в якості оцінок для теоретичних квартилів розподілу спостережень.

Інтерквартильний розмах
$$\mathrm{IQR}(X) = Q_3(X) - Q_1(X)$$
визначає ширину інтервалу, в якому міститься приблизно половина спостережень з вибірки. Розмахом вибірки визначають відстань між найбільшим та найменшими спостереженнями у вибірці:
$$\mathrm{Range}(X) = X_{(n)} - X_{(1)}.$$
$\mathrm{IQR}(X)$ та $\mathrm{Range}(X)$ є статистиками характеризації розкиду спостережень.

2 Задачі

2.1 Задача 1

В огляді, що виконаний поштовою компанією, з 200 клієнтів 172 вказали на задоволення часом доставки кореспонденції. Обчислити наближений 95 %-й довірчий інтервал для теоретичної частоти задоволених клієнтів.

Розв'язання.

Давайте зорієнтуємося в тому, що треба оцінити. Генеральною сукупністю виступають усі користувачі послуг поштової компанії, а невідомим параметром є так зване загальну частку задоволених від користування цими послугами (позначимо через $\theta \in (0, 1)$). Досліджувати всю популяцію може бути витратним в сенсі ресурсів та часу. Тому компанія проводить відбір з генеральної сукупності $n = 200$ та проводить обстеження, таким чином формуючи вибірку з відповідей опитаних. З n відібраних клієнтів, кількість задоволених послугами становить $m = 172$.

Оцінити частку генеральної сукупності θ за вибіркою можна використовуючи вибіркову частку. Обчислимо її значення:
$$\hat{\theta}_n = \frac{m}{n} = \frac{172}{200} = 0.86$$
Тут варто розуміти, що ми працюємо з *реалізацією* вибірки, тому тут $\hat{\theta}_n$ – *значення* вибіркової частки (тобто сприймати як число а не випадкову величину).

Побудуємо наближений симетричний довірчий інтервал з рівнем довіри $p = 0.95$. Квантиль в цьому випадку дорівнює $x_p = Q^{N(0,1)}((1 + p)/2) \approx 1.96$. Маючи все необхідне для підрахунку, маємо (врахувавши округлення)

$$\hat{\theta}_n - x_p\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} \leq \theta < \hat{\theta}_n + x_p\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} \Leftrightarrow 0.8119 \leq \theta < 0.9081$$

2.2 Задача 2

Компанія хоче оцінити відсоток від її клієнтів, які бажають робити покупки у Інтернеті. Для цього вирішує обчислити симетричний двосторонній 95%-й довірчий інтервал для невідомого відсотка.

1. Показати, що на підставі випадкової вибірки з 200 клієнтів, необхідний довірчий інтервал буде мати ширину, що не перевищує 13.8%.
2. Обчислити розмір вибірки, який буде гарантувати, що безвідносно теоретичного відсотку, ширина вірогідного інтервалу не перевищить 10%.

Розв'язання.

Щоб розв'язати перший пункт, потрібно спочатку знайти ширину наближеного симетричного інтервалу та зробити оцінку зверху. Легко бачити, що ширина інтервалу має вигляд

$$\left(\hat{\theta}_n + x_p\sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}}\right) - \left(\hat{\theta}_n - x_p\sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}}\right) = 2x_p\sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}}$$

Далі зауважимо, що $f(x) = x(1-x) \leq 1/4$ для всіх $x \in [0,1]$, причому рівність досягається при $x = 1/2$ (розберіться чому це так). Тому

$$2x_p\sqrt{\frac{\hat{\theta}_n(1-\hat{\theta}_n)}{n}} \leq \frac{x_p}{\sqrt{n}} \tag{4}$$

Таким чином маємо верхню оцінку на ширину довірчого інтервалу. Доведемо перший пункт, підставивши відомі значення:

$$\text{Ширина д.і.} \leq \frac{x_p}{\sqrt{n}} \approx \frac{1.96}{\sqrt{200}} \approx 0.138,$$

що і завершує першу частину розв'язку.

Щоб розв'язати другий пункт задачі, скористаємося (4), розв'язавши нерівність відносно n:

$$\frac{1.96}{\sqrt{n}} \leq 0.1 \Leftrightarrow n \geq (1.96/0.1)^2 \approx 384.16$$

Тобто, для того, щоб ширина довірчого інтервалу не перевищувала задане значення, потрібно мати вибірку обсягу принаймні з $n \geq 385$ спостережень.

