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 степенем 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: $x_1, x_2, \ldots, x_n$.

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 $x_{\min} = \min\limits_i(x_i)$, $x_{\max} = \max\limits_i(x_i)$;
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 $s = 1 + [\log_2 n]$ або $s = [10 \log_{10}(n)]$;
3. $x_i^* = \dfrac{a_i + b_I}{2}$ – центральна точка.

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

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

   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 $\hat{F}_\xi(x) = \frac{1}{n} \sum\limits_{i:b_i \le x} v_i$.
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:

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

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

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

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

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

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.

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

Теорема 3.1. Максимальний період послідовність $\{\tilde{x}_i\}_{i\geq 0}$ 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 ($\ell = 1$, $a_0 > 0$).

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

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:

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

$T_{\max} = M$.

Ускладнення лінійної змішаної формули:

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

$T_{\max} = M$.

4 Моделювання дискретних випадкових величин

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

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

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 $[1,m]$: $p_i = P\{\xi = i\} = \dfrac{1}{m}$, $i = \overline{1,m}$. 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 $\lambda > 0$.

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

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

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

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 $u_q = W_{100(1-q)}$, $w_Q = u_{1-Q/100}$.

Медіана – це квантиль рівня 0.5, тобто $u_{0.5}$.

Нижній та верхній квартилі визначаються як $u_{0.25}$ та $u_{0.75}$ відповідно.

Децилі – це квантилі $\{u_{i/10}\}_{i=1}^{9}$.

Процентилі задаються наступним чином $\{u_{i/100}\}_{i=1}^{99}$.

Інтерквантильна широта рівня q $(0 < q < 1/2)$ – це величина яка обчислюється за формулою $(u_{1-q} - u_q)$.

Інтерквартильна широта це інтерквантильна широта рівня $1/4$, а саме $(u_{0.75} - u_{0.25})$.

Ймовірнісне відхилення d_ξ визначається як половина інтерквартильної широти, тобто $d_\xi = (u_{0.75} - u_{0.25})/2$.

Інтердецильна широта – це інтерквантильна широта рівня $1/10$, а саме $(u_{0.9} - u_{0.1})$.

Інтерсекстильна широта – це інтерквантильна широта рівня $1/6$, тобто $(u_{5/6} - u_{1/6})$.

8 Характеристики положення центра значень змінної

Нехай обробляється вибірка об'єму n спостережень $x_1, x_2, \ldots, x_n$ над скалярною змінною ξ.

Математичне сподівання (теоретичне середнє) обчислюється за відомою формулою для $M\xi$. Відповідне вибіркове значення має вигляд $\bar{x}(n) = \dfrac{1}{n}\sum_{i=1}^{n} x_i$.

Середнє геометричне G_ξ визначається для випадкових величин, які з ймовірністю 1 додатні. Згідно визначення $G_\xi = \exp\{M(\ln(\xi))\}$.

Оцінка величини має має наступний вигляд $\hat{G}_\xi = \sqrt[n]{\prod_{i=1}^{n} x_i}$.

Середнє гармонічне H_ξ вводиться для випадкових величин ξ з позитивними значеннями наступним чином: $H_\xi = 1/M(1/\xi)$. Емпіричне значення середнього гармонічного має вигляд

$$\hat{H}_\xi = \frac{n}{\sum_{i=1}^{n} \dfrac{1}{x_i}}$$

Мода $x_{\mathbf{mod}}$ для неперервної випадкової величини ξ вводиться як точка максимуму функції щільності ξ. Для дискретного розподілу $\{y_i, p_i\}_{i\geq 0}$ визначається як довільне значення y_k яке приймається з

найбільшою ймовірністю. Мода може бути не єдиною. Характеристика застосовується до унімодальних розподілів. Мода визначається для неперервної випадкової величини за її гістограмою щільності, а у дискретної – за полігоном частот відповідно.

Медіана $x_{\mathbf{med}}$ – це квантиль рівня 0.5, її оцінка $\hat{x}_{\text{med}}(n)$ обчислюється на основі емпіричної функції розподілу.

9 Характеристики розсіювання значень змінної

Маємо вибірку об'єму n спостережень $x_1, x_2, \ldots, x_n$ над скалярною змінною ξ.

