

Під час аналізу даних виділяються наступні етапи: отримання вхідної інформації, безпосередньо сама обробка її, аналіз та інтерпретація результатів обробки даних.

Головне зробити правильні висновки з результатів.

Значення змінних які спостерігаються можуть бути як *кількісні* так і *якісні*. Якісні змінні поділяють на *ординальні* та *номінальні*. Ординальні змінні називають *порядковими*, а номінальні – *класифікаційними*. Обидва типи змінних приймають свої значення з деякої множини, елементи якої називають *градаціями*. Градації, які приймає як свої значення ординальна змінна, природно **впорядковані за ступенем прояву властивості**. Градації номінальної змінної такого порядку **не мають**. Серед якісних змінних виділяють *категоризовані* та *не категоризовані*.

До категоризованих змінних відносять змінні, для яких повністю визначена множина градацій та правило віднесення значення змінної, яке спостерігається, до певної градації.

Змінні ще поділяють на *дискретні* та *неперервні*.

1 Групування даних

ξ – скалярна змінна, яка досліджується.

Вибірка об'єму n : x_1, x_2, \dots, x_n .

У випадку великих об'ємів вибірок виникає бажання провести деяке перетворення їх з метою стиснення даних без суттєвої втрати вибірками інформативності, а тільки згодом проводити обробку цих перетворених даних. Як правило, його застосовують при обробці спостережень над неперервними змінними, коли об'єм вибірки перевищує 50, а над дискретними змінними, коли кількість значень m , які вони приймають, перевищує 10.

Перехід до згрупованих даних:

1. Визначити $x_{\min} = \min_i(x_i)$, $x_{\max} = \max_i(x_i)$;
2. Інтервал $[x_{\min}, x_{\max}]$ розбивають на s однакових під-інтервалів $[a_i, b_i)$, $i = \overline{1, s}$. Зазвичай $5 \leq s \leq 30$. Зазвичай $s = 1 + \lceil \log_2 n \rceil$ або $s = \lceil 10 \log_{10}(n) \rceil$;
3. $x_i^* = \frac{a_i + b_i}{2}$ – центральна точка.

v_i – кількість вимірів з вибірки що належать інтервалу $[a_i, b_i)$.

$$\{x_1, x_2, \dots, x_n\} \mapsto \{x_i^*, v_i\}_{i=1}^s \left(\sum_{i=1}^s v_i = n \right).$$

Рекомендується $v_i \geq 5$, в разі $v_i < 5$ сусідні інтервали зливаються в один.

Зауваження! При проведенні групування даних зовсім не обов'язково брати під-інтервали однакової довжини.

$F_\xi(x) = P\{\xi < x\}$ – функція розподілу, $p_\xi(x)$ – функція щільності, $\{y_i, p_i\}_{i=1}^m$ – полігон ймовірності, якщо ξ – дискретна випадкова величина, що набуває значення y_i з ймовірністю p_i , $i = \overline{1, m}$.

Оцінка характеристик по згрупованим даним:

Емпірична (вибіркова) функція розподілу $\hat{F}_\xi(x)$ буде $\hat{F}_\xi(x) = \frac{1}{n} \sum_{i: b_i \leq x} v_i$.

Емпірична (вибіркова) функція щільності $\hat{p}_\xi(x)$ буде $\hat{p}_\xi(x) = \frac{v_{i(x)}}{n(b_{i(x)} - a_{i(x)})}$, де $i(x)$ – номер підінтервалу якому належить x .

2 Моделювання змінних

Потреба в генерації спостережень над випадковими величинами із заданими функціями розподілу.

Зазвичай $\xi = g(\xi_1, \xi_2, \dots, \xi_q)$, де $\xi_1, \xi_2, \dots, \xi_q$ – найпростіші випадкові величини, як правило вони рівномірно розподілені на відрізку $[0, 1)$.

Датчик (генератор) випадкових чисел – спеціальний пристрій, який після запиту на виході дозволяє отримати реалізацію випадкової величини із заданим законом розподілу.

Класи датчиків (генераторів) випадкових чисел:

- **табличні** – таблиця, заповнена реалізаціями випадковою величини із заданим законом розподілу, зазвичай досить високої якості, але вони маю обмежений об'єм. Кількість вибірок невелика.
- **фізичні** – деякий електронний пристрій на виході якого отримують необхідну реалізацію вибірки довільного об'єму, але кожна вибірка унікальна і неповторна.
- **програмні** – програма, що формує потрібну реалізацію. Базуються на використанні рекурентних формул з деякою глибиною пам'яті: задаючи однакові початкові значення можна отримати однакові вибірки. Генератор періодичний, отримані числа “псевдовипадкові”.

3 Програмні датчики

Генератор випадкової величини з $F(x) = U([0, 1))$.