2.3 Задача 3

Спостерігаються дві незалежні послідовності випробувань Бернуллі з ймовірностями успіхів θ_i, та з n_i випробуваннями, $i = 1, 2$. Нехай $\hat{\theta}_{n_i}$ – відповідні відносні частоти успіхів. Довести, що

$$\sqrt{n_1}((\hat{\theta}_{n_1} - \hat{\theta}_{n_2}) - (\theta_1 - \theta_2)) \to^W \xi \sim N(0, \theta_1(1-\theta_1) + \rho\theta_2(1-\theta_2))$$

при $n_i \to +\infty$ так, що $n_1/n_2 \to \rho$. Побудувати довірчий інтервал для $\theta_1 - \theta_2$.

Розв'язання.

Перезапишемо

$$\sqrt{n_1}((\hat{\theta}_{n_1} - \hat{\theta}_{n_2}) - (\theta_1 - \theta_2)) = \sqrt{n_1}(\hat{\theta}_{n_1} - \theta_1) - \sqrt{n_1}(\hat{\theta}_{n_2} - \theta_2)$$

Оскільки $\hat{\theta}_{n_i}$ є асимптотично нормальними оцінками ймовірностей успіху θ_{n_i} у відповідних випробувань Бернуллі, маємо

$$\sqrt{n_i}(\hat{\theta}_{n_i} - \theta_i) \to^W \xi_i \sim N(0, \theta_i(1-\theta_i)),\ n_i \to \infty.$$

За теоремою Слуцького, має місце слабка збіжність для $\sqrt{n_1}(\hat{\theta}_{n_2} - \theta_2)$

$$\sqrt{n_1}(\hat{\theta}_{n_2} - \theta_2) = \sqrt{\frac{n_1}{n_2}} \cdot (\sqrt{n_2}(\hat{\theta}_{n_2} - \theta_2)) \to^W \sqrt{\rho} \cdot \xi_2 \sim N(0, \rho \cdot \theta_2(1-\theta_2)),\ n_i \to \infty.$$

Ясно, що $-\sqrt{n_1}(\hat{\theta}_{n_2} - \theta_2) \to^W \sqrt{\rho} \cdot \xi_2$ внаслідок симетричності стандартного нормального розподілу. Залишається перевірити, чи справді

$$\sqrt{n_1}(\hat{\theta}_{n_1} - \theta_1) - \sqrt{n_1}(\hat{\theta}_{n_2} - \theta_2) \to^W N(0, \theta_1(1-\theta_1) + \rho \cdot \theta_2(1-\theta_2))\ n_i \to \infty. \tag{5}$$

Оскільки $\hat{\theta}_{n_1}$ та $\hat{\theta}_{n_2}$ побудовані на основі двох незалежних вибірок (векторів), то ці оцінки є незалежними (так само як і перетворення від кожної). Врахуємо це у дослідженні характеристичної функції суми: для всіх $\lambda \in \mathbb{R}$

$$\begin{aligned}
&\mathbf{E}\left[\exp\left(i\lambda(\sqrt{n_1}(\hat{\theta}_{n_1} - \theta_1) - \sqrt{n_1}(\hat{\theta}_{n_2} - \theta_2))\right)\right] = \\
&= \mathbf{E}\left[\exp\left(i\lambda\sqrt{n_1}(\hat{\theta}_{n_1} - \theta_1)\right)\right] \mathbf{E}\left[\exp\left(i\lambda(-1)\sqrt{n_1}(\hat{\theta}_{n_2} - \theta_2)\right)\right] \to \\
&\to \mathbf{E}\left[\exp\left(i\lambda\xi_1\right)\right] \mathbf{E}\left[\exp\left(i\lambda\sqrt{\rho}\xi_2\right)\right] = \exp\left(-\frac{\theta_1(1-\theta_1)}{2}\lambda^2\right) \exp\left(-\frac{\rho\theta_2(1-\theta_2)}{2}\lambda^2\right) = \\
&= \exp\left(-\left(\frac{\theta_1(1-\theta_1) + \rho\theta_2(1-\theta_2)}{2}\right)\lambda^2\right)
\end{aligned}$$

У правій частині збіжності маємо характеристичну функцію нормального розподілу з нульовим середнім та дисперсією $\theta_1(1-\theta_1) + \rho \cdot \theta_2(1-\theta_2)$. Тому, за теоремою Леві про неперервність та однозначності визначення функції розподілу через характеристичну функцію доводимо збіжність (5).