Дисперсія σ^2 підраховується згідно формули $\sigma^2 = D\xi = M(\xi - M\xi)^2$. Незміщена оцінка σ^2 має вигляд $s^2(n) = \dfrac{1}{n-1}\sum\limits_{i=1}^{n}(x_i - \bar{x}(n))^2$. Деколи більш корисно представляти $s^2(n) = \dfrac{1}{n-1}\left(\sum\limits_{i=1}^{n} x_i^2 - n\bar{x^2}(n)\right)$.

Стандартне (середнє квадратичне) відхилення σ є коренем з дисперсії $\sigma = \sqrt{D\xi}$. $s(n) = \sqrt{\dfrac{1}{n-1}\left(\sum\limits_{i=1}^{n} x_i^2 - n\bar{x^2}(n)\right)}$.

Зауваження. Стандартне відхилення для деякої оцінки називають її *стандартною похибкою*. У випадку обробки нормальної вибірки $N(m, \sigma^2)$ об'єму n, стандартна похибка e_ξ оцінки її математичного сподівання $\bar{x}(n)$ визначається як $e_\xi = \sigma/\sqrt{n}$, а відповідне вибіркове значення як $\hat{e}_x i = s(n)/\sqrt{n}$.

Коефіцієнт варіації V_ξ визначається для випадкових величин у яких $M\xi \neq 0$ і підраховується як $V_\xi = \sqrt{D\xi}/M\xi \cdot 100\%$. Вибіркове значення має вигляд

$$\hat{V}_\xi(n) = \frac{s(n)}{\bar{x}(n)} \cdot 100\% = \frac{\sqrt{\dfrac{1}{n-1}\left(\sum\limits_{i=1}^{n} x_i^2 - n\bar{x^2}(n)\right)}}{\dfrac{1}{n}\sum\limits_{i=1}^{n} x_i} \cdot 100\%$$

Ймовірнісне відхилення d_ξ є половиною інтерквартильної широти, тобто $d_\xi = (u_{0.75} - u_{0.25})/2$. Емпіричне значення має вигляд $\hat{d}_\xi = (\hat{u}_{0.75} - \hat{u}_{0.25})/2$.

Розмах (широта) вибірки $x_1, x_2, \ldots, x_n$ спостережень над ξ визначається таким чином: $\hat{R}_\xi(n) = x_{\max}(n) - x_{\min}(n)$, де $x_{\max}(n)$, $x_{\min}(n)$ – найбільший та найменший значення в вибірці відповідно.

Інтервал концентрації розподілу випадкової величини ξ має такий вигляд: $(M\xi - 3\sqrt{D\xi}, M\xi + 3\sqrt{D\xi})$. Вибірковий аналог має вигляд $(\bar{x}(n) - 3s(n), \bar{x}(n) + 3s(n))$.

10 Аналіз скошеності та гостроверхості розподілу

Маємо вибірку об'єму n спостережень $x_1, x_2, \ldots, x_n$ випадкової величини ξ.

Очевидно, що якщо розподіл ξ симетричний відносно $M\xi$ то всі його непарні центральні моменти $M(\xi - M\xi)^{2k-1}$ дорівнюють нулю, якщо вони існують. В основі **коефіцієнта асиметрії** – характе-

ристики скошеності розподілу – лежить третій центральний момент

$$\beta_1 = \frac{M(\xi - M\xi)^3)}{(M(\xi - M\xi))^{3/2}}, \quad D\xi > 0$$

Вибіркове значення:

$$\hat{\beta}_1(n) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}(n))^3}{\left(\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x^2}(n)\right)\right)^{3/2}}$$

Для симетричних відносно $M\xi$ розподілів $\beta_1 = 0$. Якщо $\beta_1 < 0$ то розподіл скошений праворуч, якщо $\beta_1 > 0$ то розподіл скошений ліворуч.

При дослідженні загальної поведінки розподілу в околі моди як характеристики гостроверхості використовують **коефіцієнт ексцесу**, який базується на четвертому центральному моменті і має вигляд

$$\beta_2 = \frac{M(\xi - M\xi)^4}{(M(\xi - M\xi)^2)^2} - 3, \quad D\xi > 0$$

"-3" застосовується для того. щоб коефіцієнт ексцесу нормального розподілу був рівний 0.

Емпіричне значення

$$\hat{\beta}_2(n) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}(n))^4}{\left(\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x^2}(n)\right)\right)^{2}} - 3$$

Якщо $\beta_2 > 0$ то розподіл більш гостроверхний ніж нормальний, а якщо $\beta_2 < 0$ – більш плоский.