Лінійна змішана формула:

$$\begin{cases} x_i = \frac{\tilde{x}_i}{M} \\ \tilde{x}_i = \left(a_0 + \sum_{j=1}^{\ell} a_j \tilde{x}_{i-j} \right) \bmod M, i = 1, 2, \dots \end{cases}$$

$$\ell \geq 1, a_j \geq 0 \ (j = \overline{1, \ell}), M > 0, \ell, a_j \ (j = \overline{0, \ell}), M \in \mathbb{Z}^+, 0 \leq \tilde{x}_{\ell-j} \leq M-1, j = \overline{1, \ell}.$$

Мультиплікативний конгруентний метод: Лінійна змішана формула ($\ell = 1, a_0 = 0$).

$$\begin{cases} x_i = \frac{\tilde{x}_i}{M} \\ \tilde{x}_i = (a_1 \tilde{x}_{i-1}) \bmod M, i = 1, 2, \dots \end{cases}$$

$$0 \leq \tilde{x}_0 \leq M-1, \{\tilde{x}_i\}_{i \geq 0} \in \{0, 1, \dots, M-1\}.$$

Послідовність $\{\tilde{x}_i\}_{i \geq 0}$ періодична. T_{\max} – максимальний період. $T_{\max} \leq M$. Вигідно взяти M якомога більшим, ближчим до максимального цілого числа, наприклад найбільше просте число, що менше $\max \text{int}$.

Мультиплікативний конгруентний метод не дозволяє досягти максимального теоретично можливого періоду рівного M .

$$\lambda(M) = \begin{cases} 1, & M = 2 \\ 2, & M = 4 \\ p^{q-1}(p-1), & M = p^q, p > 2, p \in \mathbb{P}, q \geq 1 \\ \text{lcm}(\lambda(p_1^{q_1}), \lambda(p_2^{q_2}), \dots, \lambda(p_k^{q_k})), & M = p_1^{q_1} \cdot p_2^{q_2} \cdot \dots \cdot p_k^{q_k}. \end{cases}$$

Теорема 3.1. Максимальний період послідовності $\{\tilde{x}_i\}_{i \geq 0}$ мультиплікативного конгруентного методу $T_{\max} = \lambda(M)$. T_{\max} досягається при:

1. $\gcd(\tilde{x}_0, M) = 1$;
2. $a_1^{\lambda(M)} \bmod M = 1$, a_1 є первісним коренем за модулем M .

Зауваження. Якщо покласти M рівним простому числу, то $T_{\max} = M-1$. В залежності від розрядності комп'ютера найбільшим простим числом буде:

розрядність	16	32	64
max просте число	$2^{16} - 15$	$2^{32} - 5$	$2^{64} - 59$

Змішаний конгруентний метод:

Лінійна змішана формула ($\ell = 1$, $a_0 > 0$).

$$\begin{cases} x_i = \frac{\tilde{x}_i}{M} \\ \tilde{x}_i = (a_0 + a_1 \tilde{x}_{i-1}) \bmod M, i = 1, 2, \dots \end{cases}$$

Теорема 3.2. Для отримання послідовності $\{\tilde{x}_i\}_{i \geq 0}$ яка досягає свого max періоду $T_{\max} = M$, необхідно:

- $\gcd(a_0, M) = 1$;
- $(a_1 - 1) \bmod p = 0$ для всіх $p|M$, $p \in \mathbb{P}$;
- $(a_1 - 1) \bmod 4 = 0$, якщо $4|M$.

Зауваження! Вибір параметрів змішаного конгруентного методу не є гарантією високої якості вибірки. Наприклад $a_0 = a_1 = 1$.

Квадратичний конгруентний метод:

$$\begin{cases} x_i = \frac{\tilde{x}_i}{M} \\ \tilde{x}_i = (a_0 + a_1 \tilde{x}_{i-1} + a_2 \tilde{x}_{i-1}^2) \bmod M, i = 1, 2, \dots \end{cases}$$

$$T_{\max} = M.$$

Ускладнення лінійної змішаної формули:

$$\begin{cases} x_i = \frac{\tilde{x}_i}{M} \\ \tilde{x}_i = g(\tilde{x}_{i-1}, \tilde{x}_{i-2}, \dots, \tilde{x}_{i-\ell}) \bmod M, i = 1, 2, \dots \end{cases}$$

$$T_{\max} = M.$$

4 Моделювання дискретних випадкових величин

Скористаємося побудованими датчиками для $U([0, 1))$: ξ – дискретна випадкова величина $p_i = P\{\xi = y_i\}$, $i = \overline{1, m}$. $\sum_{i=1}^m p_i = 1$, отже інтервал $[0, 1)$ можна розбити на m під-інтервалів

$$\delta_1 = [0, p_1), \Delta_2 = [p_1, p_1 + p_2), \dots, \Delta_i = \left[\sum_{j=1}^{i-1} p_j, \sum_{j=1}^i p_j \right), \dots, \Delta_m = \left[\sum_{j=1}^{m-1} p_j, 1 \right)$$

Довжина інтервалу Δ_i дорівнює p_i ($i = \overline{1, m}$). Отримуємо від датчика $U([0, 1))$ значення X . Якщо $x \in \Delta_i$, то ξ прийняла значення y_i .