2.4 Задача 4

Довести, що для вибірки $X = (X_1, \ldots, X_n)$ з неперервною функцією розподілу статистика Колмогорова $\hat{\kappa}_n = \sqrt{n} \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)|$ дорівнює

$$\hat{\kappa}_n = \sqrt{n} \max_{1 \le k \le n} \max\{|k/n - F(X_{(k)})|, |F(X_{(k)}) - (k-1)/n|\}.$$

Розв'язання.

Достатньо показати, що

$$\Delta_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| = \max_{1 \le k \le n} \max\{|k/n - F(X_{(k)})|, |F(X_{(k)}) - (k-1)/n|\}.$$

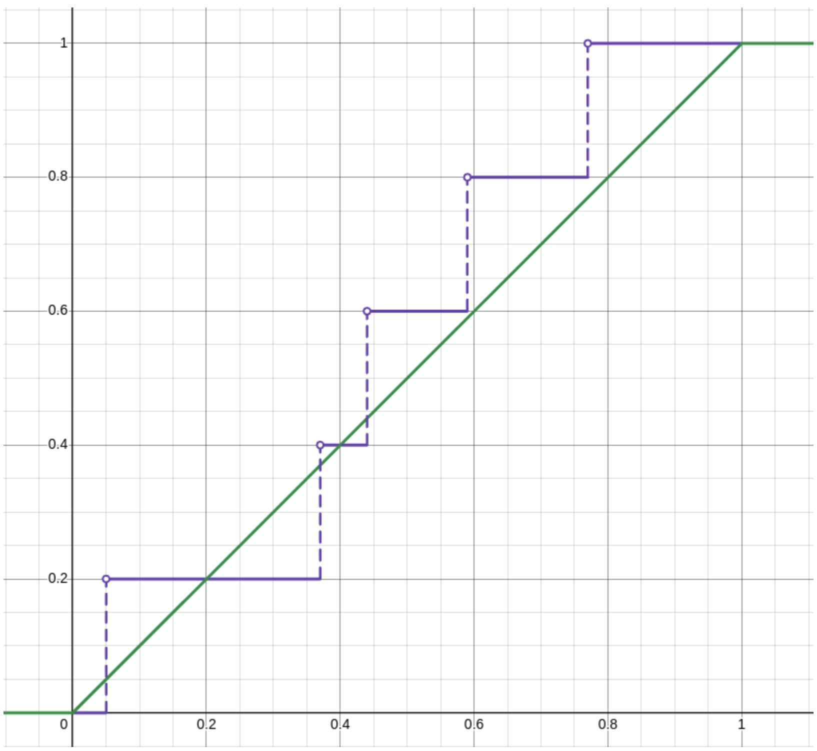

Рис. 1: Графіки теоретичної функції розподілу $U[0,1]$ (зелений) та емпіричної функції розподілу (фіолетовий) за вибіркою $X = (0.44, 0.59, 0.37, 0.77, 0.05)$ (перевірте самостійно).

Потрібно зауважити, що найбільші відхилення $|F(x) - \hat{F}_n(x)|$ спостерігаються у точках, де $\hat{F}_n(x)$ робить стрибок. У кожній точці порівнюємо відстані зліва та справа:

$$\delta_k^- = |F(X_{(k)}) - \hat{F}_n(X_{(k)}-)| = |F(X_{(k)}) - (k-1)/n|,$$
$$\delta_k^+ = |\hat{F}_n(X_{(k)}+) - F(X_{(k)})| = |\hat{F}_n(X_{(k)}) - F(X_{(k)})| = |k/n - F(X_{(k)})|.$$

Звідки обираємо найбільше відхилення серед усіх точок стрибку $\hat{F}_n(x)$.