Генерування рівномірного розподілу на $[1, m]$: $p_i = P\{\xi = i\} = \frac{1}{m}$, $i = \overline{1, m}$. x – значення датчика $U([0, 1))$, тоді ξ набуває значення $\lfloor 1 + mx \rfloor$.

5 Моделювання неперервних випадкових величин

Необхідно моделювати неперервну випадкову величину ξ із функцією розподілу $F(z)$.

Розглянемо випадок коли $F(z)$ – строго монотонна функція. Тоді у ролі реалізації ξ може виступити $F^{-1}(x)$, де x – значення датчику $U([0, 1))$, а $F^{-1}(x)$ – обернена функція розподілу до $F(z)$. Нехай η – випадкова величина, $F(\eta) = U([0, 1))$. Тоді $F^{-1}(\eta)$:

$$P\{F^{-1}(\eta) < x\} = P\{\eta < F(x)\} = F(x)$$

Приклад. ξ – випадкова величина, що має показниковий закон розподілу з параметром $\lambda > 0$.

$$F(z) = \begin{cases} 1 - e^{-\lambda z}, & z \geq 0, \\ 0, & z < 0. \end{cases}$$

$F^{-1}(y) = -\frac{\ln(1-y)}{\lambda}$, тобто $-\frac{\ln(1-\eta)}{\lambda}$ має потрібний показниковий розподіл, де η – випадкова величина з розподілом $U([0, 1))$. Оскільки $1 - \eta$ також має розподіл $U([0, 1))$, то величина $-\frac{\ln \eta}{\lambda}$, $\lambda > 0$ також має показниковий розподіл. Підсумовуючи, в ролі реалізації ξ може виступити $-\frac{\ln x}{\lambda}$, де x – випадкова величина з розподілом $U([0, 1))$.

6 Моделювання нормального розподілу з параметрами m та σ^2

Теорема 6.1. Нехай η_1 та η_2 мають розподілу $U([0, 1])$. Тоді випадкові величини

$$\begin{aligned}\xi_1 &= \sin(2\pi\eta_1)\sqrt{-2\ln\eta_2}, \\ \xi_2 &= \cos(2\pi\eta_1)\sqrt{-2\ln\eta_2},\end{aligned}$$

незалежні, нормально розподілені з параметрами 0 та 1.

Позначимо x_1, x_2 – незалежні спостереження над рівномірно розподіленою величиною на інтервалі $[0, 1)$. Тоді згідно теореми можна стверджувати, що значення

$$m + \sigma \sin(2\pi x_1)\sqrt{-2\ln x_2}, m + \sigma \cos(2\pi x_1)\sqrt{-2\ln x_2}$$

є спостереженнями на незалежними нормально розподіленими випадковими величинами з параметрами m та σ^2 .

У разі необхідності моделювання випадкових величин рівномірного розподілу на інтервалі $[a, b)$ достатньо взяти вихід x з датчика $U([0, 1])$ та отримати реалізацію випадкової величини як $a + (b - a)x$.

7 Попередня обробка даних

Попередня обробка даних проводить роботу пов'язану з отриманням попередніх висновків про змінні, які спостерігаються.

Квантилі та процентні точки розподілу.

Нехай $F(x)$ – функція розподілу випадкової величини ξ .

Квантилем рівня q розподілу (q -квантилем розподілу) неперервної випадкової величини ξ називається таке значення u_q , що визначається з рівняння:

$$F(u_q) = P\{\xi < u_q\} = q, \quad (0 < q < 1)$$

Квантилем рівня q розподілу (q -квантилем розподілу) дискретної випадкової величини ξ називається довільне значення u_q з інтервалу $[y_{i(q)}, y_{i(q)+1}]$, для границь якого справедливо

$$F(y_{i(q)}) < q, F(y_{i(q)+1}) \geq q, \quad (0 < q < 1)$$

де $\{y_i\}$ – значення, які приймає дискретна випадкова величина ξ .

Емпіричний (вибірковий) квантиль рівня q розподілу випадкової величини визначається як квантиль рівня q відповідного емпіричного (вибіркового) розподілу.

Q -процентною точкою розподілу неперервної випадкової величини ξ називається таке значення w_Q , яке є розв'язком рівняння:

$$1 - F(w_Q) = P\{\xi \geq w_Q\} = Q/100, \quad 0 < Q < 100.$$

Q -процентною точкою розподілу дискретної випадкової величини ξ називається довільне значення w_Q з інтервалу $(y_{i(Q)}, y_{i(Q)+1}]$, для границь якого справедливо

$$1 - F(y_{i(Q)}) = P\{\xi \geq y_{i(Q)}\} > \frac{Q}{100}$$

$$1 - F(y_{i(Q)+1}) = P\{\xi \geq y_{i(Q)+1}\} \leq \frac{Q}{100}$$

Ці два поняття взаємно доповнюють одне одного. Для неперервного випадку для певних розподілів справджується наступне