2.5 Задача 5

За вибіркою з неперервного розподілу:

$$2.4, 1.0, 0.7, 0.2, 1.1, 1.6, 1.1, -0.4, 0.1, 0.7$$

записати варіаційний ряд, знайти вибіркову медіану, нижній та верхній квартилі, розмах вибірки. Побудувати графік емпіричної функції розподілу $\hat{F}_n(x)$.

Розв'язання.

Працюємо з реалізацією вибірки з деякого розподілу.

Впорядкуємо значення вибірки за зростанням, таким чином отримавши варіаційний ряд:

$$-0.4, 0.1, 0.2, 0.7, 0.7, 1.0, 1.1, 1.1, 1.6, 2.4$$

Обсяг вибірки $n = 10$. Тому значення медіани – це середнє значення сусідніх елементів у варіаційному ряду по центру, тобто

$$\text{Median}(X) = (0.7 + 1)/2 = 0.85$$

Знайдемо нижні та верхні підвибірки, що менші або більші медіани:

$$X[X \leq \text{Median}(X)] = (0.7, 0.2, -0.4, 0.1, 0.7)$$
$$X[X \geq \text{Median}(X)] = (2.4, 1.0, 1.1, 1.6, 1.1)$$

Нижній та верхній квартилі відповіно дорівнюють

$$Q_1(X) = \text{Median}(X[X \leq \text{Median}(X)]) = 0.2,$$
$$Q_3(X) = \text{Median}(X[X \geq \text{Median}(X)]) = 1.1.$$

Розмах вибірки: $\text{Range}(X) = X_{(n)} - X_{(1)} = 2.4 - (-0.4) = 2.8$.

Знайдемо явно функцією розподілу:

$$\hat{F}_n(x) = \sum_{k=1}^{n} \frac{k}{n} \mathbf{1}\{X_{(k)} \leq x < X_{(k+1)}\} = \begin{cases} 0, & x \leq -0.4 \\ 1/10, & x \in (-0.4, 0.1] \\ 2/10, & x \in (0.1, 0.2] \\ 3/10, & x \in (0.2, 0.7] \\ 5/10, & x \in (0.7, 1] \\ 6/10, & x \in (1, 1.1] \\ 8/10, & x \in (1.1, 1.6] \\ 9/10, & x \in (1.6, 2.4] \\ 1, & x < 2.4 \end{cases}$$

Нижче покажемо графік емпіричної функції розподілу.

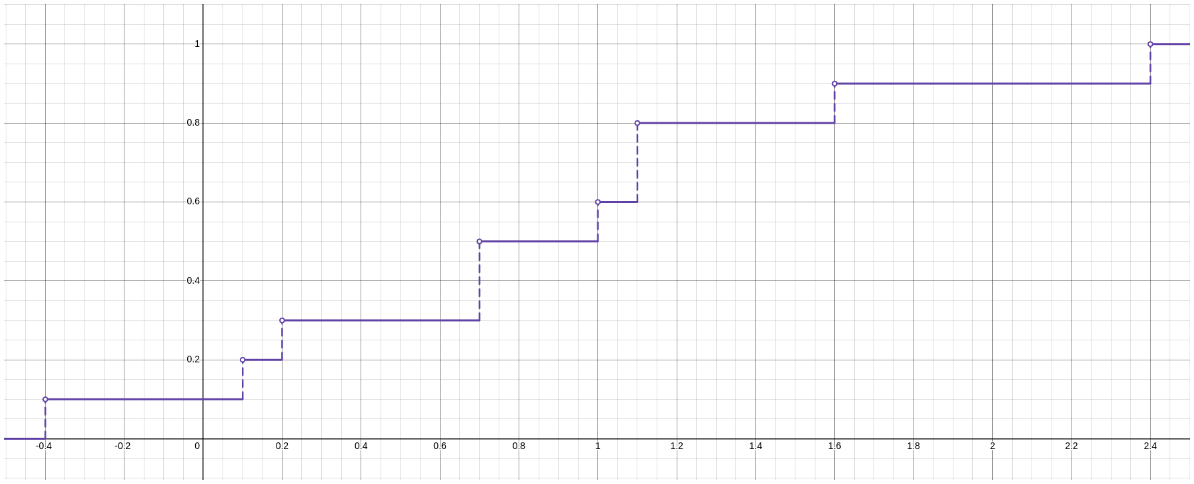

Рис. 2: Графіки темпіричної функції розподілу за спостережннями в задачі.